24e Conférence sur le Traitement Automatique des Langues Naturelles (TALN)

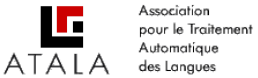
Orléans, France – 26-30 juin 2017

https://taln2017.cnrs.fr

Actes de TALN 2017, volume 2 : articles courts

Iris Eshkol, Jean-Yves Antoine (Eds.)

Sponsors :


ATALA
Association
pour le Traitement
Automatique
des Langues

Ortolang

cnrs
dépasser les frontières

eptica
MEANINGFUL CONVERSATIONS

acatus
informatique

aktan

proxem

inbenta

Orléans
Mairie

ORLÉANS
MÉTROPOLE

LOIRE&ORLEANS
ECO

FRENCH TECH
LOIRE
VALLEY

UNIVERSITE D'ORLEANS

Laboratoire
Ligérien de
Linguistique

LABORATOIRE
D'INFORMATIQUE
LIFO
FONDAMENTALE
D'ORLEANS

LI
Laboratoire d'Informatique
EA 2101

Liberté · Égalité · Fraternité
RÉPUBLIQUE FRANÇAISE
MINISTÈRE
DE LA CULTURE

Préface

Bienvenue à TALN-RECITAL-SITAL 2017

La 24e édition de la conférence TALN et 19 e édition du RECITAL se déroulent cette année à Orléans, en plein cœur de la région Centre Val de Loire, célèbre pour ses châteaux, inscrite au patrimoine mondial de l'Unesco et située à 125 km au sud de Paris. Nous accueillons des participants sur le campus boisé d'Orléans du lundi 26 au vendredi 30 juin 2017. La conférence est organisée par trois laboratoires de référence dans le domaine du TAL, de l'informatique et de la linguistique de corpus : LLL (Laboratoire Ligérien de Linguistique), LIFO (Laboratoire d'Informatique Fondamentale d'Orléans) et LI (Laboratoire Informatique) de Tours. Après les Sables d'Olonne, Marseille, Paris, la conférence arrive dans une région mondialement réputée pour sa contribution au rayonnement de la langue française. En marge de la conférence, des activités culturelles, sportives et gastronomiques seront proposées pour découvrir la région du Val de Loire tout au long de la semaine, en après-midi ainsi qu'en soirée.

La conférence accueille 150 participants. Le premier jour, lundi, est consacré à quatre ateliers thématiques : « DEFT Fouille de texte » consacré cette année à l'analyse d'opinion et langage figuratif dans des messages postés sur Twitter ; la seconde édition du « HackaTAL » consacrée au résumé automatique de description de produits à partir de leurs commentaires et la prédiction automatique de la brevetabilité de termes liés à des technologies selon leur historique ; l'atelier « Les corpus annotés du français : ressources disponibles et exploitation en TAL » proposé pour la première fois à TALN et l'atelier « DiLiTAL – Diversité linguistique et TAL » dédié au traitement automatique des langues peu dotées. Nous remercions les organisateurs de ces ateliers pour leurs propositions, leur animation et leur organisation de ces événements. Les trois jours qui suivent associent dans de mêmes séances TALN et la conférence jeunes chercheurs RECITAL. Le dernier jour, vendredi 30 juin, est dédié au salon de l'innovation (SITAL) destiné à accueillir les entreprises et les chercheurs dans ce domaine du numérique de plus en plus émergent avec le développement du big data sur les données textuelles, ceci afin d'échanger leurs idées sur les développements actuels et futurs du domaine. Dans le cadre du SITAL, trois tables rondes « TAL et Humanités numériques », « Traitement automatique de la langue biomédicale » et « TAL dans l'expérience utilisateur : analyses et outils » sont proposées. Nous remercions les animateurs des tables rondes pour leur aide dans l'organisation de ces événements. Les tables rondes sont suivies par la session des posters et de démonstrations qui permettent aux participants de discuter plus longuement sur leurs travaux et de faire connaissance avec de nouveaux logiciels développés. La conférence se termine par une grande table ronde intitulée

« TAL aujourd'hui et demain : nouvelles méthodes, nouveaux usages, nouvelles applications ». Tout au long de la conférence, les étudiants ont la possibilité de s'inscrire et de participer aux déjeuners avec des experts d'un domaine.

Nous accueillons deux conférenciers invités. Tout d'abord Shuly Wintner (Université de Haifa, Israel) spécialiste en grammaires formelles, morphologie, traduction automatique et acquisition du langage, et Laura Kallmeyer (Université de Duesseldorf, Allemagne) qui travaille depuis des années sur l'analyse syntaxique, les grammaires d'arbres adjoints et l'interface syntaxe-sémantique. Nous les remercions d'avoir accepté notre invitation.

TALN2017 et RECITAL2017 ont permis aux chercheurs de présenter leurs travaux sous forme de communications orales, de posters et de démonstrations. 122 papiers au total ont été soumis aux deux conférences, 71 papiers (58%) ont été acceptés. Sur 28 articles longs soumis, 14 articles (50%) ont été acceptés. Sur 60 articles courts proposés, 30 articles (50%) ont été acceptés dont 8 pour la présentation orale et 22 en poster. Enfin, 17 papiers ont été soumis pour la session de démonstrations dont 14 ont été acceptés. En ce qui concerne la conférence RECITAL, 17 articles ont été soumis, 13 articles ont été acceptés dont 5 pour la présentation orale et 8 en poster.

La procédure de relecture et de la prise de décision finale est une procédure complexe, c'est pourquoi nous remercions le comité de relecture de TALN et de RECITAL ainsi que le comité de programme pour le temps consacré, pour leur patience et leur bonne volonté. Nous tenons à remercier le comité permanent de la conférence (CPERM) et son président pour l'aide dans l'organisation de l'événement et dans la prise de décisions pour certaines questions. Nous remercions le comité d'organisation constitué de chercheurs de trois laboratoires LLL, LIFO et LI, d'agents CNRS, de doctorants et d'étudiants de master et de licence. C'est une équipe formidable qui a mis tout en œuvre pour pouvoir permettre le meilleur accueil et le meilleur déroulement de la conférence. Nous tenons à remercier nos partenaires et nos sponsors : ATALA, DGLFLF, Orléans Métropole, ORTOLANG, les entreprises Inbenta, Aktan, Acatus, Proxem, sans lesquels cette édition de TALN, de RECITAL et le salon SITAL n'aurait pas pu avoir lieu. Un grand merci à l'Université d'Orléans et plus particulièrement à la Faculté LLSH d'avoir prêté les locaux et d'avoir permis l'organisation de cet événement dans les meilleures conditions.

Iris Eshkol-Taravella et Jean-Yves Antoine
Co-Présidents de TALN2017

Comités

Présidente du comité d'organisation de TALN 2017

Iris Eshkol-Taravella (LLL-UMR 7270, Université d'Orléans)

Vice-président du comité d'organisation de TALN 2017

Jean-Yves Antoine (LI-EA 6300, Université de Tours)

Président du comité d'organisation de RECITAL 2017

Yannick Parmentier (LIFO-EA 4022, Université d'Orléans)

Vice-présidente du comité d'organisation de RECITAL 2017

Hélène Flamein (LLL-UMR 7270, Université d'Orléans)

Présidente du salon d'innovation de TALN 2017

Gabrielle Bosshard (Aktan, Orléans)

Vice-présidente du salon d'innovation de TALN 2017

Sandra Cestic (LLL-UMR 7270, Acatus Informatique, Orléans)

Comité d'Organisation

Ahmed Abid (LI, Université de Tours)
Catherine Aléonard (LLL, Université d'Orléans)
Flora Badin (LLL, Université d'Orléans)
Gabriel Bergounioux (LLL, Université d'Orléans)
Sylvie Billot (LIFO, Université d'Orléans)
Marwa Boulakbech (LI, Université de Tours)
Caroline Cance (LLL, Université d'Orléans)
Guillaume Cleuziou (LIFO, Université d'Orléans)

Hyun Jung Kang (LLL, Université d'Orléans)
Anaïs Lefeuvre-Halftermeyer (LIFO, Université d'Orléans)
Denis Maurel(LI, Université de Tours)
Agata Savary (LI, Université de Tours)
Emmanuel Schang (LLL, Université d'Orléans)

Étudiants du Master Linguistique, spécialité LASTIC de l'Université d'Orléans

Hélène Couderc
Laetitia Delay
Siqi Fan
Lorraine Gaspard
Mélanie Lefeuvre
Sara Masaud
Stéphanie Nogueira
Camille Pertin
Cathy Querineau
Céline Vaschalde
Jidong Xie

Comité de Programme de TALN

Jean-Yves Antoine (Université de Tours, LI)
Delphine Bernhard (LiLPa - Université de Strasbourg)
Laurent Besacier (Université de Grenoble, LIG)
Nathalie Camelin (LIUM - Université du Maine)
Benoît Crabbé (Université Paris Diderot, LLF)
Iris Eshkol-Taravella (Université d'Orléans, LLL)
Olivier Ferret (CEA LIST)
Claire Gardent (CNRS, LORIA)
Thierry Hamon (Université Paris 13, LIMSI)
Philippe Langlais (Université de Montréal, RALI)
Emmanuel Morin (Université de Nantes, LINA)
Philippe Muller (CNRS, CLLE)
Adeline Nazarenko (Université Paris-Nord, LIPN)

Comité de Relecture de TALN 2017

Stergos Afantenos, CNRS/Université Paul Sabatier
Salah Ait-Mokhtar, Xerox Research Centre Europe
Maxime Amblard, Université de Lorraine
Jean-Yves Antoine, LI, Université de Tours
Delphine Batistelli, MoDyCo, Université paris Ouest Nanterre La Défense
Frédéric Bechet, Aix Marseille Université - LIF

Delphine Bernhard, LiLPa, Université de Strasbourg
Romaric Besançon,CEA LIST
Philippe Blache, CNRS & Université de Provence
Hervé Blanchon, Laboratoire d'Informatique de Grenoble - Equipe GETALP
Florian Boudin, Université de Nantes
Annelies Braffort, LIMSI-CNRS
Nathalie Camelin, LIUM, Université du Maine
Thierry Charnois, LIPN CNRS University of PARIS 13
Guillaume Cleuziou, LIFO, Université d'Orléans
Benoit Crabbé,Paris 7 et INRIA
Béatrice Daille, Laboratoire d'Informatique Nantes Atlantique (LINA)
Marco Dinarelli, Lattice-CNRS
Iris Eshkol-Taravella, University of Orléans
Yannick Estève, LIUM - Université du Maine
Cécile Fabre,Université Toulouse 2
Karën Fort, Paris 4
Thomas François, Université Catholique de Louvain
Nathalie Friburger, LI, Université de Tours
Eric Gaussier, LIG-UJF
Natalia Grabar, STL CNRS Université Lille 3
Lamia Hadrich, MIRACL Laboratory
Nicolas Hernandez, Université de Nantes - LINA CNRS UMR 6241
Stéphane Huet, LIA - Université d'Avignon
Sylvain Kahane, Université Paris Ouest Nanterre & CNRS
Olivier Kraif, Université Stendhal Grenoble 3
Mathieu Lafourcade, LIRMM
Guy Lapalme, RALI-DIRO, Université de Montréal
Joseph Le Roux, Laboratoire d'Informatique de Paris Nord
Jean-Marc Lecarpentier, GREYC
Anaïs Lefeuvre-Halftermeyer LIFO, Université d'Orléans
Anne-Laure Ligozat, LIMSI-CNRS
Denis Maurel, LI, Université de Tours
Richard Moot, CNRS (LaBRI) & Bordeaux University
Véronique Moriceau, LIMSI-CNRS
Philippe Muller, IRIT, Toulouse University
Aurélie Névéol, CNRS
Jian-Yun Nie, Université de Montréal
Yannick Parmentier, LIFO, Universite d ?Orleans
Thierry Poibeau, LaTTiCe-CNRS
Andrei Popescu-Belis, IDIAP Research Institute
Jean-Philippe Prost, LIRMM, Université Montpellier 2
Solen Quiniou, LINA - Université de Nantes

Christian Raymond, UEB/INRIA/IRISA/INSA
Christian Retoré, Université de Montpellier ; LIRMM-CNRS
Mathieu Roche, Cirad, TETIS
Didier Schwab, Univ. Grenoble Alpes
Michel Simard, National Research Council Canada (NRC)
Kamel Smaili, LORIA
Xavier Tannier, LIMSI, CNRS, Univ. Paris-Sud, Université Paris-Saclay
Isabelle Tellier, PARIS 3, Lattice
Juan-Manuel Torres-Moreno, Laboratoire Informatique d'Avignon / UAPV
Christel Vrain, LIFO, university of Orléans
Eric Wehrli, University of Geneva
Guillaume Wisniewski, LIMSI-UPS
François Yvon, LIMSI/CNRS et Université Paris-Sud
Pierre Zweigenbaum, LIMSI-CNRS

Table des matières

Session « oral »

Annotation d'expressions polylexicales verbales en français
Marie Candito, Mathieu Constant, Carlos Ramisch, Agata Savary, Yannick Parmentier, Caroline Pasquer and Jean-Yves Antoine 1

Évaluation de mesures d'association pour les bigrammes et les trigrammes au moyen du test exact de Fisher
Yves Bestgen 10

Réseaux neuronaux profonds pour l'étiquetage de séquences
Yoann Dupont, Marco Dinarelli, Isabelle Tellier 19

Schémas Winograd en français : une étude statistique et comportementale
Pascal Amsili, Olga Seminck 28

Critères numériques dans les essais cliniques : annotation, détection et normalisation
Natalia Grabar, Vincent Claveau 36

Analyse automatique FrameNet : une étude sur un corpus français de textes encyclopédiques
Gabriel Marzinotto, Géraldine Damnati, Frédéric Béchet 44

Détection de coréférences de bout en bout en français
Godbert Elisabeth, Benoit Favre 52

Une approche hybride pour la construction de lexiques bilingues d'expressions multi-mots à partir de corpus parallèles
Nasredine Semmar, Morgane Marchand 60

Session « poster »

Une interprétation probabiliste des informations de factivité
Timothée Bernard 69

Une approche universelle pour l'abstraction automatique d'alternances morphophonologiques
Sacha Beniamine 77

Détection automatique de métaphores dans des textes de Géographie : une étude prospective
Max Beligné, Aleksandra Campar, Jean-Hugues Chauchat, Melanie Lefeuvre, Isabelle Lefort, Sabine Loudcher, Julien Velcin 86

Describing derivational polysemy with XMG
Marios Andreou, Simon Petitjean 94

Changement stylistique de phrases par apprentissage faiblement supervisé
Damien Sileo, Camille Pradel, Philippe Muller, Tim Van de Cruys 102

Amélioration de la similarité sémantique vectorielle par méthodes non-supervisées
El Moatez Billah Nagoudi, Jérémy Ferrero, Didier Schwab 110

Typologies pour l'annotation de textes non standards en français
Louise Tarrade, Cédric Lopez, Rachel Panckhurst, Geroges Antoniadis 118

Simbow : une mesure de similarité sémantique entre textes
Delphine Charlet, Géraldine Damnati 126

Adaptation au domaine pour l'analyse morpho-syntaxique
Éléonor Bartenlian, Margot Lacour, Matthieu Labeau, Alexandre Allauzen, Guillaume Wisniewski, François Yvon 134

Représentation vectorielle de sens pour la désambiguïsation lexicale à base de connaissances
Loïc Vial, Benjamin Lecouteux, Didier Schwab 142

Parcourir, reconnaître et réfléchir. Combinaison de méthodes légères pour l'extraction de relations sémantiques
Mathieu Lafourcade, Nathalie Le Brun 150

Si les souris étaient des reptiles, alors les reptiles pourraient être des mammifères ou Comment détecter les anomalies dans le réseau JDM ?

Alain Joubert, Mathieu Lafourcade, Nathalie Le Brun 158

Vers l'annotation par le jeu de corpus (plus) complexes : le cas de la langue de spécialité
Karën Fort, Bruno Guillaume, Nicolas Lefebvre, Laura Ramírez, Mathilde Regnault, Mary Collins, Oksana Gavrilova, Tanti Kristanti 165

Détection des mots non-standards dans les tweets avec des réseaux de neurones
Tian Tian, Isabelle Tellier, Marco Dinarelli, Pedro Cardoso 174

MAR-REL : une base de marqueurs de relations conceptuelles pour la détection de Contextes Riches en Connaissances
Luce Lefeuvre, Anne Condamines 183

Apprentissage en ligne interactif d'un générateur en langage naturel neuronal pour le dialogue homme-machine
Matthieu Riou, Bassam Jabaian, Stéphane Huet, Fabrice Lefèvre 192

Apports des analyses syntaxiques pour la détection automatique de mentions dans un corpus de français oral
Loïc Grobol, Isabelle Tellier, Éric de La Clergerie, Marco Dinarelli, Frédéric Landragin 200

Un nouveau besoin dans l'industrie : une aide au « rédacteur traduisant »
Claire Lemaire 209

Adaptation incrémentale de modèles de traduction neuronaux
Christophe Servan, Josep Crego, Jean Senellart 218

Détection de concepts et granularité de l'annotation
Pierre Zweigenbaum, Thomas Lavergne 226

Tri Automatique de la Littérature pour les Revues Systématiques
Christopher Norman, Mariska Leeflang, Pierre Zweigenbaum, Aurélie Névéol 234

Une approche linguistique pour la détection des dialectes arabes
Houda Saâdane, Damien Nouvel, Hosni Seffih, Christian Fluhr 242

Annotation d'expressions polylexicales verbales en français

Marie Candito[1] Mathieu Constant[2] Carlos Ramisch[3] Agata Savary[4]
Yannick Parmentier[5] Caroline Pasquer[4] Jean-Yves Antoine[4]

(1) Université Paris Diderot, LLF, Paris, France
(2) Université de Lorraine, ATILF, CNRS, Nancy, France
(3) Aix Marseille Université, CNRS, LIF, Marseille, France
(4) Université François Rabelais Tours, LI, Blois, France
(5) Université d'Orléans, LIFO, Orléans, France

marie.candito@linguist.univ-paris-diderot.fr, Mathieu.Constant@univ-lorraine.fr, carlos.ramisch@lif.univ-mrs.fr, agata.savary@univ-tours.fr, yannick.parmentier@univ-orleans.fr, caroline.pasquer@etu.univ-tours.fr, Jean-Yves.Antoine@univ-tours.fr

RÉSUMÉ

Nous décrivons la partie française des données produites dans le cadre de la campagne multilingue PARSEME sur l'identification d'expressions polylexicales verbales (Savary *et al.*, 2017). Les expressions couvertes pour le français sont les expressions verbales idiomatiques, les verbes intrinsèquement pronominaux et une généralisation des constructions à verbe support. Ces phénomènes ont été annotés sur le corpus French-UD (Nivre *et al.*, 2016) et le corpus Sequoia (Candito & Seddah, 2012), soit un corpus de 22 645 phrases, pour un total de 4 962 expressions annotées. On obtient un ratio d'une expression annotée tous les 100 tokens environ, avec un fort taux d'expressions discontinues (40%).

ABSTRACT

Annotation of verbal multiword expressions in French

We describe the French part of the annotated data produced for the multilingual PARSEME shared task on the identification of verbal multiword expressions (Savary *et al.*, 2017). The annotated verbal expressions for French are idioms, inherently reflexive verbs and a generalization over support verb constructions. These were annotated on the French-UD corpus (Nivre *et al.*, 2016) and the Sequoia corpus (Candito & Seddah, 2012) for a total of 22,645 sentences and 4,962 annotated expressions. On average, we obtain around 1 expression every 100 tokens with a high ratio of discontiguities (40%).

MOTS-CLÉS : Expressions polylexicales verbales, annotation, corpus.

KEYWORDS: Verbal multiword expressions, annotation, corpora.

1 Introduction

Les expressions polylexicales (ci-après EP), telles que ***pomme de terre***, ***tout à coup***, ***prendre*** *une* ***décision***, ***se comporter*** ou ***avoir l'air***[1], sont des objets linguistiques constitués d'au moins deux composants (éléments se comportant comme un mot par ailleurs, ou bien des éléments qui ne peuvent

1. Nous adoptons la convention de mettre en gras les composants *lexicalisés* d'une EP, c'est-à-dire ceux toujours réalisés par les mêmes lexèmes, par opposition aux modifieurs non requis et aux arguments requis mais pouvant être choisis dans une large classe sémantique, qui eux apparaîtront simplement en italique, comme *il* ***a l'air*** *pressé, ils n'****ont*** *pas* ***l'air*** *mécontents.*

apparaître de manière autonome, comme *fi* dans ***faire fi***) et caractérisés par un certain degré d'idiosyncrasie au niveau morphologique, syntaxique ou sémantique (Baldwin & Kim, 2010). Les EP verbales (EPV), c'est-à-dire ayant un verbe comme tête syntaxique, posent des défis particuliers pour la modélisation comme pour le traitement automatique, essentiellement du fait d'une variabilité syntaxique en général plus forte que pour d'autres EP. Elles peuvent avoir des interprétations idiomatiques ou littérales (p. ex. *le roi* ***ouvre*** *alors* ***les yeux*** *sur son geste* vs. *la victime parle et ouvre les yeux*). Des EPV de même structure syntaxique peuvent appartenir à des catégories différentes (p. ex. ***avoir l'air*** est une expression idiomatique, tandis que ***prendre la fuite*** est une construction à verbe support). Certains composants peuvent être partagés par plusieurs EPV, en particulier lors d'une coordination (p. ex. *ils* $\textbf{\textit{ont}}_{1,2}$ *le* $\textbf{\textit{pouvoir}}_1$ *et le* $\textbf{\textit{devoir}}_2$ *de voter*), d'une relativisation (p. ex. *il veut* $\textbf{\textit{avoir}}_1$ *la* $\textbf{\textit{perception}}_{1,2}$ *qu'*$\textbf{\textit{ont}}_2$ *les entreprises*) ou de l'imbrication complète d'une EP au sein d'une autre (p. ex. *ils* $\textbf{\textit{font}}_{1,2}$ $\textbf{\textit{l'}}_{1,2}\textbf{\textit{objet}}_{1,2}$ d'une $\textbf{évaluation}_2$). Enfin, selon la convention de tokenisation utilisée, une EPV peut parfois correspondre à un seul token (p. ex. ***contre-indiquer***).

Tandis que l'étude linguistique des EP françaises de tout type a une longue tradition (Gross, 1986; Mel'čuk *et al.*, 1988), l'annotation en corpus des EPV n'avait pas été réalisée à large échelle. Nous présentons le fruit d'un travail qui tente de combler cette lacune. Il s'agit d'une tâche d'annotation d'EPV effectuée dans le cadre du réseau européen PARSEME [2], et plus spécifiquement de sa campagne d'évaluation [3] d'outils de reconnaissance automatique d'EPV (Savary *et al.*, 2017). Cet article se concentre sur la production des données françaises.

2 Travaux antérieurs

Le traitement des EP en TAL est l'objet d'une vaste littérature, et plusieurs typologies à visée universelle ont été proposées (Sag *et al.*, 2001; Heid, 2008; Baldwin & Kim, 2010; Mel'čuk, 2010; Tutin *et al.*, 2015), mais aucune n'a été effectivement mise à l'épreuve d'une annotation multilingue. En effet, si l'annotation d'EP et plus précisément d'EPV a été l'objet de nombreuses initiatives (voir Savary *et al.* (2017) pour un résumé), la campagne PARSEME part du constat d'une grande hétérogénéité dans les types d'EP annotées, les critères retenus et les formats utilisés.

En outre, concernant plus spécifiquement le français, s'il existe une forte tradition de ressources lexicales incluant des EP, avec en particulier le DELAC (Courtois *et al.*, 1997), leur annotation en corpus est plus rare. Le French Treebank (Abeillé *et al.*, 2003) constitue un projet pionnier en la matière. Il comprend environ 20 000 phrases annotées pour la morphologie et la syntaxe, avec une part non négligeable d'EP annotées (Abeillé & Clément, 2006), dont des EPV. Plus précisément, les EPV retenues sont celles comportant un mot inexistant par ailleurs (***faire fi***), ou une syntaxe irrégulière avec en particulier les composés V N sans aucun déterminant possible (p. ex. ***faire face*** est annoté, mais pas *avoir peur*, cf. on a aussi *avoir une peur immense*). Si un mécanisme a bien été prévu pour gérer d'éventuelles discontinuités, il a été en pratique très peu utilisé. Les combinaisons clitiques-verbes ont été volontairement écartées, ainsi que les constructions à verbe support.

D'autres corpus arborés ne comprennent que des mots composés grammaticaux, contigus. C'est le cas des deux corpus arborés utilisés comme base de l'annotation PARSEME, Sequoia et French-UD. Plus récemment, Tutin *et al.* (2015) ont produit une typologie fine d'EP, incluant des EPV, et leur annotation sur un corpus français d'environ 45 000 tokens.

2. http://www.parseme.eu

3. http://multiword.sourceforge.net/sharedtask2017

3 La campagne PARSEME

La campagne PARSEME citée supra est le fruit de l'effort collaboratif de 18 équipes nationales (Savary *et al.*, 2017). Deux phases d'annotation pilote, menées pour 15 langues, chacune suivie d'un retour d'expérience et d'améliorations de la méthodologie, ont permis d'aboutir à :

— une typologie universelle d'EPV, validée par annotation pilote multilingue, qui laisse néanmoins une place pour des types spécifiques à des sous-ensembles de langues,
— un guide d'annotation sous forme d'arbres de décision fondés sur des tests linguistiques,
— une infrastructure consistant en une plate-forme d'annotation, des scripts d'homogénéisation, des outils d'insertion automatisée d'exemples d'EPV dans le guide d'annotation,
— des mesures et outils d'évaluation des systèmes d'identification d'EPV, les systèmes participants étant évalués sur l'identification seulement, et pas sur la catégorisation d'EPV.

Les corpus résultants [4] couvrent 18 langues et comportent au total près de 5,5 million de mots. Les 60 000 EPV annotées, ainsi que le guide d'annotation, sont diffusés sous deux versions de la licence libre Creative Commons (CC BY et CC BY-NC-SA) [5].

4 Schéma d'annotation

L'annotation des EPV du français repose sur le guide d'annotation [6] de la campagne PARSEME (Savary *et al.*, 2017). Une EPV y est définie comme une EP dont la forme canonique a pour tête syntaxique un verbe, et dont la distribution est celle d'un verbe, d'un syntagme verbal ou d'une phrase [7]. Les formes non canoniques (p. ex. au passif, ou avec un composant extrait) sont également annotées. Le schéma d'annotation permet d'annoter des expressions verbales imbriquées, ainsi que des expressions partageant certains éléments (p. ex. *Luc* ***prend***$_{1,2}$ *une* ***douche***$_1$ *puis un* ***bain***$_2$).

L'annotation pour le français utilise toutes les catégories prévues sauf celle des verbes à particule : les expressions verbales idiomatiques (ID, p. ex. ***avoir lieu***, ***faire partie***), les constructions à verbe support (CVS, p. ex. ***faire*** *une* ***proposition***) et les verbes intrinsèquement pronominaux (SeV, p. ex. ***s'évanouir***). Une quatrième catégorie « Autre » a également utilisée marginalement pour le cas d'EP de type verbes coordonnés (p. ex. ***aller et venir***) ou des verbes dont la structure interne est irrégulière (p. ex. ***court-circuiter***). L'annotation d'une expression s'effectue en plusieurs étapes. Premièrement, pour chaque verbe, les annotateurs repèrent d'après leurs connaissances linguistiques s'il y a une idiosyncrasie potentielle dans la composition du verbe avec un ou plusieurs autres composants. La deuxième étape filtre plus précisément quels sont les composants qui entreraient effectivement dans l'expression candidate (p. ex. est-ce que les déterminants sont à inclure car figés). C'est ensuite la troisième étape qui permet de trancher sur le statut effectif d'EPV de l'expression candidate, et sur sa catégorie. Elle est réalisée en suivant deux arbres de décision, dont nous donnons un aperçu seulement, pour insister sur la marche à suivre très précise fournie aux annotateurs.

Le premier arbre de décision fonctionne comme une série de conditions suffisantes d'EP (on sort de l'arbre dès qu'une condition est vérifiée), comme le fait qu'un des composants ne puisse exister de

4. `https://gitlab.com/parseme/sharedtask-data/tree/master`
5. En ce qui concerne les données françaises : (i) les annotations en EPV sont diffusées sous CC-BY v4, (ii) les annotations morpho-syntaxiques pré-existantes sont diffusées sous licence CC BY-NC-SA v4 pour la partie UD, et LGPL-LR pour la partie Sequoia.
6. `http://parsemefr.lif.univ-mrs.fr/guidelines-hypertext`
7. Ainsi, par exemple, les noms composés comportant des verbes (p. ex. *porte-parole*) ne sont pas des EPV.

manière autonome (p. ex. ***prendre la poudre d'escampette***), ou qu'une transformation régulière, de type substitution lexicale, modification syntaxique ou morphologique, provoque une inacceptabilité ou un changement de sens inattendu. Par exemple, la forme idiomatique de ***jeter l'éponge*** ne se passive pas alors que c'est le cas pour sa forme libre ; de même, remplacer *éponge* par *serpillière* n'est pas autorisé dans son sens idiomatique (*#jeter la serpillière*). Par ailleurs *éponge* est invariable en nombre (*#jeter les éponges*).

À ce stade, si une condition suffisante a été trouvée, le candidat est considéré comme EPV, et le deuxième arbre de décision permet de préciser son type. Par exemple, on annotera « Autre » une expression où l'on n'a pas un seul verbe fonctionnant comme tête syntaxique. Toutefois, même si une condition suffisante n'a pas été trouvée avec le premier arbre, dès lors que le candidat a la forme d'une CVS, des tests spécifiques aux CVS doivent être appliqués.

Expressions verbales idiomatiques (ID). Les expressions satisfaisant une condition suffisante d'EP sont typées ID dans les cas suivants :

- le verbe de l'expression n'a pas un et un seul dépendant syntaxique lexicalisé (p. ex. ***prendre le taureau par les cornes***, ***il est question***)
- le verbe de l'expression a un et un seul dépendant syntaxique lexicalisé, et ne satisfait pas les critères de CVS (cf. ci-dessous).

Constructions à verbes support (CVS). Une EPV candidate est annotée CVS si toutes les conditions suivantes sont satisfaites :

- Le candidat a la forme verbe v plus un dépendant nominal n direct ou *via* préposition régie.
- Le nom n a un de ses sens habituels, il a au moins un argument sémantique, et il décrit un événement ou un état (*décision*, *courage*) [8].
- Le verbe v n'ajoute aucune sémantique qui ne soit déjà présente dans le sens du nom, mis à part la sémantique des marques de flexion, et l'indication de quel est l'argument sémantique du nom qui est réalisé comme sujet du verbe. Ainsi on annotera aussi bien *donner un ordre* que *recevoir un ordre*. On n'annotera pas *X prend la responsabilité [de]* (car inchoatif), *X donne la migraine à Y* (car causatif), *X termine sa promenade* (aspectuel), alors qu'on annotera ces mêmes noms avec respectivement les verbes *avoir, avoir* et *faire*. **Il est important de noter que ce critère est à la fois une restriction de la notion habituelle de CVS, qui comprend des CVS aspectuelles ou causatives (p. ex. Gross (1993)), et une généralisation de la notion de verbe support, car on n'impose pas que le verbe support ait perdu sa sémantique habituelle**. On peut toutefois noter une faiblesse dans le traitement actuel, pour les cas où le verbe n'a pas un de ses sens habituels, mais a un sens causatif ou aspectuel, comme par exemple ***attirer l'attention***. On doit alors coder ID car le verbe est idiosyncratique, et pas CVS car le verbe n'est pas sémantiquement neutre.
- On doit pouvoir former un groupe nominal (GN) en ajoutant le sujet de v au GN de n, et ce GN étendu doit pouvoir référer à l'éventualité décrite par la version avec verbe support. Par exemple, avec *Luc* ***prend*** *une* ***décision***, on parle de *la décision de Luc*. Avec *Luc* ***donne*** *l'***ordre** *de partir à Paul*, la réduction est plus difficile telle quelle (? *L'ordre de Luc à Paul de*

8. Un relecteur fait remarquer que ce critère purement sémantique est sans doute superflu. En réalité, il a été utilisé dans la campagne comme condition nécessaire pour un nom d'avoir un argument sémantique, autre qu'un possesseur. La contrainte d'un argument sémantique écarte par exemple les noms atmosphériques : *La pluie tombe* n'est actuellement pas annoté, l'annotation se concentrant sur les cas où il y a divergence syntaxe-sémantique, avec un argument syntaxique du verbe qui est en réalité un argument sémantique du nom.

partir rapidement était un peu rude), mais acceptable si l'on pronominalise le sujet (*son ordre à Paul de partir rapidement était un peu rude*).

— Dans la forme canonique de la CVS, le sujet de v correspond à un actant de n, d'où l'impossibilité de réaliser cet actant à la fois au sein du complément nominal et comme sujet de v (**Paul reçoit la visite de Pierre à Jacques*), sauf à devoir interpréter une comparative (*Paul **fait** la **promenade** de Luc*) [9].

Verbes intrinsèquement pronominaux (seV). Les verbes pronominaux sont des combinaisons d'un verbe v et d'un clitique réflexif (en français *se*, *me*, *te*, *nous*, *vous*, que l'on note ci-dessous SE), ayant différents statuts possibles. Le phénomène existe dans de nombreuses langues, dont pour la campagne, les langues romanes, les langues slaves, l'allemand et le suédois. Un sous-guide d'annotation a été mis au point pour ces cas, sous la forme d'un arbre de décision. Ont été considérés comme des EPV les cas intrinsèquement impersonnels, c'est-à-dire pour lesquels il n'existe pas une relation régulière avec une version sans SE du verbe. N'ont donc pas été annotés les cas de vrais réfléchis ou réciproques, les moyens ou « à agent fantôme » (p. ex. *une telle vitre se casse avec un marteau*), ou les neutres (p. ex. *la branche s'est cassée d'un coup*). En revanche, nous avons annoté les cas de verbes v n'apparaissant jamais sans le clitique SE (p. ex. ***se suicider***), ou dont le clitique SE modifie de manière imprévisible le comportement de v, sur le plan syntaxique (p. ex. ***s'apercevoir** de Y*, cf. **X aperçoit Z de Y*) ou sémantique. Pour trancher ce dernier cas, difficile, le guide d'annotation utilise un critère d'implication logique : si « X v Y » $\Rightarrow$ « Y SE v » alors l'expression candidate n'est pas considérée comme une EPV. Par exemple *le clown égaye les enfants* $\Rightarrow$ *les enfants s'égayent.*

5 Méthodologie d'annotation

Une fois le guide multilingue PARSEME stabilisé, la phase d'annotation proprement dite a été réalisée sur deux mois, en parallèle pour les différentes langues, avec l'outil FLAT [10] (van Gompel & Reynaert, 2013). Six personnes ont annoté le français (les auteurs de cet article). Par manque de temps et de moyens, il n'a pas été possible de réaliser une double annotation suivie d'une adjudication : chaque portion de corpus a été annotée par une seule personne, sauf un extrait pour le calcul de l'accord (cf. section 6). Pour compenser la perte de qualité potentielle (erreurs d'inattention et incohérences du fait d'interprétations différentes des consignes d'annotation), nous avons utilisé différents outils :

— Pendant la phase d'annotation, des questions sur l'interprétation du guide ou des demandes de précision pouvaient être adressées et débattues via le gestionnaire de tickets gitlab, à différents niveaux (langue, groupe de langues, toutes les langues).

— Parallèlement, les annotateurs ont maintenu une liste de cas précis tranchés collectivement. Contrairement à la résolution de conflits via adjudication, une telle organisation ne permet pas de garantir la cohérence des différentes décisions entre elles, cohérence normalement évaluée par la capacité des annotateurs à converger sur la seule base du guide d'annotation. Elle permet bien cependant de limiter les incohérences d'annotation pour un même phénomène.

— Après l'annotation simple, nous avons utilisé un outil de repérage automatique de bruit et de silence. Il extrait la liste des EPV annotées, avec pour chaque EPV les occurrences annotées, et de possibles occurrences oubliées avec une recherche approchée bruitée. Un parcours manuel

9. À noter que ne sont pas pris en compte les rares cas où le nom prédicatif est sujet (*les applaudissements de la foule crépitèrent*, (Jousse, 2010)).

10. `http://github.com/proycon/flat`, `http://flat.science.ru.nl`

de cette liste a permis de rapidement corriger des incohérences.

— Enfin une dernière étape a utilisé la liste des cas tranchés négativement pour repérer automatiquement et supprimer des cas annotés à tort.

6 Description et évaluation du corpus

Les EPV ont été annotées sur deux corpus préexistants, annotés pour la morphologie et la syntaxe en dépendances : la partie française du corpus Universal Dependencies [11] (Nivre *et al.*, 2016), qui comprend 16 447 phrases françaises extraites au hasard de Google News, Blogger, Wikipedia et des avis de consommateurs ; et le corpus Sequoia [12] (Candito & Seddah, 2012), qui comprend 3 099 phrases issues de l'Est Républicain, de rapports de l'Agence Européenne du Médicament, de Wikipedia et d'Europarl. Pour la campagne PARSEME, les 500 premières EPV ont été réservées comme corpus d'évaluation des systèmes participants (*test*). Le restant du corpus a été considéré comme corpus d'entraînement pour les systèmes (*train*) [13].

	#Phrases	#Tokens	#EPV	#ID	#SeV	#CVS	#Autre
Complet	19 547	486 005	4 962	1 905	1 418	1 633	6
Train	17 880	450 221	4 462	1 786	1 313	1 362	1
Test	1 667	35 784	500	119	105	271	5

TABLE 1 – Statistiques sur le corpus annoté divisé en corpus d'entraînement (*train*), corpus d'évaluation (*test*) et corpus complet : nombre de phrases, de tokens, nombre total d'EPV, suivi d'un découpage par catégorie d'EPV.

La table 1 fournit la taille du corpus annoté en phrases et en tokens, et les nombres d'EPV annotées. La table 2 donne des informations sur la longueur des EPV en nombre de tokens et les discontinuités au sein des EPV. Si l'on se concentre sur le corpus complet, on constate qu'environ 4 EPV sur 10 sont des ID, 3 sur 10 sont des verbes pronominaux, et 1 sur 3 est une CVS. La catégorie Autre est marginale. On constate que le corpus de *test* est atypique : il comprend beaucoup plus de CVS en proportion, et les EPV y sont globalement plus discontinues que dans le corpus complet (seulement $42,9\%$ d'EPV continues dans le *test*, contre $60,0\%$ dans le corpus complet). Pour ce qui est de la longueur des EPV, les 3 quarts des instances comportent deux tokens ($77,4\%$), et un peu moins de 20% comportent 3 tokens. En étudiant ces mêmes indicateurs mais par type d'EPV, on constate que la plupart des EPV de longueur > 2 sont des ID ($45,5\%$ des IDs sont de longueur 3).

La même table 2 fournit des informations sur les discontinuités, calculées en nombre total de tokens apparaissant entre des composants d'EPV mais n'en faisant pas partie. Là encore, on constate de fortes disparités selon le type d'EPV. Par exemple, la proportion d'EPV sans aucune discontinuité est globalement de $60,0\%$, mais de $76,6\%$ pour les IDs, $85,8$ pour les SeV, et seulement de $18,1$ pour les CVS. Pour ces dernières, environ la moitié des annotations ont une discontinuité réduite à un seul token (souvent pour le déterminant du nom), et 8% ont une discontinuité de plus de 3 tokens.

11. version 1.4, `http://universaldependencies.org/`

12. version 7.0, `https://www.rocq.inria.fr/alpage-wiki/tiki-index.php?page=CorpusSequoia`. La tokenisation du corpus Sequoia a été automatiquement rapprochée de celle du corpus UD, en particulier pour les prépositions contractées et les mots composés grammaticaux (seuls composés annotés originellement dans Sequoia).

13. Il est apparu *a posteriori* que ce découpage correspond exactement à la partie médicale et la partie Est Républicain du corpus Sequoia. Cela explique des différences importantes de distributions des phénomènes entre *train* et *test*.

	Longueur de l'EPV			Longueur des discontinuités (en nombre de tokens)						
Corpus	Moy.	% lg=2	% lg=3	Moy.	DAM	% d=0	% d=1	% d=2	% d=3	% d>3
Tot.	2,28	77,4	18,8	0,68	0,81	60,0	26,4	7,9	2,7	2,9
Tot. ID	2,67	46,3	45,1	0,33	0,51	76,6	17,4	3,7	1,6	0,6
Tot. SeV	2,00	99,8	0,1	0,17	0,3	85,8	11,8	1,8	0,4	0,1
Tot. CVS	2,07	94,5	4,2	1,52	1,07	18,1	49,5	18,3	6,0	8,0
Train	2,29	77,1	19,2	0,65	0,80	61,9	25,0	7,5	2,8	2,8
Test	2,24	80,2	15,2	0,95	0,81	42,9	39,1	11,9	1,6	4,5

TABLE 2 – Longueur des EPV et longueur cumulée des discontinuités (en nombre de tokens) dans le corpus complet (en tout, et par type d'EPV), et dans les corpus d'entraînement (*train*) et d'évaluation (*test*). Col 1 : longueur moyenne. Col 2 et 3 : pourcentages d'EPVs de longueur 2 et 3. Col 4 et 5 : Long. moyenne et déviation absolue moyenne (DAM) des discontinuités. Col 6 : Pourcentage d'EPV sans discontinuité. Col 7 à 10 : Pourcentage d'EPV avec discontinuité de 1, 2, 3 et plus de 3 tokens.

Afin d'estimer la qualité de la méthodologie d'annotation et du corpus résultant, des extraits de corpus ont été doublement annotés. L'extrait pour le français comprend 1 000 phrases (24 666 mots) [14]. L'accord inter-annotateur (AIA) pour la tâche d'identification est évalué *via* une F-mesure, où les annotations du premier annotateur jouent le rôle de la référence. Une annotation est considérée correcte si elle couvre l'ensemble précis des éléments lexicalisés d'une EPV de la référence. Pour la catégorisation, on calcule un kappa de Cohen sur les EPV identifiées par les deux annotateurs avec les mêmes composants. On obtient pour l'extrait français $F = 0,819$ et $\kappa = 0,93$, soit un accord substantiel, parmi les 3 meilleurs dans l'ensemble des 12 langues concernées [15].

7 Conclusion

Nous avons présenté une ressource d'environ $5\,000$ instances d'expressions polylexicales verbales annotées sur environ $19\,500$ phrases en français, comprenant des expressions idiomatiques, des verbes intrinsèquement pronominaux et des constructions à verbe support. Les perspectives futures sont par exemple d'étudier la variation interne au corpus, selon les domaines des phrases annotées, ainsi que les taux d'ambiguïté. Une extension à tout type d'EP est également prévue.

Remerciements

Ce travail a été mené dans le cadre de l'Action COST PARSEME (IC1207), et du projet ANR PARSEME-FR (ANR-14-CERA-0001). Les auteurs remercient chaleureusement les organisateurs de la campagne, pour le travail mené sur le guide d'annotation, et pour toute l'infrastructure d'annotation.

14. Un des deux annotateurs de l'extrait doublement annoté a participé à la rédaction du guide.
15. Notons que la qualité des annotations a pu encore être améliorée *via* les outils de recherche de bruit et de silence (sec. 5).

Références

ABEILLÉ A. & CLÉMENT L. (2006). *Annotation morpho-syntaxique - Les mots simples, les mots composés - Corpus Le Monde.* Rapport interne, TALANA, Université Paris 7.

ABEILLÉ A., CLÉMENT L. & TOUSSENEL F. (2003). Building a treebank for french. In A. ABEILLÉ, Ed., *Treebanks*, p. 165–187. Dordrecht : Kluwer.

BALDWIN T. & KIM S. N. (2010). Multiword expressions. In N. INDURKHYA & F. J. DAMERAU, Eds., *Handbook of Natural Language Processing*, p. 267–292. Boca Raton, FL, USA : CRC Press, Taylor and Francis Group, 2 edition.

CANDITO M. & SEDDAH D. (2012). Le corpus sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *Proceedings of TALN 2012*.

COURTOIS B., GARRIGUES M., GROSS G., GROSS M., JUNG R., MATHIEU-COLAS M., SILBERZTEIN M. & VIVÈS R. (1997). *Dictionnaire électronique des noms composés DELAC : les composants NA et NN. Rapport technique.* Rapport interne, LADL, Université Paris 7.

GROSS G. (1993). Trois applications de la notion de verbe support. *L'Information Grammaticale*, **59**(1), 16–22.

GROSS M. (1986). Lexicon-grammar : The Representation of Compound Words. In *Proceedings of the 11th Coference on Computational Linguistics*, COLING '86, p. 1–6, Stroudsburg, PA, USA : Association for Computational Linguistics.

HEID U. (2008). In *Phraseology. An interdisciplinary perspective*, chapter Computational phraseology. An overview, p. 337–360. John Benjamins Publishers : Amsterdam, Netherlands.

JOUSSE A.-L. (2010). *Modèle de structuration des relations lexicales fondé sur le formalisme des fonctions lexicales*. PhD thesis, Université de Montréal et Université Paris Diderot.

MEL'ČUK I. (2010). La phraséologie en langue, en dictionnaire et en taln. In *Conférence invitée de TALN 2010*, Montréal, Canada.

MEL'ČUK I., ARBATCHEWSKY-JUMARIE N., DAGENAIS L., ELNITSKY L., IORDANSKAJA L., LEFEBVRE M.-N. & MANTHA S. (1988). *Dictionnaire explicatif et combinatoire du français contemporain : Recherches lexico-sémantiques*, volume II of *Recherches lexico-sémantiques*. Presses de l'Univ. de Montréal.

NIVRE J., DE MARNEFFE M.-C., GINTER F., GOLDBERG Y., HAJIC J., MANNING C. D., MCDONALD R., PETROV S., PYYSALO S., SILVEIRA N., TSARFATY R. & ZEMAN D. (2016). Universal dependencies v1 : A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* : European Language Resources Association (ELRA).

SAG I. A., BALDWIN T., BOND F., COPESTAKE A. & FLICKINGER D. (2001). Multiword expressions : A pain in the neck for nlp. In *In Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, Mexico City, Mexique.

SAVARY A., RAMISCH C., CORDEIRO S., SANGATI F., VINCZE V., QASEMIZADEH B., CANDITO M., CAP F., GIOULI V., STOYANOVA I. & DOUCET A. (2017). The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, Valencia, Spain : Association for Computational Linguistics.

TUTIN A., ESPERANÇA-RODIER E., IBORRA M. & REVERDY J. (2015). Annotation of multiword expressions in French. In C.-P. GLORIA, Ed., *European Society of Phraseology Conference (EUROPHRAS 2015)*, Computerised and Corpus-based approaches to phraseology : monolingual and multilingual perspectives, p. 60–67, Malaga, Spain.

VAN GOMPEL M. & REYNAERT M. (2013). FoLiA : A practical XML format for linguistic annotation - a descriptive and comparative study. *Computational Linguistics in the Netherlands Journal*, **3**, 63–81.

Évaluation de mesures d'association pour les bigrammes et les trigrammes au moyen du test exact de Fisher

Yves Bestgen
CECL, Place du Cardinal Mercier 10, B-1348 Louvain-la-Neuve, Belgique
yves.bestgen@uclouvain.be

RÉSUMÉ

Pour déterminer si certaines mesures d'association lexicale fréquemment employées en TAL attribuent des scores élevés à des n-grammes que le hasard aurait pu produire aussi souvent qu'observé, nous avons utilisé une extension du test exact de Fisher à des séquences de plus de deux mots. Les analyses ont porté sur un corpus de quatre millions de mots d'anglais conversationnel extrait du BNC. Les résultats, basés sur la courbe précision-rappel et sur la précision moyenne, montrent que le LL-simple est extrêmement efficace. IM3 est plus efficace que les autres mesures basées sur les tests d'hypothèse et atteint même un niveau de performance presque égal à LL-simple pour les trigrammes.

ABSTRACT

Using Fisher's Exact Test to Evaluate Association Measures for Bigrams and Trigrams

To determine whether some often-used lexical association measures assign high scores to n-grams that chance could have produced as frequently as observed, we used an extension of Fisher's exact test to sequences longer than two words to analyse a corpus of four million words. The results, based on the precision-recall curve and the average precision, show that simple-ll is extremely effective. MI3 is more efficient than the other hypothesis tests-based measures and even reaches a performance level almost equal to simple-ll for trigrams.

MOTS-CLÉS : Mesures d'association lexicale ; N-grammes de mots ; LL-simple ; IM3.

KEYWORDS: Lexical association measures ; Word n-grams ; Simple-LL ; MI3.

1 Introduction

L'évaluation de l'efficacité des mesures d'association lexicale pour identifier les expressions polylexicales retient l'attention de nombreux chercheurs depuis plus de vingt ans (Church & Hanks, 1990 ; Evert & Krenn, 2005 ; Ramish, 2015). Dans le cas de bigrammes, un grand nombre de mesures ont été comparées et plusieurs synthèses sont disponibles (Evert, 2008 ; Pecina, 2010). Les mesures d'association pour des n-grammes de plus de deux mots ont retenu beaucoup moins l'attention comme l'ont souligné Evert (2008), Gries (2010) et Nerima *et al.* (2010).

Cette situation a probablement conduit certains chercheurs à privilégier la fréquence de cooccurrence brute comme critère de sélection (Wermter & Hahn, 2006 ; Biber, 2009). Dans le cas des bigrammes, cependant, il est bien établi que la fréquence seule est insuffisante pour distinguer les séquences intéressantes de celles que le hasard aurait pu produire en aussi grand nombre qu'observé dans un corpus (Evert, 2008 ; Gries, 2010). Selon Ellis *et al* (2015), la fréquence devrait également être peu

efficace pour identifier des séquences de plus de deux mots. Pourtant, plus une séquence est longue, moins il est probable qu'elle se produise par hasard, même si elle est composée de mots fréquents. La question est donc de savoir quelle longueur doit avoir une séquence pour s'assurer que la fréquence brute ne sélectionne que les séquences qui ne sont probablement pas dues au hasard.

Cette question est en fait bien plus générale. Elle se pose à propos de presque toutes les mesures d'association puisque celles-ci visent principalement à mettre en évidence des séquences dont la force d'association est suffisamment supérieure à ce que le hasard pourrait produire. Si cela est évident pour les mesures d'association basées sur des tests d'hypothèses, tels que *z* ou *LL-simple*, c'est également vrai pour des mesures de taille d'effet, telles que *IM* et ses variantes, car ces mesures comparent également les fréquences observées aux fréquences attendues par le seul effet du hasard, même si elles ne dérivent pas d'un test statistique. La comparaison de la fréquence aux mesures d'association est d'autant plus importante que Biber (2009) a formulé des critiques très fortes contre l'utilisation des mesures d'association pour sélectionner les n-grammes les plus intéressants dans un corpus.

Pour essayer de répondre à cette question de l'efficacité des mesures d'association pour identifier des séquences de plus de deux mots, cette recherche propose d'employer le test exact de Fisher, le test inférentiel considéré comme le plus adéquat pour analyser la fréquence de bigrammes dans un corpus (Evert, 2008 ; Moore, 2004 ; Pedersen, 1996). Il est important de noter que cette étude vise uniquement à développer un point de vue complémentaire sur l'efficacité de certaines mesures d'association. En aucun cas, l'approche proposée ne pourra remplacer une évaluation fondée sur une liste étalon établie manuellement ou sur les bénéfices que les expressions polylexicales extraites automatiquement peuvent apporter à des applications du TAL.

La section suivante décrit l'approche employée pour étendre le test de Fisher à des n-grammes de plus de deux mots. Nous présentons ensuite une expérience, menée sur un corpus de quatre millions de mots, qui analyse les performances de six mesures d'association pour extraire des séquences de deux à quatre mots, les bigrammes servant de point de repère auquel les autres séquences sont comparées.

2 Étendre le test exact de Fisher à des n-grammes de plus de deux mots

Le test exact de Fisher est considéré comme le test inférentiel le plus adéquat pour calculer la probabilité que le hasard seul a de produire au moins autant d'occurrences d'un bigramme que le nombre effectivement observé dans un corpus (Evert, 2008 ; Jones & Sinclair, 1974 ; Moore, 2004 ; Pedersen *et al.*, 1996). La proposition de Fisher est de considérer l'ensemble des tables de contingence qu'il est possible de construire en respectant les totaux marginaux réellement obtenus et de déterminer la proportion de celles-ci qui donne lieu à un résultat au moins aussi extrême que celui observé, une table plus extrême étant une table dans laquelle la fréquence du bigramme est plus élevée que celle observée. La formule de calcul est basée sur la distribution hypergéométrique (Evert, 2004).

Son utilisation pour analyser des séquences de mots plus longues est toutefois problématique parce que cela implique d'analyser des tables de contingence à trois dimensions ou plus et que les procédures de calculs exacts pour celles-ci (Zelterman *et al.*, 1995) ne sont pas adaptées à l'étude des séquences de mots, en raison de la taille de l'échantillon et du grand nombre de tests à effectuer. Le tableau 1 présente, à gauche, la table de contingence obtenue dans le corpus qui sera analysé ci-après (5 529 378 formes graphiques) pour le trigramme *a_lot_of*. À droite, on y trouve une des très nombreuses tables

Mot_1	Mot_2	Mot_3	
		of	¬of
a	lot	1723	1357
	¬lot	5146	69102
¬a	lot	179	737
	¬lot	35244	5415890

Mot_1	Mot_2	Mot_3	
		of	¬of
a	lot	3996	0
	¬lot	19148	54184
¬a	lot	0	0
	¬lot	19148	5432902

TABLE 1 – Tables de contingence pour le trigramme *a_lot_of*

de contingence les plus extrêmes possible qui respectent l'effectif total et les totaux marginaux univariés (la fréquence totale de chaque mot). Ces tables les plus extrêmes ne sont évidemment qu'une infime fraction de toutes les tables possibles qui respectent les mêmes conditions. À titre d'exemple, Mielke, Berry et Zelterman (1994) ont calculé qu'il y a plus de 3,6 milliards de tables différentes compatibles avec une table de contingence $2\times2\times2$ dont l'effectif total est 1663 et les totaux marginaux 624, 623, et 384, soit des valeurs incomparablement plus petites que celles qui doivent être analysées pour chaque trigramme et quadrigramme du corpus.

Bestgen (2014) a proposé d'appliquer aux séquences de plus de deux mots la procédure approchée recommandée pour estimer la probabilité du test exact de Fisher lorsque l'utilisation de la formule hypergéométrique n'est pas possible (Agresti, 1992 ; Pedersen, 1996). Cette méthode repose sur une procédure de permutation de type Monte-Carlo qui génère un échantillon aléatoire des tables de contingence possibles, compte tenu des totaux marginaux, et estime la probabilité exacte sur la base de la proportion de ces tables qui produisent une valeur au moins aussi extrême que celle observée. Étant donné qu'un corpus est une longue séquence de formes graphiques, une table de contingence possible est simplement une permutation aléatoire de l'ensemble de ces formes. Toute permutation de ces formes génère donc une table de contingence pour chacun des bigrammes originaux, résolvant le problème du grand nombre de tests à effectuer. Cette procédure d'estimation peut être directement généralisée à des n-grammes plus longs en comptant ceux-ci dans chaque permutation.

Bestgen (2014) a montré que cette procédure permet une estimation presque parfaite de la probabilité du test de Fisher pour tous les bigrammes dans un corpus, avec la limitation que le nombre de permutations effectuées détermine la précision de la probabilité, celle-ci ne pouvant être plus petite que 1 divisé par le nombre de permutations effectuées. La principale faiblesse de la procédure est donc son coût-calcul.

3 Expérience

Cette expérience vise à évaluer, au moyen du test de Fisher, l'efficacité de six mesures d'association qui ont été proposées pour extraire des expressions polylexicales de corpus.

3.1 Matériel et procédure

3.1.1 Corpus

Les analyses ont été menées sur un corpus de 5 529 378 formes graphiques, dont 4 232 259 mots de conversations spontanées extraits du BNC. Ce corpus, similaire à ceux utilisés dans plusieurs études

sur les paquets lexicaux (Biber, 2009), a été choisi, car il est suffisamment grand pour extraire des expressions composées de plusieurs mots, mais pas trop grand de sorte qu'un nombre suffisant de permutations puisse être obtenu en un temps raisonnable.

3.1.2 Procédure

La version du corpus lemmatisée par CLAWS a été utilisée. Toutes les formes graphiques (mots, nombres, symboles et ponctuations) détectées par CLAWS (http ://ucrel.lancs.ac.uk/claws/) ont été considérées comme des éléments à permuter aléatoirement. Vingt millions de permutations ont été effectuées ; la plus petite probabilité qui peut être estimée est donc de 0,00000005. Dans les résultats, seules les chaînes de mots ininterrompues sont prises en compte. Étant donné que les mesures d'association calculées pour des n-grammes très rares ne sont pas fiables (Evert, 2008), nous avons analysé uniquement les séquences qui apparaissent au moins trois fois dans le corpus original.

3.1.3 Mesures d'association évaluées

Pour ces analyses, une série de mesures d'association ont été sélectionnées en fonction de leur popularité dans les études des bigrammes et des séquences plus longues : *IM* (ou information mutuelle ponctuelle) et *t*-score (Church *et al.*, 1991), z (Berry-Rogghe, 1973), *LL-simple* (Evert, 2008) et la fréquence brute (f). Nous avons également analysé *IM3*, une modification heuristique de *IM*, proposée pour réduire la tendance de celle-ci à attribuer des scores importants à des mots rares qui ne s'observent ensemble que quelquefois (Daille, 1994). Cependant, comme *IM3* viole une convention des mesures d'association en attribuant, dans certains cas, des scores positifs aux n-grammes qui se produisent moins fréquemment que le hasard le prédit (Evert, 2008, p. 1226), nous avons modifié la formule de telle sorte que ces n-grammes reçoivent un score négatif et, par conséquent, apparaissent à la fin de la liste [1]. Un problème similaire se pose pour *LL-simple* ; nous avons utilisé la version signée recommandée par Evert (2008). Les formules de ces indices pour les trigrammes et les quadrigrammes sont identiques à celles utilisées pour les bigrammes, la fréquence attendue étant calculée en utilisant l'extension habituelle de la formule pour les bigrammes (c.-à-d., en multipliant le produit des probabilités marginales de tous les mots de la séquence par la taille du corpus (Ramish, 2015, p. 64)).

4 Analyses et résultats

La principale question de recherche à laquelle cette étude tente de répondre est de déterminer si certaines mesures d'association attribuent des scores élevés à des n-grammes dont la fréquence de cooccurrence est susceptible de résulter du seul effet du hasard. Le seuil de signification a été fixé à 0,05, ce qui implique qu'on accepte le risque qu'une séquence soit déclarée erronément statistiquement significative (donc, trop fréquente) dans 5 % des cas. La procédure séquentielle de Holm (Holm, 1979 ; Howell, 2008, p. 370–371) a été utilisée pour tenir compte du grand nombre de tests effectués qui augmente la probabilité de décider de manière erronée qu'au moins un test est

1. Plusieurs modifications ont été évaluées, mais comme elles ont donné lieu à des performances presque identiques, nous avons employé la formule de IM pour tout n-gramme observé moins souvent que le hasard le prédit.

	Bigramme	Trigramme	Quadrigramme
Nbr. de n-grammes différents	96881	111164	53334
Nbr. de n-grammes significatifs	27063	70942	52825
PM du niveau de base	0.279	0.638	0.990

TABLE 2 – Statistiques pour les trois longueurs de n-gramme

statistiquement significatif. Cette procédure assure un taux d'erreur d'ensemble de 0,05 pour chaque longueur de n-grammes, autorisant des comparaisons entre celles-ci.

Étant donné que la liste des n-grammes pour lesquels le test de Fisher indique qu'ils ne sont pas dus au hasard est considérée comme la référence à atteindre par les autres mesures, les résultats ont été résumés de la manière habituelle (Biemann *et al.*, 2013, p. 40–41) par des courbes précision-rappel (PR) et par la précision moyenne (PM, *average precision*). Comme niveau de base, nous avons utilisé la précision moyenne atteinte par une mesure d'association qui classe les n-grammes au hasard (Bestgen, 2015). Le nombre de n-grammes testés, le nombre de n-grammes significatifs pour le test de Fisher et la PM du niveau de base sont donnés dans le tableau 1. On constate que presque tous les quadrigrammes sont considérés par le test inférentiel comme étant statistiquement significatifs, mais aussi que le niveau de base est beaucoup plus élevé pour les trigrammes que pour les bigrammes.

Ces résultats ont deux implications pour les analyses présentées dans la suite. Tout d'abord, celles-ci n'ont été effectuées que sur les bigrammes et les trigrammes puisque presque tous les quadrigrammes présents dans le corpus au moins trois fois passent ce test inférentiel avec succès. Cette observation répond donc à la première question posée dans cette étude : dès quatre mots, une séquence est suffisamment longue pour garantir que la fréquence brute ne sélectionne (quasiment) que des séquences qui ne sont probablement pas dues au hasard. Toutefois, ce résultat ne démontre pas qu'ordonner les séquences sur la base de la seule fréquence est une procédure efficace pour extraire les meilleures expressions composées de plusieurs mots, les quadrigrammes arrivant en tête de liste n'étant pas nécessairement les meilleurs pour un annotateur humain.

La deuxième conséquence concerne la comparaison des performances des mesures d'association pour les bigrammes et les trigrammes puisqu'une meilleure performance est attendue dans le second cas en raison du niveau plus élevé atteint par le hasard. Pour cette raison, en plus de la PM standard, nous rapportons un nouvel indice appelé *Précision moyenne corrigée pour le hasard* (PMch) obtenu en insérant dans la formule classique du Kappa (Cohen, 1960 ; Powers, 2012) la PM du niveau de base, puisqu'il s'agit du niveau de performance que devrait produire un classement aléatoire des n-grammes. La formule de cet indice est donc : $(PMo - PMnb)/(1 - PMnb)$ dans laquelle PMo représente la PM observée et $PMnb$ celle du niveau de base.

Les figures 1 et 2 présentent les courbes PR. Pour chaque mesure, la ligne commence au premier n-gramme qui ne passe pas le test inférentiel. Jusqu'à ce point, la précision est donc de 100 %. Pour faciliter la comparaison des figures, le point de départ de l'axe des *y* a été fixé à une valeur juste inférieure à la PM du niveau de base. Le tableau 2 rapporte les PM et les PM corrigées pour le hasard.

Les courbes PR pour les bigrammes indiquent que *LL-simple* est de loin la meilleure mesure d'association, atteignant un niveau de performance presque parfait. Elle est suivie par *IM3*, *z* et *t* [2]. Comme prévu, la fréquence brute est nettement moins efficace, sa courbe PR montrant qu'elle classe en tête

2. La forme distinctive de la courbe PR pour *t* est due à sa formule qui attribue des scores presque identiques aux n-grammes qui se produisent dans le corpus à la même faible fréquence et qui ont une fréquence attendue proche de 0. Ces n-grammes étant considérés comme trop fréquents par le test inférentiel, ils forment une longue série de scores PR légèrement croissants.

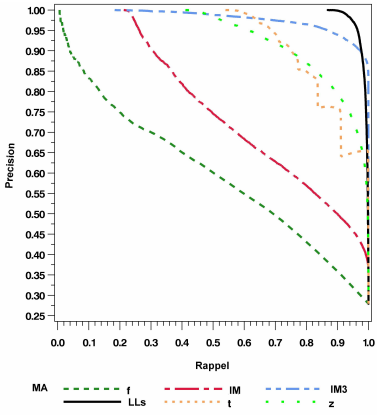


FIGURE 1 – Courbes PR pour les bigrammes

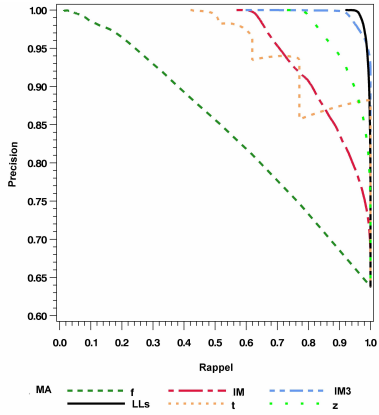


FIGURE 2 – Courbes PR pour les trigrammes

	PM		PMch	
Mesures d'association	Bigramme	Trigramme	Bigramme	Trigramme
f	0.588	0.824	0.429	0.515
IM	0.756	0.957	0.662	0.881
IM3	0.947	0.998	0.926	0.994
z	0.934	0.985	0.909	0.959
t	0.931	0.959	0.905	0.886
LL-simple	0.993	0.998	0.990	0.995

TABLE 3 – PM et PM corrigée pour le hasard (PMch)

de liste plusieurs n-grammes que le hasard pourrait avoir produit aussi fréquemment qu'observé.

Pour les trigrammes, la différence entre *LL-simple* et *IM3* est extrêmement faible, comme le confirme la PM. Ce résultat est inattendu. La performance de *Z* est légèrement moins bonne. *IM* est presque aussi efficace que *t*, un autre résultat inattendu. La fréquence brute est à nouveau la moins efficace.

La comparaison des PM corrigés des bigrammes et des trigrammes montre que *t* ne s'améliore pas du tout alors que *IM*, *IM3* et *z* s'améliorent nettement.

Un objectif de ces analyses est de déterminer si les mesures d'association classent en début de liste des n-grammes que le hasard aurait pu produire aussi souvent qu'observé. Il est donc intéressant d'analyser la position dans le classement du premier n-gramme rejeté par le test inférentiel. Il correspond dans les graphiques au point le plus à gauche des courbes PR. Il est fourni en termes de rappel et doit donc être interprété en fonction du nombre de n-grammes significatifs. Pour *LL-simple*, le premier bigramme non significatif est observé lorsque le rappel atteint 87 %, soit après le 23000e bigramme dans la liste. Pour les trigrammes, c'est au-delà de la 65000e position. Ces valeurs sont clairement meilleures que celles obtenues par les autres mesures d'association, ce qui suggère qu'il pourrait être utile de filtrer les listes de ces autres mesures d'association sur cette base.

5 Discussion et conclusion

L'objectif de cette étude est de fournir une perspective différente sur l'efficacité des mesures d'association lexicales pour l'extraction d'expressions polylexicales à partir de corpus. Les analyses mettent en évidence des différences importantes entre des mesures d'association basées sur des tests d'hypothèse. La performance presque parfaite de *LL-simple* est aussi observée pour les trigrammes, étendant à ceux-ci la conclusion d'Evert (2008) selon laquelle *LL-simple* est une bonne approximation du test exact de Fisher pour les bigrammes. *Z* est moins efficace. Le score *t* l'est encore beaucoup moins et ne s'améliore pas lorsqu'on passe des bigrammes aux trigrammes. En ce qui concerne les mesures de la taille de l'effet, *IM3* [3] est nettement plus efficace que *IM* et, pour les trigrammes, atteint une performance équivalente à *LL-simple*. Il mérite certainement une évaluation approfondie.

Le critère utilisé (c.-à-d., une mesure d'association attribue-t-elle des scores élevés à des n-grammes que le hasard aurait pu produire aussi souvent qu'observé) suggère qu'il pourrait être judicieux d'utiliser *LL-simple* pour filtrer les classements générés par d'autres mesures d'association, l'extension du test de Fisher ne pouvant être employée en raison de son coût-calcul. Cependant, avant de recommander une telle procédure, il est nécessaire de s'assurer qu'elle améliore effectivement les performances des mesures d'association dans des applications du TAL et que les séquences retenues ainsi sont considérées comme de véritables expressions par un évaluateur humain.

Une des limitations de cette étude est qu'elle a été menée sur un seul corpus, composé exclusivement de conversations et d'une taille limitée. Si l'analyse d'un corpus plus grand augmente considérablement le temps nécessaire pour obtenir les permutations, il pourrait être intéressant d'analyser des corpus plus petits, aussi semblables que possible à celui utilisé ici, afin d'estimer au moins partiellement l'impact du facteur taille.

Il serait aussi intéressant, lors de la sélection initiale des n-grammes à évaluer, de prendre en compte les catégories morphosyntaxiques des mots qui composent les séquences (Benigno & Kraif, 2016). Il est en effet possible que les performances de la fréquence brute, mais aussi de certaines mesures d'association, s'améliorent nettement si des profils syntaxiques spécifiques, comme *adjectif + préposition + nom*, sont testés.

Enfin, il faut rappeler la principale limitation de l'extension du test de Fisher aux séquences de plus de deux mots : son coût-calcul. La procédure de Monte-Carlo limite fortement la précision des probabilités les plus faibles, produisant donc un grand nombre de séquences obtenant le meilleur score (la plus faible probabilité estimable). Cette extension ne constitue donc pas un indice d'association viable, permettant de classer les séquences de la plus exceptionnelle à la moins exceptionnelle.

Remerciements

Cette recherche a bénéficié du soutien du Fonds de la recherche scientifique (crédit F.R.S.-FNRS J.0025.16) dont l'auteur est chercheur qualifié. Une partie des ressources informatiques utilisées ont été fournies par les installations de calcul intensif de l'Université catholique de Louvain (CISM/UCL) et du Consortium des Équipements de Calcul intensif en Fédération Wallonie Bruxelles (CÉCI) financé par le F.R.S.-FNRS.

3. Des analyses effectuées en utilisant la formule non corrigée ont indiqué que la correction n'améliore les performances que pour les bigrammes.

Références

AGRESTI A. (1992). A survey of exact inference for contingency tables. *Statistical Science*, **7**, 131–153.

BENIGNO V. & KRAIF O. (2016). Core vocabulary and core collocations : combining corpus analysis and native speaker judgement to inform selection of collocations in learner dictionaries. In A. ORLANDI & L. GIACOMINI, Eds., *Defining Collocations For Lexicographic Purposes : From Linguistic Theory To Lexicographic Practice (Vol. 1)*, p. 235–268. Peter Lang.

BERRY-ROGGHE G. L. M. (1973). The computation of collocations and their relevance in lexical studies. In A. J. AITKEN, R. W. BAILEY & N. HAMILTON-SMITH, Eds., *The Computer and Literary Studies*. Edinburgh University Press.

BESTGEN Y. (2014). Extraction automatique de collocations : Peut-on étendre le test exact de Fisher à des séquences de plus de 2 mots ? In *Actes de JADT 2014*, p. 79–90.

BESTGEN Y. (2015). Exact expected average precision of the random baseline for system evaluation. *The Prague Bulletin of Mathematical Linguistics*, **103**, 131–138.

BIBER D. (2009). A corpus-driven approach to formulaic language in English : Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics*, **14**, 275–311.

BIEMANN C., BILDHAUER F., EVERT S., GOLDHAHN D., QUASTHOFF U., SCHÄFER R., SIMON J., SWIEZINSKI L. & ZESCH T. (2013). Scalable construction of high-quality Web corpora. *Journal for Language Technology and Computational Linguistics*, **28**, 23–59.

CHURCH K., GALE W. A., HANKS P. & HINDLE D. (1991). Using statistics in lexical analysis. In U. ZERNIK, Ed., *Lexical Acquisition : Using On-line Resources to Build a Lexicon*, p. 115–164. Lawrence Erlbaum.

CHURCH K. & HANKS P. (1990). Word association norms, mutual information and lexicography. *Computational Linguistics*, p. 22–29.

COHEN J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**, 37–46.

DAILLE B. (1994). *Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques*. PhD thesis, Université Paris 7.

ELLIS N. C., SIMPSON-VLACH R., RÖMER U., O'DONNELL M. & WULFF S. (2015). Learner corpora and formulaic language in SLA. In S. GRANGER, G. GILQUIN & F. MEUNIER, Eds., *Cambridge Handbook of Learner Corpus Research*. Cambridge University Press.

EVERT S. (2008). Corpora and collocations. In A. LÜDELING & M. KYTÖ, Eds., *Corpus Linguistics. An International Handbook*, p. 1211–1248. Mouton de Gruyter.

EVERT S. & KRENN B. (2005). Using small random samples for the manual evaluation of statistical association measures. *Computer Speech and Language*, **19**, 450–466.

GRIES S. T. (2010). Useful statistics for corpus linguistics. In A. SÁNCHEZ & M. ALMELA, Eds., *A Mosaic of Corpus Linguistics : Selected Approaches*, p. 269–291. Frankfurt au Main, Germany : Peter Lang.

HOLM S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, **6**, 65–70.

HOWELL D. (2008). *Méthodes statistiques en sciences humaines*. Bruxelles : De Boeck Université.

JONES S. & SINCLAIR. J. (1974). English lexical collocations. a study in computational linguistics. *Cahiers de Lexicologie*, **24**, 15–61.

MIELKE P. M., BERRY K. J. & ZELTERMAN D. (1994). Fisher's exact test of mutual independence for 2×2×2 cross-classification tables. *Educational and Psychological Measurement*, **54**, 110–114.

MOORE R. C. (2004). On log-likelihood-ratios and the significance of rare events. In *Proceedings of EMNLP 2004*, p. 333–340.

NERIMA L., WEHRLI E. & SERETAN V. (2010). A recursive treatment of collocations. In *Proceedings of LREC 2010*, p. 634–638.

PECINA P. (2010). Lexical association measures and collocation extraction. *Language Resources & Evaluation*, **44**, 137–158.

PEDERSEN T. (1996). Fishing for exactness. In *Proceedings of the South Central SAS Users Group*, p. 188–200.

PEDERSEN T., KAYAALP M. & BRUCE R. (1996). Significant lexical relationships. In *Proceedings of the 13th National Conference on Artificial Intelligence*, p. 455–460.

POWERS D. M. W. (2012). The problem of kappa. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, p. 345–355.

RAMISH C. (2015). *Multiword Expressions Acquisition : A Generic and Open Framework*. Springer.

WERMTER J. & HAHN U. (2006). You can't beat frequency (unless you use linguistic knowledge) – a qualitative evaluation of association measures for collocation and term extraction. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, p. 785–792.

ZELTERMAN D., CHAN I. S. & MIELKE P. W. (1995). Exact tests of significance in higher dimensional tables. *The American Statistician*, **49**, 357–361.

Réseaux neuronaux profonds pour l'étiquetage de séquences

Yoann Dupont[1] Marco Dinarelli[1] Isabelle Tellier[1]
(1) Lattice (UMR 8094), CNRS, ENS Paris, Université Sorbonne Nouvelle,
PSL Research University, USPC, 1 rue Maurice Arnoux, 92120 Montrouge, France
yoa.dupont@gmail.com, marco.dinarelli@ens.fr,
isabelle.tellier@univ-paris3.fr

RÉSUMÉ

Depuis quelques années les réseaux neuronaux se montrent très efficaces dans toutes les tâches de Traitement Automatique des Langues (TAL). Récemment, une variante de réseau neuronal particulièrement adapté à l'étiquetage de séquences textuelles a été proposée, utilisant des représentations distributionnelles des étiquettes. Dans cet article, nous reprenons cette variante et nous l'améliorons avec une version profonde. Dans cette version, différentes couches cachées permettent de prendre en compte séparément les différents types d'informations données en entrée au réseau. Nous évaluons notre modèle sur les mêmes tâches que la première version de réseau de laquelle nous nous sommes inspirés. Les résultats montrent que notre variante de réseau neuronal est plus efficace que les autres, mais aussi qu'elle est plus efficace que tous les autres modèles évalués sur ces tâches, obtenant l'état-de-l'art.

ABSTRACT

Deep Neural Networks for Sequence Labeling

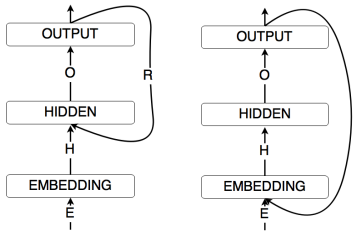
Since a couple of years, neural networks prove to be very effective on all NLP tasks. Recently, a variant of neural network particularly suited for sequence labeling has been proposed, which uses label embeddings. In this paper we propose a deep version of the above-mentioned variant of neural network, where several hidden layers allow to take into account separately the different types of information given as input to the model. We evaluate our variant on the same tasks as the basic version. Results show that our variant is not only more effective than the other neural models, but also that it outperforms all the other models evaluated on the same tasks, reaching state-of-the-art performances.

MOTS-CLÉS : Réseaux neuronaux, apprentissage artificiel, compréhension de la parole, étiquetage de séquences.

KEYWORDS: Neural Networks, Machine Learning, Spoken Language Understanding, Sequence Labeling.

1 Introduction

Ces dernières années, les réseaux neuronaux (Jordan, 1989; Elman, 1990; Hochreiter & Schmidhuber, 1997) se sont montrés très efficaces dans toutes les tâches de TAL (Mikolov *et al.*, 2010, 2011; Collobert & Weston, 2008; Collobert *et al.*, 2011; Yao *et al.*, 2013; Mesnil *et al.*, 2013; Vukotic *et al.*, 2015). Les réseaux neuronaux récurrents notamment sont très efficaces grâce à leur architecture, qui permet de prendre en compte l'information passée comme contexte, et ce sous forme de représen-



(a) Jordan (b) I-RNN

FIGURE 1 – Le RNN de Jordan et le *I-RNN* utilisé dans cet article.

tations distributionnelles. Cependant la communauté s'est rendue compte très vite que ces modèles ne sont pas particulièrement adaptés à l'étiquetage de séquences textuelles. En effet, malgré leur relative efficacité, même les modèles les plus sophistiqués tels que LSTM et GRU (Hochreiter & Schmidhuber, 1997; Cho *et al.*, 2014) ne donnent pas des résultats significativement meilleurs que ceux obtenus avec des réseaux plus simples (Vukotic *et al.*, 2015, 2016; Dinarelli & Tellier, 2016a).

Ces modèles sont très efficaces pour *"garder en mémoire"* un contexte de mots arbitrairement long. Cependant cette capacité n'a pas été explicitement exploitée pour prendre en compte un contexte au niveau des étiquettes. Dans les tâches d'étiquetage de séquences, telles que l'annotation en partie du discours ou la compréhension automatique de la parole, le contexte des étiquettes prédites précédemment est au moins aussi important que le contexte au niveau des mots. C'est pour cette raison que les modèles CRF se sont montrés très efficaces sur ce type de tâches (Lafferty *et al.*, 2001; Hahn *et al.*, 2010; Vukotic *et al.*, 2015).

La fonction de décision classique des réseaux neuronaux (*softmax*) n'est pas adaptée à l'étiquetage de séquences. Pour éliminer cette limitation, plusieurs chercheurs ont proposé des modèles neuronaux hybrides *RNN+CRF* (Huang *et al.*, 2015; Lample *et al.*, 2016; Ma & Hovy, 2016), où la couche de sortie classique est remplacée par une couche CRF neuronale.

Dans les tâches d'étiquetage de séquences telles que celles mentionnées, des modèles très efficaces peuvent être conçus en modélisant les dépendances entre étiquettes. Nous reprenons pour cela l'idée de *I-RNN* décrite dans (Dinarelli & Tellier, 2016c). Ce modèle utilise des plongements pour les étiquettes, de la même façon que les plongements pour les mots, pour apprendre les dépendances entre étiquettes et leurs interactions. Les étiquettes prédites par le modèle sont converties en index qui sont donnés en entrée du réseau à l'étape de prédiction suivante, elles sont donc transformées en plongements de la même façon que les mots. Idéalement, cette architecture peut-être vue comme une évolution du réseau de Jordan (Jordan, 1989). Dans cette évolution, la connexion récurrente du réseau relie la couche de sortie à la couche d'entrée. Un schéma de ces deux architectures est montré dans la figure 1.

Le modèle *I-RNN*, bien qu'il utilise toujours une fonction de décision locale, s'est montré très efficace sur certaines tâches d'étiquetage de séquences. Le fait d'utiliser des plongements d'étiquettes permet à ce réseau de modéliser très efficacement les dépendances entre étiquettes, et d'obtenir des résultats meilleurs non seulement que les autres RNN, mais aussi que des modèles CRF très compétitifs.

Nous reprenons le modèle *I-RNN* (Dinarelli & Tellier, 2016c) comme modèle de base pour concevoir des modèles profonds encore plus efficaces. Nous exploitons dans cet article les avantages de l'apprentissage profond en utilisant deux couches cachées différentes : i) le premier niveau est divisé en plusieurs couches cachées pour apprendre séparément la représentation interne des différentes

entrées du réseau (mots, étiquettes et autres) ; ii) le second niveau prend en entrée la concaténation de la sortie des couches cachées précédentes et produit une nouvelle sortie utilisée par la couche de sortie pour prédire la prochaine étiquette.

Pour avoir une évaluation comparable, nous évaluons nos modèles sur les mêmes tâches que celles employées par les auteurs du modèle *I-RNN* : *ATIS* (Dahl *et al.*, 1994), en anglais, et *MEDIA* (Bonneau-Maynard *et al.*, 2006), en français. Grâce à la combinaison des plongements d'étiquettes pour apprendre les dépendances entre étiquettes, et aux couches cachées profondes pour apprendre des représentations internes sophistiquées, le modèle proposé dans cet article atteint des résultats à l'état-de-l'art sur les deux tâches.

Dans la suite de l'article, nous exposons d'abord (dans la section 2) la variante *I-RNN* utilisée comme modèle de base, puis la version profonde que nous proposons (section 3). Nous décrivons ensuite les expériences effectuées pour évaluer nos modèles (section 4), et nous concluons (section 5).

2 Le modèle *I-RNN*

Dans cet article, nous partons du modèle *I-RNN* proposé par (Dinarelli & Tellier, 2016c). Ce modèle utilise une table de correspondance pour les mots E_w et une pour les étiquettes E_l. Ces deux tables sont des matrices respectivement de taille $N \times D$ et $|O| \times D$, où N est la taille du dictionnaire de mots, D est la taille des plongements, $|O|$ est le nombre d'étiquettes qui correspond aussi à la taille de la couche de sortie du réseau.

Pour rendre le modèle plus efficace, une fenêtre de taille d_w est utilisée pour les mots, alors qu'elle est de taille d_l pour les étiquettes. Nous désignons par $E_w(w_i)$ le plongement d'un mot quelconque w_i, et par $E_l(y_i)$ le plongement d'une étiquette quelconque y_i. Les entrées du réseau au niveau des mots et des étiquettes à l'étape de traitement t, sont définies respectivement comme :

$$W_t = [E_w(w_{t-d_w})...E_w(w_t)...E_w(w_{t+d_w})]$$

$$L_t = [E_l(y_{t-d_l+1})E_l(y_{t-d_l+2})...E_l(y_{t-1})]$$

où $[\,]$ désigne la concaténation de vecteurs (ou de matrices par la suite).

La couche cachée est ensuite calculée par :

$$h_t = \Phi(H[W_t L_t])$$

Φ étant la fonction d'activation *ReLU* dans cet article (Bengio, 2012). La sortie du réseau, c'est-à-dire l'étiquette prédite par le modèle, est enfin obtenue par :

$$y_t = softmax(Oh_t)$$

Grâce à l'utilisation des plongements d'étiquettes et à leur combinaison au niveau de la couche cachée, le modèle *I-RNN* peut apprendre de façon très efficace les dépendances entre étiquettes.

Comme on peut le remarquer, ce modèle n'est que idéalement récurrent, puisque les étiquettes prédites par le modèle sont converties en indexes, utilisés après pour sélectionner le plongement correspondant dans la matrice E_l. Aussi il n'y a pas de propagation en arrière à travers le temps (*BPTT*) puisque, comme dans (Dinarelli & Tellier, 2016a), le modèle est appris avec l'algorithme *Back-propagation* classique. Grâce à l'utilisation d'un contexte d'étiquette, tout cela donne une approximation d'un modèle récurrent appris avec l'algorithme *BPTT*. Comme il est montré dans (Mesnil *et al.*, 2013), ces choix ne semblent pas affecter significativement les résultats, du moins sur certains tâches.

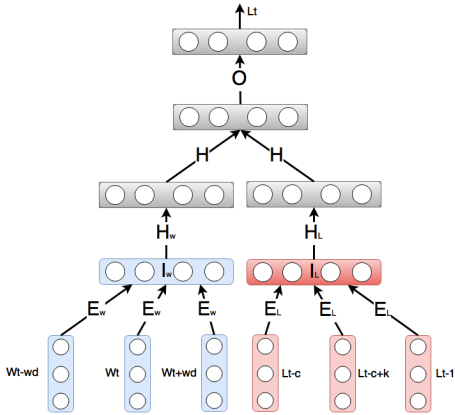


FIGURE 2 – Le *I-RNN* profond proposé dans cet article.

3 Réseaux profonds

Dans cet article nous proposons une variante profonde du modèle *I-RNN*. Cette variante exploite plusieurs niveaux de représentations internes au réseau, sous forme de couches cachées.

L'apprentissage profond pour le traitement d'images ou du signal audio (Hinton *et al.*, 2012; He *et al.*, 2015) a en effet montré qu'en utilisant plusieurs couches cachées, un réseau neuronal est capable d'apprendre des traits très fins et abstraits à partir des données. L'apprentissage profond a déjà, bien sûr, été utilisé avec succès dans les tâches de TAL. Mais quand l'entrée d'un tel réseau est constituée uniquement de mots, il est plus difficile de motiver l'utilisation des plusieurs couches cachées, car leur interprétation est problématique, au delà de l'évidence empirique constituée de meilleurs résultats. Puisque les réseaux utilisés dans cet article prennent en entrée au moins deux types différents d'informations (les mots et les étiquettes), l'utilisation de plusieurs couches cachées nous semble ici mieux justifiée.

Nous avons donc conçu un réseau neuronal profond avec deux niveaux de couches cachées. Au premier niveau chaque type d'entrée distinct (W_t et L_t, décrits plus haut) est connecté à sa propre couche cachée, respectivement à travers les paramètres H_w et H_l. La sortie de ces couches cachées est ensuite concaténée et donnée en entrée à une seconde couche cachée globale, elle-même utilisée comme entrée de la couche de sortie, comme déjà décrit plus haut. Un schéma de cette architecture profonde est fourni en figure 2.

Si d'autres entrées sont fournies au réseau (e.g. la sortie d'une convolution sur les caractères, comme cela sera proposé plus loin), celles-ci sont connectées à leur propre couche cachée dont la sortie est concaténée aux autres, avant d'être transmise en entrée à la seconde couche cachée.

La motivation de cette architecture profonde est de donner la possibilité au modèle d'apprendre une représentation interne différente pour chaque type d'entrée au premier niveau des couches cachées. Au second niveau, le réseau peut utiliser toute sa capacité de modélisation pour apprendre uniquement les interactions entre les entrées. Quand on utilise une seule couche cachée, le modèle est obligé d'apprendre à la fois une représentation interne globale pour toutes les entrées et pour leurs interactions, ce qui est un problème bien plus difficile.

MEDIA			ATIS		
Mots	**Classes**	**Étiquettes**	**Mots**	**Classes**	**Étiquettes**
Oui	-	Answer-B	i'd	-	O
l'	-	BDObject-B	like	-	O
hotel	-	BDObject-I	to	-	O
le	-	Object-B	fly	-	O
prix	-	Object-I	Delta	airline	airline-name
à	-	Comp.-payment-B	between	-	O
moins	relative	Comp.-payment-I	Boston	city	fromloc.city
cinquante	tens	Paym.-amount-B	and	-	O
cinq	units	Paym.-amount-I	Chicago	city	toloc.city
euros	currency	Paym.-currency-B			

TABLE 1 – Un exemple d'annotation pris du corpus MEDIA (gauche) et ATIS (droite).

4 Évaluation

4.1 Corpus pour la compréhension de la parole

Nous évaluons nos modèles sur les deux tâches de compréhension de la parole **ATIS** (*Air Travel Information System*) (Dahl *et al.*, 1994), en anglais, et **MEDIA** (Bonneau-Maynard *et al.*, 2006), en français.

Ces deux tâches sont celles employées pour l'évaluation des modèles auxquels nous nous comparons (Yao *et al.*, 2013; Mesnil *et al.*, 2013; Vukotic *et al.*, 2015, 2016; Dinarelli & Tellier, 2016a). Nous renvoyons les lecteurs à ces travaux pour plus de détails sur les corpus.

Un exemple comparatif d'annotation pris des deux corpus est montré dans le tableau 1.

4.2 Réglages

Notre variante de *I-RNN* profonde, appelée aussi *LD-RNN* dans (Dupont *et al.*, 2017), a été implémentée par nous-mêmes en *Octave* [1] avec le support de la bibliothèque *OpenBLAS*. [2]

Les modèles *I-RNN* sont entraînés avec la procédure suivante :

— deux modèles de langage neuronaux comme ceux de (Bengio *et al.*, 2003) sont entraînés pour générer les plongements des mots et des étiquettes. 30 et 20 époques d'entraînement, respectivement, sont utilisées pour ces deux modèles.

— les modèles *forward* et *backward* sont appris en utilisant les plongements entraînés à l'étape précédente. Ces deux modèles sont entraînés pendant 30 époques.

— le modèle bidirectionnel est appris en utilisant comme poids initiaux les modèles *forward* et *backward* entraînés à l'étape précédente. Ce dernier modèle est entraîné pendant 8 époques seulement. Grâce au fait qu'il est constitués de deux modèles appris à l'étape précédente, il est tout de suite très proche de l'optimum, souvent il est à l'optimum dès la première époque pour la tâche ATIS, alors qu'il atteint l'optimum entre la troisième et la cinquième époques sur la tâche MEDIA.

La première étape de la procédure est optionnelle, notamment sur des petites tâches comme ATIS

1. https ://www.gnu.org/software/octave/ ; Notre logiciel est décrit à la page http ://www.marcodinarelli.it/software.php et il est disponible sous requête.

2. http ://www.openblas.net

Modèle	Mesure F1		
	forward	backward	bidirectionnel
(Vukotic *et al.*, 2016) LSTM	95.12	–	95.23
(Vukotic *et al.*, 2016) GRU	95.43	–	95.53
(Dinarelli & Tellier, 2016b) E-RNN	94.73	93.61	94.71
(Dinarelli & Tellier, 2016b) J-RNN	94.94	94.80	94.89
(Dinarelli & Tellier, 2016b) I-RNN	95.21	94.64	94.75
I-RNN_{deep} Mots	94.47	94.29	94.52
I-RNN_{deep} Mots+Classes	95.40	**95.56**	**95.69**
I-RNN_{deep} Mots+Classes+CC	**95.72**	95.51	**95.67**

TABLE 2 – Comparaison des résultats sur la tâche ATIS avec la littérature, en terme de mesure F1.

Modèle	Mesure F1 / Concept Error Rate (CER)		
	forward	backward	bidirectionnel
(Vukotic *et al.*, 2015) CRF		86.00 / –	
(Hahn *et al.*, 2010) CRF		– / 10.6	
(Hahn *et al.*, 2010) ROVER×6		– / 10.2	
(Vukotic *et al.*, 2015) E-RNN	81.94 / –	– / –	– / –
(Vukotic *et al.*, 2015) J-RNN	83.25 / –	– / –	– / –
(Vukotic *et al.*, 2016) LSTM	81.54 / –	– / –	83.07 / –
(Vukotic *et al.*, 2016) GRU	83.18 / –	– / –	83.63 / –
(Dinarelli & Tellier, 2016b) E-RNN	82.64 / –	82.61 / –	83.13 / –
(Dinarelli & Tellier, 2016b) J-RNN	83.06 / –	83.74 / –	84.29 / –
(Dinarelli & Tellier, 2016b) I-RNN	84.91 / –	86.28 / –	86.71 / –
I-RNN_{deep} Mots	85.93 / 11.84	**86.80** / 10.50	**87.24 / 10.03**
I-RNN_{deep} Mots+Classes	85.63 / 11.87	86.64 / 10.41	**87.30 / 9.83**
I-RNN_{deep} Mots+Classes+CC	85.73 / 11.81	**86.83** / 10.33	**87.43 / 9.80**

TABLE 3 – Comparaison des résultats sur la tâche MEDIA avec la littérature, en terme de mesure F1 et *Concept Error Rate* (CER).

et MEDIA le pre-apprentissage des plongements n'apporte pas des gains significatifs. Cependant sur des tâches plus larges nous avons remarqué que même un modèle simple comme celui que nous utilisons pour pre-apprendre les plongements donne déjà des gains significatifs.

Puisque notre modèle, comme celui de (Dinarelli & Tellier, 2016a), n'est récurrent que grâce au fait qu'il utilise un contexte d'étiquettes sous forme de plongements, il est important d'utiliser au moins une étiquette prédite comme contexte. Nous avons optimisé la taille de ce contexte et nous avons trouvé que la meilleure taille est de 5 étiquettes passées pour les 2 tâches, bien que les différences par rapport à l'utilisation d'autres tailles ne soient pas très grandes. La taille du contexte des mots est de 7 pour MEDIA (3 à gauche, 3 à droite plus le mot courant) et de 11 pour ATIS. Ces valeurs correspondent à celles utilisées dans (Dinarelli & Tellier, 2016a), dont nous avons repris aussi les autres paramètres que nous n'allons donc pas discuter dans cet article.

4.3 Résultats

Tous les résultats de cette section sont des moyennes de 10 expériences. Les plongements des mots et des étiquettes ont été appris une seule fois pour toutes les expériences. Pour que nos résultats soit comparables à ceux obtenus avec d'autres modèles neuronaux, nous avons utilisé les mêmes réglages et les mêmes hyper-paramètres (Vukotic *et al.*, 2015, 2016; Dinarelli & Tellier, 2016a), et nous montrons aussi les résultats obtenus avec les modèles *forward*, *backward* et bidirectionnel.

Nos différents résultats ont été obtenus avec un nombre incrémental de types d'entrées : i) seulement les mots (*Mots*) ; ii) mots et classes de mots disponibles avec le corpus (*Mots+Classes*) ; iii) mots, classes et une convolution sur les caractères des mots, comme celle utilisée dans (Collobert *et al.*, 2011) (*Mots+Classes+CC*). Notre réseau figure dans les tableaux sous le nom I-RNN_{deep}. *E-RNN*,

J-RNN et *I-RNN* sont les réseau d'Elman, de Jordan et la version basique du *I-RNN* utilisée dans (Dinarelli & Tellier, 2016a) respectivement.

Les résultats obtenus sur la tâche ATIS sont montrés dans le tableau 2. Pour cette tâche, nous nous comparons avec les résultats de (Vukotic *et al.*, 2016; Dinarelli & Tellier, 2016a), qui ont aussi employé des couches cachées de type LSTM et GRU. Les résultats pour la tâche MEDIA sont montrés dans le tableau 3, dans laquelle nous nous comparons aux modèles décrits dans (Vukotic *et al.*, 2015, 2016; Dinarelli & Tellier, 2016a; Hahn *et al.*, 2010). Pour cette tâche nous montrons aussi l'évaluation en terme de *Concept Error Rate* (CER) [3], puisque certains travaux n'utilisent que cette mesure. Tous les autres résultats sont en terme de mesure *F1*.

A notre connaissance, le meilleur résultat jusqu'à présent pour la tâche ATIS était une mesure F1 de 95.53 obtenue par (Vukotic *et al.*, 2016) avec un modèle GRU. Les meilleurs résultats sur la tâche MEDIA étaient la mesure F1 de 86.71 atteinte par (Dinarelli & Tellier, 2016a) avec le modèle *I-RNN*, et la CER de 10.2 mentionnée par (Hahn *et al.*, 2010), avec une combinaison de 6 modèles individuels par *ROVER* ($ROVER \times 6$ dans le tableau) (Fiscus, 1997).

Tous nos résultats égaux ou meilleurs que l'état-de-l'art sont mis en gras dans les tableaux. Notre intuition qu'un réseau profond peut mieux apprendre les représentations internes des entrées et leurs interactions semble ainsi confirmée empiriquement. Sur ATIS, bien que nos modèles soient significativement meilleurs, tous les modèles sont relativement proches. Sur MEDIA, les écarts sont plus importants et nos modèles sont largement meilleurs que tous les autres. Le résultat le plus remarquable est la CER de 9.8 obtenue avec I-RNN_{deep} *Mots+Classes+CC*. Ce modèle améliore de 0.4 l'état-de-l'art obtenu par (Hahn *et al.*, 2010) avec le modèle $ROVER \times 6$, qui combinent 6 modèles individuels.

Au delà de ces résultats quantitatifs, une analyse simple des sorties de nos modèles montre qu'ils sont réellement capables d'apprendre efficacement les dépendances entre étiquettes : nos modèles ne font en effet aucune erreur de segmentation *BIO* (un O suivi d'un I, par exemple). Ceci suggère que l'utilisation des plongements d'étiquettes, combinée à l'utilisation de plusieurs couches cachées pour séparer l'apprentissage des représentations internes et de leurs interactions, pallie effectivement au caractère local de la fonction de décision.

5 Conclusion

Dans cet article nous avons décrit un modèle profond pour l'étiquetage de séquences dans le contexte de la compréhension automatique de la parole. Notre modèle s'appuie sur le modèle *I-RNN* introduit précédemment. Celui-ci utilise des plongements d'étiquettes pour apprendre leurs dépendances. Notre variante profonde, employant plusieurs couches cachées, permet d'apprendre séparément les représentations internes des entrées et leurs interactions. Les résultats obtenus sur les deux tâches ATIS et MEDIA sont à l'état-de-l'art, et confirment l'efficacité de cette solution.

Références

BENGIO Y. (2012). Practical recommendations for gradient-based training of deep architectures.

3. Il s'agit d'un taux d'erreur, le plus petit est le meilleur.

CoRR.

BENGIO Y., DUCHARME R., VINCENT P. & JAUVIN C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, **3**, 1137–1155.

BONNEAU-MAYNARD H., AYACHE C., BECHET F., DENIS A., KUHN A., LEFÈVRE F., MOSTEFA D., QUGNARD M., ROSSET S. & SERVAN, S. VILANEAU J. (2006). Results of the french evalda-media evaluation campaign for literal understanding. In *LREC*, p. 2054–2059, Genoa, Italy.

CHO K., VAN MERRIENBOER B., GÜLÇEHRE Ç., BOUGARES F., SCHWENK H. & BENGIO Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*.

COLLOBERT R. & WESTON J. (2008). A unified architecture for natural language processing : Deep neural networks with multitask learning. In *Proceedings ICML*, p. 160–167 : ACM.

COLLOBERT R., WESTON J., BOTTOU L., KARLEN M., KAVUKCUOGLU K. & KUKSA P. (2011). Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, **12**, 2493–2537.

DAHL D. A., BATES M., BROWN M., FISHER W., HUNICKE-SMITH K., PALLETT D., PAO C., RUDNICKY A. & SHRIBERG E. (1994). Expanding the scope of the atis task : The atis-3 corpus. In *Proceedings of HLT Workshop* : ACL.

DINARELLI M. & TELLIER I. (2016a). Etude des reseaux de neurones recurrents pour etiquetage de sequences. In *Actes de la 23eme conference sur le Traitement Automatique des Langues Naturelles*, Paris, France : Association pour le Traitement Automatique des Langues.

DINARELLI M. & TELLIER I. (2016b). Improving recurrent neural networks for sequence labelling. *CoRR*.

DINARELLI M. & TELLIER I. (2016c). New recurrent neural network variants for sequence labeling. In *Proceedings of the 17th International Conference on Intelligent Text Processing and Computational Linguistics*, Konya, Turkey : Lecture Notes in Computer Science (Springer).

DUPONT Y., DINARELLI M. & TELLIER I. (2017). Label-dependencies aware recurrent neural networks. In *Proceedings of the 18th International Conference on Computational Linguistics and Intelligent Text Processing*, Budapest, Hungary : Lecture Notes in Computer Science (Springer).

ELMAN J. L. (1990). Finding structure in time. *COGNITIVE SCIENCE*, **14**(2), 179–211.

FISCUS J. G. (1997). A post-processing system to yield reduced word error rates : Recogniser output voting error reduction (ROVER). In *Proceedings of ASRU Workshop*, p. 347–352.

HAHN S., DINARELLI M., RAYMOND C., LEFÈVRE F., LEHEN P., DE MORI R., MOSCHITTI A., NEY H. & RICCARDI G. (2010). Comparing stochastic approaches to spoken language understanding in multiple languages. *IEEE TASLP*, **99**.

HE K., ZHANG X., REN S. & SUN J. (2015). Delving deep into rectifiers : Surpassing human-level performance on imagenet classification. In *IEEE ICCV*, p. 1026–1034.

HINTON G., DENG L., YU D., MOHAMED A., JAITLY N., SENIOR A., VANHOUCKE V., NGUYEN P., DAHL T. S. G. & KINGSBURY B. (2012). Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*, **29**(6), 82–97.

HOCHREITER S. & SCHMIDHUBER J. (1997). Long short-term memory. *Neural Comput.*, **9**(8), 1735–1780.

HUANG Z., XU W. & YU K. (2015). Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv :1508.01991*.

JORDAN M. I. (1989). Serial order : A parallel, distributed processing approach. In *Advances in Connectionist Theory : Speech*. Erlbaum.

LAFFERTY J., MCCALLUM A. & PEREIRA F. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML*, p. 282–289.

LAMPLE G., BALLESTEROS M., SUBRAMANIAN S., KAWAKAMI K. & DYER C. (2016). Neural architectures for named entity recognition. *arXiv preprint*.

MA X. & HOVY E. (2016). End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016*.

MESNIL G., HE X., DENG L. & BENGIO Y. (2013). Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding. In *Interspeech 2013*.

MIKOLOV T., KARAFIÁT M., BURGET L., CERNOCKÝ J. & KHUDANPUR S. (2010). Recurrent neural network based language model. In *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, p. 1045–1048.

MIKOLOV T., KOMBRINK S., BURGET L., CERNOCKY J. & KHUDANPUR S. (2011). Extensions of recurrent neural network language model. In *ICASSP*, p. 5528–5531 : IEEE.

VUKOTIC V., RAYMOND C. & GRAVIER G. (2015). Is it time to switch to word embedding and recurrent neural networks for spoken language understanding ? In *InterSpeech*, Dresde, Germany.

VUKOTIC V., RAYMOND C. & GRAVIER G. (2016). A step beyond local observations with a dialog aware bidirectional GRU network for Spoken Language Understanding. In *Interspeech*, San Francisco, United States.

YAO K., ZWEIG G., HWANG M.-Y., SHI Y. & YU D. (2013). : Interspeech.

Schémas Winograd en français: une étude statistique et comportementale

Pascal Amsili Olga Seminck
Laboratoire de Linguistique Formelle
Université Paris Diderot
amsili@linguist.univ-paris-diderot.fr,
olga.seminck@cri-paris.org

RÉSUMÉ

Nous présentons dans cet article une collection de schémas Winograd en français, adaptée de la liste proposée par Levesque *et al.* (2012) pour l'anglais. Les schémas Winograd sont des problèmes de résolution d'anaphore conçus pour être IA-complets. Nous montrons que notre collection vérifie deux propriétés cruciales : elle est robuste vis-à-vis de méthodes statistiques simples ("Google-proof"), tout en étant largement dépourvue d'ambiguïté pour les sujets humains que nous avons testés.

ABSTRACT

Winograd schemas in French : a statistical and behavioral study

We present in this paper a collection of Winograd schemas in French, adapted from the English collection proposed by Levesque *et al.* (2012). Winograd schemas constitute anaphora resolution problems meant to be AI-complete. We show that our collection has two crucial properties : it is robust regarding simple statistical techniques of resolution ("Google-proof") and basically non ambiguous for the human participants that were tested.

MOTS-CLÉS : Résolution d'anaphores, schémas Winograd, information mutuelle, test de Turing.

KEYWORDS: Anaphora resolution, Winograd Schemas, Mutual Information, Turing test.

1 Les schémas Winograd

Les schémas Winograd [1], proposés par Levesque *et al.* (2012), se présentent comme des cas particuliers du problème de résolution anaphorique, qui ne peuvent être résolus qu'en faisant appel à un raisonnement sur des connaissances du monde. Un schéma Winograd est constitué d'une paire de discours identiques à un mot (ou une expression) près, et qui comprennent une expression anaphorique à résoudre, dont l'antécédent change d'une version à l'autre. Ainsi, dans (1), la réponse naturelle pour la question formée avec le mot faible (on appellera ce mot *special*) est *Nicolas* (R0), alors que la question formée avec le mot lourd (appelé *alternate*) appelle plutôt la réponse *son fils* (R1).

(1) Nicolas n'a pas pu soulever son fils car **il** était trop ⟨faible/lourd⟩.
Qui était trop ⟨faible/lourd⟩ ? R0 : Nicolas
R1 : son fils

1. Ce nom a été proposé en référence à une paire d'exemples proposés par Winograd (1972) pour illustrer la difficulté du problème de la compréhension du langage naturel.

L'intérêt de disposer d'une telle paire de discours quasi identiques est de garantir que la structure linguistique n'empêche aucun syntagme nominal d'être un antécédent possible pour le pronom. Un schéma Winograd regroupe donc deux questions liées que nous appellerons *items* dans la suite.

Parce qu'ils nécessitent des capacités de raisonnement et d'intégration de connaissances encyclopédiques, les schémas Winograd ont été proposés comme un nouveau test d'intelligence artificielle dans l'esprit du fameux test de Turing (1950) (où un système automatique doit convaincre un juge humain qu'il est humain en discutant avec lui). En effet, les schémas Winograd évitent deux écueils du test de Turing : il n'est pas nécessaire pour les résoudre de faire semblant d'être humain (dans le test de Turing, la machine doit prétendre avoir tous les traits d'un être humain, avoir une enfance, une couleur préférée...), et de plus les schémas ne peuvent faire l'objet du genre de stratégie d'évitement qu'une machine pourrait mettre en œuvre (changement de sujet de la conversation, réponse par une autre question, etc.).

Suite à la publication des premiers schémas Winograd (Davis *et al.*, 2015), a été organisé en 2016 le premier *Winograd Schema Challenge* (Morgenstern *et al.*, 2016). Ce concours comportait au premier tour une série de problèmes de résolution de pronoms ambigus inspirés des schémas Winograd, mais qui pouvaient comporter plus de deux antécédents possibles, et qui n'étaient pas aussi contrôlés que les items issus de schémas Winograd (2).

(2) The storekeepers stayed in town to run their stores and lived in the rooms behind **them**.
A : storekeepers
B : stores
C : rooms

Selon les règles du concours, un second tour ne comportant que des vrais schémas Winograd était prévu au cas où le taux de réussite au premier tour dépassait les 90%, mais ça n'a pas été le cas en 2016. Parmi les 5 systèmes soumis au concours, le meilleur (Liu *et al.*, 2016) a obtenu un score de 58% d'exactitude, ce qui montre la difficulté de la tâche [2]. Par la suite, les mêmes auteurs ont élaboré une nouvelle version de leur système significativement meilleure, qui a obtenu le taux de 66,7%. Plusieurs autres travaux portant sur les schémas Winograd ont été publiés en dehors du concours, mais ils ne portent que sur un sous-ensemble des schémas (entre 5 et 25%) choisis pour le rôle qu'y jouent divers phénomènes comme la causalité, la pertinence, ou les relations de discours (Bailey *et al.*, 2015; Schüller, 2014; Sharma *et al.*, 2015, *i.a.*), et leur traitement implique une phase manuelle.

Deux propriétés des schémas Winograd nous intéressent : d'une part ils doivent nécessiter de l'intelligence artificielle, ce qui veut dire en particulier qu'ils ne doivent pas pouvoir être résolus avec des méthodes statistiques simples (on parle de *Google-proofness*), et d'autre part ils doivent être faciles pour les humains. Dans cet article, nous présentons une collection de schémas en français, adaptée de la collection originale de Davis *et al.* (2015), et rapportons les résultats de deux études : la première établit que nos schémas sont résistants à un test statistique basé sur l'information mutuelle ; la seconde est une étude comportementale qui permet de montrer que les humains n'ont pas de difficulté pour traiter correctement les items de notre collection. Notre test statistique simple ne parvient pas à dépasser 55% de taux de réussite alors que les humains obtiennent un taux de 93%.

2. Le nombre d'antécédents possibles étant de 2,33 en moyenne sur les 60 problèmes, le taux de réussite théorique pour une réponse au hasard était de 42,8%.

2 Collection de schémas en français

Notre collection a été produite en prenant comme point de départ les schémas de la collection anglaise de 144 schémas [3] (Davis *et al.*, 2015). Comme ces schémas ont été traduit en japonais et partiellement en chinois, nous avons décidé de les prendre comme point de départ pour permettre une comparaison multi-langue. Le principe a été de tenter une traduction directe des schémas, mais dans la plupart des cas il a fallu procéder à une adaptation, ne serait-ce qu'à cause des propriétés de genre et de nombre : les deux antécédents doivent être compatibles avec le pronom. Nous illustrons dans cette section quelques-uns des problèmes qui se sont présentés.

Dans le cas d'un item comme (3a), la traduction directe n'est pas possible, puisque le singulier *hair* se traduit en un pluriel (*cheveux*). Pour résoudre le problème, nous avons cherché un autre mot qui pouvait correspondre en nombre, et pour cet exemple particulier nous avons remplacé *cheveux* par *savon* (3b). Cet exemple illustre aussi le fait que nous avons cherché à rendre l'exemple aussi naturel que possible, au prix d'une distance plus grande entre l'original anglais et notre version.

(3) a. The drain is clogged with hair. **It** has to be ⟨cleaned/removed⟩.
b. Il y a du savon dans le siphon de douche. Il faut **le** ⟨retirer/nettoyer⟩.

Divers schémas présentaient des problèmes d'ordre lexical, comme par exemple l'item (4a) : la traduction directe de *indiscreet* donne *indiscrète*, qui se trouve avoir (au moins) deux sens distincts en français : en plus de désigner une personne qui révèle des secrets (le sens du mot *indiscreet*), le mot désigne une personne qui cherche à découvrir des secrets (le sens du mot *nosy*). Le schéma a donc dû être entièrement adapté :

(4) a. Susan knows all about Ann's personal problems because **she** is ⟨nosy/indiscreet⟩.
b. Sylvie est au courant de tous les problèmes personnels de Marie car **elle** est ⟨curieuse/bavarde⟩.

On peut encore évoquer le cas de l'item (5a) qui n'a pas pu être adapté, à cause de la préférence en français pour une structure infinitive pour exprimer une subordonnée finale portant sur le sujet de la principale : la version (5b) est fortement dégradée.

(5) a. Mary tucked her daughter Anne into bed, so that **she** could ⟨sleep/work⟩.
b. #Mary$_i$ a mis sa fille au lit (pour/afin/de sorte) qu'**elle**$_i$ dorme.

Le travail de traduction a été réalisé par plusieurs stagiaires, séparément puis en concertation, avant une validation par les auteurs. La traduction la plus naturelle (et si possible pas trop longue) a toujours été privilégiée, et les items sur lesquels il n'y avait pas d'accord entre les traducteurs ont été exclus. La collection finale comprend 107 schémas (214 items) et peut être librement téléchargée sur le site suivant : `http://www.llf.cnrs.fr/winograd-fr`. Chaque schéma comprend un lien vers le schéma original en anglais.

3. Au moment de nos expériences, il y avait 144 schémas, mais aujourd'hui la collection compte 146 schémas.

3 Test de robustesse statistique

L'enjeu des schémas Winograd est donc qu'il doit être nécessaire pour les résoudre de raisonner sur des connaissances du monde. Cela implique qu'il devrait être impossible d'avoir de bons résultats en appliquant de simples tests statistiques (Levesque *et al.* (2012) parlent de la propriété de *Google-proofness*). Ainsi, par exemple, on peut penser qu'un item comme (6) n'est pas *Google-proof*, parce qu'il suffirait d'exploiter des statistiques de co-occurrence en corpus (ou des requêtes à un moteur de recherche) pour déduire qu'il vaut mieux associer *vite* avec *bolide* qu'avec *bus scolaire*.

(6) Le bolide a dépassé le bus scolaire, parce qu'**il** roulait très vite.

Même si notre collection a été adaptée de celle de Davis *et al.* (2015) où les items les plus douteux ont été testés pour la *Google-proofness* (d'après la documentation fournie avec la collection, au moyen de comptages de nombre de *hits* donnés par Google), notre projet était de mener ce test pour la collection française dans son ensemble, de façon rigoureuse et systématique.

Pour cela, nous avons défini un test basé sur l'information mutuelle (Shannon & Weaver, 1949). L'information mutuelle permet de mesurer la dépendance entre deux variables aléatoires, et cette mesure a déjà été utilisée pour mesurer l'association entre deux mots : si les mots x et y sont mutuellement dépendants, la probabilité de leur cooccurrence $P(x, y)$ sera plus élevée que la probabilité d'observer les deux mots ensemble par hasard : $MI(x, y)$ sera positive (Church & Hanks, 1990) :

$$MI(x, y) = log_2 \left(\frac{P(x, y)}{P(x)P(y)} \right) \quad (1)$$

Pour tester la robustesse de nos schémas, nous avons cherché une méthode permettant de choisir entre les deux réponses possibles pour chaque item. Par exemple, pour l'item (7a), le choix doit se faire entre les deux réponses possibles (7b). Nous mesurons l'information mutuelle entre les mots importants de ces deux réponses, ce qui donne (7c).

(7) a. La sculpture est tombée de l'étagère car **elle** était trop encombrée.
b. (R0, alt) la sculpture était trop encombrée
(R1, alt) l'étagère était trop encombrée
c. MI(sculpture, encombrer) = 4,23
MI(étagère, encombrer) = 10,01

La façon la plus simple d'interpréter ces scores est de choisir la réponse qui maximise le score. Pour l'item (7a), cela donne la bonne réponse (*étagère*). Cependant, cette stratégie peut aussi échouer, par exemple sur l'autre item associé au même schéma :

(8) a. La sculpture est tombée de l'étagère car **elle** était trop lourde.
b. MI(sculpture, lourd) = 2,41
MI(étagère, lourd) = 4,03

Dans ce cas, la réponse correcte est *la sculpture*, mais ce n'est pas cette réponse qui a le plus fort score. On peut penser que les scores d'information mutuelle sont trop bas, ou plus simplement que c'est l'écart entre les deux scores qui est trop petit pour être significatif. Cela nous a conduit à introduire un seuil d'écart entre les scores, que nous faisons varier pour en étudier l'impact.

Il est important de préciser que notre méthode exige que les schémas aient une certaine forme : le score MI peut être calculé quand les deux réponses possibles comprennent les deux antécédents

possibles et le mot *special* (ou l'*alternate*). La collection comprend des schémas, comme (9a), qui n'ont pas cette propriété : la paire de réponses possibles est la même pour les deux items du schéma.

(9) a. Anna a ⟨mieux/moins bien⟩ réussi l'examen que son amie Lucy car **elle** avait beaucoup révisé.

b. (R0, ⟨spe/alt⟩) Anna a beaucoup révisé
(R1, ⟨spe/alt⟩) Lucy a beaucoup révisé

Nous avons exclu de notre analyse les 30 schémas de ce type, ainsi que 2 autres schémas où le jeu de réponses est différent selon l'item associé.

Il convient de noter de plus que dans une grande partie des schémas on trouve des noms propres comme antécédents possibles, ce qui donne par construction des items *Google-proof*, puisque dans le cas général les co-occurrences de noms propres avec des noms communs sont aléatoires. Néanmoins, quand il nous était possible de mesurer les fréquences pour ces items, nous les avons inclus dans nos scores. Au total, nous avons mesuré l'information mutuelle pour 90 schémas (180 items).

Pour estimer MI nous avons utilisé les mesures de fréquence (sans lissage) du corpus FrWaC (Baroni *et al.* (2009), 1,6 millards de mots du domaine .fr d'Internet). Sauf quand la différence portait sur la flexion, nous avons utilisé les lemmes pour les mesures, et nous avons choisi les têtes dans le cas d'expressions poly-lexicales. Nous avons choisi d'utiliser un corpus à taille fixe, au lieu des comptages Google, car ces derniers ne sont pas stables (Lapata & Keller, 2005). La Figure 1 donne le taux de réussite de notre méthode pour les différents seuils. Sur les 180 items de départ, 49 n'ont pas pu recevoir un score car une des mesures de fréquence/co-occurrence était nulle.

Seuil	Nb items	Exactitude	Couverture
Aucun	131	0,55	0,40
Δ 0,5	95	0,59	0,31
Δ 1,0	73	0,62	0,25
Δ 1,5	59	0,64	0,21
Δ 2,0	38	0,68	0,14
Δ 2,5	30	0,70	0,12
Δ 3,0	25	0,68	0,09
Δ 3,5	18	0,67	0,07
Δ 4,0	15	0,60	0,05

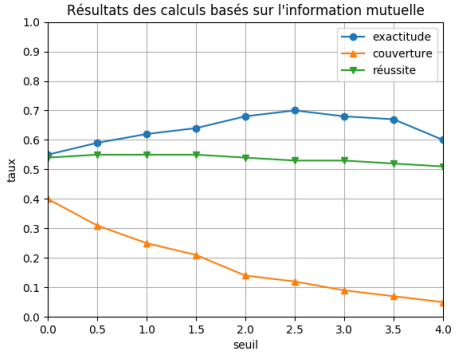


FIGURE 1 – Résultats des calculs d'information mutuelle. L'*exactitude* est le nombre de réponses correctes pour les items auxquels la méthode s'applique, alors que la *couverture* correspond au nombre de réponses correctes divisé par le nombre total d'items (180). Par *réussite* on désigne le taux de réussite théorique qui serait obtenu par une stratégie consistant à utiliser les scores d'information mutuelle quand la différence est supérieure au seuil, et à répondre au hasard dans les autres cas.

Rappelons qu'une réponse au hasard à l'ensemble des question donnerait un taux de réussite autour de 50%. Par conséquent, quand aucun seuil n'est appliqué, le taux obtenu (55%) n'est pas vraiment différent du hasard. Quand le seuil de Δ 2,5 est appliqué, le taux de réussite monte à 70%, mais le nombre d'items auxquels la méthode s'applique est extrêmement petit : moins de 15%. La courbe *réussite* montre que quel que soit le seuil choisi, on ne parvient pas à dépasser 55% de réussite globale.

On peut donc conclure que notre collection est *Google-proof*, au moins dans le sens où nous avons interprété ce terme.

4 Résolution pour les sujets humains

Pour que les schémas jouent leur rôle de test d'IA, ils ne doivent pas poser de difficulté pour les humains. Pour vérifier que notre collection était correcte de ce point de vue, nous avons lancé une expérience en ligne (sur *Ibex Farm*) où nous avons présenté les items aux participants, en leur demandant de répondre à la question. Nous avons garanti que chaque participant était exposé à un seul des deux items de chaque schéma, et contrebalancé l'ordre de présentation. Au total, nous avons testé 203 items [4] sur 22 participants. Nous avons exclu de l'analyse les réponses associées à un temps de réaction inférieur à $1''$ ou supérieur à $60''$.

La réussite moyenne de nos participants est de 93,6%, ce qui est très proche du chiffre de 92% établi pour les schémas anglais (Bender, 2015). Ce taux moyen cache une certaine disparité que l'on peut mettre en évidence en étudiant le taux moyen par item. La Figure 2 donne le nombre d'items en fonction de leur taux moyen de réussite. Si beaucoup d'items sont très faciles (145 items ont un taux élevé de 80 à 100%), il reste 19 items dont le taux est inférieur à 80%, et nous avons étudié individuellement ces items problématiques. Nous proposons de distinguer trois classes de schémas difficiles.

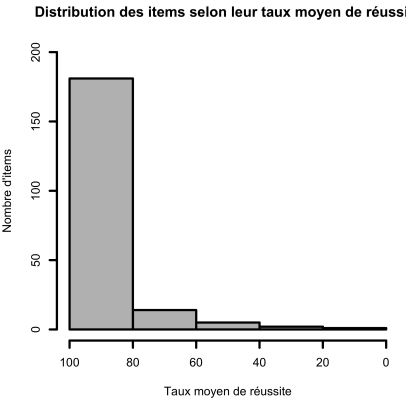


FIGURE 2 – Le taux de réussite moyen pour les participants humains est de 93,6% sur l'ensemble des items testés. L'analyse par item montre que la plupart des items ont un taux très élevé, et sont donc très faciles.

La première classe est constituée d'items ambigus. Ce genre de schéma ambigu est une conséquence du caractère artificiel de la construction des schémas Winograd : en cherchant un schéma, on perd de vue que dans une situation où seule une version est présentée, l'item peut être ambigu. Le schéma (10) illustre ce cas (les deux items sont mal réussis : *special* : 50%, *alternate* : 18%).
L'ambiguïté concerne 7 items sur les 19 étudiés.

(10) Pierre et Marc sont poursuivis pour diffamation. Pierre a écrit dans leur livre plusieurs faux témoignages que Marc a colportés. **Il** aurait dû être plus ⟨prudent/honnête⟩.

Une deuxième classe d'items à problème comprend des items où un connecteur discursif joue un rôle important. Dans le schéma (11) le connecteur *alors que* introduit une opposition. Or, on constate que cette opposition n'est pas toujours interprétée : la version *special* du schéma est réussie à 70% et l'*alternate* à 75%. Ce problème de connecteur concerne 6 items.

4. Le nombre d'items est différent de celui la collection finale, car cette expérience a été conduite avant la toute dernière version de notre collection.

(11) Les pompiers sont arrivés ⟨avant/après⟩ les policiers alors qu'**ils** venaient de plus loin.

Une dernière catégorie concerne des items qui ont l'air d'être « trop construits », pas naturels. Par exemple le schéma (12) pour lequel le *special* a un taux de réussite à 50% et l'*alternate* 67%. Nous avons classé au total 6 items dans cette dernière catégorie.

(12) Pierre jouait aux cartes avec Adam qui menait au score. Si la chance d'Adam n'avait pas tourné **il** aurait ⟨perdu/gagné⟩.

En excluant ces items difficiles de notre collection, nous obtiendrions certainement un taux de réussite supérieur à 93,6%, mais ce score en tant que tel, compte tenu des erreurs inhérentes à ce genre d'étude comportementale, peut déjà être considéré comme très proche d'un score parfait.

5 Conclusion

Nous étudions dans cet article la collection de schémas Winograd en français que nous avons adaptée de la collection anglaise (Levesque *et al.*, 2012; Davis *et al.*, 2015). Cette étude montre que nos schémas sont *Google-proof* : un test statistique simple ne permet pas de dépasser un taux de réussite de 55%, à comparer avec le taux de 93,6% obtenu par les participants à une étude comportementale sur les mêmes données.

À partir de ces résultats nous pouvons envisager de produire une collection encore plus *Google-proof* et encore plus naturelle pour les humains, en ôtant les items repérés comme à la limite dans les deux cas et en ajoutant de nouveaux schémas, sans faire appel à la traduction. Nous envisageons aussi de profiter de cette occasion pour préparer des items en français pour le prochain *Winograd Schema Challenge*.

Remerciements

Nous remercions nos stagiaires du cursus Linguistique Informatique à l'Université Paris Diderot, Sarah Ghumundee, Biljana Knežević, et Nicolas Bénichou, pour leur aide dans la préparation des items et le calcul des scores d'information mutuelle. Nous remercions également les étudiants de la Licence Frontières du Vivant au Centre de Recherches Interdisciplinaires, Ryan Hunt, Dara Nguyen et Hugo Taquet pour leur étude comportementale préliminaire. Nous remercions aussi les relecteurs de TALN 2017 qui nous ont aidé à améliorer cet article, ainsi que les relecteurs et les participants de l'atelier CORBON associé à EACL 2017 (Valence) où une version préliminaire de ce travail a été présentée (Amsili & Seminck, 2017). Ce projet est soutenu en partie par l'École Doctorale Frontières du Vivant — Programme Bettencourt.

Références

AMSILI P. & SEMINCK O. (2017). A Google-proof collection of French Winograd Schemas. In *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017)*, p. 24–29, Valencia : EACL Association for Computational Linguistics.

BAILEY D., HARRISON A., LIERLER Y., LIFSCHITZ V. & MICHAEL J. (2015). The winograd schema challenge and reasoning about correlation. In *In Working Notes of the Symposium on Logical Formalizations of Commonsense Reasoning*.

BARONI M., BERNARDINI S., FERRARESI A. & ZANCHETTA E. (2009). The wacky wide web : a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, **43**(3), 209–226.

BENDER D. (2015). Establishing a human baseline for the winograd schema challenge. In *MAICS*, p. 39–45.

CHURCH K. W. & HANKS P. (1990). Word association norms mutual information, and lexicography. *Computational Linguistics, Volume 16, Number 1, March 1990*.

DAVIS E., MORGENSTERN L. & ORTIZ C. (2015). A collection of winograd schemas. `http://www.cs.nyu.edu/faculty/davise/papers/WinogradSchemas/WS.html` – Web page collecting 144 Winograd pairs, with comments and references.

LAPATA M. & KELLER F. (2005). Web-based models for natural language processing. *ACM Transactions on Speech and Language Processing (TSLP)*, **2**(1), 3.

LEVESQUE H., DAVIS E. & MORGENSTERN L. (2012). The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

LIU Q., JIANG H., LING Z.-H., ZHU X., WEI S. & HU Y. (2016). Combing context and commonsense knowledge through neural networks for solving winograd schema problems. *arXiv preprint arXiv :1611.04146*.

MORGENSTERN L., DAVIS E. & ORTIZ JR. C. L. (2016). Planning, executing, and evaluating the winograd schema challenge. *AI Magazine*, **37**(1), 50–54.

SCHÜLLER P. (2014). Tackling winograd schemas by formalizing relevance theory in knowledge graphs. In *Fourteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

SHANNON C. E. & WEAVER W. (1949). The mathematical theory of information.

SHARMA A., VO N. H., ADITYA S. & BARAL C. (2015). Towards addressing the winograd schema challenge-building and using a semantic parser and a knowledge hunting module. In *Proceedings of Twenty-Fourth International Joint Conference on Artificial Intelligence. AAAI*.

TURING A. M. (1950). Computing machinery and intelligence. *Mind*, **59**(236), 433–460.

WINOGRAD T. (1972). Understanding natural language. *Cognitive psychology*, **3**(1), 1–191.

Critères numériques dans les essais cliniques : annotation, détection et normalisation

Natalia Grabar[1] Vincent Claveau[2]
(1) CNRS, Univ. Lille, UMR 8163 - STL - Savoirs Textes Langage, F-59000 Lille, France
(2) CNRS, IRISA, Rennes, France
natalia.grabar@univ-lille3.fr, vincent.claveau@irisa.fr

RÉSUMÉ
Les essais cliniques sont un élément fondamental pour l'évaluation de nouvelles thérapies ou techniques de diagnostic, de leur sécurité et efficacité. Ils exigent d'avoir un échantillon convenable de la population. Le défi consiste alors à recruter le nombre suffisant de participants avec des caractéristiques similaires pour garantir que les résultats des essais sont bien contrôlés et dus aux facteurs étudiés. C'est une tâche difficile, effectuée essentiellement manuellement. Comme les valeurs numériques sont une information très fréquente et importante, nous proposons un système automatique qui vise leur extraction et normalisation.

ABSTRACT
Numerical criteria in clinical trials : annotation, detection and normalization.

Clinical trials are very important for the evaluation of new therapies or diagnosis procedures, of their security and efficency. They must rely on convenient sample of the population. the challenge is then to recruit sufficient number of participants with similar characteristics to guarantee that the results are controlled and due tu the factors studied. This is a difficult task, mainly performed manuelly. Because numerical values are frequent and provide important information, we propose an automatic system that aims their extraction and normalization.

MOTS-CLÉS : Valeurs numériques, extraction d'information, normalisation, domaine médical.

KEYWORDS: Numerical values, information extraction, normalization, medical area.

1 Introduction

Les essais cliniques sont un élément fondamental pour l'évaluation de nouvelles thérapies ou techniques de diagnostic, de leur sécurité et efficacité. Ils se basent sur l'inférence statistique et exigent d'avoir un échantillon convenable de la population. Le défi consiste donc à recruter un nombre suffisant de participants avec des caractéristiques spécifiques pour garantir que les résultats des essais sont bien contrôlés et dus aux facteurs étudiés et non pas à d'autres facteurs ou au hasard. Pour cette raison, les essais cliniques doivent définir précisément la population d'intérêt en indiquant les critères d'inclusion et d'exclusion. Ce sont par exemple l'âge, le genre, l'histoire médicale, les traitements, les biomarqueurs, etc. La source principale d'information sur les patients est le dossier patient, qui peut être en version papier ou électronique. Ces dossiers contiennent essentiellement du texte non structuré en langue naturelle. Pour cette raison, ce sont surtout les opérateurs humains (ie, investigateurs principaux et assistants de recherche clinique) qui sont capables de détecter efficacement

les patients éligibles (Campillo-Gimenez *et al.*, 2015). Il va sans dire que c'est une tâche laborieuse et longue, qui constitue un écueil dans le recrutement des patients. Il est en effet commun que les personnes qui mènent les essais cliniques n'arrivent pas à recruter les patients à temps et en un coût raisonnable (Fletcher *et al.*, 2012). Ainsi, à cause de la complexité croissante des protocoles et du processus de recrutement, presque la moitié de retards dans les essais cliniques sont dus à la difficulté de recrutement, alors que le pourcentage des essais qui terminent les recrutement à temps est très bas : 18 % en Europe, 17 % en Asie, 15 % en Amérique Latine, 7 % aux Etats-Unis (Center Watch, 2013). S'il existe des travaux qui proposent de développer les systèmes de recrutement électronique (Cuggia *et al.*, 2011), il reste difficile de générer une représentation comparable des informations contenues dans le texte libre des dossiers cliniques et les critères d'éligibilité (Embi *et al.*, 2005; Olasov & Sim, 2006; Ross *et al.*, 2010; Tu *et al.*, 2011; Pressler *et al.*, 2012; Shivade *et al.*, 2014). Pour cette raison, ces systèmes exploitent essentiellement les données structurées, en laissant de côté le texte narratif, et reposent ensuite sur l'expertise manuelle. Nous pensons que les méthodes de TAL peuvent être efficacement exploitées pour cette tâche. Nous proposons de nous concentrer sur les informations numériques ou quantifiées, qui sont fréquentes et importantes pour le recrutement des patients.

Il existe quelques travaux sur l'extraction des informations numériques à partir des données non structurées. Dans la langue générale, comme les textes de la Toile, un système pour l'extraction et approximation de valeurs numériques (comme la hauteur ou le poids) a été proposé (Davidov & Rappaport, 2010). Il repose sur des patrons de relation et WordNet. Les termes similaires sont d'abord obtenus grâce à WordNet et les valeurs numériques liées sont extraites. Ensuite, les objets extraits sont comparés. Ce système montre une précision moyenne exacte de 0,84 et inexacte de 0,72. Un autre travail propose deux systèmes d'extraction qui demande toutefois une supervision manuelle (Madaan *et al.*, 2016). Différents cas sont pris en charge, comme la durée de vie, l'inflation, la production d'électricité. Plus proche de nos besoins, quelques études sont à noter dans le domaine médical. A partir d'un petit ensemble de données cliniques annotées en français, un modèle CRF est entraîné pour reconnaître trois types d'entités (concepts, valeurs, unités). Ensuite, un système à base de règles est créé pour calculer les relations entre ces entités (Bigeard *et al.*, 2015). Des travaux similaires ont aussi été proposés pour l'anglais sur des données cliniques (Sarath *et al.*, 2016) et sur des rapports de radiologie cardiaque (Nath *et al.*, 2016).

Ces différents travaux obtiennent des résultats difficilement comparables entre eux du fait des différences de nature des textes traités, des informations recherchées et des protocoles d'évaluation. Il nous a ainsi semblé nécessaire de développer un jeu de données qui soit spécifiquement dédié à notre cas d'usage final (section 2). Nous nous intéressons ainsi aux protocoles des essais cliniques en anglais. Ils comportent en effet une très grande variété d'expressions numériques et de critères quantifiés. En plus de l'extraction, pour laquelle nous reprenons des techniques similaires à l'état de l'art (section 3), nous proposons une méthode pour aider à la normalisation des unités (section 4).

2 Données et annotations

2.1 Essais cliniques

211 438 protocoles d'essais cliniques, en anglais, ont été collectés (en décembre 2016 à partir du site `www.clinicaltrials.gov`). Parmi les différentes sections composant ces protocoles, nous nous intéressons uniquement aux critères d'inclusion et d'exclusion, ce qui constitue un corpus de

κ/phase	A1 vs. A2	A1 vs. A3	A2 vs. A3	A1 vs. A4
phase 1 global	0.78	0.51	0.47	-
détails (C, U, V)	0.72, 0.58, 0.83	0.34, 0.58, 0.57	0.22, 0.56, 0.7	
phase 2 global	-	-	-	0.92
détails (Q, T)				0.84 0.93

TABLE 1 – Accord inter-annotateur (κ de Cohen) sur les deux phases d'annotation

plus de 2 millions de phrases. Trois phrases tirées de ce corpus sont données ci-après :

— *Absolute neutrophil count $\geq$ 1,000 cells/μl at time of enrollment.*
— *Exclude if T3 uptake is less than 19% ; T4 less than 2.9 ((g/dL) ; free T4 index is less than 0.8.*
— *Elevated bilirubin within the past two years.*

2.2 Annotations de référence

1 500 phrases ont été sélectionnées aléatoirement et annotées chacune par plusieurs annotateurs. L'annotation a eu lieu en deux temps. Dans un premier temps, trois annotateurs (A1, A2, A3), avec des profils différents, ont annoté les phrases en repérant les valeurs (V) et unités (U) exprimées, ainsi que les concepts (C) auquel ces mesures se rapportent. Dans un second temps, deux annotateurs (A1, A4) ont annotés ces mêmes phrases en repérant les critères exprimés avec des quantificateurs (Q) non numériques (*normal, low, clinically important...*), ainsi que les expressions de validité temporelle (T). A2 est médecin, A1, A3, A4 sont informaticiens ; A1 et A2 ont une bonne expérience des données cliniques.

Ainsi, dans les exemples précédents, les annotations seraient :

— C : *Absolute neutrophil count, T3 uptake, T4, free T4 index, bilirubin* ;
— V : *$\geq$ 1,000, less than 19, less than 2.9, less than 0.8* ;
— U : *cells/μl, %, g/dL* ;
— Q : *Elevated* ;
— T : *at time of enrollement, within the past two years*

Pour évaluer l'accord inter-annotateur, nous utilisons la mesure du κ de Cohen (Artstein & Poesio, 2008) pour chaque paire d'annotateurs. Le résultats sont reportés dans le tableau 1 pour chacune des phases d'annotation (C, V, U d'une part et Q, T d'autre part). de manière globale et en détail pour chacun des types d'annotations. Ils montrent un très fort agrément sur la phase 2 entre les deux annotateurs. Concernant la phase 1, les résultats sont plus contrastés : A1 et A2 obtiennent le plus fort agrément, ce qui souligne l'intérêt d'une connaissance du domaine médical, en tant que médecin ou informaticien. Un examen des cas de désaccord montre qu'une grande part est due à l'annotation des concepts sur lesquels portent les mesures, notamment pour les accords impliquant A3. Les valeurs numériques sont en revanche bien reconnues, quelle que soit la formation initiale de l'annotateur.

Sur la base des annotations de A1, A2 et A4, une annotation finale est obtenue par discussion des désaccords jusqu'à l'obtention d'un consensus. Ces données annotées et le guide d'annotation sont accessibles à `people.irisa.fr/Vincent.Claveau/Corpus`.

3 Annotation automatique

Dans cette section, nous décrivons l'utilisation du jeu de données pour entraîner des modèles CRF pour détecter, dans de nouveaux textes cliniques, des concepts mesurés (C), soit précisément (avec valeur V et unité U), soit avec un quantificateur (Q), et le temps associés à cette mesure (T).

3.1 Méthodologie

Pour produire ces modèles d'annotation automatique, nous utilisons les champs aléatoires conditionnels (Conditional Random Fields, CRF) (Lafferty *et al.*, 2001). Ces modèles graphiques non dirigés permettant d'apprendre les distributions de probabilités d'annotations y sachant les observations x. Très populaires en TAL par leur capacité à prendre en compte l'aspect séquentiel du texte et à tirer parti de descriptions riches du texte, ils sont des outils standard en extraction d'information, reconnaissance d'entités nommées, étiquetage, etc. (Wang *et al.*, 2006; Pranjal *et al.*, 2006; Constant *et al.*, 2011; Raymond & Fayolle, 2010, inter alia).

Dans notre cas, nous adoptons un schéma d'étiquetage BIO : les CRF doivent assigner une étiquette à chaque mot selon qu'il dénote les éléments recherchés (C, V, U, T, Q) ou non (O) ; ces éléments pouvant être composés de plusieurs mots, nous adoptons le schéma d'annotation BIO (par exemple, B-C indique le premier mot d'un concept, I-C les mots suivants de ce concept, O pour un mot n'appartenant à aucune des cinq informations recherchées).

Les textes dont décrits par les mots-formes, les lemmes, les parties du discours (ces informations et la tokénisation sont obtenues par Treetagger (Schmid, 1994)), des indices graphémiques (présence de majuscules, de chiffres, de symboles...). Nous autorisons les CRF à regarder un contexte de 4 mots avant et après le mot courant. L'implémentation CRF que nous utilisons est Wapiti (Lavergne *et al.*, 2010). Pour mesurer les performances de cette tâche d'annotation automatique, nous évaluons les résultats obtenus sur une partie du jeu annoté manuellement (cf. section 2.2) en termes de taux d'erreur mot (pourcentage de mots recevant une étiquette différente de celle du jeu de référence) et de F-mesure (Sebastiani, 2002).

3.2 Résultats

La figure 1 présente les résultats en taux d'erreurs mots et la F-mesure de chacune des catégories (concept, valeur, unité, quantificateur, temps) en fonction de la quantité de données d'apprentissage utilisée. Pour chaque taille de jeu d'entraînement, le phrases d'entraînement sont tirées au hasard, le reste des données annotées sert de jeu de test, et ce processus est répété cinq fois puis les résultats moyennés pour éviter tout effet lié à la variabilité des petits jeux d'entraînement. Le taux d'erreurs global décroît continuellement, ce qui suggère que des données d'apprentissage supplémentaires permettraient d'améliorer encore les résultats. L'analyse des résultats par catégorie permet de comprendre que ce sont les concepts et les quantificateurs qui bénéficieraient le plus de ces données supplémentaires puisque leur rappel est bas et la détection des autres catégories atteint un plateau.

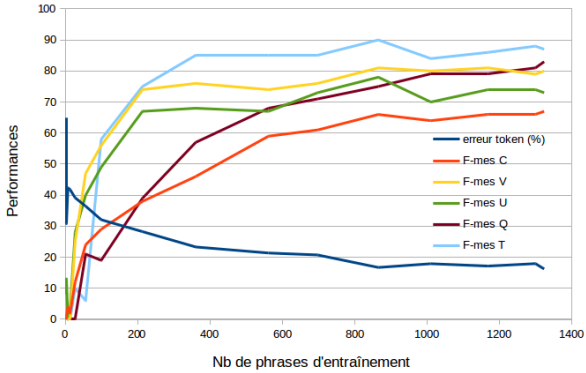


FIGURE 1 – Performance de l'annotation par CRF (globalement en terme de taux d'erreur mot, et par catégorie en terme de F-mesure) selon la quantité de données d'entraînement (nombre de phrases).

4 Vers une normalisation des unités

Une fois que les critères d'éligibilité recherchés sont détectés, il est nécessaire de les normaliser pour pouvoir les utiliser pour sélectionner des patients. C'est notamment le cas pour les unités qui apparaissent sous de nombreuses variantes :

— variation typographique (*cells/mL* vs. *cells / ml*) ;
— flexion (*cell/mL* vs. *cells/mL*) ;
— acronymie (*xULN* vs. *time Upper Limit of Normal*) ;
— variation paradigmatique (*cigarettes per day* vs. *packs per week*).

4.1 Détection des variations

Nous proposons une approche pour identifier, pour une expression d'unité donnée, toutes ses variations possibles et les ordonner par similarité. L'idée est d'utiliser les concepts associés à ces unités avec des techniques issues de la recherche d'information. Plus précisément, notre approche est la suivante. Tout d'abord, un modèle CRF, appris sur les données de référence, est utilisé pour annoter l'ensemble du corpus d'essais cliniques (i.e. les 2 millions phrases des critères d'éligibilité). Ensuite, chaque unité détectée est collectée ; celles avec exactement la même graphie sont groupées. Les concepts détectés avec ces unités sont décomposés en n-grammes et pondérés (TF-IDF). Chaque graphie d'unité est ainsi représentée par un vecteur de n-grammes (de concepts). La proximité entre deux graphies d'unités est alors définie par la proximité entre leurs vecteurs, et mesurée par un cosinus.

4.2 Résultats

Vingt graphies d'unités ont été tirées au hasard parmi celles détectées par CRF sur les essais cliniques. Pour chacune, nous évaluons la précision de la liste de ses n plus proches voisins (ppv) selon la similarité cosinus. Deux types de précision sont données : la précision 'unité identique' n'accepte

	P@1 (%)	P@10 (%)	P@1_freq (%)
unité identique	50	45	45
unité équivalente	100	84.5	95

TABLE 2 – Précisions des plus-proches voisins (moyenne sur 20 expressions d'unités choisies aléatoirement) au rang 1 et 10, et précision sur le voisin le plus fréquent parmi les 10 plus proches. La précision est calculée selon deux critères : le voisin est une variante linguistique de la même unité (acronyme, flexion...) ou une unité équivalente (incluant donc les variantes paradigmatiques).

comme valides que les variantes linguistiques d'une même unité (e.g. *cm*^{3} et *cubed centimeters*), alors que la précision 'unité équivalente' accepte les variantes avec les mêmes dimensions (au sens physique, comme *cm*^{3} and *liters*) et suppose donc qu'une conversion est possible et serait nécessaire pour normaliser plus avant. Pour évaluer l'intérêt de cette approche pour normaliser les unités, nous mesurons aussi la précision de l'expression la plus fréquente parmi les 10 ppv (P@1_freq).

Les résultats sont fournis dans le tableau 2. Les bonnes précisions obtenues montrent l'intérêt de cette approche simple pour collecter les variantes. Plus intéressant, ce ne sont pas seulement de simples variantes linguistiques qui sont capturées, mais aussi les variantes paradigmatiques. La très forte P@1_freq en 'unité équivalente' montre le potentiel de cette approche pour aider à normaliser les mesures cliniques détectées avec une unité standard. Voici quelques exemples d'expressions d'unités ainsi regroupées ; celle soulignée est la requête, les autres sont celles retrouvées, par ordre de similarité cosinus décroissante, par notre approche et la plus fréquente est en gras :

— pack of cigarette per day : **cigarette per day** ; cigarette a day ; cigarette daily ; cigarette or equivalent per day ; cigarette/day ; pack per day ; cig/day ; pack/year ; pipe per day ; pack year ; cigarette per week ;

— kgs : pound ; Kg ; lbs ; % of ideal body weight ; lb ; % of ideal weight ; **kg** ; % body weight ; % of their ideal body weight ; % of body weight ; kilogram.

5 Conclusion et perspectives

Nous proposons une approche automatique pour extraire les valeurs numériques et les informations associées (unité, concept, qualificateur, temporalité) à partir des critères d'inclusion des essais cliniques. Les données traitées sont en anglais. Une annotation de référence est obtenue grâce aux annotations effectuées par plusieurs annotateurs. L'accord inter-annotateur va de 0,47 à 0,93, selon les étapes d'annotation et l'expérience des annotateurs. Le système d'extraction exploite ces données de référence et les CRF, et montre une bonne performance : entre 0,65 (concepts) et 0,88 (temps). Notons que le taux d'erreurs diminue avec l'augmentation des données de référence utilisées. Une deuxième étape de notre approche propose de normaliser les unités extraites et montre une précision entre 45 et 100 %. Nous pensons que notre travail peut permettre de traiter et de proposer une représentation des critères d'inclusion, ce qui peut fournir une bonne base pour le recrutement des patients pour les essais cliniques.

Nous avons plusieurs perspectives à ce travail. Nous allons construire un système de conversion des unités extraites en unités standard, comme par exemple *cell/mm3* au lieu de *cell/cm3*, ou *ml* au lieu de *l* ou μl. Concernant les annotations de référence, nous allons préparer un jeu d'annotations plus grand. En effet, nous avons montré qu'il est possible d'obtenir une amélioration des résultats

avec des données d'entraînement plus volumineuses. Le modèle CRF et les règles de conversion seront exploités pour transformer les informations extraites en une représentation formelle des critères d'inclusion et d'exclusion. Selon les travaux de l'état de l'art, il apparaît que ceci correspond actuellement à un défi réel lorsque l'on travaille avec les données en langue naturelle. Finalement, le modèle d'annotation automatique de même que les systèmes de normalisation et de conversion seront appliqués aux essais cliniques et aux dossiers patients pour détecter les patients candidats et les proposer au recrutement. Les hôpitaux en France et au Brésil seront les lieux de test de ce système. Ainsi, nous allons pouvoir appliquer et tester notre approche dans un contexte multilingue. Plus spécifiquement, le modèle d'annotation sera adapté à ces deux langues (français et portugais brésilien), tandis que nous pensons que la présentation des informations numériques respecte les conventions similaires dans différentes langues dans le domaine médical.

Remerciements

Ce travail a été partiellement réalisé et financé dans le cadre du projet Franco-Brésilien CNRS-CONFAP FIGTEM et a bénéficié d'une aide de l'État attribuée au labex COMIN LABS et gérée par l'Agence Nationale de la Recherche au titre du programme « Investissements d'avenir » portant la référence ANR-10-LABX-07-01. Nous tenons également a remercier Guillaume Bouzillé (Univ. Rennes 1), Lucas Oliveira (PUCPR) et Claudia Moro (PUCPR) pour leur aide dans la préparation des données.

Références

ARTSTEIN R. & POESIO M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, **34**(4).

BIGEARD E., JOUHET V., MOUGIN F., THIESSARD F. & GRABAR N. (2015). Automatic extraction of numerical values from unstructured data in EHRs. In *MIE (Medical Informatics in Europe) 2015*, Madrid, Spain.

CAMPILLO-GIMENEZ B., BUSCAIL C., ZEKRI O., LAGUERRE B., LE PRISÉ E., DE CREVOISIER R. & CUGGIA M. (2015). Improving the pre-screening of eligible patients in order to increase enrollment in cancer clinical trials. *Trials*, **16**(1), 1–15.

CENTER WATCH (2013). *State of the Clinical Trials Industry : A Sourcebook of Charts and Statistics*. Rapport interne, Center Watch.

CONSTANT M., TELLIER I., DUCHIER D., DUPONT Y., SIGOGNE A. & BILLOT S. (2011). Intégrer des connaissances linguistiques dans un CRF : Application à l'apprentissage d'un segmenteur-étiqueteur du français. In *Traitement Automatique du Langage Naturel (TALN'11)*, Montpellier, France.

CUGGIA M., BESANA P. & GLASSPOOL D. (2011). Comparing semi-automatic systems for recruitment of patients to clinical trials. *International Journal of Medical Informatics*, **80**(6), 371–88.

DAVIDOV D. & RAPPAPORT A. (2010). Extraction and approximation of numerical attributes from the web. In *48th Annual Meeting of the Association for Computational Linguistics*, p. 1308–1317.

EMBI P., JAIN A., CLARK J. & HARRIS C. (2005). Development of an electronic health record-based clinical trial alert system to enhance recruitment at the point of care. In *AMIA Symposium*, p. 231–35.

FLETCHER B., GHEORGHE A., MOORE D., WILSON S. & DAMERY S. (2012). Improving the recruitment activity of clinicians in randomised controlled trials : A systematic review. *BMJ Open*, **2**(1), 1–14.

LAFFERTY J., MCCALLUM A. & PEREIRA F. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning (ICML)*.

LAVERGNE T., CAPPÉ O. & YVON F. (2010). Practical very large scale CRFs. In *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, p. 504–513 : Association for Computational Linguistics.

MADAAN A., MITTA A., MAUSAM, RAMAKRISHNAN G. & SARAWAGI S. (2016). Numerical relation extraction with minimal supervision. In *Thirtieth AAAI Conference on Artificial Intelligence*.

NATH C., ALBAGHDADI M. & JONNALAGADDA S. (2016). A natural language processing tool for large-scale data extraction from echocardiography reports. *PLoS One*, **11**(4), 153749–64.

OLASOV B. & SIM I. (2006). Ruleed, a web-based semantic network interface for constructing and revising computable eligibility rules. In *AMIA Symposium*, p. 1051.

PRANJAL A., DELIP R. & BALARAMAN R. (2006). Part Of speech Tagging and Chunking with HMM and CRF. In *Proceedings of NLP Association of India (NLPAI) Machine Learning Contest*.

PRESSLER T., YEN P., DING J., LIU J., EMBI P. & PAYNE P. (2012). Computational challenges and human factors influencing the design and use of clinical research participant eligibility pre-screening tools. *BMC Med Inform Dec Mak*, **12**, 47.

RAYMOND C. & FAYOLLE J. (2010). Reconnaissance robuste d'entités nommées sur de la parole transcrite automatiquement. In *Actes de la conférence Traitement Automatique des Langues Naturelles*, Montréal, Canada.

ROSS J., TU S., CARINI S. & SIM I. (2010). Analysis of eligibility criteria complexity in clinical trials. In *Summit on Translational Bioinformatics*, p. 46–50.

SARATH P. R., MANDHAN S. & NIWA Y. (2016). Numerical atrribute extraction from clinical texts. *CoRR*, **1602.00269**.

SCHMID H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proc of International Conference on New Methods in Language Processing*, p. 44–49.

SEBASTIANI F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, **34**(1), 1–47.

SHIVADE C., RAGHAVAN P., FOSLER-LUSSIER E., EMBI P., ELHADAD N., JOHNSON S. & LAI A. (2014). A review of approaches to identifying patient phenotype cohorts using electronic health records. *J Am Med Inform Assoc*, **21**(2), 221–30.

TU S., PELEG M., CARINI S., BOBAK M., ROSS J., RUBIN D. & SIM I. (2011). A practical method for transforming free-text eligibility criteria into computable criteria. *J Biomed Inform*, **44**(2), 239–50.

WANG T., LI J., DIAO Q., WEI HU Y. Z. & DULONG C. (2006). Semantic event detection using conditional random fields. In *IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW '06)*.

Analyse automatique FrameNet : une étude sur un corpus français de textes encyclopédiques

Gabriel Marzinotto[1, 2] Géraldine Damnati[1] Frédéric Béchet[2]
(1) Orange Labs, Lannion, France
(2) Aix-Marseille Université, Marseille, France
{gabriel.marzinotto,geraldine.damnati}@orange.com
frederic.bechet@lif.univ-mrs.fr

RÉSUMÉ

Cet article présente un système d'analyse automatique en cadres sémantiques évalué sur un corpus de textes encyclopédiques d'histoire annotés selon le formalisme FrameNet. L'approche choisie repose sur un modèle intégré d'étiquetage de séquence qui optimise conjointement l'identification des cadres, la segmentation et l'identification des rôles sémantiques associés. Nous cherchons dans cette étude à analyser la complexité de la tâche selon plusieurs dimensions. Une analyse détaillée des performances du système est ainsi proposée, à la fois selon l'angle des paramètres du modèle et de la nature des données.

ABSTRACT

FrameNet automatic analysis : a study on a French corpus of encyclopedic texts

This article presents an automatic frame analysis system evaluated on a corpus of French encyclopedic history texts annotated according to the FrameNet formalism. The chosen approach relies on an integrated sequence labeling model which jointly optimises frame identification and semantic role segmentation and identification. The purpose of this study is to analyze the task complexity from several dimensions. Hence we provide detailed evaluations from a feature selection point of view and from the data point of view.

MOTS-CLÉS : Analyse en cadres sémantiques, étiquetage de séquence, textes encyclopédiques.

KEYWORDS: Semantic frame analysis, sequence labeling, encyclopedic texts.

1 Introduction

L'extraction d'informations structurées dans des textes est un préalable qui favorise l'accès aux connaissances qu'ils contiennent par des outils de Traitement Automatique du Langage. Dans cette étude, nous nous intéressons au cas particulier de textes encyclopédiques historiques et nous nous plaçons dans le contexte de la représentation sémantique FrameNet. Dans l'approche FrameNet initiée par l'Institut ICSI de Berkeley (Baker *et al.*, 1998), un cadre sémantique (*Frame*) peut être évoqué par des unités lexicales (les triggers ou cibles). Par exemple, le cadre "Commitment" peut être évoqué par "promettre", "promesse", "s'engager" et le cadre "Becoming_aware" peut être déclenché par "découvrir" et "découverte". Les unités lexicales (UL) qui déclenchent un cadre peuvent être nominales ou verbales. Par ailleurs, un cadre englobe la définition des participants et des propriétés qui peuvent lui être attachés : ce sont les *Frame Elements* (FE). Ils sont spécifiques à chaque cadre et

sont nommés par des labels explicites. Par exemple, dans la phrase suivante, l'agent de l'action de découvrir, est représenté par le FE "Cognizer" qui a deux instances :

> [le premier Européen]`Cognizer` à avoir [découvert]`Becoming_aware`
> [Mammoth Cave]`Phenomenon` était [John Houchin]`Cognizer`, [en 1797]`Time`.

Les cadres peuvent être liés entre eux par des relations (Fillmore *et al.*, 2004) (ex : inheritence, using, ...) auquel cas les FE peuvent être mis en correspondance. Dans cette étude, nous réalisons une analyse "à plat" sans mettre les cadres en relation. Si les ressources linguistiques décrivant ces cadres sont de plus en plus nombreuses pour la langue anglaise, leur constitution pour le français n'en est qu'au début avec les contributions du projet ASFALDA qui s'est attaché à produire des ressources sur la base de FrameNet pour le français (Djemaa *et al.*, 2016). Pour notre part, nous avons constitué le corpus CALOR Béchet *et al.* (2017) annoté en cadres sémantiques sur des textes encyclopédiques issus de différentes sources, dans le domaine de l'histoire, décrit plus en détail à la section 3.1.

L'analyse en cadres sémantiques a pour objectif de repérer dans des documents des instances de cadres avec l'ensemble des rôles permettant de les caractériser, et se fait classiquement en deux étapes. La première est une étape de désambiguïsation afin d'identifier un cadre étant donnée la présence d'un déclencheur potentiel (UL) La seconde consiste à identifier les rôles sémantiques (FE) et est le plus souvent traitée de façon séquentielle comme l'enchaînement d'une étape de détection de segment et de classification de ce segment (Johansson *et al.*, 2012; Lechelle & Langlais, 2014). Le système Semafor (Das *et al.*, 2014) constitue à l'heure actuelle une référence dans le domaine. Dans Semafor, l'étape d'identification de cadre, étant donné un déclencheur, est réalisée à l'aide d'un classifieur probabiliste par Maximum d'Entropie. Ensuite, l'étape de labélisation des rôles sémantiques est réalisée à l'aide d'un modèle log-linéaire conditionnel qui catégorise des segments (labélisation des FE). Dans cette étape, les segments candidats sont obtenus à partir de l'analyse en dépendance et correspondent aux sous-arbres produits. De cette façon, le modèle arrive à gérer à la fois la classification et la segmentation.

Dans cette étude, nous adoptons une approche plus intégrée où les étapes de désambiguïsation, de détection des FE et de labélisation des FE se font de façon simultanée à l'aide de modèles d'étiquetage de séquences de type Conditional Random Fields (CRF). Notre modélisation du problème n'impose pas la contrainte que les FE soient toujours la projection d'une tête dans l'arbre de dépendances, ce qui rend le système robuste aux erreurs d'analyse en dépendance. Nous avons décidé de travailler avec les CRF car ce sont des modèles simples qui ne nécessitent pas de grandes puissances de calcul, ni de gros volumes de données en comparaison avec les méthodes neuronales, mais qui sont suffisamment performants pour nous permettre la mise en place de nos expériences contrastives.

La section 2 présente en détail cette approche avec un focus sur le choix des paramètres des modèles. La section 3 présente un ensemble d'expériences visant à montrer l'influence du choix des paramètres, l'impact de la complexité intrinsèque des textes, et l'influence des données d'apprentissage.

2 Analyse en cadres comme une tâche d'étiquetage de séquence

Les CRF ont été utilisés dans de nombreuses tâches du TALN ainsi que les modèles neuronaux de type RNN ou LSTM (Hakkani-Tür *et al.*, 2016; Tafforeau *et al.*, 2016). Nous avons choisi dans cet article d'utiliser des modèles CRF en se focalisant sur des comparaisons entre différents systèmes de traits et différents corpus pour entraîner et tester nos systèmes. Nous nous intéressons également à évaluer les effets du choix du corpus d'apprentissage en considérant des textes de différents styles (encyclopédique, encyclopédique adressé aux enfants, etc.) et qui traitent de sujets différents

(archéologie, histoire, etc.). Vu que l'apprentissage se fait sur des corpus de taille moyenne et que notre objectif est de faire une étude comparative de la tâche, et non pas d'arriver aux meilleures performances possibles du système final, nous avons décidé de travailler avec les modèles CRF, car ils sont plus simples, plus rapides en apprentissage et présentent moins de paramètres à régler.

Apprendre un seul CRF avec tous les exemples de phrases annotées résulterait en un très grand nombre d'étiquettes, ce qui peut être rédhibitoire lorsqu'on augmente le nombre de cadres. Une autre possibilité est de modéliser chaque cadre (sens) avec un CRF, mais cela nous obligerait à mettre en place un modèle de désambiguïsation en amont, de manière à savoir quel est le CRF qui doit être appliqué à chaque unité lexicale. Pour éviter ces problèmes nous avons décidé de modéliser chaque UL avec un CRF, cela permet de faire en même temps la désambiguïsation de cadres, la détection et la sélection des rôles sémantiques. Ce choix n'est pas nécessairement optimal dans le sens où il disperse les données d'apprentissage et ne permet pas le partage d'information entre des UL qui se ressemblent. Néanmoins il permet de passer à l'échelle lorsqu'on augmente le nombre de cadres.

Ainsi, pour analyser en cadres une nouvelle phrase, nous allons d'abord extraire les UL de la phrase qui apparaissent dans notre liste de 145 UL possibles. Pour chaque phrase il y aura autant d'applications de CRF qu'il y a d'UL, puis une étape de post-traitement permet de vérifier la cohérence des résultats d'étiquetages. Ici nous vérifions que les étiquettes mises sur les rôles sémantiques sont compatibles et rattachables aux types de cadres sémantiques prédits par les CRF. Il est possible en effet que le CRF prédise un rôle sémantique qui ne fasse pas partie des rôles possibles du cadre mais qui ferait partie des rôles d'un autre cadre qui pourrait être déclenché par la même UL. Dans notre modèle tous les rôles qui ne sont pas rattachables à leur cible sont systématiquement rejetés.

Pour extraire des caractéristiques pertinentes à la tâche nous avons évalué plusieurs types de paramètres et de codages inspirés de la littérature (Das *et al.*, 2014; Michalon *et al.*, 2016). Une sélection incrémentale a été faite pour ne retenir que les 5 paramètres les plus pertinents pour chaque token :

— son lemme, le lemme du mot parent, sa partie du discours (POS), la distance linéaire à la cible et les deux derniers niveaux du chemin de dépendances entre le mot et la cible.

La distance linéaire à la cible est le nombre de tokens entre le token courant et l'UL qui déclenche le cadre (cible). Ce nombre est négatif si le token est avant la cible, ou positif s'il est après. Le chemin de dépendance vers la cible se construit comme la concaténation des dépendances entre le token courant et la cible. L'analyse syntaxique est réalisée à l'aide de l'analyseur MACAON (Nasr *et al.*, 2010) qui construit des arbres syntaxiques avec un jeu de dépendances très similaire à celui du French TreeBank (Abeillé *et al.*, 2003; Abeillé & Barrier, 2004). Dans le cas général la cible qui déclenche un cadre n'est pas nécessairement la racine de l'arbre de dépendance de la phrase, cela implique que le chemin de dépendances entre un token et une cible est composé des dépendances non pas seulement de fils à parent (relations ascendante), mais aussi de parent à fils (relations descendantes). Nous faisons cette distinction de manière explicite en codant les chemins ascendants et descendants avec des symboles différents. Par ailleurs, nous avons observé que les chemins syntaxiques très longs étaient difficiles à modéliser. Pour contourner ce problème nous avons étudié la simplification de ces chemins en limitant leur longueur maximale, c'est-à-dire, lorsque le chemin de dépendances d'un token vers la cible du cadre sémantique dépasse une certaine longueur, nous allons le représenter avec un chemin plus court qui garde la plus grand quantité d'information possible. Dans nos expériences, nous avons obtenu que la simplification qui produisait les meilleures performances consiste à garder les deux dépendances du chemin les plus proches de la cible, qui sont souvent les plus pertinents.

3 Evaluation

3.1 Protocole expérimental

Nous avons réalisé toutes nos expériences sur le corpus CALOR. Il est constitué de documents issus de 4 sources différentes : le portail Wikipédia sur l'Archéologie (WA, 201 documents), le portail Wikipédia sur la Première Guerre Mondiale (WGM, 355 documents), des textes issus de Vikidia (VKH, 183 documents), l'encyclopédie en ligne pour enfants, à partir de deux portails (Préhistoire et Antiquité) et des textes historiques de ClioTexte (https ://clio-texte.clionautes.org/) sur la Première Guerre Mondiale (CTGM, 16 documents). Annoter un corpus en cadres sémantiques n'est pas une tâche facile à aborder car le nombre de cadres et d'unités lexicales (UL) porteuses de sens que l'on pourrait définir est énorme. Dans le cas de FrameNet (Baker *et al.*, 1998), le dictionnaire des cadres sémantiques pour l'anglais, propose 1222 cadres possibles et 13615 UL à ce jour. Pour cette raison, un corpus annoté en cadres n'est souvent étiqueté que sur une sélection des cadres et UL les plus pertinents. Les UL en dehors de cette sélection restent sans annotation et une UL sélectionnée apparaissant dans un texte avec un sens qui n'est pas prévu dans notre dictionnaire de cadres sémantiques simplifié, nous lui attribuons un cadre spécial « OTHER ». Sur le corpus CALOR, 21.398 occurrences de cadres sémantiques ont été annotées, déclenchées par une des 145 UL présentes dans notre liste de UL traitables. Au total, 53 cadres sémantiques différents ont été annotés, auxquels s'ajoute le cadre OTHER.

Lorsqu'une phrase est étiquetée en cadres sémantiques, il y a 4 sous-tâches qui se développent, parfois de façon implicite. Nous les avons incluses dans notre protocole car elles permettent d'évaluer très précisément les systèmes d'analyse en cadres sémantiques. Ce sont les tâches de : détection de cibles (DC) qui revient à décider si une UL doit être associée à OTHER ou non ; sélection du bon cadre (SC) pour chaque cible détectée ; détection des segments qui constituent des rôles sémantiques (DR) ; sélection des types de rôles sémantiques (SR). Même si l'ensemble de ces tâches est réalisé par un seul modèle intégré nous présentons les différents niveaux d'évaluation, avec un accent plus particulier sur le SR, sous-tâche qui est, de façon générale, la plus difficile de l'analyse en cadres sémantiques.

Le corpus a été divisé en cinq parties de sorte qu'aucun document ne soit jamais sous divisé et de sorte que la distribution des cadres soit la plus homogène possible entre chaque partie. Pour chaque expérience nous mesurons la précision, le rappel et la F-mesure moyennés entre les 5-Folds ainsi que l'écart type des mesures de performances sur les 5 folds.

3.2 Évaluation globale et influence des paramètres

Dans le tableau 1 nous montrons les performances du meilleur système développé à partir des 5 caractéristiques les plus pertinentes pour la tâche. Sur ce corpus les tâches DC et SC ont une complexité assez basse car nous traitons un nombre de cadres limité. Sur ces deux tâches, notre système CRF augmente la précision de 5 points par rapport à un système naïf qui choisirait la classe majoritaire. Comme ces sont des tâches simples dans notre corpus, nous arrivons à des performances élevées et assez proches car la proportion d'UL pouvant conduire à plusieurs cadres différents est assez faible (seulement 12 UL). La détermination de la catégorie OTHER demeure la principale difficulté à ce niveau. Par ailleurs, la tâche de SR qui est la plus complexe, présente un taux de précision acceptable (82.2%) étant donné le nombre de rôles possibles (150 au total), mais les performances en termes de rappel sont à peine de 51.2%. La performance élevée du système en termes de précision est due au fait d'avoir un modèle CRF pour chaque UL, car ceci diminue le nombre d'étiquettes (et le nombre de confusions) possibles au moment des prédictions.

	Précision	Rappel	Fmesure
Détection de Cible (DC)	96.4 ± 0.2	96.4 ± 0.2	96.4 ± 0.1
Sélection de Cadre (SC)	95.3 ± 0.4	95.2 ± 0.2	95.3 ± 0.2
Détection des Roles (DR)	89.7 ± 0.5	55.9 ± 0.7	68.8 ± 0.5
Sélection des Roles (SR)	82.2 ± 0.6	51.2 ± 0.7	63.1 ± 0.6

TABLE 1 – évaluation par niveaux avec la meilleure configuration (CRF à 5 paramètres)

S'il est difficile de comparer avec les résultats obtenus par le système SEMAFOR (données en anglais en plus grande quantité, nombre de cadres modélisés supérieur,...), notons cependant que notre évaluation (SR) correspondrait à la tâche *Argument Identification*, avec la configuration *full parsing* car nous ne fixons pas de valeurs Oracle dans les étapes intermédiaires, et l'évaluation *partial matching* car nous ne comptons pas les erreurs de frontière sur les rôles sémantiques. Dans ces conditions le meilleur système évalué dans (Das *et al.*, 2014) conduit à une F-mesure de 50.24.

Pour chaque cadre, ses rôles sémantiques peuvent être interprétés comme des réponses à certaines questions que l'on peut poser sur le cadre. Par exemple, pour le cadre *Deciding* nous avons :

> `Cognizer` (*qui est l'agent ?*) prend une `Decision` (*quoi ?*) parmi `Possibilities` (*parmi quoi ?*) parce que `Explanation` (*pour quelle raison ?*) à un `Time` (*quand ?*) et dans un `Place` (*où ?*).

Ceci permet de regrouper les rôles sémantiques de différents cadres et de leur donner une interprétation simple qui aide à analyser quelles sont les questions génériques pour lesquelles notre système est capable de trouver le plus grand nombre de réponses correctes. En évaluant nos résultats de cette manière nous observons que les questions *à quoi, de quand* ont des performances excellentes, ceci est dû au fait que ces questions sont fortement reliées à une préposition. Les questions les plus fréquentes sont *qui est l'agent, quoi* liées aux sujets et COD dans la syntaxe, avec des F-mesures avoisinant les 70% ensuite nous avons les *quand, où, qui* liées aux CCT, CCL et COD. Pour tous ceux-ci, nous avons des performances à peu près équivalentes, de l'ordre de 55%. Les sujets et COD sont plus faciles à détecter, car leurs chemins de dépendances sont souvent plus simples et le nombre d'exemples d'apprentissage est plus grand. Les questions pour lesquelles nous obtenons les performances les plus basses sont *dans quelle circonstance, avec quelle conséquence, de quelle manière* ce sont des questions qui ont une énorme variabilité au niveau syntaxique, sont moins fréquentes et ne sont pas ancrées à une préposition spécifique.

Dans le tableau 2, nous cherchons à mesurer l'impact de chaque caractéristique sur les performances. L'analyse est faite sur la tâche de SR. Le chemin de dépendances simplifié et la partie du discours (POS) sont les caractéristiques les plus importantes pour améliorer les performances de notre système. Par ailleurs, nous voyons que la précision est plus affectée par les lemmes, alors que le rappel est affecté par les POS, le chemin de dépendances et la distance linéaire à la cible. Le lemme du mot parent dans l'analyse en dépendances permet aussi d'augmenter la précision de notre système. En effet lorsque deux compléments ont des chemins de dépendances similaires (par exemple « dans le journal » et « pendant la guerre »), ils sont faciles à classer grâce à leur tête syntaxique.La pertinence de ce paramètre est liée au fait que l'analyse en dépendances a été faite en suivant une convention similaire à celle du French Treebank (Abeillé *et al.*, 2003; Abeillé & Barrier, 2004) et donc en considérant les prépositions comme tête des sous-arbres. Pour la dernière ligne du tableau, seuls les paramètres Lemme, POS et distance linéaire sont utilisés, et nous pouvons constater une perte de 4.5 points de F-mesure par rapport au système qui se sert de l'analyse en dépendances.

Paramètres	Précision	Rappel	F-mesure
Tous les paramètres	82.2 ± 0.6	51.2 ± 0.7	**63.1** ± 0.6
Tous sauf Chemin Dépendance	82.5 ± 0.8	47.2 ± 0.7	*60.0* ± 0.6
Tous sauf Partie du Discours (POS)	83.0 ± 1.0	47.1 ± 1.0	*60.1* ± 1.0
Tous sauf Distance Linéaire	82.2 ± 0.7	48.6 ± 0.7	61.1 ± 0.7
Tous sauf Lemme	80.2 ± 0.6	50.9 ± 0.7	62.2 ± 0.6
Tous sauf Lemme Parent	81.0 ± 0.9	51.0 ± 0.8	62.6 ± 0.8
Tous sauf Analyse en Dépendance	80.8 ± 1.2	45.9 ± 0.7	*58.6* ± 0.7

TABLE 2 – Effets de l'élimination de chaque paramètre sur les performances

3.3 Influence de la complexité des textes

Chaque phrase a une complexité inhérente, qui est due à divers facteurs. D'une façon très simpliste, une phrase plus longue est souvent plus complexe et difficile à traiter. Si nous n'observons pas d'influence sur la sélection des cadres (SC), la longueur des phrases s'avère très importante pour la tâche de SR. Nous avons observé en effet une perte de précision de plus de 7 points et une perte en rappel de plus de 22 points entre les phrases du premier décile (8 mots par phrase en moyenne) et les phrases du dernier décile (50 mots par phrase en moyenne), avec une décroissance monotone du rappel sur les 10 déciles. Ceci est dû au fait que les phrases très longues ont souvent plus de rôles sémantiques et des rôles sémantiques plus rares. De façon analogue, chaque UL a une complexité inhérente, qui dépend du fait que ce soit un verbe ou un substantif, et de la position qu'elle occupe dans l'arbre de dépendances de la phrase. Une cible est dite « racine » lorsqu'elle constitue la racine de l'arbre de dépendance de sa phrase, et « non racine » dans le cas contraire. En analysant le détail

Type de Cible	Nb Cibles	Nb FE	Précision	Rappel	Fmesure
Verbe Racine	5389	13592	**85.4** ± 0.3	**68.2** ± 1.4	**75.9** ± 0.9
Verbe non Racine	8532	19496	83.0 ± 0.9	51.3 ± 1.2	63.4 ± 0.8
Nom Racine	279	252	72.2 ± 7.6	50.2 ± 6.9	59.0 ± 6.5
Nom non Racine	7198	13538	75.4 ± 2.1	34.2 ± 0.8	47.0 ± 0.9
Total	21398	46878	82.2 ± 0.6	51.2 ± 0.7	63.1 ± 0.6

TABLE 3 – Résultats de la sélection de rôles par type de cible

des performances par Unité Lexicale, on observe de grandes disparités dans les résultats, avec 8 UL qui produisent une F-mesure supérieure à 75%, et 8 qui conduisent à une F-mesure inférieure à 25%. La quantité de données d'apprentissage n'est pas le seul facteur explicatif. Parmi les UL qui ont plus de 1000 occurrences dans le corpus, 2 UL nominales ont des performances très moyennes autour de 40% (*combat* et *attaque*) alors que les deux UL (*provenir* et *contenir*) qui produisent les meilleurs résultats (F-mesure supérieure à 80%) n'ont que 200 échantillons dans le corpus. Dans le tableau 3 nous montrons que la position de la cible dans l'arbre de dépendance a un impact important sur le rappel, avec une différence de plus de 15 points entre le cas des cibles « racine » et « non racine ». Les cibles « non racine » présentent des chemins plus compliqués et moins fiables vers leurs rôles sémantiques. Lorsqu'on compare les UL nominales avec les UL verbales, il y a une différence d'environ 10 points sur la précision et d'environ 17 points pour le rappel. Même si les cibles nominales ont moins de rôles sémantiques associés (2.3 rôles sémantiques par cadre verbal contre 1.8 par cadre nominal en moyenne) elles demeurent plus complexes à traiter, car les chemins de dépendance vers leurs rôles sémantiques sont très variables. Il faut aussi prendre en compte le fait que les UL nominales sont plus rares et ont moins de données d'apprentissage associées.

3.4 Influence des données d'apprentissage

Vue la complexité de l'annotation manuelle des cadres sémantiques, la génération de nouvelles ressources n'est pas toujours possible. Pour extraire ces cadres sur des documents d'un nouveau domaine ou issus d'une nouvelle source, nous nous intéressons à évaluer les performances des modèles appris sur des données d'autres sources annotées. Comme dans cette expérience nous nous intéressons à évaluer l'impact de la similarité thématique sur les performances du système, nous avons réduit notre jeux de données aux 54 UL qui étaient présentes dans nos 4 corpus. Pour cette

	Taille App.	Précision	Rappel	Fmesure
80% CTGM	304	83.1 ± 9.7	15.1 ± 2.0	25.5 ± 3.2
80% WA	3264	78.6 ± 8.1	26.1 ± 4.6	39.1 ± 5.9
40% WGM	2918	77.1 ± 8.4	32.2 ± 5.4	45.2 ± 5.8
80% WGM	5836	80.3 ± 7.1	37.8 ± 4.6	51.3 ± 4.8
80% WGM + 80% WA + 80%VKH	9413	78.6 ± 7.7	39.0 ± 5.2	52.0 ± 5.7
80% WGM + 80% CTGM	6140	79.8 ± 5.4	39.9 ± 3.5	53.1 ± 3.6
80% ALL	9717	79.3 ± 5.9	41.2 ± 2.3	54.1 ± 2.7

TABLE 4 – Effets de la constitution du corpus d'apprentissage

expérience nous considérons que nos documents issus de CTGM sont une nouvelle source. Cliotexte regroupe des textes historiques (discours, déclarations, ...) qui ne correspondent pas exactement à un style encyclopédique. Nous proposons diverses répartitions du corpus d'apprentissage et nous mesurons les performances des systèmes pour chaque configuration. Dans le tableau 4 nous montrons que pour une même taille de corpus d'apprentissage et un style fixe (données issues de Wikipedia, 80%WA vs. 40%WGM) les performances obtenues avec un corpus du même domaine thématique (40%WGM) sont supérieures par 6 points de F-mesure. Nous arrivons à des performances moyennes rien qu'avec un apprentissage fait avec un corpus du même domaine (80%WGM), sans avoir utilisé aucune données annotées de CTGM. De plus, à partir du moment où WGM est inclus dans le corpus d'apprentissage, l'ajout de 3500 exemples hors domaine n'a pas eu d'impact important alors que le simple ajout de 304 exemples de cadres issus du CTGM, augmente les performances de 2 points de F-mesure. Ceci met en évidence le fait qu'il est toujours utile d'annoter quelques exemples des phrases de la même source, pour franchir les différences de vocabulaire et de style.

4 Conclusion

Dans cet article nous avons présenté la tâche d'analyse en cadres sémantiques comme un problème d'étiquetage de séquences que nous avons abordé à l'aide de modèles CRF. Nous avons effectué diverses expériences faites sur le corpus CALOR constitué de données encyclopédiques annotées en cadres sémantiques, montrant des performances encourageantes à partir de données d'apprentissage de taille moyenne. Les résultats obtenus révèlent une grande variabilité des performances en fonction des types d'unité lexicale (verbales ou nominales), des types de rôles sémantiques (relations directes ou circonstancielles) mais également en fonction de la complexité intrinsèque des phrases considérées (longueur, structure de dépendance). Dans nos futurs travaux, nous allons explorer des modélisations par étiquetage de séquences à l'aide de modèles neuronaux RNN, LSTM ; et nous allons nous intéresser également au partage d'information entre les rôles des différents cadres sémantiques et UL, pour pouvoir mieux traiter les cas des cibles et cadres sémantiques peu fréquents.

Références

ABEILLÉ A. & BARRIER N. (2004). Enriching a french treebank. In *LREC*.

ABEILLÉ A., CLÉMENT L. & TOUSSENEL F. (2003). Building a treebank for french. *Treebanks*, p. 165–187.

BAKER C. F., FILLMORE C. J. & LOWE J. B. (1998). The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98, p. 86–90, Stroudsburg, PA, USA : Association for Computational Linguistics.

BÉCHET F., DAMNATI G., HEINECKE J., MARZINOTTO G. & NASR A. (2017). Calor-frame : un corpus de textes encyclopédiques annoté en cadres sémantiques. *Atelier ACor4French - Les corpus annotés du français, TALN*.

DAS D., CHEN D., MARTINS A. F., SCHNEIDER N. & SMITH N. A. (2014). Frame-semantic parsing. *Computational linguistics*, **40**(1), 9–56.

DJEMAA M., CANDITO M., MULLER P. & VIEU L. (2016). Corpus annotation within the french framenet : methodology and results. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, **X**, 3794–3801.

FILLMORE C. J., BAKER C. F. & SATO H. (2004). Framenet as a "net". In *LREC*.

HAKKANI-TÜR D., TUR G., CELIKYILMAZ A., CHEN Y.-N., GAO J., DENG L. & WANG Y.-Y. (2016). Multi-domain joint semantic frame parsing using bi-directional rnn-lstm. In *Proceedings of The 17th Annual Meeting of the International Speech Communication Association*.

JOHANSSON R., HEPPIN K. F. & KOKKINAKIS D. (2012). Semantic role labeling with the swedish framenet. In *LREC*, p. 3697–3700.

LECHELLE W. & LANGLAIS P. (2014). Utilisation de représentations de mots pour l'étiquetage de rôles sémantiques suivant framenet. In *TALN*.

MICHALON O., RIBEYRE C., CANDITO M. & NASR A. (2016). Deeper syntax for better semantic parsing. In *Coling 2016 - 26th International Conference on Computational Linguistics*, Osaka, Japan.

NASR A., BÉCHET F. & REY J.-F. (2010). Macaon : Une chaîne linguistique pour le traitement de graphes de mots. In *Traitement Automatique des Langues Naturelles - session de démonstrations*, Montréal.

TAFFOREAU J., BECHET F., ARTIERE T. & FAVRE B. (2016). Joint syntactic and semantic analysis with a multitask deep learning framework for spoken language understanding. *Interspeech 2016*, p. 3260–3264.

Détection de coréférences de bout en bout en français

Elisabeth Godbert Benoit Favre
Aix-Marseille Université, LIF-CNRS UMR 7279, 163 avenue de Luminy, 13288 Marseille Cedex 9
Benoit Favre@lif.univ-mrs.fr, Elisabeth.Godbert@lif.univ-mrs.fr

RÉSUMÉ

Notre objectif est l'élaboration d'un système de détection automatique de relations de coréférence le plus général possible, pour le traitement des anaphores pronominales et les coréférences directes. Nous décrivons dans cet article les différentes étapes de traitement des textes dans le système que nous avons développé : (i) l'annotation en traits lexicaux et syntaxiques par le système Macaon ; (ii) le repérage des mentions par un modèle obtenu par apprentissage sur le corpus ANCOR ; (iii) l'annotation sémantique des mentions à partir de deux ressources : le DEM et le LVF ; (iv) l'annotation en coréférences par un système à base de règles. Le système est évalué sur le corpus ANCOR.

ABSTRACT

End-to-end coreference resolution for French.

We aim at developing a general coreference resolution system, for processing pronoun anaphora as well as direct coreferences. We describe in this article the different processing steps of the developped system : (i) extraction of lexical and syntactic features using the Macaon system ; (ii) mention detection with a supervised sequence classifier trained on the ANCOR corpus ; (iii) semantic mention tagging with two resources : DEM and LVF ; (iv) coreference prediction with a rule-based system. The system is evaluated on the ANCOR corpus.

MOTS-CLÉS : expression référentielle, mention, détection automatique de relations de coréférence.

KEYWORDS: referential expression, mention, automatic detection of coreferences.

1 Introduction

La résolution des anaphores participe à l'interprétation sémantique des textes : elle met en relation les expressions qui font référence à une même entité du discours. Les expressions référentielles, également appelées mentions, sont les noms propres, les groupes nominaux (définis ou indéfinis) et les pronoms.
Considérons l'exemple : *J'ai perdu mon portable dans le bus 45, où puis-je espérer le récupérer ? il (/cet appareil) m'a coûté cher, j'espère le retrouver ; malheureusement l'écran est un peu rayé.* Les mentions peuvent être liées par différents types de relations de coréférence. Nous utilisons la terminologie définie dans (Muzerelle *et al.*, 2013) : anaphore pronominale *mon portable ... le* ; relation directe *mon portable ... ce portable* ; relation indirecte *mon portable ... cet appareil* ; relation associative *mon portable...l'écran*.

Pour identifier les antécédents des pronoms (*portable...le...il...le*) et trouver les coréférences (*portable...portable / appareil...écran*), on utilise classiquement le genre, le nombre, la position syntaxique, la distance des mots et des informations sémantiques sur les éventuelles mentions coréférentes (Lee *et al.*, 2013; Poesio *et al.*, 2010). Les systèmes développés pour traiter les anaphores portent en majorité sur l'anglais, on en trouve une description détaillée dans (Poesio *et al.*, 2010). Dans les systèmes développés pour le français, on peut citer les travaux de Trouilleux (2001) qui traite les anaphores pronominales et RefGen (Longo & Todirascu, 2010).

L'évaluation d'un système de détection des coréférences peut être faite manuellement, ce qui est très coûteux. Elle peut être faite automatiquement si l'on dispose d'un corpus de référence, et, à ce titre, le corpus ANCOR est une ressource très intéressante, constituée d'un ensemble de 460 dialogues transcrits annotés manuellement en mentions et en relations de coréférence et relations anaphoriques, comptant environ 500 000 mots au total (Muzerelle *et al.*, 2013).

Notre objectif est l'élaboration d'un système de détection automatique de relations de coréférence intra- et inter-phrases pour le français, de bout en bout, le plus général possible, pour le traitement des anaphores pronominales et des coréférences directes. Nous décrivons dans cet article les différentes étapes du traitement des textes sur lesquels nous travaillons :

1. annotation en traits lexicaux et syntaxiques par le système Macaon (Nasr *et al.*, 2011) ;
2. repérage des mentions par un modèle obtenu par apprentissage sur le corpus ANCOR ;
3. annotation sémantique des mentions à partir de deux ressources : le DEM et le LVF (Dubois & Dubois-Charlier, 2010) , (Dubois & Dubois-Charlier, 1997) ;
4. annotation en coréférences par un système à base de règles.

Nous décrivons pour finir l'évaluation mise en oeuvre par alignement sur le corpus ANCOR.

Nos corpus d'application sont des corpus de dialogues oraux transcrits ou dialogues en tchat. Dans ce type de textes, les relations anaphoriques pronominales sont nombreuses. Par exemple, sur le corpus de dialogues ANCOR, une étude distributionnelle sur les mentions a compté un total de 115 672 mentions tous types confondus, dont 51,1% de mentions pronominales et 48,9% de mentions nominales (Désoyer *et al.*, 2015).

2 Extraction des traits

L'extraction de traits consiste en l'analyse syntaxique des phrases, la détection de mentions, puis la classification sémantique des noms et verbes.

2.1 Repérage des mentions

Nous avons utilisé le corpus ANCOR pour construire un modèle de repérage des mentions par apprentissage automatique. Les mentions sont les groupes nominaux et les pronoms, hors pronoms réfléchis. Le corpus a été partitionné en *train, dev* et *test*, le développement du système étant réalisé grâce aux deux premières parties et l'évaluation sur la dernière.

La chaîne de traitement Macaon (Nasr *et al.*, 2011) génère une analyse en parties de discours et en dépendances syntaxiques à partir du corpus brut. Le Lefff permet ensuite de faire une analyse

morphologique des formes et d'en déduire leur lemme (Sagot, 2010). Ces traits servent de base au repérage des mentions.

Les mentions du corpus ANCOR sont projetées sur les mots à l'aide d'un encodage IOB, en applatissant les entités enchassées selon la segmentation maximale. Puis un modèle de séquence CRF (Kudo, 2005) est entraîné à partir des traits suivants :

— 1,2,3-grammes de mots, lemmes et parties de discours ;
— étiquette de dépendance du mot ;
— gouverneur (mot, lemme, partie de discours et étiquette de dépendance) ;
— longueur du constituant dont le gouverneur est la tête ;
— profondeur et hauteur dans l'arbre syntaxique du parent commun entre le mot et son suivant, et le mot et son précédent.

Les segments obtenus sont corrigés à l'aide de l'arbre syntaxique de manière à faire réapparaître les entités enchassées.

2.2 Classification sémantique des noms et des verbes

La sémantique des entités dont on parle est un trait important à prendre en compte. Soit les exemples :

1. *Mon fils a eu un problème avec son bus ce matin ; il est passé avec 30 minutes de retard*
2. *Mon fils a eu un problème avec son bus ce matin ; il est arrivé en retard au collège, il a été puni.*
3. *J'ai perdu mon portable dans le bus 45, où puis-je espérer le récupérer ?*
4. *Téléphonez au Service des objets trouvés, ils vous diront s'il a été rapporté.*

Si l'on sait identifier quelles expressions dénotent une personne ou un groupe de personnes, et lesquelles dénotent une entité inanimée, certaines anaphores pronominales pourront être résolues facilement : par exemple l'objet de *punir* ne peut être qu'un humain ou un animal, et l'objet de *rapporter* doit être une entité inanimée. Il est donc utile de disposer d'un typage sémantique des expressions nominales, et d'un typage de chaque pronom dans le contexte où il apparaît.
Pour acquérir des données sémantiques sur les noms, la plupart des systèmes développés pour l'anglais utilisent WordNet ; d'autres travaux (par exemple (Ji & Lin, 2009)) mettent en oeuvre des processus d'apprentissage sur des N-grams du Web, sur les catégories de Wikipedia, ou sur des corpus annotés manuellement, mais la fiabilité des résultats d'apprentissage n'est pas totale, et les rares corpus annotés ont une couverture limitée (Poesio *et al.*, 2010). Les pronoms quant à eux peuvent être typés sémantiquement à partir des verbes dont ils sont actants, si l'on applique des préférences sélectives comme celles que l'on trouve par exemple, pour l'anglais, dans VerbNet (Kipper-Schuler, 2005).

Dans le but d'élaborer pour le français un système de traitement des anaphores qui puisse être utilisé quel que soit le domaine d'application, nous faisons le choix du DEM (*Dictionnaire Électronique des Mots*) et du LVF (*Les Verbes Français*) pour l'acquisition de données sémantiques (Dubois & Dubois-Charlier, 2010) (Dubois & Dubois-Charlier, 1997) : en effet leur couverture du français est maximale et des données sémantiques peuvent en être extraites directement, sans apprentissage.

Classification des noms Le DEM est une base de données qui répertorie tous les mots du français (145 198 entrées). On y trouve pour chaque entrée son sens, des propriétés catégorielles, morphologiques, sémantiques, syntaxiques. Chaque nom peut y avoir plusieurs entrées, qui correspondent à ses

différents sens. Nous avons extrait du DEM tous les noms communs qui y sont présents, et choisi de ne garder que les deux premiers sens de chaque nom, pour ne pas trop en brouiller la sémantique. Nous avons extrait pour chaque nom les informations pertinentes qui donnent :

— son appartenance à l'une des cinq classes : *Tout, Animé, NonAnimé, Humain, Animal* ;

— des informations sur le ou les domaines sémantiques dans lesquels ce nom est utilisé.

Plusieurs opérations de filtrage et de croisement de ces informations ont permis d'attribuer à chaque nom une entrée unique, avec sa classe et ses domaines de rattachement (87 037 entrées).
La classification a été établie à partir des cinq classes *Tout, Animé, NonAnimé, Humain, Animal* ; *Tout* est la réunion des classes *Animé* et *NonAnimé*, *Animé* est la réunion des classes *Humain* et *Animal*. Quelques autres classes ont été définies, par exemple *Véhicule*. Pour les noms polysémiques, on utilise si nécessaire la réunion de deux classes.
Ci-dessous un extrait des informations obtenues (LOC = locatif/lieu ; LAN = langue ; LIT = littérature).

nom	classe	domaine
acadien	Humain+NonAnime	LOC, LAN
académicien	Humain	LIT
académie	NonAnime	LIT

Par ailleurs, nous avons établi une liste de noms propres, où chacun est associé à une classe sémantique : villes, pays, prénoms, etc. Cette liste contient entre autres les noms propres de nos corpus d'application. Nous obtenons finalement une classification sémantique dans une taxinomie d'une quinzaine de classes.

Typage sémantique des verbes à partir du LVF Le LVF donne pour chaque verbe ses constructions syntaxiques accompagnées de la nature de ses actants, à l'interface syntaxe-sémantique (26 610 entrées). Chaque verbe peut y avoir plusieurs entrées. L'extraction d'informations dans plusieurs champs, puis le filtrage de ces informations nous ont permis d'obtenir un typage sémantique des actants sujet et complément d'objet direct. Nous ne gardons qu'une entrée pour chaque verbe, qui couvre les usages les plus courants de ce verbe. Nous avons obtenu une table de 12 484 verbes, où chacun est associé au type de ses actants sujet et objet.

verbe	sujet	compl-objet-direct
attendre	Animé+Véhicule	Tout
atterrir	Tout	
attester	Humain	
attraper	Animé	Tout

3 Traitement des anaphores

Les coréférences sont extraites à partir d'un ensemble de règles sur l'arbre syntaxique et l'ensemble des traits sémantiques et morphologiques extraits. Comme dans la plupart des systèmes de résolution de coréférences, nous cherchons pour chaque mention son coréférent le plus proche dans le texte qui

précède.

Résolution d'anaphores pronominales
On commence par identifier, via les dépendances, les pronoms utilisés dans des formes impersonnelles (*il y a*, *il faut*) ou les pronoms qui reprennent un segment phrastique (- *Louis est arrivé.* - *Oui, je le sais*). Ces pronoms sont ignorés dans les traitements qui suivent.
Ensuite, pour chaque pronom en position d'actant sujet ou objet d'un verbe, les dépendances permettent de trouver le verbe dont il est actant ; on en déduit son type sémantique, à partir de la table de typage des verbes. Pour les pronoms en position de complément indirect ou prépositionnel, on affecte le type NonAnimé à *en* et *y*, et le type Animé à *lui, elle, eux, elles*.
Pour chaque pronom on cherche, parmi les mentions candidates, celle qui est la meilleure au vu des contraintes du système, qui prennent en compte le genre, le nombre, la classe sémantique, la distance, ainsi que des éléments de la théorie du liage (Chomsky, 1981). La relation de coréférence peut être établie avec un nom qui est l'antécédent du pronom, ou un autre pronom dont il est coréférent. La recherche se fait en plusieurs passes, chacune d'elles remontant plus ou moins loin dans le dialogue et étant plus ou moins contrainte : dans les premières passes on impose un accord sur le genre, le nombre et le type sémantique, dans les suivantes on assouplit progressivement ces contraintes. Cette méthode est semblable aux méthodes classiques qui utilisent des filtres linguistiques pour écarter des candidats, et un ordonnanceur qui met en oeuvre des préférences.

Le cas des cataphores est très particulier (*Vous pouvez le décrire ce portefeuille*) : si l'annotation syntaxique en dépendances est correcte, la relation de coréférence entre le pronom et son coréférent est directement tirée des dépendances : les deux mots ont même étiquette et même gouverneur.

Recherche de coréférences nominales
On effectue une recherche de coréférents pour chaque expression nominale définie construite sur un nom commun ou un nom propre. Les expressions définies sont de deux types : celles qui introduisent une nouvelle entité et celles qui font référence à une entité déjà introduite. Notre recherche ne porte que sur les coreférences directes entre expressions de même tête lexicale (*mon portable...ce nouveau portable*). Elle se fait en considérant, outre la tête nominale, la présence éventuelle de modifieurs, pour ne pas mettre en relation par exemple *mon nouveau portable* et *mon ancien portable*. Mais notre recherche à ce niveau est très incomplète car il faudrait savoir a priori quels modifieurs sont compatibles et lesquels ne le sont pas. Dans l'état actuel de notre travail, nous considérons que deux expressions peuvent être mises en relation si l'ensemble des modifieurs de l'une est inclus dans l'ensemble des modifieurs de l'autre ; le cas le plus simple étant que l'une des deux expressions n'ait pas de modifieur (*mon portable...ce nouveau portable*).

4 Évaluation

L'évaluation et la comparaison des systèmes de recherche de coréférences sont complexes car il faut prendre en compte (Poesio *et al.*, 2010) :

— les performances des éventuels pré-traitements pour l'annotation en genre, nombre, syntaxe, etc ;
— les performances de l'éventuel reconnaisseur d'entités nommées et de l'identification des mentions ;

— le type de texte que l'on traite : manuel technique, article de journal, dialogue, etc.

Différentes métriques ont été définies pour les campagnes d'évaluation. Nous utilisons la métrique Blanc (M.Recasens & Hovy, 2010), qui est la plus récemment définie et dont la vocation est de considérer conjointement les liens de coréférence et de non-coréférence. Les calculs se font sur les chaînes de coréférences de chaque mention. L'évaluation de notre système a été faite en utilisant comme référence la partie *test* du corpus ANCOR, qui contient 23 079 mentions.

Le corpus ANCOR contient des coréférences *directes*, *indirectes* et *associatives*. Notre système ne traite que les coréférences *directes*, mais l'évaluation a été faite sur les coréférences directes et indirectes pour que les résultats soient comparables [1].

Tout d'abord, sur le *test*, le détecteur de mentions obtient un F-score de 88.94% comme précisé dans la table 1a. Ces performances sont données sur les arbres syntaxiques prédits (ainsi que tous les autres traits). Ensuite, la prédiction des liens de coréférence obtient un score Blanc de 65.70% comme décrit dans la table 1b. Ces performances sont données sur les mentions prédites par le détecteur susmentionné. Pour mettre ce résultat en perspective, on peut se référer aux travaux sur BART dans le projet SENSEI (Kabadjov & Stepanov, 2015) qui a obtenu un score Blanc de 15.04% sur le corpus ANCOR (provoqué par un faible rappel).

Metrique	**Perf.**
Précision	90.05%
Rappel	87.86%
F-score	88.94%

(a) Détecteur de mentions

Système	**F-score**
Toutes les mentions	65.70%
Pronoms uniquement	64.05%

(b) Détecteur de coréférences

TABLE 1: Performances sur l'ensemble de *test* du corpus ANCOR.

Pour explorer la généralité du système créé, nous avons effectué une évaluation manuelle supplémentaire sur le corpus DECODA de service client dans le domaine des transports (Bechet *et al.*, 2012), uniquement sur les pronoms clitiques. Les prédictions du système ont donné un taux de réussite de 0.70, ce qui tend à montrer que l'approche peut être appliquée à de nouveaux domaines.

5 Discussion

Dans les prédictions de notre système, la plupart des erreurs viennent des éléments suivants :

— les cas où le pronom ne peut pas être typé de façon fine : en particulier s'il est sujet de *avoir* ou *être* ou s'il n'a pas de lien de dépendance vers un verbe (problème souvent dû à une disfluence ; les disfluences sont très nombreuses dans les dialogues) ;
— des expressions figées, qui ne sont pas repérées dans les prétraitements ;
— des cas d'homonymie sur lesquels une erreur de lemmatisation produit une erreur d'analyse syntaxique.

Nous pensons que notre système est général dans le sens qu'il a une grande couverture lexicale et que les procédures de recherche de coréférences sont classiques et devraient s'appliquer à tous les types de textes. Néanmoins, ces procédures ont été définies en partant de l'étude de corpus de dialogues ;

1. Les références non supportées par notre systèmes sont donc considérées fausses

elles sont donc probablement plutôt adaptées à ce type de textes.
Notre travail sur le traitement des relations directes peut encore être amélioré. Quelques critères de sélection d'antécédents n'ont pas encore été exploités et vont être étudiés, en particulier le parallélisme dans les rôles syntaxiques des mentions. D'un autre côté, l'ajout d'une mesure de saillance n'a pas été probant, alors qu'il est mis en avant dans de nombreux travaux, par exemple dans (Lappin & Leass, 1994) et (Mitkov, 2002) qui utilisent la saillance mais aucune connaissance sémantique de sens commun. Par ailleurs, nous étudions actuellement l'ajout de la sémantique des adjectifs, mais cela semble n'apporter qu'une faible amélioration.

Le traitement des relations indirectes serait envisageable en intégrant plus de données sémantiques, en particulier pour prendre en compte la synonymie ou l'hyperonymie. Actuellement, nous traitons quelques cas de synonymie spécifiques à nos corpus (*le bus 42... le 42*). Par contre le traitement des relations associatives semble hors de portée d'un système de règles simple.

6 Remerciements

Ce travail a été financé par l'Agence Nationale pour la Recherche au sein des projets suivants : ANR-15-CE23-0003 (DATCHA), ANR-16-CONV-0002 (ILCB), ANR-11-LABX-0036 (BLRI) et ANR-11-IDEX-0001-02 (A*MIDEX).

Références

BECHET F., MAZA B., BIGOUROUX N., BAZILLON T., EL-BEZE M., DE MORI R. & ARBILLOT E. (2012). Decoda : a call-centre human-human spoken conversation corpus. In *LREC*, p. 1343–1347.

CHOMSKY N. (1981). *Lectures on Government and Binding*. Foris Publications.

DUBOIS J. & DUBOIS-CHARLIER F. (1997). *Les Verbes Français.* Larousse-Bordas.

DUBOIS J. & DUBOIS-CHARLIER F. (2010). La combinatoire lexico-syntaxique dans le dictionnaire électronique des mots. les termes du domaine de la musique à titre d'illustration. *Langages 3/*, p. 31–56.

DÉSOYER A., LANDRAGIN, F. TELLIER I., LEFEUVRE A. & ANTOINE J. (2015). Les coréférences à l'oral : une expérience d'apprentissage automatique sur le corpus ANCOR. *TAL*, **55 (2)**, 97–121.

JI H. & LIN D. (2009). Gender and animacy knowledge discovery from web-scale n-grams for unsupervised person mention detection. In *Proceedings of the Pacific Asia Conference on Language, Information and Computation*, p. 220–229.

KABADJOV M. & STEPANOV E. (2015). *D4.2 – The SENSEI Discourse Analysis Tools*. Rapport interne.

KIPPER-SCHULER K. (2005). *VerbNet : A broad-coverage, comprehensive verb lexicon*. PhD. Thesis. Computer and Information Science Dept., University of Pennsylvania. Philadelphia, PA.

KUDO T. (2005). Crf++ : Yet another crf toolkit. *Software available at http ://crfpp. sourceforge. net*.

LAPPIN S. & LEASS H. (1994). An algorithm for pronominal anaphora resolution. *Computational Linguistics vol. 20 (4)*, p. 535–561.

LEE H., CHANG A., PEIRSMAN Y., CHAMBERS N., SURDEANU M. & JURAFSKY D. (2013). Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics vol. 39 (4)*, p. 885–916.

LONGO L. & TODIRASCU A. (2010). RefGen : a Tool for Reference Chains Identification. In *International Multiconference on Computer Science and Information Technology*, p. 447–454.

MITKOV R. (2002). *Anaphora Resolution*. Longman.

M.RECASENS M. & HOVY E. (2010). Blanc : Implementing the rand index for coreference evaluation. *Natural Language Engineering*, **17 (4)**, 485–510.

MUZERELLE J., LEFEUVRE A., ANTOINE J.-Y., SCHANG E., MAUREL D., VILLANEAU J. & ESHKOL I. (2013). ANCOR : premier corpus de français parlé d'envergure annoté en coréférence et distribué librement. In *Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*, p. 555–563.

NASR A., BECHET F., REY J., & LE ROUX J. (2011). Macaon :a linguistic tool suite for processing word lattice. In *49th Annual Meeting of the Association for Computational Linguistics : demonstration session*.

POESIO M., PONZETTO S. & VERSLEY Y. (2010). Computational models of anaphora resolution : A survey. *http ://wwwusers.di.uniroma1.it/ ponzetto/pubs/poesio10a.pdf*.

SAGOT B. (2010). The lefff, a freely available and large-coverage morphological and syntactic lexicon for french. In *7th international conference on Language Resources and Evaluation (LREC 2010)*.

TROUILLEUX F. (2001). *Identification des reprises et interprétation automatique des expressions pronominales dans des textes en français*. Thèse. Université Blaise Pascal, Clermont-Ferrand.

Une approche hybride pour la construction de lexiques bilingues d'expressions multi-mots à partir de corpus parallèles

Nasredine Semmar[1] Morgane Marchand[2]
(1) CEA, LIST, Laboratoire Vision et Ingénierie des Contenus, 91191 Gif-sur-Yvette, France
(2) eXenSa, 41 rue Périer, 92120 Montrouge, France
nasredine.semmar@cea.fr, morgane.marchand@exensa.com

RESUME

Les expressions multi-mots jouent un rôle important dans différentes applications du Traitement Automatique de la Langue telles que la traduction automatique et la recherche d'information interlingue. Cet article, d'une part, décrit une approche hybride pour l'acquisition d'un lexique bilingue d'expressions multi-mots à partir d'un corpus parallèle anglais-français, et d'autre part, présente l'impact de l'utilisation d'un lexique bilingue spécialisé d'expressions multi-mots produit par cette approche sur les résultats du système de traduction statistique libre Moses. Nous avons exploré deux métriques basées sur la co-occurrence pour évaluer les liens d'alignement entre les expressions multi-mots des langues source et cible. Les résultats obtenus montrent que la métrique utilisant un dictionnaire bilingue amorce de mots simples améliore aussi bien la qualité de l'alignement d'expressions multi-mots que celle de la traduction.

ABSTRACT

A hybrid approach to build bilingual lexicons of multiword expressions from parallel corpora
Multiword expressions play an important role in numerous applications of Natural Language Processing such as machine translation and cross-language information retrieval. This paper presents, on the one hand, a hybrid approach to build bilingual lexicons of multiword expressions from an English-French parallel corpus, and on the other hand, the impact of using a domain-specific bilingual lexicon of multiword expressions on the results of the open source statistical machine translation Moses. We have explored two metrics based on co-occurrence to evaluate alignment links between multiword expressions of source and target texts. The obtained results show that scoring based on a bilingual dictionary improves the quality of both alignment and translation.

MOTS-CLES : Lexique bilingue, alignement d'expressions multi-mots, programmation linéaire.
KEYWORDS: Bilingual lexicon, multiword expressions alignment, linear programming.

1 Introduction

Une expression multi-mots est une combinaison de mots pour laquelle les propriétés syntaxiques ou sémantiques de l'expression entière ne peuvent pas être obtenues à partir de ses parties (Sag et al., 2002). Contrairement à l'alignement de mots simples qui est désormais une tâche bien maîtrisée plus particulièrement pour les langues à écriture latine, l'alignement d'expressions multi-mots continue à

susciter de nombreux travaux de recherche (DeNero, Klein, 2008; Lefever et al., 2009; Bouamor et al., 2012; Makcen et al., 2013). La plupart de ces travaux commencent tout d'abord par identifier les expressions multi-mots dans chaque partie du corpus parallèle (Ramisch, 2014), ensuite, utilisent différentes approches d'alignement pour les apparier. Les approches pour l'extraction monolingue d'expressions multi-mots peuvent être: (1) symboliques en reposant sur des patrons morpho-syntaxtiques (Okita et al., 2010; Dagan, Church, 1994), (2) statistiques en utilisant des mesures d'association pour classer les expressions multi-mots candidates (Vintar, Fisier, 2008), et (3) hybrides combinant (1) et (2) (Seretan, Wehrli, 2007; Semmar, Laib, 2010). Pour identifier les correspondances entre expressions multi-mots dans différentes langues, plusieurs travaux font appel à des outils d'alignement de mots simples pour guider l'alignement d'expressions multi-mots. L'outil de création de lexiques bilingues de termes techniques de Dagan et Church (1994) identifie ces termes sur la base d'un étiqueteur morpho-syntaxique. Ensuite, la liste de candidats trouvés est filtrée manuellement. L'application de filtres à base de catégories grammaticales permet une réduction importante du bruit dans les sorties par l'exclusion de candidats constitués de mots vides. Malgré leur simplicité, les approches symboliques restent difficiles à appliquer lorsque les données ne sont pas étiquetées morpho-syntaxiquement. Une autre limite de cette approche est que la définition de patrons d'extraction d'expressions multi-mots est dépendante de la langue. D'autres utilisent les algorithmes d'apprentissage statistique (Okita et al., 2010). Une hypothèse largement suivie pour acquérir des expressions multi-mots bilingues est qu'une expression multi-mots dans une langue source garde la même structure syntaxique que son équivalente dans une langue cible donnée (Seretan, Wehrli, 2007). Or, cette hypothèse n'est pas toujours vérifiée puisque certaines expressions multi-mots ne se traduisent pas forcément par des expressions ayant la même structure syntaxique (Bouamor et al., 2012). Nous présentons, dans cet article, une approche hybride pour l'acquisition d'un lexique bilingue d'expressions multi-mots à partir d'un corpus parallèle anglais-français. Contrairement aux travaux cités précédemment, notre approche permet de réaliser les tâches d'extraction des expressions multi-mots et d'alignement en une seule étape.

La suite de l'article est organisée comme suit : dans la section 2, nous présentons notre approche d'alignement d'expressions multi-mots, nous décrivons en particulier les métriques que nous avons utilisées pour évaluer les liens d'alignement. Puis nous décrivons dans la section 3 les résultats obtenus. La section 4 conclut notre étude et présente nos travaux futurs.

2 Notre approche pour l'alignement d'expressions multi-mots

Comme nous l'avons mentionné précédemment, la plupart des approches d'alignement d'expressions multi-mots se déroulent en deux étapes. Tout d'abord les fragments de phrases représentant potentiellement des expressions sont extraits. Ensuite les potentielles expressions repérées sont associées deux à deux. Cela a l'inconvénient de propager d'éventuelles erreurs. Nous décrivons dans cette section une approche hybride permettant de réaliser les tâches d'extraction des expressions multi-mots et d'alignement en une seule étape sous la forme d'un problème d'optimisation (Marchand, Semmar, 2011). Ce problème d'optimisation peut être formulé comme suit : une bi-phrase consiste en deux phrases *e* et *f* en deux langues différentes qui sont traduction l'une de l'autre. Chacune de ces phrases peut être vue comme une séquence ordonnée de mots. On désigne par e_{ij} le fragment de phrase contenu entre les espaces *i* et *j* de la phrase *e*. f_{kl} identifie le fragment de phrase contenu entre les espaces *k* et *l* de la phrase *f*. Un lien est une paire de fragments (un bi-fragment) alignés que l'on note (e_{ij}, f_{kl}). Chaque e_{ij} peut être lié avec plusieurs f_{kl} et réciproquement. Un alignement *a* de la bi-phrase (*e*, *f*) est une segmentation des deux phrases en fragments contigus ainsi que l'ensemble des liens entre ces fragments. Nous utilisons une fonction réelle Φ pour donner des scores aux liens selon s'ils sont plus ou moins probables.

$$\phi : \{e_{ij}\} \times \{f_{kl}\} \rightarrow R$$

Le score d'un alignement *Φ(a)* est le produit des scores des liens qui le composent.

$$\phi(a) = \prod_{(e_{ij}, f_{kl}) \in a} \phi(e_{ij}, f_{kl})$$

Formellement, cette fonction n'a pas d'autres contraintes que celle d'être réelle. Dans la pratique on va choisir une fonction qui donne une idée sur la pertinence d'aligner ensemble tels ou tels fragments. Plus le score associé est élevé plus la pertinence de l'alignement est grand. On cherche donc l'alignement (segmentation + liens) qui va maximiser le score décrit ci-dessus. Cependant, trouver un alignement qui maximise ce score est un problème NP-complet (DeNero, Klein, 2008). C'est pourquoi on utilise une forme linéarisée du problème afin de pouvoir lui appliquer des méthodes de programmation linéaire en nombre entiers et de trouver une solution approchée rapidement. Pour écrire le problème sous forme de programme linéaire en nombres entiers nous devons introduire des variables binaires. Ces variables sont identifiées par des majuscules au contraire des fragments de phrases qui sont dénotés par des minuscules. Les variables E_{ij} et F_{kl} indiquent si les fragments e_{ij} et f_{kl} font partie de la segmentation considérée. Les variables binaires A_{ijkl} indiquent quant à elles s'il existe un lien entre les fragments e_{ij} et f_{kl}. On utilise $W_{ijkl} = log(\Phi(e_{ij}, f_{kl}))$ afin de pouvoir linéariser le score d'un alignement. Le problème se présente sous forme ci-dessous:

$$\begin{cases} \max \sum_{i,j,k,l} W_{i,j,k,l} A_{i,j,k,l} & \\ \forall\, x : 1 \leq x \leq |e| & \sum_{i,j:i<x\leq j} E_{i,j} = 1 \quad (1) \\ \forall\, y : 1 \leq y \leq |f| & \sum_{k,l:k<y\leq l} F_{k,l} = 1 \quad (2) \\ \forall\, i, j & \sum_{k,l} A_{i,j,k,l} \geq E_{i,j} \quad (3) \\ \forall\, k, l & \sum_{i,j} A_{i,j,k,l} \geq F_{k,l} \quad (4) \\ \forall\, i, j, k, l & 2 \cdot A_{i,j,k,l} \leq E_{i,j} + F_{k,l} \quad (5) \end{cases}$$

Avec les contraintes suivantes sur les variables:

$$\begin{cases} 0 \leq i < |e|, & 0 < j \leq |e|, \quad i < j \\ 0 \leq k < |f|, & 0 < l \leq |f|, \quad k < l \end{cases}$$

Les contraintes (1) et (2) signifient qu'un mot fait partie d'un et un seul fragment dans la segmentation choisie des phrases. La contrainte (3) assure que chaque fragment de la segmentation choisie de *e* est lié à au moins un fragment de la segmentation choisie de *f*. De même pour la contrainte (4) pour *f*. Un fragment peut tout à fait être lié à plusieurs fragments. Enfin, la contrainte (5) assure que si un lien existe entre e_{ij} et f_{kl} (c'est à dire que $A_{ijkl} = 1$), alors e_{ij} et f_{kl} sont dans les segmentations choisies de *e* et *f*. En effet, nous ne voulons lier entre eux que des fragments existants réellement. Bien que notre approche s'inspire des travaux de DeNero et Klein (2008), la formulation

des contraintes (3) à (5) nous différencie. En effet, DeNero et Klein (2008), ne considèrent que les alignements bijectifs alors que nous prenons en compte les alignements surjectifs. C'est un premier pas vers la non contiguïté des expressions. Comme mentionné plus haut, pour mettre ce problème sous forme linéaire nous avons été obligé de nous restreindre aux fragments continus. Hors certaines expressions sont en plusieurs parties. Cette surjectivité des alignements vise à contrebalancer en partie cette restriction en autorisant un fragment à être lié à plusieurs fragments pas forcément contigus. Nous avons utilisé la boite à outils GLPK[1] pour résoudre ce programme linéaire en nombres entiers. L'objectif étant de trouver comment attribuer un score pertinent aux bi-fragments. Ce programme fonctionne avec n'importe quelle fonction de score Φ à valeur réelle. Nous décrivons dans les sections suivantes les deux métriques que nous avons utilisées.

2.1 Co-occurrence

Pour estimer la distribution de probabilité d'alignement, nous avons utilisé une métrique fondée sur la co-occurrence. Cette métrique utilise l'information selon laquelle la traduction d'une partie d'une phrase se trouve forcément quelque part dans la phrase associée. Pour calculer la distance en terme de co-occurence, on utilise un corpus d'entrainement bilingue aligné phrase à phrase. Pour chacun des fragments, on vérifie s'il est ou non présent dans chaque phrase. Le score entre deux fragments e_{ij} et f_{kl} est calculé selon la formule suivante :

$$\phi_c(e_{ij}, f_{kl}) = \frac{\sum_{s' \in S} N_{s'}(e_{ij}) \times N_{s'}(f_{kl})}{\sum_{s \in S} N_s(e_{ij}) + N_s(f_{kl}) - N_s(e_{ij}) \times N_s(f_{kl})}$$

$N_s(e_{ij})$ vaut 1 si le fragment e_{ij} de la première langue est présent dans la phrase s du corpus S et 0 sinon. $N_s(f_{kl})$ se comporte de la même façon pour l'autre langue. Ce score calcule le nombre de présence commune des deux fragments divisé par le nombre total d'apparition de l'un ou de l'autre. Si e_{ij} ou f_{kl} n'apparaissent pas dans le corpus, le score vaut zéro. Ainsi si un fragment du corpus test est inconnu du corpus d'apprentissage son score sera nul et le score de n'importe quel alignement de la phrase le contenant sera nul également.

2.2 Dictionnaire bilingue

Nous voulons que notre alignement fragment à fragment respecte autant que possible les alignements mot à mot procurés par un dictionnaire bilingue amorce. Le dictionnaire anglais-français utilisé dans cette étude est composé de 243539 entrées. Ce dictionnaire n'est certainement pas un lexique amorce type mais nous avons utilisé uniquement les entrées qui sont sous la forme de mots simples (unitermes). Nous supposons que la présence d'une paire de mots dans ce dictionnaire renforcera la probabilité d'alignement des bi-fragments qui les contiennent. Ce dictionnaire peut aussi nous donner des informations d'alignement négatives lorsqu'un mot n'a pas de traduction. Nous prenons en compte ces informations pour pénaliser légèrement la création de liens qui n'étaient pas à la base dans le dictionnaire. Le score dictionnaire se calcule alors avec la formule suivante :

[1] GLPK est disponible sur http://www.gnu.org/s/glpk/.

$$\phi(e_{ij}, f_{kl}) = \frac{a \cdot R_1 + b \cdot R_0}{a \cdot R_1 + b \cdot R_0 + c \cdot N_1 + d \cdot N_0}$$

Avec, R_1: nombre de liens respectés; R_0: nombre de non-liens respectés; N_1: nombre de liens non-respectés; N_0: nombre de non-liens non respectés. Les coefficients *a*, *b*, *c* et *d* peuvent être adaptés dans le but d'ajuster l'importance relative des quatre termes. Nous avons analysé un petit corpus qui nous a permis de choisir empiriquement d'utiliser les valeurs suivantes : $a = b = c = 1$ et $d = ½$. Notons que le score dictionnaire est utilisé pour renforcer ou affaiblir le score co-occurrence.

3 Résultats expérimentaux et discussion

Nous avons évalué les résultats de notre approche d'alignement d'expressions multi-mots pour le couple de langues anglais-français selon deux stratégies différentes : (1) une évaluation manuelle comparant les résultats de notre approche d'alignement d'expressions multi-mots et ceux de l'outil Giza++ (Och, Ney, 2000) par rapport à un alignement de référence; (2) une évaluation automatique qui consiste à étudier l'impact de l'utilisation du lexique bilingue produit par notre approche d'alignement d'expressions multi-mots sur la qualité de traduction du système Moses[2].

3.1 Evaluation intrinsèque

L'évaluation manuelle a été réalisée en utilisant un lexique bilingue de référence construit à partir de 1992 phrases parallèles extraites du corpus Europarl (European Parliament Proceedings). Ce lexique composé de 7807 entrées a été construit manuellement à l'aide de l'outil Yawat (Germann, 2008) par deux annotateurs. Pour les métriques d'évaluation, nous avons utilisé celles du protocole défini lors de la conférence HLT/NAACL 2003 (Mihalcea et al., 2003). Le tableau 1 résume les performances des différentes approches.

Métrique utilisée	**Précision**	**Rappel**	**F-mesure**
Baseline (Giza++)	0,83	0,37	0,51
Co-occurrence	0,61	0,63	0,61
Co-occurrence + Dictionnaire bilingue	0,85	0,54	0,66

TABLE 1 : Performances des différentes approches d'alignement d'expressions multi-mots

Les résultats obtenus avec uniquement la métrique co-occurrence montrent que quand la structure des phrases est similaire dans les deux langues, l'alignement est efficace même si la traduction n'est pas mot à mot (Figure 1). La segmentation se fait mot à mot ou bien fragment par fragment en fonction de ce qui est plus fréquent dans le corpus.

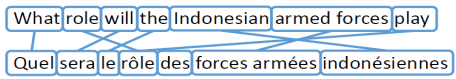


FIGURE 1: Exemple d'un alignement correct

[2] http://www.statmt.org/moses/

Ces résultats montrent aussi que la formulation surjective du problème permet de détecter les expressions en deux parties. Nous pouvons en effet voir que le mot français « rôle » est lié à la fois à « role » et « play », ce qui aurait été impossible avec les hypothèses de DeNero et Klein (2008), puisqu'elles ne permettent pas à un fragment d'une langue d'être aligné avec plusieurs fragments de l'autre langue. Dans ce cas précis, l'expression est partiellement reconnue car « role » et « play » sont deux mots pleins ayant une distribution de présence significative sur le corpus d'entrainement. Nous avons aussi constaté que l'ajout de l'information provenant du dictionnaire bilingue amorce améliore les résultats d'alignement. Dans l'exemple suivant (Figure 2), le dictionnaire donne les liens « be/être », « decided/décidé » et « there/y ». Ces liens permettent à notre approche d'alignement de reconstruire le bi-segment complet « is to be decided on there/doit y être décidé ». De plus, les liens « programme/programme » et « concrete/concret » sont renforcés. Notons que le lexique produit par notre approche d'alignement ne contient pas que les expressions multi-mots mais peut contenir des fragments de textes. Par exemple, le fragment « doit y être décidé » n'est pas une expression multi-mots mais un segment phrastique.

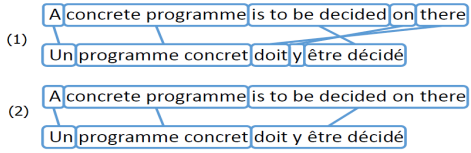


FIGURE 2: Exemples d'améliorations : (1) Métrique co-occurence, (2) Métrique co-occurence + dictionnaire bilingue

3.2 Evaluation extrinsèque

La non disponibilité d'un protocole commun permettant d'évaluer les résultats d'alignement d'expressions multi-mots nous a conduit à réaliser une évaluation extrinsèque dans laquelle nous étudions l'impact de l'utilisation du lexique bilingue produit par notre approche d'alignement dans le système Moses. Pour ce faire, nous avons réalisé des expérimentations sur deux corpus parallèles anglais-français (Table 2): Europarl et Emea (European Medicines Agency Documents). Ces deux corpus ont été extraits de la base libre de corpus parallèles OPUS (Tiedemann, 2012).

N° du run	Apprentissage (nombre de phrases)	Développement (nombre de phrases)
1	150K+10K (Europarl+Emea)	2K+0,5K (Europarl+Emea)
2	150K+20K (Europarl+Emea)	2K+0,5K (Europarl+Emea)
3	150K+30K (Europarl+Emea)	2K+0,5K (Europarl+Emea)

TABLE 2 : La taille (en nombre de phrases) de l'ensemble des corpus bilingues utilisés pour l'apprentissage et le développement du système de traduction Moses

Nous avons effectué trois runs avec, pour chacun d'entre eux, un test avec des données du domaine et un autre hors-domaine. Pour cela, nous avons extrait de manière aléatoire un ensemble de 500 paires de phrases à partir du corpus Europarl comme un corpus correspondant au domaine et 500 autres paires de phrases à partir du corpus Emea comme un corpus hors-domaine. Ces tests ont pour but de montrer l'impact que peut avoir le lexique du domaine produit par notre approche d'alignement d'expressions multi-mots à partir du corpus parallèle spécialisé (Emea) sur le modèle

de traduction du système Moses. Ce lexique est constitué en utilisant les deux métriques : co-occurrence et co-occurrence + dictionnaire bilingue. Nous comparons les résultats de traduction du système Moses après l'injection de ces deux lexiques spécialisés avec les résultats de ce système après l'injection du corpus ayant servi à produire ce lexique spécialisé (Emea) dans les données d'apprentissage (Europarl). Comme les entrées du dictionnaire bilingue sont sous forme de lemmes, nous avons d'une part, lemmatisé le corpus d'apprentissage à l'aide de la plate-forme LIMA (Besançon et al., 2010), et d'autre part, estimé le modèle de traduction sur des lemmes plutôt que sur des formes de surface. Les performances de Moses ont été évaluées en utilisant le score BLEU (Papineni et al., 2002) sur les deux ensembles de test pour les trois runs sachant que nous considérons une référence par phrase, les résultats obtenus sont présentés dans le tableau 3 (Table 3).

N° du run	**Domaine (Europarl)**			**Hors-Domaine (Emea)**		
	Baseline (Giza++)	**Co-occurrence**	**Co-occurrence + Dictionnaire bilingue**	**Baseline (Giza++)**	**Co-occurrence**	**Co-occurrence + Dictionnaire bilingue**
1	32,62	32,69	32,71	22,96	23,03	23,06
2	33,81	33,88	33,89	23,30	23,37	23,39
3	34,25	34,30	34,32	24,55	24,59	24,62

TABLE 3 : Scores BLEU de Moses selon l'approche d'alignement d'expressions multi-mots utilisée

Tout d'abord, nous constatons que le score BLEU obtenu est satisfaisant compte tenu de la taille du corpus d'apprentissage utilisé. Ce score varie en fonction du type du jeu de test. Le corpus de test du Domaine (Europarl) rapporte des scores BLEU plus élevés que le corpus de test Hors-Domaine (Emea) dans les trois configurations (alignement avec Giza++, alignement avec notre approche en utilisant la métrique fondée sur la co-occurrence, alignement avec notre approche en utilisant une combinaison des deux métriques : co-occurrence et dictionnaire bilingue). Les résultats obtenus montrent que le score fondé sur la combinaison des deux métriques produit les meilleures performances pour tous les runs.

4 Conclusion

Dans cet article, nous avons présenté une approche hybride pour l'acquisition d'un lexique bilingue d'expressions multi-mots à partir d'un corpus parallèle permettant de réaliser les tâches d'extraction des expressions multi-mots et d'alignement en une seule étape. Nous avons, en particulier, montré que l'ajout d'un lexique bilingue spécialisé (construit à l'aide de cette approche) au corpus d'apprentissage du système de traduction statistique Moses améliore la qualité de traduction aussi bien des textes du domaine de spécialité que les textes du domaine général. Dans nos expérimentations, le modèle de traduction a été estimé sur des lemmes plutôt que sur des formes de surface. De même, nos diverses métriques calculent des probabilités jointes. Nos travaux futurs s'orientent, d'une part, vers l'utilisation d'un modèle de génération pour produire les formes de surface adéquates à partir des résultats de traduction présentés en lemmes dans cette étude, et d'autre part, vers l'extraction des probabilités conditionnelles associées qui sont les primitives de la probabilité jointe afin de pouvoir intégrer les scores fournies par nos métriques dans la table de traduction de Moses.

Références

BESANÇON R., DE CHALENDAR G., FERRET O., GARA F., LAIB M., MESNARD O., SEMMAR N. (2010). LIMA: A multilingual framework for linguistic analysis and linguistic resources development and evaluation. Actes de *LREC 2010.*

BOUAMOR D., SEMMAR N., ZWEIGENBAUM P. (2012). Identifying bilingual Multi-Word Expressions for Statistical Machine. Actes de *LREC 2012.*

DAGAN I., CHURCH K. (1994). Termight : Identifying and translating technical terminology. Actes de *the 4th Conference on ANLP*, pages 34–40, Stuttgart, Germany.

DENERO J., KLEIN D. (2008). The complexity of phrase alignment problems. Actes de *the 46th annual meeting of the association for computational linguistics on human language technologies.*

GERMANN U. (2008). Yawat: Yet Another Word Alignment Tool. Actes de *ACL 2008.*

LEFEVER E., MACKEN L., HOSTE V. (2009). Language-independent bilingual terminology extraction from a multilingual parallel corpus. Actes de *the 12th Conference of the European Chapter of the Association for Computational Linguistics, Greece*.

MACKEN L., LEFEVER E., HOSTE V. (2013). TExSIS: Bilingual terminology extraction from parallel corpora using chunk-based alignment. *Terminology, Volume 19, Issue 1*, pages: 1-30.

MARCHAND M., SEMMAR N. (2011). A Hybrid Multi-Word Terms Alignment Approach Using Word Co-occurrence with a Bilingual Lexicon. Actes *de 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC'11).*

MIHALCEA R., PEDERSEN T. (2003). An evaluation exercise for word alignment. Actes de *the HLT-NAACL 2003 Workshop on Building and using parallel texts: data driven machine translation and beyond*, pages: 1-10.

OCH F. J., NEY H. (2000). Improved Statistical Alignment Models. Actes de *the 38th Annual Meeting of the Association for Computational Linguistics,* pages: 440-447.

OKITA T., GUERRA M., ALFREDO GRAHAM Y., WAY A. (2010). Multi-word expression sensitive word alignment. Actes de *4th International Workshop on Cross Lingual Information Access*.

PAPINENI K., ROUKOS S., WARD T., ZHU W. J. (2002). Bleu: a method for automatic evaluation of machine translation. Actes de *the 40th Annual meeting of the Association for Computational Linguistics,* pages: 311-318.

RAMISCH C. (2014). *Multiword Expressions Acquisition: A Generic and Open Framework*. Springer.

SAG I., BALDWIN T., FRANCIS BOND F., COPESTAKE A., FLICKINGER D. (2002). Multiword expressions: a pain in the neck for NLP. Actes de *CICLing 2002*.

SEMMAR N., LAIB M. (2010). Using a Hybrid Word Alignment Approach for Automatic Construction and Updating of Arabic to French Lexicons. Actes de *Workshop on Language Resources and Human Language Technologies for Semitic Languages co-located with the seventh international conference on Language Resources and Evaluation (LREC 2010).*

SERETAN V., WEHRLI E. (2007). Collocation translation based on sentence alignment and parsing. *Actes de TALN 2007.*

TIEDEMANN J. (2009). News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. *N. Nicolov and K. Bontcheva and G. Angelova and R. Mitkov (eds.) Recent Advances in Natural Language Processing, Volume V*, pages: 237-248.

VINTAR S., FISIER D. (2008). Harvesting multi-word expressions from parallel corpora. Actes de *LREC 2008.*

Une interprétation probabiliste des informations de factivité

Timothée Bernard[1, 2]
(1) Laboratoire de linguistique formelle (U. Paris Diderot & CNRS), Paris
(2) Sémagramme (Inria Nancy - Grand Est), Nancy
timothee.bernard@protonmail.com

RÉSUMÉ

Nous présentons une nouvelle formalisation de la *factivité*, la dimension représentant le degré de croyance qu'une source – l'auteur ou tout autre agent mentionné dans un texte – accorde à une éventualité donnée. Nous insistons sur l'aspect *dynamique* de cette notion ainsi que sur ses interactions avec la *structure discursive*. Nous montrons comment une interprétation en termes d'*ensembles de probabilités* permet de s'affranchir des principaux problèmes que posait la formalisation utilisée dans les travaux précédents au calcul d'une factivité cohérente à l'échelle du texte dans sa totalité.

ABSTRACT

Factuality information as sets of probabilities

We present a new formalisation of *factuality*: the dimension representing the degree of belief of a source (e.g. the author or any other agent mentionned in a text) for a given eventuality. We lay emphasis on the *dynamic* aspect of this notion and on its interactions with *discourse structure*. We show how an interpretation in terms of *sets of probabilities* solves the main problems arising out of the previous formalisation on the computation of coherent values of factuality at the discourse level.

MOTS-CLÉS : factivité, modalité, degrés de croyance, analyse discursive.

KEYWORDS: factuality, modality, degrees of belief, discourse analysis.

1 Introduction

Les événements mentionnés dans un texte ne sont pas tous décrits comme réels (*il pleut*), mais apparaissent parfois seulement hypothétiques (*il a probablement plu*) ou non réalisés (*il ne pleut pas*). C'est le degré de certitude associé à chaque événement que, suivant la terminologie employée dans (Saurí, 2008; Saurí & Pustejovsky, 2012), nous allons appeler sa *factivité*.

La capacité à déterminer la factivité des différents événements mentionnés dans un texte, ou au moins à détecter un certain nombre de marqueurs de négation ou d'incertitude, s'avère crucial pour un très grand nombre d'applications liées au traitement automatique des langues (Morante & Sporleder, 2010; Farkas *et al.*, 2010). Les domaines concernés sont aussi variés que la recherche d'information en corpus biomédical (Kim *et al.*, 2009), la détection d'opinion circulant sur les réseaux sociaux (Soni *et al.*, 2014) ou la reconnaissance d'implication textuelle (Hickl & Bensley, 2007).

Dans (Saurí, 2008) est décrit un corpus, FactBank, dans lequel chaque mention d'éventualité est annotée avec les valeurs de factivité que lui accordent l'auteur et les différents agents dont il est question dans le texte. Ce corpus a servi à l'élaboration d'un algorithme, *De Facto*, permettant de

calculer ces valeurs de factivité. Cependant, la notion de factivité considérée dans (Saurí, 2008) est *statique* et essentiellement *lexicale* ; aucun phénomène discursif, ni les connaissances du monde n'y sont considérés, bien que ces éléments influent sur le jugement du lecteur (de Marneffe *et al.*, 2012).

À partir de la formalisation de (Saurí, 2008), un paradigme pour l'élaboration d'un algorithme de calcul de factivité dynamique s'appuyant sur la structure discursive a été proposé (Danlos & Rambow, 2011). Ces travaux ont par exemple inspiré l'analyse des connecteurs *en réalité* et *en effet* présentée dans (Danlos, 2012). Toutefois, nous pensons que la factivité telle que définie dans (Saurí, 2008), liée à des considérations pratiques sur l'échelle phrastique, est peu adaptée à l'analyse discursive et qu'une reformalisation de cette dimension et de ses valeurs possibles permettrait une modélisation à la fois plus exacte et plus simple de sa nature dynamique, ouvrant la voie à la conception d'outils automatiques de traitement de la langue naturelle qui en tireraient profit.

Nous commençons par présenter en section 2 la notion de factivité. En section 3, nous montrons que la formalisation qui en est donnée dans (Saurí, 2008) n'est pas adaptée à l'analyse discursive. C'est pourquoi nous proposons en section 4 une nouvelle définition, fondée sur une interprétation en termes de probabilité et s'affranchissant des problèmes précédents.

2 Factivité

Telle qu'étudiée dans (Saurí, 2008), la factivité est une propriété des mentions d'éventualités (une notion large, regroupant événements – s'inscrivant dans le temps – et états – permanents ou temporaires) présentes dans un texte. La factivité est liée à la concordance qu'a une éventualité avec le monde réel, c'est-à-dire indique si l'éventualité est exprimée comme correspondant à une situation réalisée (1a), à une situation irréalisée (1b), ou à une situation de statut incertain (1c).

(1) a. Fred est parti au Brésil.
b. Fred n'est pas parti au Brésil.
c. Fred est peut-être parti au Brésil.

Une valeur de factivité est toujours définie relativement à un agent et un moment donné ; cette conjonction (agent, temps) est la *source* de la factivité. La source de factivité la plus courante dans un texte est l'auteur au moment de l'énonciation (notée Wr chez (Saurí, 2008) ainsi que dans ce texte), cependant celui-ci peut exprimer la factivité issue d'autres agents et à d'autres instants. Par exemple, deux agents peuvent être la source d'une factivité différente sur un même événement comme la pluie en (2a), ou encore, un même agent peut être la source d'une factivité différente à deux moments distincts sur un unique événement comme la venue de Fred en (2b).

(2) a. Jeanne affirme qu'il va pleuvoir alors que Fred est sûr du contraire.
b. Fred m'avait assuré qu'il viendrait, mais il dit finalement qu'il restera chez lui.

Pour exprimer la factivité, (Saurí, 2008) utilise les deux échelles <*certitude, probabilité, possibilité*> et <*impossibilité, improbabilité, incertitude*> de (Horn, 1972), représentant chacune trois valeurs de *modalité* (de la plus forte à la plus faible) pour deux valeurs de *polarité* (respectivement positive et négative). En introduisant des valeurs sous-spécifiées, (Saurí, 2008) combine alors 4 valeurs de modalité épistémique – CT (« certainty », PR (« probability »), PS (« possibility ») et U (« underspecified »)

– et 3 valeurs de polarité – + (« positive »), − (« negative ») et u (« underspecified ») – pour former les 8 valeurs de factivité présentées au tableau 1, les 4 valeurs restantes (PSu, PRu, U+ et U−) n'étant pas considérées linguistiquement pertinentes. Suivant la convention de (Saurí, 2008), le fait qu'une éventualité e possède une valeur de factivité val pour une source s est ici noté $f(e, s) = val$.

valeur	interprétation	valeur	interprétation
CT+	factuel	PR+	probable
CT−	contre-factuel	PR−	improbable
PS+	possible	CTu	certitude non spécifiée
PS−	non certain	Uu	incertitude totale ou factivité non précisée

TABLE 1 – Les valeurs de factivité définies dans (Saurí, 2008).

La plupart de ces 8 valeurs ne requièrent pas d'explication ; précisons seulement que la valeur CTu correspond aux cas où la source est certaine de l'issue sans que celle-ci ne soit précisée, comme c'est le cas pour Marie à propos de la venue de Fred en (3a), et que la valeur Uu s'applique lorsque la source (toujours Marie dans les exemples suivants) ne connaît pas la factivité de l'événement (3b-i), n'est pas au courant de l'événement (3b-ii) ou lorsque la factivité n'est pas précisée (3b-iii).

(3) (exemples adaptés de (Saurí, 2008))

a. Marie sait si Fred est $\underline{\text{venu}}_e$. $f(e, Marie) =$ CTu

b. i. Marie ne sait pas si Fred est $\underline{\text{venu}}_e$. $f(e, Marie) =$ Uu

ii. Marie ne sait pas que Fred est $\underline{\text{venu}}_e$. $f(e, Marie) =$ Uu

iii. Marie sait que Paul a dit que Fred était $\underline{\text{venu}}_e$. $f(e, Marie) =$ Uu

Il est vrai que suivant la confiance que Marie accorde à Paul, nous pouvons déduire de la phrase (3b-iii) différentes croyances de Marie sur la venue du Fred. Par exemple, si l'on sait que Marie croit Paul sur parole, alors on peut déduire de cette phrase que Marie croit que Fred est venu (PR+) ; si au contraire Marie pense que Paul a menti, on en déduit l'inverse (PR-). Bien que les *verbes de dire* (*dire*, *affirmer*, etc.) sont généralement interprétés à l'aide de valeurs de factivité positive (de Marneffe *et al.*, 2012), nous pensons que traiter de tels phénomènes pragmatiques nécessiterait une représentation de la confiance entre agents que ni (Saurí, 2008) ni le travail présent ne cherchent à modéliser. Dire que (3b-iii) exprime une factivité Uu pour Marie sur l'éventualité *Fred était venu*, c'est affirmer qu'intrinsèquement cette phrase ne précise pas l'attitude de Marie vis-à-vis de cette éventualité.

3 Analyse discursive

Le travail de (Saurí, 2008) se base essentiellement sur une étude lexicographique, d'expressions appelées *factuality marker* regroupant notamment les auxiliaires modaux, les marqueurs de négation ou de modalité épistémique ainsi que l'ensemble des *Event selecting predicates* (« ESP » ; prédicats de natures diverses se construisant avec un argument dénotant une éventualité sur laquelle est projetée une valeur de factivité ; ex : *to try*, *a promise*). Les interactions par enchâssement de ces différents marqueurs sont analysées et synthétisées dans l'algorithme De Facto. Bien que dans (Saurí, 2008) l'auteur ait décidé de travailler uniquement à l'échelle phrastique, elle souligne l'aspect dynamique de la factivité. En effet, deux phrases d'un même texte peuvent projeter des valeurs de factivité différentes sur la même éventualité ; la seconde venant préciser (4) voire corriger (5) la première.

(4) a. Marie m'a dit qu'il pleuvra$_e$ demain. $f(e, Wr) =$ Uu
b. Je n'y$_e$ crois pas vraiment. $f(e, Wr) =$ PR-

(5) a. Il pleuvra$_e$ peut-être. $f(e, Wr) =$ PS+
b. En fait, il pleuvra$_e$ même à coup sûr. $f(e, Wr) =$ CT+

Dans les deux exemples précédents, les valeurs de factivité peuvent être déterminées en analysant chaque phrase indépendamment, la seconde faisant autorité. Mais cela n'est pas toujours le cas, à cause notamment de la *révision dynamique des informations de factivité* (Danlos, 2011). C'est ce phénomène discursif qui est à l'œuvre dans le discours (6). Alors que sortie de son contexte, la seconde phrase correspond à un engagement de Jane et en aucun cas à un engagement de l'auteur, on infère grâce à l'enchaînement discursif une relation de type $Evidence$ (Asher & Lascarides, 2003) et par là même on comprend que l'auteur adhère aux propos de Jane [1]. S'appuyant sur cette logique, (Danlos & Rambow, 2011) décrit un paradigme dynamique calculant conjointement la factivité des éventualités et la structure discursive du texte analysé. Il permet ainsi de déterminer pour (6) la valeur de factivité $f(e, Wr) =$ CT+ intuitivement attendue.

(6) Les voisins sont partis en vacances. Jane m'a dit que leurs volets sont fermés$_e$.

Cependant, à l'instar de (Saurí, 2008), ce paradigme ne suffit pas, de manière générale, à calculer une valeur de factivité unique pour une éventualité et une source données. En effet, lorsqu'une éventualité est mentionnée plusieurs fois dans le même segment discursif, ou dans une structure discursive sans révision ou correction à son sujet, plusieurs valeurs de factivité pourront lui être attribuées pour une même source, sans qu'il soit précisé comment calculer de valeur au niveau global du texte. Il manque donc un mécanisme général de composition des valeurs de factivité.

De plus, l'ensemble de valeurs de factivité choisi par (Saurí, 2008) pose déjà un problème théorique. En effet, alors que la phrase (7) projette alternativement $f(e, Wr) =$ PS+ et $f(e, Wr) =$ PR−, il semble que la valeur globale de cette factivité ne corresponde à aucune des valeurs proposées par (Saurí, 2008), et en particulier ni à PS+, ni à PR−. C'est pour remédier aux différentes lacunes mentionnées dans cette section que nous proposons une reformalisation de la notion de factivité.

(7) Il est possible qu'il pleuve$_e$ mais je n'y$_e$ crois pas.

4 Reformalisation de la factivité

4.1 Factivité en tant qu'ensemble de probabilités

Rappelons que pour (Saurí, 2008), la factivité est une propriété des *mentions* d'éventualité et non des éventualités elles-mêmes. Or, pour capturer la complexité des phénomènes discursifs, il apparaît naturel de s'intéresser à ce que *représente* la factivité, c'est-à-dire à la *croyance* qu'une source a envers une éventualité donnée.

Ainsi, considérons pour toute source s (un agent à un certain moment) et toute éventualité e, la probabilité $P_s(e) \in [0, 1]$, probabilité que e corresponde au monde réel selon s. *A priori*, nous ne

1. La relation pragmatique *Evidence* permet de faire le lien entre les deux phrases du discours en (6) : l'auteur exprime qu'il déduit l'absence des voisins du fait que leurs volets sont fermés, information qu'il tient de Jane. On remarque qu'ici *Jane m'a dit que* est paraphrasable par *Jane m'a appris que*, qui est factif, tout comme *savoir* ou *découvrir* (Karttunen, 1971).

savons rien d'autre que $P_s(e) \in [0,1]$, mais les différentes mentions de e dans le texte, ainsi que les relations discursives qui le composent, vont nous permettre de restreindre cet ensemble de valeurs. C'est cet ensemble que nous proposons d'appeler factivité. Notons que $f(e,s)$ est alors lié à la probabilité $P_s(e)$ – qui elle est inconnue – par la relation $P_s(e) \in f(e,s) \subseteq [0,1]$ et que sa valeur ne s'applique plus à une mention en particulier mais au texte tout entier [2]. La factivité $f(e,s)$ représente l'imprécision ou l'incertitude du texte sur $P_s(e)$ [3].

Nous interprétons donc les valeurs de factivité comme des sous-ensembles de l'intervalle $[0,1]$. Nous verrons dans la suite du texte comment estimer les ensembles correspondant aux différentes valeurs primitives, mais avant cela, nous présentons les deux opérations, combine et overwrite à partir desquelles s'effectue le calcul de la factivité à l'échelle discursive.

4.2 Opérations sur la factivité

Reprenons l'exemple (7) (répété ci-dessous en (8)). D'après la première mention (α), la probabilité que l'auteur accorde à e se situe dans une certaine plage caractérisée par la possibilité – notons la $f_\alpha(e, Wr)$. D'après la seconde mention (β), cette probabilité se situe dans une autre plage, caractérisée cette fois-ci par l'improbabilité – notons la $f_\beta(e, Wr)$. La probabilité $P_s(e)$ est donc dans l'intersection de ces deux plages de valeurs [4], que de plus l'on déduit être non vide :

$$f(e, Wr) = f_\alpha(e, Wr) \cap f_\beta(e, Wr)$$

(8) (Il est possible qu'il pleuve$_e$)$_\alpha$ mais (je n'y$_e$ crois pas)$_\beta$.

Plus généralement, nous considérons que la factivité d'une éventualité e pour une source s, exprimée par n mentions, est l'intersection des n valeurs de factivité individuelles correspondantes. Ainsi, le calcul séquentiel de la factivité est amené à utiliser l'opération combine, qui précise la valeur courante par intersection avec une nouvelle valeur X :

$$\text{combine}(e, s, X) : f(e,s) \leftarrow (f(e,s) \cap X)$$

D'autre part, à travers certains connecteurs dont ceux de correction tels que *(quoique) non* ou *après tout*, la langue permet d'exprimer des corrections sur ce qui a été dit précédemment dans le discours, comme à propos de la pluie en (9). Il ne s'agit pas alors d'une précision sur la factivité comme dans les cas précédents, mais d'un véritable *écrasement* de la valeur courante, d'où la nécessité d'une seconde opération. C'est à cet effet que nous définissons overwrite :

$$\text{overwrite}(e, s, X) : f(e,s) \leftarrow X$$

2. Contrairement à la logique modale par mondes possibles utilisée pour modéliser la notion de croyance (Hintikka, 1962), dans le système de (Saurí, 2008) – et le nôtre – la vérité d'une proposition ne varie pas et les valeurs de factivité $f(e,s)$ sont les objets de base de la théorie. De même, il ne s'agit pas non plus de logique floue (Hájek, 1998), bien que celle-ci permette de définir des valeurs de vérité pouvant se situer n'importe où dans l'intervalle $[0,1]$. Notre proposition traite de *degrés de croyance* et non de *degrés de vérité* et serait plutôt à rapprocher du *continuum model* présenté dans (de Haan, 1997).

3. Imprécision qui peut être réduite à une unique valeur, comme dans *La probabilité qu'il pleuve$_e$ est très exactement de 85 chances sur 100*.

4. Dans l'exemple (8), nous n'attribuons pas à la relation de contraste lexicalisée par *mais* d'effet sur $f(e, Wr)$. En conséquence, c'est la même factivité qui est exprimée dans la phrase suivante, où les deux termes sont inversés : *Je ne crois pas qu'il pleuve$_e$ mais c'$_e$est possible*. S'il existe une différence de sens entre les deux phrases, elle ne se situe pas au niveau de $f(e, \overline{Wr})$ mais au niveau pragmatique, *via* l'effet asymmétrique de *violation d'attente* de la relation de contraste (Lakoff, 1971). Il est probable que cet effet puisse aussi s'exprimer – au moins partiellement – en termes de factivité (la factivité de l'attente contredite par le second conjoint), nous laissons cependant cet aspect à des recherches futures.

(9) Il pleuvra$_e$ probablement. Quoique non, il ne devrait pas pleuvoir$_e$.

Les deux fonctions combine et overwrite opèrent à partir de valeurs primitives exprimées au niveau lexical pour chaque mention. Ces valeurs primitives sont décrites dans la section suivante.

4.3 Valeurs primitives

Dans (Saurí, 2008), 8 valeurs de factivité sont distinguées sur la base de tests de co-prédication. La formalisation de la factivité que nous proposons permet de donner une explication à ces compatibilités et incompatibilités. En effet, d'après notre interprétation, une factivité $f(e, s)$ contient nécessairement au moins une valeur : $P_s(e)$, la probabilité que l'éventualité e corresponde au monde réel d'après s. Par définition, $f(e, s)$ ne peut donc être réduite à l'ensemble vide – du moins dans un discours cohérent. La possibilité de co-prédication entre deux expressions nous informe donc que les valeurs de factivité qu'elles expriment ont une intersection non nulle (10), et à l'inverse qu'il s'agit de deux ensembles disjoints lorsque toute co-prédication est impossible (11).

(10) a. (Je crois qu'il pleuvra$_e$ demain)$_\alpha$ mais (ce$_e$ n'est pas sûr)$_\beta$.
b. $f_\alpha(e, Wr) \cap f_\beta(e, Wr) \neq \emptyset$

(11) a. # (Je crois qu'il pleuvra$_e$ demain)$_\alpha$ mais (ça$_e$ n'arrivera probablement pas)$_\beta$.
b. $f_\alpha(e, Wr) \cap f_\beta(e, Wr) = \emptyset$

Les 8 valeurs CT±, PR±, PS±, CTu et Uu distinguées dans (Saurí, 2008) correspondent donc à autant de parties de $[0, 1]$, qu'il est possible d'estimer. Notons que ce choix de valeurs avait été fait à partir de considérations pratiques d'annotation et guidé par une approche purement statique de la factivité. La formalisation que nous proposons offre la possibilité de moduler cet ensemble de valeurs primitives en fonction de la finesse de l'analyse attendue, tout en restant dans le même cadre théorique. Une modification en particulier nous paraît nécessaire pour rendre compte de certains discours. Celle-ci consiste à réserver la valeur de sous-spécification Uu aux mentions où *aucune information* de factivité n'est effectivement donnée pour la source concernée, comme pour $s = Wr$ dans (12a-i) [5]. Une nouvelle valeur ¬CT est alors utilisée pour les cas exprimant *l'absence de certitude* de la source (12a-ii). La continuation avec une mention de factivité CT+ (12b), possible pour Uu mais non pour ¬CT, est l'un des tests permettant de discriminer ces deux valeurs. En termes d'ensembles de probabilités, celui qui représente l'absence totale d'information est Uu $= [0, 1]$. Concernant ¬CT, affirmons seulement pour l'instant ¬CT $\subseteq [0, 1] \setminus (\text{CT}+ \cup \text{CT}-)$.

(12) a. i. Fred dit qu'il a$_e$ le permis. $f(e, Wr) =$ Uu
ii. Je ne sais pas si Fred a$_e$ le permis. $f(e, Wr) =$ ¬CT
b. Comme il a$_e$ le permis, je vais lui demander de m'emmener en ville. $f(e, Wr) =$ CT+

À l'aide de tests de co-prédication, nous pouvons étendre les tableaux de compatibilité donnés dans (Saurí, 2008). Dans la table 2, le symbole + à l'intersection de la ligne « CT± » et de la colonne « PR = » indique que les valeurs CT+ et PR+ (ainsi que CT− et PR−) sont compatibles ; le symbole − à l'intersection de la ligne « CT± » et de la colonne « ¬CT » indique que ni CT+ ni CT− ne sont compatibles avec ¬CT.

5. Il s'agit aussi de la valeur par défaut : dans (12a-i) par exemple, on a $f(e, Fred) =$ CT+, $f(e, Wr) =$ Uu, mais aussi pour toute autre source s, $f(e, s) =$ Uu.

Interprétées en termes d'intersections, ces compatibilités ne suffisent pas à déterminer sans ambiguïté à quel ensemble correspond chaque valeur. La figure 1 présente cependant une solution possible, exhibant des propriétés de symétrie : les valeurs de polarité $+$ ou $-$ sont les symétriques des valeurs de polarité opposée (par exemple $\text{PR}+ = \{(1-x) \mid x \in \text{PR}-\}$) alors que les valeurs non polarisées sont symétriques par rapport à la probabilité 0.5. Soulignons que ce choix, bien que cohérent, est arbitraire et est donné ici à titre illustratif ; une étude psycholinguistique des locuteurs natifs permettrait de déterminer avec plus de rigueur un ensemble de valeurs.

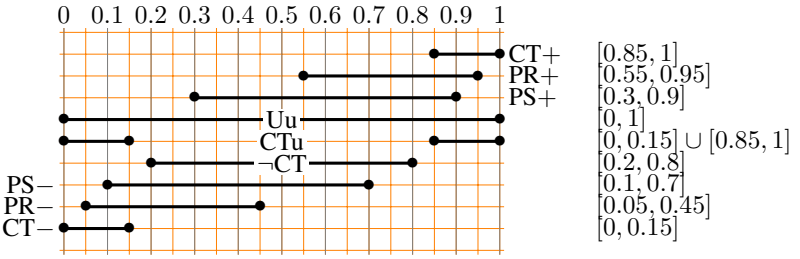


FIGURE 1 – Une interprétation ensembliste des valeurs primitives.

5 Conclusion et perspectives

Nous avons montré que la formalisation de la factivité proposée dans (Saurí, 2008) n'était pas adaptée à l'analyse discursive. Nous avons alors décrit une nouvelle formalisation fondée sur une interprétation probabiliste qui explique les compatibilités et incompatibilités entre valeurs de factivité, jusque là simplement observées.

La principale nouveauté de cette formalisation est de définir une valeur de factivité *unique* pour chaque paire éventualité-source, cohérente à l'échelle discursive. De plus, cette valeur n'est plus restreinte à un ensemble fixé de valeurs catégoriques ; notre cadre théorique permet d'exprimer naturellement par intersection des combinaisons de valeurs primitives.

L'un des prochains objectifs de notre recherche est d'adapter De Facto et d'implémenter le paradigme présenté dans (Danlos & Rambow, 2011) pour étudier en corpus les effets de la structure discursive sur les informations de factivité, en particulier le phénomène de révision dynamique introduit par (Danlos, 2011).

	PS=	PS≠	PR=	PR≠	CT=	CT≠	CTu	¬CT	Uu
PS±	+	+	+	+	+	-	+	+	+
PR±			+	-	+	-	+	+	+
CT±					+	-	+	-	+
CTu							+	-	+
¬CT								+	+
Uu									+

TABLE 2 – Compatibilités des valeurs primitives.

Références

ASHER N. & LASCARIDES A. (2003). *Logics of Conversation*. Cambridge University Press.

DANLOS L. (2011). Analyse discursive et informations de factivité. In *Actes de la 18e conférence sur le Traitement Automatique des Langues Naturelles*, volume 1, p. 445–456, Montpellier, France.

DANLOS L. (2012). Formalisation des conditions d'emploi des connecteurs en réalité et (et) en effet. In *Actes du 2ème Congrès Mondial de Linguistique Française*, p. 493–508, Lyon, France.

DANLOS L. & RAMBOW O. (2011). Discourse Relations and Propositional Attitudes. In *Proceedings of CID 2011*, Agay, France.

DE HAAN F. (1997). *The Interaction of Modality and Negation : a Typological Study*. Outstanding Dissertations in Linguistics. New York, USA : Garland.

DE MARNEFFE M.-C., MANNING C. D. & POTTS C. (2012). Did It Happen ? The Pragmatic Complexity of Veridicality Assessment. *Computational Linguistics*, **38**(2), 301–333.

FARKAS R., VINCZE V., MÓRA G., CSIRIK J. & SZARVAS G. (2010). The CoNLL-2010 Shared Task : Learning to Detect Hedges and Their Scope in Natural Language Text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning — Shared Task*, CoNLL '10 : Shared Task, p. 1–12, Stroudsburg, PA, USA : Association for Computational Linguistics.

HICKL A. & BENSLEY J. (2007). A discourse commitment-based framework for recognizing textual entailment. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, p. 171–176, Prague, Czech Republic : Association for Computational Linguistics.

HINTIKKA J. (1962). *Knowledge and Belief - An Introduction to the Logic of the Two Notions*. Ithaca : Cornell University Press.

HORN L. R. (1972). *On the Semantic Properties of Logical Operators in English*. Ph.D. Thesis, Yale University.

HÁJEK P. (1998). *Metamathematics of Fuzzy Logic*. Springer Science & Business Media.

KARTTUNEN L. (1971). Some observations on factivity. *Paper in Linguistics*, **4**(1), 55–69.

KIM J.-D., OHTA T., PYYSALO S., KANO Y. & TSUJII J. (2009). Overview of BioNLP'09 Shared Task on Event Extraction. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing : Shared Task*, BioNLP '09, p. 1–9, Stroudsburg, PA, USA : Association for Computational Linguistics.

LAKOFF R. (1971). If's, and's and but's about conjunction. In C. FILLMORE & D. T. LANGENDOEN, Eds., *Studies in Linguistic Semantics*, p. 115–150. New York : Holt, Rinehart and Winston.

MORANTE R. & SPORLEDER C. (2010). *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*. Upsala, Sweden : University of Antwerp.

SAURÍ R. (2008). *A Factuality Profiler for Eventualities in Text*. Ph.D. Thesis, Brandeis University, Waltham, MA, USA.

SAURÍ R. & PUSTEJOVSKY J. (2012). Are You Sure That This Happened ? Assessing the Factuality Degree of Events in Text. *Computational Linguistics*, **38**(2), 261–299.

SONI S., MITRA T., GILBERT E. & EISENSTEIN J. (2014). Modeling Factuality Judgments in Social Media Text. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume Volume 2 : Short Papers, p. 415–420, Baltimore, Maryland, USA.

Un algorithme universel pour l'abstraction automatique d'alternances morphophonologiques

Sacha Beniamine
(1) Laboratoire de Linguistique Formelle, Labex EFL, Université Paris Diderot 7, France
sbeniamine@linguist.univ-paris-diderot.fr

RÉSUMÉ

Cet article présente un algorithme implémenté pour l'inférence de patrons d'alternances morphophonologiques entre mots-formes. Il est *universel* au sens où il permet d'obtenir des classifications comparables d'une langue à l'autre sans préjuger des types d'alternances. Les patrons constituent une première étape pour les travaux quantitatifs dans l'approche Mot et Paradigme de la morphologie.

ABSTRACT

A universal algorithm for the automatical abstraction of morphophonological alternations.

We present an implemented algorithm for the inference of morphophonological alternation patterns between word-forms. It is *universal* in that it leads to comparable classifications across languages without expectations on the shape of the alternations it searches for. Alternation patterns are a necessary first step for the quantitative study of morphology in the Word and Paradigm framework.

MOTS-CLÉS : Flexion, morphophonologie, linguistique quantitative, typologie.

KEYWORDS: Inflection, morphophonology, quantitative linguistics, typology.

1 Introduction

L'étude quantitative et typologique de la morphologie présuppose l'extraction des données morphophonologiques à partir de lexiques. Les travaux existant en morphologie quantitative reposent sur des grammaires élaborées manuellement (Ackerman & Malouf, 2013; Stump & Finkel, 2013) difficilement comparables d'une langue à l'autre, ou sur des heuristiques fixant des biais linguistiquement motivés qui ne se prêtent pas aux comparaisons typologiques (Albright & Hayes, 2002).

Cet article présente un algorithme pour l'abstraction de patrons morphophonologiques permettant de caractériser des alternances entre mots-formes. Il est implémenté et distribué sous licence libre [1]. Il est *universel* au sens où il permet d'obtenir des classifications comparables d'une langue à l'autre sans préjuger des types d'alternances. Il a été développé dans le cadre de l'approche « mot et paradigme » pour l'analyse des paradigmes de flexion et peut être utilisé pour décrire d'autres alternances morphophonologique entre mots-formes.

Nous utilisons le vocabulaire standard en morphologie flexionnelle (Matthews, 1991). Un paradigme est l'ensemble des formes d'un lexème, structuré en fonction des propriétés morphosyntaxiques exprimées par ces formes (par exemple « présent première personne » ou PRS.1SG). On appelle *case*

1. Documentation et téléchargement : `http://drehu.linguist.univ-paris-diderot.fr/qumin/`. Le programme est écrit en Python3.

du paradigme un ensemble de propriétés à exprimer. Dans ce contexte, un système flexionnel peut être représenté sous la forme d'une table à double entrée, où chaque ligne représente le paradigme d'un lexème et chaque colonne une case de paradigme.

Les modèles « Mot et Paradigme » (WP) suivent la tradition grammaticale européenne et considèrent le mot et non le morphème comme l'unité fondamentale de la morphologie (Hockett, 1954; Robins, 1959). Ces modèles décrivent traditionnellement les systèmes de flexion en se fondant sur l'analogie à un petit nombre de paradigmes exemplaires. Les développements récents mettent au centre de l'attention les relations implicatives entre cases du paradigme (Wurzel, 1989) et la mesure quantitative de la fiabilité de ces implications (Ackerman *et al.*, 2009; Sims, 2010; Ackerman & Malouf, 2013; Blevins, 2016; Ackerman & Malouf, 2016). La formulation précise de ces implications nécessite d'identifier des patrons d'alternance entre formes de surface (Bonami, 2014).

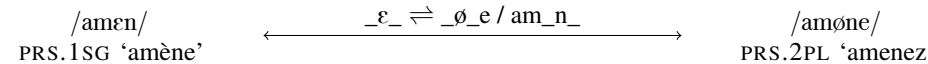


FIGURE 1 – Un patron d'alternance

Les patrons d'alternances tels que nous les formulons ici prennent la forme d'alternances bidirectionnelles du type « X *alterne avec* Y *dans le contexte* Z », noté $X \rightleftharpoons Y/Z$, comme illustré en figure 1. X et Y constituent l'alternance et Z constitue le contexte d'application du patron.

L'une des questions centrales étudiées dans ce cadre est celle du problème de remplissage des cases de paradigme (« Paradigm Cell Filling Problem » ou PCFP). Étant donnée la distribution zipfienne des mots, comment les locuteurs qui n'ont rencontré que certaines formes d'un lexème peuvent-ils produire les formes inconnues ? Ackerman *et al.* (2009) évaluent la difficulté de ce problème à travers l'entropie des distributions de patrons d'alternance dans les systèmes flexionnels. Les résultats de ce type de calcul dépendent crucialement de la façon dont les patrons sont inférés.

Nous présentons d'abord (§2) les stratégies existantes pour extraire des alternances morphophonologiques. Puis (§3) nous présentons le problème comme une recherche d'alignements entre formes qui soient optimaux à la fois localement à ces formes et globalement à travers le système. L'algorithme (§4) procède en trois étapes : aligner localement les formes, généraliser les patrons, sélectionner les patrons globalement. Enfin nous évaluons la qualité des généralisations capturées par les patrons (§5) dans le cadre d'un problème du type PCFP idéalisé.

2 État de l'art

Deux traditions ont contribué à l'inférence automatique à grande échelle de règles morphologiques à même de capter les implications paradigmatiques. En linguistique, le *Minimal Generalization Learner* (MGL, Albright & Hayes 2002, 2003, 2006) constitue le premier modèle quantitatif des tests psycholinguistiques d'élicitations morphologiques dits *wug tests*, et a inspiré des travaux qui s'appuient sur des biais linguistiquement motivés pour inférer efficacement des patrons d'alternance dans un ensemble de données spécifique. En TAL, le PCFP se pose dans le contexte de la complétion automatique de lexiques et s'est traduit par la formulation d'une tâche de réinflection.

Le MGL est un programme qui infère incrémentalement, à partir de couples de formes, un ensemble

de patrons d'alternance unidirectionnels. Il infère d'abord des patrons spécifiques à une seule paire de formes, puis crée des patrons de généralité croissante au fil de la découverte de nouvelle formes. Pour chaque paire de formes dans les données d'apprentissage, on obtient un ensemble de patrons de généralité variable à même de les dériver. L'ensemble de ces règles sont scorées pour l'application aux formes inconnues. Les patrons générés par le MGL sont capables de découvrir un changement unique dans les formes d'entrées, cherchant dans l'ordre un changement suffixal, préfixal puis interne.

Bonami & Boyé (2014); Bonami & Luís (2014) combinent l'idée directrice du MGL et celle d'Ackerman *et al.* (2009), en évaluant la prédictibilité au sein des paradigmes au moyen de l'entropie conditionnelle sur la base de patrons d'alternance bidirectionnels. Un seul patron est alors associé à toute paire de cases afin de fournir une classification des lexèmes. Ce choix permet d'appliquer ces programmes à des larges ensembles de données et nombre d'alternances, ce qui n'était pas envisageable pour le MGL. Pour identifier les alternances, ces modèles s'appuient sur des biais adaptés à un ou plusieurs ensembles de données. Ces stratégies nécessitent la conception d'un nouveau programme *ad hoc* pour chaque nouveau système. Elles ne se prêtent donc pas à la typologie quantitative.

L'extraction de lexiques à partir de corpus produit des paradigmes incomplets en raison de la distribution zipfienne des formes. Des efforts se sont donc tournés vers l'inférence des formes manquantes en s'appuyant sur la structure implicative des paradigmes (Durrett & DeNero, 2013; Ahlberg *et al.*, 2014; Nicolai *et al.*, 2015). La campagne d'évaluation SIGMORPHON (Cotterell *et al.*, 2016) a formalisé une famille de tâches de TAL dites de « réinflection » répondant à cette manifestation concrète du PCFP. Le problème qui nous intéresse le plus consiste à prédire une forme cible à partir d'une forme déjà fléchie. Les approches par réseaux neuronaux sont les plus performantes (Kann & Schütze, 2016). D'autres systèmes, suivant Durrett & DeNero (2013), extraient des opérations d'édition de chaînes puis entraînent des transducteurs à les appliquer. Enfin certains reposent sur des heuristiques linguistiques et extraient soit des segmentations affixales (Taji *et al.*, 2016) soit des « paradigmes abstraits », règles morphologiques formulées pour l'ensemble du paradigme (Sorokin, 2016; Ahlberg *et al.*, 2014). Le meilleur modèle parvient à 95% d'exactitude en s'entraînant uniquement sur des paires de formes, montrant que l'information nécessaire à la résolution du PCFP peut être extraite de données éparses.

Cependant, aucun de ces systèmes ne permet d'extraire le type de représentations dont nous avons besoin, c'est-à-dire des patrons d'alternance entre deux cases, interprétables de façon transparente comme des règles linguistiques, calculables pour toutes les paires de cases sur de larges lexiques, ne s'appuyant pas sur des données externes à la paire de case concernée, et fournissant une classification catégorique des lexèmes selon leur comportement morphophonologique. Nous proposons donc une méthode à même d'inférer automatiquement de telles représentations.

3 Les patrons dépendent d'alignements entre formes

Déterminer le patron qui relie une paire de formes revient à trouver un alignement optimal entre ces formes : si deux segments alignés ensemble sont identiques, ils font partie du contexte, sinon ils font partie de l'alternance. Cet alignement ne peut pas, dans une perspective multi-lingue, être fixé à l'avance. Différents types d'alternances requièrent des stratégies d'alignements distinctes. Les alternances suffixales, par exemple dans la flexion verbale du français (figure 2), sont captées par un alignement à gauche. Une alternance préfixale nécessiterait un alignement à droite. Certaines alternances discontinues nécessitent un alignement qui ne peut être pré-défini (flexion verbale de

l'arabe, figure 2).

Français : 'amène' ⇌ 'amenez'

PRS.1SG	a	m	ɛ	n	
	|	|	|	|	|
PRS.2PL	a	m	ø	n	e

Arabe : 'il a écrit' ⇌ 'il écrit'

PFV.M.3SG			k	a ː	t	a	b	a
	|	|	|	|	|	|	|	|
IPF.M.3SG	j	u	k	a ː	t	i	b	u

FIGURE 2 – Les alignements optimaux dépendent du type d'alternance.

Il existe parfois plusieurs alignements plausibles (Bonami, 2014, pp.104-106). La figure 3 présente trois alignements des formes imaginaires /baba/ et /ba/ et trois patrons qui offrent respectivement une analyse préfixale, suffixale ou infixale. Seul l'examen d'autres formes permet de décider de l'analyse préférable. Pour chacun des systèmes A, B et C de la figure 3, une seule des trois analyses, respectivement infixe, préfixe et suffixe, permet de rendre compte de l'ensemble des lexèmes.

Alignement					Patron	A.		B.		C.	
	b	a	b	a		SG	PL	SG	PL	SG	PL
	b	a	b	a		ba	baba	ba	baba	ba	baba
A.Infixe	b	_	_	a	ϵ ⇌ ab / b_a	ri	rabi	ri	bari	ri	riba
B.Préfixe	_	_	b	a	ϵ ⇌ ba / _ba	su	sabu	su	basu	su	suba
C.Suffixe	b	a	_	_	ϵ ⇌ ba / ba_	ne	nabe	ne	bane	ne	neba

FIGURE 3 – Trois alignements et systèmes imaginaires.

4 Méthode

Notre système prend en entrée des paradigmes dont les formes sont transcrites en notation phonétique ainsi qu'une spécification de la valeur des phonèmes utilisés en traits. Il calcule, pour chaque paire de cases possible parmi l'ensemble des cases du paradigme, un ensemble de patrons d'alternance qui relient les paires de formes. Malgré la complexité de cette combinatoire, les besoins de l'analyse linguistique de type Mot et Paradigme imposent de calculer les patrons pour toutes les paires de cases. Cependant chaque patron se fonde strictement sur la connaissance des formes des deux cases qu'il relie.

L'algorithme procède en trois étapes : produire pour chaque paire de formes l'ensemble des alignements localement optimaux et en déduire des patrons élémentaires ; généraliser les patrons élémentaires en fusionnant les alternances structurellement identiques ; choisir les patrons selon leur pouvoir descriptif global pour l'ensemble des lexèmes.

Nous commençons par trouver pour chaque paire de formes un ensemble d'alignements qui minimisent une distance d'édition. Nous considérons deux mesures de distance. La plus simple est la distance de Levenshtein, qui maximise les segments identiques alignés ensemble. Les trois alignements linguistiquement motivés entre /ba/ et /baba/ sont aussi ceux qui présentent la plus petite distance de Levenshtein. Considérons cependant l'alternance en Tchèque entre le nominatif singulier « čivava » prononcé /tʃɪvava/ et le génitif pluriel du nom « čivav » (chihuahua), prononcé /tʃɪvaf/. L'alignement intuitif pour un linguiste, en raison du dévoisement final, aligne [f] avec le deuxième

[v]. Cependant, il existe un deuxième alignement, non linguistiquement fondé, qui suppose un infixe /-av-/ et présente la même distance de Levenshtein.

Afin d'éviter ce type d'alignement fallacieux, et d'évaluer leur impact, nous implémentons également une seconde distance d'édition pondérée par la similarité phonologique (Albright & Hayes, 2006). (Frisch *et al.*, 2004) proposent une similarité phonologique fondée sur la proportion de classes naturelles partagées entre deux segments. En phonologie, une classe naturelle est un ensemble de segments phonologiques partageant un ou plusieurs traits (Chomsky & Halle, 1968). Les segments d'une même classe se comportent identiquement relativement à certains processsus phonologiques. Soit $C(a)$ l'ensemble des classes naturelles d'un segment a, la similarité phonologique peut se définir $\text{sim}(a,b) = \frac{|C(a) \cap C(b)|}{|C(a) \cup C(b)|}$. Nous fixons le coût de la substitution à $1 - \text{sim}(a,b)$. Le coût de l'insertion est un paramètre que nous fixons à $\frac{1}{3}$ du coût moyen de la substitution dans le système phonologique. Soit I l'ensemble des phonèmes connus et $I \times I$ le produit cartésien sur cet ensemble, le coût moyen d'une substitution dans ce système est $\frac{1}{I \times I} \cdot \sum_{(a,b) \in I \times I} 1 - \text{sim}(a,b)$.

Nous obtenons un ensemble de patrons concurrents pour chaque paire de formes. L'alignement de /ba/ et /baba/ produit l'ensemble des trois patrons : {$\epsilon \rightleftharpoons$ ba / _ba, $\epsilon \rightleftharpoons$ ba / ba_, $\epsilon \rightleftharpoons$ ab / b_a}. Ces patrons sont spécifiques aux formes dont ils proviennent, car leurs contextes contiennent des segments phonologiques spécifiques.

La seconde étape consiste à fusionner les patrons présentant la même alternance structurelle de façon à capter des généralisations sur les contextes d'application et sur les alternances. Les contextes sont alors exprimés sous la forme d'expressions régulières où les classes de caractères, entre crochets, correspondent à des classes naturelles de segments phonologiques et les quantificateurs « ?, +, * » rendent compte de la longueur des séquences concernées. Nous fusionnons enfin les patrons dont les alternances constituent différentes instanciations d'une même opération phonologique régulière.

Alternances	Contextes							Lexèmes
ɛ _ ⇌ ə_õ /			a	m	_	n	_	AMENER
ɛ _ ⇌ ə_õ /	p	ʁ	o	m	_	n	_	PROMENER
ɛ _ ⇌ ə_õ /				s	_	v ʁ	_	SEVRER
	{p} ?	{ʁ} ?	{o,a} ?	{m,s}	_	{n,v,ʁ}+	_	(i)
	{p,ʁ,o,a}*			{m,s}	_	{n,v,ʁ}+	_	(ii)
ɛ _ ⇌ ə_õ /	[ɛɔɔøœabdflpstvzəʁ]*			[flmnsvz]	_	[lmnvzʁ] +	_	(iii)

TABLE 1 – Généralisation du contexte de trois patrons

La généralisation des contextes suit la procédure décrite par Albright & Hayes (2002), adaptée pour une généralisation n-par-n plutôt que 2 par 2. Étant donné un ensemble de patrons partageant une même alternance structurelle, on souhaite déterminer ce que ces contextes ont en commun. Pour cela, on aligne les segments à travers les contextes, puis les segments alignés sont traduits sous la forme d'une expression régulière.

La table 1 présente trois patrons d'alternance dont les contextes doivent être fusionnés. (i) Les séquences sont d'abord alignées du côté des blancs : à gauche du premier blanc, les séquences « am », "pʁom et « s » sont alignées à gauche. Entre les deux blancs, les séquences « n »,« n » et « vʁ » sont alignées au centre. (ii) Les ensembles de segments alignés facultatifs consécutifs sont ensuite fusionnés. (iii) Pour chaque ensemble de segments alignés, on détermine l'ensemble des segments qui partagent les mêmes traits phonologiques. On en déduit une expression régulière qui caractérise

Langue	Lexique	Traits phonologiques
Français	Vlexique (Bonami *et al.*, 2014)	Dell 1973
Anglais	CELEX2 (Baayen *et al.*, 1995)	Halle & Clements 1983
Chatino	Oto-Manguean Inflectional Class Database (Feist & Palancar, 2015)	Campbell 2014
portugais	Dictionnaire de prononciation (Veiga *et al.*, 2013)	Bonami & Luís 2014
Arabe	Unimorph (Kirov *et al.*, 2016)	Hayes 2012

TABLE 2 – Sources pour les ensembles de données utilisés dans l'évaluation.

les trois contextes.

Nous obtenons plusieurs patrons concurrents pour une même paire de formes, parmi lesquels il nous faut choisir celui qui est le plus approprié au vu du reste du système. Pour cela, nous attribuons des scores aux patrons. Intuitivement, un patron est meilleur s'il est souvent correct et un patron général est préférable à un patron spécifique. La précision d'un patron pour son application dans une direction $C_1 \rightarrow C_2$ est le rapport entre le nombre de formes appartenant à la case C_1 dont il dérive correctement la forme pour la case C_2 sur le nombre de formes appartenant à la case C_1 auxquelles ce patron est applicable. Sa couverture est le rapport entre le nombre de formes de la case C_1 auxquelles il est applicable et le nombre de formes total. Les patrons, qui sont bidirectionnels, sont scorés par la moyenne harmonique des couvertures et précisions dans les deux sens d'application. La généralisation du patron infixe de la figure 3 produira un patron du type $\epsilon \rightleftharpoons \text{ab}/C_V$. Il aura un score de 1, car il dérive toutes les formes correctement. Les deux autres patrons ne sont pas généralisables et obtiendront un score inférieur.

Pour chaque paire de formes, nous choisissons parmi l'ensemble des patrons qui les dérivent correctement celui dont le score est le meilleur. Ce patron peut avoir été inféré à partir d'autres paires de formes. Ceci permet d'obtenir des patrons robustes à bas coût computationnel.

5 Évaluation

Notre algorithme vise à la comparaison quantitative de systèmes flexionnels. La pertinence des patrons inférés se jugera donc *in fine* aux généralisations trans-linguistiques qu'ils permettront de dégager. Cependant, leur capacité à capter des généralisations internes à un système peut d'ores et déjà être évaluée au travers d'une tâche de prédiction.

Nous avons mené une évaluation pour les systèmes verbaux du français, de l'anglais, du portugais européen, du zenzontepec chatino et de l'arabe classique. Les tables de paradigmes sont issues de lexiques flexionnels. Au besoin, nous avons transcrit automatiquement les formes en alphabet phonétique international depuis d'autres notations (zenzontepec chatino, anglais, arabe). La table 2 présente leur sources ainsi que celles des traits distinctifs. La première partie de la table 3 rapporte la taille des paradigmes.

Nous avons mené une validation croisée en 10 plis. L'ensemble des patrons est chaque fois appris sur les formes d'entraînement. Puis, sur ces mêmes données, nous déterminons les classes de patrons applicables (Bonami, 2014), c'est-à-dire les ensembles de patrons susceptibles d'être appliqués à chaque forme. Nous calculons la probabilité conditionnelle $P(\text{patron}|\text{classe})$ d'après les fréquences relatives des patrons et des classes ainsi que la probabilité simple $P(\text{patron})$ pour servir de repli. Lors de l'évaluation, nous considérons des formes uniques, dont nous connaissons la case de paradigme, et

	Chatino	Français	portugais	Arabe	Anglais
Lexèmes d'entraînement	352	4688	1797	576	5450
Lexèmes d'évaluation	39	521	199	64	605
Cases de paradigme	4	51	69	109	8
Alignement à gauche (suffixe)	23.39%	94.35%	93.92%	40.53%	94.03%
Alignement à droite (préfixe)	54.12%	23%	18.04%	24.27%	31.09%
Albright & Hayes 2002	53.67%	94.35%	93.67%	42%	94.01%
Distances d'édition simples	**57.37%**	**94.55%**	**92.77%**	**82.42%**	**94.18%**
Distances d'édition pondérées	**57.45%**	**94.53%**	**92.86%**	**82.58%**	**94.18%**

TABLE 3 – Résultats de l'évaluation : pourcentage d'exactitude moyenne.

tentons de prédire la forme cible dans la case cible. À partir de l'ensemble des patrons applicables à cette forme, nous choisissons le patron qui maximise $P(\text{patron}|\text{classe})$ si la classe est connue, sinon celui qui maximise $P(\text{patron})$.

Nous comparons les résultats du même algorithme en faisant varier la stratégie d'alignement (table 3). Afin de comparer notre proposition à des alternatives plus simples, nous évaluons les alignements fixes à gauche ou à droite ainsi que l'alignement décrit dans Albright & Hayes (2002). Nous évaluons également les deux alignements fondés sur des distances d'édition décrits dans cet article. De façon générale, les résultats sont nettement meilleurs sur les lexiques de taille importante (français, portugais, anglais) que sur les plus petits échantillons (chatino, arabe). Comme attendu, les stratégies d'alignement fixe sont efficaces exclusivement pour certains ensembles de données. Seul l'alignement par distances d'éditions produit de bons résultats sur l'ensemble des langues. Nos résultats ne montrent pas de différence importante entre les deux distances d'édition. La distance de Levenshtein a l'avantage de la simplicité. Elle produit initialement plus de patrons concurrents, ce qui conduit à un temps d'exécution plus long. Cependant l'algorithme est robuste à ce bruit, la sélection des patrons se faisant *in fine* selon leur adaptation à l'ensemble du paradigme.

6 Conclusion

Nous avons présenté un algorithme permettant d'inférer des patrons d'alternances à partir de paires de formes. Cet algorithme est robuste et applicable à travers les langues sans présumer du type d'alternances. Il peut constituer le fondement d'une analyse quantitative des alternances morphophonologiques, aussi bien dans le domaine de la flexion, pour lequel il a été développé, que pour l'étude des changements diachroniques, dialectaux, et des alternances dérivationnelles.

Remerciements

Nous remercions les trois relecteurs anonymes pour leur nombreuses remarques, Olivier Bonami pour ses conseils et suggestions, Enrique Palancar pour la préparation des données du chatino et Kenza Ouldhamouda pour la validation manuelle de la transcription des données de l'arabe.

Références

ACKERMAN F., BLEVINS J. P. & MALOUF R. (2009). *Parts and wholes : Patterns of relatedness in complex morphological systems and why they matter*, In J. P. BLEVINS & J. BLEVINS, Eds., *Analogy in Grammar : Form and Acquisition*, p. 54–82. Oxford University Press.

ACKERMAN F. & MALOUF R. (2013). Morphological organization : The low conditional entropy conjecture. *Language*, **89**(3), 429–464.

ACKERMAN F. & MALOUF R. (2016). Word and pattern morphology : An information-theoretic approach. *Word Structure*, **9**(2), 125–131.

AHLBERG M., FORSBERG M. & HULDEN M. (2014). Semi-supervised learning of morphological paradigms and lexicons. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, Gothenburg, Sweden 26–30 April 2014*, p. 569–578.

ALBRIGHT A. & HAYES B. (2002). Modeling english past tense intuitions with minimal generalization. In *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning - Volume 6*, MPL '02, p. 58–69, Stroudsburg, PA, USA : Association for Computational Linguistics.

ALBRIGHT A. & HAYES B. (2003). Rules vs. analogy in english past tenses : A computational/experimental study. *Cognition*, **90**, 119–161.

ALBRIGHT A. & HAYES B. (2006). Modeling productivity with the gradual learning algorithm : The problem of accidentally exceptionless generalizations. *Gradience in grammar : Generative perspectives*, p. 185–204.

BAAYEN R., PIEPENBROCK R. & GULIKERS L. (1995). Celex2.

BLEVINS J. P. (2016). *Word and Paradigm Morphology*. Oxford University Press (OUP).

BONAMI O. (2014). La structure fine des paradigmes de flexion. Mémoire d'habilitation U. Paris Diderot.

BONAMI O. & BOYÉ G. (2014). De formes en thèmes. In F. VILLOING, S. LEROY & S. DAVID, Eds., *Foisonnements morphologiques. Etudes en hommage à Françoise Kerleroux*, p. 17–45. Presses Universitaires de Paris Ouest.

BONAMI O., CARON G. & PLANCQ C. (2014). Construction d'un lexique flexionnel phonétisé libre du français. In F. NEVEU, P. BLUMENTHAL, L. HRIBA, A. GERSTENBERG, J. MEINSCHAEFER & S. PRÉVOST, Eds., *Actes du quatrième Congrès Mondial de Linguistique Française*, p. 2583–2596.

BONAMI O. & LUÍS A. R. (2014). Sur la morphologie implicative dans la conjugaison du portugais : une étude quantitative. In J.-L. LÉONARD, Ed., *Morphologie flexionnelle et dialectologie romane. Typologie(s) et modélisation(s).*, number 22 in Mémoires de la Société de Linguistique de Paris, p. 111–151. Leuven : Peeters.

CAMPBELL E. (2014). *Aspects of the phonology and morphology of Zenzontepec Chatino, a Zapotecan language of Oaxaca, Mexico*. PhD thesis, University of Texas at Austin.

CHOMSKY N. & HALLE M. (1968). *The Sound Pattern of English*. Studies in language. New York, NY : Harper & Row.

COTTERELL R., KIROV C., SYLAK-GLASSMAN J., YAROWSKY D., EISNER J. & HULDEN M. (2016). The sigmorphon 2016 shared task—morphological reinflection. In *Proceedings of the 14th Annual SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, p. 10–22, Berlin, Germany : Association for Computational Linguistics.

DELL F. (1973). *Les règles et les sons*. Collection Savoir. Hermann.

DURRETT G. & DENERO J. (2013). Supervised learning of complete morphological paradigms. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 1185–1195, Atlanta, Georgia : Association for Computational Linguistics.

FEIST T. & PALANCAR E. L. (2015). Oto-manguean inflectional class database. University of Surrey.

FRISCH S. A., PIERREHUMBERT J. B. & BROE M. B. (2004). Similarity avoidance and the ocp. *Natural Language & Linguistic Theory*, **22**(1), 179–228.

HALLE M. & CLEMENTS G. N. (1983). *Problem Book in Phonology : A Workbook for Introductory Courses in Linguistics and in Modern Phonology*. A Bradford book. A Bradford.

HAYES B. (2012). Spreadsheet with segments and their feature values. Distributed as part of course material for Linguistics 120A : Phonology I at UCLA. These are the features used in FeaturePad.

HOCKETT (1954). two models of grammatical description.

KANN K. & SCHÜTZE H. (2016). Med : The lmu system for the sigmorphon 2016 shared task on morphological reinflection. *ACL 2016*, p.62.

KIROV C., SYLAK-GLASSMAN J., QUE R. & YAROWSKY D. (2016). Very-large scale parsing and normalization of wiktionary morphological paradigms. In N. C. C. CHAIR), K. CHOUKRI, T. DECLERCK, S. GOGGI, M. GROBELNIK, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK & S. PIPERIDIS, Eds., *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France : European Language Resources Association (ELRA).

MATTHEWS P. H. (1991). *Morphology*. Cambridge University Press, second edition. Cambridge Books Online.

NICOLAI G., CHERRY C. & KONDRAK G. (2015). Inflection generation as discriminative string transduction. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 922–931, Denver, Colorado : Association for Computational Linguistics.

ROBINS R. H. (1959). In defence of WP. *Transactions of the Philological Society*, **58**(1), 116–144.

SIMS A. (2010). Probabilistic paradigmatics : Principal parts, predictability and (other) possible pieces of the puzzle. In *14th International Morphology Meeting, Budapest*, p. 13–16.

SOROKIN A. (2016). Using longest common subsequence and character models to predict word forms. *ACL 2016*, p.54.

STUMP G. & FINKEL R. (2013). *Morphological Typology : From Word to Paradigm*. Cambridge Studies in Linguistics. Cambridge University Press.

TAJI D., ESKANDER R., HABASH N. & RAMBOW O. (2016). The columbia university-new york university abu dhabi sigmorphon 2016 morphological reinflection shared task submission. *ACL 2016*, p.71.

VEIGA A., CANDEIAS S. & PERDIGÃO F. (2013). Generating a pronunciation dictionary for european portuguese using a joint-sequence model with embedded stress assignment. *Journal of the Brazilian Computer Society*, **19**(2), 127–134.

WURZEL W. (1989). *Inflectional Morphology and Naturalness*. Studies in Natural Language and Linguistic Theory. Springer.

Détection automatique de métaphores dans des textes de Géographie : une étude prospective

Max Beligné[1, 2] Aleksandra Campar[1] Jean-Hugues Chauchat[1] Mélanie Lefeuvre[1] Isabelle Lefort[2] Sabine Loudcher[1] Julien Velcin[1]
(1) Université de Lyon, Université Lyon 2, Laboratoire ERIC (EA 3083), France
(2) Université de Lyon, Université Lyon 2, Laboratoire EVS (UMR 5600), France
max.beligne, isabelle.lefort, sabine.loudcher, julien.velcin,
jean-hugues.chauchat@univ-lyon2.fr

RÉSUMÉ
Cet article s'intègre dans un projet collaboratif qui vise à réaliser une analyse longitudinale de la production universitaire en Géographie. En particulier, nous présentons les premiers résultats de l'application d'une méthode de détection automatique de métaphores basée sur les modèles de thématiques latentes. Une analyse détaillée permet de mieux comprendre l'impact de certains choix et de réfléchir aux pistes de recherche que nous serons amenés à explorer pour améliorer ces résultats.

ABSTRACT
Automatic detection of metaphors in Geographical research papers : a prospective study

This paper relates to a collaborative project that aims at carrying out a diachronic analysis of Geographical research papers. In particular, we present the first results we achieve by applying a method for automatic metaphor detection based on latent topic models. A more in-depth analysis is conducted in order to get a better understanding of the impact of our choices and to discuss further lines of research.

MOTS-CLÉS : Fouille de textes, métaphores, géographie, humanités numériques.

KEYWORDS: Text Mining, metaphore, geography, digital humanities.

1 Introduction

Objet de questionnement depuis les années soixante-dix, le champ de recherche des métaphores en Géographie n'a toutefois jamais été exploré de façon systématique. A partir des études fragmentées sur la question (Bachimon, 1979; Gaudin & Pumain, 2000; Ascher, 2005; Debarbieux, 2014), ce trope est reconnu comme un élément important de la logique argumentative. Par exemple, Bachimon (1979) montre l'importance de la métaphore organiciste dans la Géographie vidalienne afin de la rapprocher d'une science plus objective : la biologie. Dans cette métaphore, les villages peuvent être présentés comme des cellules : « Au lieu d'être dispersée en molécules, la population y forme un noyau et ce rudiment d'organisation suffit pour donner prise sur elle » (de La Blache, 1903). De la même manière, les régions peuvent devenir des organes et les pays se transformer en êtres vivants que les géographes doivent *disséquer*.

Si la Géographie vidalienne est maintenant dépassée, l'ensemble des autres études démontrent que les métaphores sont encore très présentes dans la Géographie mais que les systèmes référentiels mis en jeu ont fortement varié au cours du temps : organicisme, mécanicisme, théâtre... Désireux d'étudier cette évolution de manière quantifiée sur un large corpus, des chercheurs en Géographie et en Informatique travaillent ensemble sur une méthode de détection automatique des métaphores.

Pour répondre à l'objectif de cette recherche, la méthode doit non seulement détecter les métaphores mais aussi déterminer les systèmes référentiels mis en jeu, ce qui nous renvoie à la dimension conceptuelle des métaphores mise en évidence par Lakoff & Johnson (1980), avec pour chaque métaphore une projection d'un domaine source sur un domaine cible. Par exemple, quand des géographes parlent de « pays qui jouent un rôle sur la scène internationale », ils projettent le mot *scène* de son domaine source du *théâtre* dans le domaine cible de la *géopolitique*. Nous cherchons donc une méthode capable de détecter les métaphores mais également de trouver les domaines conceptuels sources utilisés.

Dans un premier temps, un état des lieux global sur la question de la détection de métaphore a été réalisé (Beligné, 2017). Si nous nous concentrons dans un second temps exclusivement sur les méthodes permettant de détecter les domaines sources, celles-ci sont beaucoup moins nombreuses. Ce constat méthodologique se retrouve dans la synthèse réalisée par Shutova (2015). Plusieurs types de méthodologies peuvent être utilisées de façon plus ou moins complémentaire pour détecter à la fois les métaphores et leurs domaines sources : des ressources lexicales (Roy *et al.*, 2006; Bracewell *et al.*, 2013), de la classification non supervisée (Gandy *et al.*, 2013; Shutova & Sun, 2013), de l'apprentissage supervisé (Bracewell *et al.*, 2014) des modèles thématiques (Heintz *et al.*, 2013) et du plongement de mots (Su *et al.*, 2017).

L'étude de Roy *et al.* (2006) est la plus proche de l'objectif attendu de cette recherche. Cependant, la méthode de Heintz *et al.* (2013) nous a semblé être une bonne extension et automatisation de leur travail grâce à une recherche automatique des thématiques latentes (et non plus par l'établissement manuel d'une centaine de lexies pour chaque domaine conceptuel étudié). De plus, l'utilisation des thématiques trouvées à l'échelle du texte et des phrases sert d'indice de détection des métaphores ce qui généralise le travail de Roy *et al.* puisqu'ils utilisent implicitement cet indice en recherchant dans un texte très spécialisé (sur la finance) des phrases métaphoriques de domaines très éloignés de ce thème principal (comme la météorologie).

De plus, la méthode de Heintz *et al.* (2013) répond à plusieurs contraintes fortes du projet :

- Elle ne repose pas sur un apprentissage automatique supervisé car notre projet étant dans sa phase initiale, nous n'avions pas de corpus préalablement annoté pouvant servir de base à un apprentissage supervisé.
- Elle détecte tous les types de métaphores. Les méthodes citées dans l'état des lieux à base de classification non supervisée ou de plongement de mots sont spécialisées sur un type de métaphore comme les métaphores nominales ou verbales. Dans le cadre de notre étude générale sur la Géographie, ces réductions à un type de métaphore ne sont pas pertinentes.
- Elle est facilement utilisable en français car notre objectif final est de travailler sur la Géographie française. Beaucoup de méthodes sont développées en anglais et utilisent des ressources lexicales qui ne sont pas forcément disponibles avec la même qualité en français, ce qui les rend difficilement utilisables dans cette langue.
- Son efficacité semble relativement bonne : F-score de 59% obtenu par Heintz.

Nous voulons évaluer en détail cette méthode qui répond le plus à l'objectif et aux contraintes de notre projet. Nous avons choisi dans un premier temps de travailler sur un petit corpus annoté en anglais

pour comparer nos résultats sans le changement de la variable langue. Notons cependant que nous constituons pour la suite en parallèle un corpus de taille importante d'articles en français qui couvre les quelques cent trente dernières années. Ici toutefois, la priorité a été donnée au travail d'évaluation pour pouvoir réaliser et présenter dans cet article un premier test détaillé de la méthode retenue.

2 Constitution d'un corpus annoté

Nous avons fixé un objectif de taille du corpus à environ 1500 phrases annotées manuellement de façon à avoir suffisamment de représentativité pour l'évaluation des résultats mais sans que le corpus ne soit trop long à annoter pour des contraintes de ressources. Les articles ont été sélectionnés dans divers sites en ligne [1] pour varier les sujets et donc les façons d'écrire et d'utiliser les métaphores en Géographie. Cependant, cet effort de sélection des articles et de représentativité du champ disciplinaire n'a pas été très poussé car il s'agissait, dans une première étape, de tester une méthode de détection automatique sur un petit corpus et non de généraliser tout de suite notre démarche.

Deux annotateurs ont annoté manuellement et de manière indépendante les phrases en suivant la *Metaphor Identification Procedure* (MIP). Cette méthode a été choisie car elle a été construite pour essayer de limiter les biais d'interprétations personnelles qui sont une difficulté importante dans la reconnaissance de métaphore (Group, 2007) et elle est reconnue internationalement (Shutova, 2015). Après une lecture attentive en entier du texte, il s'agit d'établir pour chaque unité lexicale son sens en contexte. Si cette unité lexicale possède dans d'autres contextes un sens plus basique (soit plus concret, soit plus relié à l'action corporelle, soit plus précis, soit plus historiquement ancien), alors l'unité lexicale est annotée comme étant métaphorique. En cas de doute sur les sens plus basiques, le dictionnaire Oxford Dictionary [2] est utilisé.

La méthode imposant de définir un concept cible, nous avons choisi *Geography*. Un important travail de clarification a été réalisé en amont pour mieux définir ce terme. Après discussion, il a été décidé de retenir une définition restreinte selon trois domaines : les éléments naturels touchant à la topographie ou à la géomorphologie (montagne, mer, désert, etc.), les éléments sociaux (population, groupe, collectivité, etc.) et les éléments territoriaux (pays, villes, lieux, etc.). Les annotateurs ont gardé seulement les métaphores avec un concept cible en lien avec ces éléments.

Pour chaque métaphore trouvée, les annotateurs ont déterminé son concept source. Il n'y a pas eu de méthodologie fixée à l'avance pour la détermination des concepts mais un travail continu de catégorisation. Les géographes ont défini préalablement une liste de concepts sources à partir de leurs connaissances sur la question. Ensuite, au fur et à mesure de l'avancée de l'annotation manuelle et des nouvelles métaphores trouvées, la liste des concepts sources s'est étoffée pour atteindre au final 42 concepts.

Finalement, le corpus est constitué de 17 articles, soit 1527 phrases dont 365 contiennent au moins une métaphore géographique ; les articles ont entre 29 et 264 phrases ; au total le corpus contient 41620 mots. Le degré d'accord entre les deux annotateurs est mesuré grâce au test du kappa (Cohen, 1960). Le taux K qui mesure l'accord entre deux annotateurs est, dans notre cas, égal à 0,63. En comparaison de celui obtenu par Heintz($K = 0,48$), l'accord a été plus fort entre nos deux annotateurs. Une hypothèse à vérifier pour expliquer cette situation concerne l'importance en amont de la clarification

1. Hypergeo, Cybergeo, Science Direct, Wiley online library
2. https ://en.oxforddictionaries.com/

initiale de ce qu'on entendait par "Géographie" qui a pu contribuer à une meilleure détection et donc à l'augmentation de l'accord inter-annotateurs.

3 Méthode de détection automatique de métaphores

Le principe de la méthode consiste à sélectionner les phrases qui vont comporter à la fois le concept cible Géographie et un des 42 concepts sources. Deux filtres sont ensuite appliqués pour augmenter les chances qu'il s'agisse bien d'une métaphore. Les paragraphes suivant détaillent l'ensemble de la méthode. Pour un développement plus formel, nous renvoyons les lecteurs intéressés à l'article originel de Heintz.

Pour chaque concept source, quatre mots spécifiques (mots graines dans la suite de cet article : cf. partie gauche de la table 2 pour quelques exemples) sont choisis en collaboration avec les géographes et les annotateurs. Pour le concept cible, il est difficile de réduire la Géographie à quelques mots graines. Pour faire face à cette difficulté, le nombre de ces mots est augmenté à 9 (alors que Heintz se limitait à 4) pour représenter au mieux l'ensemble des domaines de définition préalablement mentionnés : éléments naturels, sociaux et territoriaux (cf table 2).

L'objectif est d'associer à chaque concept les thématiques généralement utilisées pour parler de ce concept. Pour trouver ces thématiques, et afin d'avoir un spectre large de thématiques, la méthode LDA (*Allocation de Dirichlet Latente*) est utilisée pour trouver 100 thématiques (100 étant le nombre retenu par Heintz) sur la moitié des articles en anglais de Wikipédia choisis aléatoirement. La table 1 nous montre l'exemple de trois de ces thématiques. Chacune est une distribution de probabilités sur un dictionnaire de mots. Les cinq mots les plus représentatifs de ces trois thématiques sont indiqués à titre d'exemple avec les probabilités correspondantes. Chaque thématique est ensuite alignée avec le

t_1 : music : 0.0212 / opera : 0.0084 / orchestra : 0.0083 / piano : 0.0057 / symphony : 0.0049
t_2 : company : 0.0083 / bank : 0.0081 / tax : 0.0066 / market : 0.006 / financial : 0.0056
t_3 : city : 0.007 / population : 0.0065 / century : 0.0054 / world : 0.0051 / government : 0.005

TABLE 1 – Les cinq premiers mots les plus probables de trois thématiques différentes parmi les 100 thématiques extraites à partir de Wikipedia

concept qui maximise la somme des probabilités des mots graines correspondants lorsque celle-ci dépasse un seuil fixé Z (le score). Puis, on ne conserve qu'un nombre maximum de $N(C)$ thématiques pour les concepts cibles et $N(S)$ pour les concepts sources. Après plusieurs expérimentations, les seuils suivants sont retenus : $Z = 0,0075$, $N(S) = 3$ et $N(C) = 5$. Par rapport aux seuils de Heintz, nous avons seulement baissé un peu le seuil Z pour obtenir un meilleur alignement. La table 2 donne quelques exemples d'alignements réussis concept-thématique(s).

A partir des probabilités d'apparition de chaque thématique, il est possible de calculer la somme des probabilités d'apparition des thématiques alignées de chaque concept pour chaque phrase. Si cette somme dépasse un certain seuil, alors la phrase est considérée comme relevant du concept en question. Ce seuil peut prendre des valeurs différentes pour les concepts source ($Z_{rel,S}$) et le concept cible ($Z_{rel,C}$).

Enfin, chaque phrase qui relève à la fois du concept cible *Geography* et d'un concept source est

Concept	Mots graines	Score	Thématique
Music	rhythm, music, concert, jazz	0.0291	music, opera, orchestra, piano...
		0.0145	album, band, song, released...
Economy	economy, money, financial, bank	0.0199	company, bank, tax, market...
Geography	mountain, river, desert, climate, population, society, territory, environment, land	0.0459	river, lake, estonia, county...
		0.0397	city, population, century, world...
		0.0361	population, municipality, people...
		0.0259	district, rural, province, country...
		0.0194	town, village, area, river...

TABLE 2 – Quelques exemples d'alignements réussis concept-thématique(s)

sélectionnée comme étant métaphorique pour l'algorithme. Par exemple, la phrase suivante « Several Caucasian ethnic groups, particularly the Ossetians, the Circassians, and the Armenians, have played major roles on a vastly wider stage. » relève à la fois du concept source *Theatre* (« played major roles on a vastly wider stage ») et du concept cible *Geography* (« Caucasian ethnic groups. . . »).

Pour finir, pour chaque phrase sélectionnée, un premier filtre enlève les phrases qui ont moins de 2 mots après la suppression des mots-outils [3] car l'information disponible pour l'algorithme est jugée trop faible pour être significative. Ensuite, un deuxième filtre vérifie si la thématique source la plus importante de la phrase fait partie des 10 premières thématiques de l'article ; si c'est le cas, la phrase n'est pas gardée car l'utilisation aussi fréquente d'une thématique est souvent reliée à un sens littéral et non pas métaphorique [4]. Enfin, contrairement à la méthode originelle de Heintz, nous n'avons pas cherché à déterminer explicitement les mots supports de la métaphore.

4 Résultats et discussion

D'un point de vue quantitatif, les résultats obtenus sont encourageants compte tenu de la difficulté de la tâche. Une précision de 0,60 et un rappel de 0,19 sont obtenus après plusieurs expériences en baissant de manière significative les seuils ($Z_{rel,C}$ = 0.03 et $Z_{rel,S}$ = 0.01). Avec des seuils plus hauts ($Z_{rel,C}$ = 0.06 et $Z_{rel,S}$ = 0.04), le gain est minime en précision (+0.05) alors que le rappel chute de manière importante (-0.15). Toutefois, une analyse qualitative des résultats donne un autre éclairage. En comparant les concepts cibles annotés et ceux retenus par l'algorithme, une seule correspondance parfaite entre l'annotation et la détection est trouvée. Il s'agit de la phrase : « The county of hungry is inhabited by roughly - 9,919,128 as of July 2014 and this population is spread through a mix of Rural- 3308351 ppl city- 1701464,0 ppl Urban- 6747428 ppl who are all inside of this 93,028 sq km country ». La métaphore a bien été détectée manuellement dans l'expression « the county of hungry » (concept *Food*) et l'a été également par la détection automatique [5].

Dans tous les autres cas, les concepts sources détectés manuellement sont différents de ceux de la

3. Les mots-outils utilisés sont ceux fournis dans la librairie MALLET utilisée pour extraire les thématiques.
4. On retrouve ici l'idée de Roy de chercher un thème secondaire éloigné du thème principal de l'article.
5. Un retour au texte originel montre qu'il s'agit en plus d'une faute. Le contexte laisse penser qu'il s'agit en réalité de l'expression « the county of Hungary » ce qui démontre qu'il est nécessaire d'accorder plus d'importance à la qualité et à la sélection des textes sources.

détection automatique. L'analyse des erreurs permet de mettre en évidence le rôle à double tranchant des seuils $Z_{rel,C}$ et $Z_{rel,S}$ très bas : il suffit alors de peu de mots pour associer une phrase à un concept. Ainsi, des associations fausses peuvent rapidement apparaître comme, par exemple, dans cette phrase qui est associée par erreur au concept source *Music* : « The urbanisation density threshold is 1,019 people/sq km ». Le terme *density* explique cette erreur car il peut renvoyer à la densité sonore.

Par conséquent, l'algorithme trouve dans l'expérimentation menée des métaphores pour de mauvaises raisons. Cette situation provient de la méthode de détection manuelle employée MIP qui « suit une approche maximale, et non minimale, de telle sorte qu'une large gamme de mots puissent être considérés comme communiquant un sens métaphorique d'après leur utilisation en contexte » (Desalle, 2012) . Avec des seuils bas, la méthode automatique affecte plus de phrases au champ de la Géographie ; or beaucoup de phrases *géographiques* peuvent être considérées comme métaphoriques avec la méthode de détection manuelle employée, ce qui améliore les résultats de manière erronée.

En remontant les seuils, le rappel devient faible. Dans son étude, Heintz contourne ce problème en prenant une autre base évaluative. En effet, notre étude a pour objectif de trouver toutes les métaphores dans des textes, alors que l'étude de Heintz prend comme base d'évaluation les métaphores les mieux notées par l'algorithme (obtenues avec des seuils de détection élevés) mélangées avec un ensemble de phrases détectées comme non métaphoriques. Dans notre cas, des seuils hauts se traduisent par un nombre important de métaphores non détectées par l'algorithme. Dans le cas de l'étude de Heintz, des seuils hauts effectuent au contraire une présélection efficace l'évaluation.

Si on accepte de remonter les seuils, deux éléments sont déterminants pour l'efficacité de la méthode : tout d'abord, l'alignement des concepts et des thématiques. Si certains alignements ont bien fonctionné comme le montre la table 2, seuls 26 concepts sources sur les 43 définis ont été associés à au moins une thématique. Une perspective de recherche pour faire face à ce problème est d'augmenter le nombre initial de thématiques demandées à l'algorithme. Ainsi, certains concepts non associés précédemment à une thématique pourraient peut-être être associés à certaines nouvelles thématiques. Toutefois si le seuil maximal de 3 thématiques par concept source est maintenu ($N(S) = 3$), certaines thématiques assez représentatives d'un concept peuvent n'être plus associées à ce concept car elles se retrouvent en quatrième ou cinquième position. Si l'on augmente ce seuil maximal de trois, le risque est d'affecter certaines thématiques finalement peu représentatives d'un concept à ce concept. Il y a donc un travail sur ces nouveaux seuils (nombre général de thématiques, nombre de thématiques par concept) à effectuer.

Ensuite, le deuxième élément déterminant est le choix des mots graines. Plus ces derniers sont précis et représentatifs d'un concept, meilleure est la détection. Un problème identifié de notre étude vient de l'aspect trop large du concept cible *Geography*. La Géographie pouvant a priori traiter de tout, il est difficile de choisir des mots graines qui fonctionnent pour un grand ensemble d'articles. Par exemple, les métaphores d'un article parlant d'une rivière vont avoir beaucoup plus de chance d'être trouvées (car *river* a été choisi comme mot graine et a été bien aligné : voir table 2) que les métaphores d'un article sur le changement climatique (car *climate* n'a pas été bien aligné). Avec un concept cible plus précis, ce problème serait beaucoup moins important. Il s'agit là d'un axe important d'amélioration possible. Il est possible, une fois l'algorithme en place, de l'exécuter plusieurs fois de suite avec des concepts cibles beaucoup plus précis. Il serait ainsi possible de couvrir un domaine aussi large que la Géographie, tout en conservant une efficacité maximale de la méthode ici testée.

5 Conclusion et perspectives

L'application de la méthode de Heintz a constitué un bon point de départ pour notre projet. Certes, notre objectif final n'est pas encore atteint car nous n'avons pas produit de vues diachroniques de la répartition des métaphores sur notre corpus. Cependant, cet objectif est entièrement conditionné à une bonne détection des métaphores. En effet, une fois les métaphores correctement détectées, leur représentation dans le temps ne pose pas de véritable problème comme le montre l'article de Roy *et al.* (2006). La méthode de Heintz nous a donc permis d'aborder frontalement l'aspect le plus problématique de notre projet, à savoir la détection automatique des métaphores et de leurs domaines sources.

Deux pistes de recherches importantes pour l'efficacité de l'algorithme sont mises en évidence : l'alignement thématiques-concepts en travaillant sur le nombre général de thématiques et le nombre de thématiques par concept ; l'amélioration des mots-graines en exécutant l'algorithme de manière successive sur des concepts cibles plus précis. Si les pistes de recherche évoquées ci-dessus ne sont pas concluantes, il est possible d'opter pour un changement de perspective plus radical.

Cette idée provient des analyses qualitatives effectuées mais aussi de la réflexion des géographes. En effet, une de leurs conclusions, à la suite de cette étude, est que toutes les métaphores annotées manuellement avec la méthode employée ne les intéressent pas. Certaines métaphores sont usées et n'opèrent plus vraiment de transfert conceptuel comme par exemple l'expression « le lit de la rivière ». Les métaphores les plus intéressantes, les métaphores vives (Ricoeur, 1975), sont celles qui activent intentionnellement un transfert conceptuel. Citons à titre d'exemple une métaphore que les géographes aimeraient pouvoir détecter : « [La recherche de modèle en Géographie] consiste à élaguer ce que les faits ont de touffu et de contingent pour en saisir l'essentiel ou, mieux encore, la structure » (Brunet, 1972). Il y a consciemment, et ceci est marqué par la double présence des termes « élaguer » et « touffu » dans la phrase, une métaphore liée à la foresterie. Nous pensons que la méthode ici développée est particulièrement intéressante pour capter ce type de métaphore que l'on peut qualifier de « filée sur une même phrase » et qui peut jouer un rôle important d'intégration conceptuelle comme l'a montré Gréa (2002).

Ce changement de perspective modifie de manière importante l'objectif initial assigné à cette méthode. Il ne s'agit plus de détecter toutes les métaphores mais un type particulier de métaphore. Ainsi, il serait possible de remonter les seuils pour renforcer l'efficacité de la méthode sur cette tâche spécifique de détection. Il est probable qu'en voulant assigner une tâche trop large et ambitieuse à cette méthode, à savoir la détection de toutes les métaphores, nous avons diminué son efficacité spécifique. Ce changement de cap, s'il devait se confirmer, devra s'accompagner de réflexions importantes autant du côté des géographes que des informaticiens. Les premiers devront trouver des critères les plus précis et objectivés possibles pour définir les métaphores qui les intéressent le plus. Suite à ce travail, les informaticiens devront tester à nouveau cette méthode sur ce type particulier de métaphores et réfléchir à d'autres méthodes de détection pour les autres types de métaphores intéressant les géographes.

Références

ASCHER F. (2005). La métaphore est un transport. *Cahiers internationaux de sociologie*, (1), 37–54.

BACHIMON P. (1979). Physiologie d'un langage. l'organicisme aux débuts de la géographie humaine. *Espaces Temps*, (13), 75–103.

BELIGNÉ M. (2017). Des espaces dans la détection de métaphores. *Nouvelles Perspectives en Sciences Sociales*, **12**(2), 69–99.

BRACEWELL D. B., TOMLINSON M. T. & MOHLER M. (2013). Determining the conceptual space of metaphoric expressions. In *International Conference on Intelligent Text Processing and Computational Linguistics*, p. 487–500 : Springer.

BRACEWELL D. B., TOMLINSON M. T., MOHLER M. & RINK B. (2014). A tiered approach to the recognition of metaphor. In *International Conference on Intelligent Text Processing and Computational Linguistics*, p. 403–414 : Springer.

BRUNET R. (1972). Organisation de l'espace et cartographie de modèles [les villes du massif central]. *Espace géographique*, **1**(1), 43–48.

COHEN J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, **20**(1), 27–46.

DE LA BLACHE P. V. (1903). Tableau de la géographie de la france. In *Histoire de France depuis les origines la Révolution*. Hachette.

DEBARBIEUX B. (2014). Enracinement – ancrage – amarrage : raviver les métaphores. *L'Espace géographique*, **43**(1), 68–80.

DESALLE Y. (2012). *Réseaux lexicaux, métaphore, acquisition : une approche interdisciplinaire et inter-linguistique du lexique verbal*. PhD thesis, Université Toulouse le Mirail-Toulouse II.

GANDY L., ALLAN N., ATALLAH M., FRIEDER O., HOWARD N., KANAREYKIN S., KOPPEL M., LAST M., NEUMAN Y. & ARGAMON S. (2013). Automatic identification of conceptual metaphors with limited knowledge. In *AAAI*.

GAUDIN J.-P. & PUMAIN D. (2000). Quelques métaphores, au miroir des analyses spatiales : réseaux de villes et réseaux de pouvoir. *Revue européenne des sciences sociales. European Journal of Social Sciences*, (XXXVIII-117), 105–119.

GRÉA P. (2002). Intégration conceptuelle et métaphore filée. *Langue française*, p. 109–123.

GROUP P. (2007). Mip : A method for identifying metaphorically used words in discourse. *Metaphor and symbol*, **22**(1), 1–39.

HEINTZ I., GABBARD R., SRINIVASAN M., BARNER D., BLACK D. S., FREEDMAN M. & WEISCHEDEL R. (2013). Automatic extraction of linguistic metaphor with lda topic modeling. In *Proceedings of the First Workshop on Metaphor in NLP*, p. 58–66 : Citeseer.

LAKOFF G. & JOHNSON M. (1980). Conceptual metaphor in everyday language. *The journal of Philosophy*, **77**(8), 453–486.

RICOEUR P. (1975). *La métaphore vive*. Paris : Seuil.

ROY T., FERRARI S. & BEUST P. (2006). Étude de métaphores conceptuelles à l'aide de vues globales et temporelles sur un corpus. *Actes de Traitement Automatique des Langues Naturelles (TALN), Leuven, Belgique*, **1**, 580–589.

SHUTOVA E. (2015). Design and evaluation of metaphor processing systems. *Computational Linguistics*, **41**(4), 579–623.

SHUTOVA E. & SUN L. (2013). Unsupervised metaphor identification using hierarchical graph factorization clustering. In *HLT-NAACL*, p. 978–988.

SU C., HUANG S. & CHEN Y. (2017). Automatic detection and interpretation of nominal metaphor based on the theory of meaning. *Neurocomputing*, **219**, 300–311.

Describing derivational polysemy with XMG

Marios Andreou Simon Petitjean
Heinrich Heine Universität Düsseldorf, Düsseldorf, Germany
Marios.Andreou@uni-duesseldorf.de , petitjean@phil.uni-duesseldorf.de

RÉSUMÉ

Décrire la polysémie dérivationnelle avec XMG.

Dans cet article, nous modélisons et testons les approches monosémique et polysémique à la multiplicité des sens en morphologie dérivationnelle, en utilisant la sémantique des cadres (Frame semantics) et XMG. Pour illustrer nos hypothèses et propositions, nous utilisons des exemples de nominalisations déverbales avec le suffixe *-al* sur des verbes de changement de possession en anglais (par exemple, *rental*, *disbursal*). Dans notre implementation XMG, nous montrons que le sens sous-spécifié des affixes ne peut pas toujours être réduit à un simple sens unitaire et qu'en conséquence l'approche polysémique est plus judicieuse que l'approche monosémique. Nous introduisons également des contraintes sur les potentiels référents. Ces contraintes prennent la forme de contraintes de type et spécifient quels arguments de la base verbale sont compatibles avec le référent de la forme dérivée. L'introduction de contraintes de type rend certaines dérivations impossibles en raison d'échecs d'unification entre frames causées par des types incompatibles.

ABSTRACT

In this paper, we model and test the monosemy and polysemy approaches to derivational multiplicity of meaning, using Frame Semantics and XMG. In order to illustrate our claims and proposals, we use data from deverbal nominalizations with the suffix *-al* on verbs of change of possession (e.g. *rental*, *disbursal*). In our XMG implementation, we show that the underspecified meaning of affixes cannot always be reduced to a single unitary meaning and that the polysemy approach to multiplicity of meaning is more judicious compared to the monosemy approach. We also introduce constraints on the potential referents of derivatives. These constraints have the form of type constraints and specify which arguments in the frame of the verbal base are compatible with the referential argument of the derivative. The introduction of type constraints rules out certain readings because frame unification only succeeds if types are compatible.

MOTS-CLÉS : morphologie, dérivation, XMG, frame semantics.

KEYWORDS: frame semantics, XMG, word formation, derivation.

1 Introduction

A characteristic property of the outcomes of word formation processes is polysemy (Lieber, 2004; Rainer, 2014). In this paper, we model and test the monosemy and polysemy approaches to derivational multiplicity of meaning, using Frame Semantics and XMG. In order to illustrate our claims and proposals we use data from deverbal nominalizations with the suffix *-al*. Following recent advance-

ments in the study of the semantics of derived words in Frame Semantics (Plag *et al.*, forthcoming), we focus on a well defined class of bases and examine attestations in context. In particular, in the present paper we examine *-al* nominals that are based on verbs of change of possession (e.g. *rental*, *disbursal*).

What is a frame? A frame is a general format of mental representations of concepts which is also applicable to linguistic phenomena. It is a recursive attribute-value structure that provides information about a referent (Petersen, 2007; Kallmeyer & Osswald, 2013; Löbner, 2013). Word formation in Frame Semantics is generally treated in terms of referential shifts (Plag *et al.*, forthcoming). For example, as we will see in the analysis, the suffix *-al* can target particular arguments of the base verb and shift reference from the original referent (i.e. causation event) to a new referent (e.g. theme).

What is XMG? XMG (eXtensible MetaGrammar, (Crabbé *et al.*, 2013)) is a modular and extensible tool used to generate various types of linguistic resources from an abstract and compact description. This description, the metagrammar, relies on the concepts of logic programming and constraints. XMG comes with a system of dimensions, allowing to separate the different levels of linguistic description (e.g. syntax and semantics), and providing dedicated languages adapted to the structures the user wishes to generate. In this work, the dimension we used is the <frame> dimension, proposed in Lichte & Petitjean (2015), where semantic frames can be described using typed feature structures descriptions.

The rest of this paper is structured as follows: In section 2 we examine which readings are possible in context for *-al* nominals which are based on verbs of change of possession. We then provide a modeling of multiplicity of meaning using XMG. In section 3 we focus on monosemy approaches and in section 4 we offer a modeling of the polysemy approach. Section 5 concludes the paper.

2 -al on verbs of change of possession

In this paper we focus on the suffix *-al* on verbs of "change of possession" (Levin, 1993). The class of verbs of change of possession comprises "give" verbs (e.g. *give*, *rent*), "contribute" verbs (e.g. *submit*, *betroth*), verbs of "future having" (e.g. *promise*, *bequeath*), verbs of "providing" (e.g. *provide*, *redress*), verbs of "obtaining" (e.g. *retrieve*, *accrue*), verbs of "exchange" (e.g. *exchange*, *swap*), and "berry" verbs (e.g. *nest*, *clam*).

Verbs of change of possession describe caused possession of the kind 'x causes y to have z', in which x is the agent, y is the recipient, and z is the theme (Goldberg, 1995, 2006; Jackendoff, 1990; Rappaport Hovav & Levin, 2008). The general schema for verbs of change of possession in Frame Semantics is given in Figure 1 in the form of an attribute-value matrix.

In Figure 1, we model verbs of change of possession as causation events that involve three arguments; an agent, a theme, and a recipient. Given that verbs of change of possession are causation events, they come with a bipartite structure that comprises a CAUSE and an EFFECT (for more on the modeling of events in Frame Semantics see Plag *et al.*, forthcoming; Kallmeyer & Osswald, 2012). A typical causation event involves a relationship between situations and entities in which a particular entity (e.g. an agent) causes another entity (e.g. theme) to go from an initial situation to a result situation. In the case of verbs of change of possession, there is a start, i.e. an initial state, in which the agent is the possessor of the theme, and an end, i.e. a result state, in which the recipient is the possessor of the theme. As captured by the constraint "$6 \neq 7$", the initial state and the result state are not equal (Plag

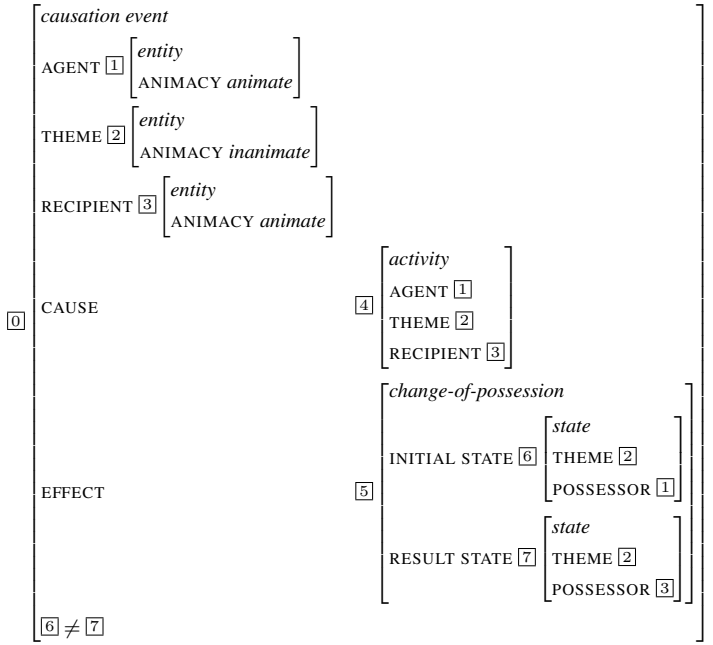


Figure 1: Verbs of change of possession

et al., forthcoming).

The suffix *-al* on verbs of change of possession derives event, result, and inanimate theme noun readings. Consider the following examples from the Corpus of Contemporary American English (COCA):

(1) a. Event reading: This results in delays in the **disbursal** and utilization of funds – especially at the Gram Panchayat level.
b. Result reading: At the same time as it emerged that Fitzroy was terminally ill with 'a rapid consumption', Henry learned of Margaret Douglas's **betrothal** to Thomas Howard.
c. Inanimate theme reading: The room was technically full of locals, people from Bianca' s life before she headed West, friends who crossed the bridge searching for more affordable **rentals** in Williamsburg or Long Island City.

In the first case, i.e. event reading, the nominalization lexicalizes the event denoted by the verb. This type of nominalization is also referred to as 'transpositional' in that the nominalization 'transposes' (recategorizes) the word from verb to noun without altering the sense of the verbal base. Thus, *disbursal* in (1-a) can be paraphrased as "event of disbursing". In the case of result readings, the nominalization denotes "the outcome of verb-ing". Thus, *betrothal* in (1-b) can be paraphrased as "the outcome of betrothing". Finally, *-al* nominalizations may lexicalize the inanimate patient reading, that is, "the thing verb-ed, the thing affected by verb-ing". In (1-c), *rental* can be paraphrased as "the

thing that someone rented".[1]

The foregoing discussion suggests that the suffix *-al* can manipulate the frame of a verb of change of possession and target certain arguments of it. In particular, it can target the *causation event* argument (i.e. 0), the *result state* argument (i.e. 7), and the *theme* argument (i.e. 2). Thus, the referent of a form derived by *-al* can be identified with some of arguments of the verbal base, but not all of them. Observe, for instance, that the referent of *-al* derivatives is never the agent (i.e. 1), the recipient (i.e. 3), the cause (activity, i.e. 4), the effect (change-of-possession, i.e. 5) or the initial state (i.e. 6).

3 Modeling monosemy

There are two approaches to multiplicity of meaning in derivation: monosemy and polysemy. In this section we examine the monosemy approach. Under this approach, multiplicity of meaning is reduced by assigning a more vague meaning to an affix. Thus, more concrete meanings of affixes derive from a general highly underspecified meaning.

In order to apply the monosemy approach to *-al*, we have to reduce multiplicity of meaning by identifying meanings that are shared by all *-al* derivatives. As we saw in the previous section, *-al* derivatives denote (a) eventualities (e.g. event 'transpositional' readings), and (b) entities (e.g. inanimate theme readings). Thus, the abstract core meaning of *-al* can be characterized as 'eventuality or entity having to do with X' (with 'X' denoting the base).

Monosemy approaches to the semantics of derivation are confronted with two problems to which we turn our attention. The first problem is that it is very hard to establish a unitary meaning for an affix. In particular, the aim of monosemy approaches is to reduce multiplicity of meaning by postulating a unitary abstract meaning. Forms derived by *-al*, however, denote both eventualities and entities. Thus, the disjunction 'eventuality or entity' that is needed in order to capture the multiplicity of meaning of *-al* derivatives reveals that the desirable underspecified meaning of affixes cannot always be reduced to a single unitary meaning.

The second problem with the monosemy approach to the semantics of derivation is (massive) over-generation. As we saw earlier, the abstract meaning for *-al* informs us that *-al* forms denote both eventualities and entities. This gives rise to the question: *What kind of predictions follow from the abstract meaning 'eventuality or entity having to do with X'*? This particular formulation of the abstract meaning of *-al* leads one to expect that *-al* derivatives could in principle denote all 'entities' and all 'eventualities'. Our data, however, suggests that not all 'entities' and not all 'eventualities' can be denoted by *-al* derivatives. For instance, the referents of *-al* derivatives may be inanimate themes (e.g. *houses* in the case of *rentals*) but not recipients.

The first XMG implementation we propose makes the modeling of the monosemy approach possible by using a scope-over constraint, expressing the fact that the reference of *-al* can be identified with any of the elements of the verbal base.

We first need to describe the frame given in Figure 1, by means of a XMG reusable abstraction which we will name rent. This abstraction describes the class of verbs of change of possession:

> **class** rent

[1] Other meanings exist and may be lexicalized. For example, the referent of *boat rental* could be the place where boats are rented or the company that rents boats.

```
export ?X0 ?X1 ?X2 ?X3 ?X4 ?X5 ?X6 ?X7
declare ?X0 ?X1 ?X2 ?X3 ?X4 ?X5 ?X6 ?X7
{<frame>{
  ?X0[causation,
    agent: ?X1[entity, animacy:[animate]],
    theme: ?X2,
    recipient : ?X3[entity],
    cause: ?X4[ activity ,
                agent:?X1,
                theme:?X2,
                recipient :?X3[entity, animacy:[animate]]
                ],
    effect : ?X5[change_of_possession,
                 initial −state: ?X6[ initial_state ,
                                      theme:?X2[entity],
                                      possessor :?X1],
                 result −state: ?X7[ result_state ,
                                     theme:?X2[entity],
                                     possessor :?X3] ] ]
                 }
}
```

where the first lines define the set of unification variables which can be used within the class (**declare**) and outside of it (**export**). **<frame>** means that the description belongs to the Frame Semantics dimension. The structure described in the frame dimension, labeled by **?**X0, is a straightforward translation of the one in Figure 1, where all variables **?**X0 ,..., **?**X7 stand for the boxed numbers from 0 to 7.

To define the scope-over relation mentioned earlier, we can use a new abstraction (a class we will name al_nominal). This class, as its name suggests, models the semantics of *-al* derivatives, which for the purposes of this study are based on verbs of change of possession.

```
class al_nominal
import rent[]
declare ?Ref
{
  <frame>{
    [al−lexeme,
      m−base:[event,
              sem:?X0]
      ref :?Ref
    ]
    ;
    ?X0 >∗ ?Ref;
  }
}
```

With **import** rent[], i.e. we make the structure defined in the class rent available in the current class, together with its variables (we can refer to all variables **?**X0 ,..., **?**X7 in the current class as

they are exported by rent). The operator >∗ means that there must be a "path" (as it would be in a graph representation[2] of the frame) from the root ?X0 to the semantic reference ?Ref. Concretely, the compiler will try to generate structures where the reference is identified with another label, starting with the whole frame (?X0), and then exploring all of its subparts, recursively. As we do not specify any other constraint here, the reference will be able to unify with every possible subpart. This means that readings such as initial state, which should be ruled out, are also generated when this first version of the metagrammar is executed.

In this section we modeled the first approach to multiplicity meaning, i.e. monosemy, and showed that it leads to massive overgeneration. In the next section we focus on the second approach to multiplicity of meaning, i.e. polysemy.

4 Modeling polysemy

Under the polysemy approach, there is multiplicity of meaning in word formation patterns. Given the architecture of Frame Semantics, this multiplicity of meaning can be expressed in an inheritance hierarchy of lexeme formation rules (Riehemann, 1998; Koenig, 1999; Plag *et al.*, forthcoming). Thus, attested readings of words of a given morphological category result from indexation of particular elements (e.g. arguments) of the semantic representation of the verb, combined with inheritance mechanisms.

An open question is how we can model the polysemy approach in XMG and constrain possible readings. We suggest that there are two ways to tackle this issue. First, via a fully specified (and explicit) rule, which will replace the scope over relation in the previous class al_nominal:

```
{?X0=?Ref | ?X2=?Ref | ?X7=?Ref}
```

where | and = are respectively the disjunction and the unification operators, `?X0`, `?X2` and `?X7` respectively correspond to the boxed numbers 0, 2 and 7 of Figure 1, and `?Ref` is a variable representing the semantic reference.

Under this approach, possible readings are considered as generalizations over already attested derivatives. Thus, agent, recipient, and initial state readings are ruled out since they are not part of the possible readings in the fully-specified-rule; the rule models readings that are already attested in *-al* derivatives.

The second way to model the polysemy approach in XMG is the introduction of an underspecified rule with constraints on types. In this case, the referent of an *-al* nominal can have three possible types: causation, result state, or entity.

```
?X0 >∗ ?Ref;
{ ?Ref[ result_state ] | ?Ref[causation]
| ?Ref[entity , animacy:[inanimate]] }
```

Here, the first line is once again the scope over relation, but of course, in this case, only the structures where no type constraint is violated will eventually be generated.

[2] An attribute-value matrix can be seen as a directed graph in which every attribute-value pair is an edge labeled by the attribute and pointing to the node representing the value.

In the second line, we express the fact that the referent of an *-al* derivative can have any of the three types previously stated. In the case of entity, only the theme should be a possible referent. We, therefore, add information about animacy (here, inanimate), which makes the reference of *-al* derivatives incompatible with frames of type animate, such as the agent and the recipient. This is in accordance with findings in the literature on possible constraints on animacy (see Lieber (2016) on *-al* and Kawaletz & Plag (2015) on the suffix *-ment*). When the referent of an *-al* derivative is a state, the type result_state is given to prevent unification with the initial state frame (of type initial_state). This way, agent, recipient, and initial state readings are ruled out because frame unification only succeeds if types are compatible. The type constraints (for example incompatibility of event and entity) are also specified in the metagrammar.

5 Conclusion

In a nutshell, our analysis allows one to model and test the various theoretical approaches to a long-standing issue in word formation. The XMG implementation revealed that the underspecified meaning of affixes cannot always be reduced to a single unitary meaning and that the monosemy approach leads to massive overgeneration. As we showed, the polysemy approach to multiplicity of meaning is more judicious compared to the monosemy approach.

In order to deal with overgeneration, we proposed to introduce constraints on the potential referents of *-al* derivatives. These constraints have the form of type constraints and specify which arguments in the frame of the verbal base are compatible with the referential argument of the derivative. The introduction of type constraints rules out certain readings because frame unification only succeeds if types are compatible.

In the present paper, we focused on *-al* derivatives which are based on verbs of change of possession. The next step is to apply the proposed analysis to other affixes and classes of bases as well. This will allow us to identify which constraints are specific to particular classes or affixes, and which constraints are shared across classes or affixes. Of course, the main advantages of the metagrammatical framework will become more obvious as the linguistic resource grows: for example, inheritance will help sharing information across classes with similar behaviors.

Acknowledgements

The authors gratefully acknowledge financial support by the Deutsche Forschungsgemeinschaft (DFG Collaborative Research Centre 991: 'The Structure of Representations in Language, Cognition, and Science', Project C08 'The semantics of derivational morphology: A frame-based approach')

CRABBÉ B., DUCHIER D., GARDENT C., LE ROUX J. & PARMENTIER Y. (2013). XMG : eXtensible MetaGrammar. *Computational Linguistics*, **39**(3), 1–66.

GOLDBERG A. (1995). *Constructions: A Construction Grammar approach to argument structure*. Chicago, IL: University of Chicago Press.

GOLDBERG A. (2006). *Constructions at Work. The Nature of Generalizations in Language*. Oxford: Oxford Univ. Press.

JACKENDOFF R. (1990). *Semantic structures*. Cambridge, MA: MIT Press.

KALLMEYER L. & OSSWALD R. (2012). A frame-based semantics of the dative alternation in Lexicalized Tree Adjoining Grammars. In C. PIÑÓN, Ed., *Empirical Issues in Syntax and Semantics 9*, p. 167–184. ISSN 1769-7158.

KALLMEYER L. & OSSWALD R. (2013). Syntax-driven semantic frame composition in lexicalized tree adjoining grammars. *Journal of Language Modelling*, **1**(2), 267–330.

KAWALETZ L. & PLAG I. (2015). *Predicting the Semantics of English Nominalizations: A Frame-Based Analysis of -ment Suffixation*, In L. BAUER, L. KÖRTVÉLYESSY & P. ŠTEKAUER, Eds., *Semantics of Complex Words*, p. 289–319. Springer International Publishing: Cham.

KOENIG J. (1999). *Lexical Relations*. Stanford Monographs in Linguistics. Cambridge University Press.

LEVIN B. (1993). *English Verb Classes and Alternations*. The University of Chicago Press.

LICHTE T. & PETITJEAN S. (2015). Implementing semantic frames as typed feature structures with XMG. *Journal of Language Modelling*, **3**(1), 185–228.

LIEBER R. (2004). *Morphology and lexical semantics*. Cambridge: Cambridge University Press.

LIEBER R. (2016). *English nouns: The ecology of nominalization*. Cambridge: Cambridge University Press.

LÖBNER S. (2013). *Understanding semantics*. London: Arnold, 2nd, revised edition edition.

PETERSEN W. (2007). Representation of concepts as frames. In J. SKILTERS, Ed., *The Baltic international yearbook of cognition, logic and communication*, volume 2, p. 151–170.

PLAG I., ANDREOU M. & KAWALETZ L. (forthcoming). A frame-semantic approach to polysemy in affixation. In O. BONAMI, G. BOYÈ, G. DAL, H. GIRAUDO & F. NAMER, Eds., *The lexeme in descriptive and theoretical morphology*. Berlin: Language Science Press.

RAINER F. (2014). Polysemy in derivation. In R. LIEBER & P. ŠTEKAUER, Eds., *The Oxford handbook of derivational morphology*, Oxford Handbooks in linguistics, p. 338–353. Oxford: Oxford University Press.

RAPPAPORT HOVAV M. & LEVIN B. (2008). The English dative alternation: The case for verb sensitivity. *Journal of Linguistics*, **44**, 129–167.

RIEHEMANN S. Z. (1998). Type-based derivational morphology. *Journal of Comparative Germanic Linguistics*, **2**(1), 49–77.

Changement stylistique de phrases par apprentissage faiblement supervisé

Damien Sileo[1, 2] Camille Pradel[1,*] Philippe Muller[2,*] Tim Van de Cruys[2,*]
(1) Synapse Développement, 5 Rue du Moulin Bayard, 31000 Toulouse
(2) IRIT, Université Paul Sabatier 118 Route de Narbonne 31062 Toulouse
(*) Contributions égales
damien.sileo@synapse-fr.com, camille.pradel@synapse-fr.com,
philippe.muller@irit.fr, tim.van-de-cruys@irit.fr

RÉSUMÉ

Plusieurs tâches en traitement du langage naturel impliquent de modifier des phrases en conservant au mieux leur sens, comme la reformulation, la compression, la simplification, chacune avec leurs propres données et modèles. Nous introduisons ici une méthode générale s'adressant à tous ces problèmes, utilisant des données plus simples à obtenir : un ensemble de phrases munies d'indicateurs sur leur style, comme des phrases et le type de sentiment qu'elles expriment. Cette méthode repose sur un modèle d'apprentissage de représentations non supervisé (un auto-encodeur variationnel), puis sur le changement des représentations apprises pour correspondre à un style donné. Le résultat est évalué qualitativement, puis quantitativement sur le jeu de données de compression de phrases Microsoft, avec des résultats encourageants.

ABSTRACT

Textual Style Transfer using Weakly Supervised Learning

Several natural language processing tasks, such as sentence paraphrasing, compression, or simplification, consist of sentence modifications that aim to preserve the global sentence meaning. Most existing methods rely on specific data and models tuned towards a particular task. We introduce a general method that is capable of tackling those problems using simpler data : a set of sentences paired with their stylistic features, such as their lengths for compression. The method relies on unsupervised representation learning with a variational auto-encoder, and then changing the input text representation to match a given style. The method is evaluated both qualitatively and quantitatively on Microsoft's dataset for sentence compression, with encouraging results..

MOTS-CLÉS : auto-encodeur variationnel, apprentissage faiblement supervisé, changement stylistique.

KEYWORDS: variational auto-encoder, weakly supervised learning, style transfer.

1 Introduction

La génération textuelle est une tâche centrale pour l'interaction entre un système intelligent et ses utilisateurs (réponse d'un agent conversationnel, résumé de texte, génération d'article...). Lors de cette interaction, il est désirable de contrôler les générations afin qu'elles respectent des contraintes imposées par le contexte. On peut ainsi vouloir agir sur la longueur d'une phrase générée, son niveau

de langue, sa politesse, sa polarité, et d'autres caractéristiques dé-corrélables, au moins en partie, de la sémantique. On les regroupera sous le terme fédérateur de "style". La transformation de textes pour modifier leur style et résoudre ce problème constitue un domaine actif de recherche, où sont notamment utilisés des modèles et données spécifiques (Pitler, 2010) (Shardlow, 2014) (Xu *et al.*, 2012). Les données utilisées sont des couples alignés de phrases du style original et du style "cible".

Prenons l'exemple de la contraction de phrase. On peut vouloir passer automatiquement de la phrase source *They also, by law, have to be held in Beirut* à une phrase contractée dite cible : *They have to be held in Beirut.* Cependant, les données sous forme de telles paires alignées sont parfois trop peu nombreuses pour bien apprendre directement la tâche, et leur création est coûteuse.
Nous nous plaçons ici dans un cadre à la fois unificateur aux problèmes de changement de style et nécessitant une plus faible supervision. Au lieu de phrases alignées de deux styles différents, la méthode proposée se contente d'un ensemble de phrases et d'indicateurs sur leur style. Ces indicateurs peuvent être liés à l'origine des phrases (par exemple l'année d'écriture) ou calculés (comme la longueur des phrases). Notre modèle utilise un indicateur comme signal pour modifier les générations et correspondre à une valeur voulue.
Pour cela, nous introduisons d'abord des modèles neuronaux de génération de phrases (section 2), puis proposons une méthode de changement de style (section 3).

2 Modèles génératifs conditionnés

Nous présentons d'abord les modèles capables d'apprendre une représentation de phrase qui permettra des modifications préservant certaines propriétés : les auto-encodeurs variationnels récurrents.

2.1 Auto-encodeurs récurrents

Les modèles de langue définissent une distribution de probabilité sur des séquences de mots $x = w_1...w_n$.

$$p(x) = p(w_1...w_n) = \prod_{i=1}^{n} p(w_i|w_1...w_{i-1}) \tag{1}$$

Capable de prédire la probabilité d'un mot sachant ce qui précède, un tel modèle de langue est un modèle génératif (Bengio *et al.*, 2003).

Afin de générer des phrases correspondant précisément à un message, on suppose que p est conditionné par une représentation latente $z \in \mathbb{R}^d$ de la phrase (d est la dimension de l'espace latent).

$$p(x|z) = \prod_{i=1}^{n} p(w_i|w_1...w_{i-1}, z) \tag{2}$$

z est censé capturer les dimensions sémantiques et stylistiques de la phrase.

Ici $p(x|z)$ est paramétré par un réseau de neurones récurrents conditionné par z, et z est obtenu par encodage de la phrase par $z = q(x)$ où q est un autre réseau de neurones récurrent. C'est l'architecture seq2seq (Sutskever *et al.*, 2014), constituée de cet encodeur et d'un décodeur. Ainsi, dans le paradigme des auto-encodeurs, on apprend un modèle génératif en minimisant l'erreur de reconstruction des données. L'encodeur change la phrase en un signal qui aide le décodeur à la restituer.

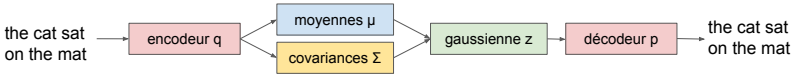


FIGURE 1 – Auto-encodeur variationnel

2.2 Changement de style dans l'espace latent : auto-encodeurs variationnels

Étant donné que z capture les propriétés de la phrase, le changement de style voulu est une trajectoire dans $\mathbb{R}^d$. (Bowman *et al.*, 2016) a montré que dans le cadre d'auto-encodeurs "classiques", les translations dans l'espace latent entre deux phrases cohérentes peuvent produire des phrases incohérentes. Or, on souhaite que les déplacements dans l'espace latent aient du sens. On se place donc dans le cadre variationnel où z est stochastique. La représentation z d'un texte est encore un vecteur, mais stochastique : ses composantes peuvent s'écarter de leur moyenne, et donc la représentation est plus robuste aux déplacements. Le cadre variationnel a été par exemple utilisé dans le cadre de la traduction (Zhang & Xiong, 2016) ou pour la génération de textes (Bowman *et al.*, 2016). On dit désormais que z est une variable aléatoire $z \sim \mathcal{N}(0, I)$, ici une gaussienne multivariée. L'encodeur apprend alors à conditionner cette variable en estimant le postérieur $P(z|x)$ par $q(z|x)$. Cela permet de déterminer un tirage qui sera présenté au décodeur. Concrètement, l'encodeur prédit ici pour chaque x les moyennes μ de z ainsi qu'une matrice de covariance Σ qui définissent le tirage. La figure 1 montre une vue d'ensemble du système. Toujours fixer $\Sigma = 0$ ramène au cas de la section 2.1.

On a alors $q(z|x) = \mathcal{N}(z - \mu, \Sigma)$. Le décodeur p estime la probabilité $p(x|z)$ de la phrase à partir d'un tirage de z .

Dans ce cadre, on ne peut pas calculer simplement la vraisemblance des données, mais il existe la borne inférieure suivante (Kingma & Welling, 2014) :

$$\mathcal{L}(x, \Theta) = \mathrm{E}_{q(z|x)}[\log p(x|z)] - \mathrm{KL}(q(z|x)||p(z)), \text{ où } \Theta \text{ désigne les paramètres de } p \text{ et } q \qquad (3)$$

KL désigne la divergence de Kullback-Leibler qui mesure l'écart entre deux distributions de probabilités. Maximiser $\mathcal{L}$ permet de maximiser la vraisemblance des données, et ses deux termes sont interprétables : le premier correspond au terme de reconstruction et le second à une régularisation qui rapproche le postérieur du prior. C'est le modèle des auto-encodeurs variationnels (VAE). L'optimisation est possible par descente de gradient.

3 Procédure de changement style

On introduit maintenant une variable y qui correspond à une valeur de style. On suppose un modèle graphique où z conditionne à la fois x et y, comme le montre la figure 2 Enlever y ramène au VAE classique.

La variable y peut être le nombre de mots d'une phrase, son sentiment, son temps verbal, etc. Pour trouver z, on optimise maintenant $P(z|y, x)$, la probabilité de la représentation z sachant un texte x et le style y qu'on veut lui imposer. On a d'après Bayes, et en conditionnant par x :

$$P(z|y, x) \propto P(y|z, x)P(z|x) \qquad (4)$$

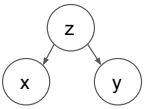


FIGURE 2 – Modèle graphique proposé

Comme y ne dépend que de z si z est connu, on a donc $P(y|z,x) = P(y|z)$. Supposons que y dépend linéairement de μ et que le modèle linéaire $\theta\mu$ s'écarte de y avec une variance ϵ^2. On approxime alors $P(y|z)$ avec $r(y|z) = \mathcal{N}(y - \theta\mu, \epsilon^2)$ où μ est la moyenne de z. Par ailleurs, $q(z|x) = \mathcal{N}(z - \mu, \Sigma)$, décrit dans 2.2, estime $P(z|x)$. D'où :

$$P(z|y,x) \propto r(y|z)q(z|x) \tag{5}$$

Il suffit alors d'estimer r et q pour trouver la représentation z désirée correspondant à la fois à x et y, en maximisant $P(z|y,x)$. Nous estimons d'abord q puis r sur des données d'entraînement d'après la procédure présentée dans la section suivante.

3.1 Apprentissage

Nous prenons comme point de départ un ensemble de textes $\mathcal{U}$, un ensemble $\mathcal{S}$ de textes pour lesquels $y_s, s \in \mathcal{S}$ est défini, et un autre ensemble $\mathcal{T}$ de textes à changer selon $y_t, t \in \mathcal{T}$. Ces ensembles peuvent se chevaucher. Nous entraînons un auto-encodeur variationnel afin d'apprendre q et p avec une descente de gradient sur les textes $x_u, u \in \mathcal{U}$. On extrait ensuite les représentations latentes z_s des textes $s \in \mathcal{S}$ de par les moyennes μ_s et les covariances Σ_s . Afin de déterminer r, on entraîne un modèle linéaire θ pour prédire y_s à partir de μ_s, par maximisation de la vraisemblance $\prod_{i \in \mathcal{S}} \mathcal{N}(y_i - \theta\mu_i, \epsilon^2)$ en réservant une partie de $\mathcal{S}$ pour la validation. L'estimation de y_n par $\theta\mu_n$ exhibe une variance empirique ϵ^2 estimée sur les données de validation, et θ est fixé pour la suite.

3.2 Inférence

Pour changer les styles de textes $t \in \mathcal{T}$, afin qu'ils correspondent à y_t, on cherche à maximiser la vraisemblance de leurs représentations z_t :

$$\mathcal{L} = \prod_{i \in \mathcal{T}} \mathcal{N}(y_i|\theta z_i, \epsilon^2)\mathcal{N}(z_i - \mu_i, \Sigma_i) \tag{6}$$

La fonction de coût du changement de style est alors la suivante (log-vraisemblance) :

$$C(z,y,\mu,\Sigma) = \sum_{i \in \mathcal{T}} ((y_i - \theta z_i)^2 + \epsilon^2 (z_i - \mu_i)^T \Sigma_i^{-1} (z_u - \mu_i)) \tag{7}$$

On retrouve une régularisation Tikhonov, de coefficients dictés par le VAE. Intuitivement, z_i s'éloigne de la représentation d'origine μ_i pour qu'elle ait plus de chances de correspondre à un style y_i, éloignement davantage pénalisé sur les dimensions où la variance prédite est faible, qui sont plus importantes pour garder le sens de la phrase. La représentation $\tilde{z}$ d'un texte qui minimise C a donc subi un changement de style et le décodage $\tilde{x}$ de $\tilde{z}$ par p renvoie une phrase correspondant au style y.

4 Évaluation

L'apprentissage non supervisé utilise 22M posts en anglais du réseau social Reddit ayant entre 4 et 40 symboles et ne contenant pas d'hyperliens.

On s'évalue sur une tâche de compression de phrase avec le jeu de données Microsoft (Toutanova & Brockett, 2016) composé de phrases sources alignées avec des phrases raccourcies par des humains. On garde les 3423 phrases dont la cible a moins de 16 mots et les phrases sources de moins de 40 mots correspondant à $x_t, t \in \mathcal{T}$.

4.1 Configuration

On utilise un vocabulaire de 32k mots et le tokenizer Treebank (Bird *et al.*, 2009) ; [unk] dénote les mots hors vocabulaire.

Les réseaux récurrents q et p sont des "Recurrent Highway Network" (Zilly *et al.*, 2016) de dimension 600 et profondeur 5. Les représentations des mots sont initialisées avec ConceptnetNumberbatch (Speer & Chin, 2016) qui sont de dimension 300. z est de dimension $d = 64$. Les représentations des mots en entrée et sortie sont les mêmes, une projection affine étant utilisée sur la dernière couche du réseau (Press & Wolf, 2017). L'optimisation est réalisée avec l'algorithme Adam (Kingma & Ba, 2015), le taux d'apprentissage fixé à 10^{-4} et des "batch" de taille 128 sur 2 epochs. On utilise un "word dropout" (Bowman *et al.*, 2016) et un minimum d'information de 4 pour la divergence KL (Kingma *et al.*, 2016). Le décodeur génère 2000 phrases par exemple dans l'analyse qualitative où les deux meilleures sont sélectionnées au sens de la vraisemblance obtenue. Dans l'analyse quantitative, les exemples étant bien plus nombreux, seulement 50 phrases sont générées pour chaque exemple et la meilleure est retenue.

On prend pour y le nombre de mots par phrase, et on apprend θ à partir d'une version normalisée de y et de $800k$ exemples dont 10% servent à la validation et à la détermination de ϵ.

4.2 Résultats

La table 1 montre des phrases x choisies comme entrées, une longueur y qu'on veut leur imposer et pour chacune deux phrases générées par la procédure 3.1 avec ces deux entrées. Les générations conservent globalement le sens de la phrase source tout en rapprochant le nombre de mots de l'objectif. Le nombre de mots souhaité n'étant pas atteint, on exagère l'objective pour compresser/allonger davantage. Les générations sachant x et y peuvent être assez différentes tout en restant correctes. Pour le deuxième exemple, on essaie d'allonger la phrase et constate que le doute exprimé par *i don't know* est complété par des idées génériques qui déforment peu l'idée d'origine. Ceci dit, lorsque les idées exprimées figurent trop rarement dans le jeu de données, le modèle est incapable d'encoder et décoder la phrase sans perte, même sans changement de longueur comme le montre le troisième exemple. *sky* et *blue* sont dans le vocabulaire mais il est difficile d'encoder la relation dans z.

Afin d'évaluer la compression, on génère des phrases à partir des phrases sources, qu'on tronque pour qu'elles soient de la même longueur que la phrase cible du jeu de données. On compare la génération fournie par notre modèle puis tronquée avec la compression terrain. Une mesure de rappel au niveau des mots permet d'estimer si l'information de la phrase source a pu être conservée. Cette

TABLE 1 – Exemples de changement de style : compression et allongement.

phrase source (x)	y	phrases générées ($\tilde{x}$)
yes, i totally agree with you	1	yeah , i know .
	1	agreed , it is .
i don't know	18	i do n't think it counts .
	18	i do n't know they did .
the sky is blue	4	the [unk] is [unk]
	4	[unk] [unk] [unk] [unk]

TABLE 2 – résultats compression et intervalles de confiance

	rappel	longueur des générations
$z = 0$	5.26 ± 0.16	10.91 ± 0.015
$z = \mu$	6.65 ± 0.15	11.14 ± 0.029
$z = \tilde{z}$	6.88 ± 0.15	11.09 ± 0.029

métrique n'est pas optimale pour évaluer le système(Toutanova & Brockett, 2016) mais est simple et interprétable. La baseline $z = 0$ correspond à une génération de phrase non conditionnée par la phrase source. Le cas $z = \mu$ correspond à une génération conditionnée par la phrase source x. Enfin, le cas $z = \tilde{z}$ correspond à la génération conditionnée par x et la longueur y de la phrase cible en y retranchant l'écart type de y pour compresser davantage. Les résultats présentés à la table 2 démontrent que le modèle compression raccourcit les phrases en conservant des mots présents dans les phrases cibles donc a priori importants. Bien sûr les résultats sont plus faibles que des modèles dédiés à la tâche mais notre approche est très générale et évaluée sur des données différentes de celles de l'apprentissage de l'auto-encodeur.

5 Travaux connexes

Ce travail s'inscrit dans le domaine des changement de style de texte (Shardlow, 2014; Xu *et al.*, 2012), mais avec une approche radicalement différente, plus générale et souple. Des analogies sont possibles avec les méthodes de transfert de style appliquées au traitement d'image (Gatys *et al.*, 2015), où on rapproche les textures d'une image à celles que l'on désire.

Les VAE conditionnés ont été utilisés dans d'autres travaux. (Kingma *et al.*, 2014) considère dans le modèle M2 que y est une variable latente concaténée à z, et z est utilisée pour prédire y dans le modèle M1 mais sans interprétation probabiliste et dans l'optique d'apprentissage semi-supervisé.

Le modèle le plus proche du notre est (Hu *et al.*, 2017), pré-publié récemment et qui s'attaque au même problème avec une procédure moins découplée (prise en compte de nouveaux indicateurs bien plus coûteuse) et en attachant le style à la variable latente z. Leur évaluation ne porte pas sur une tâche établie comme la compression mais sur le changement de polarité de reviews de films. Enfin,(Suzuki *et al.*, 2016) utilise un modèle graphique comparable pour l'apprentissage multimodal, cadre aussi exploré par (Sohn *et al.*, 2015). Notre approche est la seule applicable au texte, et qui permette si simplement d'ajouter de nouveaux changements de style.

6 Conclusion

On a introduit un cadre faiblement supervisé de changement de style unifiant plusieurs tâches, et montré des éléments de faisabilité. L'approche est très générale et peut aider à comprendre les représentions latentes. Des applications concrètes peuvent maintenant être envisagées avec de meilleurs modèles d'apprentissage non supervisé ou en utilisant des heuristiques pour sélectionner les meilleures phrases générées par le modèle.

Références

BENGIO Y., DUCHARME R., VINCENT P. & JANVIN C. (2003). A Neural Probabilistic Language Model. *The Journal of Machine Learning Research*, **3**, 1137–1155.

BIRD S., KLEIN E. & LOPER E. (2009). *Natural Language Processing with Python*. O'Reilly Media, Inc.

BOWMAN S. R., VILNIS L., VINYALS O., DAI A. M., JOZEFOWICZ R. & BENGIO S. (2016). Generating sentences from a continuous space. In *Proceedings of the Twentieth Conference on Computational Natural Language Learning*, p. 10–21, Berlin, Germany : Association for Computational Linguistics.

GATYS L. A., ECKER A. S. & BETHGE M. (2015). A Neural Algorithm of Artistic Style. *arXiv preprint*, p. 1–16.

HU Z., YANG Z., LIANG X., SALAKHUTDINOV R. & XING E. P. (2017). Controllable Text Generation. *arXiv preprint*, p. 1–10.

KINGMA D. P. & BA J. (2015). Adam : A Method for Stochastic Optimization. In *International Conference on Learning Representations*.

KINGMA D. P., REZENDE D. J., SHAKIR M. & WELLING M. (2014). Semi-supervised Learning with Deep Generative Models. In *Advances in Neural Information Processing Systems 27*, p. 3581–3589.

KINGMA D. P., SALIMANS T., JOZEFOWICZ R., CHEN X., SUTSKEVER I. & WELLING M. (2016). Improved variational inference with inverse autoregressive flow. In *Advances in Neural Information Processing Systems 29*, p. 4743–4751.

KINGMA D. P. & WELLING M. (2014). Auto-Encoding Variational Bayes. In *International Conference on Learning Representations*.

PITLER E. (2010). *Methods for Sentence Compression*. Rapport interne, University of Pennsylvania.

PRESS O. & WOLF L. (2017). Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics : Volume 2, Short Papers*, p. 157–163, Valencia, Spain : Association for Computational Linguistics.

SHARDLOW M. (2014). A Survey of Automated Text Simplification. *International Journal of Advanced Computer Science and Applications*, (Special Issue on Natural Language Processing), 58–70.

SOHN K., LEE H. & YAN X. (2015). Learning Structured Output Representation using Deep Conditional Generative Models. In *Advances in Neural Information Processing Systems*, p. 3483–3491.

SPEER R. & CHIN J. (2016). An Ensemble Method to Produce High-Quality Word Embeddings. *arXiv preprint*, p. 1–12.

SUTSKEVER I., VINYALS O. & LE Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27*, p. 3104–3112.

SUZUKI M., NAKAYAMA K. & MATSUO Y. (2016). Joint Multimodal Learning With Deep Generative Models. *arXiv preprint*, p. 1–12.

TOUTANOVA K. & BROCKETT C. (2016). A Dataset and Evaluation Metrics for Abstractive Compression of Sentences and Short Paragraphs. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP-16)*, p. 340–350.

XU W., RITTER A., DOLAN B., GRISHMAN R. & CHERRY C. (2012). Paraphrasing for style. In *Proceedings of COLING 2012*, p. 2899–2914, Mumbai, India : The COLING 2012 Organizing Committee.

ZHANG B. & XIONG D. (2016). Variational Neural Machine Translation. p. 521–530.

ZILLY J. G., SRIVASTAVA R. K., KOUTNÍK J. & SCHMIDHUBER J. (2016). Recurrent Highway Networks. *arXiv preprint*, p. 1–12.

Amélioration de la similarité sémantique vectorielle par méthodes non-supervisées

El Moatez Billah Nagoudi[1] Jérémy Ferrero[2] Didier Schwab[2]
(1) Laboratoire d'Informatique et de Mathématique LIM, Université Echahid Hamma Lakhdar d'ElOued, Algérie
(2) LIG-GETALP, Univ. Grenoble Alpes, Grenoble, France.
e.nagoudi@lagh-univ.dz, jeremy.ferrero@imag.fr, didier.schwab@imag.fr

RÉSUMÉ

Mesurer la similarité sémantique est à la base de nombreuses applications. Elle joue un rôle important dans divers domaines tels que la recherche d'information, la traduction automatique, l'extraction d'information ou la détection de plagiat. Dans cet article, nous proposons un système fondé sur le plongement de mots (*word embedding*). Ce système est destiné à mesurer la similarité sémantique entre des phrases en arabe. L'idée principale est d'exploiter la représentation des mots par des vecteurs dans un espace multidimensionnel, afin de faciliter leur analyse sémantique et syntaxique. Des pondérations dépendant de la fréquence inverse en documents et de l'étiquetage morpho-syntaxique sont appliquées sur les phrases examinées, afin d'améliorer l'identification des mots qui sont plus importants dans chaque phrase. La performance de notre système est confirmée par la corrélation de Pearson entre nos scores de similarité assignés et les jugements humains sur un corpus de référence de l'état de l'art sur des phrases en arabe.

ABSTRACT

Improved the Semantic Similarity with Weighting Vectors

Semantic textual similarity is the basis of countless applications and plays an important role in diverse areas, such as information retrieval, plagiarism detection, information extraction and machine translation. This article proposes an innovative word embedding-based system devoted to calculate the semantic similarity between sentences. The main idea is to exploit the word representations as vectors in a multidimensional space to capture the semantic and syntactic properties of words. IDF weighting and Part-of-Speech tagging are applied on the examined sentences to support the identification of words that are highly descriptive in each sentence. The performance of our proposed system is confirmed through the Pearson correlation between our assigned semantic similarity scores and human judgments on a dataset of the state of the art on arabic sentences.

MOTS-CLÉS : Similarité sémantique de phrases, Représentation vectorielle, Pondération de vecteurs, Arabe.

KEYWORDS: Semantic Sentences Similarity, Word Embedding, Vectors Weighting, Arabic.

1 Introduction

La similarité entre textes est une tâche qui trouve des applications dans divers domaines de recherche tels que la recherche d'information, la traduction automatique, la classification des textes, l'extraction

d'information ou la détection de plagiat.

Dans cet article, nous concentrons notre étude sur le problème de mesure de similarité sémantique entre les phrases en arabe. Notre étude compare plusieurs méthodes de calcul de représentations vectorielles d'une phrase à partir de vecteurs de mots, utilisant chacune différentes pondérations, dépendant par exemple de la fréquence inverse en documents (IDF) ou de la partie du discours (POS) des mots de cette phrase. L'idée sous-jacente est d'améliorer la représentation vectorielle d'une phrase en identifiant les mots qui sont les plus importants dans cette phrase.

Cet article est organisé comme suit, dans la section 2 nous présentons les modèles les plus utilisés pour représenter les mots dans un espace vectoriel. Dans la section 3, nous présentons un aperçu général sur le système proposé. Nous proposons quatre variantes de notre système de mesure de similarité dans la section 4. Enfin, la section 5 présente et discute les résultats de chacune de nos approches.

2 Modèles de représentation vectorielle continue de mots

La représentation des mots par vecteurs dans un espace multidimensionnel, en prenant en compte le contexte des mots, permet de conserver les propriétés sémantiques et syntaxiques de la langue (Mikolov *et al.*, 2013a). Cette technique se fonde sur la construction d'un modèle sémantique qui projette les termes d'une langue dans un espace dans lequel certains liens sémantiques entre ces termes peuvent être observés et mesurés. Dans la littérature, plusieurs techniques ont été proposées pour construire des modèles vectoriels (voir Schwab (2005) pour un état de l'art plus complet) mais dernièrement, ces techniques sont principalement issues des modèles de langue neuronaux (Bengio *et al.*, 2003).

Parmi les plus utilisés, on peut citer word2vec (Mikolov *et al.*, 2013c) ou GloVe (Pennington *et al.*, 2014). Mikolov *et al.* (2013c) utilisent un réseau de neurones récurrent (Mikolov *et al.*, 2010) pour apprendre une représentation vectorielle des mots. Les poids du réseau formé sont utilisés comme des vecteurs pour représenter les mots. Ce modèle, appelé *"Word2Vec"*, est capable de capturer des régularités sémantiques et syntaxiques. Ainsi, les angles entre les projections des mots sont corrélés aux relations qui les relient, par exemple : singulier/pluriel, féminin/masculin ou pays/capitale. Grâce à cette observation, il est possible d'exploiter ces relations avec des opérations arithmétiques simples sur ces vecteurs. Il existe deux variantes de *"Word2Vec"*, la première appelée Sac-de-Mots Continue CBOW (*Continuous Bag-Of-Words*) (Mikolov *et al.*, 2013a), et la seconde, Skip-gram, (Mikolov *et al.*, 2013b). L'objectif de l'approche CBOW est de prédire un mot à partir de son contexte d'apparition en utilisant une fenêtre de mots. Soit une suite de mots $S = w_1, w_2, \ldots, w_i$, le modèle CBOW permet de prédire les mots w_k de leurs co-textes $(w_{k-l}, , w_{k-1}, w_{k+1}, , w_{k+l})$. En revanche, la deuxième approche Skip-gram permet de prédire, pour un mot donné, le co-texte dont il est issu.

3 Description du système

Dans les travaux de Mikolov *et al.* (2013c), les approches de Collobert & Weston (2008), Turian *et al.* (2010), Mnih & Hinton (2009) et Mikolov *et al.* (2013c) sont évaluées et comparées. Ils montrent ainsi que CBOW et Skip-gram sont beaucoup plus efficaces en termes de rapidité et de précision pour

apprendre un modèle que les techniques précédentes.

C'est d'ailleurs pour cette raison que nous avons adopté dans notre travail le modèle CBOW. Par conséquent, dans le but de mesurer la similarité sémantique entre les phrases en arabe, nous avons utilisé le modèle CBOW pour représenter les mots arabes[1] (Zahran *et al.*, 2015). Ce modèle a été appris à partir de différentes sources comme le Wikipedia Arabe (WikiAr, 2006), le corpus arabe de la BBC et de CNN (Saad & Ashour, 2010) et le corpus "open parallel" (Tiedemann, 2012) représentant quelques $5,8$ milliards de mots.

Les performances du modèle CBOW dépendent essentiellement de la tâche et de la bonne définition des paramètres d'apprentissage. Après quelques tests sur des corpus échantillons, nous avons choisi de conserver les paramètres suivant : 300 pour la taille des vecteurs, 5 pour la taille de la fenêtre de contexte, 10^{-5} pour le seuil de sous-échantillonnage des mots fréquents, 10 en exemples négatifs dans l'apprentissage et 100 pour le nombre de mots les moins fréquents à supprimer.

3.1 Similarité entre mots

L'utilisation d'un modèle de représentation vectorielle continue de mots permet d'observer et de mesurer les relations sémantiques entre ces mots. Dans ce sens, nous avons utilisé le modèle CBOW (Zahran *et al.*, 2015) afin d'identifier les relations sémantiques entre deux mots w_i et w_j comme par exemple la synonymie ou le changement de nombre et de genre. La similarité entre w_i et w_j est obtenue en comparant leur représentation vectorielle v_i et v_j en utilisant la similarité cosinus. Par exemple, soit les trois mots: "الجامعة" *(l'université)*,"المساء" (*le soir*) et "الكلية" (*la fac*), la similarité entre eux sera :

$$Sim(\text{المساء},\text{الجامعة}) = Cos(v(\text{المساء}), v(\text{الجامعة})) = 0,13$$

$$Sim(\text{الكلية},\text{الجامعة}) = Cos(v(\text{الجامعة}), v(\text{الكلية})) = 0,72$$

L'interprétation est que les mots "الكلية" (*la fac*) et "الجامعة" (*l'université*) sont sémantiquement plus proches que les mots "المساء" (*le soir*) et "الجامعة" *(l'université*).

3.2 Similarité entre phrases

Soit deux phrases, $S_1 = w_1, w_2, ..., w_i$ et $S_2 = w'_1, w'_2, ..., w'_j$, leurs vecteurs étant respectivement $(v_1, v_2, ..., v_i)$ et $(v'_1, v'_2, ..., v'_j)$. Pour mesurer la similarité sémantique entre S_1 et S_2, nous effectuons la somme (pondérée) des vecteurs de leurs mots. Nous avons utilisé quatre pondérations différentes durant cette somme : une pondération unitaire où tous les vecteurs ont le même poids, une pondération des mots avec leur fréquence inverse en documents, une pondération des mots dépendant de leur étiquetage morpho-syntaxique et une pondération mixte, combinaison des deux précédentes. La figure 1 illustre un aperçu global de notre procédure de calcul de similarité entre deux phrases dans notre système.

[1] https://sites.google.com/site/mohazahran/data

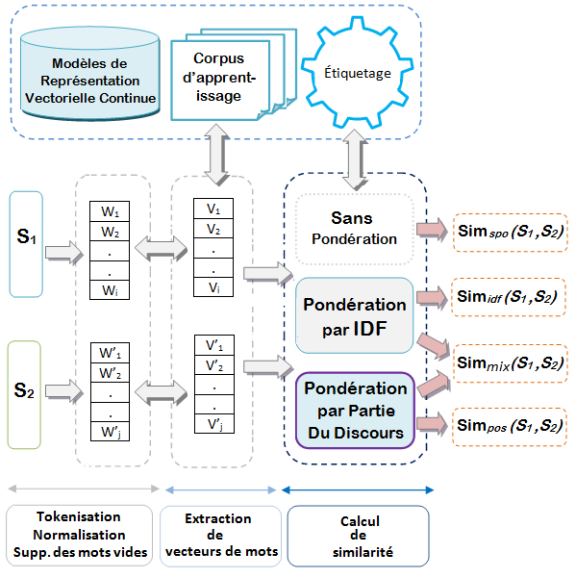


Figure 1: L'architecture du système proposé

4 Méthodes proposées

Dans ce qui suit, nous expliquons nos quatre méthodes proposées pour calculer la représentation vectorielle d'une phrase. Dans tous les cas, il s'agit d'une somme pondérée des vecteurs des mots de la phrase. Nous effectuons une somme de vecteurs de mots afin de prouver qu'une approche intuitive, optimisant une somme de vecteurs de mots en pondérant individuellement chacun des mots en fonction de leur importance dans la phrase, peut être tout aussi performante que des approches utilisant des réseaux de neurones calculant directement un plongement pour une phrase, un paragraphe ou un document, sensible elles à la tâche ou au contexte.

Nous comparons quatre pondérations, une pondération unitaire où tous les mots ont le même poids, une pondération basée sur la fréquence inverse en document des mots, une pondération basée sur la partie du discours des mots et une pondération basée sur la combinaison des deux pondérations précédentes.

Pondération unitaire Il s'agit de la méthode la plus naïve, celle généralement utilisée pour calculer la représentation vectorielle d'une phrase. Il s'agit de la somme des vecteurs des mots composants la phrase, soit $V = \sum_{k=1}^{i} v_k$ où v_k est le vecteur du kième mot de la phrase.

Cependant, nous pensons qu'il y a des mots plus importants que d'autres dans une phrase, des mots qui ont plus de chance d'être conservés même lors d'une paraphrase ou d'une reformulation, méritant donc ainsi un poids plus important dans la représentation de cette phrase. Nous faisons l'hypothèse que l'importance de ces mots peut être déterminée par leur rôle dans la phrase ou par leur rareté dans l'usage de leur langue.

Pondération avec la fréquence inverse en document La fréquence inverse en documents (IDF) est une mesure statistique qui permet d'évaluer l'importance d'un mot dans un corpus (Salton & Buckley, 1988). C'est une mesure très utilisée dans le domaine de la recherche d'information (Turney & Pantel, 2010). L'idée sous-jacente est que les mots les moins fréquents sont considérés comme plus discriminants. Notre pondération donne ainsi un poids plus important aux mots moins fréquents.

Ainsi, la représentation vectorielle d'une phrase devient alors $V = \sum_{k=1}^{i} idf(w_k) * v_k$ où v_k est le vecteur du $k^{\text{ième}}$ mot de la phrase et $idf(w)$ la fonction qui donne l'IDF du mot w.

Pour calculer les pondération IDF de chaque mot, nous avons utilisé le corpus arabe de la BBC et CNN[2] (Saad & Ashour, 2010) comme corpus d'apprentissage. La pondération IDF de chaque mot w_k est obtenue en appliquant la formule $idf(w_k) = log(\frac{S}{WS})$ où S représente le nombre total de phrases dans le corpus et WS est le nombre de phrases contenant le mot w_k.

Pondération avec l'étiquetage morpho-syntaxique Cette idée, inspirée de Schwab (2005) et plus récemment de Ferrero *et al.* (2017), consiste à donner un poids au vecteur d'un mot en fonction de son étiquette morpho-syntaxique.

Ainsi, la représentation vectorielle d'une phrase devient alors $V = \sum_{k=1}^{i} Pos_weight(Pos_{w_k}) * v_k$ où v_k est le vecteur du $k^{\text{ième}}$ mot de la phrase et $Pos_weight(Pos_w)$ est la fonction qui retourne le poids de la partie du discours du mot w.

Alors que Schwab (2005) et Ferrero *et al.* (2017) utilisent des approches supervisées, nous reprenons ici l'aproche de Lioma & Blanco (2009) qui est entièrement non-supervisée. Le principe est d'étiqueter morpho-syntaxiquement le corpus de test en utilisant l'analyseur morpho-syntaxique de Gahbiche-Braham *et al.* (2012), puis de calculer pour chacune des parties du discours sa fréquence inverse en documents et d'en déduire le poids à attribuer à chaque partie du discours. La pondération pos de chaque partie du discours pos_i est obtenue en appliquant la formule $idf(pos_i) = log(\frac{S}{PS})$ où S représente le nombre total de phrases dans le corpus et PS est le nombre de phrases contenant la partie du discours pos_i.

Pondération mixte Cette dernière pondération utilise à la fois la pondération IDF et la pondération en parties du discours. Ainsi, avec cette pondération, la représentation vectorielle d'une phrase devient alors $V = \sum_{k=1}^{i} idf(w_k) * Pos_weight(Pos_{w_k}) * v_k$ où $Pos_weight(Pos_w)$ est la fonction qui retourne le poids de la partie du discours correspondant au mot w et $idf(w)$ la fonction qui donne l'IDF du mot w.

5 Expériences et résultats

Principe Afin d'évaluer les performances de nos méthodes, nous avons utilisé quatre corpus de tests tirés de la sous-tâche 1 de la tâche 1 (détection de similarité textuelle sémantique sur des paires en langue arabe) de la campagne d'évaluation SemEval 2017 (Agirre *et al.*, 2017)[3], comprenant les trois corpus d'essai suivant :

[2] https://sourceforge.net/projects/ar-text-mining/files/Arabic-Corpora/
[3] http://alt.qcri.org/semeval2017/task1/index.php?id=data-and-tools

- *Microsoft Research Paraphrase Corpus* (MSR-Paraphrase) composé de 510 paires ;

- *Microsoft Research Video Description Corpus* (MSR-Video) composé de 368 paires ;

- *WMT2008 Development Dataset* (SMT-europarl) composé de 203 paires ;

ainsi que le corpus d'évaluation composé de 250 paires.

Dans chacun de ces corpus, cinq annotateurs ont donné un score entre 0 et 5 aux paires de phrases, un score de 0 indiquant que le sens des deux phrases est totalement différent et un score de 5 indiquant que les deux phrases ont exactement le même sens. Le score final de chaque paire est la moyenne des scores des cinq annotateurs.

Évaluation Les performances de nos quatre approches ont donc été évaluées sur plus de 1338 paires de phrases. Nous avons calculé la corrélation de Pearson entre les scores de similarité sémantique résultants de nos méthodes et les jugements humains. Les résultats sont présentés dans le Tableau 1.

Corpus	MSRpar	MSRvid	SMTeuroparl	STS 2017	Global
Pondération unitaire	0,6745	0,7233	0,6233	0,5957	0,6653
Pondération avec IDF	0,7432	0,7820	0,7110	0,7309	0,7467
Pondération avec POS	0,7446	0,7951	0,7317	0,7425	0,7562
Pondération mixte	**0,7523**	**0,8276**	**0,7460**	**0,7646**	**0,7745**
Baseline de SemEval 2017	/	/	/	0,6045	/
BIT (gagnant de SemEval 2017)	/	/	/	0,7543	/

Table 1: Performances de nos quatre méthodes ainsi que ceux des organisateurs et des vainqueurs de SemEval 2017.

Ces résultats montrent que sur chacun des corpus considéré, la pondération unitaire est nettement moins performante que les autres. La pondération par fréquence inverse en documents et celle basée sur les parties du discours obtiennent des résultats comparables (respectivement $+8,14\%$ et $+9,09\%$ par rapport à la pondération unitaire), même si la seconde est légèrement meilleure. Enfin, la combinaison de ces deux pondérations donne les meilleurs résultats ($+10,92\%$ vis-à-vis de la pondération unitaire).

En ce qui concerne la campagne d'évaluation SemEval 2017 (Agirre *et al.*, 2017), la dernière pondération serait arrivée en tête avec une corrélation de $0,7646$ contre 0.7543 pour l'équipe gagnante, BIT (Wu *et al.*, 2017). À l'heure où nous écrivons ces lignes, nous n'avons pas pu lire en détail ce travail mais nous savons qu'il s'agit d'une approche utilisant intensivement WordNet alors que notre méthode n'exploite pas de telles données et pourrait être utilisée avec des langues qui n'en disposent donc pas.

6 Conclusion

Dans cette article, nous avons présenté un système d'estimation de la similarité sémantique entre phrases en arabe basée sur l'utilisation de représentations continues de mots. Nous avons montré qu'identifier les mots les plus importants d'une phrase permettait d'améliorer fortement les perfor-

mances d'un tel système. En effet, une pondération des vecteurs des mots qu'elle se fasse en se basant sur la fréquence inverse en documents ou en fonction de la partie du discours est bien plus efficace que si la pondération est la même pour chaque mot. Les performances des systèmes non-supervisés proposés sont confirmées par la corrélation de Pearson avec des jugements humains sur quatre corpus différents. Une pondération combinant à la fois la fréquence inverse en documents et la parties du discours des mots aurait remporté la récente tâche 1 de la campagne d'évaluation SemEval 2017. Les pistes de travail que nous poursuivons actuellement cherchent à combiner ces résultats à ceux d'autres techniques classiques dans le domaine (fingerprint, n-gram, etc.) ou à ceux d'autres ressources (bases lexicales pour enrichir des phrases de synonymes, génériques, voire traductions dans d'autres langues).

Références

AGIRRE E., CER D., DIAB M., INIGO LOPEZ-GAZPI & SPECIA L. (2017). Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of Semeval 2017*.

BENGIO Y., DUCHARME R., VINCENT P. & JAUVIN C. (2003). A neural probabilistic language model. *Journal of machine learning research*, **3**(Feb), 1137–1155.

COLLOBERT R. & WESTON J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, p. 160–167: ACM.

FERRERO J., BESACIER L., SCHWAB D. & AGNÈS F. (2017). Using word embedding for cross-language plagiarism detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, p. 415–421, Valencia, Spain: Association for Computational Linguistics.

GAHBICHE-BRAHAM S., BONNEAU-MAYNARD H., LAVERGNE T. & YVON F. (2012). Joint segmentation and pos tagging for arabic using a crf-based classifier. In *LREC*, p. 2107–2113.

LIOMA C. & BLANCO R. (2009). Part of speech based term weighting for information retrieval. In *European Conference on Information Retrieval*, p. 412–423: Springer.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013a). Efficient estimation of word representations in vector space. In *In: ICLR: Proceeding of the International Conference on Learning Representations Workshop Track*, p. 1301–3781.

MIKOLOV T., KARAFIÁT M., BURGET L., CERNOCKỲ J. & KHUDANPUR S. (2010). Recurrent neural network based language model. In *Interspeech*, volume 2, p.3̃.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. S. & DEAN J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, p. 3111–3119.

MIKOLOV T., YIH W.-T. & ZWEIG G. (2013c). Linguistic regularities in continuous space word representations. In *Hlt-naacl*, volume 13, p. 746–751.

MNIH A. & HINTON G. E. (2009). A scalable hierarchical distributed language model. In D. KOLLER, D. SCHUURMANS, Y. BENGIO & L. BOTTOU, Eds., *Advances in Neural Information Processing Systems 21*, p. 1081–1088. Curran Associates, Inc.

PENNINGTON J., SOCHER R. & MANNING C. D. (2014). Glove: Global vectors for word representation. In *EMNLP*, volume 14, p. 1532–1543.

SAAD M. K. & ASHOUR W. (2010). Osac: Open source arabic corpora. In *6th ArchEng Int. Symposiums, EEECS*, volume 10.

SALTON G. & BUCKLEY C. (1988). Term-weighting Approaches in Automatic Text Retrieval. In *Information Processing and Management*, volume 24, p. 513–523, Tarrytown, NY, USA: Pergamon Press, Inc.

SCHWAB D. (2005). *Approche hybride-lexicale et thématique-pour la modélisation, la détection et l'exploitation des fonctions lexicales en vue de l'analyse sémantique de texte*. PhD thesis, Université Montpellier II.

TIEDEMANN J. (2012). Parallel data, tools and interfaces in opus. In *LREC*, volume 2012, p. 2214–2218.

TURIAN J., RATINOV L. & BENGIO Y. (2010). Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, p. 384–394: Association for Computational Linguistics.

TURNEY P. D. & PANTEL P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, **37**, 141–188.

WIKIAR (2006). Arabic wikipedia corpus, http://linguatools.org/tools/corpora/wikipedia-monolingual-corpora/, (accessed january 21,2017).

WU H., HUANG H., JIAN P., GUO Y. & SU C. (2017). Bit at semeval-2017 task 1: Using semantic information space to evaluate semantic tex- tual similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval 2017)*.

ZAHRAN M. A., MAGOODA A., MAHGOUB A. Y., RAAFAT H., RASHWAN M. & ATYIA A. (2015). Word representations in vector space and their applications for arabic. In *International Conference on Intelligent Text Processing and Computational Linguistics*, p. 430–443: Springer.

Typologies pour l'annotation de textes non standard en français

Louise Tarrade[1] Cédric Lopez[1] Rachel Panckhurst[2] Georges Antoniadis[3]
(1) Viseo R&D, 4 avenue doyen Louis Weil, 38000 Grenoble, France
(2) Praxiling UMR 5267 CNRS & Université Paul-Valéry Montpellier 3, France
(3) LIDILEM, Université Grenoble Alpes, Grenoble, France
prenom.nom@viseo.com, rachel.panckhurst@univ-montp3.fr,
georges.antoniadis@univ-grenoble-alpes.fr

RESUME

La tâche de normalisation automatique des messages issus de la communication électronique médiée requiert une étape préalable consistant à identifier les phénomènes linguistiques. Dans cet article, nous proposons deux typologies pour l'annotation de textes non standard en français, relevant respectivement des niveaux morpho-lexical et morpho-syntaxique. Ces typologies ont été développées en conciliant les typologies existantes et en les faisant évoluer en parallèle d'une annotation manuelle de tweets et de SMS.

ABSTRACT

Typologies for the annotation of non-standard French texts.
The task of automatic normalization of messages resulting from mediated electronic communication requires a preliminary step consisting of identifying linguistic phenomena. In this paper, two typologies for annotating non-standard texts in French, from both morpholexical and morphosyntactic perspectives are proposed. We have elaborated these from existing typologies and modified them according to manual annotation of tweets and SMS.

MOTS-CLÉS : typologie, SMS, tweets, normalisation
KEYWORDS: typology, SMS, tweets, normalization

1 Introduction

Ces dernières années ont été marquées par l'avènement des messages textuels courts, notamment les tweets et les SMS. Les outils de traitement automatique de la langue sont généralement conçus pour traiter du texte « standard », or ces messages courts font l'objet d'écrits présentant des variantes de graphie qui s'éloignent de la langue standardisée, ce qui implique un traitement automatique difficile. De nombreuses applications nécessitent pourtant l'analyse de tels textes, par exemple l'extraction d'informations médicales à partir des SMS des patients (Stenner *et al.*, 2012), l'analyse d'opinions et de sentiments dans les tweets (Vinodhini & Chandrasekaran, 2012), ou encore la synthèse vocale (Ill & Ford, 2011). Afin de pallier cette difficulté il est nécessaire de transformer les écrits « non standard » en écrits « standard », *i.e.* de normaliser, ou transcoder (Beaufort & *al.* 2010). Un premier pas vers le développement d'un outil de normalisation consiste à dresser une typologie fine des phénomènes linguistiques apparaissant dans les écrits non standard. Une telle typologie permettra d'annoter des textes et ainsi d'identifier les phénomènes les plus fréquents dans un corpus

donné. Nous pourrons alors ensuite focaliser le travail sur le traitement des phénomènes, en fonction de leur fréquence, avec des modes de traitement adaptés.

Dans cet article, nous discutons des typologies déjà existantes (section 2), puis nous nous focalisons sur la construction d'une typologie représentant les phénomènes linguistiques présents dans les tweets et les SMS (section 3).

2 Travaux antérieurs

Le développement de typologies dans le contexte du discours numérique médié (désormais DNM) (Panckhurst, 2017) a logiquement suscité l'attention des chercheurs cette dernière décennie. Une des premières typologies répertoriant les variations graphiques et les aspects morpho-lexicaux de la DNM a été établie par (Anis, 2004). La typologie de (Fairon *et al.*, 2006) consiste en une classification générale de l'écriture SMS, incluant donc des phénomènes liés à l'orthographe phonétique, la morphosyntaxe, à la syntaxe et au discours. Nous avons besoin d'augmenter le degré de granularité considéré dans ces typologies afin de distinguer, par exemple, le cas des abréviations sémantisées (t→te/tu) (Roche *et al.* 2016), des squelettes consonantiques (dsl→désolé), tous deux classés parmi les abréviations. Plus récemment, (Cougnon *et al.*, 2013) proposent une typologie des stratégies réductionnelles de l'écriture SMS, afin de « *présenter les variations graphiques présentes dans l'écrit SMS* » (Cougnon *et al.*, 2013). Dans cette typologie, les phénomènes n'appartenant pas à des stratégies réductionnelles (les variations orthographiques) sont considérés, et l'on y inclut également des « erreurs » telles qu'une mauvaise utilisation du temps, l'inversion indicatif/impératif, l'accord du participe passé, *etc*., que nous réservons à une typologie d'ordre morpho-syntaxique. Enfin, la typologie publiée dans (Panckhurst, 2009), modifiée très récemment dans (Roche *et al.*, 2016) porte exclusivement sur les phénomènes de néographie de l'écriture SMS, en tenant compte des typologies précédentes. Son originalité repose entre autres sur le fait qu'elle distingue des phénomènes tels que la substitution, la réduction, la suppression ou l'ajout.

Ces différentes typologies ont principalement été développées dans un but descriptif et se sont focalisées sur les SMS. Dans la pratique, l'annotation manuelle fondée sur les typologies antérieures implique des choix subjectifs et/ou multiples. À partir des typologies citées précédemment, notre objectif est de 1) dresser une typologie dont les classes ont une intersection minimale afin de faciliter la tâche d'annotation manuelle, 2) étendre la typologie à des phénomènes apparaissant dans twitter, pour lequel il semble qu'aucune typologie n'ait été établie.

3 Conception et présentation de la typologie

Notre typologie concilie les typologies de (Roche *et al.*, 2016), (Fairon *et al.*, 2006), (Cougnon *et al.*, 2013) et (Anis, 2004) avec l'objectif d'une annotation dans le cadre d'une tâche de normalisation automatique. Nous exposons dans cette section notre méthodologie qui a conduit à la construction de notre typologie.

3.1 Vers une typologie modifiée pour le TAL

(Roche *et al.*, 2016) considèrent deux phénomènes possibles de suppression graphique : *typographie et ponctuation* (l'absence de ponctuation finale par exemple) et *signes diacritiques* (la suppression

des signes diacritiques) ; or, nous considérons la catégorie *signes diacritiques* comme étant plus appropriée à la définition de la substitution (remplacement de la graphie ou une partie de la graphie par une autre), et l'absence de ponctuation comme relevant plutôt du niveau syntaxique. De fait, notre typologie conserve trois catégories principales : substitution, réduction et ajout, que nous présenterons plus précisément dans la section suivante.

Dans un souci de démarcation plus nette des frontières entre les phénomènes, nous avons choisi de ne pas conserver la distinction effectuée par (Roche *et al.*, 2016) entre les variations graphiques et phonétisées. En effet, il nous semble important de repérer les modifications de graphies qui altèrent la prononciation du mot standard ou non, dans le cadre d'un éventuel recours à l'aspect phonétique du texte lors de la tâche de normalisation ; or, dans la typologie de (Roche *et al.,* 2016), la distinction phonétisé/graphique ne semble pas avoir cette vocation. Par exemple, le cas « mwa » (au lieu de « moi ») correspond à une substitution graphique avec variation dans la typologie de (Roche *et al.*, 2016) ; dans notre typologie, nous le considérons comme un cas de substitution de la graphie partielle d'un mot avec correspondance phonologique. Nous avons également réorganisé certains phénomènes : c'est le cas, par exemple, des abréviations de consonnes doubles ou de chute du « e » instable, de fin muette d'un mot ou de son début muet, répertoriés comme des cas de réductions par (Roche *et al.*, 2016), mais que nous préférons assimiler à des cas de substitutions graphiques. De la même façon, la réduction phonétisée entière (*c→ces*) nous paraît être un cas de substitution d'un mot par une lettre unique avec correspondance phonologique, étant l'initiale de celui-ci.

La tâche de normalisation automatique à laquelle nous aspirons nous impose d'avoir une typologie la plus couvrante possible, avec un niveau de description des phénomènes assez fin ; c'est pourquoi ont été intégrés à la typologie des phénomènes tels que l'écrasement décrit par (Anis, 2004) et (Fairon, 2006), mais aussi les possibilités de réduction, d'ajout ou de substitution autres que les phénomènes décrits dans notre typologie, qui peuvent s'assimiler à la majorité des notions d'omission, d'adjonction et de confusion décrites par (Cougnon *et al.*, 2013) dans la catégorie des erreurs orthographiques (*enregirstré→enregistré* ou *descentre→descendre* par exemple). Le phénomène d'hyper-segmentation (*toute fois→toutefois*), peu souvent inclus dans les différentes typologies, nous a paru important à ajouter. Ensuite, tout comme (Anis, 2004) qui proposait dans sa typologie les anglicismes, nous avons décidé de les inclure sous le prisme plus global du *code-switching*, dans sa plus large acception. Au même titre, nous avons ajouté les cas de mots en verlan. Les néologismes et jargons sont également présents dans notre typologie.

Des éléments apparaissant quasi exclusivement dans les tweets ont également été ajoutés, notamment les pointeurs, *i.e.* les mentions et les hashtags, qui peuvent par ailleurs jouer un rôle syntaxique. Cela nous a conduit à considérer le niveau morpho-syntaxique (*i.e.* incluant le niveau syntaxique). Nous considérons donc une seconde typologie, n'ayant pas pour but de décrire exhaustivement les phénomènes morpho-syntaxiques présents dans ce type de texte, mais d'apporter des précisions concernant le niveau morpho-lexical, sur lequel porte principalement notre typologie. Ceci permettra d'affiner les premières annotations et d'apporter un niveau supplémentaire d'information qui constituera une précieuse aide dans l'élaboration de l'outil de normalisation automatique.

Disposant d'un corpus de tweets obtenus grâce à l'API twitter et du corpus de SMS *88milSMS* (http://88milsms.huma-num.fr/ et https://hdl.handle.net/11403/comere/cmr-88milsms), nous avons annoté 1 000 tweets et 1 000 SMS afin de tester notre typologie et de la faire évoluer le cas échéant. À partir de ces annotations, portant sur les phénomènes décrits dans cette typologie, nous avons réadapté cette dernière en ajoutant les phénomènes de contraction et de compactage, ou en faisant la distinction entre les suppressions de signes diacritiques entraînant une approximation phonologique

ou non. Effectivement, l'annotation des SMS et des tweets a révélé que la suppression des signes diacritiques pouvait entrainer une modification de la prononciation du mot dans 43% des cas. Ces évolutions mineures ont été considérées dans une deuxième étape d'annotation manuelle qui nous a déjà permis d'observer, par exemple, une légère différence de proportion dans les phénomènes présents au niveau morpho-lexical dans les tweets et les SMS (Figure 1) (Tarrade & Lopez, 2017).

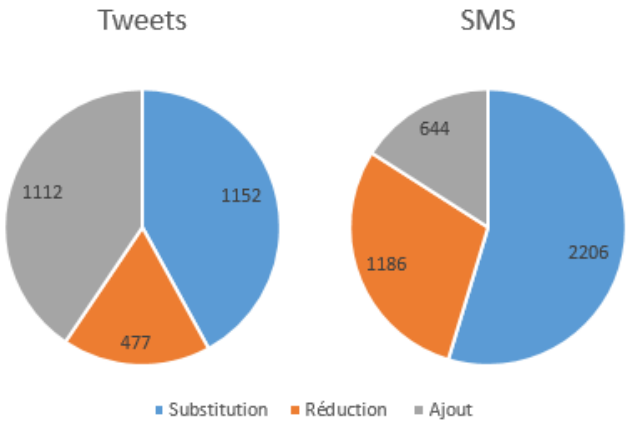


FIGURE 1: Répartition des phénomènes morpho-lexicaux en nombre d'annotations (Tarrade & Lopez, 2017)

Nous proposons ainsi deux typologies développées et mises à l'épreuve dans une tâche d'annotation manuelle pour deux types de textes différents (tweets, SMS), couvrant les niveaux morpho-lexicaux et morpho-syntaxiques.

3.2 Typologies

Nous décrivons dans cette section les catégories de nos typologies relevant respectivement du niveau morpho-lexical (Figures 2, 3 et 4) et morpho-syntaxique (Figure 5). Les exemples proviennent soit de (Roche *et al.*, 2016) et de (Fairon, 2006) pour les catégories communes, soit des corpus de tweets et de SMS que nous avons annotés.

3.2.1 Substitution (typologie morpho-lexicale)

Concernant la définition de la substitution, nous élargissons celle de (Panckhurst, 2009) « La substitution correspond à un remplacement de la graphie ou une partie de la graphie par une autre », en tenant compte du fait que les graphies de substitution peuvent ou non préserver la prononciation du mot, cette distinction étant indispensable dans l'éventualité où l'on souhaiterait traiter des cas de substitution à l'aide d'une étape de phonétisation. Les phénomènes de substitution peuvent donc concerner la graphie complète d'un mot ou groupe de mots, ou une partie de leur graphie. Il peut y avoir soit une correspondance phonologique entre la forme standard et la forme concernée par le phénomène, soit une approximation phonologique.

Substitution	graphie complète d'un mot ou groupe de mot	correspondance phonologique	lettre		c->s'est/ses/sais/sait, g->j'ai, i->y, k->qu'à/cas, l->elle, m->aime, n->haine, o->eau/au, q->cul, r->aire/air, s->est-ce, u->eu
				initiale	b->bé (aussi interjection), c->c'est/ces, d->des/dé/dès, e->eu, h->hache, j->j'y, o->oh, p->paix/pet, t->t'es/tes, v->vais
			chiffre		7->cet
			symbole		+ ->plus
		approximation phonologique			k->que, c->se/ça, z->je
	graphie partielle d'un mot	correspondance phonologique	lettre		bo->beau, Dcédé->décédé, ossi->aussi
			chiffre		2main->demain
			symbole		+sieurs->plusieurs
			graphie		bizes->bises, bisoux->bisous, mwa-moi
				consonne double	come->comme
				chute du "e" instable	douch->douche
				lettre muette enlevée	fou->fous, pa->pas, peu->peut, vou->vous
				lettre muette ajoutée	peux->peu, as->a
				initiale muette	ôtel->hôtel
			signe diacritique		voila->voilà, tantôt->tantot
		approximation phonologique			nan->non, kikou->coucou(avec graphie du 'c' modifiée), pô->pas, ui->oui
			signe diacritique		ca->ça, appelé->appele
	typographie				est ce que->est-ce que
		casse d'un caractère			suzanne -> Suzanne
		élision			m en->m'en
	rébus	caractère spécial			de grandes @ (oreilles)
		emoji			Merci pour vos RTs & 😷
	écrasement				jsuis->chuis (précédemment : agglutination+compactage)
	code-switching				yes -> oui, screen->écran
	néologisme/jargon				catche (compris), walou (rien)
	verlan				wam->moi
	contraction				p'tit->petit
	autre				chzton->chaton, les->la

FIGURE 2: Typologie morpho-lexicale : les cas de substitution

Toujours dans un souci d'adaptation à notre future tâche de normalisation automatique, il nous a paru nécessaire que la granularité de notre typologie soit assez fine. C'est pour cette raison que, comme dans la typologie proposée par (Fairon, 2006), une catégorie est dédiée aux différents types de modifications d'une graphie (ne correspondant pas à une autre possibilité du même niveau) ayant conservée une correspondance phonologique avec le mot d'origine. Sont alors distingués des cas particuliers tels que la substitution d'une consonne doublée par une consonne simple, la chute du « e » instable, de l'initiale muette ou de la lettre muette (qui peut également être ajoutée). Ces éléments sont classés pour la plupart comme des phénomènes de réduction dans les typologies de (Anis, 2004) et de (Roche *et al.*, 2016). Nous préférons les considérer comme des substitutions, car nous pensons cette méthode de classement plus adaptée à la tâche de normalisation, dans l'éventualité où la forme phonétique des mots serait utilisée dans celle-ci, par exemple.

Le phénomène de code-switching est considéré ici dans une large acception, comme c'est également le cas des néologismes ou jargons. Ces phénomènes, tout comme l'utilisation du verlan, sont considérés en tant que phénomènes de substitution. Effectivement, même s'ils n'obéissent pas à la définition stricte donnée précédemment de la substitution, ils correspondent toutefois à une substitution au sens large, car ils sont fréquemment utilisés à la place d'autres termes équivalents en français standard.

3.2.2 *Réduction (typologie morpho-lexicale)*

Réduction	Abrégement morpho-lexical	troncation	apocope		ordi -> ordinateur
			aphérèse		tain -> putain
		sigle/acronyme			"tlm"-> tout le monde, lol->laughing out loud
	abréviation sémantisée				b->ben (aussi interjection)/bien/bon, c->ce, d->de/du/dans, e->eh/et/est/être, g->gueule, h->heure, i->il/île, j->je/jeux, l->le/lesbienne/la, m->me/même/mais/mes/mètre, n->ne, p->page/peux/peut, q->que/qui, 't->tu/te/ton, u->universitaire, v->vœux/veut/veux/va/vas, w->week[-end]
	squelette consonantique				dsl->désolé, pr->pour
	agglutination				jattends -> j'attends
		avec élision			j'vois -> je vois
	compactage				jsuis->jesuis (précédemment : agglutination)
	abréviation				espr->espère
		consonne			dc->donc, cdlt->cordialement, RT->retweet
	autre				chaon->chaton

FIGURE 3: Typologie morpho-lexicale : les cas de réduction

La réduction « correspond à un enlèvement de certains caractères et résulte nécessairement en un nombre inférieur de caractères » (Panckhurst, 2009). Par ailleurs, d'un point de vue de la tâche de normalisation, il ne nous a pas paru pertinent de conserver une distinction entre les sigles et les acronymes. Cependant, préciser les catégories d'agglutination et d'abréviation nous semble approprié ; nous avons ajouté les cas d'agglutination avec élision de la première voyelle, ou encore les cas d'abréviations ne formant pas de squelettes consonantiques mais étant tout de même composées uniquement de consonnes du mot d'origine.

3.2.3 *Ajout (typologie morpho-lexicale)*

Ajout	phonétisé	avec variation			oki->ok, ouaip->ouais
		liaison			zèt->êtes, namour->amour
	allongement	ponctuation			*!!!!!!!!!!!!!!!
		lettre			suuuuuper->super
	smiley				:-)
	emoji				Bonne nuit à tous mes Amis
	onomatopée/interjection				snif, bof
	hyper-segmentation	mot	espace		toute fois->toutefois
			ponctuation		ti-pe->type (avant : i->y (graphie))
		voyelle non élidée			que il->qu'il
	pointeur	hashtag			#Ronaldinho ne s'entraînait pas [...]
		mention			Pr @NicolasSarkozy la création de [...]
	symbole				la *star*
	autre				chzat->chat

FIGURE 4: Typologie morpho-lexicale : les cas d'ajout

L'ajout peut être défini comme l'augmentation du nombre de caractères ou d'éléments tels que des smileys, des emoji, des mentions ou des hashtags, et plus généralement des symboles quelconques. De façon plus fine que les typologies précédentes, nous spécifions si l'ajout concerne, par exemple, une voyelle non-élidée, une séparation du mot par une espace ou par une ponctuation (par exemple dans « S.U.P.E.R »).

3.2.4 Typologie morpho-syntaxique

La typologie morpho-syntaxique (Figure 5) a pour but principal d'apporter une couche d'information nécessaire lors de l'annotation. Deux phénomènes sont relatifs aux tweets : c'est le cas de la troncation de texte, fréquente dans les tweets limités à 140 caractères, et de la catégorie *sans rôle syntaxique*. Cette catégorie nous paraît indispensable ici, puisque comme le soulignent (Kaufmann & Kalita, 2010) les *hashtags* et les mentions ont pour particularité de ne pas toujours jouer un rôle syntaxique dans la phrase.

Niveau morpho-syntaxique					
	sans rôle syntaxique				Moi quand la bourse va arriver #CROUS, RT @Guigabriel92 : [...]
	typographie et ponctuation				guillemets, ponctuations finales, etc.
	conversion				sms-moi qud tu arriv
	Inversion participe passé/infinitif				Ben sa va arrivé->Ben ça va **arriver**
	Inversion des mots grammaticaux				pas comme sa-> pas comme **ça**
	accord	genre			quel annee -> **quelle** année
		nombre			je t'enverrai les photo -> je t'enverrai les **photos**
		personne			moi j'attend 18:15->moi j'**attends** 18:15
	ellipse	mot grammatical			à 17h y a mec dans le bus -> à 17h **il** y a **un** mec dans le bus
		mot lexical			il avec le chat
	Répétition				tu tu vois
	troncation du texte				RT @PlanBatiment : Le rêve pour Philippe Pelletier, la rénovation des bâtiments scolaires. Un @PlanBatiment scolaire ! Envie de porter cette ...
	Autre				

FIGURE 5: Typologie morpho-syntaxique

4 Conclusion

Dans ce travail, nous avons défini deux typologies pour l'annotation du français non standard, l'une morpho-lexicale et l'autre morpho-syntaxique, testées sur 1 000 tweets et 1 000 SMS vis-à-vis de la tâche de normalisation automatique. Les résultats des annotations ainsi que les différentes observations qui en découlent sont présentés de manière plus détaillée dans (Tarrade & Lopez, 2017). La prochaine étape, avant de commencer le développement de l'outil de normalisation, est d'identifier les phénomènes les plus représentés dans les différents corpus. Nous pourrons alors développer un outil de normalisation automatique de textes en français, prenant en compte le type de message et établissant des priorités au niveau des phénomènes. Dans une autre perspective, il serait également intéressant d'observer dans quelle mesure les phénomènes décrits dans ces typologies sont susceptibles de s'appliquer à d'autres langues.

Références

ANIS J., DE FORNEL, M., & FRAENKEL B. (organisateurs, 2004). La communication électronique : Approches linguistiques et anthropologiques. Colloque international, EHESS, Paris, 5-6 février 2004.

BEAUFORT R., ROEKHAUT S., COUGNON L. A., & FAIRON C. (2010, July). A hybrid rule/model-based finite-state framework for normalizing SMS messages. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 770-779). Association for Computational Linguistics.

COUGNON L.-A., ROEKHAUT S., & BEAUFORT R. (2013). Typologies de variation graphique dans l'écrit SMS. S. Baddeley, F. Jejcic et C. Martinez (éd.)*, L'orthographe en quatre temps*, 20, 129-148.

FAIRON C., KLEIN J.-R., & PAUMIER S. (2006). *Le langage SMS : étude d'un corpus informatisé à partir de l'enquête « Faites don de vos SMS à la science »* (p. 31-47). Louvain-la-Neuve : UCL, Presses Univ. de Louvain.

ILL, E. T. G. & FORD, C. S. (2011). *U.S. Patent Application No. 12/983,946*.

KAUFMANN M. & KALITA J. (2010). Syntactic normalization of twitter messages. In *International Conference on Natural Language Processing*.

PANCKHURST (2017), « Entre linguistique et informatique. Des outils de traitement automatique du langage naturel écrit (TALNE) à l'analyse du discours numérique médié (DNM) », habilitation à diriger des recherches, Université Paris-Est.

PANCKHURST R. (2009). Short Message Service (SMS) : typologie et problématiques futures., in : *Polyphonies, pour Michelle Lanvin*. Sous la dir. de Teddy Arnavielle. Université Paul-Valéry Montpellier 3, p. 33-52.

ROCHE M., VERINE B., LOPEZ C. & PANCKHURST R. (2016). « La néographie dans un grand corpus de SMS français : 88milSMS ». In : *La neología en las lenguas románicas Recursos, estrategias y nuevas orientaciones*, Actes du colloque *Cineo 2015*, 22-24 octobre, Salamanque. Sous la dir. de Joaquín García Palacios, Goedele De Sterck, Daniel Linder, Nava Maroto, Miguel Sánchez Ibáñez et Jesús Torres del Rey. Studien zur romanischen Sprachwissenschaft und interkulturellen Kommunikation. Frankfurt, Peter Lang.: DOI: http://dx.doi.org/10.3726/978-3-631-69859-4, p.279-302.

STENNER S. P., JOHNSON K. B., & DENNY J. C. (2012). PASTE: patient-centered SMS text tagging in a medication management system. *Journal of the American Medical Informatics Association*, *19*(3), 368-374.

TARRADE L. & LOPEZ C. (2017). Corpus de tweets et de SMS annotés pour l'observation de phénomènes linguistiques en français « non standard ». *TALN 2017.*

VINODHINI, G., & CHANDRASEKARAN, R. M. (2012). Sentiment analysis and opinion mining: a survey. *International Journal*, *2*(6), 282-292.

Simbow : une mesure de similarité sémantique entre textes

Delphine Charlet, Géraldine Damnati
Orange Labs, Lannion, France
{delphine.charlet,geraldine.damnati}@orange.com

RÉSUMÉ

Cet article décrit une mesure de similarité sémantique non-supervisée qui repose sur l'introduction d'une matrice de relations entre mots, dans un paradigme de mesure cosinus entre sacs de mots. La métrique obtenue, apparentée à *soft*-cosinus, tient compte des relations entre mots qui peuvent être d'ordre lexical ou sémantique selon la matrice considérée. La mise en œuvre de cette métrique sur la tâche qui consiste à mesurer des similarités sémantiques entre questions posées sur un forum, a remporté la campagne d'évaluation SemEval2017. Si l'approche soumise à la campagne est une combinaison supervisée de différentes mesures non-supervisées, nous présentons dans cet article en détail les métriques non-supervisées, qui présentent l'avantage de produire de bons résultats sans nécessiter de ressources spécifiques autres que des données non annotées du domaine considéré.

ABSTRACT

Simbow : a semantic similarity metric between texts

This article describes an unsupervised similarity metric which relies on the introduction of a word relation matrix, in a cosine bag of words paradigm. The obtained *soft*-cosine metric takes into acount word relations that can be lexical or semantic depending on the matrix under consideration. The implementation of this metric in a task consisting of computing semantic similarities between questions from a forum, won the SemEval2017 evaluation campain. If the approach submitted to the campain is a supervised combination of several unsupervised metrics, we present in detail in this article the unsupervised metrics, which already yield good results without needing any specific ressources other than unannotated in domain data.

MOTS-CLÉS : similarité sémantique entre textes, relations sémantiques, questions de forums.

KEYWORDS: semantic textual similarity, semantic relations, forum questions.

1 Introduction

Les forums permettent à leurs participants de poser des questions et d'interagir avec les autres pour obtenir des réponses pertinentes. La popularité des forums montre la capacité de ce type d'interactions à produire des réponses pertinentes aux questions. Leur popularité est telle que le premier réflexe d'un internaute, lorsqu'il se pose une question, est de faire appel à son moteur de recherche préféré, pour vérifier si une question similaire n'a pas déjà été posée et résolue. Pour répondre de façon pertinente à cette requête, il faudrait pouvoir mesurer une similarité entre la nouvelle question posée et les questions déjà postées sur le forum, qui tienne compte de la sémantique et pas seulement du nombre de mots en communs entre les questions, sur quoi s'appuient les moteurs de recherche usuels.

Dans le cadre de la campagne d'évaluation des technologies sémantiques SemEval 2017, la tâche

Community Question Answering, est consacrée à l'analyse des questions et leurs fils de discussions dans les forums (Nakov *et al.*, 2017). Une sous-tâche traite précisément de la similarité entre questions : il s'agit pour une question, dite originale, de trier par similarité sémantique décroissante les 10 questions, dites relatives, remontées par un moteur de recherche sur ce forum. Cette tâche peut être vue comme une pure tâche de similarité sémantique entre textes, sur des données textuelles bruitées générées par des utilisateurs, à la différence d'une autre tâche de similarité textuelle (Agirre *et al.*, 2016), qui traite de textes courts bien formés. Dans la tâche *Community Question Answering*, le corpus d'étude est issu du forum en anglais Qatar Living, dans lequel les expatriés occidentaux discutent de tous les aspects de la vie quotidienne au Qatar (où pratiquer le basket, comment faire pour embaucher une nounou,...). Cette tâche est apparue dans SemEval en 2016 (Nakov *et al.*, 2016), et s'est poursuivie en 2017. Les approches proposées en 2016 étaient principalement des fusions tardives de mesures de similarités supervisées ou non. De nombreuses mesures non-supervisées étaient basées sur le comptage d'éléments en commun entre les questions, ces éléments pouvant être des n-grams de lettres ou de mots, ou des composants de plus haut niveau, comme des entités nommées, des roles sémantiques ou des portions d'arbres syntaxiques (e.g. (Franco-Salvador *et al.*, 2016)). Les plongements de mots ont également été beaucoup utilisés (e.g. (Mihaylov & Nakov, 2016)), souvent simplement moyennés au niveau de la question, et utilisé dans une mesure en cosinus ou en entrée d'un classifieur neuronal.

Les hypothèses de travail que nous avons prises pour cette tâche sont les suivantes : nous avons considéré que les données de forum étaient trop bruitées pour obtenir des sorties fiables de nos outils d'analyse linguistiques, et nous avons souhaité nous focaliser sur la similarité sémantique entre textes. C'est pourquoi nous n'avons utilisé aucune analyse de méta-données (issues des dates, profils utilisateurs etc...), afin d'obtenir des résultats qui pourraient se généraliser à d'autres tâches de similarité sémantique de textes. Ainsi, nous avons développé des mesures de similarité sémantiques entre textes, sans ressources externes, presque sans pré-traitements linguistiques, reposant uniquement sur la disponibilité d'un gros corpus non-annoté représentatif des données.

Contrairement à l'article à paraître lors de la conférence SemEval (Charlet & Damnati, 2017), où nous décrivons en détail l'optimisation de la combinaison supervisée de nos métriques non-supervisées, nous nous concentrons ici sur l'analyse des métriques non supervisées. La suite de cet article est organisée de la façon suivante : dans la section 2, la mesure de similarité non-supervisée est présentée, tandis que des mesures alternatives sont présentées dans la section 3. Les résultats sont présentés et discutés dans la section 4.

2 Mesure de similarité *soft*-cosine

2.1 Principe

Dans une approche classique en sacs de mots, les textes sont représentés par un vecteur de poids (e.g. TF-IDF) de taille N où N est la taille du lexique considéré. La mesure de similarité en cosinus entre ces 2 vecteurs est directement liée au nombre de mots en commun entre les 2 textes :

$$cos(X, Y) = \frac{X^t.Y}{\sqrt{X^t.X}\sqrt{Y^t.Y}} \text{ où } X^t.Y = \sum_{i=1}^{n} x_i y_i \quad (1)$$

Si il n'y a aucun mot en commun entre X et Y (i.e. aucun indice i pour lequel à la fois x_i et y_i sont

non nuls), la mesure vaut 0. Or les textes peuvent être sémantiquement similaires, même si ils n'ont aucun mot en commun, mais si les mots qu'ils contiennent sont eux-mêmes sémantiquement liés. C'est pourquoi nous proposons d'introduire les relations entre mots dans la mesure de similarité, en introduisant dans la formule du cosinus une matrice de relations :

$$cos(X,Y) = \frac{X^t.M.Y}{\sqrt{X^t.M.X}\sqrt{Y^t.M.Y}} \text{ où } X^t.M.Y = \sum_{i=1}^{n}\sum_{j=1}^{n} x_i m_{i,j} y_j \quad (2)$$

où M est une matrice dont l'élément $m_{i,j}$ traduit une relation entre le mot i et le mot j. Avec une telle métrique, la similarité entre 2 textes est non nulle dès que les textes ont des mots liés entre eux, même si ils n'ont aucun mot en commun. Introduire la matrice M également au dénominateur est nécessaire pour conserver le rôle de normalisation du dénominateur et une similarité réflexive égale à 1. On notera que si les mots n'ont de relations qu'avec eux-mêmes ($m_{i,i} = 1$ et $m_{i,j} = 0\ \forall i, j, i \neq j$), M devient la matrice identité, et on retrouve la mesure en cosinus classique.

Nous avons d'abord exploré cette mesure de similarité dans le contexte de la segmentation thématique de journaux télévisés (Bouchekif *et al.*, 2016) en utilisant les relations sémantiques entre mots pour améliorer le calcul de la cohésion sémantique entre passages consécutifs. Cette mesure avait été également proposée par d'autres chercheurs (e.g. (Sidorov *et al.*, 2014)), et dénommée *soft*-cosinus, où la matrice de relations était fondée par exemple sur une distance d'édition entre n-grams. Dans notre travail, nous explorons différents types de relations entre mots pour le calcul de M.

2.2 Relations sémantiques

Les plongements de mots, comme par exemple l'approche `word2vec` (Mikolov *et al.*, 2013) ont connu un immense succès ces dernières années. Ils permettent, entre autres, d'obtenir des relations sémantiques entre mots, simplement à partir d'une mesure de similarité (par exemple en cosinus) entre les vecteurs représentatifs des mots. Dans ce travail, 2 plongements sont calculés, avec l'outil `word2vec`, dans la configuration *cbow* : l'un à partir d'une version du Wikipedia en anglais, l'autre à partir du corpus non-annoté de 100 millions de mots disponible pour la campagne d'évaluation (questions et fils de discussions du forum Qatar Living). La dimension des vecteurs est fixée à 300 (les variantes de dimensions testées n'ont pas eu d'influence significative) et la fréquence minimale des mots à 50. A partir de ces plongements, M peut être calculée de différentes façons. Après avoir exploré différentes variantes, les meilleurs résultats ont été obtenus en considérant, pour traduire la relation entre deux mots w_i (représenté par v_i) et w_j (représenté par v_j) :

$$m_{i,j} = max(0, cosine(v_i, v_j))^2 \quad (3)$$

Le seuillage à 0 est motivé par l'observation que les cosinus négatifs entre mots sont difficiles à interpréter et ne semblent pas pertinents. Le passage au carré accentue la dynamique des relations sémantiques, en insistant sur les plus fortes et en diminuant les plus faibles.

2.3 Relations à partir de distances d'édition

Une similarité d'édition entre mots peut être calculée, à partir d'une distance d'édition comme celle de Levenshtein : elle permet de pallier certaines petites erreurs de typographie, fréquentes dans les textes produits par les réseaux sociaux. Nous définissons la matrice de similarité d'édition comme étant $m_{i,i} = 1$ et pour $i \neq j$:

$$m_{i,j} = \alpha * \left(1 - \frac{Levenshtein(w_i, w_j)}{max(||w_i||, ||w_j||)}\right)^{\beta} \quad (4)$$

où $||w||$ est le nombre de lettres du mot w, α est un facteur de pondération par rapport aux éléments diagonaux, et β un facteur pour amplifier la dynamique des scores. Les expériences sur le corpus $train$ et dev nous ont mené à fixer $\alpha = 1.8$ et $\beta = 5$.

3 Autres mesures

3.1 Mesure par appariement de mots

La spécificité de l'approche en soft-cosinus est de tenir compte des relations de tous les mots d'un texte par rapport à tous les mots d'un autre texte, ce qui contraste avec des approches d'alignement d'un texte par rapport à un autre, où chaque mot d'un texte est associé à un seul mot, au maximum, d'un autre texte. Ainsi, une mesure contrastive a été développée, dans une approche d'appariement entre les mots de chaque texte. Dans cette approche, chaque mot x_i de poids non nul du vecteur X est apparié à son plus proche voisin d'indice $vois(i)$ de poids non nul du vecteur Y :

$$vois(i) = argmax_{j, y_j \neq 0} m_{i,j}$$

En restant dans un paradigme inspiré du cosinus, une fois cet appariement réalisé, on propose la mesure de similarité de Y par rapport à X :

$$simvois(X, Y) = \frac{\sum_i x_i m_{i,vois(i)} y_{vois(i)}}{\sqrt{\sum_i x_i^2}\sqrt{\sum_j y_j^2}}$$

Le dénominateur assure que $simvois(X, X) = 1$ car le plus proche voisin d'un mot est toujours lui-même, et $m_{i,i} = 1$. Cette mesure est non-symétrique car la relation de plus proche voisin ne l'est pas. On la symétrise alors en calculant :

$$simvois_{sym}(X, Y) = 0.5 * (simvois(X, Y) + simvois(Y, X))$$

3.2 Mesure entre représentations moyennes des textes

A partir des plongements de mots, une façon très simple de représenter un texte consiste à calculer la moyenne, pondérée ou non, des plongements de mots présents dans le texte. Ensuite, la mesure de similarité entre textes est calculée par exemple par une mesure de similarité en cosinus entre les vecteurs représentatifs de chaque texte. Cette mesure sans pondération des mots a été très fréquemment utilisée par les participants des campagnes d'évaluation de similarité sémantique entre textes de SemEval. L'introduction d'une pondération permet des améliorations significatives de la représentation (e.g. (Arora *et al.*, 2017)), et l'approche de similarité cosinus entre vecteurs moyens pondérés constitue une base de départ robuste pour les représentations de texte. Nous choisissons ici de pondérer les mots par leur poids TF-IDF de leur représentation en sacs de mots, où x_i est le poids du terme i dans le texte X. Si $v_{i,k}$ est la k-ième composante du mot i dans l'espace des plongements de mots, la k-ième composante du vecteur représentatif $\tilde{X}$ du texte X est obtenue par :

$$\tilde{X}_k = \frac{1}{\sum_i x_i} \sum_i x_i v_{i,k}$$

La mesure $wavg - w2v(X, Y)$ consiste alors en une similarité en cosinus entre $\tilde{X}$ et $\tilde{Y}$.

4 Evaluation

4.1 Protocole expérimental

Nous présentons ici des évaluations de la tâche de similarité question-question (Task3-subtaskB de SemEval 2017). Etant donnée une nouvelle question, dite originale, il s'agit de trier par similarité sémantique décroissante les 10 questions, dites relatives, remontées par un moteur de recherche sur le forum Qatar Living. Une description détaillée du corpus et de la métrique d'évaluation se trouve dans (Nakov *et al.*, 2017). L'évaluation est faite avec la mesure MAP (Mean-Average-Precision), mesure standard en recherche d'information. Elle vaut 100% si, pour chaque question originale, toutes les questions relatives jugées pertinentes dans la vérité-terrain ont un score de similarité sémantique supérieur aux questions jugées non-pertinentes. On présente les résultats sur 3 corpus : dev (50 questions originales $\times$ 10 questions relatives), $test2016$ (70 questions originales $\times$ 10 questions relatives) et $test2017$ (88 questions originales $\times$ 10 questions relatives).

Il faut noter que le programme d'évaluation MAP utilisé dans la campagne est sensible au nombre de questions originales qui n'ont aucune question relative pertinente : le programme d'évaluation leur attribue une MAP de 0. Ainsi, une évaluation Oracle (où toutes les questions relatives labellisées pertinentes dans la vérité-terrain ont un score de similarité de 1, et toutes les questions relatives labellisées non-pertinentes dans la vérité-terrain ont un score de similarité de 0) ne donne pas un MAP de 100%, mais un MAP qui correspond au taux de questions originales qui ont au moins une question relative labellisée pertinente. Cette borne supérieure du MAP vaut 86.0% pour le dev, 88.6% pour le $test2016$ et seulement 67.0% pour le $test2017$. De plus, pour le $test2017$, le nombre moyen de questions relatives labellisées pertinentes par question originale (pour les questions originales ayant au moins une relative pertinente) n'est que de 2.7, tandis qu'il est de 3.7 pour $test2016$. C'est pourquoi le $test2017$ est un corpus plus difficile que le $test2016$.

4.2 Pré-traitements des données

Des pré-traitements très simples sont appliqués sur les textes : suppression de la casse, de la ponctuation, des mots-outils, remplacement des urls et des images par des termes génériques "_url_" et "_img_". En ce qui concerne les poids TF-IDF de la représentation sacs de mots, nous proposons une approche particulière, où les IDF sont estimés sur le gros corpus non-annoté du domaine du forum *Qatar Living*. Les "documents" considérés sont à la fois les titres et les paragraphes du forum (commentaires inclus).

4.3 Évaluation des mesures de similarité non-supervisées

Le tableau 1 présente les performances MAP des différentes mesures de similarités non-supervisées. Nous ne présentons que les résultats des similarités mesurées entre les textes $sujet + corps$ de la question originale et des questions relatives, configuration optimale pour la tâche. A titre de comparaison, la performance du moteur de recherche qui produit les 10 questions relatives, ainsi que

la performance du système qui a gagné la compétition SemEval2016 (Franco-Salvador *et al.*, 2016) sont également présentées.

similarités	dev	$test2016$	$test2017$
moteur de recherche	71.35	74.75	41.85
gagnant SemEval2016	-	77.33	-
mesures non supervisées			
baseline_token_cos	62.22	68.54	40.88
baseline_pp_cos	67.49	71.05	42.80
baseline_pp_cos_tfidf	69.41	75.53	44.37
cos_{Mrel} relations WP	72.25	77.11	45.38
cos_{Mrel} relations QL	**75.24**	**77.96**	45.27
cos_{Mlev} Levenshtein	70.02	76.34	46.10
$simvois_{sym}$ relations QL	74.40	77.47	45.67
wavg-word2vec sur QL	73.31	75.77	**46.99**
combinaison supervisée des mesures			
soumission SemEval2017	**77.30**	**79.77**	**47.22**

TABLE 1 – Performance MAP pour les mesures de similarité non-supervisées

Nous prenons comme point de départ des évaluations le système dit *baseline_token_cos*, tel que défini dans les évaluations de similarité sémantique de phrases de SemEval2015-Task2 (Agirre *et al.*, 2015) : c'est une simple similarité en cosinus entre sacs de $tokens$, sans pondération (un $token$ étant défini comme une séquence de caractères non espace entre 2 caractères d'espace, avec des poids valant 1 ou 0). Les performances obtenues avec *baseline_pp_cos*, qui est aussi un cosinus entre vecteurs binaires, mais sur les termes obtenus après pré-traitements, rappelle l'importance de pré-traitements appropriés. Ensuite, *baseline_pp_cos_tfidf* confirme l'influence d'une pondération pertinente des termes.

Les résultats suivants montrent une amélioration significative des performances lorsqu'on introduit une matrice M dans le calcul de la similarité en cosinus. Lorsqu'il s'agit de relations sémantiques, une différence significative est observée sur dev entre les relations estimées sur un corpus générique (WP : Wikipedia) et sur un corpus spécialisé du domaine (QL : Qatar Living). Cette différence est plus faible pour $test2016$ et même négative pour $test2017$. A l'inverse, l'utilisation de la matrice M basée sur une distance d'édition Levenshtein améliore le plus sur $test2017$ tandis que ces gains sont marginaux pour dev et $test2016$. L'approche $simvois_{sym}$ qui ne considère que le plus proche voisin d'un mot dans l'autre phrase donne également des performances intéressantes, et très proches de l'approche soft-cosine qui tient compte des relations entre tous les mots. Finalement, la mesure $wavg - w2v$ est moins performante que le *soft*-cosinus avec relations sémantiques sur dev et $test2016$ mais donne de bons résultats sur $test2017$. Les particularités de ce corpus, où les appariements pertinents sont largement minoritaires parmi les propositions du moteur de recherche, font que les résultats sont beaucoup moins contrastés. On notera que la mesure cos_{Mrel} avec relations sémantiques sur le corpus spécialisé aurait gagné la compétition de 2016. La combinaison des trois mesures, via une régression logistique optimisée sur le $train$ selon un processus décrit en (Charlet & Damnati, 2017), produit les meilleurs résultats quel que soit le corpus considéré, comme le montre la dernière ligne du tableau.

4.4 Expériences contrastives

Nous présentons dans cette section quelques expériences contrastives sur le corpus dev afin de mesurer plus précisément l'influence des différents éléments de la chaine, depuis les pré-traitements jusqu'à l'introduction d'une matrice de relation sémantique. Le tableau 2 présente l'influence des pré-traitements sur les performances. Dans l'approche simple qui consiste à évaluer un cosinus entre les sacs de mots, nous pouvons voir que la lemmatisation dégrade les performances lorsque les sacs de mots ne sont pas pondérés. En revanche la mise en œuvre d'une pondération appropriée (estimée sur le corpus QL lemmatisé) apporte un gain substantiel, qui est accru par le passage à la lemmatisation. Lorsque l'on introduit une matrice de relations, apprise sur le corpus QL également lemmatisé, on observe une amélioration importante des résultats par rapport au cosinus, mais non significative par rapport à l'approche non lemmatisée.

similarité	**lemme**	**TF-IDF**	**MAP on** dev
cosinus	-	-	67.49
cosinus	✓	-	65.46
cosinus	-	✓	69.41
cosinus	✓	✓	71.67
soft-cosinus	-	✓	75.24
soft-cosinus	✓	✓	75.85

TABLE 2 – Influence des pré-traitements sur les performances

5 Conclusion

Dans ce travail, nous avons exploré une extension de la mesure de similarité en cosinus entre sacs de mots, par l'introduction d'une matrice M de relations entre mots dans un *soft*-cosinus. Le calcul de M est non-supervisé et peut être issu de plongements de mots. Ce *soft*-cosinus donne de bonnes performances sur la tâche de similarité entre questions de la campagne SemEval2017-Task3. Une simple combinaison par régression logistique des différentes mesures de similarité non-supervisées s'est classée première de la compétition officielle. Nous souhaitons approfondir l'étude des similarités basées sur le *soft*-cosinus dans 2 directions : d'une part inclure de nouvelles relations dans M, par exemple à partir de rôles sémantiques, et d'autre part, étudier comment cette matrice M, efficacement initialisée de façon non-supervisée grâce aux plongements de mots, pourrait être ensuite apprise pour des tâches spécifiques.

Références

AGIRRE E., BANEA C., CARDIE C., CER D., DIAB M., GONZALEZ-AGIRRE A., GUO W., LOPEZ-GAZPIO I., MARITXALAR M., MIHALCEA R., RIGAU G., URIA L. & WIEBE J. (2015). Semeval-2015 task 2 : Semantic textual similarity, english, spanish and pilot on interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, p. 252–263, Denver, Colorado : Association for Computational Linguistics.

AGIRRE E., BANEA C., CER D. M., DIAB M. T., GONZALEZ-AGIRRE A., MIHALCEA R., RIGAU G. & WIEBE J. (2016). Semeval-2016 task 1 : Semantic textual similarity, monolingual and

cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016*, p. 497–511.

ARORA S., LIANG Y. & MA T. (2017). A simple but tough-to-beat baseline for sentence embeddings. In *Proceedings of ICLR 2017*, Toulon, France.

BOUCHEKIF A., DAMNATI G., CHARLET D., CAMELIN N. & ESTÈVE Y. (2016). Title assignment for automatic topic segments in TV broadcast news. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, March 20-25, 2016*, p. 6100–6104.

CHARLET D. & DAMNATI G. (2017). Simbow at semeval-2017 task 3 : Soft-cosine semantic similarity between questions for community question answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, SemEval '17, Vancouver, Canada : Association for Computational Linguistics.

FRANCO-SALVADOR M., KAR S., SOLORIO T. & ROSSO P. (2016). Uh-prhlt at semeval-2016 task 3 : Combining lexical and semantic-based features for community question answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, p. 814–821, San Diego, California : Association for Computational Linguistics.

MIHAYLOV T. & NAKOV P. (2016). Semanticz at semeval-2016 task 3 : Ranking relevant answers in community question answering using semantic similarity based on fine-tuned word embeddings. *Proceedings of SemEval*, p. 879–886.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv :1301.3781*.

NAKOV P., HOOGEVEEN D., MÀRQUEZ L., MOSCHITTI A., MUBARAK H., BALDWIN T. & VERSPOOR K. (2017). SemEval-2017 task 3 : Community question answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, SemEval '17, Vancouver, Canada : Association for Computational Linguistics.

NAKOV P., MÀRQUEZ L., MOSCHITTI A., MAGDY W., MUBARAK H., FREIHAT A. A., GLASS J. & RANDEREE B. (2016). Semeval-2016 task 3 : Community question answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016*, p. 525–545.

SIDOROV G., GELBUKH A. F., GÓMEZ-ADORNO H. & PINTO D. (2014). Soft similarity and soft cosine measure : Similarity of features in vector space model. *Computación y Sistemas*, **18**(3).

Adaptation au domaine pour l'analyse morpho-syntaxique

Éléonor Bartenlian Margot Lacour Matthieu Labeau Alexandre Allauzen
Guillaume Wisniewski François Yvon
LIMSI, CNRS, Univ. Paris-Sud, Université Paris-Saclay, 91 405 Orsay, France
prénom.nom@limsi.fr

RÉSUMÉ

Ce travail cherche à comprendre pourquoi les performances d'un analyseur morpho-syntaxiques chutent fortement lorsque celui-ci est utilisé sur des données hors domaine. Nous montrons à l'aide d'une expérience jouet que ce comportement peut être dû à un phénomène de masquage des caractéristiques lexicalisées par les caractéristiques non lexicalisées. Nous proposons plusieurs modèles essayant de réduire cet effet.

ABSTRACT

Domain Adaptation for PoS tagging

This work aims at understanding the performance drop observed when a PoS-tagger is applyed on out-domain data. We design a toy experiment showing that this observation may result from the fact that lexicalized features are masking non lexicalized features. We propose several improvements of history-based model aiming at reducing this phenomenon.

MOTS-CLÉS : Analyse morpho-syntaxique, adaptation au domaine, modèles statistiques, UGC.

KEYWORDS: PoS tagging, Domain Adaptation, UGC.

1 Introduction

L'étiquetage morpho-syntaxique *(PoS tagging)* est une tâche qui est généralement considérée comme résolue : à partir du moment où suffisamment de données du domaine considéré sont disponibles, les performances des méthodes statistiques sont telles que les seules erreurs de prédiction restantes résultent d'erreurs d'annotation ou de cas particulièrement ambigus (Manning, 2011). Pourtant dès que l'on change de domaine ou de type de documents, les performances chutent de manière spectaculaire (Martínez Alonso *et al.*, 2016). Les raisons généralement avancées pour expliquer cette chute sont l'augmentation du nombre de mots hors vocabulaire (HV) qui n'ont jamais été observés sur le corpus d'apprentissage et l'incapacité des modèles statistiques à généraliser les connaissances extraites du corpus d'apprentissage à ceux-ci (Foster, 2010; Seddah *et al.*, 2012).

L'objectif de ce travail est de fournir une première explication de cette observation : nous montrons par plusieurs expériences, qu'une raison possible peut être liée à un phénomène de *masquage* des caractéristiques non lexicalisées (robustes au changement de domaine mais avec un faible pouvoir prédictif) par les caractéristiques lexicalisées (permettant d'obtenir de très bonnes performances sur des données du même domaine, mais avec un faible pouvoir de généralisation). Nous proposons et évaluons également plusieurs méthodes cherchant à limiter cet effet.

caractéristiques lexicalisées	caractéristiques non lexicalisées
◇ mot à la i^{e} position	◇ trois dernières lettres des i^{e}, $(i \pm 1)^{\text{e}}$ mots
◇ mots aux positions $i-2$, $i-1$, $i+1$, $i+2$	◇ dernière lettre des i^{e} et $(i-1)^{\text{e}}$ mot
◇ conjonction du i^{e} mot et de la $(i-1)^{\text{e}}$ étiquette	◇ étiquettes des $(i-1)^{\text{e}}$ et $(i-2)^{\text{e}}$ mot
	◇ conjonction des deux étiquettes précédentes

TABLE 1 – Patron des caractéristiques utilisées dans notre analyseur morpho-syntaxique. Nous distinguons 6 patrons lexicalisés (reposant sur la connaissance du token) de 8 patrons non lexicalisées.

Le reste de cet article est organisé de la manière suivante : nous commencerons par mettre en évidence l'impact du changement de domaine sur les performances d'un analyser morpho-syntaxique (§2) ; nous chercherons ensuite à expliquer pourquoi les analyseurs morpho-syntaxiques ne sont pas robustes à un changement de domaine (§3) avant de proposer et d'évaluer trois approches visant à résoudre ces problèmes (§4).

2 Évaluation hors domaine d'un analyseur morpho-syntaxique

Nos expériences utilisent un analyseur morpho-syntaxique à base d'historique (Black *et al.*, 1992; Tsuruoka *et al.*, 2011). Dans ces modèles, la prédiction d'une séquence d'étiquettes morpho-syntaxiques est modélisée sous la forme d'une suite de problèmes de décision, consistant chacun à prédire l'étiquette d'une observation. Chaque décision est prise par un classifieur multi-classe utilisant comme descripteurs des informations extraites de la structure d'entrée, ainsi que les décisions prises antérieurement. Nous utilisons, dans toutes nos expériences, un perceptron moyenné comme classifieur multi-classe (Collins, 2002). Les caractéristiques utilisées sont décrites dans le Tableau 1 [1]. Ce sont des caractéristiques simples et indépendantes de la langue qui sont, à notre connaissance, utilisées dans tous les étiqueteurs morpho-syntaxiques. Une description détaillée de ce modèle est présentée dans (Wisniewski *et al.*, 2014b,a).

Pour mesurer l'impact du changement de domaine sur les performances de l'analyse morpho-syntaxique, nous avons mené une série d'expériences sur le français. Les paramètres de l'analyseur sont estimés sur le corpus d'apprentissage du projet UD [2] qui a pour objectif de fournir, pour de nombreuses langues, des corpus annotés en PoS, morphologie et dépendances dans un schéma commun. Ses performances sont évaluées sur un corpus du domaine (la partie test du corpus UD) ainsi que sur quatre corpus hors domaine :

— deux corpus de contenu généré par des utilisateurs (*User Generated Content*, UGC) collectés sur les forums de discussion du site MARMITON [3] et du jeu MINECRAFT [4]. La collecte et l'analyse de ces corpus sont présentées dans (Martínez Alonso *et al.*, 2016).

— deux corpus de tweets, un parlant de football, FOOT et l'autre de catastrophe naturelle, NATDIS. Ces deux corpus sont décrits plus précisément dans (Adda *et al.*, 2017).

Ces corpus diffèrent du corpus d'apprentissage sur plusieurs points : ils sont composés de phrases

1. Les entrées sont également transformées : tous les nombres, les URL et les mentions sont remplacées par un même token
2. http://universaldependencies.org/
3. http://www.marmitton.org
4. http://www.minecraft.com

MARMITON	① Que faire de facile avec un reste de boeuf pour une fondue ? ? ? ② Qu' avez -vous dans le vôtre ?
MINECRAFT	③ tu fais kit artisan ④ Apr'es tu veux alli sale noob de Azarid ...
FOOT	⑤ Ça va faire un petit Juventus - Barça , excellent aha 🙂 ⑥ Ça c' est de l' attaque ! La BBC 😜 #HalaMadrid #RealMadrid http ://t.co/iCknoGm9mF
NATDIS	⑦ A Lire #Liban Une vague de chaleur exceptionnelle frappe le Liban , 40 degrés à midi http ://t.co/moDWsrreaT ⑧ Au @Radiojournal_RC 7h Un nouveau tremblement de terre sème la terreur au Népal http ://t.co/UPtZpytQ0q

TABLE 2 – Extrait des différents corpus hors domaine considérés.

corpus	#phrases	# mots	% mots HV	% erreur	% erreur HV	% erreur ambigu
UD	298	6 829	5,8%	4,4%	18,0%	5,6%
FOOT	743	13 985	26,2%	16,0%	37,2%	10,4%
NATDIS	622	1 233	23,4%	8,8%	18,3%	6,9%
MARMITON	285	2 058	18,5%	13,2%	35,7%	11,1%
MINECRAFT	234	900	38,7%	40,2%	69,6%	23,9%

TABLE 3 – Caractéristiques des corpus considérés pour l'évaluation et taux d'erreurs obtenus selon le type de mots.

plus courtes, comportant de nombreuses idiosyncrasies et fautes d'orthographe et diffèrent par leur style, leur genre, leur registre et les thématiques qu'ils abordent. La Figure 2 montre des extraits de ces corpus. Les étiquettes de ces corpus ont été converties manuellement vers le schéma de l'UD [5].

Les performances obtenues par notre modèle sur un corpus du domaine (cf. Tableau 3) sont proches de celles de l'état de l'art ((Straka *et al.*, 2016) obtient, par exemple, un taux d'erreur de 2,92%), bien que le modèle d'apprentissage soit extrêmement simple et qu'aucun effort particulier n'ait porté sur le choix et la conception des caractéristiques [6].

Comme on pouvait s'y attendre, changer de domaine a un fort impact négatif sur la qualité de la prédiction : l'analyseur réalise entre 2 et 4 fois plus d'erreurs sur les corpus hors domaine que sur le test du domaine. Pour mieux comprendre l'impact du changement de domaine, nous proposons d'évaluer, les performances de l'analyseur sur différents types de mots :

— les mots hors vocabulaire (HV), c'est-à-dire ceux qui n'apparaissent jamais dans le corpus d'apprentissage ;
— les mots ambigus, c'est-à-dire ceux qui *i)* apparaissent plus d'une fois dans le corpus d'apprentissage et *ii)* n'apparaissent pas toujours avec la même étiquette.

Les résultats de cette évaluation sont donnés dans le Tableau 3. Sans surprise, le nombre de mots hors

5. Ces corpus ont été étiquetés en suivant les conventions du FTB qui définit un jeu d'étiquettes plus riche que celui de l'UD. La conversion est donc déterministe et ne pose pas de problèmes particuliers.

6. Comme souligné dans de nombreux travaux (Manning, 2011), cette évaluation est toutefois optimiste puisqu'elle prend en compte des occurrences dont l'étiquette peut être prédite de manière déterministe (par exemple les symboles de ponctuation).

corpus	#exemples train/test/HV	Δ erreur train	Δ erreur test	Δ erreur HV	# mises à jour lex/delex
UD	5 917/638/88	-18,84%	-13.26%	1,36%	10 476 / 55 364
SPMRL	10 440/1 769/145	-20,06%	-17,00%	0,14%	6 674 / 81 554

TABLE 4 – Résultats de notre expérience jouet : nous reportons, pour les différents corpus, la différence entre la précision du modèle lexicalisé et celle du modèle non lexicalisé.

vocabulaire augmente fortement dès que l'on change de domaine. On observe également, de manière plus surprenante, que globalement les étiquettes des mots HV et des mots ambigus sont nettement moins bien prédites sur le corpus hors domaine que sur le corpus du domaine. Cette observation, similaire à celle faite par (Barrett *et al.*, 2007), pourrait indiquer que la structure des phrases (qui conditionne les mots et étiquettes au voisinage d'un mot HV) change d'un domaine à l'autre et ne permet plus de désambiguïser les mots HV.

3 L'effet de masquage des caractéristiques lexicalisées

Il est surprenant que les performances sur les mots HV soient beaucoup plus faibles que celles sur les mots observés en apprentissage. Il est bien connu que, pour l'analyse morpho-syntaxique, la caractéristique la plus utile est la connaissance du mot, puisque la plupart des mots ne sont pas ou peu ambigus. Ainsi, sur le corpus UD, un analyseur prédisant l'étiquette la plus fréquente associée à chaque mot ne fait que 11.0% d'erreur, les mots HV étant considérés comme des erreurs. Toutefois, les modèles de l'état de l'art comportent tous des caractéristiques décrivant la structure (étiquettes voisines) et une approximation plus ou moins grossière de la « morphologie » des mots (à minima les suffixes et préfixes), afin d'être capable de *généraliser* les observations faites sur le corpus d'apprentissage aux mots HV. Par exemple, un mot inconnu se trouvant entre un déterminant *(une)* et un nom *(voiture)* sera très probablement un adjectif [7] ; les caractéristiques de forme aident à prédire qu'un mot se terminant en *-ment* sera très certainement un adverbe.

Les résultats présentés dans le Tableau 3 montrent que ces caractéristiques n'arrivent pas à jouer leur rôle. Nous faisons l'hypothèse que cela est la conséquence du très fort pouvoir prédictif des caractéristiques *lexicalisées* : comme expliqué dans le paragraphe précédent, connaître le mot dont on cherche à prédire l'étiquette permet de déterminer celle-ci de manière quasi déterministe et, l'on peut supposer qu'il suffit d'observer une fois un mot pour « retenir » l'étiquette qui lui est associée. Mais, une fois que l'étiquette d'un mot est correctement prédite (ce qui est, à priori, possible à partir d'une unique caractéristique lexicale), celui-ci ne participe plus à l'apprentissage puisque dans les modèles à base de correction d'erreur, comme le perceptron, les paramètres ne sont mis-à-jour lorsqu'une prédiction est erronée et dans les modèles maximisant la vraisemblance (log-loss), comme les réseaux neuronaux ou les CRF, la norme du gradient est proportionnelle à l'erreur. L'estimation du poids des autres caractéristiques sera alors moins fiable, puisqu'étant moins souvent mis à jour, leur variance sera plus grande. Tout se passe comme si les caractéristiques lexicales *masquaient* les autres. Une observation similaire a été faite par (Klein & Manning, 2002) qui en faisant le lien avec le phénomène *explaining-away* (Pearl, 1988), montre comment certaines caractéristiques peuvent être ignorées lors de l'apprentissage. Les expériences menées par (Pécheux *et al.*, 2015) mettent également en évidence

7. Cela pourrait également être un nom comme dans « un chien voiture »

ce phénomène en étudiants l'impact de contraintes appliquées lors de l'apprentissage d'un analyseur morpho-syntaxique.

Pour confirmer cette hypothèse, nous proposons l'expérience jouet suivante : est-il possible d'identifier si un mot se terminant en `-ent` est un verbe (conjugué à la troisième personne du pluriel) ou un adverbe, à partir de la connaissance du mot et de l'étiquette du mot précédent ? En réduisant le nombre de caractéristiques et d'étiquettes, nous espérons limiter les interactions entre caractéristiques et obtenir des modèles plus facilement interprétables.

Plus précisément, nous extrayons des corpus UD et SPMRL (Seddah *et al.*, 2013) français tous les mots se terminant en `-ent` ainsi que l'étiquette (oracle) du mot précédent. Ces mots sont des verbes, des noms, des adjectifs ou des adverbes. Nous considérons ensuite deux perceptrons moyennés essayant d'identifier ces classes à partir de deux jeux de caractéristiques différents : le premier considère des caractéristiques *lexicalisées* (l'étiquette du mot précédent et le mot courant), le second uniquement des caractéristiques non lexicalisées (l'étiquette du mot précédent).

Les résultats, présentés Table 4, confirment notre hypothèse. Les caractéristiques lexicalisées ont un fort pouvoir prédictif (le taux d'erreur global du modèle lexicalisé est nettement plus faible que celui du modèle non lexicalisé), mais elles ne permettent pas de généraliser les connaissances apprises aux mots qui n'ont pas été vus en apprentissage : le modèle non lexicalisé obtient de meilleures performances que le modèle lexicalisé si l'on restreint l'évaluation aux mots qui n'ont pas été vus en apprentissage. La comparaison du nombre de mises à jour effectuées pour les deux modèles semble indiquer que cette observation est bien le résultat du phénomène de *masquage* décrit précédemment.

4 Méthode d'adaptation au domaine

Nous proposons dans cette section deux améliorations de notre analyseur morpho-syntaxique visant à améliorer les performances de celui-ci sur des données hors domaine.

Méthodes La première méthode, SPEC, repose sur une spécialisation du perceptron utilisé dans notre analyseur morpho-syntaxique. Elle consiste simplement à utiliser deux perceptrons différents pour prendre les décisions successives lors de l'inférence : un premier classifieur, utilisant l'ensemble des caractéristiques de la Table 1 sera utilisé pour prédire les étiquettes des mots observés en apprentissage ; un second, n'utilisant que les caractéristiques non lexicalisées prédira les étiquettes des mots HV. Ces deux classifieurs sont appris successivement sur l'ensemble des données d'apprentissage. Nous espérons par cette spécialisation des modèles mieux estimer les paramètres liées aux caractéristiques de structure et de morphologie que nous supposons plus robuste au changement de domaine que les caractéristiques reposant sur les mots, tout en continuant à exploiter les caractéristiques lexicalisées pour les mots vus en apprentissage.

La seconde méthode, DROPOUT-T, est directement inspirée de la méthode proposée par (Kouw *et al.*, 2016) : partant du constat que les différences entre domaines se traduisent souvent par des changements dans la probabilité d'observation des caractéristiques (typiquement une caractéristique très pertinente n'apparaît qu'en apprentissage et pas en test), les auteurs proposent d'apprendre un modèle plus robuste en *bruitant* les données d'apprentissage. C'est une approche classique en apprentissage statistique (Van Der Maaten *et al.*, 2013; Wang *et al.*, 2013) permettant de réduire le sur-apprentissage.

	UD	FOOT	NATDIS	MARMITON	MINECRAFT
BASELINE	4,4% 18,0%/5,5%	16,0% 37,2%/10,4%	8,8% 18,3%/6,9%	13,2% 35,7%/11,1%	40,2% 69,6%/23,9%
SPEC	4,3% 21,0%/5,3%	16,3% 37,9%/10,9%	9,5% 22,8%/6,4%	13,9% 37,5%/11,5%	40,1% 72,4%/22,6%
DROPOUT-T	4,9% 20,5%/6,1%	16,2% 37,5%/11,0%	8,7% 18,3%/7,0%	12,8% 35,7%/10,6%	38,3% 68,7%/26,0%
DROPOUT-A	4,3% 17,7%/5,6%	16,0% 37,1%/10,6%	8,6% 19,2%/6,6%	12,4% 32,0%/10,9%	37,6% 68,7%/23,4%
NEURONALE	4,2%	15,0%	8,9%	12,3%	41,8%

TABLE 5 – Précision obtenue par les différentes méthodes d'adaptation au domaine proposées. Les taux d'erreur sur les mots HV et les mots ambigus sont détaillés.

Plus précisément, (Kouw *et al.*, 2016) propose de supprimer aléatoirement certaines caractéristiques lors de l'apprentissage afin d'encourager le modèle à utiliser les autres caractéristiques et à estimer « correctement » leur poids. En pratique, chaque caractéristique f d'une observation est supprimée avec une probabilité :

$$\max\left\{0, 1 - \frac{p_{test}}{p_{train}}\right\} \quad (1)$$

où p_{test} est la probabilité d'observer la caractéristique f sur les données de test et p_{train} la probabilité d'observer f sur les données d'apprentissage. Intuitivement, plus une caractéristique est typique des observations du domaine d'apprentissage ($p_{train} \gg p_{test}$), plus elle aura des chances d'être souvent supprimée ; à la limite, une caractéristique n'apparaissant jamais dans les données de test devrait être supprimée avec probabilité 1. Pour mesurer l'intérêt de cette approche, nous considérons également la méthode DROP-A dans laquelle la probabilité de supprimer une caractéristique est arbitrairement fixée à 0,3.

La troisième méthode, NEURONAL, repose sur un étiqueteur neuronal. Cette approche permet de s'affranchir des problèmes liés aux mots HV en modélisant directement la séquence de caractères composant le mot. Elle construit la représentation vectorielle continue d'un mot à partir de la séquence de ses caractères, à l'aide d'une couche de convolution. La prédiction de l'étiquette est réalisée avec une couche cachée classique, qui prend en entrée les représentations vectorielle du mot considéré et de ses voisins. Une description plus détaillée de cette approche est faite dans (Labeau *et al.*, 2015).

Évaluation expérimentale Les résultats de ces deux approches sont décrits dans le tableau 5. Aucun des méthodes proposées ne parvient à réduire significativement l'impact du changement de domaine. De manière plus surprenante, sur plusieurs corpus, les résultats obtenus par les approches proposées sont très proches de ceux obtenus par un simple perceptron. Nous pensons que cette observations peut-être liées à deux raisons : *i)* des différences systématiques dans la préparation (p. ex. la segmentation en mots) et l'annotation des données, une hypothèse également étudiée par (Rabary *et al.*, 2015) *ii)* une mauvaise compréhension ou modélisation des caractéristiques robustes au changement de domaine (p. ex. est-ce que la structure des phrases est conservée lors du changement de domaine). Nous menons actuellement des expériences pour mettre en évidence l'impact de ces deux raisons.

Remerciements

Ces travaux ont été en partie financés par l'Agence Nationale de la Recherche (projet PARSITI, ANR-16-CE33-0021). Nous remercions les relecteurs pour leurs commentaires et suggestions.

Références

ADDA G., CREGO J.-M., LACROIX O., LI G., MOSTEFA D., VLADIMIR P., WISNIEWSKI G. & FRANÇOIS Y. (2017). *A multilingual corpus of tweets for domain adaptation*. Research report, LIMSI-CNRS ; ELDA ; Systran.

BARRETT L., GREENBERG D. F. & SCHWARTZ M. (2007). A syntactic feature counting method for selecting machine translation training corpora. *Language and Computers : Studies in Practical Linguistics*, **60**, 1–19.

BLACK E., JELINEK F., LAFFERTY J., MAGERMAN D. M., MERCER R. & ROUKOS S. (1992). Towards history-based grammars : Using richer models for probabilistic parsing. In *Proceedings of the Workshop on Speech and Natural Language*, HLT'91, p. 134–139, Stroudsburg, PA, USA : Association for Computational Linguistics.

COLLINS M. (2002). Discriminative training methods for hidden markov models : Theory and experiments with perceptron algorithms. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, p. 1–8 : Association for Computational Linguistics.

FOSTER J. (2010). "cba to check the spelling" : Investigating parser performance on discussion forum posts. In *Human Language Technologies : The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, p. 381–384, Los Angeles, California : Association for Computational Linguistics.

KLEIN D. & MANNING C. D. (2002). Conditional structure versus conditional estimation in nlp models. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, p. 9–16, Stroudsburg, PA, USA : Association for Computational Linguistics.

KOUW W. M., VAN DER MAATEN L. J., KRIJTHE J. H. & LOOG M. (2016). Feature-level domain adaptation. *Journal of Machine Learning Research*, **17**(171), 1–32.

LABEAU M., LÖSER K. & ALLAUZEN A. (2015). Non-lexical neural architecture for fine-grained pos tagging. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, p. 232–237, Lisbon, Portugal : Association for Computational Linguistics.

MANNING C. D. (2011). Part-of-speech tagging from 97% to 100% : Is it time for some linguistics ? In *Proceedings of the Computational Linguistics and Intelligent Text Processing, 12th International Conference*, CICLing 2011, p. 171–189 : Springer.

MARTÍNEZ ALONSO H., SEDDAH D. & SAGOT B. (2016). From noisy questions to minecraft texts : Annotation challenges in extreme syntax scenario. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, p. 13–23, Osaka, Japan : The COLING 2016 Organizing Committee.

PEARL J. (1988). *Probabilistic Reasoning in Intelligent Systems : Networks of Plausible Inference*. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc.

PÉCHEUX N., ALLAUZEN A., LAVERGNE T., WISNIEWSKI G. & YVON F. (2015). Oublier ce qu'on sait, pour mieux apprendre ce qu'on ne sait pas : une étude sur les contraintes de type dans les modèles crf. In *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles*, p. 37–48, Caen, France : Association pour le Traitement Automatique des Langues.

RABARY C., LAVERGNE T. & NÉVÉOL A. (2015). Etiquetage morpho-syntaxique en domaine de spécialité : le domaine médical. In *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles*, p. 508–514, Caen, France : Association pour le Traitement Automatique des Langues.

SEDDAH D., SAGOT B., CANDITO M., MOUILLERON V. & COMBET V. (2012). The French Social Media Bank : a treebank of noisy user generated content. In *Proceedings of COLING 2012*, p. 2441–2458, Mumbai, India : The COLING 2012 Organizing Committee.

SEDDAH D., TSARFATY R., KÜBLER S., CANDITO M., CHOI J. D., FARKAS R., FOSTER J., GOENAGA I., GOJENOLA GALLETEBEITIA K., GOLDBERG Y., GREEN S., HABASH N., KUHLMANN M., MAIER W., NIVRE J., PRZEPIÓRKOWSKI A., ROTH R., SEEKER W., VERSLEY Y., VINCZE V., WOLIŃSKI M., WRÓBLEWSKA A. & DE LA CLERGERIE E. V. (2013). Overview of the SPMRL 2013 shared task : A cross-framework evaluation of parsing morphologically rich languages. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, p. 146–182, Seattle, Washington, USA : Association for Computational Linguistics.

STRAKA M., HAJIČ J. & STRAKOVÁ J. (2016). UDPipe : trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Paris, France : European Language Resources Association (ELRA).

TSURUOKA Y., MIYAO Y. & KAZAMA J. (2011). Learning with lookahead : Can history-based models rival globally optimized models ? In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, CoNLL'11, p. 238–246, Portland, Oregon, USA : Association for Computational Linguistics.

VAN DER MAATEN L., CHEN M., TYREE S. & WEINBERGER K. Q. (2013). Learning with marginalized corrupted features. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, p. I–410–I–418 : JMLR.org.

WANG S., WANG M., WAGER S., LIANG P. & MANNING C. D. (2013). Feature noising for log-linear structured prediction. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, p. 1170–1179, Seattle, Washington, USA : Association for Computational Linguistics.

WISNIEWSKI G., PÉCHEUX N., GAHBICHE-BRAHAM S. & YVON F. (2014a). Cross-lingual part-of-speech tagging through ambiguous learning. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 1779–1785, Doha, Qatar : Association for Computational Linguistics.

WISNIEWSKI G., PÉCHEUX N., KNYAZEVA E., ALLAUZEN A. & YVON F. (2014b). Cross-lingual pos tagging through ambiguous learning : First experiments (apprentissage partiellement supervisé d'un étiqueteur morpho-syntaxique par transfert cross-lingue) [in french]. In *Proceedings of TALN 2014 (Volume 1 : Long Papers)*, p. 173–183, Marseille, France : Association pour le Traitement Automatique des Langues.

Représentation vectorielle de sens pour la désambiguïsation lexicale à base de connaissances

Loïc Vial, Benjamin Lecouteux, Didier Schwab
GETALP – LIG – Univ. Grenoble Alpes
{loic.vial, benjamin.lecouteux, didier.schwab}
@univ-grenoble-alpes.fr

RÉSUMÉ

Dans cet article, nous proposons une nouvelle méthode pour représenter sous forme vectorielle les sens d'un dictionnaire. Nous utilisons les termes employés dans leur définition en les projetant dans un espace vectoriel, puis en additionnant les vecteurs résultants, avec des pondérations dépendantes de leur partie du discours et de leur fréquence. Le vecteur de sens résultant est alors utilisé pour trouver des sens reliés, permettant de créer un réseau lexical de manière automatique. Le réseau obtenu est ensuite évalué par rapport au réseau lexical de WordNet, construit manuellement. Pour cela nous comparons l'impact des différents réseaux sur un système de désambiguïsation lexicale basé sur la mesure de Lesk. L'avantage de notre méthode est qu'elle peut être appliquée à n'importe quelle langue ne possédant pas un réseau lexical comme celui de WordNet. Les résultats montrent que notre réseau automatiquement généré permet d'améliorer le score du système de base, atteignant quasiment la qualité du réseau de WordNet.

ABSTRACT

Sense Embeddings in Knowledge-Based Word Sense Disambiguation

In this paper, we develop a new way of creating sense vectors for any dictionary, by using an existing word embeddings model, and summing the vectors of the terms inside a sense's definition, weighted in function of their part of speech and their frequency. These vectors are then used for finding the closest senses to any other sense, thus creating a semantic network of related concepts, automatically generated. This network is hence evaluated against the existing semantic network found in WordNet, by comparing its contribution to a knowledge-based method for word sense disambiguation. This method can be applied to any other language which lacks such semantic network, as the creation of word vectors is totally unsupervised, and the creation of sense vectors only needs a traditional dictionary. The results show that our generated semantic network improves greatly the WSD system, almost as much as the manually created one.

MOTS-CLÉS : Représentation vectorielle de sens, Désambiguïsation Lexicale.

KEYWORDS: Sense Embeddings, Word Sense Disambiguation.

1 Introduction

Dans le Traitement Automatique des Langues (TAL), la Désambiguïsation Lexicale (DL) consiste à attribuer le sens le plus probable à un mot donné dans un document, à partir d'un inventaire prédéfini de sens. Les méthodes état de l'art sont souvent basées sur des approches supervisées, comme l'évoque Navigli (2009), où les modèles sont appris sur de grands corpus de mots annotés. Les modèles appris sont alors utilisés pour annoter les mots avec un sens dans un contexte donné. *A contrario*, les méthodes non supervisées ont l'avantage de nécessiter beaucoup moins de ressources pour fonctionner : elles n'utilisent pas de corpus préalablement annoté pour l'apprentissage. Il en découle en général une meilleure couverture sur les données de test, car elles sont moins tributaires des sens observés lors de l'apprentissage. Un autre avantage des méthodes non supervisées est qu'elles sont beaucoup plus généralisables et ne sont pas limitées à des langues très dotées telles que l'anglais. Cet aspect est important à relever, car l'annotation de corpus en sens est extrêmement lourde : peu de langues disposent de ce type de ressource. [1]

Pour ces raisons, nos travaux s'articulent autour des méthodes non supervisées et présentent une approche originale améliorant les performances existantes, basée sur des vecteurs de sens.

Les méthodes récentes et état de l'art de représentation vectorielle de mots, telles que Mikolov *et al.* (2013) (Word2Vec), Pennington *et al.* (2014) (GloVe) ou Levy & Goldberg (2014) (vecteurs basés sur des interdépendances), ont montré des gains intéressants dans de nombreuses tâches du TAL telles que la traduction automatique, le calcul de similarité entre mots ou encore en désambiguïsation lexicale, où ils ont été employés dans des méthodes récentes telles que Chen *et al.* (2014) ou Yuan *et al.* (2016).

Dans cet article nous présentons une méthode de représentation vectorielle pour des sens : où des vecteurs vont représenter les sens présents dans la base lexicale *Princeton WordNet* (Miller, 1995). La méthodologie utilisée pour créer ces vecteurs est décrite dans la section 2. Ces vecteurs de sens sont ensuite évalués en comparant leurs performances en tant que réseau lexical appliqué à un système de DL à base de connaissances dans la section 3.1.

Ces vecteurs de sens sont utilisés pour étendre les définitions des sens présents dans *WordNet* en concaténant les définitions des sens proches, d'une manière similaire au Lesk étendu de Banerjee & Pedersen (2002), sauf qu'au lieu de considérer que deux sens sont proches car ils partagent une relation syntaxique ou sémantique dans le réseau *WordNet*, nous considérons que deux sens sont proches en fonction de leur similarité cosinus. Dans la partie relative aux expérimentations nous explorons comment exploiter au mieux cette similarité.

2 Création de vecteurs de sens

Notre modèle pour représenter sous forme vectorielle les sens d'un dictionnaire repose sur leur définition, ainsi qu'un modèle pré-existant de vecteurs de mots. Dans la pratique, la méthode est relativement similaire à celle présentée par Ferrero *et al.* (2017) qui crée des vecteurs de phrases pour la détection de plagiat inter-lingue. Nos vecteurs de sens sont calculés comme la somme normée de tous les vecteurs des termes présents dans la définition du sens donné. Ces vecteurs sont pondérés

1. Moins de dix langues possèdent au moins un corpus annoté en sens via WordNet, listées sur `http://globalwordnet.org/wordnet-annotated-corpora/`, en date du 16 avril 2017.

en fonction de leur partie du discours (*Part Of Speech*, ou POS) : nom, verbe, adjectif ou adverbe, et également pondérés par leur *Inverse Document Frequency* (ou IDF) : l'inverse de leur nombre d'occurrence dans le dictionnaire entier. Plus formellement nous avons :

— $D(S) = \{w_0, w_1, w_2, \ldots, w_n\}$ la définition du sens S dans le dictionnaire
— $pos(w_n) = \{n, v, a, r\}$ la partie du discours du terme w_n (nom, verbe, adjectif, ou adverbe)
— $weight(pos)$ le poids associé à une partie du discours
— $idf(w_n)$ la valeur IDF de w_n

La définition du vecteur du sens S, notée $\phi(S)$ correspond à :

$$\phi(S) = \sum_{i=0}^{n} (\phi(w_n) \times weight(pos(w_n)) \times idf(w_n))$$

$\phi(S)$ est ensuite normé afin d'avoir la même taille que les vecteurs de mots (qui sont eux aussi normés). Les poids attribués aux POS sont les même que ceux utilisés par Ferrero *et al.* (2017).

Nous avons créé cinq modèles de vecteurs de sens, tous basés sur le vocabulaire et les définitions de *WordNet* 3.0, mais exploitant des modèles de vecteurs de mots différents. Les cinq modèles utilisés sont :

1. Un modèle pré-entraîné et proposés par Mikolov *et al.* (2013) [2]. Ce modèle a été entraîné sur environ 100 milliards de mots issus du corpus de nouvelles de Google. La taille du vocabulaire est d'environ de 3 millions de mots et les vecteurs ont une dimension de 300
2. Un modèle de Pennington *et al.* (2014)'s GloVe [3], entraîné sur 42 milliards de mots issus de *Common Crawl*. Le vocabulaire est de 2 millions de mots et les vecteurs ont une taille de 300.
3. Un modèle de Levy & Goldberg (2014) [4]. L'apprentissage a été effectué sur Wikipedia, le vocabulaire est de 175 000 mots et la taille des vecteurs 300.
4. Le meilleur des modèles de prédiction de Baroni *et al.* (2014) [5]. La taille du vocabulaire est de 300 000 et la taille des vecteurs est de 400.
5. Et finalement, le meilleur modèle à base de comptage également créé par Baroni *et al.* (2014)[5] de dimension 500 et une taille de vocabulaire identique au précédent modèle.

Dans la section expérimentale, nous comparons les performances de chacun de ces modèles dans une tâche de désambiguïsation lexicale en anglais. *WordNet* 3.0 est composé de 206 941 sens, et nous avons donc créé un vecteur pour chacun d'eux. Tous nos modèles de vecteurs de sens sont rendus publics sur la page de l'article, afin de reproduire les expériences, à l'adresse suivante : `https://github.com/getalp/WSD-TALN2017-Vialetal`.

3 Intégration à un système de désambiguïsation lexicale

Les vecteurs de sens ainsi générés sont évalués sur une tâche de désambiguïsation lexicale (DL). L'idée est d'exploiter le modèle en lieu et place d'un réseau lexical. Pour cela nous comparons les

2. https://code.google.com/archive/p/word2vec/
3. https://nlp.stanford.edu/projects/glove/
4. https://levyomer.wordpress.com/2014/04/25/dependency-based-word-embeddings/
5. http://clic.cimec.unitn.it/composes/semantic-vectors.html

sens entre eux via la distance cosinus, et selon un seuil δ déterminé empiriquement, les sens les plus proches sont sélectionnés. Les sens sélectionnés vont alors contribuer à l'algorithme de DL au niveau de sa mesure de similarité locale, dans un mode de fonctionnement similaire à celui présenté par Banerjee & Pedersen (2002) à la différence qu'eux exploitent le réseau lexical de *WordNet*.

3.1 Algorithme de désambiguisation

Le système de DL que nous proposons est composé de deux éléments : un algorithme local qui calcule un score de similarité pour une paire de sens, et un algorithme global qui va chercher la meilleure combinaison de sens à l'échelle du document, en utilisant l'algorithme local.

L'algorithme global est basé sur une heuristique permettant de ne pas explorer exhaustivement l'ensemble des combinaisons possibles, ce qui engendrerait des coûts calculatoires trop élevés.

L'heuristique utilisée dans notre système exploite un algorithme de recherche à base de coucous, comme celui présenté dans Vial *et al.* (2017). Les auteurs montrent que cet algorithme global est actuellement un des plus efficace connu pour la tâche de désambiguïsation.

Notre implémentation réutilise les meilleurs paramètres présentés dans Vial *et al.* (2017) (un seul coucou, le *Levy location* à 5 et le *Levy scale* à 0,5).

Dans l'ensemble de nos expériences nous avons utilisé un grand nombre d'itérations (300 000) afin de réduire l'effet aléatoire d'une expérience à une autre. De plus, nous exécutons en moyenne une dizaine d'exécutions pour en extraire un score moyen. De cette manière nous assurons un écart-type inférieur à 0,1 : les résultats sont ainsi stables et reproductibles.

L'algorithme local est l'élément central de notre système et c'est pour l'amélioration de ce dernier qu'est utilisé le réseau lexical. Comme système étalon nous utilisons une mesure de Lesk standard. Cet algorithme retourne, comme mesure de similarité, le nombre de mots communs à deux définitions de sens. Formellement, si nous notons $D(S) = \{w_1, w_2, \ldots, w_n\}$ la définition de S, alors la mesure de Lesk entre deux sens S_1 et S_2 notée $Lesk(S_1, S_2)$ est la suivante :

$$Lesk(S_1, S_2) = |D(S_1) \cap D(S_2)|$$

Comme second système étalon et dans l'objectif de mesurer la qualité de notre réseau lexical généré automatiquement, nous utilisons la mesure de Lesk étendue de Banerjee & Pedersen (2002).

Cette mesure considère à la fois la définition du sens cible mais également les définitions de tous les sens qui ont une relation avec le sens cible (hyponyme, etc.). Si l'on note $rel(S)$ l'ensemble des sens reliés à S à travers un lien explicite dans *WordNet*, alors la mesure de Lesk étendue entre les sens S_1 et S_2 notée $ExtLesk(S_1, S_2)$ est la suivante :

$$ExtLesk(S_1, S_2) = \left|(D(S_1) \cup_{r \in rel(S_1)} D(r)) \cap (D(S_2) \cup_{r \in rel(S_2)} D(r))\right|$$

3.2 Extension de définitions grâce aux vecteurs de sens

Notre méthode utilisant les vecteurs de sens décrits dans la section 2 est d'étendre l'algorithme de Lesk de la même manière que Banerjee & Pedersen (2002) qui exploite les relations entre les sens, lors du calcul de la similarité entre deux sens.

Étant donné que nos vecteurs de sens sont créés à partir des définitions d'un dictionnaire, les termes utilisés dans ces définitions sont très importants. Pour cette raison, quand nous voulons capturer les sens proches d'un sens S, nous ne prenons pas en compte automatiquement tous les sens qui ont une forte similarité cosinus avec S. En effet, si nous faisions cela, nous considérerions aveuglément comme "proche" tous les sens dont les définitions sont formées de mots similaires à ceux la définition de S. Ainsi nous filtrons également les sens proches du lemme du sens S, dans l'espace des vecteurs de mots. Ce filtrage assure de capturer des sens proches de l'*idée* du lemme capturée par le modèle de vecteurs de mots.

Dans notre approche nous avons eu la volonté de pouvoir effectuer des similarités entre vecteurs de sens et vecteurs de mots. Ceci est rendu possible par le fait que les vecteurs de sens sont une somme de vecteurs de mots, et donc sont de même dimension.

Si l'on note $rel(S, \delta_1, \delta_2)$ l'ensemble des sens liés à S. δ_1 est le seuil de sélection sur la similarité cosinus avec le vecteur du lemme, δ_2 est un seuil de sélection sur la similarité avec le vecteur du sens. La fonction $rel(S, \delta_1, \delta_2)$ est calculée de la manière suivante :

$$rel(S_1, \delta_1, \delta_2) = \{S_2 \in Senses \mid cosine(\phi(lemma(S_1)), \phi(S_2)) > \delta_1, cosine(\phi(S_1), \phi(S_2)) > \delta_2\}$$

Finalement, le calcul pour notre nouvel algorithme local de DL que nous appellons $VecLesk(S_1, S_2, \delta_1, \delta_2)$, peut être décrit de la manière suivante :

$$VecLesk(S_1, S_2, \delta_1, \delta_2) = \left|(D(S_1) \cup_{r \in rel(S_1, \delta_1, \delta_2)} D(r)) \cap (D(S_2) \cup_{r \in rel(S_2, \delta_2, \delta_2)} D(r))\right|$$

Dans la section suivante nous évaluons ce nouvel algorithme local sur deux tâches de DL en le comparant avec deux systèmes étalons : l'algorithme de Lesk classique, et de Lesk étendu.

4 Évaluation

Afin d'expérimenter notre nouveau modèle de vecteurs de sens, nous l'exploitons pour effectuer une extension lexicale de définitions de dictionnaire avec pour objectif d'améliorer l'algorithme local de Lesk dans notre système de DL.

Notre extension considère les sens proches d'un autre en fonction d'un seuil de similarité δ_1 sur la distance cosinus avec le lemme du sens cible, et un filtrage supplémentaire sur le vecteur de sens cible, avec un seuil δ_2.

Ces deux seuils ont été paramétrés sur un corpus de développement. Ici, nous avons utilisé la tâche 7 de SemEval 2007 (Navigli *et al.*, 2007) pour fixer le meilleur ensemble de paramètres puis testé sur la tâche 13 de SemEval 2015. Nous avons échangé les corpus dans un second temps. Ce choix a été motivé pour présenter des résultats sur deux corpus de même nature, la tâche de SemEval 2007 étant très couramment utilisé dans les articles traitant de DL.

Les cinq modèles présentés dans la section 2 sont évalués séparément. Les paramètres δ_1 et δ_2 ont été estimés en les faisant varier dans les intervalles $[0.5, 0.9]$ avec un pas de 0.1.

Les résultats des meilleurs jeux de paramètres sur SemEval 2007 et SemEval 2015 sont présentés dans le tableau Table 1. La comparaison de notre méthode par rapport aux deux systèmes étalons est présentée dans le tableau Table 2.

Modèle	baroni_c		baroni_p		deps		glove		word2vec	
Paramètres	δ_1	δ_2	δ_1	δ_2	δ_1	δ_2	δ_1	δ_2	δ_1	δ_2
Meilleurs sur SemEval 2007	0.6	0.6	0.5	0.5	0.6	0.8	0.5	0.6	0.5	0.6
Meilleurs sur SemEval 2015	0.5	0.8	0.5	0.6	0.6	0.8	0.5	0.7	0.5	0.6

TABLE 1 – Estimation des paramètres δ_1 et δ_2 sur SemEval 2007 et SemEval 2015.

Système	SemEval 2007 F1 score	SemEval 2015 F1 score
Chen *et al.* (2014)	*75.80%* [6]	
Lesk	68.70%	50.65%
Lesk étendu	78.01%	61.42%
VecLesk (baroni_c)	**75.29**%	58.02%
VecLesk (baroni_p)	73.52%	53.46%
VecLesk (deps)	73.02%	56.40%
VecLesk (glove)	73.00%	**59.01**%
VecLesk (word2vec)	73.30%	57.00%

TABLE 2 – Comparaison de nos résultats sur SemEval 2007 et SemEval 2015 pour chacun des modèles de vecteurs de mots par rapport aux étalons Lesk et Lesk étendu. Les paramètres δ_1 et δ_2 utilisés sont ceux estimés dans le précédent tableau sur **l'autre tâche**, pas celle qui est testée.

Les résultats montrent que notre extension améliore très significativement les résultats obtenus par la mesure de Lesk. L'amélioration, sur SemEval 2007, est d'environ +5% pour la pire combinaison et de +7% pour la meilleure. Sur SemEval 2015 les améliorations oscillent entre +3% et +9%. Cependant, notre extension n'atteint pas les résultats du Lesk étendu ; ceci montre que notre réseau lexical est un peu moins performant qu'un réseau créé manuellement (tel que *WordNet*).

Un point intéressant à noter est la différence de score obtenue entre les différents modèles de vecteurs de mots. Les meilleurs résultats sur SemEval 2007 utilisent la méthode de Baroni *et al.* (2014) basée sur le comptage. Ceci tend à montrer que les méthodes de création de vecteurs de mots plus "ancienne" sont parfois meilleures que les modèles prédictifs. Cependant, sur SemEval 2015, c'est la méthode GloVe qui obtient les meilleurs résultats. Ces fluctuations dans les résultats sont peut être dues aux différentes natures des méthodes de création des vecteurs de mots, ou bien aussi dues aux différents corpus sur lesquels ils ont été entrainés. En tout cas, nous montrons que quelque soit le modèle sous-jacent utilisé, notre méthode améliore systématiquement le score de Lesk.

Par ailleurs si l'on compare notre système à la meilleure méthode état de l'art à notre connaissance exploitant des ressources similaires aux nôtres (un dictionnaire et des corpus non annotés), nous observons que nos résultats sont les meilleurs sur les systèmes de DL à base de connaissances. Par ailleurs nous souhaitons souligner que les résultats présentés par Chen *et al.* (2014) sont biaisés par le fait que les paramètres optimisant leur système sont estimés sur le corpus de test. À noter que nous obtenons un score de 77.08% sur SemEval 2007 avec le meilleur jeu de paramètres appris sur cette même tâche.

6. Un biais est présent dans l'article référencé étant donné que les auteurs estiment leur paramètre $\delta \in [-0.1, 0.3]$ comparable au nôtre, mais directement durant la phase de test. Leurs scores varient entre 72.10% et 75.80% en fonction de δ.

5 Conclusion

Dans cet article nous avons présenté un modèle de vecteurs de sens, représentant tous les sens issus d'un dictionnaire. Ce modèle est construit en se basant sur les définitions du dictionnaire et sur un modèle de vecteurs de mots existant. Notre méthode consiste à sommer les vecteurs de mots présents dans la définition, pondérés en fonction de leur partie du discours ainsi que de leur fréquence.

Nous avons créé cinq modèles de vecteurs de sens, chacun reposant sur un modèle de vecteurs de mots différents. Nos modèles sont distribués librement sur la page de notre article.

Ces modèles sont ensuite utilisés comme un réseau lexical pour améliorer un système de DL basé sur l'algorithme de Lesk. Nous utilisons une méthode similaire à Banerjee & Pedersen (2002) avec comme grande différence l'utilisation d'un réseau lexical construit totalement automatiquement.

Les résultats que nous obtenons sur deux tâches de DL montrent une amélioration systématique du score par rapport à la mesure de Lesk classique (d'environ +3% à +9% selon le modèle).

Nous avons appliqué notre méthode uniquement sur la base lexicale WordNet pour évaluer nos modèles sur des tâches très connues en DL sur la langue anglaise. Cependant, notre méthode peut être appliquée très facilement à d'autres langues où les ressources sont moindres : un dictionnaire classique ainsi que des corpus non annotés sont les ressources suffisantes pour ainsi créer un système de désambiguïsation lexicale performant.

Références

BANERJEE S. & PEDERSEN T. (2002). An adapted lesk algorithm for word sense disambiguation using wordnet. In *CICLing 2002*, Mexico City.

BARONI M., DINU G. & KRUSZEWSKI (2014). Don't count, predict ! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 238–247, Baltimore, Maryland : Association for Computational Linguistics.

CHEN X., LIU Z. & SUN M. (2014). A unified model for word sense representation and disambiguation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 1025–1035, Doha, Qatar : Association for Computational Linguistics.

FERRERO J., BESACIER L., SCHWAB D. & AGNÈS F. (2017). CompiLIG at SemEval-2017 Task 1 : Cross-Language Plagiarism Detection Methods for Semantic Textual Similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval 2017)*, Vancouver, Canada.

LEVY O. & GOLDBERG Y. (2014). Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2 : Short Papers*, p. 302–308.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. S. & DEAN J. (2013). Distributed representations of words and phrases and their compositionality. In C. BURGES, L. BOTTOU, M. WELLING, Z. GHAHRAMANI & K. WEINBERGER, Eds., *Advances in Neural Information Processing Systems 26*, p. 3111–3119. Curran Associates, Inc.

MILLER G. A. (1995). Wordnet : A lexical database. *ACM*, **Vol. 38**(No. 11), p. 1–41.

NAVIGLI R. (2009). Wsd : a survey. *ACM Computing Surveys*, **41**(2), 1–69.

NAVIGLI R., LITKOWSKI K. C. & HARGRAVES O. (2007). Semeval-2007 task 07 : Coarse-grained english all-words task. In *SemEval-2007*, p. 30–35, Prague, Czech Republic.

PENNINGTON J., SOCHER R. & MANNING C. D. (2014). Glove : Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, p. 1532–1543.

VIAL L., TCHECHMEDJIEV A. & SCHWAB D. (2017). Comparison of global algorithms in word sense disambiguation.

YUAN D., RICHARDSON J., DOHERTY R., EVANS C. & ALTENDORF E. (2016). Semi-supervised word sense disambiguation with neural models. In *COLING 2016*.

Actes de TALN 2017, volume 2 : articles courts

Parcourir, reconnaître et réfléchir. Combinaison de méthodes légères pour l'extraction de relations sémantiques

Mathieu Lafourcade[1] Nathalie Le Brun[2]

(1) LIRMM, 860 rue de St Priest, 34095 Montpellier cedex 5, France
(2) Imagin@t, 34400 Lunel, France
lafourcade@lirmm.fr, imaginat@imaginat.name

RESUME

La capture de relations sémantiques entre termes à partir de textes est un moyen privilégié de constituer/alimenter une base de connaissances, ressource indispensable pour l'analyse de textes. Nous proposons et évaluons la combinaison de trois méthodes de production de relations lexico-sémantiques.

ABSTRACT

Browse, recognize and think. Combination of light methods applied to semantic relations extraction

Extracting semantic relations from texts is a good way to build and supply a knowledge base, an indispensable resource for text analysis. We propose and evaluate the combination of three ways of producing lexical-semantic relations.

MOTS-CLES : cooccurrences, inférences, patrons et relations sémantiques, combinaison

KEYWORDS: cooccurrences, inferences, semantic schemas and relations, combining.

1. Introduction

Les relations sémantiques, qu'elles soient ontologiques (hyperonymes, hyponymes, parties/tout), lexicales (synonymes), ou encore qu'elles relèvent des rôles sémantiques (agent, patient, instrument, manière, lieux, etc.) sont d'un intérêt majeur pour la quasi-totalité des applications du TALN où le système doit « comprendre » de quoi il retourne : traduction automatique, indexation, résumé, détection de textes similaires, etc. La création de procédures pour produire ce type de ressources répond donc à de multiples besoins en TALN.

Les approches peuvent être variées, manuelles comme pour WordNet, (Miller, 1995), plus ou moins automatiques (BabelNet, Navigli and Ponzetto, 2010), ou encore contributives (Lafourcade, *et al.*, 2015). De nombreuses méthodes automatiques d'extraction de relations à partir de textes ont été proposées, mais les performances restent très inégales. En effet, certaines approches se veulent très précises, ce qui suppose une analyse lexico-sémantique profonde, donc coûteuse. La vitesse de traitement des textes, et donc le taux d'extraction de nouvelles relations, s'en trouve

ralenti. A l'inverse, on trouve des méthodes statistiques qui n'incluent quasiment aucun traitement linguistique du texte d'entrée.

Nous proposons dans cet article d'évaluer les bénéfices d'une approche combinant plusieurs stratégies. Cette approche s'inscrit dans le cadre d'un apprentissage sans fin (never ended learning) au sein du projet JeuxDeMots (JDM). L'idée est de mettre en place des boucles d'extraction/exploitation où un système d'extraction automatique joue le rôle de contributeur au sein du réseau. Les joueurs, par leur activité ludique ou de contribution directe, valident ou invalident les contributions. Ainsi, il est possible de façon holistique d'évaluer l'évolution des performances de notre système Schémas – Inférences- Cooccurrences (SIC), qui alimente le réseau et l'exploite pour mener sa tâche à bien.

Dans ce qui suit, nous présentons rapidement les travaux précédents en extraction automatique de relations sémantiques, en nous limitant à ceux dont l'approche est en rapport avec notre proposition. Ensuite, nous détaillons trois stratégies d'extraction ainsi que la façon dont nous les combinons. Enfin, nous détaillons et discutons les résultats obtenus.

2. Etat de l'art

L'utilisation de schémas lexico-sémantiques a été proposée par (Hearst, 1992) pour extraire des relations de synonymie et d'hyperonymie dans des textes. Les schémas sont par exemple de la forme « A est un type de B ». (Herbelot et Copestake, 2006) ont utilisé ce genre de schémas afin d'extraire des relations dans le domaine de la biologie à partir de pages Wikipédia, avec une précision excellente (88%) mais un rappel assez faible (20%).

Dans (Ruiz-Caasado, 2005 et 2007) il est fait état d'apprentissage automatique de tels schémas de façon à extraire des relations également depuis Wikipédia, et les insérer dans Wordnet ; on note là aussi la faiblesse des performances concernant le rappel. L'approche consistant à faire appel à de l'apprentissage automatique de schémas à partir de textes a aussi été explorée par (Snow, *et al.*, 2004), également pour identifier des relations d'hyperonymie ou d'hyponymie. Dans (Girju, *et al.*, 2003) une approche supervisée vise à déterminer des contraintes sémantiques pour extraire des relations de méronymie. Les contraintes sont issues de la relation *part-of* de Wordnet et servent de données d'entraînement. Dans (Ramadier et Lafourcade, 2016) une approche similaire est présentée, à ceci près qu'un grand nombre de types de relations sémantiques sont identifiés et que les contraintes sémantiques sont déterminées manuellement.

De nombreux auteurs tentent d'extraire des relations depuis Wikipédia en exploitant les informations de structure des pages. Par exemple, (Sumida et Torisawa, 2008) ont utilisé cette stratégie sur Wikipédia en japonais pour extraire 1.4 millions de relations d'hyponymie avec une précision de 0.75. De façon similaire, (Ponzetto and Strub, 2007) exploitent les liens de catégorie de Wikipédia pour identifier des relations d'hyperonymie. Les travaux de (Pachenko, 2013) présentent une analyse en profondeur de fonctions d'évaluation de relations sémantiques entre termes. Une des conclusions est qu'aucune des mesures d'évaluation de la relation sémantique ne surpasse les autres. Ce sont ces différentes mesures de similarité qui conditionnent la proposition de telles ou telles relations sémantiques, ontologiques pour la plupart (hyperonymes, co-hyponymes, etc.)

On remarquera que peu de travaux sur la question s'appuient sur l'utilisation de bases de connaissances pour extraire de nouvelles relations sémantiques dans une approche de type apprentissage permanent en boucle. Par contre, quand il s'agit d'apprentissage automatique, de telles bases sont utilisées pour l'entraînement. De plus, la plupart des approches se limitent à des

relations ontologiques, comme l'hyperonymie (isa), la synonymie (syn) et la méronymie ou l'holonymie (has parts / is part of). Des relations comme cause/conséquence, caractéristiques, lieux, agent, patient et instrument (pour des verbes) ne font que rarement l'objet d'extraction.

Dans l'approche que nous expérimentons, nous nous focalisons sur du texte pur, en français exclusivement, issu de Wikipédia ou autre, sans exploiter la structure du document source. Nous souhaitons aussi extraire des informations de textes non encyclopédiques (comme des romans, par exemple). Pour chacune des méthodes d'extraction, nous utilisons, à des degrés différents, le réseau lexico-sémantique JeuxDeMots.

3. Combinaison de méthodes d'extraction de relations

Nous présentons trois méthodes relativement simples d'extraction de relations sémantiques entre paires de termes. Nous indiquons ensuite les grandes lignes de la combinaison de ces méthodes.

3.1. Cooccurrences et relations

La première méthode pour détecter des relations sémantiques est de construire un réseau de cooccurrences. La méthode tient compte des termes composés et comprend un prétraitement sur le texte, qui consiste à :

- Remplacer un terme par son lemme, quand le terme est un verbe conjugué et uniquement cela. Par exemple, le segment : « les poules dorment » deviendra « les poules dormir ». Par contre le segment « les poules couvent » reste inchangé car *couvent* peut être une forme du verbe *couver* mais aussi le substantif.
- Repérer les occurrences de termes composés par confrontation avec le réseau JDM. Les espaces sont remplacés par des soulignés, ce qui permet d'éviter leur segmentation.
- Les ponctuations sont conservées mais sont décollées des termes qui les précèdent (*chat, => chat* ,).
- Les majuscules ne sont pas modifiées ;
- En dehors de ce qui est mentionné ci-dessus, ni *pos tagging* ni analyse syntaxique ne sont réalisés.

Les termes composés sont identifiés par comparaison avec ceux existant dans le réseau JDM. En cas de conflit (par exemple, un segment A B C avec deux mots composés A_B et B_C), on applique une priorité à droite (on obtiendra A B_C). Ensuite, une segmentation est réalisée à partir des caractères espaces. Une fenêtre de k mots est utilisée pour établir les relations de cooccurrences, avec un poids décroissant de k (mot adjacent) à 1 (mot à une distance de k termes). Nous avons utilisé une fenêtre de 10 mots, afin de maximiser le rappel, ce qui est l'objet de cette méthode.

Nous utilisons la base de connaissances JeuxDeMots comme support pour la détermination des lemmes et catégories morphosyntaxiques mais aussi pour l'identification approximative des types de relations. Plus précisément, nous avons des règles de ce type, qui exploitent les parties du discours :

- Si X r_pos Verbe & Y r_pos Adv → X r_manner Y ;
- Si X r_pos Nom & Y r_pos Adj →X r_caracY ;
- Par défaut, si X est en cooccurrence avec Y → X r_assoc Y

Les règles sont strictes et doivent être entendues comme : si X est un verbe et uniquement un verbe et si Y est uniquement un adverbe alors X sera lié à Y par une relation de manière. Par exemple, soit la phrase suivante : « le chat attrapa rapidement le rat noir ». La phase de prétraitement nous fournit le texte : « le chat attraper rapidement le rat noir » (nous n'indiquons pas les poids). Ces relations sont pondérées par le poids de la cooccurrence entre les deux termes.

• chat r_assoc attraper • attraper r_manner rapidement • attraper r_assoc rat • attraper r_assoc noir • rapidement r_assoc rat • …	• attraper r_assoc rat • rat r_assoc noir • chat r_assoc rat • chat r_assoc noir • rapidement r_assoc chat

3.2. Schéma lexico-sémantique avec contraintes

Dans (Ramadier et Lafourcade, 2016) la méthode de (Hearst, 1992) et (Herbelot et Copestake, 2006) qui exploite des schémas lexico-sémantiques, a été étendue de façon à ce que les termes vérifient des relations sémantiques issues du réseau lexical de JeuxDeMots. Par exemple :

- X du Y avec X r_isa artéfact & Y r_isa personne → Y r_own X (le fusil du soldat)
- X du Y avec X r_isa partie du corps & Y r_isa personne → Y r_part X (le bras du soldat)
- X du Y avec X r_isa personne & Y r_isa lieu_humain → Y r_lieu X (la fille du coron)

Bien sûr, certains schémas ne sont pas associés à des contraintes, par exemple :

- X est localisé dans le/la/mes/un/une/des/ Y → X r_lieu Y
- X est un type de Y → X r_isa Y
- X fait partie de Y → X r_holo Y
- X se compose de Y → X r_has_parts Y

Une relation entre deux termes sera pondérée par le nombre de fois ou elle a été découverte par des schémas différents dans des segments textuels distincts.

3.3. Induction et Abduction sur un réseau lexico-sémantique

Dans (Zarrouk, *et al.*, 2014) et (Zarrouk and Lafourcade, 2015), des approches de production de nouvelles relations lexico-sémantiques par inférence ont été proposées. Ces approches sont strictement endogènes sur le réseau sémantique JeuxDeMots et se basent sur la déduction et plusieurs formes d'abduction. Aucun texte n'est donc utilisé pour cette méthode.

3.4. Combinaison de relations

La combinaison de deux méthodes consiste à ne retenir que les relations sémantiques trouvées conjointement par chacune des deux méthodes. Bien que la méthode par cooccurrences produise

des relations sous-spécifiées (relation de type *idées associées* entre les termes), qui n'ont pas d'équivalences dans les deux autres méthodes, ces relations neutres sont utilisées comme suit :

X r_t Y + X r_assoc Y → X r_t Y	Une relation neutre (de type *idées associées* r_assoc) valide une relation typée (r_t) pour la même paire de termes.

Nous combinons les approches deux à deux car une combinaison des trois approches, bien qu'augmentant la précision, réduirait trop drastiquement le nombre de relations retenues. On retiendra donc les relations produites par au moins deux des trois méthodes. Le poids de la combinaison est la moyenne géométrique des relations.

4. Expérimentation et Discussion

Nous avons fondé notre expérimentation sur le réseau lexical JeuxDeMots pour l'approche par inférences. Pour les deux autres approches, qui sont basées sur des textes, nous avons utilisé un corpus comprenant Wikipédia en français (pour les schémas et les cooccurrences) et l'œuvre d'Émile Zola via Wikisource (pour les cooccurrences). Le choix de ce corpus pour les cooccurrences s'explique par la volonté de ne pas nous limiter à des textes de nature encyclopédique, mais d'enrichir la collecte de relations sémantiques en exploitant les avantages de la littérature romanesque : offrir (1) une plus grande diversité de relations entre les termes, et (2) plus d'informations de sens commun et de la vie quotidienne.

Les relations extraites sont les suivantes : synonymie (pour les verbes, noms, adjectifs, adverbes), agent, patient, instrument, manière, a pour lieu (pour les verbes), hyperonymie, hyponymie, instance, caractéristique, a pour lieu, est un lieu pour, parties, tout, cause, conséquence (pour les noms). Certaines relations sont clairement symétriques (comme l'hyperonymie et l'hyponymie) cependant le niveau de renseignement est très différent, en particulier au niveau des poids. Dans le cas du réseau JeuxDeMots, ces relations sont volontairement non symétrisées.

La *productivité* est la capacité d'une approche à produire des relations. Nous utilisons cette mesure en lieu et place du traditionnel rappel, que nous ne sommes pas en mesure d'évaluer. En effet, nous ne disposons pas des moyens humains pour déterminer l'ensemble des relations qu'il faudrait extraire sur notre corpus. Ce travail est lourd dans la mesure où il faut lire chacun des textes. De plus, l'accord inter-annotateur n'est généralement pas très élevé (moins de 50% en moyenne). Nous prenons comme référence l'approche par inférence et nous lui attribuons une productivité de 1. En pratique, l'approche par inférence a produit environ 60 millions de relations (qui restent potentielles tant que non validées) entre le 1er novembre 2016 et le 30 mars 2017. La *précision* est la mesure classique du rapport entre les relations évaluées comme justes et l'ensemble des relations proposées. Dans notre approche, on cherche bien entendu à maximiser ce critère critique, tout en restant raisonnablement productif. Enfin, la *pertinence* est le rapport entre les relations pertinentes et les relations justes. Décider si une relation est pertinente reste relativement subjectif, mais en général les personnes interrogées (par crowdsourcing et GWAP) tombent globalement d'accord. Par exemple, la relation : *souris r_has_part atomes*, est correcte mais peu pertinente, car la propriété d'être constitué d'atomes est universelle. La pertinence a un

rapport avec la spécificité d'une relation, en général, plus une relation est spécifique à une classe de termes, plus elle est pertinente.

L'évaluation a été réalisé par deux méthodes conjointes 1) validation manuelle d'un échantillon aléatoire et 2) croisements de réponses de joueurs via JeuxDeMots. Ces évaluations se font en continu (pour les chiffres ci-dessous sur la période de novembre 2016 à mars 2017), les résultats affichés ci-dessous sont ceux de fin mars 2017. La validation manuelle consiste à proposer la relation à un joueur (jeu Askit, http://jeuxdemots.org/askit.php), ce dernier devant se prononcer sur sa validité et sa pertinence. La méthode par croisement fait appel au jeu classique JeuxDeMots où des parties sont offertes aux joueurs avec comme mot-cible le terme premier de la relation trouvée et comme consigne le type de la relation. Par exemple, si la relation *: rat r_carac noir* a été extraite, alors des parties seront proposées pour le terme *rat* et la relation *r_carac*. Si parmi les réponses se trouve le terme *noir*, alors la relation est validée. Cette approche équivaut à interroger des personnes afin de voir si la relation extraite émerge ou pas. Les joueurs peuvent passer s'ils ne savent pas. Nous avons sélectionné en priorité les relations dont le poids correspond au 2^{e} quartile (soit les 50% ayant les poids les plus élevés).

	Schémas (S)	**Inférences (I)**	**Cooccurrences (C)**
Productivité	0.37 (22 M)	1 (60 M)	3,16 (190 M)
Précision	93 %	65 %	12 %
Pertinence	88 %	75 %	47 %

Tableau 1 : Evaluation des méthodes isolées. Nous prenons la méthode I comme référence pour le nombre de relations produite en 5 mois (60 millions pour I).

La méthode d'extraction par schémas (S) lexico-sémantiques avec contraintes produit très peu de relations fausses mais elle est relativement lente. Les 7% d'erreur correspondent à l'impossibilité d'appliquer des contraintes, quand au moins un des deux termes de la relation n'est pas suffisamment renseigné. Les relations extraites sont pertinentes, ce qui est normal puisqu'elles sont directement prélevées telles quelles dans des textes. La méthode par cooccurrence, comme attendu, est très productive, très rapide, et fort peu précise (beaucoup de déchets). Les relations correctes sont pertinentes une fois sur deux. Enfin, la méthode par inférence (qui ne s'appuie que sur la base de connaissances JeuxDeMots) montre des performances correctes. Les erreurs viennent essentiellement de l'impact de la polysémie, qui vient perturber les inférences déductives et abductives. La différence de productivité entre S et C s'explique essentiellement par la différence de vitesse des deux méthodes (S lente et précise, C rapide et floue).

La combinaison des méthodes deux à deux (tableau 2) consiste à ne retenir une relation que si elle est proposée par les deux méthodes. Il s'agit donc d'une intersection entre les propositions des deux méthodes.

	S+I	**S+C**	**I+C**
Productivité	0.22	0.35	0.78
Précision	96 %	94 %	87 %
Pertinence	93 %	84 %	88 %

Tableau 2 : Evaluation des méthodes par paires. S correspond à la méthode fondée sur les Schéma lexico-sémantiques, I à l'approche par Inférences, et C à l'approche par Cooccurrences.

Nous constatons que pour chaque paire la productivité baisse par rapport à celle obtenue pour chacune des méthodes prise isolément, ce qui est un résultat attendu. La combinaison S+I a une très faible productivité ce qui indique que peu de relations sont produites à la fois par les méthodes S et I.

L'approche par union des trois méthodes (tableau 3), permet non seulement d'obtenir une augmentation de productivité de 28 % par rapport à la méthode I (prise comme référence), mais surtout d'augmenter la précision et la pertinence. Si nous prenons comme référence la méthode I, la combinaison des approches permet de renforcer la précision avec la méthode S et la pertinence avec S et C. Combiner S+C permet (toujours en référence à I) de rajouter des relations qui n'ont pu être inférées. On rappelle que C est appliqué sur un corpus de textes plus large et plus général que la méthode S.

	(S+I) ∪ (S+C) ∪ (I+C)
Productivité	1.28
Précision	99.4 %
Pertinence	96 %

Tableau 3 : Évaluation de l'approche retenant les relations proposées par au moins 2 méthodes (union). La précision de 99.4 est supérieure à la meilleure précision du tableau 2, car en effet la combinaison de méthodes rend cela possible (union des paires de méthodes).

5. Conclusion

Nous avons présenté trois méthodes d'identification de relations sémantiques entre termes. L'une d'elles est strictement endogène par rapport à une base de connaissances (le réseau JeuxDeMots dans notre cas), les deux autres se fondent sur des textes, mais exploitent également des informations sémantiques externes aux textes. Globalement, ces méthodes sont légères. Individuellement, elles ont des défauts (productive mais imprécise / précise mais peu productive). Leur union, en compensant leurs défauts respectifs, permet d'améliorer significativement productivité, pertinence, et précision.

L'approche proposée est générique et peut être appliquée à d'autres bases de connaissances, dans d'autres langues ou pour des domaines de spécialités. Toutefois dans ce dernier cas, il faut disposer d'un corpus de textes suffisamment important afin d'être capable d'utiliser des schémas lexico-sémantiques et d'extraire des cooccurrences dans des proportions statistiquement significatives.

Les relations strictement sémantiques (et non lexicales) étant indépendantes des langues, une piste d'amélioration du processus serait d'adapter cette approche à l'extraction de relations depuis des textes de langues différentes, en ayant recours soit à un processus de traduction, soit au réseau JDM multilingue actuellement à l'étude.

Références

GIRJU, R., BADULESCU, A., and MOLDOVAN, D. (2003) *Learning semantic constraints for the automatic discovery of part-whole relations.* In Proc. Conf. North American Chapter of the Association for Computational Linguistics on Human Language Technology , NAACL '03, pages 1–8. Association for Computational Linguistics, 2003.

HEARST, M. A. (1992) *Automatic acquisition of hyponyms from large text corpora.* In Proc . 14th Conf. on Computational Linguistics, COLING '92, pages 539–545. Association for Computational Linguistics, 1992.

HERBELOT, A. and COPESTAKE, A. (2006) *Acquiring ontological relationships from Wikipédia using RMRS.* In Proc. ISWC 2006 Workshop on Web Content Mining with Human Language Technologies , 2006.

LAFOURCADE, M. (2007) *Making people play for Lexical Acquisition.* In Proc. SNLP 2007, 7th Symposium on Natural Language Processing. Pattaya, Thailande, 13-15 December 2007, 8 p.

LAFOURCADE, M., LE BRUN N., and JOUBERT A. (2015) *Games with a Purpose (GWAPS)*, ISBN: 978-1-84821-803-1 July 2015, Wiley-ISTE, 158 p.

MILLER, G. A. (1995) *Wordnet: A lexical database for English.* Communications of the ACM , 38(11):39–41, November 1995.

NAVIGLI, R.and PONZETTO, S. P. (2010*) Babelnet: Building a very large multilingual semantic network.* In Proc. 48th Annual Meeting of the Association for Computational Linguistics , ACL'10, pages 216–225, 2010.

PANCHENKO, A. (2013) *Similarity Measures for Semantic Relation Extraction*. PhD Dissertation, Université catholique de Louvain & Bauman Moscow State Technical University, 193 p.

RAMADIER, L. ET LAFOURCADE, M. (2016) *Patrons sémantiques pour l'extraction de relations entre termes - Application aux comptes rendus radiologiques.* In 23rd French Conference on Natural Language Processing (JEP-TALN-RECITAL 2016), Paris, France, 4-8 July 2016, 6 p.

RUIZ-CASADO M., ALFONSECA E., AND CASTELLS P. (2005) *Automatic extraction of semantic relationships for wordnet by means of pattern learning from Wikipédia*. In Proc. 10th Int. Conf. Natural Language Processing and Information Systems, NLDB'05, pages 67–79. Springer, 2005

SNOW R., JURAFSKY D., AND ANDREW Y. NG. (2004) *Learning syntactic patterns for automatic hypernym discovery.* In Advances in Neural Information Processing Systems (NIPS), 8 p. 2004.

SUMIDA, A. AND TORISAWA, K. (2008) *Hacking Wikipédia for hyponymy relation acquisition.* In Proc. of IJCNLP 2008 , pages 883–888, 2008.

ZARROUK, M., LAFOURCADE, M., and JOUBERT A. (2014). About Inferences in a Crowdsourced Lexical-Semantic Network, *EACL 2014 (14th Conference of the European Chapter of the Association for Computational Linguistics)*, Gothenburg (Sweden), April 2014

ZARROUK, M. and LAFOURCADE, M. (2014) *Inferring Knowledge with Word Refinements in a Crowdsourced Lexical-Semantic Network.* In proc of the the 25th International Conference on Computational Linguistics (COLING 2014), Dublin, Irlande, 9 p.

Si les souris étaient des reptiles, alors les reptiles pourraient être des mammifères ou Comment détecter les anomalies dans un réseau lexical ?

Mathieu Lafourcade[1] Alain Joubert[1] Nathalie Le Brun[2]
(1) LIRMM, 860 rue de St Priest, 34095 Montpellier cedex 5, France
(2) Imagin@t, 34400 Lunel, France
lafourcade@lirmm.fr, joubert@lirmm.fr, imaginat@imaginat.name

RESUME

La correction des erreurs dans une collection de données est un problème délicat. Elle peut être réalisée manuellement par un expert, ou en utilisant des méthodes de crowdsourcing, ou encore automatiquement au moyen d'algorithmes. Nous présentons ici des méthodes automatiques permettant de détecter les erreurs potentielles « secondaires » induites par les mécanismes automatiques d'inférences de relations, lorsqu'ils s'appuient sur des relations erronées « initiales » détectées manuellement. Des résultats encourageants, mesurés sur le réseau JeuxDeMots, nous invitent à envisager également des stratégies qui permettraient de détecter automatiquement les relations erronées « initiales », ce qui pourrait conduire à une détection automatique de la majorité des erreurs présentes dans le réseau.

ABSTRACT

If mice were reptiles, then the reptiles could be mammals, or How to detect errors in a lexical network?

Correcting errors in a data set is a critical issue. This task can be either hand-made by experts, or by crowdsourcing methods, or automatically done using algorithms. We present here automatic methods for detecting potential "secondary" errors that would result from automatic inference mechanisms when they rely on an "initial" error manually detected. We measured encouraging results on the JeuxDeMots network; they also invite us to consider strategies that would automatically detect erroneous "initial" relations, which could lead to the automatic detection of the majority of errors in the network.

MOTS-CLES : Réseau JeuxDeMots, détection des relations erronées, inférence de relations

KEYWORDS: JeuxDeMots network, detection of erroneous relations, inference of relations

1. Introduction

Toute collection de données est nécessairement entachée d'erreurs et, selon le domaine et les applications concernés, leur nombre est plus ou moins tolérable. Les réseaux lexicaux ne font pas exception à la règle : dans le réseau JeuxDeMots (JDM), le taux d'anomalies est relativement faible (largement inférieur à 1%),). Minimiser ce taux d'erreur reste toutefois une priorité, et suppose, pour optimiser le taux de correction, des stratégies de détection efficaces.

Les anomalies sont de natures variées. Elles peuvent porter soit sur les termes, comme les fautes d'orthographe (ex : *théatre*/*théâtre*), soit sur les relations entre termes (ex : *Milou est_un être_humain* ou *Dalida idée_associée Samson*, par confusion entre *Dalida* et *Dalila*).

Actuellement, la détection des anomalies est essentiellement réalisée manuellement par des joueurs/contributeurs *via* leur activité d'enrichissement du réseau à travers l'interface Diko (le dictionnaire contributif d'associations lexicales du projet JDM). Mais comme elle est fortuite, elle ne peut en aucun cas prétendre à l'exhaustivité, d'où la nécessité de mettre au point une vraie méthode de détection.

Dans un premier temps, nous regarderons d'où peuvent provenir les anomalies, puis nous présenterons une méthode qui permet de signaler comme fausses un certain nombre de relations, à charge pour le validateur humain d'effectuer les corrections correspondantes.

2. D'où proviennent les anomalies dans le réseau JDM ?

Le mode de construction et les caractéristiques du réseau JDM tels que décrits dans (Lafourcade *et al.*, 2015) le rendent vulnérable à deux grands types d'erreurs :

- Les anomalies « initiales » introduites par les joueurs et contributeurs, volontairement ou non. Selon notre expérience, il s'agit essentiellement d'erreurs involontaires, car l'intérêt de saisir volontairement des informations erronées est très limité. Effectivement, de par le principe de construction du réseau JDM, une relation n'est créée (ou renforcée si elle existait déjà) que si les deux joueurs l'ont proposée en réponse à une même consigne. Le fait que les parties soient jouées de façon anonyme et asynchrone interdit toute communication entre les deux joueurs d'une même partie. Saisir des incohérences n'a donc qu'une probabilité extrêmement faible d'avoir des conséquences sur les données enregistrées. Cela entraine une forte diminution du risque, sans toutefois l'annuler totalement.

- Les anomalies « secondaires » induites par les mécanismes automatiques d'inférences (déduction, induction, abduction) à partir des anomalies initiales. Effectivement, le réseau JDM peut être densifié automatiquement par inférences à partir des relations déjà existantes (Lafourcade *et al.*, 2014) ; si certaines de ces relations « initiales » sont erronées, alors les mécanismes d'inférences vont générer des relations potentiellement erronées. Quand les relations inférées automatiquement sont jugées incertaines par le système, (à savoir qu'elles sont détectées par les algorithmes présentés ci-après), alors leur validation est soumise à un processus de votes majoritaires et/ou un avis d'expert, ce qui réduit fortement le risque d'enregistrer des relations erronées.

Prenons un exemple : nous avons la relation *souris is_a mammifère*. Supposons que soit apparue ensuite la relation erronée « initiale » : *souris is_a reptile*. Le système dispose d'une série de paires d'hyperonymes incompatibles ; il « sait » donc que *mammifère* et *reptile* sont

incompatibles, tout comme *mammifère* et *poisson* ou *mammifère* et *insecte*, par exemple. Il en déduit alors que le terme *souris* est polysémique, et donc apparaissent les raffinements : *souris* > *mammifère* et *souris* > *reptile*. A partir de ce dernier raffinement, par des mécanismes de déduction/induction, le système peut générer de nouvelles relations qui seront probablement erronées puisque le raffinement *souris* > *reptile* est erroné.

3. Comment détecter et corriger les anomalies

Plusieurs auteurs se sont penchés sur la problématique de la détection/correction automatique des erreurs dans le domaine du TALN. (Boudin et Hernandez, 2012) proposent des méthodes de détection/correction automatique des erreurs d'annotation de syntaxe dans le French Treebank, qui s'appuient sur une approche présentée précédemment par (Dikinson et Meurers, 2003). Concernant la détection/correction des erreurs sémantiques, citons les travaux de (Ben Othmane Zribi *et al.*, 2007) pour la langue arabe. (Bouraoui et al, 2009) ont analysé les différents types d'erreurs que l'on rencontre en expression écrite, afin de réaliser une typologie générale des erreurs.

Avant toute chose, signalons que, dans le réseau JDM, corriger une anomalie ne signifie pas effacer la relation concernée, mais plus exactement la *négativer*, c'est-à-dire lui affecter un poids négatif. En effet, il peut être intéressant d'avoir l'information qu'une relation n'est pas vraie, plutôt que de ne pas avoir d'information sur cette relation. Par exemple, la relation *autruche agent voler* pondérée négativement signifie qu'une *autruche* ne peut pas *voler*, alors que l'absence de relation signifierait que l'on ne sait pas si une *autruche* peut *voler* ou non. De plus, en négativant une relation, on s'assure qu'elle ne puisse pas réapparaitre, ce qui pourrait être le cas en la supprimant. Notre détecteur automatique d'erreurs va préparer ce processus de correction : en contribuant par un vote négatif sur une relation suspecte, il va la signaler comme telle au validateur humain, qui tranchera en décidant de la négativer (ou pas…).

3.1. Anomalies initiales

Elles semblent difficiles à détecter par des mécanismes endogènes. Si, par exemple deux joueurs sur une même partie ont proposé la relation *souris is_a reptile*, le système n'a pas actuellement les moyens de détecter qu'il s'agit d'une relation erronée. Il peut juste en déduire que le terme *souris* est polysémique. Ces relations sont actuellement signalées par les contributeurs/joueurs puis corrigées manuellement par un expert. Des mécanismes exogènes, basés sur des connaissances extérieures au réseau JDM (Wikipédia, Babelnet …), pourraient être envisagés, en particulier en ce qui concerne les fautes d'orthographe sur les termes.

3.2. Anomalies secondaires

Quand une anomalie est détectée (par un contributeur), comment retrouver les inférences fausses qu'a pu faire le système à partir de cette erreur ? Il s'avère que ces relations fausses que le système a inférées ont été créées soit par déduction, soit, dans une moindre mesure, par induction (Lafourcade *et al.*, 2014).

Déduction

Le mécanisme de la déduction est le suivant : Soit *R* une relation sémantique quelconque. Si *A is_a B* et *B R C* , alors *A R C* possible (sauf exception ou polysémie de B). Donc, si *A is_a B* est fausse, il se peut que *A R C* le soit aussi. Il faut signaler comme fausses les relations sortantes de A qui proviennent des propriétés de B, excepté celles qui proviennent des hyperonymes de A.

Ex : *souris is_a reptile* => inférences sur *souris* basées sur les propriétés de *reptile*
Comment détecter, et donc négativer, ces inférences fausses ? On sait que *souris is_a mammifère*. Dans les relations de *souris*, il faut signaler comme potentiellement fausses celles qui proviennent des propriétés de *reptile* à l'exclusion de celles qui seraient en commun entre *reptile* et *mammifère* :

reptile r_agent pondre des œufs, ce qui n'est pas une relation partagée avec *mammifère* => donc, la relation *souris agent pondre_des_œufs* (inférée par déduction) est à signaler comme fausse.
a contrario, *reptile has_part vertèbres*, et *mammifère has_part vertèbres* => donc, la relation *souris has_part vertèbres* est à conserver.

Le problème ne se limite pas aux relations portant sur le seul terme pour lequel l'anomalie a été décelée. Par processus de déduction, le système a pu inférer de nouvelles relations erronées (ex : *souris agent pondre_des_œufs*), et à partir de ces relations erronées, inférer de nouvelles relations erronées. La relation erronée partant de *souris*, tout spécifique de *souris* a pu être infecté par application de la déduction. Il sera donc nécessaire de rechercher également des relations « secondaires » potentiellement erronées déduites des spécifiques de *souris.*

Par exemple, comme *souris_blanche is_a souris*, la relation *souris_blanche agent pondre_des_œufs* a pu être inférée. Comment détecter cette nouvelle relation erronée ? La relation *souris agent pondre_des_œufs* ayant reçu un vote négatif à l'étape précédente le système de détection va en tenir compte et signaler également cette relation comme suspecte en l'affectant d'une contribution négative.

Les anomalies « secondaires » ont pu également se propager dans la chaine des génériques du terme initial. Si *A is_a B* est erronée, dans la recherche des relations « secondaires » potentiellement erronées, il est nécessaire qu'un validateur remonte dans la chaine des génériques de B pour vérifier la validité des différentes relations de généricité entre A et les termes rencontrés. Dès que le validateur rencontrera une relation valide, il ne sera plus nécessaire de continuer à remonter.

- ex : *souris is_a reptile* est erronée. Donc avons-nous :
- *reptile is_a sauropside => souris is_a sauropside* ? réponse : relation erronée, on la négative et on continue
- *sauropside is_a vertébré => souris is_a vertébré* ? réponse : relation valide. On peut donc arrêter le processus de recherche des relations « secondaires » potentiellement erronées trouvées par déduction/induction à *souris is_a sauropside*.

Induction

Mécanisme de l'induction : Si *A is_a B* et *A R C*, alors *B R C* possible (sauf exception ou polysémie de A) Donc, si *A is_a B* est erronée, il se peut que *B R C* le soit aussi. Il faut donc signaler comme fausses (en les affectant d'un vote négatif) les relations sortantes de B qui proviennent des propriétés de A, à l'exclusion de celles provenant des hyponymes de B.
Ex : *souris is_a reptile* => inférences sur *reptile* en se basant sur les propriétés de *souris*
Comment détecter, et donc négativer, ces inférences fausses ?

En procédant par comparaison avec des hyponymes bien renseignés de *reptile* : on sait que *tortue is_a reptile*. Dans les relations de *reptile*, il faut détecter comme faux ce qui provient des propriétés de *souris*, à l'exclusion de ce qui provient des propriétés de *tortue*.

- *souris has_part poils*, ce que n'a pas *tortue* ; donc, la relation *reptile has_part poils* (inférée par induction) est à signaler comme fausse.
- a contrario, *souris has_part tête*, et *tortue has_part tête* ; donc, la relation *reptile has_part tête* est à conserver.

Algorithmes

Nous synthétisons les considérations ci-dessus sous la forme des deux algorithmes suivants :

```
fonction ListeRelations erronéesParDéduction (Terme A, Z)
    // pré-requis : A is_a Z est une relation erronée
    // résultat : retourne la liste des relations sortantes de A, présentes dans le réseau JDM,
    //                 qui sont potentiellement erronées du fait que A is_a Z est erronée
Début
    LR = new ListeRelations()
    LT = hyper (A, Z)          // liste des hyperonymes de A non hyperonymes de Z
    Pour chaque B ∈ LT
    Faire   LR = LR ∪ déduitErronées (A, B, Z)
            // liste des relations sortantes de A potentiellement erronées
    FinPour
    Retourne LR
Fin
```

```
fonction ListeRelations erronéesParInduction (Terme A, Z)
    // pré-requis : A is_a Z est une relation erronée
    // résultat : retourne la liste des relations sortantes de Z, présentes dans le réseau JDM,
    //                 qui sont potentiellement erronées du fait que A is_a Z est erronée
Début
    LR = new ListeRelations()
    Pour chaque relation de type ARC
    // Parcours sur l'ensemble des relations sortantes de A
    Faire   Si ZRC > 0        // si elle est sortante de Z, il se pourrait qu'elle soit erronée
            Alors   Si erronéeParInduction (A, C, Z)   // ZRC potentiellement erronée ?
                    Alors LR = LR ∪ ZRC
                    FinSi
            FinSi
    FinPour
    Retourne LR
Fin
```

Pour ces anomalies secondaires, une question subsiste : les mécanismes endogènes mis en œuvre détectent-ils tout ? Ou plus exactement quelle proportion d'anomalies est corrigée par ces mécanismes ? De plus, ces mécanismes de correction ne risquent-ils pas d'introduire des erreurs ?

3.3. Résultats

Pour évaluer les performances de notre système de détection de relations fausses, nous avons (sur une copie locale du réseau lexical JeuxDeMots) artificiellement ajouté des hyperonymes faux à des termes. Nous disposons dans JDM d'une liste de paires d'hyperonymes incompatibles, c'est-à-dire tels qu'un terme ne peut pas avoir les deux termes d'une paire comme génériques, à moins d'être polysémique. Par exemple : *poisson – mammifère ; insecte – reptile ; animal – plante ; avion – navire ; homme – femme ; voiture – avion ; train – bateau, etc*. Nous avons sélectionné 250 termes ayant comme hyperonyme un des génériques ci-dessus mais pas le second, que nous avons donc rajouté. Nous avons ensuite lancé sur les 250 termes les mécanismes d'inférences. Nous avons ensuite appliqué nos deux algorithmes de détection de relations fausses afin de trouver les relations à éliminer.

Les mécanismes d'inférences ont produit 4500 nouvelles relations que nous avons évaluées *via* l'application en ligne Askit (http://jeuxdemots/askit.php). Nous avons retenu les 3600 qui ont été évalués au moins deux fois. Le tableau suivant présente les résultats de l'évaluation globalement et pour quelques exemples de couples de génériques incompatibles.

% relations trouvées par les 2 algorithmes	**global**	**poisson / *mammifère**	**insecte / *reptile**	**animal / *plante**	**avion / *navire**	**homme / *femme**
% correctement trouvées	97.6	98.2	95.6	98.4	96.1	99.2
% faux positifs	2.4	1.8	4.4	1.6	3.8	0.8
% faux négatifs	0.52	0.25	0.67	0.35	0.7	0.3

Les termes avec un astérisque (*) sont les génériques faux introduits afin de produire des inférences erronées. Globalement, nos algorithmes retrouvent 97 % des relations fausses ayant été inférées à partir d'un générique erroné. Ils « loupent » 3% de relations fausses, et trouve 0.5% de faux négatifs (ils supposent fausses des relations qui sont vraies). L'analyse des faux négatifs nous indique qu'il s'agit soit d'exceptions soit de conclusions légitimes vu l'état de complétion du réseau lexical (des relations importantes peuvent manquer).

On remarque des écarts assez importants selon les couples de génériques incompatibles. Manifestement, les différences entre *homme* et *femme* font que les inférences fausses sont plus facilement détectables qu'entre *insecte* et *reptile*. On peut supposer que le réseau est beaucoup plus abondamment et précisément renseigné concernant l'espèce humaine, que des domaines de spécialité comme la zoologie, et que d'une manière générale, les performances diminuent avec le degré de spécialisation du domaine, en raison d'une moindre quantité d'information dans le réseau.

4. Conclusion

Bien que le taux d'anomalies dans le réseau JDM soit faible, nous pouvons le diminuer encore par des mécanismes endogènes automatiques de détection des relations erronées. Ces mécanismes, relativement simples, permettent de détecter une proportion importante des relations potentiellement erronées « secondaires » inférées à partir de relations « initiales » fausses. Par ailleurs, comme ces relations signalées comme suspectes sont invalidées dès lors qu'un expert a vérifié qu'elles étaient bien erronées, elles ne peuvent plus donner lieu à de nouvelles inférences fausses et donc servir de terreau à la naissance et à la propagation de nouvelles erreurs, ce qui joue aussi en faveur d'une diminution globale du taux d'erreurs du réseau. Les erreurs "initiales" sont en revanche plus délicates à détecter de manière automatique. Cependant, l'exploitation des listes de génériques incompatibles est une piste intéressante dans la mesure où elle permet de donner l'alerte lorsqu'un terme *a priori* monosémique a deux termes incompatibles pour hyperonymes.

Pour conclure, la synergie entre détection manuelle exercée par les joueurs/contributeurs et méthodes de détection automatiques contribue à maintenir dans le réseau JDM un taux d'erreurs raisonnablement bas.

Références

BEN OTHMANE ZRIBI C., MEJRI H., BEN AHMED M. (2007). Un analyseur hybride pour la détection et la correction des erreurs cachées sémantiques en langue arabe, *TALN 2007*, Toulouse, 5–8 juin 2007

BOUDIN F., HERNANDEZ N. (2012). Détection et correction automatique d'erreurs d'annotation morpho-syntaxique du French TreeBank. *TALN 2012*, Juin 2012, Grenoble. pp.281-291.

BOURAOUI J.-L., BOISSIERE P., MOJAHID M., VIGOUROUX N., LAGARRIGUE A., VELLA F., NESPOULOUS J.-L. (2009) Problématique d'analyse et de modélisation des erreurs en production écrite. Approche interdisciplinaire. *TALN 2009*, Senlis, 24-26 juin 2009

DICKINSON M., MEURERS W.D. (2003). Detecting errors in part-of-speech annotation. *EACL 2003 (10th Conference of the European Chapter of the Association for Computational Linguistics)*, pp.107–114, Budapest, Hungary

LAFOURCADE M., ZARROUK M., JOUBERT A. (2014). About Inferences in a Crowdsourced Lexical-Semantic Network, *EACL 2014 (14th Conference of the European Chapter of the Association for Computational Linguistics)*, Gothenburg (Sweden), April 2014

LAFOURCADE M., LE BRUN N., JOUBERT A. (2015). *Games with a Purpose (GWAPS)*, ISBN: 978-1-84821-803-1 July 2015, Wiley-ISTE, 158 p.

Vers l'annotation par le jeu de corpus (plus) complexes : le cas de la langue de spécialité

Karën Fort[1] Bruno Guillaume[2] Nicolas Lefèbvre[2] Laura Ramírez[3]
Mathilde Regnault[3] Mary Collins[3] Oksana Gavrilova[3] Tanti Kristanti[3]
(1) STIH - EA 4509, Université Paris-Sorbonne, France
(2) Sémagramme - Inria Nancy Grand Est, France
(3) Université Paris-Sorbonne, France
(1,3) prenom.nom@paris-sorbonne.fr, (2) prenom.nom@inria.fr

RÉSUMÉ

Nous avons précédemment montré qu'il est possible de faire produire des annotations syntaxiques de qualité par des participants à un jeu ayant un but. Nous présentons ici les résultats d'une expérience visant à évaluer leur production sur un corpus plus complexe, en langue de spécialité, en l'occurrence un corpus de textes scientifiques sur l'ADN. Nous déterminons précisément la complexité de ce corpus, puis nous évaluons les annotations en syntaxe de dépendances produites par les joueurs par rapport à une référence mise au point par des experts du domaine.

ABSTRACT

Towards (more) complex corpora annotation using a game with a purpose : the case of scientific language.

We previously showed that participants in a game with a purpose can produce quality dependency syntax annotations. We present here the results of an experiment aiming at evaluating their production on a more complex corpus of scientific texts on DNA. We precisely describe the complexity of the corpus, then we evaluate the annotations produced by the players as compared to a gold corpus, annotated and adjudicated by experts of the domain.

MOTS-CLÉS : annotation en syntaxe de dépendances, *crowdsourcing*, jeux ayant un but.

KEYWORDS: dependency syntax annotation, crowdsourcing, games with a purpose.

1 Introduction

La création de ressources langagières de qualité, indispensables au développement et à l'évaluation des outils de traitement automatique des langues (TAL), représente un coût élevé. Ce coût est rarement rendu public, mais l'un des rares cas chiffré est le corpus arboré de Prague, dont la construction a été évaluée à 600 000 dollars (Böhmová *et al.*, 2001).

Les jeux ayant un but ont montré leur efficacité dans la production de lexiques et de corpus annotés, en termes à la fois de quantité et de qualité produites (Chamberlain *et al.*, 2013; Lafourcade *et al.*, 2015). Plus intéressant encore, au-delà de la mise en œuvre des connaissances du monde et de la langue des locuteurs, comme dans `JeuxDeMots` (Lafourcade & Joubert, 2008) ou `Phrase Detectives` (Poesio *et al.*, 2013), les jeux permettent de former les participants à des tâches

complexes, par exemple le repliement de protéines dans FoldIt (Khatib *et al.*, 2011) ou l'annotation en syntaxe de dépendances dans ZombiLingo (Guillaume *et al.*, 2016).

Nous souhaitons étudier l'évolution de la qualité de la production des joueurs lorsque la complexité de la tâche est accrue. Pour ce faire, nous nous proposons de faire annoter en syntaxe de dépendances des corpus réputés plus difficiles, notamment : i) un corpus de langue de spécialité, ii) un corpus de parole transcrite et iii) un corpus de contenu généré par l'utilisateur (extrêmement bruité, de type forum). Il n'existe à notre connaissance aucune expérience de ce type relatée dans la littérature, que ce soit pour le français ou pour une autre langue.

Nous présentons ici la première expérience que nous avons menée sur un corpus de langue de spécialité, en l'occurrence des textes portant sur l'ADN, donc touchant au biomédical et aux biotechnologies. Dans un premier temps, nous détaillons la création de ce corpus, puis les conditions de l'expérience, avant de présenter les résultats obtenus et de discuter les biais potentiels de l'étude.

2 Construire un corpus de langue de spécialité

2.1 Sélection des textes

Le corpus utilisé dans cette expérience a été constitué en utilisant le moteur de recherche du site de Creative Commons [1]. Notre but était de créer un corpus i) librement disponible, ii) en langue de spécialité, iii) à partir d'échantillons suffisamment hétérogènes pour représenter la diversité de la langue de spécialité choisie et iv) d'environ 800 phrases, afin d'être comparable aux autres corpus de la plateforme.

Le corpus ADN comprend au final 22 textes en français sur le thème de l'ADN, soit en tout 792 phrases et 25 690 tokens. Il est constitué principalement d'extraits ou de résumés d'articles de recherche (18 textes, soit 21 132 tokens), complétés par deux articles de vulgarisation, un cours et un texte didactique. Le corpus est encodé en UTF-8 : les erreurs d'encodage et les césures ont été corrigées. Il est balisé en XML et librement disponible, sous licence CC BY-NC-SA [2].

2.2 Évaluation de la complexité du corpus

2.2.1 Méthodologie

Si le corpus créé semble d'évidence complexe, encore faut-il s'en assurer et tenter de préciser en quoi. De nombreuses études linguistiques ont porté sur la complexité de la langue, cependant, bien peu fournissent des métriques réellement utilisables dans notre cas, c'est-à-dire sans information sur le temps de lecture (Vasishth, 2003), sur les mouvements oculaires (Lee *et al.*, 2007) et dans un but qui n'est pas de simplifier le texte (Seretan, 2012; Brouwers *et al.*, 2014). Au vu des informations à notre disposition, nous avons décidé de nous concentrer sur la complexité lexicale et syntaxique des phrases en faisant l'hypothèse qu'elles sont un prédicteur suffisamment fiable de la complexité au sens large.

Pour la syntaxe, nous nous sommes largement inspirés de (Blache, 2010). Parmi les mesures listées

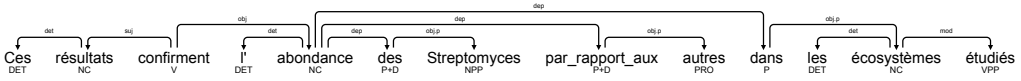
1. Disponible ici : https://search.creativecommons.org/.
2. Voir : https://zombilingo.org/informations#export.

dans cet article, nous nous inspirons de deux qui peuvent être calculées automatiquement à partir de structures en dépendances syntaxiques : i) la profondeur et ii) la longueur moyenne des dépendances. L'idée de considérer la profondeur des structures en dépendances était déjà présente dans les travaux de Gibson (Gibson, 1998). Morrill et Gavarró (Morrill & Gavarró, 2004) ont montré la corrélation de cette mesure avec les difficultés de compréhension liées à l'aphasie.

Nous comparons, selon ces dimensions, le corpus ADN à trois autres corpus utilisés dans le même dispositif, dont un est un extrait du *Corpus journalistique issu de l'Est Républicain* (ATILF, 2011) et deux corpus issus de la Wikipédia. Un de ces deux corpus a pour thème le football, l'autre les Pokémon. Ils ont été mis en jeu à l'occasion d'événements liés à ces sujets (respectivement, Euro 2016 et sortie en France de Pokémon Go).

Techniquement, nous calculons les indices de complexité syntaxique comme suit :

Profondeur. La profondeur quantifie la complexité en suivant l'*hypothèse de la dépendance incomplète*. Pour une phrase donnée, on calcule le nombre maximum de dépendances incomplètes quand on parcourt la phrase. Intuitivement, cela correspond au nombre d'éléments que l'on doit garder en mémoire en lisant la phrase pour pouvoir la comprendre. Pour calculer cette mesure, on ne tient ici pas compte des relations de ponctuation. Ainsi, dans l'analyse en dépendances ci-dessous, la profondeur est de 3 (maximum atteint deux fois entre *abondance* et *des* et entre *des* et *Streptomyces*).



Les phrases sont analysées par deux analyseurs syntaxiques, Talismane (Urieli, 2013) et FrDep-Parse (Guillaume & Perrier, 2015), et nous sélectionnons le maximum des valeurs obtenues sur chacune des deux analyses. Pour finir, nous calculons une moyenne de cette mesure pour l'ensemble des phrases du corpus.

Longueur moyenne. Dans (Blache, 2010), le *coût d'intégration* est défini comme le nombre de référents du discours entre deux têtes rattachées par une dépendance. Ici, nous donnons une approximation de cette mesure en considérant la distance moyenne (en nombre de tokens) entre la tête et le gouverneur pour une relation donnée.

Enfin, pour estimer la complexité lexicale des corpus considérés, nous comptabilisons les tokens n'apparaissant pas dans un lexique standard des formes fléchies du français (*token inconnus*). Nous utilisons pour cela les résultats produits par MElt (Denis & Sagot, 2010), qui préfixe d'une étoile les lemmes quand la forme fléchie correspondante n'existe pas dans le lexique Le*fff* (Sagot, 2010) [3]. Nous calculons également le nombre de tokens inconnus différents.

2.2.2 Analyse des résultats

Les tableaux 1 et 2 présentent les tailles des corpus, ainsi que les résultats des mesures de ces indices de complexité syntaxique et lexicale.

Les indices de complexité syntaxique utilisés montrent que le corpus ADN présente une plus grande profondeur moyenne des dépendances (4,20, Pokémon atteignant 4,06) et des relations pour la plupart plus longues que pour les autres corpus, ou de longueurs proches du maximum (pour A_OBJ

3. Nous avons retiré les cardinaux (en chiffres et en lettres) du décompte, car MElt les préfixe d'une étoile.

	Prof.	Longueur moyenne								
	moy	mod	mod.rel	suj	obj	de_obj	a_obj	p_obj.o	coord	dep.coord
ADN	**4,20**	**3,92**	**7,09**	2,70	**2,90**	**2,70**	2,40	**2,41**	**6,44**	2,23
Foot	3,92	3,30	5,61	**4,19**	2,83	1,50	2,35	1,87	4,59	2,29
Pokémon	4,06	3,43	5,81	3,92	2,42	2,10	1,77	1,76	5,69	**2,94**
Est Rép.	3,40	3,43	5,79	3,55	2,59	2,16	**2,51**	2,00	4,32	2,22

TABLE 1 – Indices de complexité syntaxique.

et DEP.COORD). Seule exception, les sujets, pour lesquels ADN présente la plus petite longueur de relation [4].

	Nb de phrases	Nb de mots	Tokens par phrases	% Tokens inconnus	Nb de tokens inconnus différents
ADN	771	25 160	**32,63**	4,92 %	**624**
Foot	825	21 439	25,99	3,17 %	413
Pokémon	804	24 280	30,20	**6,19 %**	485
Est Rép.	3 536	73 174	20,69	4,55 %	~~3 326~~ 540 (800 phrases)

TABLE 2 – Indices de complexité lexicale.

En ce qui concerne la complexité lexicale, les résultats sont là encore assez nets, avec une longueur moyenne de phrase supérieure pour ADN (32,63 tokens par phrase). La proportion de tokens inconnus par phrase (4,92 %) n'est supérieure que pour Pokémon (6,19 %), ce qui s'explique par le nombre très élevé de noms de Pokémons (inconnus du lexique) dans le corpus. Par ailleurs, le corpus ADN présente le plus grand nombre de tokens inconnus différents (624). Pour le corpus Est Rép. le nombre de tokens inconnus différents est bien supérieur (3 326) du fait de la taille du corpus. Si on sélectionne aléatoirement 800 phrases dans ce corpus pour obtenir une taille comparable, ce nombre devient alors inférieur (540). Ces mesures de complexité confirment donc notre intuition de départ : le corpus ADN est plus complexe que les autres corpus annotés précédemment par le jeu.

3 Conditions de l'expérience

3.1 ZombiLingo

Le jeu utilisé dans cette expérience, ZombiLingo, est présenté en détail dans (Guillaume *et al.*, 2016). Il a permis de faire produire plus de 300 000 annotations en syntaxe de dépendances à plus de 1 000 joueurs [5] sur des corpus en français, qui sont librement disponibles sur le site du jeu [6].

Afin de remobiliser les joueurs, nous lançons régulièrement des défis, lors desquels les participants jouent (donc annotent) pendant deux semaines des corpus sur un thème (le football, les Pokémons, etc.). Un classement spécifique est affiché et une publicité adaptée est réalisée sur les réseaux sociaux.

4. On peut faire l'hypothèse que cela est lié au style scientifique, mais cela reste à vérifier et n'est pas l'objet de notre expérience.

5. Données à la date du 17 avril 2017.

6. Voir : https://zombilingo.org/informations#export.

Ainsi, le défi ADN a eu lieu du 14 au 30 novembre 2016, mobilisant 232 joueurs, qui ont produit 25 039 annotations.

3.2 Évaluation

Un sous-corpus ADN-EVAL du corpus ADN a été constitué pour évaluer les annotations produites par le jeu. ADN-EVAL contient 39 phrases parmi les 792 du corpus ADN, soit 1 245 tokens (environ 5 % du corpus). Ces phrases ont été annotées sur l'outil WebAnno (Yimam *et al.*, 2013) par des étudiantes en Master 2 de linguistique et informatique à l'université Paris-Sorbonne. La majorité a été annotée par deux annotatrices, puis les annotations divergentes ont fait l'objet d'une adjudication par un chercheur spécialiste de la syntaxe (B. Guillaume). Certaines annotations n'ont été réalisées que par une seule annotatrice (par manque de temps) et ont été directement corrigées par ce même chercheur.

Dans le jeu, les corpus de référence sont utilisés de trois façons différentes (voir figure 1) :

— Pour la **formation** : avant de commencer à annoter une relation, un joueur doit suivre une formation et s'entraîner sur des phrases de référence issues du corpus $REF_{Form.\&Ctrl}$ pour lesquelles on lui donne la bonne réponse en cas d'erreur.

— Pour le **contrôle** : en cours de jeu, de temps en temps, une phrase de référence issue du même corpus est proposée au joueur. S'il se trompe, la correction lui est indiquée. Cela permet de contrôler que les joueurs n'ont pas oublié les consignes.

— Pour l'**évaluation** : les phrases d'évaluation sont traitées comme le reste du corpus pendant le déroulement du jeu. C'est *a posteriori* que le corpus ADN-EVAL est utilisé pour l'évaluation.

On peut utiliser la même partie du corpus pour la formation et le contrôle mais bien évidemment, l'évaluation se fait sur une partie différente du corpus de référence.

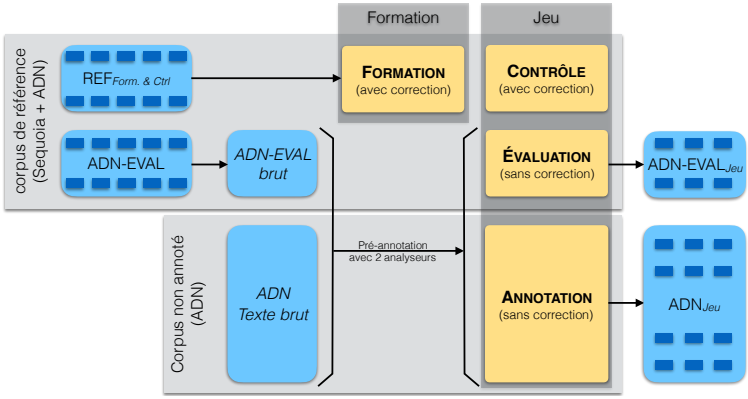


FIGURE 1 – Méthodologie d'évaluation.

	# jouable	densité	FrDep-Parse	Talismane	ZombiLingo
Tout sauf PONCT	292	8,7	0,772	0,858	**0,859**
MOD	155	1,6	0,623	0,759	**0,773**
COORD	28	4,7	0,521	0,474	**0,718**
OBJ	13	13,5	0,857	0,872	**0,901**
OBJ.P	13	17,6	0,931	**0,978**	0,944
DEP.COORD	12	9,5	0,838	0,877	**0,883**
DET	12	31,3	0,946	**0,978**	0,958
P_OBJ.O/P_OBJ.AGT	11	12,0	0,720	**0,815**	0,774

TABLE 3 – F-mesure pour le corpus ADN-EVAL complet, puis par relation.

4 Résultats obtenus

4.1 Annotations réalisées

Le tableau 3 présente le résultat de l'évaluation des phrases annotées par les deux parseurs FrDep-Parse (Guillaume & Perrier, 2015) et Talismane (Urieli, 2013), puis par ZombiLingo. La densité est le nombre moyen de joueurs qui ont donné leur avis par item jouable. Nous observons que le jeu produit des annotations de qualité similaire au parseur Talismane (0,859 de F-mesure). Afin d'interpréter plus finement ce résultat, nous le détaillons par relation (pour les relations où le nombre d'items jouables est supérieur à 10) [7]. Les scores sont plus élevés pour les coordinations, en particulier en ce qui concerne la relation COORD (+0,197) pour la partie gauche de la coordination. On observe également un meilleur score sur la relation MOD, mais la différence (+0,014) n'est pas significative compte-tenu de la taille du corpus de référence. Cela correspond à l'intuition que l'annotation des coordinations ou des rattachements prépositionnels est difficile pour les analyseurs automatiques, moins pour les joueurs.

Dans le format FTB/Sequoia, sur lequel est basé le jeu, la relation MOD est utilisée dans des contextes différents (entre noms et adjectifs, entre verbes et adverbes, entre noms, entre verbes et prépositions, etc.). En revanche, le lien entre noms et prépositions dans le cas d'un groupe prépositionnel modifiant un nom est codé avec la relation DEP [8]. Il est important de noter que la relation MOD représente plus de la moitié des items jouables (155 sur 292), le score sur cette relation représente donc une part importante du score global.

Dans (Guillaume *et al.*, 2016), nous avions utilisé une partie du corpus Sequoia comme corpus d'évaluation. Cette expérience avait montré que, sur un corpus non-spécialisé, la qualité des annotations obtenues avec le jeu était meilleure que celles obtenues avec les analyseurs, dès lors que les annotations étaient jouées par un nombre assez grand de joueurs. Sur un corpus de langue de spécialité, et bien que la densité d'annotation soit nettement supérieure à 1 partout, nous ne retrouvons pas des résultats aussi nets.

7. La relation P_OBJ.AGT n'est pas gérée par Talismane, nous l'avons donc regroupée avec la relation P_OBJ.O pour avoir des données comparables entre les deux parseurs.

8. Dans les données, il y a 13 cas où l'export contient une relation MOD entre un nom et une préposition alors que la bonne solution est la relation DEP. Nous avons considéré ces cas comme corrects car c'est un changement systématique lié à un choix d'annotation dans le format FTB/Sequoia.

4.2 Discussion

L'analyse de la complexité du corpus réalisée pour cette expérience est limitée par les indices dont nous disposions. Les indices choisis nous semblent cependant suffisants pour confirmer l'intuition selon laquelle ce corpus est plus difficile à annoter que les corpus `Wikipédia` utilisés précédemment. Une analyse plus fine de la difficulté d'annotation devrait prendre en compte plus de paramètres et notamment la complexité de la tâche d'annotation telle que définie dans (Fort *et al.*, 2012).

D'autre part, l'expérience présente un biais lié à la petite taille du corpus de référence utilisé (39 phrases, 1 245 tokens). Un travail d'annotation est en cours pour obtenir une référence sur un corpus de plus grande taille pour permettre une analyse plus fiable de la qualité produite.

En ce qui concerne le profil des joueurs, une première étude générale (non spécifique à ce corpus) a été réalisée (Fort *et al.*, 2017), qui montre que ceux-ci ont un niveau d'études élevé, puisque 75 % des gros joueurs ont au moins un niveau Master. Pour autant, ils ne sont probablement pas spécialistes du biomédical ou des biotechnologies (40 % des gros joueurs viennent du domaine du TAL ou de la linguistique).

5 Conclusion

Cette étude présente des résultats partiels concernant l'annotation en syntaxe de dépendances à l'aide d'un jeu ayant un but dans le cas d'un corpus complexe. Si pour l'instant, le jeu n'a pas permis d'obtenir de meilleurs annotations que les analyseurs, on observe que certaines relations comme la coordination sont mieux annotées par les joueurs, ce qui est très prometteur.

Nous pensons cependant qu'il est possible d'aller plus loin et d'améliorer ces résultats, notamment en ce qui concerne les relations MOD et DEP. Plusieurs joueurs nous ont en effet fait part de leur difficulté à jouer la relation MOD. Nous avons donc prévu de modifier la présentation de ces relations dans le jeu pour les rendre plus intuitives pour les joueurs. Nous comptons séparer les différents usages de la relation MOD et regrouper les relations MOD et DEP quand elles introduisent un groupe prépositionnel. Ces modifications sont en cours et une nouvelle version du jeu va voir le jour en juin 2017.

Références

ATILF (2011). Corpus journalistique issu de l'est républicain. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

BLACHE P. (2010). Un modèle de caractérisation de la complexité syntaxique. In *Traitement Automatique des Langues Naturelles*, p. 1–10, Montréal, Canada.

BÖHMOVÁ A., HAJIČ J., HAJIČOVÁ E. & HLADKÁ B. (2001). The prague dependency treebank : Three-level annotation scenario. In A. ABEILLÉ, Ed., *Treebanks : Building and Using Syntactically Annotated Corpora*. Kluwer Academic Publishers.

BROUWERS L., BERNHARD D., LIGOZAT A.-L. & FRANÇOIS T. (2014). Syntactic sentence simplification for french. In *The 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR 2014)*.

CHAMBERLAIN J., FORT K., KRUSCHWITZ U., LAFOURCADE M. & POESIO M. (2013). Using games to create language resources : Successes and limitations of the approach. In I. GUREVYCH & J. KIM, Eds., *The People's Web Meets NLP*, Theory and Applications of Natural Language Processing, p. 3–44. Springer Berlin Heidelberg.

DENIS P. & SAGOT B. (2010). Exploitation d'une ressource lexicale pour la construction d'un étiqueteur morphosyntaxique état-de-l'art du français. In *Traitement Automatique des Langues Naturelles : TALN 2010*, Montréal, Canada.

FORT K., GUILLAUME B. & LEFÈBVRE N. (2017). Who wants to play Zombie ? A survey of the players on ZOMBILINGO. In *Games4NLP 2017 - Using Games and Gamification for Natural Language Processing*, Symposium Games4NLP, Valence, Espagne.

FORT K., NAZARENKO A. & ROSSET S. (2012). Modeling the complexity of manual annotation tasks : a grid of analysis. In *Actes de International Conference on Computational Linguistics (COLING)*, p. 895–910, Mumbai, Inde.

GIBSON E. (1998). Linguistic complexity : locality of syntactic dependencies. *Cognition*, **68**, 1–76.

GUILLAUME B., FORT K. & LEFEBVRE N. (2016). Crowdsourcing complex language resources : Playing to annotate dependency syntax. In *Actes de International Conference on Computational Linguistics (COLING)*, Osaka, Japon.

GUILLAUME B. & PERRIER G. (2015). Dependency Parsing with Graph Rewriting. In *Proceedings of IWPT 2015, 14th International Conference on Parsing Technologies*, p. 30–39, Bilbao, Espagne.

KHATIB F., COOPER S., TYKA M. D., XU K., MAKEDON I., POPOVIĆ Z., BAKER D. & PLAYERS F. (2011). Algorithm discovery by protein folding game players. *Proceedings of the National Academy of Sciences*.

LAFOURCADE M. & JOUBERT A. (2008). JeuxDeMots : un prototype ludique pour l'émergence de relations entre termes. In *Journées internationales d'Analyse statistique des Données Textuelles (JADT)*, Lyon, France.

LAFOURCADE M., LEBRUN N. & JOUBERT A. (2015). *Jeux et intelligence collective : résolution de problèmes et acquisition de données sur le web*. Collection science cognitive et management des connaissances. ISTE.

LEE Y., LEE H. & GORDON P. C. (2007). Linguistic complexity and information structure in korean : Evidence from eye-tracking during reading. *Cognition*, **104**(3).

MORRILL G. & GAVARRÓ A. (2004). On aphasic comprehension and working memory load. In *Categorial Grammars*, p. 259–287, Montpellier, France.

POESIO M., CHAMBERLAIN J., KRUSCHWITZ U., ROBALDO L. & DUCCESCHI L. (2013). Phrase detectives : Utilizing collective intelligence for internet-scale language resource creation. *ACM Trans. Interact. Intell. Syst.*, **3**(1), 3 :1–3 :44.

SAGOT B. (2010). The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. In *7th international conference on Language Resources and Evaluation (LREC 2010)*, La Valette, Malte.

SERETAN V. (2012). Acquisition of syntactic simplification rules for french. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*.

URIELI A. (2013). *Robust French syntax analysis : reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. PhD thesis, Université de Toulouse II le Mirail, France.

VASISHTH S. (2003). Quantifying processing difficulty in human sentence parsing : The role of decay, activation, and similarity-based interference. In *Proceedings of Eurocogsci 03 : The European Cognitive Science Conference*.

YIMAM S. M., GUREVYCH I., ECKART DE CASTILHO R. & BIEMANN C. (2013). Webanno : A flexible, web-based and visually supported system for distributed annotations. In *Actes de Annual Meeting of the Association for Computational Linguistics : System Demonstrations*, p. 1–6, Sofia, Bulgarie : Association for Computational Linguistics.

Détection des mots non-standards dans les tweets avec des réseaux de neurones

Tian Tian[1, 2] Marco Dinarelli[1] Pedro Cardoso[2] Isabelle Tellier[1]
(1) Lattice (UMR 8094), CNRS, ENS Paris, Université Sorbonne Nouvelle,
PSL Research University, USPC, 1 rue Maurice Arnoux, 92120 Montrouge, France
(2) Synthesio, 8-10 rue Villedo, 75001 Paris
ttian@synthesio.com, marco.dinarelli@ens.fr, pedro@synthesio.com,
isabelle.tellier@univ-paris3.fr

RÉSUMÉ

Dans cet article, nous proposons un modèle pour détecter dans les textes générés par des utilisateurs (en particulier les tweets), les mots non-standards à corriger. Nous utilisons pour cela des réseaux de neurones convolutifs au niveau des caractères, associés à des "plongements" (embeddings) des mots présents dans le contexte du mot courant. Nous avons utilisé pour l'évaluation trois corpus de référence. Nous avons testé différents modèles qui varient suivant leurs plongements pré-entrainés, leurs configurations et leurs optimisations. Nous avons finalement obtenu une F1-mesure de 0.972 en validation croisée pour la classe des mots non-standards. Cette détection des mots à corriger est l'étape préliminaire pour la normalisation des textes non standards comme les tweets.

ABSTRACT

Detecting non-standard words in tweets with neural networks

In this paper, we propose a model for detecting non-standard words to be corrected in user-generated texts (such as tweet). We use character-level convolutional neural networks, associated with word embeddings for the current word's context. We conducted our evaluations with three reference corpora. We tested different pretrained word embedding models, different configurations and different optimizations for our various models. We finally obtained a F1-measure of 0.972 in cross validation for the non-standard words class. This non-standard word detection is a first step towards the normalisation of tweets.

MOTS-CLÉS : mots non-standards, réseaux de neurones, modèle convolutionnel, plongements.

KEYWORDS: non-standard words, neural networks, convolutional model, embeddings.

1 Introduction

Les contenus générés par des utilisateurs sur le web, par exemple via les tweets, sont devenus une source de données importante. Ces textes, sous la forme de petits messages d'au maximum 140 caractères, véhiculent souvent des opinions sur l'actualité ou sur des produits de consommation. Ils sont donc particulièrement intéressants à analyser automatiquement dans le cadre de la fouille d'opinion. Cependant, les outils traditionnels de Traitement Automatique des Langues (TAL) se comportent

rarement bien face à ces textes, courts et souvent écrits dans une langue non standard. Par exemple, le Stanford POS Tagger, étiqueteur morpho-syntaxique appris avec le Penn TreeBank [1] (Toutanova & Manning, 2000) atteint 96.86% d'exactitude sur les textes journalistiques, mais seulement 81.3% sur les tweets (Ritter *et al.*, 2011). Ce phénomène se retrouve aussi pour d'autres tâches de TAL, comme l'analyse syntaxique (Plank *et al.*, 2014; Kong *et al.*, 2014). La *normalisation de texte* est la tâche qui consiste à transformer *a minima* un texte "non standard" pour le faire se rapprocher le plus possible d'un texte "standard" sans modifier son sens. C'est devenu une étape de prétraitement nécessaire pour certains textes, avant d'essayer de leur appliquer d'autres étapes de TAL (Han *et al.*, 2013). Elle opère pour l'instant essentiellement au niveau lexical, par l'identifiaction et le remplacement des mots (ou "tokens") non standards. Beaucoup de travaux montrent qu'une telle normalisation appliquée sur des tweets peut améliorer le résultat de l'extraction des entités nommées qu'ils contiennent (Nguyen *et al.*, 2016; Liu *et al.*, 2012). Dans certains cas, la détection des unités non-standards peut à elle seule améliorer ce résultat (Li & Liu, 2015).

Notre travail dans cet article s'inspire de certaines expériences de (Li & Liu, 2015). Il vise au final à améliorer l'extraction des entités nommées dans les tweets après une phase de normalisation. Nous nous focalisons en particulier sur le pré-traitement lexical permettant de rendre les tweets un peu mieux formés, en remplaçant les mots non-standards qui y figurent par une forme standard. Nous essayons ici, dans un premier temps, de détecter dans un tweet ces mots non-standards avant de proposer une liste de candidats pour les corriger.

Nous proposons pour cela un modèle de réseau de neurones convolutifs au niveau des caractères, associé à des "plongements" de mots, pour déterminer si un mot est non-standard (à corriger). Nous avons utilisé pour nous évaluer des corpus de tweets en anglais dans lesquels les mots non standards ont été manuellement corrigés et avons mené diverses expériences avec divers protocoles (en validation croisée, ou avec un autre corpus d'évaluation).

Cet article est organisé de la manière suivante :

— la section 2 détaille la tâche de normalisation des tweets ainsi que l'état de l'art ;
— la section 3 présente les corpus annotés utilisés dans cet article ;
— la section 4 explique en détail le modèle de réseau de neurones convolutifs utilisé ;
— la section 5 contient les résultats obtenus avec les différentes configurations ;
— la section 6 conclut sur ces expériences et évoque quelques perspectives pour l'avenir.

2 Etat de l'art

2.1 Notion de mots non-standard

Les mots non-standards (non-standard words, NSW) sont des mots à corriger dans le contexte où ils apparaissent par un mot présent dans un vocabulaire. Ils peuvent contenir des fautes d'orthographe (comme "dis" pour "this" en anglais), des acronymes (comme "lol" pour "laughing out loud"), des abréviations (comme "u" pour "you"), etc. Cette notion de NSW se distingue de celle de mots hors vocabulaire (out-of-vocabulary, OOV). En effet, les OOV sont des mots absents d'un dictionnaire de référence (par exemple celui de GNU Aspell pour l'anglais [2]), mais ne sont pas nécessairement pour autant mal écrits et à corriger : les entités nommées ou noms de produits sont souvent OOV.

1. Stanford Pos Tagger : http ://nlp.stanford.edu/software/tagger.shtml
2. GNU Aspell : http ://aspell.net/

Ainsi, un mot comme "xanax" (nom d'un médicament contre l'anxiété), ne se trouve pas dans les dictionnaires. Ce n'est cependant pas un NSW parce qu'il n'y a pas moyen de le corriger, il est déjà sous sa forme correcte. Les NSW des exemples précédents sont des OOV, mais il y a aussi des NSW qui figurent dans un vocabulaire, mais ne conviennent pas dans leur contexte, par exemple "wit" dans la phrase "I'll go wit you" (la forme correcte est "with") (Han, 2014). Le contexte joue donc un rôle très important pour ranger les mots dans ces différentes classes. Enfin, du bruit peut aussi exister dans les tweets, comme "bahahahaha" à la fin d'un message, pour exprimer un sentiment. Ce token est OOV parce qu'il n'est pas présent dans les dictionnaires usuels, mais il n'existe pas non plus de forme correcte correspondante. Il ne s'agit donc pas d'un NSW. Certaines fois, il est difficile de juger si un token est NSW ou SW comme "footie", "y'all" et "yous" (Baldwin *et al.*, 2015).

2.2 Normalisation des tweets

La normalisation lexicale de texte peut se décomposer en trois sous-tâches :

1. Trouver les NSW de ce texte. Dans un corpus de tweets, le taux de NSW est d'environ 10%. Cette sous-tâche se traduit souvent par une classification binaire (même si, comme nous venons de le voir, la situation est souvent plus compliquée à cause du bruit). Il s'agit donc à cette étape de décider, pour chaque token d'un texte, s'il est à corriger ou pas.
2. Pour un NSW dans un texte, proposer une liste de candidats pour le corriger. Si un token peut être considéré comme candidat à sa propre correction, alors cette sous-tâche peut être exécutée pour tous les tokens d'un texte, sans passer par la première sous-tâche comme dans (Jin, 2015) et (Mausam *et al.*, 2012). (Han & Baldwin, 2011) a proposé de générer les corrections des NSW en se basant sur les similarités morpho-phonétiques. (Supranovich & Patsepnia, 2015) a, lui, fait appel à la distance d'édition (ou distance de Levenshtein) (Miller *et al.*, 2009).
3. Pour un NSW et une liste de candidats de correction proposée, trouver le candidat qui a la plus forte probabilité d'être la forme correcte du NSW. Cela peut être réalisé en calculant des mesures de similarité entre le NSW et chacun de ses candidats comme dans (Jin, 2015), ou par la distance d'édition (Miller *et al.*, 2009) comme (Supranovich & Patsepnia, 2015).

Pour la première sous-tâche, (Han & Baldwin, 2011; Supranovich & Patsepnia, 2015; Leeman-Munk *et al.*, 2015) ont utilisé un classifieur pour détecter les NSW. (Supranovich & Patsepnia, 2015) a utilisé les champs conditionnels aléatoires (Conditional Random Fields, CRF) tandis que (Leeman-Munk *et al.*, 2015) ont utilisé un réseau de neurones pour le faire. (Berend & Tasnádi, 2015) a quant à lui exploité les CRF pour la classification et la correction des NSW.

La correction de tweets a fait l'objet d'un challenge dans le workshop WNUT (Workshop on Noisy User-generated Text (Baldwin *et al.*, 2015)). Les meilleurs résultats ont été obtenus par (Leeman-Munk *et al.*, 2015), qui ont atteint un résultat de $0,98$ de F1-mesure pour la classe NSW sur le corpus de développement du workshop, et le même résultat sur le corpus d'évaluation. Dans cet article, nous essayons de reprendre la méthode qu'ils ont utilisée pour détecter dans des tweets les mots non-standards (NSW), considéré comme une tâche de classification binaire.

3 Les corpus annotés

Nous avons utilisé trois corpus annotés : 2577 tweets provenant de (Li & Liu, 2014) d'une part, les corpus de développement et d'évaluation du workshop WNUT d'autre part (Han & Baldwin, 2011;

Baldwin *et al.*, 2015). Le tableau 1 fournit les statistiques de base de ces trois corpus.

nom de corpus	nombre de tweets	# tokens	# NSW	# SW	taille de vocabulaire
2577 tweets	2577	44385	3942	40443	8270
WNUT dev	2950	40560	4274	36286	15108
WNUT eval	1698	29421	2776	26645	10765

TABLE 1 – Corpus statistiques

Ces trois corpus sont tokenisés puis annotés avec la correction manuelle de NSW. Pour la convenance des annotateurs, les annotations (les corrections de tokens) sont en minuscules. Nous avons donc utilisé uniquement l'annotation pour distinguer les NSW des SW pour notre tâche de détection des NSW : si la forme de surface d'un token dans le corpus est différente de sa correction (la différence de majuscules et de minuscules sont ignorées), ce token est considéré comme un NSW.

D'après les statistiques du tableau 1, les mots non-standards sont minoritaires, ils constituent respectivement 8.9%, 10.5% et 9.4% du nombre total des tokens de ces trois corpus.

4 Les modèles de réseaux de neurones convolutifs

Les modèles de réseaux de neurones convolutifs (CNN) introduits dans (Collobert *et al.*, 2011) sont utilisés pour extraire les propriétés de haut niveaux des vecteurs de mots. Dans (Ma & Hovy, 2016), un modèle convolutif similaire au niveau des caractères est aussi utilisé pour représenter des propriétés plus fines. Dans cet article, nous proposons un modèle de réseau de neurones convolutifs au niveau des caractères, associé à des plongements. La figure 1 montre le détail de la représentation du mot "Mercure" par la convolution des caractères comme dans (Ma & Hovy, 2016).

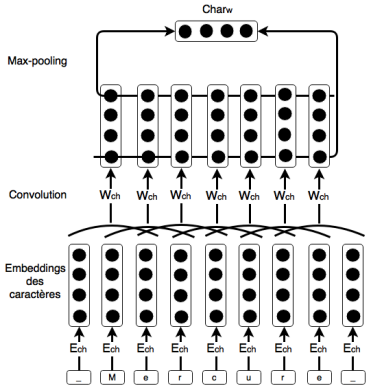


FIGURE 1 – structure du CNN sur les caractères utilisé

Dans cette figure, les E_{ch} désignent les plongements (embeddings) des caractères (un vecteur pour

représenter un caractère, de taille quatre sur le dessin). Les symboles "_" avant le "M" et après le "e" du mot servent à le compléter pour la fenêtre de taille trois autour d'un caractère. Les W_{ch} sont les poids de la matrice correspondant à un triplet de caractères. Enfin le max-pooling prend uniquement la valeur la plus élevée de chaque composante du vecteur pour arriver à une représentation du mot "Mercure" ($Char_w$). Dans notre modèle, la taille de ce vecteur est en réalité 300 (comme la configuration de modèle word2vec de Google).

Nous avons ensuite ajouté à la représentation du mot cible ainsi obtenue son contexte dans une fenêtre couvrant deux mots avant et deux mots après lui, représentés avec des plongements de mots classiques ($Emb(W_i)$, décrits plus loin). Les débuts et la fin de tweets sont marqués par <s> <s> et </s> </s> pour constituer une suite de mots de taille constante (5-grammes). Ainsi, un tweet de N mots sera présenté par N 5-grammes. Par exemple, pour un tweet avec six tokens : "I txtd you last night." il y a six 5-grammes qui représentent cette séquence de texte : "<s> <s> I txtd you", "<s> I txtd you last", "I txtd you last night", "txtd you last night .", "you last night . </s>" et "last night . </s> </s>". La structure complète du modèle est illustrée dans le schéma 2.

Emb(W_{-2})	Emb(W_{-1})	Emb(W_0)	Emb(W_1)	Emb(W_2)	représentation du mot
Dropout(0.2)					
Couche cachée (dimension = 128, activation = ReLU)					
Dropout(0.2)					
Couche de sortie (dimension = 2, activation = softmax), loss = categorical_crossentropy					

TABLE 2 – Structure du réseau

5 Experiences et résultats

Notre baseline est une méthode simple par dictionaire induit sur les données d'entraînement : pour chaque token des données d'apprentissage, s'il est au moins une fois marqué comme NSW, il est stocké dans un ensemble (la casse est toujours ignorée comme convenu). Dans la phase de test, si le token est dans cet ensemble, il sera taggé comme NSW, et comme SW sinon. Cette baseline reconnaît donc uniquement les NSW déjà présents dans les données d'apprentissage ; les autres tokens des données de test seront taggés comme SW.

Nous avons procédé à diverses expériences avec notre réseau de neurones convolutif. Le tableau 3 montre les résultats de la validation croisée en 5 blocs avec le corpus WNUT dev et avec le corpus WNUT dev plus le corpus de 2577 tweets. Le tableau 4 montre les résultats d'évaluation obtenus sur WNUT test par un modèle appris sur WNUT dev et sur WNUT dev plus 2577 tweets. Ces derniers résultats sont donc comparables avec les travaux de (Leeman-Munk *et al.*, 2015).

Les conditions d'expérimentation sont les suivantes :

— Logiciel d'implémentation : Keras [3], Theano [4]

— Initialisation des paramètres : Xavier initialisation (Glorot & Bengio, 2010)

3. Keras : https ://keras.io/

4. Theano : http ://deeplearning.net/software/theano/

— Plongements pré-entrainés des mots ($Emb(W_i)$) : Google word2vec (Mikolov *et al.*, 2013) versus *interne*.

Le modèle *interne* est appris avec word2vec sur des textes plus longs et mieux formés que les tweets (news, forums, blogs, etc.) mais portant sur des domaines similaires (produits de luxe, automobiles et musique). La taille du vocabulaire de ce modèle est d'environ deux millions de tokens différents, tandis qu'il y a trois millions de tokens différents dans le modèle word2vec de Google. Certaines entités nommées comme les dates, heures, unités de longueur, de volumes, etc. sont remplacées par un token artificiel (par exemple "__DATE__" pour les dates). Les tokens qui ne varient que suivant la casse sont aussi normalisés par leur forme la plus courante. Par exemple, le token "nissan" apparaît sous trois formes : tout en minuscule, tout en majuscule et "Nissan". La forme "Nissan" a le plus grand nombre d'occurrences dans le corpus *interne*, les formes "NISSAN" et "nissan" sont donc remplacées par "Nissan". Le modèle de word2vec *interne* est ainsi appris sur un vocabulaire restreint. En revanche, le modèle word2vec de Google (Mikolov *et al.*, 2013) a gardé toutes les formes pour tous les tokens. Donc les tokens "NISSAN", "Nissan" et "nissan" y ont des représentations vectorielles différentes (parce que leurs contextes sont aussi différents).

Nous avons testé les configurations suivantes :

— Corpus d'apprentissage : WNUT dev versus WNUT dev + 2577 tweets

— Optimisers : RMS prop (Tieleman & Hinton, 2012), adadelta (Zeiler, 2012), adagrad (Duchi *et al.*, 2011), sgd (Bergstra & Bengio, 2012)

Dans le tableau 3, les expériences en validation croisée avec 5 plis sont effectuées d'abord avec seulement le corpus de WNUT dev en premières lignes, puis sur les corpus 2577 tweets et WNUT dev mélangés (les dernières lignes). Pour les expériences sur le corpus WNUT eval, le modèle est appris avec seulement le corpus de WNUT dev pour les premières lignes, puis avec les corpus 2577 tweets et WNUT dev ensemble.

Training data	Word2vec	Optimizer name	précision NSW	rappel NSW	f1 measure NSW
WNUT dev	baseline		0.6382	0.5733	0.6092
	google	rmsprop	0.8859	0.7721	0.8242
		adadelta	0.8813	0.7976	0.8372
		adagrad	0.8941	0.7819	0.8341
		sgd	0.8823	0.8127	0.8459
	interne	rmsprop	0.9047	0.8013	0.8496
		adadelta	0.911	0.8022	0.853
		adagrad	0.9157	0.8062	0.8574
		sgd	0.9022	0.824	0.8612
WNUT dev+2577	baseline		0.3012	0.5388	0.3864
	google	rmsprop	0.9727	0.9501	0.9613
		adadelta	0.9813	0.9576	0.9693
		adagrad	0.9791	0.9566	0.9677
		sgd	0.9812	0.9597	0.9703
	interne	rmsprop	0.9808	0.9545	0.9674
		adadelta	0.9841	0.9620	**0.9729**
		adagrad	0.9876	0.9583	0.9727
		sgd	0.9853	0.9602	0.9725

TABLE 3 – Résultat en validation croisée

Nous pouvons observer dans ces résultats :

1. sans surprise, plus le corpus d'entraînement est grand, meilleurs sont les résultats : WNUT

Training data	Word2vec	Optimizer name	précision NSW	rappel NSW	f1 measure NSW
WNUT dev	baseline		0.3999	0.3542	0.3757
	(Leeman-Munk *et al.*, 2015)		0.9783	0.9930	0.9736
	google	rmsprop	0.7385	0.7965	0.7664
		adadelta	0.6952	0.8307	0.7569
		adagrad	0.7835	0.7954	0.7894
		sgd	0.7964	0.8188	0.8075
	interne	rmsprop	0.6023	0.8260	0.6966
		adadelta	0.7961	0.7821	0.7890
		adagrad	0.9157	0.8062	0.8574
		sgd	0.8931	0.7673	0.8254
WNUT dev+2577	baseline		0.3943	0.6459	0.4897
	google	rmsprop	0.6160	0.8606	0.7181
		adadelta	0.7000	0.8455	0.7659
		adagrad	0.8139	0.7893	0.8014
		sgd	0.7859	0.8184	0.8018
	interne	rmsprop	0.6254	0.8289	0.7129
		adadelta	0.8257	0.7849	0.8048
		adagrad	0.9038	0.7478	0.8145
		sgd	0.8769	0.7648	**0.8170**

TABLE 4 – Résultat d'évaluation

dev seul produit des modèles moins performants que WNUT dev+2577 tweets.

2. plongements : les plongements de Google sont moins efficaces que ceux développés en *interne* (nous avons aussi testé sans plongements pré-entrainés, le résultat n'est pas meilleur)
3. optimiser : le meilleur optimiser varie selon les expériences. En général, RMS prop donne le résultat le moins bon. Le sgd donne souvent un meilleur résultat.
(nous avons testé aussi decay=learn rate/nb d'epochs, le résultat n'est pas meilleur)

6 Conclusion et perspectives du travail

Nous proposons dans cet article un modèle de réseau de neurones convolutif pour détecter dans des tweets les mots non-standards (NSW) à corriger. Nos expériences montrent que la quantité de données d'apprentissage influe sensiblement sur le résultat. Le modèle word2vec *interne* et l'optimiser SGD produisent le meilleur classifeur de NSW, avec une F1-mesure d'environ $0,83$ pour cette classe sur le corpus d'évaluation. Ce travail servira de base pour une étape de normalisation ultérieure des tweets.

Références

BALDWIN T., DE MARNEFFE M.-C., HAN B., KIM Y.-B., RITTER A. & XU W. (2015). Shared tasks of the 2015 workshop on noisy user-generated text : Twitter lexical normalization and named entity recognition. In *Proceedings of the Workshop on Noisy User-generated Text*, p. 126–135, Beijing, China : Association for Computational Linguistics.

BEREND G. & TASNÁDI E. (2015). Uszeged : Correction type-sensitive normalization of english tweets using efficiently indexed n-gram statistics. In *Proceedings of the Workshop on Noisy User-generated Text*, p. 120–125, Beijing, China : Association for Computational Linguistics.

BERGSTRA J. & BENGIO Y. (2012). Random search for hyper-parameter optimization. *J. Mach. Learn. Res.*, **13**, 281–305.

COLLOBERT R., WESTON J., BOTTOU L., KARLEN M., KAVUKCUOGLU K. & KUKSA P. (2011). Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, **12**, 2493–2537.

DUCHI J., HAZAN E. & SINGER Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.*, **12**, 2121–2159.

GLOROT X. & BENGIO Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *In Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS'10). Society for Artificial Intelligence and Statistics*.

HAN B. (2014). *Improving the utility of social media with Natural Language Processing*. PhD thesis, The University of Melbourne.

HAN B. & BALDWIN T. (2011). Lexical normalisation of short text messages : Makn sens a #twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies - Volume 1*, HLT '11, p. 368–378, Stroudsburg, PA, USA : Association for Computational Linguistics.

HAN B., COOK P. & BALDWIN T. (2013). Lexical normalization for social media text. *ACM Trans. Intell. Syst. Technol.*, **4**(1), 5 :1–5 :27.

IOFFE S. & SZEGEDY C. (2015). Batch normalization : Accelerating deep network training by reducing internal covariate shift. In F. R. BACH & D. M. BLEI, Eds., *ICML*, volume 37 of *JMLR Workshop and Conference Proceedings*, p. 448–456 : JMLR.org.

JIN N. (2015). Ncsu-sas-ning : Candidate generation and feature engineering for supervised lexical normalization. In *Proceedings of the Workshop on Noisy User-generated Text*, p. 87–92, Beijing, China : Association for Computational Linguistics.

KONG L., SCHNEIDER N., SWAYAMDIPTA S., BHATIA A., DYER C. & SMITH N. A. (2014). A dependency parser for tweets. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 1001–1012, Doha, Qatar : Association for Computational Linguistics.

LEEMAN-MUNK S., LESTER J. & COX J. (2015). Ncsu_sas_sam : Deep encoding and reconstruction for normalization of noisy text. In *Proceedings of the Workshop on Noisy User-generated Text*, p. 154–161, Beijing, China : Association for Computational Linguistics.

LI C. & LIU Y. (2014). Improving text normalization via unsupervised model and discriminative reranking. In *ACL (Student Research Workshop)*, p. 86–93.

LI C. & LIU Y. (2015). *Improving named entity recognition in tweets via detecting non-standard words*, In *ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics*

and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference, volume 1, p. 929–938. Association for Computational Linguistics (ACL).

LIU F., WENG F. & JIANG X. (2012). A broad-coverage normalization system for social media language. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics : Long Papers - Volume 1*, ACL '12, p. 1035–1044, Stroudsburg, PA, USA : Association for Computational Linguistics.

MA X. & HOVY E. (2016). End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016*.

MAUSAM, SCHMITZ M., BART R., SODERLAND S. & ETZIONI O. (2012). Open language learning for information extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, p. 523–534, Stroudsburg, PA, USA : Association for Computational Linguistics.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013). Efficient estimation of word representations in vector space. *CoRR*, **abs/1301.3781**.

MILLER F. P., VANDOME A. F. & MCBREWSTER J. (2009). *Levenshtein Distance : Information Theory, Computer Science, String (Computer Science), String Metric, Damerau ?Levenshtein Distance, Spell Checker, Hamming Distance*. Alpha Press.

NGUYEN V. H., NGUYEN H. T. & SNASEL V. (2016). Text normalization for named entity recognition in vietnamese tweets. *Computational Social Networks*, **3**(1), 10.

PLANK B., HOVY D., MCDONALD R. T. & SØGAARD A. (2014). Adapting taggers to twitter with not-so-distant supervision. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference : Technical Papers, August 23-29, 2014, Dublin, Ireland*, p. 1783–1792.

RITTER A., CLARK S., MAUSAM & ETZIONI O. (2011). Named entity recognition in tweets : An experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, p. 1524–1534, Stroudsburg, PA, USA : Association for Computational Linguistics.

SUPRANOVICH D. & PATSEPNIA V. (2015). Ihs_rd : Lexical normalization for english tweets. In *Proceedings of the Workshop on Noisy User-generated Text*, p. 78–81, Beijing, China : Association for Computational Linguistics.

TIELEMAN T. & HINTON G. (2012). Lecture 6.5—RmsProp : Divide the gradient by a running average of its recent magnitude. COURSERA : Neural Networks for Machine Learning.

TOUTANOVA K. & MANNING C. D. (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora : Held in Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics - Volume 13*, EMNLP '00, p. 63–70, Stroudsburg, PA, USA : Association for Computational Linguistics.

ZEILER M. D. (2012). ADADELTA : an adaptive learning rate method. *CoRR*, **abs/1212.5701**.

MAR-REL : une base de marqueurs de relations conceptuelles pour la détection de Contextes Riches en Connaissances

Luce Lefeuvre Anne Condamines
CLLE-ERSS, Toulouse, France

RÉSUMÉ

Les marqueurs de relation conceptuelle sont un moyen efficace de détecter automatiquement en corpus des *Contextes Riches en Connaissances*. Dans le domaine de la terminologie ou de l'ingénierie des connaissances, les *Contextes Riches en Connaissances* peuvent être utiles pour l'élaboration de ressources termino-ontologiques. Bien que la littérature concernant ce sujet soit riche, il n'existe pas de recensement systématique ni d'évaluation à grande échelle des marqueurs de relation conceptuelle. Pour ces raisons notamment, nous avons constitué une base de marqueurs pour les relations d'hyperonymie, de méronymie, et de cause, en français. Pour chacun de ces marqueurs, son taux de précision est proposé pour des corpus qui varient en fonction du domaine et du genre textuel.

ABSTRACT

MAR-REL : a conceptual relation markers database for Knowledge-Rich Contexts extraction
Markers of conceptual relation are an efficient means to automatically retrieve *Knowledge-Rich Contexts* from corpora. Those contexts can be useful when building terminological or ontological resources. Although this subject has been actively researched, there exists neither an inventory nor a broad study of the markers of conceptual relation. For those reasons, we built a French markers database for the hyperonym, the meronym and the causal relations. Each marker is associated with his accuracy rate depending on domain and text genre.

MOTS-CLÉS : Marqueur de relation conceptuelle, Contextes Riches en Connaissances, variation, domaine, genre textuel, ressources terminologiques, corpus spécialisés.

KEYWORDS: Marker of conceptual relation, knowledge-rich context, variation, domain, text genre, terminological resources, specialized corpora.

1 Introduction

En terminologie ou en ingénierie des connaissances, les linguistes s'appuient sur des corpus spécialisés pour constituer des ressources termino-ontologiques ou des réseaux de termes. Des outils informatiques peuvent aider, accompagner l'exploration de ces corpus, en détectant notamment des portions de textes supposées « riches en connaissances » car contenant des informations conceptuelles sur les termes. Ce type de portion de texte, ou *Contexte Riche en Connaissances* (Meyer, 2001) est défini comme un « *context indicating at least one item of domain knowledge that could be useful for conceptual analysis* ». Il est considéré comme pertinent car il peut constituer le point de départ de définitions, ou améliorer la connaissance d'un domaine. L'une des méthodes pour

repérer ce type de contexte consiste à définir des marqueurs de relations conceptuelles. Utilisés en ingénierie des connaissances, en lexicographie, ou en terminologie, ces éléments linguistiques permettent de repérer des relations existant entre deux termes.

Dans le cadre du projet ANR CRISTAL[1] (Contextes RIches en connaissanceS pour la TrAduction terminoLogique), nous avons constitué une base de marqueurs français permettant d'identifier en corpus spécialisé les relations d'hyperonymie, de méronymie et de cause. Afin de caractériser leur fonctionnement linguistique, nous avons mené une analyse à grande échelle en corpus variant du point de vue du domaine (Volcanologie vs. Cancer du sein) et du genre textuel (scientifique vs. vulgarisé). Les résultats obtenus à la suite de cette analyse, en particulier la précision, sont associés à chaque candidat-marqueur de la base, et constituent une indication quant à la capacité du candidat-marqueur à indiquer la relation attendue.

Nous définissons en section 2 les marqueurs de relation conceptuelle, en nous appuyant sur les travaux existants. En section 3, nous exposons notre méthodologie de recensement et d'évaluation des candidats-marqueurs de relation. Enfin, nous présentons les résultats obtenus en section 4, et discutons des suites possibles de ce travail en section 5.

2 Contextes Riches en Connaissances et marqueurs de relations conceptuelles

La notion de *Contexte Riche en Connaissances* fait souvent écho à la notion de marqueurs de relation conceptuelle, utilisés en intelligence artificielle pour l'élaboration de ressources terminologiques notamment. Constitués d'éléments lexico-syntaxiques, voire typographiques ou dispositionnels (Auger, Barrière, 2008), ils mettent en évidence le lien existant entre deux termes, et sont généralement représentés sous la forme d'un triplet « Terme 1 - Marqueur - Terme 2 », dans lequel le marqueur exprime la relation existant entre les deux termes. Par exemple, la relation d'hyperonymie (générique - spécifique) peut être indiquée par le marqueur **[X est un Y + caractéristiques]** (« *Le cancer est une maladie caractérisée par la prolifération incontrôlée de cellules* ») ; et la relation de méronymie (ou partie - tout) peut être indiquée par le marqueur **[X être {formé/constitué} de DET Y]** (« *le volcan primitif est en majorité constitué de coulées d'andésites* »). Bien que potentiellement, il existe une multitude de relations pertinentes à étudier, les marqueurs examinés concernent principalement trois relations : l'hyperonymie, la méronymie, et la cause (Marshman et al. 2012 ; Nuopponen, 2014). Considérées comme structurantes et universelles, ces trois relations apportent des éléments de connaissance sur les termes d'un domaine.

De nombreux travaux s'attachent ainsi à décrire les marqueurs de ces trois relations. Dans le domaine de la lexicologie, des recherches ont été menées autour des différentes formes de la définition (Flowerdew, 1992 ; Pearson, 1996 ; Rebeyrolle, 2000). Dans le domaine du Traitement Automatique des Langues, les travaux de Hearst (1992) ont servi de modèle pour automatiser la détection de la relation d'hyperonymie, et donc pour construire *in fine* des taxonomies ou des ontologies de manière automatique ou semi-automatique (Morin, 1999, Malaisé et al., 2004, entre autre). Les principes méthodologiques de ces travaux ont par ailleurs été repris dans de nombreux travaux s'intéressant à la mise en œuvre automatique des marqueurs de relation (Aussenac-Gilles, Séguéla, 2000 ; Condamines, Rebeyrolle, 2001, Lafourcade, Ramadier, 2016). Concernant la relation de méronymie, Girju et son équipe (2003), dans le domaine de l'extraction d'information,

[1] Projet ANR CRISTAL [ANR-12-CORD-0020], http://www.projet-cristal.org

se sont penchés sur l'expression de la méronymie en anglais et sur des méthodes efficaces pour extraire des contextes contenant cette relation. Enfin, certains chercheurs ont souligné l'intérêt d'étudier les énoncés causaux en discours et la pertinence d'intégrer la relation de cause aux ressources terminologiques (Garcia, 1998 ; Barrière, 2001).

D'autres travaux, plus rares, s'intéressent à la variation de ces marqueurs selon le genre textuel, le domaine, ou la langue (Pearson, 1998 ; Condamines, 2000, 2006, 2008 ; Marshman, 2008 ; Marshman, L'Homme, 2008). Ces travaux montrent que la productivité et la répartition des marqueurs varie parfois fortement d'un domaine ou d'un genre à l'autre. Ils soulignent la nécessité de prendre en compte la variation dans la description des marqueurs de relation, afin d'en étudier la « portabilité » d'un corpus à un autre (Marshman, L'Homme, 2006).

Bien que la littérature sur ce sujet soit abondante, il n'existe pas de base de données recensant l'ensemble des marqueurs des relations d'hyperonymie, de méronymie et de cause, ni d'analyse systématique à grande échelle de ces marqueurs. Nous avons donc constitué cette base de données et avons analysé le fonctionnement de chaque candidat-marqueur dans des corpus variant en fonction du domaine et du genre.

3 Recensement et évaluation des marqueurs de relation

Notre travail s'est déroulé en deux étapes :

1. Élaboration de la liste des candidats-marqueurs en français et en anglais pour les relations d'hyperonymie, de méronymie et de cause,
2. Analyse des occurrences des candidats-marqueurs français en corpus.

Nous détaillons dans la suite chacune de ces étapes.

3.1 Recensement des candidats-marqueurs de relation

D'une manière générale, nous avons procédé de la même façon pour élaborer les listes des marqueurs des trois relations étudiées. La base de marqueurs de relation a été construite en trois phases :

1. À partir des travaux existants et dans la lignée des travaux mentionnés en section 2, nous avons relevé les marqueurs français des relations d'hyperonymie, de méronymie, et de cause,
2. Nous avons ensuite évalué la première liste obtenue, sur la base de l'introspection, afin d'y ajouter des modifications si nécessaire,
3. Enfin, nous avons uniformisé le codage linguistique des candidats-marqueurs retenus.

La liste finale résulte à la fois d'un travail introspectif et de données issues de l'observation de corpus. Le tableau suivant recense le nombre de candidats-marqueurs obtenus pour chaque relation :

Relation conceptuelle	Nombre de candidats-marqueurs recensés
Hyperonymie	33
Méronymie	95
Cause	192

TABLE 1. Nombre de candidats-marqueurs recensés par relation conceptuelle.

3.2 Évaluation en corpus

La seconde phase de notre travail a concerné l'analyse à grande échelle des candidats-marqueurs, en prenant en compte les différents paramètres de variation que nous avons listés plus haut. Notre corpus traite ainsi de deux domaines : la volcanologie, qui appartient aux Sciences de la Terre, et le cancer du sein chez la femme, qui appartient aux Sciences du Vivant. Ces deux domaines scientifiques sont très spécialisés et font l'objet de beaucoup de vulgarisation. Nous avons donc pu constituer pour chaque domaine une partie scientifique et une partie vulgarisée. Les corpus scientifiques sont constitués de textes issus de revues spécialisées, écrits par des experts à destination d'experts du domaine ou de domaines connexes. Les corpus vulgarisés sont constitués de textes issus de revues ou de sites internet de vulgarisation ; ils sont écrits par des experts du domaine ou par des auteurs ayant des connaissances avancées dans le domaine et sont à destination du grand public. Le tableau 2 ci-dessous synthétise ces informations.

	Cancer du sein	Volcanologie
Corpus scientifique	200 000 mots	400 000 mots
	2002 – 2008	1980 - 2012
Corpus vulgarisé	200 000 mots	400 000 mots
	2001 - 2008	1980 - 2002

TABLE 2. Constitution du corpus

Nous avons extrait de ce corpus les contextes comportant les candidats-marqueurs recensés. Pour chaque candidat-marqueur de chaque relation, nous avons annoté le contexte comme suit :

- « Oui » : la relation est présente

« *Un dynamisme explosif, extrusif et / ou intrusif* ***a généré*** *des cônes stromboliens, des necks basaltiques* » (volcanologie, scientifique).
Le candidat-marqueur « Det X générer Det Y » lie les termes « *dynamisme explosif, extrusif et / ou intrusif* » d'une part et « *cônes stromboliens* » et « *necks basaltiques* » d'autre part par la relation de cause.

- « Non » : le candidat-marqueur n'indique pas la relation attendue

« *Mais notre but* ***est un*** *autre volcan* ***très*** *actif et dangereux* » (volcanologie, vulgarisation)
Le candidat-marqueur testé « Y être DET X très Adj » n'indique pas la relation d'hyperonymie attendue entre « *but* » et « *volcan* ».

- « Plutôt oui » : le candidat-marqueur exprime la relation conjointement avec un autre élément.

« *Trop de repos ou un manque d'activité peuvent* ***diminuer*** *l'oxygénation des tissus musculaires* » (cancer du sein, vulgarisation)
La nominalisation « oxygénation » associée au candidat-marqueur « diminuer » nous permet d'interpréter la relation comme causale. Deux éléments du triplet sont ainsi présents.

- « Plutôt non »: la relation est difficile à interpréter ; ou alors les éléments en relation ne nous intéressent pas dans l'optique de construire des ressources termino-ontologiques (ce ne sont pas des termes du domaine par exemple).

« *Cette découverte* ***motive*** *son élection à l'Académie des sciences* » Relation de cause (volcanologie, vulgarisation)
Il ne nous semble pas pertinent d'intégrer les éléments en relation à une ressource terminologique liée au domaine de la volcanologie.

- « Indéterminé »: nous ne pouvons évaluer la relation (par manque d'indices linguistiques ou par manque de connaissances sur le domaine).

« *Hormones hypophysaires : Ce sont des hormones* ***sécrétées par*** *l'hypophyse, glande cérébrale située juste sous le cerveau* » (cancer du sein, vulgarisation)
Les candidats-termes « *hormones* » et « *hypophyse* » peuvent être reliés par une relation de cause ou une relation de fonction. Aucun indice linguistique ne nous permet de statuer pour l'une ou l'autre des relations.

Environ 9000 contextes ont été annotés selon ces critères.

4 Variabilité du fonctionnement des candidats-marqueurs

Comptabilisant ensemble les « oui » et « plutôt oui », nous avons calculé la précision de chaque candidat-marqueur. Étant donné le nombre d'occurrences total d'un candidat-marqueur, nous avons calculé le ratio du nombre de cas dans lesquels la relation était présente sur le nombre d'occurrence total. Ce rapport, exprimé en pourcentage, correspond à la formule suivante :

$$Précision = \left(\frac{\text{Nombre d'occurrences dans lesquelles la relation est présente}}{\text{Nombre d'occurrences total}}\right) \times 100$$

Nous présentons dans la suite deux types de résultats : une analyse quantitative comparant les taux de précision des candidats-marqueurs de méronymie selon le domaine, et une analyse qualitative de la variation selon le domaine également. À partir de ces observations, nous proposons quelques pistes de réutilisabilité des candidats-marqueurs.

4.1 Analyse quantitative

Le calcul des taux de précision des candidats-marqueurs nous a permis de comparer la répartition des candidats-marqueurs au sein des sous-corpus, ainsi que leur taux de précision. Deux hypothèses ont émergé : le domaine influence les types de relations présentes en corpus ; tandis que le genre textuel influence l'apparition de certains candidats-marqueurs. À travers l'exemple des relations de méronymie, nous illustrons les phénomènes de variation observés.

Le tableau ci-dessous présente la répartition et la précision des candidats-marqueurs de méronymie selon les sous-relations auxquelles ils appartiennent et selon le domaine :

	VOLCANOLOGIE		CANCER	
Sous-relation	Nb Occ	Précision	Nb Occ	Précision
1- Réunion compacte	1	100%	na	na
2- Fusion	10	90%	1	**100%**
3- Décomposition	52	90%	6	50%
4- Non-organisation	18	83%	22	**100%**
5- Inclusion	187	80%	123	60%
6- Rapprochement	4	75%	11	0%
7- Organisation	20	75%	2	50%
8- Expression du Lieu / Temps	126	71%	52	65%
9- Type de parties	40	70%	10	60%
10- Parties de genre différent	14	64%	na	na
11- Jonction	21	57%	9	**67%**
12- Parties de même genre	22	41%	27	**70%**
Total	515	75%	263	63%

TABLE 3. Nombre d'occurrences et précision des candidats-marqueurs par sous-relation selon le domaine.

D'une manière générale les candidats-marqueurs de la relation de méronymie ont un taux de précision plus élevé dans le domaine de la volcanologie (75%) que dans le domaine du cancer du sein (63%). Deux cas sont possibles quant à l'influence du domaine sur la précision selon les sous-relations : soit le taux de précision des candidats-marqueurs est supérieur dans le domaine de la volcanologie, soit il est supérieur dans le domaine du cancer du sein :

- Les candidats-marqueurs des sous-relations *Parties de même genre*, *Jonction*, *Fusion*, et *Non-Organisation* (numérotées 2, 4, 11, 12) ont des taux de précision plus élevés dans le domaine du cancer du sein
- Les candidats-marqueurs des sous-relations *Parties de genre différent, Rapprochement*, *Réunion compacte*, *Inclusion, Organisation, Décomposition, Type de parties, Expression du lieu/temps* ont des taux de précision plus élevés dans le domaine de la volcanologie.

La tendance générale se vérifie donc au niveau des sous-relations puisque 8 sur 12 d'entre elles ont des candidats-marqueurs dont les taux de précision plus élevés dans le domaine de la volcanologie. Dans la suite, nous illustrons certains des phénomènes de variation observés, en établissant un lien avec les thématiques traitées dans les corpus d'étude.

4.2 Analyse qualitative

Nous présentons dans la suite la variation d'un point de vue linguistique, en nous penchant plus particulièrement sur les candidats-marqueurs ayant un taux de précision plus élevé dans le domaine de la volcanologie. Les thèmes abordés dans ce corpus concernent les types de volcanisme, mais également les roches, les laves, le magma. La composition, mais également la décomposition des éléments chimiques ou des roches lors d'un phénomène volcanique sont souvent étudiées. Le volcan

est souvent décrit comme un *édifice* ayant une *structure*, au sein de laquelle sont organisées différentes parties. Cela peut ainsi expliquer la précision plus élevée des sous-relations énoncées plus haut dans le domaine de la volcanologie. Par exemple, les candidats-marqueurs exprimant la *Décomposition* permettent de détecter des contextes comme le (1) ci-dessous :

(1) Le magma basique, en contact avec le magma acide plus froid, ***se fragmente en*** *micropillows (ou pillows) et en flammes. (volcanologie, scientifique)*

Le contexte (1) illustre l'expression de la décomposition d'un tout, *le magma basique*, en parties, les *micropillows* et les f*lammes*. Le marqueur testé ici est **[DET X se fragmenter en DET Y]**. Le terme *fragmentation* est présent parmi les candidats-termes du domaine de la volcanologie, mais pas parmi ceux du cancer du sein. Cela est également une indication sur la présence de tels marqueurs dans le corpus de volcanologie. Il est intéressant de noter ici que certains éléments lexicaux présents dans les marqueurs sont des candidats-termes du domaine. Autrement dit, l'un des éléments du candidat-marqueur, un verbe ou une nominalisation, est à la fois candidat-terme du domaine et candidat-marqueur. Dans ce cas, les marqueurs sont moins soumis à la polysémie, ce qui permet d'interpréter la relation attendue.

5 Conclusion et perspectives

La construction de ressources terminologiques ou ontologiques est l'un des champs explorés par le TAL ou l'ingénierie des connaissances. L'intérêt d'utiliser des marqueurs de relation pour détecter de manière automatique ou semi-automatique des Contextes Riches en Connaissances a de nombreuses fois été souligné. Notre contribution est d'avoir constitué la base MAR-REL[2], qui rassemble les principaux candidats-marqueurs français des relations d'hyperonymie, de méronymie et de cause. Chacun des candidats-marqueurs a été évalué en corpus variant selon deux domaines (volcanologie et cancer du sein) et deux genres textuels (scientifique vs vulgarisé).

La caractérisation du fonctionnement linguistique des marqueurs, en termes de stabilité ou de variation, permet d'émettre l'hypothèse qu'il est possible de réutiliser certains résultats pour des corpus ayant des caractéristiques proches de ceux de notre étude. Ainsi par exemple, le marqueur de méronymie [DET X se fragmenter en DET Y] pourrait donner des résultats intéressants dans des corpus liés à l'urbanisme ou dont les thématiques sont liées à la description de structures.

Compte-tenu de la taille de nos corpus, les résultats obtenus nécessitent cependant d'être nuancés. Un travail sur des corpus variés pourrait permettre de mener une autre étude contrastive, qui fournirait une description linguistique encore plus étendue des marqueurs de relation. Enfin, dans l'objectif de donner une perspective plurilingue à cette base, le même travail est en cours pour l'anglais et pour l'italien. Une étude des marqueurs de relation en espagnol est également envisagée.

[2] La base est constituée de deux parties : les marqueurs décrits de manière lexico-syntaxique, et leur traduction en langage UIMA (effectuée par Lingua&Machina). Nous indiquerons l'adresse à laquelle l'ensemble de la base MAR-REL sera disponible lors de la Conférence si notre communication est acceptée.

Références

AUGER A., BARRIÈRE C. (2008). Pattern-based approaches to semantic relation extraction: A state-of-the-art. *Terminology* 14, 1–19.

AUSSENAC-GILLES N., SÉGUÉLA P., (2000). Les relations sémantiques : du linguistique au formel. *Cahiers de Grammaire, « Sémantique et Corpus »* 25, 175–198.

BARRIERE C. (2001). Investigating the causal relation in informative texts, *Terminology* 7, 135–154.

CONDAMINES A. (2000). Chez dans un corpus de sciences naturelles: un marqueur de relation de relation méronymique?. *Cahiers de lexicologie* 77, 165-187.

CONDAMINES A. (2006). Avec et l'expression de la méronymie: l'importance du genre textuel. in G. Kleiber, C. Schnedecker et A. Thyssen (Eds) : *La relation «Partie - Tout».* Leuven : Peeters, 633–650.

CONDAMINES A. (2008). Taking genre into account when analysing conceptual relation patterns. *Corpora* 8, 115–140.

CONDAMINES A., REBEYROLLE J. (2000). Construction d'une base de connaissances terminologiques à partir de textes: expérimentation et définition d'une méthode. in J. Charlet, M. Zacklad, G. Kassel & D. Bourigault (Eds) : *Ingénierie des connaissances, évolutions récentes et nouveaux défis*, Eyrolles, 225-241

FLOWERDEW J. (1992). Definitions in Science Lectures. *Applied Linguistics* 13, 203–221.

GARCIA D. (1998). *Analyse automatique des textes pour l'organisation causale des actions : réalisation du système informatique COATIS.* Thèse de Doctorat en Informatique. Université Paris 4.

GIRJU R., BADULESCU A., MOLDOVAN D. (2003). Learning semantic constraints for the automatic discovery of part-whole relations. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*-Volume 1, Association for Computational Linguistics,1–8.

HEARST M.A. (1992). Automatic acquisition of hyponyms from large text corpora. *Proceedings of COLING-92*, 539-545.

Lafourcade M., Ramadier, L. (2016) Semantic Relation Extraction with Semantic Patterns Experiment on Radiology Reports. 10th edition of the Language Resources and Evaluation Conference (LREC 2016), 23-28 May 2016, Portorož (Slovenia), 6 p.

MALAISÉ V., ZWEIGENBAUM P., BACHIMONT B. (2004). Repérage et exploitation d'énoncés définitoires en corpus pour l'aide à la construction d'ontologie. Actes de *TALN'04*, *Conférence sur le Traitement Automatique des Langues Naturelles*, 269–278.

MARSHMAN E. (2008). Expressions of uncertainty in candidate knowledge-rich contexts: A comparison in English and French specialized texts. *Terminology* 14, 124–151.

MARSHMAN E., L'HOMME M.-C. (2006). Portabilité de la relation causale : étude sur deux corpus spécialisés, *Corpus et dictionnaires de langues de spécialité* : *Actes des Journées du CRTT*, 87-110.

MARSHMAN E., L'HOMME M.-C., SURTEES V. (2008). Portability of cause-effect relation markers across specialised domains and text genres: a comparative evaluation. *Corpora* 3, 141–172.

MARSHMAN E., GARIÉPY JL., HARMS C (2012). Helping language professionals relate to terms: Terminological relations and termbases. *The Journal of Specialised Translation* 18, 30-56.

MEYER I. (2001). Extracting knowledge-rich contexts for terminography: A conceptual and methodological framework. in Bourigault, D., Jacquemin, C., L'Homme, M.-C. (Eds) : *Recent Advances in Computational Terminology*, Amsterdam/Philadephia : John Benjamins, 279–302.

MORIN E. (1999). *Extraction de liens sémantiques entre termes à partir de corpus de textes techniques.* Thèse de doctorat en Informatique, Université de Nantes.

NUOPPONEN A. (2014). Tangled Web of Concept Relations. Concept relations for ISO 1087-1 and ISO 704. *Terminology and Knowledge Engineering 2014*, 1-10.

PEARSON J. (1998). *Terms in context*. Amsterdam/Philadephia : John Benjamins.

PEARSON J. (2000). Une tentative d'exploitation bi-directionnelle d'un corpus bilingue. *Cahiers de Grammaire, « Sémantique et Corpus »* 25, 53–69.

REBEYROLLE J. (2000). *Forme et fonction de la définition en discours*. Thèse de doctorat en Sciences du Langage, Université Toulouse 2.

Apprentissage en ligne interactif d'un générateur en langage naturel neuronal pour le dialogue homme-machine

Matthieu Riou Bassam Jabaian Stéphane Huet Fabrice Lefèvre
CERI-LIA, Université d'Avignon, France
matthieu.riou@alumni.univ-avignon.fr, {bassam.jabaian, stephane.huet, fabrice.lefevre}@univ-avignon.fr

RÉSUMÉ

Récemment, de nouveaux modèles à base de réseaux de neurones récurrents ont été proposés pour traiter la génération en langage naturel dans des systèmes de dialogue (Wen *et al.*, 2016a). Ces modèles demandent une grande quantité de données d'apprentissage ; or la collecte et l'annotation de ces données peuvent être laborieuses. Pour répondre à cette problématique, nous nous intéressons ici à la mise en place d'un protocole d'apprentissage en ligne basé sur un apprentissage par renforcement, permettant d'améliorer l'utilisation d'un modèle initial appris sur un corpus plus restreint généré par patrons. Dans cette étude exploratoire, nous proposons une approche basée sur un algorithme de bandit contre un adversaire, afin d'en étudier l'intérêt et les limites.

ABSTRACT

On-line Interactive Learning of Natural Language Neural Generation for Human-machine Dialogue.

Recently, some propositions have emerged to handle natural language generation in spoken dialog systems with recurrent neural network models (Wen *et al.*, 2016a). Those models require a huge amount of learning data, which complicates the data collection and annotation. To address this issue, we propose an online-learning protocol, based on a reinforcement learning approach. We learn a model on a reduced corpus produced with templates, and adapt it online. For this first experiment, we propose an approach using an adversarial bandit algorithm, studying its advantages and limitations.

MOTS-CLÉS : génération en langage naturel, réseau de neurones récurrent, bandit contre un adversaire, adaptation à l'utilisateur.

KEYWORDS: natural language generation, recurrent neural network, adversarial bandit, on-line learning, user adaptation.

1 Introduction

Dans un système de dialogue oral, le composant de génération en langage naturel (NLG pour *Natural Language Generation*) vise à produire une phrase à partir d'un acte de dialogue (AD) du système, proposé par le gestionnaire de dialogue. Par exemple, l'acte $inform(name = bar_metropol, type = bar, area = north, food = french)$ peut générer la phrase « *Bar Metropol est un bar au nord de la ville qui propose de la cuisine française* ». Pour produire les réponses, les systèmes utilisent traditionnellement des patrons et des règles rédigés par des experts et associés à des ressources linguistiques (Rambow *et al.*, 2001). Ce type de modèles NLG produit des phrases de bonne qualité

mais répétitives et seulement pour des tâches bien spécifiées. Plusieurs approches stochastiques basées sur les données ont conduit à remettre en cause ces méthodes car elles présentent notamment deux défauts : le manque de généralisation pour des domaines ouverts et la répétition fréquente de phrases identiques et de structures rigides.

Oh et Rudnicky ont proposé d'utiliser un modèle de langue n-grammes pour surgénérer un ensemble de phrases candidates, dans lequel est retenue la forme finale (Oh & Rudnicky, 2002). Mairesse et Young ont étendu ce modèle en introduisant des facteurs construits sur une représentation sémantique peu profonde pour construire des n-grammes de segments (Mairesse & Young, 2014). Plus récemment Wen *et al.* ont proposé plusieurs modèles basés sur les RNN (*Recurrent Neural Networks*) (Wen *et al.*, 2015a,c,b). Les évaluations par des humains ont montré que ces systèmes peuvent générer des phrases de bonne qualité linguistique et plus variées que les patrons. L'utilisation de modèles neuronaux récurrents encodeur-décodeur a aussi été étudiée pour construire des systèmes de dialogue de bout en bout (Serban *et al.*, 2016). Dans ce cadre, les RNN se chargent de la NLG mais aussi de la compréhension et du processus de décision ; cette étude est toutefois effectuée dans un contexte différent de cet article en considérant des systèmes conversationnels non dirigés par le but.

D'une manière générale, les modèles stochastiques impliquent un travail important pour la production de corpus pour les nouveaux domaines. Wen *et al.* présentent une approche incrémentale pour gérer l'adaptation de domaine d'un modèle de génération à base de RNN (Wen *et al.*, 2016b). Ils recourent à des données contrefaites synthétisées à partir d'un ensemble hors-domaine pour ajuster leur modèle sur un ensemble réduit de phrases du domaine. Il est aussi envisageable d'employer des méthodes d'extension de corpus (Manishina *et al.*, 2016) mais elles ne permettent pas une adaptation simultanée aux préférences de l'utilisateur.

Dans cet article, nous tentons de réduire le coût de la production de nouvelles données, en adaptant un modèle initial, non pas à une nouvelle tâche mais afin de générer des phrases avec une plus grande diversité. Dans cette logique, une approche par renforcement basée sur un algorithme de bandit contre un adversaire est appliquée (Auer *et al.*, 2002). Si cette approche a déjà été utilisée dans les systèmes de dialogue pour la compréhension du langage (Ferreira *et al.*, 2015, 2016), ici nous proposons un protocole pour adapter le modèle RNN aux nouvelles phrases, différentes de l'ensemble d'apprentissage, en prenant en compte le coût requis de l'utilisateur pour les fournir au système.

2 Le problème de l'interaction en ligne

Les modèles NLG neuronaux peuvent donner de bons résultats (Wen *et al.*, 2015c) mais nécessitent une quantité importante de données d'apprentissage annotées pour pouvoir produire des sorties diversifiées et de bonne qualité. Plusieurs exemples de chaque phrase pour chaque acte de dialogue sont en particulier nécessaires pour entraîner le modèle. Afin de réduire le coût de collecte du corpus, nous proposons dans cet article un protocole d'apprentissage en ligne.

Nous procédons en deux étapes. Tout d'abord, un corpus initial, constitué de références générées par patrons, est utilisé pour entraîner un modèle de génération. Puis, le modèle est utilisé pour générer des phrases, en interaction vocale avec l'utilisateur, auquel il peut demander de produire un énoncé meilleur ou différent. À cette étape du développement, nous considérons l'utilisateur comme averti (il peut être un des développeurs ou un collègue), puisqu'il pourrait être hasardeux de laisser n'importe quel utilisateur accéder à une telle fonctionnalité sans un moyen de contrôle efficace de l'adaptation du modèle. Cette difficulté sera traitée dans de futurs travaux.

Afin de réduire l'effort demandé à l'utilisateur et éviter des actions inutiles, nous proposons d'utiliser un algorithme de bandit contre un adversaire pour décider si le système devrait ou non demander l'expertise de l'utilisateur, selon le gain et le coût estimés de cette action.

Concrètement afin de collecter de nouvelles données pour son modèle, le système va devoir décider à chaque tour de parole s'il : 1. doit demander à l'utilisateur une alternative à sa proposition, 2. peut utiliser la transcription automatique de l'entrée utilisateur ou demander un traitement correctif additionnel.

2.1 Cas statique

Une fois que le système génère l'énoncé, le système peut choisir une action (à partir d'une distribution de probabilité) parmi un ensemble $\mathcal{I}$ de M actions. Dans une configuration préliminaire, nous prenons $M = 3$ et $\mathcal{I}$ est défini comme : $\mathcal{I} := \{\text{Skip, AskDictation, AskTranscription}\}$. Soit $i \in \mathcal{I}$ l'indice de l'action. Nous supposons que l'effort de l'utilisateur $\phi(i) \in \mathbb{R}^+$, peut être associé au temps nécessaire pour effectuer l'action i. Les **efforts estimés** associés à chaque action sont :

• **Skip** : n'appliquer aucune mise à jour au modèle. Le coût de cette action est nul ($\phi(\text{Skip}) = 0$).
• **AskDictation** : affiner le modèle en considérant un énoncé alternatif proposé par l'utilisateur et transcrit automatiquement par un système de reconnaissance ($\phi(\text{AskDictation}) = 1$).
• **AskTranscription** : demander à l'utilisateur de transcrire la correction ou l'énoncé alternatif ($\phi(\text{AskTranscription}) = 1 + l$, avec l la taille de l'énoncé proposé).

Nous estimons ensuite le **gain** de l'action choisie $g(i) \in [0, 1]$ comme suit :

• **Skip** : rien n'est appris, le gain est de 0 ($g(\text{Skip}) = 0$).
• **AskDictation** : le gain est calculé en considérant la marge restante du score BLEU (Papineni *et al.*, 2002) de l'énoncé généré par le système, en utilisant l'énoncé proposé par l'utilisateur comme référence $\text{BLEU}_{gen/prop}$. Pour tenir compte des erreurs potentielles ajoutées par le système de reconnaissance vocal, cette mesure est pénalisée par l'estimation d'un WER et d'un SER :

$$g(\text{AskDictation}) = (1 - \text{BLEU}_{gen/prop}) \times (1 - \text{WER}) \times (1 - \text{SER})$$

Le WER (*Word Error Rate*) donne le nombre moyen d'erreurs de transcription au niveau des mots, tandis que le SER (*Slot Error Rate* (Wen *et al.*, 2015c)) évalue la présence des slots (concepts, valeurs) de l'AD et pénalise les énoncés générés qui ne contiennent pas les informations sémantiques requises.
• **AskTranscription** : cette action suppose de demander à l'utilisateur de transcrire manuellement l'énoncé pour éviter les erreurs de transcription automatique. Donc, l'estimation du gain ne tient compte que du score BLEU de l'énoncé généré par le système, en utilisant la phrase proposée par l'utilisateur comme référence ($g(\text{AskTranscription}) = 1 - \text{BLEU}_{gen/prop}$).

Enfin, une fonction de perte $l(i) \in [0, 1]$ est définie afin que le système maximise la mesure de gain $g(i)$ et minimise l'effort de l'utilisateur $\phi(i)$:

$$l(i) = \underbrace{\alpha(1 - g(i))}_{\text{amélioration du système}} + \underbrace{(1 - \alpha)\frac{\phi(i)}{\phi_{max}}}_{\text{effort de l'utilisateur}} \quad (1)$$

α est un scalaire qui pondère le gain par rapport au coût, permettant au système de s'adapter aux préférences de l'utilisateur. ϕ_{max} correspond à l'effort maximal (ici il peut être atteint par AskTranscription et vaut 41, la taille maximale d'une phrase étant fixée à 40 mots) pour normaliser l'effort de l'utilisateur entre 0 et 1.

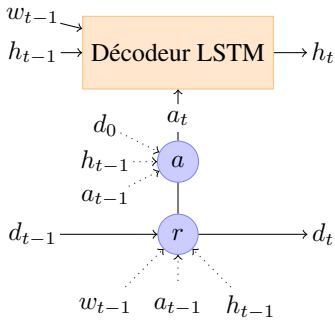


FIGURE 1 – LSTM à contexte combiné

2.2 Cas du bandit contre un adversaire

Nous considérons le scénario suivant pour le problème de bandit contre un adversaire : à chaque itération, le système produit une phrase puis choisit une action $i_t \in \mathcal{I}$. Une fois que l'action i_t est effectuée, le système calcule l'estimation du gain $g(i_t)$, l'effort de l'utilisateur $\phi(i_t)$ et la perte $l(i_t)$. Le but de l'algorithme de bandit est alors de trouver $i_1, i_2, \ldots$, afin que pour chaque t, le système minimise la perte $l(i_t)$.

Nous découpons l'apprentissage en bloc de n itérations chacun. À la fin de chaque bloc, nous ajoutons les énoncés proposés par l'utilisateur à notre corpus d'apprentissage et nous mettons à jour le modèle sur ce corpus étendu. En même temps, nous calculons la fonction de perte pour chaque choix de bandit et nous mettons à jour sa politique. Cette mise à jour du modèle NLG et de la politique décidant de l'action au bout de plusieurs itérations, plutôt qu'à chaque énoncé collecté, permet de réduire le temps de mise à jour du modèle et d'améliorer sa robustesse.

3 Expériences

3.1 Le système de génération

Le modèle NLG utilisé est le LSTM (*Long short-term memory*) à contexte combiné décrit plus en détail dans (Riou *et al.*, 2017). Ce modèle combine deux modèles neuronaux proposés par Wen *et al.*, le modèle LSTM conditionné sémantiquement (SCLSTM) et l'encodeur-décodeur RNN avec attention. Chacun de ces systèmes traite différemment l'information sémantique représentée par un AD pour produire la phrase. La cellule de contrôle (*reading gate*) du SCLSTM permet de filtrer l'AD en ne conservant à chaque étape que l'information restante qui n'a pas été encore traitée. Au contraire, le mécanisme d'attention permet de sélectionner dans l'AD au complet l'information à considérer spécifiquement à l'étape suivante, mais perd la progression dans l'information déjà traitée. L'objectif de notre système est de combiner les avantages des deux systèmes, en utilisant une cellule de contrôle r et un mécanisme d'attention a pour traiter séquentiellement l'AD d_t restant à générer (Figure 1), en prenant en compte l'information h_{t-1} portée dans la couche cachée du LSTM, le mot précédemment généré w_{t-1}, l'attention portée a_{t-1} à un temps donné et l'information d_0 de l'AD en cours de lecture. Ainsi, ce modèle permet de sélectionner dans l'information encore non traitée, celle à considérer spécifiquement à l'étape suivante. Cette information sélectionnée est ensuite envoyée dans un décodeur LSTM pour produire le mot suivant.

3.2 Cadre expérimental

Les expériences ont été menées sur le corpus *SF restaurant* décrit dans (Wen *et al.*, 2015c) et librement accessible en ligne [1]. Ces données contiennent 5 191 phrases, pour 271 actes de dialogue (AD) distincts. Le corpus associe à chaque acte une phrase générée par patron et plusieurs phrases proposées par des annotateurs humains, chaque phrase étant délexicalisée [2].

Le LSTM à contexte combiné a été implémenté en utilisant la bibliothèque Tensorflow [3]. Ce système a ensuite été entraîné sur le corpus séparé en 3 parties suivant un ratio 3 :1 :1 : apprentissage, validation et test, en utilisant uniquement les références proposées par les annotateurs humains.

3.3 Comparaison des systèmes de génération

Une évaluation a été conduite avec deux métriques objectives calculées à partir des générations de référence, le score BLEU-4 (Papineni *et al.*, 2002) et le taux d'erreur SER. Le BLEU valide la génération de phrases, notamment la grammaticalité, tandis que le SER se concentre spécifiquement sur le contenu sémantique. Pour chaque exemple, nous produisons 20 hypothèses et ne gardons pour l'évaluation que les 5 meilleures selon le score donné par le modèle NLG. Des références multiples pour chaque AD sont obtenues en groupant les phrases délexicalisées du même AD et en les relexicalisant ensuite.

Le LSTM à contexte combiné obtient un score BLEU de 71,1%, voisin des deux autres systèmes initiaux (72,2% pour le SCLTSM et 69,7% pour l'encodeur-décodeur), mais le taux d'erreur SER est divisé par trois par rapport aux autres systèmes, en passant de 0,77% et 0,65% pour le SCLSTM et l'encodeur-décodeur, à 0,24% pour le LSTM à contexte combiné. Cela veut dire qu'il propose une meilleure couverture des concepts à exprimer et donc moins d'omissions ou d'erreurs de concepts, ce qui est le principal but recherché pour un module NLG.

3.4 Évaluation de l'apprentissage en ligne

Nous avons utilisé le même corpus, mais avec un découpage en apprentissage, validation et test suivant un ratio 2 :1 :1. Le modèle NLG utilisé est encore le LSTM à contexte combiné. Pour initialiser le modèle, nous l'entraînons en utilisant les références générées par patron du corpus d'apprentissage. Le corpus de validation a permis de décider l'arrêt de la phase d'apprentissage (*early stopping*). Ensuite, nous avons simulé un apprentissage en ligne en réutilisant le corpus d'apprentissage, mais cette fois en apprenant sur les références proposées par des annotateurs humains. Le modèle ainsi que la politique de bandit ont été mis à jour toutes les 400 phrases. Le WER a été simulé en insérant de manière aléatoire des erreurs (confusion, insertion et suppression) dans les exemples du corpus, jusqu'à atteindre un taux global de WER prédéfini.

Le modèle initial, entraîné sur les références générées par patron, atteint un haut score BLEU de 80,2% quand celui-ci est calculé à partir de références générées par patrons, mais qui est réduit à seulement 39,7% en refaisant les calculs à partir des seules références proposées par des annotateurs humains. Cela montre bien la grande diversité possible des réponses dans une conversation.

La Figure 2 montre l'évolution du score BLEU en fonction du coût d'apprentissage de la politique du bandit, avec un WER simulé (fixé à 5%). Le score BLEU est obtenu en testant le modèle sur les

1. https://www.repository.cam.ac.uk/handle/1810/251304

2. La délexicalisation remplace les formes de surface des concepts par des variables, *inform(name=la mimosa, food=mediterranean)* devenant « *$name sert des plats $food* ».

3. https://www.tensorflow.org

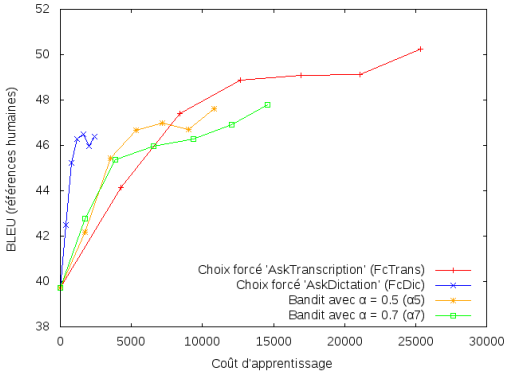
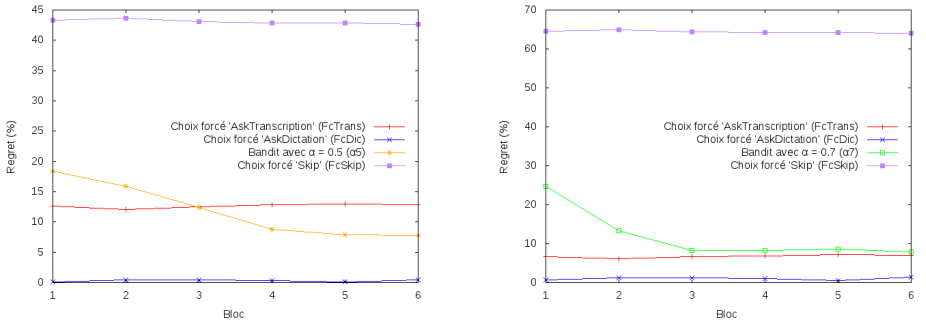


FIGURE 2 – Évolution du score BLEU en fonction du coût cumulé.

références du corpus de test proposées par des annotateurs humains, tandis que le coût d'apprentissage est calculé en sommant les coûts de chaque choix effectué par le bandit durant l'apprentissage en ligne. La taille d'un bloc a été fixée à 400 itérations. Nous avons testé deux configurations limites en forçant le choix 'AskDictation' (FcDic) et en forçant le choix 'AskTranscription' (FcTrans). De plus, nous avons testé le bandit avec deux valeurs différentes pour α : 0,5 ($\alpha 5$) et 0,7 ($\alpha 7$). La seconde valeur réduit l'influence du coût d'apprentissage, permettant au système de demander plus d'effort à l'utilisateur. Chaque courbe est composée de sept points. Le premier correspond au score BLEU du modèle par rapport aux références proposées par des annotateurs humains, avant l'apprentissage en ligne. Les six suivants sont calculés après chaque bloc de 400 phrases. Le coût est cumulé sur tous les blocs précédents. Nous pouvons observer que le bandit réduit avec succès le coût d'apprentissage jusqu'à une certaine quantité de données apprises ; après un coût cumulé de 7 500, AskTranscription devient la meilleure configuration. Une fois toutes les données d'apprentissage utilisées, $\alpha 5$ et $\alpha 7$ atteignent tous deux 47,6% de BLEU, entre les 46,4% pour FcDic et 50,3% pour FcTrans. Le coût pour AskTranscription est bien plus élevé que pour AskDictation. Par conséquent, le bandit apprend mieux au départ que FcTrans, en équilibrant entre les deux choix ; après les deux premiers blocs, l'augmentation du score BLEU ralentit, jusqu'à ce que les courbes d'$\alpha 5$ et $\alpha 7$ passent toutes les deux sous celle FcTrans. Une plus grande valeur d'α tend à favoriser les actions de type Ask sur Skip, et AskTranscription sur AskDictation, ce qui permet notamment de s'adapter aux préférences de l'utilisateur, y compris en cours d'utilisation.

Une autre façon de visualiser l'apprentissage du bandit est d'observer la variation du regret, présentée dans la Figure 3. Le regret représente la marge du bandit pour améliorer son apprentissage. Pour cela, on considère la récompense estimée du bandit ($r(i) = 1,0 - l(i)$). On peut alors calculer le regret comme la différence entre la meilleure récompense possible parmi les trois choix $r_{opti}(i)$ et la récompense estimée $r(i)$. Les six points des courbes correspondent à la somme des regrets sur chaque bloc, exprimée en pourcentage de la somme des récompenses optimales. Nous pouvons observer que le bandit diminue effectivement son regret au cours de l'apprentissage, ce qui vérifie bien l'apprentissage de sa politique. Il n'arrive cependant pas à passer en dessous des 10% de regret. On peut aussi observer que le choix AskDictation possède un très faible regret, ce qui est attendu, puisqu'il permet d'obtenir bien souvent des phrases correctes en demandant moins d'effort à l'utilisateur. Mais il faut noter aussi un effet de moyenne qui cache les exemples possédant beaucoup d'erreurs de transcription, et donc un fort regret. Pour permettre au bandit de détecter ces exemples



(a) $\alpha = 0,5$ (b) $\alpha = 0,7$

FIGURE 3 – Évolution du regret lors de 2 apprentissages simulés, avec $\alpha = 0,5$ et $0,7$.

où le choix AskDictation est à éviter, mais aussi de mieux savoir lorsqu'il est plus intéressant de ne rien demander à l'utilisateur, il devient nécessaire de prendre en compte le contexte de l'exemple traité dans la décision du bandit. Ainsi, tout en gardant le même protocole d'apprentissage en ligne et ses avantages (équilibrage entre différents choix et adaptation à l'utilisateur), nous espérons pouvoir améliorer nos résultats en prenant en compte le contexte de la phrase générée (estimation du nombre d'erreurs lors de la génération, diversité des hypothèses générées par le système NLG) à l'aide d'un bandit contextuel (Auer *et al.*, 2002).

4 Conclusions et perspectives

Dans cet article, nous avons proposé un nouveau protocole pour adapter un modèle initial de génération de langage naturel neuronal à l'aide d'un apprentissage en ligne. Dans une expérience simulée, un algorithme de bandit a permis d'équilibrer de manière automatique l'évolution des performances du système avec le coût induit pour l'utilisateur. Ces résultats ne sont encore que préliminaires et le travail doit être poursuivi. Tout d'abord, les capacités d'apprentissage du bandit doivent être augmentées par la prise en compte du contexte lors de ses décisions, à l'aide d'un bandit contextuel. Une évaluation avec des utilisateurs doit également être menée pour avoir une évaluation plus fiable des coûts et montrer l'intérêt pratique de notre approche.

Remerciements

Ce travail a été partiellement financé par le Labex BLRI (ANR-11-LABX-0036).

Références

AUER P., CESA-BIANCHI N., FREUND Y. & SCHAPIRE R. E. (2002). The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, **32**(1), 48–77.

FERREIRA E., JABAIAN B. & LEFÈVRE F. (2015). Zero-shot semantic parser for spoken language understanding. In *INTERSPEECH*.

FERREIRA E., REIFFERS-MASSON A., JABAIAN B. & LEFÈVRE F. (2016). Adversarial bandit for online interactive active learning of zero-shot spoken language understanding. In *ICASSP*.

MAIRESSE F. & YOUNG S. (2014). Stochastic language generation in dialogue using factored language models. *Computational Linguistics*, **40**(4), 763–799.

MANISHINA E., JABAIAN B., HUET S. & LEFÈVRE F. (2016). Automatic corpus extension for data-driven natural language generation. In *LREC*.

OH A. H. & RUDNICKY A. I. (2002). Stochastic natural language generation for spoken dialog systems. *Computer Speech & Language*, **16**(3–4), 387–407.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). Bleu : a method for automatic evaluation of machine translation. In *ACL*.

RAMBOW O., BANGALORE S. & WALKER M. (2001). Natural language generation in dialog systems. In *HLT*.

RIOU M., JABAIAN B., HUET S. & LEFÈVRE F. (2017). Online adaptation of an attention-based neural network for natural language generation. In *INTERSPEECH*.

SERBAN I. V., SORDONI A., BENGIO Y., COURVILLE A. & PINEAU J. (2016). Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI Conference on Artificial Intelligence*.

WEN T.-H., GAŠIĆ M., MRKŠIĆ N., BARAHONA L. M. R., SU P.-H., ULTES S., VANDYKE D. & YOUNG S. (2016a). Conditional generation and snapshot learning in neural dialogue systems. In *EMNLP*.

WEN T.-H., GAŠIĆ M., KIM D., MRKŠIĆ N., SU P.-H., VANDYKE D. & YOUNG S. (2015a). Stochastic language generation in dialogue using recurrent neural networks with convolutional sentence reranking. In *SIGDIAL*.

WEN T.-H., GAŠIĆ M., MRKŠIĆ N., ROJAS-BARAHONA L. M., SU P.-H., VANDYKE D. & YOUNG S. (2015b). Toward multi-domain language generation using recurrent neural networks. In *NIPS Workshop on Machine Learning for Spoken Language Understanding and Interaction*.

WEN T.-H., GAŠIĆ M., MRKŠIĆ N., ROJAS-BARAHONA L. M., SU P.-H., VANDYKE D. & YOUNG S. (2016b). Multi-domain neural network language generation for spoken dialogue systems. In *NAACL-HLT*.

WEN T.-H., GAŠIĆ M., MRKŠIĆ N., SU P.-H., VANDYKE D. & YOUNG S. (2015c). Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. In *EMNLP*.

Apports des analyses syntaxiques pour la détection automatique de mentions dans un corpus de français oral

Loïc Grobol[1], Isabelle Tellier[1], Éric de la Clergerie[2],
Marco Dinarelli[1], Frédéric Landragin[1]
(1) Lattice (UMR 8094), CNRS, ENS Paris, Université Sorbonne Nouvelle,
PSL Research University, USPC, 1 rue Maurice Arnoux, 92120 Montrouge, France
(2) ALMAnaCH, INRIA, 2 rue Simone Iff, 75589 Paris, France
loic.grobol@gmail.com, isabelle.tellier@sorbonne-nouvelle.fr,
eric.de_la_clergerie@inria.fr, marco.dinarelli@ens.fr, frederic.landragin@ens.fr

RÉSUMÉ

Cet article présente trois expériences de détection de mentions dans un corpus de français oral : ANCOR. Ces expériences utilisent des outils préexistants d'analyse syntaxique du français et des méthodes issues de travaux sur la coréférence, les anaphores et la détection d'entités nommées. Bien que ces outils ne soient pas optimisés pour le traitement de l'oral, la qualité de la détection des mentions que nous obtenons est comparable à l'état de l'art des systèmes conçus pour l'écrit dans d'autres langues. Nous concluons en proposant des perspectives pour l'amélioration des résultats que nous obtenons et la construction d'un système *end-to-end* pour lequel nos expériences peuvent servir de base de travail.

ABSTRACT

Experiences in using deep and shallow parsing to detect entity mentions in oral French

We present three experiments in detecting entity mentions in the corpus of oral French ANCOR, using publicly available parsing tools and state-of-the-art mention detection techniques used in coreference detection, anaphora resolution and Entity Detection and Tracking systems. While the tools we use are not specifically designed to deal with oral French, our results are comparable to those of state-of-the-art end-to-end systems for other languages. We also mention several ways to improve our results for future work in developing an end-to-end coreference resolution system for French, to which these experiments could be a baseline for mention detection.

MOTS-CLÉS : corpus de dialogues, coréférence, détection de mentions, analyse syntaxique, chunks, apprentissage automatique.

KEYWORDS : Dialogue corpus, Coreference, Mention detection, Parsing, Chunking, Machine Learning.

1 Introduction

La reconnaissance automatique des chaînes de coréférence dans un texte est une tâche complexe et fondamentale pour avoir accès au contenu informationnel d'un long document. Elle a donné lieu à de grandes campagnes d'évaluation, comme celles proposées par MUC (*Message Understanding Conference*), ACE (*Automatic Content Extraction*), SemEval (*Semantic Evaluation*) ou CoNLL (*Computationnal Natural Language Learning*) et à de nombreux travaux. Cette tâche est souvent présentée comme revenant à partitionner l'ensemble des mentions d'un texte. Mais, avant de regrouper entre elles les mentions coréférentes, il faut les identifier. C'est sur cette étape préliminaire que nous nous concentrons dans cet article.

Commençons par décrire plus précisément l'objet de notre étude : on appelle *mention* toute expression

faisant référence à une entité[1]. Les mentions sont généralement des groupes nominaux ou des pronoms, mais tous les groupes nominaux et tous les pronoms ne sont pas des mentions. Ainsi, *il* dans « **il** y a pas encore de texte » et *la bienvenue* dans « la question là est la bienvenue et » sont non-référentiels et ne sont donc pas considérés comme des mentions. Détecter des mentions ne saurait donc se réduire à identifier les groupes nominaux et les pronoms. Les mentions peuvent également être imbriquées les unes dans les autre, voir par exemple dans la figure 1 « [le langage parlé et courant [des gens]] ».

On appelle *chaîne de coréférences* l'ensemble des mentions faisant référence à une même entité. On admet qu'une chaîne de coréférences peut n'être constituée que d'une seule mention[2] et on dira dans ce cas qu'il s'agit d'un *singleton*. La figure 1 montre un exemple de chaîne de coréférences (en gras) et des exemples de mentions (entre crochets).

— euh [on] dit quelquefois que [**la langue française**] se dégrade qu'est-ce que [vous] en pensez ?
— [**la langue française**] se dégrade dans [le langage parlé et courant [des gens]] [**elle**] se dégrade ?

FIG. 1: Mentions et chaînes de coréférences (corpus ANCOR).

On appelle *système complet d'identification de chaînes de coréférences* ou *système* end-to-end tout système capable de repérer les mentions dans un texte brut et de les regrouper en chaînes de coréférences. De tels systèmes ont déjà été conçus pour le français par (Trouilleux, 2001) et (Longo, 2013) mais leur portée reste modeste par rapport à celles de systèmes comme BART (Versley *et al.*, 2008) ou dcoref (Lee *et al.*, 2013), faute notamment de corpus en français annoté en chaînes de coréférences.

Un système — CROC — réalisant le regroupement des mentions en chaînes a déjà été réalisé pour le français par (Désoyer *et al.*, 2015) avec des performances comparables à l'état de l'art en anglais. Cependant, CROC s'appuyait sur des mentions de référence déjà repérées et portant des annotations morphosyntaxiques et sémantiques. Pour compléter ce travail, il reste donc à détecter les mentions dans le texte brut — c'est l'objet des travaux que nous présentons — et à y ajouter les annotations utilisées par CROC.

La détection des mentions n'est pas une priorité dans l'état de l'art : peu de travaux lui sont spécifiquement consacrés, beaucoup de systèmes *end-to-end* se contentent d'heuristiques simples, et certains travaux d'identification de chaînes de coréférence n'ont été développés qu'en tenant compte de mentions de référence.

De fait, les résultats obtenus par des systèmes de détection de mentions relativement simples sont globalement satisfaisants, ou en tout cas de suffisamment bonne qualité compte tenu de l'efficacité des systèmes de détection de coréférences qu'ils alimentent. Ces derniers, comportant généralement des détecteurs d'expressions non-référentielles, sont par ailleurs peu influencés par la précision de la détection de mentions, l'objectif est donc simplement de maximiser le rappel. La phase de détection des mentions des systèmes existants se résume donc bien souvent à la détection des groupes nominaux et des pronoms, tâche somme toute relativement simple pour des applications au langage écrit.

Dans notre cas, il serait cependant imprudent de négliger cette tâche. Le corpus avec lequel nous travaillons est en effet un corpus d'oral transcrit, et il n'est pas si évident que les outils d'analyse existants obtiennent des performances suffisantes pour lui appliquer directement les méthodes classiques. En effet, des travaux comme (Bove, 2008) ou (Tellier *et al.*, 2014) montrent que des techniques conçues pour l'écrit peuvent voir leurs performances significativement dégradées dans des applications directes au traitement de l'oral. Notre objectif est donc ici de déterminer dans quelle mesure les techniques classiques de détection de mentions peuvent s'appliquer au traitement du français oral et de produire une base de travail pour des recherches futures sur cette question.

1. Au sens de (Recasens, 2010)

2. Ce n'est pas le parti pris notamment pour les jeux de données des tâches MUC, CoNLL-2011 et CoNLL-2012, mais les singletons étant en fait très majoritaires parmi les chaînes de coréférences, en tenir compte permet d'améliorer significativement la détection des autres chaînes de coréférences (de Marneffe *et al.*, 2015).

2 Détection des mentions

2.1 Méthodes de détection

Nous présentons ici trois méthodes de détection des mentions dans un texte non-annoté. Nous les avons étudiées séparément, mais elle ne sont en rien mutuellement exclusives — (dos Santos & Carvalho, 2011) combine par exemple une extraction depuis une analyse syntaxique avec un système de détection d'entités nommées. Nous ne prétendons pas ici à l'exhaustivité : on pourra se reporter à (Poesio *et al.*, 2016) pour plus de détails.

2.1.1 Extraction à partir d'une analyse syntaxique

Cette méthode est utilisée par la majorité des systèmes *end-to-end* existants, généralement complétée par des systèmes de détection d'entités nommées et d'expressions non-référentielles. Elle consiste à extraire d'une analyse syntaxique les constituants pouvant être des mentions. Cette classification peut être réalisée à l'aide de règles écrites à la main (dos Santos & Carvalho, 2011), (Kummerfeld *et al.*, 2011) ou apprise automatiquement (Uryupina & Moschitti, 2013).

Cette procédure est raisonnablement fiable (voir tableau 1) et particulièrement adaptée au traitement de corpus arborés, comme ceux issus d'OntoNotes. Elle est cependant plus difficile à mettre en œuvre dans le cadre d'un système *end-to-end*, qui devient de ce fait dépendant de l'existence d'analyseurs syntaxiques pour la langue traitée et doit composer avec leurs éventuelles erreurs d'analyse.

2.1.2 Extraction à partir d'une analyse de surface

Les inconvénients de l'utilisation d'analyses syntaxiques évoqués dans la section 2.1.1 peuvent être — dans une certaine mesure — évités en utilisant non plus une analyse syntaxique complète, mais une analyse syntaxique de surface de type *chunking* (Abney, 1992). C'est le parti pris par (Soraluze *et al.*, 2012), qui utilise des outil de *chunking* et de détection de propositions dont les résultats sont ensuite traités par des règles écrites manuellement. L'avantage de cette approche est la plus grande disponibilité de *chunkers* de qualité et leur plus grande robustesse. L'inconvénient majeur est que ce type d'analyse, qui ignore les imbrications de constituants, ne permet pas de récupérer directement tous les types de mentions, et nécessite un traitement a posteriori — par exemple des règles de regroupement de *chunks*.

2.1.3 Apprentissage automatique de la détection des frontières de mentions

Les méthodes évoquées précédemment peuvent donner de bons résultats, mais sont entièrement dépendantes de l'existence de ressources linguistiques adaptées. Il existe cependant une autre possibilité : détecter directement dans le texte brut les frontières de mentions, comme le ferait un *chunker* ou un système de *NER*.

Cette technique a été étudiée entre autres par (Jing *et al.*, 2003), (Florian *et al.*, 2010), et (Qian *et al.*, 2007) (Nguyen *et al.*, 2016) pour la tâche de reconnaissance et de suivi d'entités (EDT) issue des campagnes ACE, qui ne traite de la détection de chaînes de coréférences que pour une classe restreinte d'entités. Elle permet de ramener le problème de la détection de mentions à une tâche d'étiquetage de séquences, pour laquelle des méthodes efficaces sont connues et ont été étudiées en profondeur : MEMM pour (Florian *et al.*, 2010), CRF pour (Qian *et al.*, 2007) et plus récemment réseaux de neurones récurrents pour (Nguyen *et al.*, 2016). Elle a en revanche le défaut classique de la modélisation de frontières par des étiquettes : la prise en compte d'entités imbriquées complique significativement les modèles (voir par exemple (Finkel & Manning, 2009)).

2.2 Mesures

Contrairement aux mesures de qualité de la détection des chaînes de coréférences, la mesure de la qualité de la détection des mentions tend à faire consensus et est en général donnée par le triplet précision/rappel/F_1-

mesure. La priorité est souvent donnée (par exemple dans (Lassalle, 2015)) au rappel au détriment de la précision, considérant que les faux positifs seront de toute façon détectés pendant la phase de détection des chaînes de coréférences. Il reste toutefois à déterminer ce que signifie « être correctement identifié » pour une mention trouvée par un système. (Recasens & Hovy, 2011) suggère deux interprétations :

Correspondance stricte une mention système correspond à une mention de référence si elle recouvre exactement le même empan de texte.

Correspondance indulgente une mention système S correspond à une mention de référence G si

1. S est inclus dans G
2. La tête syntaxique de G appartient à S

3 Expérimentations et résultats

Nous présentons ici les différentes expériences de détection des mentions que nous avons menées. Le tableau 1 présente (dans sa seconde partie) les résultats de ces expériences, dont les modalités sont détaillées par la suite, et — a titre indicatif — ceux de systèmes de référence. Notons cependant que ces systèmes n'ayant pas été développés dans les mêmes conditions, ni même pour les mêmes langues, une comparaison strictement quantitative aurait peu de sens. De même, nos expériences utilisant un apprentissage automatique n'ont été évaluées que sur 10 % d'ANCOR (le reste ayant servi à l'apprentissage), leurs résultats sont donc difficilement comparables avec ceux de nos autres expériences.

Tab. 1: Comparaison de systèmes de détection de mentions (correspondances strictes)

Système	Corpus	P(%)	R(%)	F_1(%)	Méthode
(Lassalle, 2015)	CoNLL-2012 (en)[1]	43,77	97,97	60,05	2.1.1[3, 2]
(Kummerfeld *et al.*, 2011)	OntoNotes (en)[1]	56,97	69,77	62,72	2.1.1[2]
(Ogrodniczuk *et al.*, 2014)	NKJP	66,04	63,99	65,00	2.1.2
(Soraluze *et al.*, 2012)	EPEC	76,85	78,59	77,58	2.1.2
(Uryupina & Moschitti, 2013)	CoNLL-2012 (ar)[1, 5]	66,0	66,1	66,1	2.1.1[4, 2]
(Uryupina & Moschitti, 2013)	CoNLL-2012 (zh)[1, 5]	68,9	71,3	70,1	2.1.1[4, 2]
(Uryupina & Moschitti, 2013)	CoNLL-2012 (ar)[1, 5]	31,07	90,67	46,28	2.1.1[3, 2]
(Nguyen *et al.*, 2016)	ACE 2005[5]	83,7[6]	81,8[6]	82,7[6]	2.1.3
Analyse syntaxique	ANCOR	57,28	77,07	65,72	2.1.1
Analyse syntaxique	ANCOR[5]	48,66	65,46	55,82	2.1.1
Chunks	ANCOR	45,99	36,78	40,87	2.1.2
Apprentissage	ANCOR (test)	89,09	20,04	32,72	2.1.3
Apprentissage	ANCOR (test)[5]	89,09	88,61	88,85	2.1.3

1. Ne tient pas compte des singletons
2. Utilise une analyse syntaxique de référence
3. Optimisé pour le rappel
4. Optimisé pour la précision
5. Sans mentions imbriquées
6. Tient compte de la catégorisation des mentions

3.1 Corpus

Nos travaux ont été effectués sur le corpus ANCOR (Muzerelle *et al.*, 2013), premier et actuellement seul corpus français annoté en chaînes de coréférences. Ce corpus a pour particularité d'être construit à partir d'oral conversationnel transcrit, ce qui pose un défi supplémentaire pour l'utilisation des méthodes évoquées dans la section 2.1, toutes développées pour le traitement du langage écrit. Une autre particularité notable par rapport à une grande partie des corpus utilisés dans l'état de l'art est qu'il n'existe à l'heure actuelle aucune analyse syntaxique de référence pour ANCOR. De ce fait, il ne nous est pas possible d'évaluer les performances en correspondance indulgente des méthodes que nous avons expérimentées.

3.2 Extraction depuis des analyses syntaxiques

Nous utilisons dans cette approche les analyses syntaxiques fournies par Talismane (Urieli, 2013), un analyseur statistique en dépendances pour le français. La méthode que nous utilisons est la suivante :

1. Repérer dans les sorties de l'analyseur les mots des catégories suivantes : noms communs (nc), noms propres (npp), clitiques sujets (cls), clitiques objets (clo), pronoms personnels (pro), pronoms relatifs (prorel), adjectifs numéraux (num). Ces mots seront les têtes des futurs candidats-mentions.
2. Pour chaque tête ainsi repérée, déterminer son extension maximale, en suivant récursivement les relations de dépendances.

Les résultats obtenus sont les suivants : $P \approx 57{,}28\%$, $R \approx 77{,}07\%$ et $F_1 \approx 65{,}72\%$.

Si la précision obtenue n'est pas entièrement satisfaisante, le rappel est comparable à l'état de l'art, même si la diversité des conventions d'annotation des mentions rend la comparaison difficile. Une analyse des erreurs produites par cette méthode indique qu'un mauvais traitement des disfluences est la plus grande cause de faux positifs, il paraît donc raisonnable de supposer que des heuristiques de rectification de frontières *a posteriori* amélioreraient significativement à la fois la précision et le rappel. De même, utiliser un analyseur syntaxique plus performant pour l'oral serait sans doute très profitable, nos résultats étant entièrement dépendants de la qualité de l'analyse syntaxique. La plupart des travaux cités dans le tableau 1 utilisent d'ailleurs des analyses syntaxiques manuelles, ou au moins corrigées manuellement plutôt que des analyses automatiques brutes. Cela dit, il reste parmi les candidats-mentions ainsi repérés un certain nombre d'expressions non-référentielles, notamment des *ils* impersonnels, qui rendent de toute façon nécessaire l'ajout en sortie d'un filtre adapté, utilisant par exemple ILIMP (Danlos, 2005).

3.3 Extraction depuis les chunks

Pour cette approche, nous utilisons les chunks nominaux (NP) détectés par SEM (Tellier *et al.*, 2012) directement comme candidats-mentions. Plus spécifiquement, nous utilisons le mode de détection des seuls NP, qui permet en réalité de détecter aussi des NP qui seraient inclus dans des chunks prépositionnels. Les modèles utilisés sont les modèles standards fournis avec SEM, appris sur le French Treebank (Abeillé *et al.*, 2003).

Les résultats obtenus par cette méthode sont les suivants : $P \approx 45{,}99\%$, $R \approx 36{,}78\%$, $F_1 \approx 40{,}87\%$.

Ces résultats, de prime abord assez décevants, sont toutefois à considérer en tenant compte des remarques faites dans la section 2.1.2 : les chunks ne tenant par définition pas compte des imbrications, ils ne peuvent représenter qu'une fraction des mentions et doivent nécessairement être traités et combinés pour reconstituer l'ensemble des mentions — et ce même dans l'hypothèse d'un chunking parfait. D'ailleurs, la qualité du chunking utilisé ici est loin d'être garantie, pour mémoire, (Tellier *et al.*, 2014) évaluait aux alentours de 77 % seulement la F_1-mesure du chunking de obtenu par les modèles standards de SEM sur un échantillon du corpus oral ESLO, à partir duquel est construit la plus grande partie d'ANCOR. Il n'est donc pas aberrant d'imaginer que la qualité des chunks nominaux que nous utilisons ici est largement sous-optimal, et que — là encore — l'utilisation de modèles spécifiquement appris pour le traitement de l'oral comme ceux décrits dans (Tellier *et al.*, 2014) pourraient grandement améliorer ces performances.

Du fait de ces limitations, il ne semble donc pas pertinent pour l'instant de comparer ces résultats avec ceux de travaux similaires, comme par exemple (Soraluze *et al.*, 2012).

3.4 Détection directe des frontières

La dernière méthode que nous avons expérimentée a consisté à apprendre directement le repérage des frontières de mentions, en le traitant comme une tâche d'étiquetage de séquences. Spécifiquement, en considérant chaque tour de parole du corpus comme une séquence de tokens, nous avons entraîné un modèle pour l'étiqueteur à CRF linéaire Wapiti (Lavergne *et al.*, 2010) associant à chaque token une étiquette du schéma BILOU (*Begin*, *Inside*, *Last*, *Outside*, *Unique*) utilisé classiquement pour l'extraction d'information par annotation, par exemple par (Nguyen *et al.*, 2016) pour la reconnaissance de mentions. Comme nous l'avons précisé en section 2.1.3, cette méthode ne permet pas non plus de prendre en compte les mentions imbriquées. Nous avons donc fait le choix de ne conserver pour cette expérience que les mentions les plus larges et d'ignorer leurs sous-mentions.

La procédure d'apprentissage a été réalisée sur une partition arbitraire du corpus ANCOR, suivant la répartition classique de 80 %/10 %/10 % respectivement pour les sous-corpus d'apprentissage (pour l'entraînement de l'étiqueteur), de développement (utilisé comme critère d'arrêt de l'apprentissage) et d'évaluation. En plus des formes des tokens, nous avons également utilisé comme traits pour l'apprentissage les parties du discours et les frontières de chunks et d'entités nommées déterminées par SEM.

Les résultats sont les suivants : $P \approx 89{,}09\%$, $R \approx 88{,}61\%$ et $F_1 \approx 88{,}85\%$ en ne tenant compte que des mentions maximales. On obtient $R \approx 20{,}04\%$ et $F_1 \approx 32{,}72\%$ en tenant compte de l'ensemble des mentions.

Ces résultats sont raisonnablement satisfaisants même si la comparaison avec l'état de l'art est délicate, les travaux en détection de coréférences prenant habituellement en compte toutes les mentions — y compris imbriquées — et les systèmes de suivi d'entités tenant en général compte de la catégorisation des entités en plus de leur repérage dans leurs évaluations.

4 Conclusion et perspectives

La détection automatique de chaînes de coréférences en français n'en est encore qu'à ses débuts, la mise à disposition de corpus annotés étant relativement récente. Le défi que représente le développement d'un système *end-to-end* adapté à cette nouvelle langue et au contexte spécifique de l'oral impose l'étude approfondie de cette tâche parfois négligée. Les expériences que nous avons exposées ici fournissent une base de travail pour des travaux futurs et montrent que des méthodes classiques exploitant des analyses syntaxiques peuvent donner des résultats encourageants, même dans ce nouveau contexte. Il reste à présent à bâtir sur ces fondations un système de détection des mentions capable de s'interfacer avec un système de détection de coréférences pour disposer d'un système *end-to-end* pour le français — par exemple en le nourrissant de traits issus des analyses syntaxiques utilisées pour la détection. La précision de la détection des mentions devra également être améliorée — et ce quelle que soit la méthode employée — en filtrant *a posteriori* les expressions non-référentielles erronément détectées.

Dans cette perspective, des travaux futurs devront probablement exploiter des ressources linguistiques plus adaptées — notamment au français oral — et employer des méthodes plus puissantes que celles que nous avons expérimentées ici, qu'il s'agisse de règles écrites à la main ou d'apprentissage automatique. Il serait ainsi intéressant de combiner plusieurs de ces approches dans un système hybride, ou de les appliquer en cascade — par exemple en appliquant des analyseurs syntaxiques directement sur les mentions maximales détectées par apprentissage. On pourrait également envisager de traiter conjointement détection des mention et détection des coréférences.

5 Remerciements

Cette recherche s'insère dans le programme « Investissements d'Avenir » géré par l'Agence Nationale de la Recherche ANR-10-LABX-0083 (Labex EFL).

Ce travail a par ailleurs bénéficié du soutien de l'ANR DEMOCRAT (Description et modélisation des chaînes de référence : outils pour l'annotation de corpus et le traitement automatique), projet ANR-15-CE38-0008.

Références

Abeillé A., Clément L. & Toussenel F. (2003). *Building a Treebank for French*, In A. Abeillé, Ed., *Treebanks : Building and Using Parsed Corpora*, p. 165–187. Springer Netherlands : Dordrecht.

Abney S. P. (1992). *Parsing By Chunks*, In R. C. Berwick, S. P. Abney & C. Tenny, Eds., *Principle-Based Parsing : Computation and Psycholinguistics*, p. 257–278. Springer Netherlands : Dordrecht.

Bove R. (2008). *Automatic speech parsing : study of disfluencies*. PhD thesis, Université de Provence - Aix-Marseille I.

Danlos L. (2005). ILIMP : Outil pour repérer les occurrences du pronom impersonnel il. In *Actes de la 12ème conférence sur le Traitement Automatique des Langues Naturelles*, p. 123–132, Dourdan, France : Association pour le Traitement Automatique des Langues.

de Marneffe M.-C., Recasens M. & Potts C. (2015). Modeling the lifespan of discourse entities with application to coreference resolution. *Journal of Artificial Intelligence Research*, **52**, 445–475.

Désoyer A., Landragin F., Tellier I., Lefeuvre A. & Antoine J.-Y. (2015). Coreference Resolution for Oral Corpus : a machine learning experiment with ANCOR corpus. *Traitement Automatique des Langues*, **55**(2), 97–121.

dos Santos C. N. & Carvalho D. L. (2011). Rule and tree ensembles for unrestricted coreference resolution. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning : Shared Task*, CONLL Shared Task '11, p. 51–55, Stroudsburg, PA, USA : Association for Computational Linguistics.

Finkel J. R. & Manning C. D. (2009). Nested named entity recognition. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing : Volume 1 - Volume 1*, EMNLP '09, p. 141–150, Stroudsburg, PA, USA : Association for Computational Linguistics.

Florian R., Pitrelli J. F., Roukos S. & Zitouni I. (2010). Improving mention detection robustness to noisy input. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, p. 335–345, Stroudsburg, PA, USA : Association for Computational Linguistics.

Jing H., Florian R., Luo X., Zhang T. & Ittycheriah A. (2003). HowtogetaChineseName(Entity) : Segmentation and combination issues. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, EMNLP '03, p. 200–207, Stroudsburg, PA, USA : Association for Computational Linguistics.

Kummerfeld J. K., Bansal M., Burkett D. & Klein D. (2011). Mention detection : Heuristics for the ontonotes annotations. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning : Shared Task*, CONLL Shared Task '11, p. 102–106, Stroudsburg, PA, USA : Association for Computational Linguistics.

Lassalle E. (2015). *Structured learning with latent trees : a joint approach to coreference resolution*. PhD thesis.

Lavergne T., Cappé O. & Yvon F. (2010). Practical very large scale CRFs. In *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, p. 504–513 : Association for Computational Linguistics.

Lee H., Chang A., Peirsman Y., Chambers N., Surdeanu M. & Jurafsky D. (2013). Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, **39**(4), 885–916.

Longo L. (2013). *Vers des moteurs de recherche "intelligents" : un outil de détection automatique de thèmes. Méthode basée sur l'identification automatique des chaînes de référence*. PhD thesis, Université de Strasbourg.

Muzerelle J., Lefeuvre A., Antoine J.-Y., Schang E., Maurel D., Villaneau J. & Eshkol I. (2013). Ancor, premier corpus de français parlé d'envergure annoté en coréférence et distribué librement. In *Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles*, p. 555–563, Les Sables d'Olonne, France : Association pour le Traitement Automatique des Langues.

Nguyen T. H., Sil A., Dinu G. & Florian R. (2016). Toward mention detection robustness with recurrent neural networks. *CoRR*, **abs/1602.07749**.

Ogrodniczuk M., Wójcicka A., Głowińska K. & Kopeć M. (2014). *Detection of Nested Mentions for Coreference Resolution in Polish*, In A. Przepiórkowski & M. Ogrodniczuk, Eds., *Advances in Natural Language Processing : 9th International Conference on NLP, PolTAL 2014, Warsaw, Poland, September 17-19, 2014. Proceedings*, p. 270–277. Springer International Publishing : Cham.

Poesio M., Stuckardt R. & Versley Y. (2016). *Anaphora Resolution : Algorithms, Resources, and Applications*. Theory and Applications of Natural Language Processing. Springer Berlin Heidelberg.

Qian D., Li W., Yuan C., Lu Q. & Wu M. (2007). Applying machine learning to Chinese entity detection and tracking. In *Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing '07, p. 154–165, Berlin, Heidelberg : Springer-Verlag.

Recasens M. (2010). *Coreference : Theory, Annotation, Resolution and Evaluation.* PhD thesis.

Recasens M. & Hovy E. (2011). Blanc : Implementing the Rand index for coreference evaluation. *Nat. Lang. Eng.*, **17**(4), 485–510.

Soraluze A., Arregi O., Arregi X., Ceberio K. & de Ilarraza A. D. (2012). Mention detection : First steps in the development of a Basque coreference resolution system. In J. Jancsary, Ed., *Proceedings of KONVENS 2012*, p. 128–136 : ÖGAI.

Tellier I., Duchier D., Eshkol I., Courmet A. & Martinet M. (2012). Apprentissage automatique d'un chunker pour le français. In *Actes de la 19e conférence sur le Traitement Automatique des Langues Naturelles*, p. 431–438, Grenoble, France : Association pour le Traitement Automatique des Langues.

Tellier I., Eshkol-Taravella I., Dupont Y. & Wang I. (2014). Peut-on bien chunker avec de mauvaises étiquettes POS ? In *Actes de la 21e conférence sur le Traitement Automatique des Langues Naturelles*, p. 125–136, Marseille, France : Association pour le Traitement Automatique des Langues.

Trouilleux F. (2001). *Referential links identification and automatic interpretation of pronominal expressions in French texts*. PhD thesis, Université Blaise-Pascal, Clermont-Ferrand.

Urieli A. (2013). *Robust French syntax analysis : reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. PhD thesis, Université de Toulouse II le Mirail.

Uryupina O. & Moschitti A. (2013). Multilingual mention detection for coreference resolution. In *IJCNLP*, p. 100–108.

Versley Y., Ponzetto S. P., Poesio M., Eidelman V., Jern A., Smith J., Yang X. & Moschitti A. (2008). BART : A modular toolkit for coreference resolution. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies : Demo Session*, HLT-Demonstrations '08, p. 9–12, Stroudsburg, PA, USA : Association for Computational Linguistics.

Un nouveau besoin dans l'industrie : une aide au « rédacteur traduisant »

Claire Lemaire[1]
(1) ILCEA4, Université Grenoble Alpes, 38058 Grenoble, France
claire.lemaire1@univ-grenoble-alpes.fr

RESUME

Nous mettons en relief, grâce à une expérimentation avec questionnaires et corpus parallèle, une situation nouvelle en entreprise de rédaction multilingue, pour laquelle il n'existe pas de technologie TAL dédiée. Nous suggérons de tirer profit de cette situation inédite de rédacteur traduisant, pour utiliser l'expertise du rédacteur pendant le processus de traduction et nous préconisons de développer une TA permettant une édition en cours de processus.

ABSTRACT

A new need in companies: writing assistance for "translating writers".
We highlight, thanks a experiment through questionnaires and parallel corpus, a new situation of multilinguistic writing in French companies, for which no dedicated NLP technology is provided. We propose to take advantage of this unprecedented situation of having a "translating writer", to use the writer expertise during the translation process and we suggest to create a new MT that allow editing during the process.

MOTS-CLES : TAL et multilinguisme, TAO, TA, TAFD.
KEYWORDS: NLP and multilinguism, CAT, MT, DBMT.

Introduction

Les technologies de TAL actuelles sont prévues, soit pour des rédacteurs monolingues qui n'effectuent pas de traduction, soit pour des traducteurs qui ne rédigent pas le texte devant être traduit. Nous pensons qu'il est temps de s'intéresser à une production bilingue. Nous justifions notre approche par une enquête de terrain révélant une situation de plus en plus fréquente dans les entreprises où il est demandé aux cadres de produire des documents bilingues, essentiellement en français et en anglais.

Dans une première partie nous présentons deux enquêtes effectuées auprès de cadres en entreprise et auprès de traducteurs professionnels sur leurs usages en outils d'aide à la traduction. Dans une deuxième partie, nous décrivons les différents outils et leurs différents types d'utilisateurs, et la naissance d'un nouveau type d'utilisateur industriel. Dans la dernière partie, nous discutons de fonctionnalités qui pourraient répondre à ce nouveau besoin.

1 Questionnaires

Les questionnaires sont largement utilisés pour étudier des pratiques de traducteurs et depuis longtemps : Slocum a effectué un sondage sur la TA (Slocum, 1985), Vasconcellos un autre sur les usages de la TA (Vasconcellos, 1993), Lagoudaki a observé l'apport de la traduction automatique (Lagoudaki, 2008), Chan s'est penché sur ce qu'apporte une certification en traduction (Chan, 2010), Dam et Zethsen ont recueilli des commentaires de traducteurs sur leur statut social (Dam, Zethsen, 2010), Frérot et Karagouch ont étudié l'adéquation entre formation universitaire et besoins du monde professionnel (Frérot, Karagouch, 2016). Les questionnaires sur les pratiques linguistiques des salariés ont été utilisés par un certain nombre de chercheurs : Harzing et Pudelko ont eu recours aux questionnaires sur les pratiques linguistiques en entreprise pour pointer les barrières langagières (Harzing, Pudelko, 2012) et proposé des solutions pour pallier le grave manque à gagner engendré par des déficiences linguistiques en entreprise (Tenzer, Pudelko, Harzing, 2014).

1.1 Objectif visé

L'objectif visé par ces questionnaires est de modifier un outil d'aide à la traduction dédié à des traducteurs professionnels de façon à ce qu'il soit utilisable par des traducteurs non professionnels. Ces travaux ont été financés par un éditeur de logiciel, lequel a également financé plusieurs autres études : ainsi il a par exemple intégré une fonctionnalité de TA à son logiciel déjà commercialisé depuis plusieurs années de mémoire de traduction. Plus précisément, l'objectif de l'éditeur est de toucher, au sein d'une entreprise ayant déjà des traducteurs professionnels utilisant son outil, des collègues cadres traduisant, afin de leur proposer son logiciel.

1.2 Méthode utilisée

Nous avons envoyé des questionnaires à des cadres travaillant en entreprise, n'exerçant pas dans le domaine de la traduction et n'ayant pas non plus suivi de formation en traduction, devant traduire dans le cadre de leur profession au moins une fois par semaine et utilisant des technologies de TAL. Les cadres sélectionnés travaillent avec au moins deux langues parmi l'allemand, l'anglais et le français. Ainsi, 34 cadres ont été sélectionnés. Le questionnaire est composé de 5 sections visant principalement à récolter des informations sur les usages des outils en général et les fonctionnalités en particulier des technologies de traduction. Parallèlement, nous avons envoyé un questionnaire semblable à un échantillon de traducteurs professionnels représentant 8,5 % des traducteurs de la SFT[1], soit 10 traducteurs, travaillant entre l'allemand, l'anglais ou le français.

1.3 Résultats obtenus

L'outil principal utilisé par le panel de traducteurs est SDL Trados pour 71 % des traducteurs ; puis en deuxième position, viennent Transit de Star et Wordfast (d'Yves Champollion) avec chacun 11 % d'utilisateurs. En comparaison, la première technologie utilisée par les experts métier en entreprise est Google Translate, avec 45 %, suivi de Word (Bing) avec 14 % et Reverso avec 13 %.

[1] Société française des traducteurs, syndicat, qui entre autres, répertorie les traducteurs en France

La majorité des traducteurs professionnels interrogés (70 %) n'utilisent pas d'outil secondaire. Cependant, 18 % des répondants utilisent SDL Trados comme second outil et 6 % utilisent Wordfast. Les experts métier, quant à eux, sont plus nombreux à utiliser un outil secondaire, à savoir un dictionnaire en ligne pour 30 %, puis trois des outils cités précédemment : Google Translate à 20 %, puis Linguee et Reverso à 13 % chacun. La traduction automatique semble ainsi omniprésente chez les experts-métier alors qu'elle n'est qu'anecdotique chez les traducteurs professionnels spécialisés.

En dehors de l'outil principal, les fonctionnalités les plus utilisées chez les cadres sont le concordancier à 74 %, puis le correcteur orthographique et grammatical à 70 %, puis les dictionnaires bilingues, également à 70 %, suivis de la traduction automatique (TA) à 52 % et enfin des dictionnaires auto-suggest à 39 %. Chez les traducteurs, les fonctionnalités les plus utilisées sont le correcteur orthographique et grammatical, puis les dictionnaires bilingues, chacun à 90 %, puis le concordancier, employé à 70 %, puis les dictionnaires auto-suggests 40%, et enfin la fonctionnalité de traduction automatique, utilisée par 10 % des traducteurs interrogés.

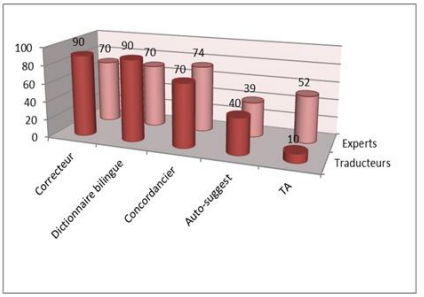


Figure 1 : Pourcentage d'utilisation des fonctionnalités par les cadres (« experts ») et les traducteurs

Le résultat sur les fonctionnalités le plus flagrant est donc, comme pour l'outil principal, la différence d'utilisation de la TA, avec 52 % chez les cadres contre 10 % chez les traducteurs professionnels. La comparaison entre les autres fonctionnalités montre peu de différence entre les pratiques des experts et celles des traducteurs interrogés. En effet, les trois premières fonctionnalités sont les mêmes, les correcteurs orthographiques et grammaticaux, les dictionnaires bilingues, et les concordanciers.

Dans un second temps, pour affiner l'analyse des différences entre les deux populations, nous avons été dans une entreprise en immersion pendant trois mois pour chercher des textes bilingues rédigés par ces cadres. Nous avons ainsi découvert un nouveau type d'utilisateur : un cadre ayant un bon niveau dans une autre langue et à qui il est demandé de produire directement des documents bilingues, sans avoir recours à un traducteur, sans soutien linguistique et sans temps alloué spécifiquement à la tâche de traduction.

2 Outils existants pour le rédacteur traduisant

2.1 Deux paradigmes

Lorsqu'il s'agit de produire des traductions de textes, les systèmes existant aujourd'hui peuvent être classés selon deux paradigmes : (1) les outils qui proposent plusieurs traductions de phrases plus

ou moins proches de celle que l'utilisateur cherche à traduire (diffusion) ou à comprendre (acquisition) selon la situation ; (2) ceux qui proposent à l'utilisateur une traduction personnalisée et le plus souvent unique du segment venant d'être saisi. Notons cependant que la plupart des systèmes ont une option permettant de proposer plusieurs possibilités de traduction, surtout lexicales, lorsque le contexte ne permet pas au système de désambiguïser avec sûreté. Ces deux types d'outils ont un paradigme commun de logique séquentielle constitué de deux étapes : une production dans une langue 1, puis une production dans une langue 2.

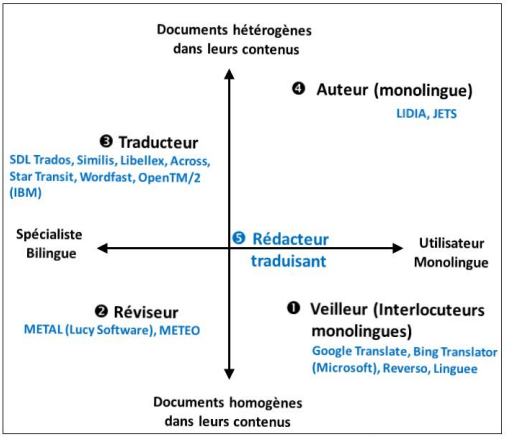


Figure 2 : Les utilisateurs déjà connus et le rédacteur traduisant

2.2 Quatre types d'utilisateurs

Si l'on classe maintenant les systèmes existants selon les utilisateurs, d'après Blanchon (Blanchon, 2004), nous avons : (1) La traduction du veilleur monolingue ou la traduction de dialogues oraux finalisés[2] (qui correspond à une traduction 100% automatique, c'est-dire la TA). Les situations pour lesquelles cette traduction est prévue sont la veille ou la navigation sur Internet par exemple. (2) et (4) La traduction du réviseur bilingue et la traduction de l'auteur monolingue qui produisent respectivement des traductions « brutes », destinées à être révisées, et des traductions ne nécessitant pas de révision, en faisant appel à l'auteur au moyen de désambiguïsation interactive en langue source (qui correspond à une traduction 75-90% automatique). (3) La traduction du traducteur. Elle est constituée d'outils destinés à un utilisateur humain bilingue (qui correspond à une traduction 15-50% automatique). En 2017, à notre connaissance, les outils conçus continuent de s'adresser à un utilisateur soit absolument monolingue, soit bilingue spécialisé en traduction. Nous avons situé par rapport au schéma de Blanchon, quelques outils : Google Translate, Bing Translator (Microsoft), METAL (Lucy Software), Meteo, SDL Trados 3, Similis, Libellex, Across, Star Transit, Wordfast, OpenTM/2 (IBM)4, Lidia et Jets.

[2] c'est-à-dire des systèmes prévus pour des interlocuteurs monolingues dans différentes langues, tels que des systèmes de réservations de billets d'avion par exemples
[3] ou SDL Studio ou SDLX5
[4] anciennement TM/2

Notre cas d'étude présente un cinquième type d'utilisateur, ni absolument monolingue, ni bilingue spécialisé en traduction, devant traiter des documents au contenu varié mais tout de même issus d'un même domaine de spécialité.

3 Outil à créer pour le rédacteur traduisant

Cette situation nouvelle de traduction par le rédacteur, sans budget ou temps supplémentaire, et sans que le rédacteur soit en contact avec le lecteur en langue cible, présente un atout majeur. En effet, contrairement au traducteur classique, qui n'a pas le droit de modifier le texte source, le rédacteur traduisant en est maître. Or s'il est connu que le besoin de certaines informations nécessaires au lecteur cible n'apparaît qu'au cours de la traduction, posant le problème de la désambiguïsation, il est tout à fait nouveau de pouvoir agir sur le texte source en fonction du résultat obtenu en langue cible. Ainsi le rédacteur traduisant pourrait, après une première TA, découvrir au vu du résultat, qu'il lui faut retoucher le texte source pour : (1) ajouter des informations extralinguistiques, c'est-à-dire désambiguïser ; (2) ajouter d'autres informations linguistiques apparues manquante par l'alignement (conséquent à la TA) et reprendre des fautes d'orthographe ou de grammaire : c'est-à-dire améliorer la rédaction. Puis le rédacteur pourra effectuer une deuxième TA.

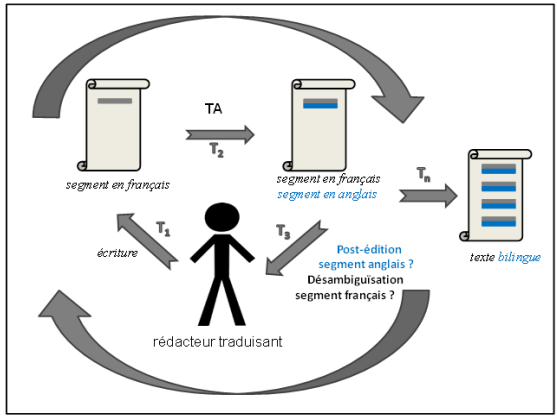


Figure 3 : Pré-édition et post-édition itérative de chaque segment

3.1 Désambiguïsation interactive

Avec cette pré-édition et post-édition permettant d'agir à la fois sur le texte source et sur le texte cible, le rédacteur traduisant pourrait lever les ambiguïtés au fur et à mesure du processus de traduction, non pas préalablement au transfert de l'ensemble du texte mais au vu d'un résultat partiel (segment par segment), de façon itérative jusqu'à ce qu'il soit satisfait. Le rédacteur-traduisant améliore d'autant plus facilement le texte en langue cible et celui en langue source, qu'il s'agit de son domaine d'expertise. En effet, le traduisant possède une excellente connaissance de toutes les informations extralinguistiques que peut fournir le rédacteur, et le rédacteur prend conscience, au vu des résultats de la TA, des besoins en information que ce processus de traduction requiert. Nous sommes convaincu qu'un utilisateur même ne connaissant pas parfaitement la langue d'arrivée pourra ainsi résoudre un nombre considérable d'erreurs. Le cadre endosse ainsi tour à tour le rôle de

rédacteur et de traducteur, mais en général pas celui de « réviseur ». Nous entendons ici par réviseur, un senior très compétent dans le domaine. Sa langue maternelle est la langue cible, et il maîtrise la langue source au moins au niveau C1. Dans notre cas, le profil type du rédacteur traduisant fait qu'il ne peut pas être un réviseur.

À notre connaissance, le seul outil de TA avec désambiguïsation interactive est une maquette ancienne, LIDIA-1, développée en Ariane-G5, un système de développement et d'exploitation d'applications propriétaire, constitué de 5 langages de programmation[5] conçus pour le développement de technologies de la traduction (Blanchon, 1994) et (Boitet, Blanchon, 1994). Le rédacteur aidait le système par une désambiguïsation de tous les termes qui étaient susceptibles de poser un problème au cours de la transposition dans les autres langues. Il s'agissait d'une désambiguïsation par étapes, qui avait lieu avant la fin de l'étape d'analyse (passage à une structure abstraite multiniveau de Vauquois) et donc avant l'étape de transfert (quelle que soit la ou les langues cible). La désambiguïsation reposait sur un processus de clarification, fondé sur un mécanisme de reconnaissance de patrons. À chaque patron, était associée une méthode de production de dialogue en langue naturelle et demandait ainsi à l'utilisateur traduisant de fournir des informations supplémentaires permettant de lever les ambiguïtés du texte en langue source. Il fallait alors prévoir toutes les ambiguïtés possible à l'avance et les lever avant la phase de transfert.

Nous préconisons de construire un système beaucoup plus simple, avec une seule paire de langue (dans les deux sens) et non plus avec une désambiguïsation interactive de tout le texte en une fois en amont d'un unique passage en TA, mais avec une désambiguïsation d'un seul segment, avant chaque passage en TA. L'environnement technologique consisterait également en un moniteur interactif lié à un système de gestion de base de données et à un traitement de textes. Le processus de traduction automatique se décomposerait également en trois étapes successives, classiques en TA : l'analyse de la langue source, le transfert et la génération en langue cible. L'unité de traduction en revanche ne dépasserait pas le segment. L'analyse de la langue source resterait monolingue. Lors de la phase de transfert, les traitements grammaticaux seraient également séparés des traitements lexicaux, et la spécialisation des grammaires aurait lieu en fonction de la typologie, et la spécialisation des dictionnaires aurait lieu en fonction du domaine. La génération en langue cible serait également monolingue. En fonction du résultat, et de la prise de conscience de certaines ambiguïtés, il est en revanche possible en plus de la désambiguïsation interactive, de modifier le segment source. Certaines ambiguïtés ne seraient donc pas traiter informatiquement, mais seraient annulées par la reformulation d'une phrase univoque en entrée du système.

3.2 Aide à la rédaction

En outre, il s'agirait d'agir sur des fonctionnalités de : (1) déclenchement systématique par défaut d'un affichage des synonymes possibles des mots source et des mots cible par info-bulle ; (2) déclenchement systématique de la détection de langue et l'affichage par défaut de la correction grammaticale et orthographique au cours de la frappe ; (3) création automatique de dictionnaires auto-suggest, en les créant systématiquement par défaut, que les utilisateurs les demandent ou non. De nombreux outils de traitement de textes possèdes déjà la plupart de ces fonctionnalités, pour le rédacteur traduisant, qui les maîtrise mal car n'étant pas spécialiste de la rédaction, comme un universitaire où un traducteur, il faudrait les activer par défaut. Par ailleurs, les travaux de Hong-Thai Nguyen (Nguyen, 2009), mais également de Phap-Cong Huyn (Huynh, 2010), et plus récemment de Ying Zhang (Zhang, 2016), montrent que ces efforts ergonomiques sont étudiés et partiellement

[5] ATEF, ROBRA, EXPANS/TRANSF, SYGMOR et TRACOMPL

développés (dans l'outil « aximag »), en particulier avec l'introduction de « mini-dictionnaires », un pour chaque segment, préparé à l'avance. À l'instar de ce que proposent des outils tels que Leslie ou Antidote, l'outil du rédacteur traduisant proposerait une cohérence lexicale, d'ajouter des virgules, de couper les phrases en deux, ou encore de mettre des parenthèses, suivant les résultats sémantiques obtenus lors de la phase de transfert du segment.

Par ailleurs nous proposons de créer deux nouvelles fonctionnalités. Ainsi, il s'agirait de créer des fonctionnalités de : (1) recherche d'information lexicologique, c'est-à-dire une recherche de vocabulaire de spécialité sur Internet directement en fonction des ajouts de l'utilisateur aux dictionnaires auto-suggest. Les outils linguistiques sont généralistes, or un cadre en entreprise utilise un vocabulaire récurrent ; (2) d'affichage, c'est-à-dire de présentation bilingue du document avec affichage ou masquage d'une langue ou de l'autre. Il s'agirait de concevoir un mode d'affichage d'une seule langue. Il serait ainsi plus aisé pour le rédacteur-traducteur de vérifier ce que le lecteur anglophone verra apparaître en anglais, bien qu'il n'y ait pas de mise en regard du texte ; puis de vérifier ce que le lecteur francophone verra apparaître.

Conclusion

Les résultats de notre expérimentation à l'aide de questionnaires et d'un corpus de traduction nous a fait découvrir une situation tout à fait inédite de traduction par un traducteur ayant la possibilité de retoucher le texte source. Nous proposons de tirer profit de cette situation de rédacteur traduisant, pour utiliser l'expertise de l'utilisateur et de l'injecter dans le processus de traduction par deux moyens : une désambiguïsation interactive et une modification du texte source. Nous proposons également une aide à la rédaction bilingue mettant en évidence les manques de traduction et autres lacunes lexicales, orthographiques et grammaticales par de la recherche d'information, et un nouvel affichage en plus d'autres fonctionnalités traditionnelles.

Les perspectives de recherches sont la création d'un « proof of concept » sur le même domaine de spécialité, le même couple de langue et le même sens de traduction du système de désambiguïsation interactive évoqué et des fonctionnalités d'aide à la rédaction bilingues proposées. Ensuite, si les tests sont concluant, bien que nous ayons l'intuition qu'il serait impossible d'intégrer l'ensemble des fonctionnalités proposées à un traitement de texte (par exemple la désambiguïsation interactive, pour des raisons techniques, au niveau de la performance en particulier), les recherches pourraient se poursuivre par une étude permettant de choisir si l'ensemble ou partie de ces fonctionnalités seraient à intégrer à l'outil déjà commercialisé ou de créer un nouvel outil dédié pour les cadres.

Références (style *Titre sans numéro*)

BLANCHON H. (2004). Comment définir, mesurer et améliorer la qualité, l'utilisabilité et l'utilité des systèmes de TAO de l'écrit et de l'oral. Une bataille contre le bruit, l'ambiguïté, et le manque de contexte. HDR, Université Joseph Fourier, Grenoble.

BLANCHON H. (1994). *LIDIA-1 : une première maquette vers la TA Interactive" pour tous".* Thèse de doctorat, Université Joseph-Fourier-Grenoble I, Grenoble.

BOITET C., BLANCHON H. (1994). Promesses et problèmes de la « TAO pour tous ». Après LIDIA-1, une première maquette". *Langages. Le traducteur et l'ordinateur, sous la direction de Jean-René Ladmiral, 116*, pp. 20–47.

CABRÉ M. T. (2007). Constituer un corpus de textes de spécialité. *Cahiers du CIEL*, 37–56.

CHAN A. L. (2010). Perceived benefits of translator certification to stakeholders in the translation profession: A survey of vendor managers. *Across Languages and Cultures, 11*(1), 93–113.

DAM H., ZETHSEN K. (2010). Translator status: Helpers and opponents in the ongoing battle of an emerging profession. *Target, 22*(2), 194–211.

FREROT C., KARAGOUCH L. (2016). Outils d'aide à la traduction et formation de traducteurs : vers une adéquation des contenus pédagogiques avec la réalité technologique des traducteurs. *ILCEA Revue de l'Institut des langues et cultures d'Europe, Amérique, Afrique, Asie et Australie, 27* , pp. 120–140.

HARZING A.-W., PUDELKO M. (2012). Language competencies, policies and practices in multinational corporations: A comprehensive review and comparison of Anglophone, Asian, Continental European and Nordic MNCs. *Journal of World Business, 48*(1).

HUYNH C.-P. (2010). *Des suites de test pour la TA à un système d'exploitation de corpus alignés de documents et métadocuments multilingues, multiannotés et multimédia.* Thèse de doctorat, Université Joseph Fourier.

KOSKINEN K. (2008). *Translating Institutions: An Ethnographic Study of EU Translation.* Manchester: St Jerome Publishing.

LAGOUDAKI, E. (2009b). *Expanding the possibilities of translation memory systems: from the translators wishlist to the developers design.* Thèse de doctorat, Imperial College London.

NGUYEN H.-T. (2009). *Des systèmes de TA homogènes aux systèmes de TAO hétérogènes. Interface homme-machine U.* Thèse de doctorat, Université Joseph-Fourier - Grenoble I.

SALDANHA G., O'BRIEN S. (2013). *Research methodologies in translation studies.* London: Routledge.

SLOCUM J. (1985). A survey of machine translation: its history, current status, and future prospects. *Computational linguistics 11.1 (1985): 1-17., 11*(1), pp. 1–17.

STURGE K. (2004). *The Alien Within: Translation into German during the Nazi Regime*. Munich: Ludicium.

TENZER H., PUDELKO M., HARZING A.-W. (2014). The impact of language barriers on trust formation in multinational teams. *Journal of International Business Studies, 45*(5), 508–535.

VASCONCELLOS M. (1993). The present state of machine translation usage technology; or: How do I use thee? Let me count the ways. *MT Summit IV, July 20–22*, (pp. 47–62). Kobe, Japan.

YIN R. (2009). *Case Study Research: Design and Methods* (éd. 4e). Los Angeles, London, etc.: Sage.

ZHANG Y. (2016). *Modèles et outils pour des bases lexicales" métier" multilingues et contributives de grande taille, utilisables tant en traduction automatique et automatisée que pour des services dictionnairiques variés*. Thèse de doctorat, Université Grenoble Alpes.

Adaptation incrémentale de modèles de traduction neuronaux

Christophe Servan Josep Crego Jean Senellart

SYSTRAN, 5 Rue Feydeau, 75002 Paris, France
prenom.nom@systrangroup.com

RÉSUMÉ

L'adaptation au domaine est un verrou scientifique en traduction automatique. Il englobe généralement l'adaptation de la terminologie et du style, en particulier pour la post-édition humaine dans le cadre d'une traduction assistée par ordinateur. Avec la traduction automatique neuronale, nous étudions une nouvelle approche d'adaptation au domaine que nous appelons "spécialisation" et qui présente des résultats prometteurs tant dans la vitesse d'apprentissage que dans les scores de traduction. Dans cet article, nous proposons d'explorer cette approche.

ABSTRACT

Incremental adaptation of neural machine translation models

Domain adaptation is a key feature in Machine Translation. It generally encompasses terminology and style adaptation, especially for human post-editing workflows in Computer Assisted Translation (CAT). With Neural Machine Translation (NMT), we study a new approach of domain adaptation that we call "specialization" and which is showing promising results both in the learning speed and in adaptation accuracy. In this paper, we propose to explore this approach.

MOTS-CLÉS : Traduction Automatique Neuronale, Adaptation au Domaine, étude préliminaire.

KEYWORDS: Neural Machine Translation, Domain Adaptation, preliminary study.

1 Introduction

Suite aux avancées de la traduction automatique neuronale (TAN) ces dernières années (Kalchbrenner & Blunsom, 2013; Sutskever *et al.*, 2014; Bahdanau *et al.*, 2014), nous observons que les performances obtenues dans les campagnes d'évaluations sont comparables, voire meilleures, que les approches de traduction automatique statistiques (TAS) (Cettolo *et al.*, 2015; Bojar *et al.*, 2015, 2016). Ces succès obtenus nous permettent d'explorer des problématiques qui se posaient déjà avec la TAS. L'une de ces problématiques concerne l'adaptation au domaine.

Malgré la quantité de données disponibles, ces dernières sont généralement hors-domaine. L'idée sous-jacente est de mieux utiliser les données spécifiques à un domaine, conjoitement aux données hors-domaine. Intuitivement, la première idée consiste à combiner les données hors-domaine à celles du domaine. Or, outre la question de la durée d'apprentissage (qui peut durer jusqu'à plusieurs semaines), se pose surtout celle de la représentativité des données du domaine par rapport aux données hors-domaine.

Contributions Cet article propose une méthode permettant d'améliorer la représentativité des données du domaine par rapport aux données hors domaine en se basant sur les précédents travaux de réentrainement initialement proposé par (Luong & Manning, 2015). Plus généralment, nous présentons une étude de cette approche d'adaptation incrémentale s'insérant dans le cadre expérimental de la traduction assistée par ordinateur.

Les expériences sont tiennent compte d'un domaine spécialisé : des transcriptions de présentations (issues de TED). La tâche de traduction du cadre expérimental sont de l'anglais vers le français et de l'anglais vers l'allemand. Les données sont issues de campagnes d'évaluations ou de dépôts libres de droits et seront mis à disposition afin d'assurer la reproductibilité des résultats.

Cet article débute par une étude de l'état de l'Art, suivit de la description de l'approche. Dans la section 4 nous présentons notre protocole d'évaluation, contenant une étude des résultats et une discussion autour de l'approche. Enfin nous conclurons cet article.

2 État de l'Art

Ces dernières années, les techniques d'adaptation au domaine pour la traduction ont été particulièrement étudiées. Ces approches peuvent être appliquées à trois différents niveaux : avant, pendant ou après la phase d'apprentissage initiale. Dans le cadre d'une traduction assistée par ordinateur (TAO), la plupart des approches sont des pré-traitements ou des post-traitements de la phase d'apprentissage initiale.

Ainsi, les pré-traitements concernent généralement les données d'apprentissage. Citons en exemple la sélection de données proposée par Lü *et al.* (2007) et améliorée par Gao & Zhang (2002) et bien d'autres (Moore & Lewis, 2010; Axelrod *et al.*, 2011). Elle est connue pour être efficace (Lambert *et al.*, 2011; Cettolo *et al.*, 2014) malgré son principal défaut : les mots et séquences de mots rares ne sont pas identifiés et sont donc rejetés (Wuebker *et al.*, 2014). En cela, l'approche permettant l'extraction de séquences de mots rares (Gascó *et al.*, 2012) palie en partie à ce problème (Wuebker *et al.*, 2014). Cependant, ce type d'approche ne nous permet pas d'adapter un modèle de traduction générique déjà entrainé.

Lors de l'apprentissage, il existe des approches permettant de pondérer des données d'apprentissage au niveau du corpus (du document) ou de la phrase (Shah *et al.*, 2010). Mais le principal inconvénient de ce type de pré-traitement est la nécessité de relancer la phase d'apprentissage pour que celle-ci soit efficace. Or, pour des besoins pratiques, cette étude focalise sur la possibilité de profiter d'un modèle déjà entrainé afin de l'adapter à un domaine particulier. Cependant, l'idée de caractériser les données d'apprentissage selon un domaine particulier, n'est pas nouvelle. Cette idée se rapproche des travaux proposés par Sennrich *et al.* (2016a) qui visent à adapter le modèle de traduction en fonction du niveau de politesse dans la phrase. Dans ces travaux ils utilisent un marqueur de niveau de language, ce qui rend toute adaptation postérieure à l'apprentissage d'un nouveau domaine compliquée. Enfin, Zhang *et al.* (2016) proposent l'utilisation de marqueur afin de caractériser les thématiques de traduction.

L'approche, dite de « spécialisation », étudiée ici, appartient à la famille de techniques de post-traitement qui ont pour but la mise-à-jour ou l'optimisation de paramètres d'un modèle de traduction existant. Elle est dérivée de celle proposée par Doulamis *et al.* (2000) pour le traitement de l'image et appliquée à la traduction une première fois avec succès par Luong & Manning (2015). Ces derniers proposent d'entrainer un modèle puis de ré-entrainer ce même modèle sur un corpus entier spécifique à un domaine particulier à travers une douzaine de cycles supplémentaires. Cette approche paraît

adaptée à la problématique d'une mise-à-jour rapide d'un modèle. A notre connaissance, il n'existe pas de travaux qui ont étudié l'impact de cette approche et son comportement. Nous proposons donc d'étudier cette approche dans le cadre d'une mise-à-jour rapide du modèle en faisant varier les quantités de données.

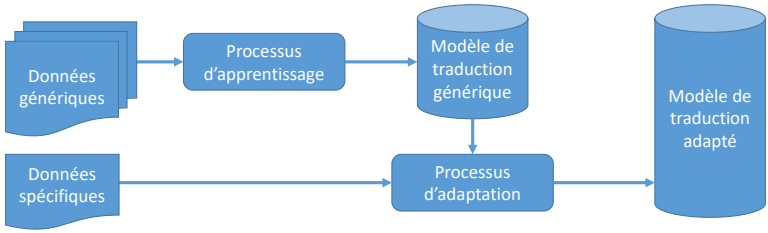


FIGURE 1 – Le modèle de traduction générique est appris sur les données génériques, puis, ce modèle générique est ré-entrainé avec des données spécifiques au domaine afin d'obtenir un modèle adapté au domaine (ou à la tâche).

3 Approche

L'approche de spécialisation vise à adapter de manière directe ou incrémentale un modèle de traduction neuronal générique en s'inspirant des travaux sur l'adaptation incrémentale en traduction automatique statistique (Blain *et al.*, 2012; Cettolo *et al.*, 2014). Ces approches s'insèrent dans un cadre applicatif particulier : la traduction assistée par ordinateur. Dans ce cadre, le traducteur corrige ou post-édite les sorties d'un système de traduction automatique. Ces travaux ont montré l'intérêt d'ajouter de manière incrémentale de nouvelles informations dans un modèle de traduction fondé sur les paires de segments. Ces informations peuvent être d'ordre lexical ou stylistique, ce qui peut déboucher éventuellement sur une adaptation au style du traducteur.

L'idée principale de l'approche est donc de spécialiser un modèle générique déjà entrainé sur des données génériques (i.e. : venant de plusieurs domaines différents). La figure 1 présente le cycle d'adaptation que nous utilisons lors de nos expériences.

Lors de ce processus de spécialisation, sont réestimées les probabilités conditionnelles $p(y_1, \ldots, y_m | x_1, \ldots, x_n)$ où $(x_1, \ldots, x_n)$ est la séquence d'entrée de taille n et $(y_1, \ldots, y_m)$ est la séquence de sortie correspondante de taille m. L'adaptation a donc comme point de départ l'état du réseau de neurones, d'une manière similaire à une pré-configuration des paramètres du modèle.

Le résultat est un modèle adapté (ou spécialisé) à une tâche ou un domaine particulier.

4 Expériences

Ne disposant pas de données post-éditées et afin d'être le plus reproductible possible, nous proposons d'étudier l'approche de spécialisation sur les données issues de l'évaluation IWSLT 2015 (Cettolo *et al.*, 2015), utilisant les traductions de transcriptions de vidéos de TED. Les paires de langues concernées sont l'anglais vers le français et l'anglais vers l'allemand. La mesure de comparaison

Corpus	Nbr lignes	Nbr mots (en)	Nbr mots (fr)	Nbr lignes	Nbr mots (en)	Nbr mots (de)
BOOKS	120 K	2,9 M	3,0 M	46,1 K	1,1 M	1,1 M
Europarl V.7	2 M	55,6 M	64,1 M	1,9 M	53,6 M	51,0 M
IT	371 K	4,3 M	5,3 M	218 K	2,7 M	2,7 M
JRC	603 K	16,2 M	18,4 M	601 K	16,8 M	15,2 M
News-Commentary V.11	203 K	5,1 M	6,4 M	220 K	5,5 M	5,6 M
Total Générique	3,3 M	84,3 M	97,4 M	3,0 M	80 M	75,8 M
WIT^3	217,9 K	4,5 M	4,9 M	192,7 K	4,0 M	3,6 M
IWSLT2014 (test)	1305	25 K	29 K	1305	23 K	25 K

TABLE 1 – Détail des données d'apprentissage pour les tâches anglais-français et anglais-allemand. Les cinq premiers corpus sont associés au domaine générique. Le corpus WIT^3 est le corpus du domaine spécifique sur lequel nous souhaitons adapter le modèle générique. Le corpus de test est celui de l'évaluation IWSLT2014. Les notations K et M correspondent respectivement à 10^3 et 10^6.

utilisée dans cette étude est le score BLEU (Papineni *et al.*, 2002), bien que contreversée, elle reste la mesure étalon lors des dernières campagnes d'évaluation (Bojar *et al.*, 2016; Cettolo *et al.*, 2016).

L'approche de traduction automatique neuronale (TAN) utilisée dans cet article est fondée sur une approche séquence-vers-séquence (Sutskever *et al.*, 2014) combinée à une approche d'attention (Luong *et al.*, 2015), à travers l'outil *OpenNMT* (Klein *et al.*, 2017).

Les données sont pré-traitées en utilisant une méthode de découpage en sous-mots issue de l'algorithme de compression de données : « *byte pair encoding* »(Sennrich *et al.*, 2016b) avec 30 000 opérations. Nous avons conservé les 32 000 mots plus fréquents pour les langues source et cible. Le modèle utilise une architecture récurrente (bi-LSTM) de 4 couches cachées avec 800 cellules chacune. Lors de l'apprentissage, le mini-batch est fixé à 64 avec une régularisation de type « dropout » fixée à 0,3. Les modèles génériques sont appris sur 18 cycles avec un taux d'apprentissage à 1,0 auquel il est appliqué une dégradation de facteur 0,7 à partir du cycle 10. Le système générique, utilisant ces paramètres, a mis plus de 12 jours à être appris avec une carte NVidia GeForce GTX 1080.

4.1 Données

Les données utilisées sont celles mises à disposition pour l'évaluation WMT 16 (Bojar *et al.*, 2016) pour Europarl et News Commentary. Les données « IT » viennent de la fusion des corpus Ubuntu, KDE4 et KDE-doc disponibles sur le site de OPUS [1](Tiedemann, 2012). De même, les corpus JRC et BOOKS sont également issus de cette même source. Le corpus WIT^3 (Cettolo *et al.*, 2012) est une collection de transcriptions de vidéos de TED avec leurs traductions. Ces données sont régulièrement utilisées pour les évaluations IWSLT d'où sont issues nos données de test (IWSLT2014). L'ensemble des corpus utilisé est détaillé dans le tableau 1.

4.2 Résultats

Le tableau 2 présente l'ensemble des résultats obtenus lors de nos expériences. Nous proposons donc d'étudier le comportement des modèles réentrainés avec 5 cycles d'apprentissage supplémentaires en faisant varier les quantités de données de réentrainement. Ainsi, nous avons découpé le corpus WIT^3 selon cinq tailles différentes : 500, 1 000, 5 000, 10 000 lignes et enfin le corpus entier. Ceci afin de mesurer l'impact de ces quantités de données sur les résultats.

1. http://opus.lingfil.uu.se

Modèles	*en-fr*	*en-de*	temps d'apprentissage
générique	33,82	20,07	12 jours
spécialisé (WIT^3-0,5K)	33,26	19,00	5 min
spécialisé (WIT^3-1K)	34,87	20,30	8 min
spécialisé (WIT^3-5K)	36,95	21,51	15 min
spécialisé (WIT^3-10K)	37,42	22,38	17 min
spécialisé (WIT^3-entier)	**39,53**	**25,53**	168 min (≃2h45)
générique+WIT^3-entier	35,60	22,68	12 jours

TABLE 2 – Scores de BLEU pour la spécialisation avec différentes tailles du corpus spécifique *WIT*^{3}. A titre indicatif, les temps d'apprentissage et de spécialisation sont indiqués dans la dernière colonne.

Nous observons ici deux systèmes de base, le premier est un système (noté *générique*) appris sur les données que nous avons classées comme génériques, puis un second système appris avec ces mêmes données génériques que nous avons concaténés avec les données spécialisées (noté *générique+WIT*^{3}). Le second obtient de meilleurs résultats sur notre corpus de test IWSLT2014 aussi bien en français qu'en allemand. A partir du premier modèle, nous proposons donc d'appliquer l'approche de spécialisation décrite dans la partie 3 en utilisant les corpus de tailles différentes.

Ainsi, la spécialisation avec 500 lignes n'apporte rien, au contraire, les scores de BLEU dans les deux langues décroient : -0,6 et -1 point respectivement pour l'anglais vers le français et l'allemand. On observe une première progression des scores lorsqu'on spécialise le modèle avec 1000 lignes. Dans la tâche de traduction vers le français, le modèle spécialisé améliore de 1 point de BLEU. Dans la tâche vers l'allemand, la progression est plus timide avec +0,3 point de BLEU. Les améliorations sont plus marquées lorsqu'on spécialise avec 5000 lignes pour les deux tâches. Ainsi, les résultats passent de 33,82 à 36,87 pour le français et de 20,07 à 21,51 pour l'allemand. La progression est toujours présente avec une quantité de corpus supplémentaire (10 000 lignes) les scores sont encore améliorés de 0,5 point pour le français et de 0,8 point pour l'allemand. Enfin, les meilleurs scores sont obtenus avec l'utilisation du corpus en entier, où le score du modèle générique français est amélioré de presque 6 points BLEU et l'allemand de 5,5 points BLEU.

Lorsque nous comparons ces résultats de la spécialisation avec les résultats obtenus avec le modèle incluant l'ensemble des données de WIT^3, nous observons que ce dernier est dépassé en termes de score dès lors que nous ajoutons 5 000 lignes, soit moins de 3 pour cent du corpus initial pour les deux paires de langue.

4.3 Discussion

Source :	So your 20s are the time to educate yourself about your body ...
Reference :	Donc la vingtaine est le moment de s' informer sur son corps ...
générique	Donc , vos années 20 sont l' occasion de vous éduquer sur votre corps ...
spécialisé (WIT^3-0.5K)	Donc , vos 20 s sont le moment d' apprendre à votre corps ...
spécialisé (WIT^3-1K)	Donc vos 20èmes sont le moment d' être éduqués sur votre corps ...
spécialisé (WIT^3-5K)	Donc , vos 20 ans sont le moment d' apprendre votre corps ...
spécialisé (WIT^3-10K)	Donc , vos 20 ans sont le moment d' éduquer votre corps ...
spécialisé (WIT^3-entier)	Donc vos 20 ans est le moment d' éduquer votre corps ...
générique+WIT^3-entier	Donc vos 20s sont le moment de vous éduquer sur votre corps ...

TABLE 3 – Exemples de traductions correspondant à différents points. Le rouge, bleu et vert sont des traductions notées respectivement *mauvaise*, *acceptable* et *bonne* .

Malgré les performances obtenues avec cette approche, il est nécessaire d'avoir une quantité minimale de corpus. Ainsi, l'idée d'utiliser quelques phrases ou un paragraphe afin de spécialiser le modèle ne semble pas suffisant. Au contraire, utiliser une si petite quantité de données semble perturber le

modèle au point de donner de moins bons résultats que si l'on ne fait rien (utilisation du modèle générique). Reste cependant l'idée de réutiliser des données du corpus d'apprentissage en plus des données spécifiques, mais nous n'avons pas eu le temps de faire ce type d'expériences.

Le tableau 3 présente un exemple tiré du corpus de test. Nous avons pris un exemple dans les sorties qui illustre bien la progression de la spécialisation ou l'adaptation des modèles à la tâche de traduction. Ainsi, nous pouvons voir les améliorations dues aux différentes quantités de données qui ont été ajoutées. Même si les sorties ne correspondent pas exactement à la référence, nous considérons la traduction comme bonne selon des critères syntaxiques et sémantiques.

De plus, un point particulier à mettre en avant est le gain de temps que permet cette approche. Outre le fait des performances, cette approche permet de s'affranchir de l'obligation de réentrainer depuis le début un modèle de traduction adapté à une tâche ou un domaine particulier. Lorsqu'on se met dans un cadre particulier de la traduction assistée par ordinateur, ce cadre nécessite une adaptation à un domaine particulier en un temps relativement court. De fait, cette approche autorise une mise-à-jour journalière du modèle de traduction et répond à cette contrainte. En effet, il faut moins de vingt minutes (pour traiter 5 000 et 10 000 lignes, voir le tableau 2) afin d'adapter et de spécialiser un modèle générique et obtenir des performances intéressantes. Le gain de temps et d'utilisabilité du modèle ne sont que plus intéressants.

Enfin, nous remarquons que cette approche pourrait s'apparenter à une méthode d'optimisation du modèle, telle que présentée par Och (2003) pour les modèles de traduction à base de paires de segments. Cette phase de spécialisation pourrait, à l'avenir, se focaliser sur une métrique particulière telle que le BLEU, le TER ou encore le METEOR.

5 Conclusion

Dans cet article nous avons proposé une courte étude sur une approche d'adaptation au domaine des modèles de traduction neuronaux : la spécialisation. Cette approche d'adaptation propose des améliorations significatives en termes de scores BLEU qui sont perceptibles dans les sorties. Ces gains permettent de s'affranchir d'un entrainement long et couteux (plusieurs jours, voire plusieurs semaines contre moins de vingt minutes) afin d'obtenir des performances significatives.

Dans le cadre d'un processus de traduction assistée par ordinateur, ce processus de spécialisation pour les modèles de traduction neuronaux semble particulièrement pertinent, permettant une rapide mise-à-jour des modèles. L'une des perspectives de ces permiers travaux est d'étudier et de mesurer, dans un cadre *in vivo*, l'impact de cette approche et vérifier quels seraient les gains obtenus du point de vue des traducteurs. Toujours dans ce cadre, nous souhaitons étudier l'impact d'une adaptation incrémentale sur les performances de traduction.

Enfin cette approche de spécialisation ouvre l'étude à d'autres approches classiques pour l'adaptation au domaine telles que la pondération de corpus ou de phrase lors de l'apprentissage (Shah *et al.*, 2010, 2011). Nous pourrons également explorer l'utilisation de marqueurs de domaine tels qu'ils ont été introduit par Sennrich *et al.* (2016a); Zhang *et al.* (2016), ce qui permettrait d'améliorer la représentativité des données dans le cadre de ce type d'approche. Enfin, cette phase de spécialisation pourrait se focaliser sur une métrique automatique afin d'optimiser le modèle selon cette même métrique.

Références

AXELROD A., HE X. & GAO J. (2011). Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, p. 355–362, Edinburgh, Scotland, UK. : Association for Computational Linguistics.

BAHDANAU D., CHO K. & BENGIO Y. (2014). Neural machine translation by jointly learning to align and translate. *CoRR*, **abs/1409.0473**. Demoed at NIPS 2014 : http ://lisa.iro.umontreal.ca/mt-demo/.

BLAIN F., SCHWENK H. & SENELLART J. (2012). Incremental adaptation using translation information and post-editing analysis. In *IWSLT*, p. 229–236.

BOJAR O., CHATTERJEE R., FEDERMANN C., GRAHAM Y., HADDOW B., HUCK M., JIMENO YEPES A., KOEHN P., LOGACHEVA V., MONZ C., NEGRI M., NEVEOL A., NEVES M., POPEL M., POST M., RUBINO R., SCARTON C., SPECIA L., TURCHI M., VERSPOOR K. & ZAMPIERI M. (2016). Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation*, p. 131–198, Berlin, Germany.

BOJAR O., CHATTERJEE R., FEDERMANN C., HADDOW B., HUCK M., HOKAMP C., KOEHN P., LOGACHEVA V., MONZ C., NEGRI M., POST M., SCARTON C., SPECIA L. & TURCHI M. (2015). Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, p. 1–46, Lisbon, Portugal.

CETTOLO M., BERTOLDI N., FEDERICO M., SCHWENK H., BARRAULT L. & SERVAN C. (2014). Translation project adaptation for mt-enhanced computer assisted translation. *Machine Translation*, **28**(2), 127–150.

CETTOLO M., GIRARDI C. & FEDERICO M. (2012). Wit^3 : Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, p. 261–268, Trento, Italy.

CETTOLO M., NIEHUES J., STUKER S., BENTIVOGLI L., CATTONI R. & FEDERICO M. (2015). The iwslt 2015 evaluation campaign. In *the proceedings of the International Workshop on Spoken Language Translation (IWSLT)*.

CETTOLO M., NIEHUES J., STUKER S., BENTIVOGLI L., CATTONI R. & FEDERICO M. (2016). The iwslt 2016 evaluation campaign. In *the proceedings of the International Workshop on Spoken Language Translation (IWSLT)*.

DOULAMIS A. D., DOULAMIS N. D. & KOLLIAS S. D. (2000). On-line retrainable neural networks : improving the performance of neural networks in image analysis problems. *IEEE Transactions on Neural Networks*, **11**(1), 137–155.

GAO J. & ZHANG M. (2002). Improving language model size reduction using better pruning criteria. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, p. 176–182, Stroudsburg, PA, USA : Association for Computational Linguistics.

GASCÓ G., ROCHA M.-A., SANCHIS-TRILLES G., ANDRÉS-FERRER J. & CASACUBERTA F. (2012). Does more data always yield better translations ? In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, p. 152–161.

KALCHBRENNER N. & BLUNSOM P. (2013). Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*.

KLEIN G., KIM Y., DENG Y., SENELLART J. & RUSH A. (2017). OpenNMT : Open-Source Toolkit for Neural Machine Translation. *ArXiv e-prints*. 1701.02810.

LAMBERT P., SCHWENK H., SERVAN C. & ABDUL-RAUF S. (2011). Investigations on translation model adaptation using monolingual data. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, p. 284–293, Edinburgh, Scotland : Association for Computational Linguistics.

LÜ Y., HUANG J. & LIU Q. (2007). Improving statistical machine translation performance by training data selection and optimization. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, p. 343–350.

LUONG T. & MANNING C. D. (2015). Stanford neural machine translation systems for spoken language domains. In *IWSLT2015*, Da Nang, Vietnam.

LUONG T., PHAM H. & MANNING C. D. (2015). Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, p. 1412–1421, Lisbon, Portugal.

MOORE R. & LEWIS W. (2010). Intelligent Selection of Language Model Training Data. In *ACL (Short Papers)*, p. 220–224, Uppsala, Sweden.

OCH F. J. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, p. 160–167.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). Bleu : a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, p. 311–318, Philadelphia, Pennsylvania, USA.

SENNRICH R., HADDOW B. & BIRCH A. (2016a). Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics.*

SENNRICH R., HADDOW B. & BIRCH A. (2016b). Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 86–96, Berlin, Germany : Association for Computational Linguistics.

SHAH K., BARRAULT L. & SCHWENK H. (2010). Translation model adaptation by resampling. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, WMT '10, p. 392–399.

SHAH K., BARRAULT L. & SCHWENK H. (2011). Parametric weighting of parallel data for statistical machine translation. In *The 5th International Joint Conference on Natural Language Processing*, Chiang Mai (Thialand).

SUTSKEVER I., VINYALS O. & LE Q. V. (2014). Sequence to sequence learning with neural networks. In *NIPS*.

TIEDEMANN J. (2012). Parallel data, tools and interfaces in opus. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012)*.

WUEBKER J., NEY H., MARTÍNEZ-VILLARONGA A., GIMÉNEZ A., JUAN A., SERVAN C., DYMETMAN M. & MIRKIN S. (2014). Comparison of data selection techniques for the translation of video lectures. In *AMTA*.

ZHANG J., LI L., WAY A. & LIU Q. (2016). Topic-informed neural machine translation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics : Technical Papers*, p. 1807–1817, Osaka, Japan.

Détection de concepts et granularité de l'annotation

Pierre Zweigenbaum[1] Thomas Lavergne[2]
(1) LIMSI, CNRS, Université Paris-Saclay, 91405 Orsay, France
(2) LIMSI, CNRS, Univ. Paris-Sud, Université Paris-Saclay, 91405 Orsay, France
pz@limsi.fr, lavergne@limsi.fr

RÉSUMÉ

Nous nous intéressons ici à une tâche de détection de concepts dans des textes sans exigence particulière de passage par une phase de détection d'entités avec leurs frontières. Il s'agit donc d'une tâche de catégorisation de textes multiétiquette, avec des jeux de données annotés au niveau des textes entiers. Nous faisons l'hypothèse qu'une annotation à un niveau de granularité plus fin, typiquement au niveau de l'énoncé, devrait améliorer la performance d'un détecteur automatique entraîné sur ces données. Nous examinons cette hypothèse dans le cas de textes courts particuliers : des certificats de décès où l'on cherche à reconnaître des diagnostics, avec des jeux de données initialement annotés au niveau du certificat entier. Nous constatons qu'une annotation au niveau de la « ligne » améliore effectivement les résultats, mais aussi que le simple fait d'appliquer au niveau de la ligne un classifieur entraîné au niveau du texte est déjà une source d'amélioration.

ABSTRACT

Concept detection and annotation granularity

We address here the detection of concepts in texts with no requirement for first spotting entities and their boundaries. This is thus a multilabel text categorization task, with a dataset with text-level annotations. We hypothesize that a finer-grained annotation, typically at the sentence level, should improve the performance of an automatic concept detector trained on these data. We examine this hypothesis in the case of specific short texts : death certificates in which we aim to detect diagnoses, with datasets initially annotated at the level of full certificates. We observe that a "line"-level annotation does improve concept detection ; but also that simply applying at the sentence level a classifier trained at the text level is already a source of improvement.

MOTS-CLÉS : Extraction d'information ; catégorisation de textes ; détection de concepts ; granularité de l'annotation ; diagnostics.

KEYWORDS: Information extraction; text categorization; concept detection; annotation granularity.

1 Introduction

La détection de concepts dans des textes consiste à y relever la présence de concepts d'un référentiel termino-ontologique. Citons par exemple *Absence acquise de poumon* (Z90.2 dans la CIM-10, Classification internationale des maladies, version 10, de l'OMS (OMS, 1993)) ou encore *Pneumonies, sans précision* (J18.9 dans la CIM-10). Ces concepts fournissent une représentation des informations utiles dans divers systèmes d'information (Zweigenbaum, 2017). Par exemple, dans le domaine médical, les « descripteurs » du thésaurus MeSH sont employés pour l'indexation et la recherche de

notices bibliographies d'articles scientifiques dans la base MEDLINE, et les classes de la Classification internationale des maladies sont employées pour caractériser les séjours hospitaliers à des fins médico-économiques ou pour compiler à l'échelle nationale et internationale les statistiques de mortalité à partir des textes des certificats de décès.

La détection de concepts est généralement abordée comme une étape suivant la détection d'entités nommées (Nouvel *et al.*, 2015). Il s'agit alors d'une tâche de normalisation d'entités, aussi appelée plus récemment liage d'entités (Zheng *et al.*, 2015). Cette tâche demande de déterminer, pour chaque expression décrivant une entité, déjà définie par ses frontières et généralement un type, le concept le plus approprié parmi ceux d'un référentiel donné (typiquement une terminologie, ontologie ou base de données). Les corpus annotés indiquent donc pour chaque mention d'une entité le concept attendu pour celle-ci.

Nous nous plaçons dans un contexte différent où la détection de concepts est définie comme une tâche de catégorisation de textes. Étant donné un texte, sans indication préalable d'entités et de leurs frontières, il s'agit de déterminer les concepts mentionnés dans ce texte. Les corpus annotés fournissent alors simplement la liste des concepts associés à chaque texte, sans indication de localisation. D'un côté, supprimer l'évaluation des frontières simplifie la tâche. D'un autre côté, l'absence d'indication sur la position des concepts dans les annotations fournies rend la tâche plus difficile : on ne sait pas a priori quels mots sont révélateurs des concepts à trouver, ce qui donne moins d'indices pour l'entraînement. De plus, par rapport aux tâches courantes de catégorisation de textes, qui considèrent quelques classes (Sebastiani, 2002), la détection de concepts s'attaque souvent à des milliers, voire des dizaines de milliers de classes (Tsatsaronis *et al.*, 2015) : de l'ordre de 13 000 dans la CIM-10 et de 27 000 dans MeSH. Parmi les types de méthodes employées pour aborder cette tâche, les deux plus fréquentes sont celles fondées sur des dictionnaires ou terminologies existants (Aronson & Lang, 2010), et celles fonctionnant par apprentissage supervisé (Lin & Wilbur, 2007).

Nous nous intéressons dans cet article à la granularité des annotations que l'on peut employer pour détecter des concepts dans cette situation. Par rapport à un corpus annoté au niveau des textes, un corpus annoté à un niveau plus fin, par exemple des énoncés, aide-t-il à mieux détecter les concepts de chaque texte ? Nous partons d'un corpus avec deux niveaux d'annotation et comparons les performances obtenues dessus en détection de concepts pour tester les deux hypothèses suivantes :

— Lorsque l'on utilise des méthodes par apprentissage supervisé, leur entraînement sur une annotation au niveau de granularité plus fin devrait améliorer la détection de concepts par rapport à leur entraînement sur une annotation au niveau des textes.
— Les méthodes à base de dictionnaire ne devraient pas être affectées par ces différences d'annotation.

2 Données et méthodes

2.1 Cas d'usage et données : codage CIM-10 dans des certificats de décès

Nous utilisons dans ces expériences les données d'entraînement de la tâche CLEF eHealth 2017 (`https://sites.google.com/site/clefehealth2017/task-1`), qui fournissent des annotations aux deux niveaux recherchés. Ces données concernent des certificats de décès français, textes courts auxquels sont associés des concepts désignant les maladies ou événements ayant mené au décès (tableau 1). Ces concepts sont choisis parmi les quelque 13 000 classes de la Classification internatio-

Grain	Segment annoté	Concepts
ligne	pneumopathie nosocomiale	J18.9
ligne	insuffisance respiratoire chronique	J96.1
ligne	insuffisance cardiaque chronique, pneumonectomisé pour cancer pulmonaire, lymphome non hodgkinien en abstention thérapeutique	C85.9, I50.9, C34.9, Z90.2
certif.	pneumopathie nosocomiale XYZT insuffisance respiratoire chronique XYZT insuffisance cardiaque chronique, pneumonectomisé pour cancer pulmonaire, lymphome non hodgkinien en abstention thérapeutique	J18.9, J96.1, I50.9, Z90.2, C34.9, C85.9

TABLE 1 – Exemple de certificat (*certif.*) et de ses trois énoncés (« lignes »)

nale des maladies, version 10 (codes CIM-10, (OMS, 1993)). Un certificat se compose de un à cinq énoncés, un par ligne dans le formulaire officiel, chaque énoncé rapportant un ou plusieurs maladies ou autres événements. On constate que ces énoncés sont essentiellement des listes de maladies, qui n'emploient naturellement pas nécessairement les mêmes termes que les libellés des classes de la CIM-10, comme le montre l'exemple de J18.9 et Z90.2, présents dans les première et troisième ligne du tableau 1, dont les libellés sont donnés dans l'introduction.

	Total		Entraînement		Test	
	certificat	ligne	certificat	ligne	certificat	ligne
Certificats	64 980	64 980	61 521	61 521	3 459	3 459
Segments annotés	64 980	192 641	61 521	182 643	3 459	9 998
Concepts	262 809	257 772	249 249	244 474	13 560	13 298

TABLE 2 – Jeux de données comparables avec des annotations au niveau du certificat ou de la ligne

La version d'origine du corpus contient des annotations au niveau des certificats entiers (ligne *certif.* du tableau 1 et colonnes *certificat* du tableau 2), que les organisateurs (Lavergne *et al.*, 2016) ont reportées au niveau des lignes (lignes ou colonnes *ligne* des tableaux) [1]. Le tableau 2 donne les statistiques des corpus résultants, contenant près de 65 000 certificats composés de près de 193 000 lignes, que nous avons divisés en deux parties pour nos expériences : entraînement ($\sim 95\ \%$) et test ($\sim 5\ \%$). Chaque partie est donc annotée au niveau du certificat ou de la ligne (rangée *Segments annotés*). On constate que le nombre de concepts (rangée *Concepts*) est légèrement inférieur ($\sim -2\ \%$) dans la version *ligne* : certains concepts se sont perdus lors du report (semi-automatique) des certificats aux lignes, ce qui pourra légèrement défavoriser les traitements effectués à partir de ces lignes. Le corpus total contient 3 290 concepts différents en version *certificat* et 3 220 en version *ligne*, ce qui va dans le même sens.

Les organisateurs fournissent aussi des dictionnaires employés par le CépiDc [2] dans le processus réel de codage des certificats de décès. Nous utilisons ici celui de 2011 (les résultats sont similaires avec celui de 2012, également fourni ; les certificats traités vont de 2006 à 2012). Il contient 156 937 entrées pour 6 093 concepts différents. Ce dictionnaire contient tous les concepts mentionnés dans le corpus sauf 249 dans la version *certificat* et 245 dans la version *ligne* (7,6 % des concepts différents).

En préalable à tout traitement, nous supprimons les mots outils et les ponctuations, racinisons, passons

1. Constatant que certains certificats s'étaient perdus dans la version annotée au niveau des certificats, nous les avons aussi supprimés de la version *ligne*.

2. Centre d'épidémiologie des causes médicales de décès, Unité Inserm US10, `http://www.cepidc.inserm.fr/`.

en minuscules et supprimons les diacritiques des mots restants dans chaque ligne. Dans les données par certificat, toutes les lignes d'un même certificat sont d'abord regroupées en un texte et séparées par un symbole arbitraire (séquence *XYZT* absente du corpus) servant de séparateur d'énoncé.

Méthode à base de dictionnaire La projection de dictionnaire consiste à rechercher dans les textes des occurrences de termes connus associés à des concepts. Étant donné une liste de concepts avec pour chacun un ou plusieurs termes, nous en effectuons une détection efficace (gauche-droite, plus longue séquence) en encodant le dictionnaire dans un transducteur à états finis. Le même traitement initial qu'aux textes est appliqué aux entrées de dictionnaire. De ce fait, la recherche exacte d'une séquence de mots ainsi normalisés fournit une recherche souple de termes, robuste à des variations de casse, de désinence, de diacritiques et de mots outils. Le dictionnaire peut proposer plusieurs concepts pour la même expression, qui est alors ambiguë. Dans ce cas nous testons deux méthodes : *tous* renvoie tous les concepts associés, *nonamb* ne renvoie pas de concept si l'expression est ambiguë.

Une méthode hybride : calibrage du dictionnaire Certains des concepts détectés par le dictionnaire peuvent être erronés. Pour réduire ce type d'erreurs, nous entraînons un classifieur binaire (OUI/NON) (un SVM) qui décide si un concept est pertinent ou pas. L'entraînement est effectué sur notre jeu d'entraînement, en prenant comme attributs le concept proposé par le dictionnaire et le sac de mots de la ligne ou du texte d'entrée. Lors de l'application au jeu de test, le classifieur entraîné est appliqué à chaque concept proposé par le dictionnaire. S'il le classe comme un NON, nous supprimons ce concept de la sortie produite.

Détection de concepts par apprentissage supervisé Nous entraînons un classifieur multiclasse pour décider des concepts associés à un texte ou une ligne d'entrée. Après des expériences initiales sur notre jeu d'entraînement, nous avons opté pour les attributs suivants : (i) le sac de mots de l'entrée, normalisé comme expliqué plus haut ; (ii) le sac de trigrammes de caractères de l'entrée, également calculé après normalisation des mots ; cela fournit de la robustesse aux fautes d'orthographe et autres variantes terminologiques ; (iii) l'année de codage du certificat : cela permet de prendre en compte des variations de règles de codage au fil des années. Après divers essais réalisés en validation croisée sur le jeu d'entraînement, nous avons constaté que ces attributs fournissaient les meilleurs résultats, et que la plupart du temps un SVM linéaire était meilleur que plusieurs autres classifieurs (régression logistique, forêt aléatoire, k plus proches voisins, etc.).

Nos expériences initiales se sont réalisées sur des classifieurs monoétiquette, qui ne produisent qu'un concept pour chaque entrée. Cependant, comme dans la situation réelle chaque entrée peut mener à plusieurs concepts, cette solution n'est pas optimale. Nous avons donc également entraîné un classifieur multiétiquette, beaucoup plus gourmand computationnellement, car il entraîne un classifieur par concept. Par exemple, l'implémentation que nous avons utilisée (scikit-learn, LinearSVM, métaclassifieur OneVsRestClassifier) a pris trois heures pour l'entraînement (182 643 exemples, de l'ordre de 3 000 classes) et 25 minutes pour le test dans la version *ligne* du corpus (9 998 exemples).

Les méthodes présentées peuvent sembler simples ; en pratique, l'union du dictionnaire calibré et du SVM mono-étiquette détient les meilleurs résultats publiés (Zweigenbaum & Lavergne, 2016) hors soumissions officielles sur le corpus de test de CLEF eHealth 2016 (Névéol *et al.*, 2016), et le SVM multi-étiquette seul ou combiné au dictionnaire a obtenu les meilleurs résultats hors soumissions officielles sur les corpus de test français (ligne et certificat) de CLEF eHealth 2017 (Zweigenbaum & Lavergne, 2017).

Granularité de l'entraînement, de l'application et de l'évaluation Pour l'entraînement d'un classifieur ou le calibrage du dictionnaire, nous disposons d'annotations au niveau des certificats

entiers ou de leurs lignes (voir les tableaux 1 et 2). C'est le premier paramètre (*entraînement*) que nous faisons varier dans nos expériences. Nous pouvons ensuite appliquer la détection de concepts au niveau d'un certificat entier ou de chaque ligne d'un certificat. C'est notre deuxième paramètre (*test*). Enfin, dans le cas où l'application se fait sur chaque ligne, l'évaluation peut se faire par comparaison à la référence de chaque ligne, ou en regroupant d'abord les concepts proposés pour toutes les lignes d'un certificat et en les comparant à la référence du certificat. C'est notre troisième paramètre (*éval*).

3 Résultats

La simple projection de dictionnaire (tableau 3, rangées avec – dans la colonne *Cal*) ne change pas avec le jeu d'entraînement. En revanche, l'évaluation au niveau de chaque ligne de certificat est plus exigeante que celle effectuée au niveau des certificats entiers : un concept peut être incorrect pour une ligne mais s'avérer correct pour le certificat du fait qu'il apparaît dans une autre ligne du même certificat. On le constate pour le dictionnaire brut, mais aussi pour ses variantes avec calibrage (rangées avec *c-t* dans la colonne *Cal*). Le calibrage sur les certificats a des résultats similaires à celui sur les lignes. Il apporte dans les deux cas une grande amélioration de la précision et de la F-mesure. Sa combinaison à la méthode *nonamb* est encore plus précise, mais un bien meilleur rappel est obtenu par la méthode *tous*, dont le calibrage apporte la meilleure F-mesure.

			test=certificat					test=ligne				
Entr.	Méth.	Cal.	Sys.	TP	P	R	F1	Sys.	TP	P	R	F1
certif	nonamb	–	10 812	8 863	82,0	65,4	72,7	10 866	8 677	79,9	65,3	71,8
certif	nonamb	c-t	8 927	8 465	**94,8**	62,4	75,3	8 983	8 347	**92,9**	62,8	74,9
certif	tous	–	18 466	10 690	57,9	**78,8**	66,8	18 591	10 450	56,2	**78,6**	65,5
certif	tous	c-t	10 711	9 787	91,4	72,2	**80,7**	10 774	9 640	89,5	72,5	**80,1**
ligne	nonamb	–	10 812	8 863	82,0	65,4	72,7	10 867	8 677	79,9	65,3	71,8
ligne	nonamb	c-t	8 368	7 999	**95,6**	59,0	73,0	8 843	8 293	**93,8**	62,4	74,9
ligne	tous	–	18 466	10 690	57,9	**78,8**	66,8	18 590	10 450	56,2	**78,6**	65,5
ligne	tous	c-t	9 949	9 196	92,4	67,8	78,2	10 548	9 575	90,8	72,0	**80,3**

TABLE 3 – Calibrage du dictionnaire sur des certificats (13 560 concepts) et des lignes (13 298 concepts). Le gras indique les meilleurs résultats pour la précision (P), le rappel (R) et la F-mesure (F1) (en pourcentages). Entr. = entraînement. Méth. = méthode du dictionnaire (voir le texte). Cal. = Type de calibrage : – = aucun, c-t = concept+mots. Sys. = nb de concepts prédits. TP = vrais positifs.

Les expériences réalisées en classification supervisée confirment que si l'objectif est de détecter les concepts de chaque énoncé-ligne (colonnes *éval=ligne* en partie droite du tableau 4) plutôt qu'au niveau du texte-certificat entier, il vaut mieux disposer d'une annotation au niveau de chaque énoncé. Ce point est très clair en classification monoétiquette (colonne éval = ligne, +18pt de F-mesure de 59 à 77 %) ; dans notre jeu de données, la classification multiétiquette est nettement plus robuste et ne perd que 2 points de F-mesure (de 87 à 85 %) du fait d'une diminution de précision (-5 pt de 88 à 83 %) à rappel relativement constant.

Si l'objectif est de détecter globalement les concepts présents dans le texte (colonnes éval = certificat du tableau 4), ces expériences montrent qu'il reste intéressant de passer par l'énoncé de deux façons :

— Un effort d'annotation des énoncés est le plus payant. En classification monoétiquette, entraîner un classifieur sur les énoncés ainsi annotés apporte un très large gain par rapport à un

		test=certificat éval=certificat			test=ligne éval=certificat ←			éval=ligne		
Entr.	Méth.	P	R	F1	P	R	F1 ←	P	R	F1
certif	mono	**85,5**	21,8	34,7	70,7	**51,6**	**59,6** ←	68,8	51,7	59,0
ligne	mono	80,7	20,6	32,8	**91,4**	67,2	**77,5** ←	89,9	**67,6**	77,1
certif	multi	85,4	79,9	82,5	**85,7**	**86,7**	**86,2** ←	83,4	86,5	84,9
ligne	multi	85,4	61,3	71,4	**90,3**	**86,3**	**88,2** ←	88,0	86,2	87,1

TABLE 4 – Classification supervisée (SVM linéaire : LinearSVC de scikit learn) sur les certificats et sur les lignes. Le gras note les meilleurs résultats par rangée en précision (P), rappel (R) et F-mesure (F1) (en pourcentages).

entraînement et test sur les textes (rangée ligne / mono, colonne test=ligne, éval=certificat : +42 points de F-mesure de 35 à 77 %).
De même, en classification multiétiquette, l'entraînement sur des lignes annotées a des résultats supérieurs à l'entraînement et test sur les textes (+5 points de F-mesure de 83 à 88 %).

— Même sans disposer d'annotations au niveau des énoncés, entraîner un classifieur monoétiquette sur les textes annotés puis l'appliquer individuellement sur les énoncés augmente mécaniquement le nombre de concepts proposés et ainsi le rappel, au prix d'une baisse de précision (rangée certif / mono, colonne test=ligne, éval=certificat : +30pt de rappel de 22 à 52 %, +25 points de F-mesure de 35 à 60 %). Le rappel est en effet limité par la classification monoétiquette, qui ne propose qu'un concept par unité textuelle.
Bien que n'ayant pas cette limitation en rappel, le classifieur multiétiquette bénéficie lui aussi d'une application sur chaque énoncé plutôt que sur un texte entier : il augmente un peu sa précision et fortement son rappel (+7 pt de rappel de 80 à 87 %, +3 pt de F-mesure de 83 à 86 %). Nous supposons que c'est lié à deux phénomènes. D'une part, la représentation choisie étant en sac de mots, l'application à des énoncés individuels réduit le risque de confusion dû à des mots non connectés dans le texte, ce qui devrait augmenter la précision et peut-être le rappel. Par exemple, dans un certificat qui contient sur trois lignes différentes les expressions *insuffisance respiratoire terminale* (J96.9), *insuffisance cardiaque* (I50.9), et *insuffisance rénale chronique sévère* (N18.9), l'application à chaque ligne du classifieur trouve chacun des trois concepts attendus alors que son application globale au certificat entier ne propose pas le troisième, et propose à la place du premier le concept J96.1 (*insuffisance respiratoire chronique*). D'autre part, le classifieur multiétiquette a appris sur des textes entiers à déterminer le nombre de concepts à proposer, et est ainsi biaisé vers un nombre de concepts plus grand que pour un énoncé individuel, ce qui est susceptible d'augmenter le rappel au détriment éventuel de la précision.

Nous insistons sur une condition qui nous semble importante dans la situation étudiée : chaque concept peut être détecté localement, c'est-à-dire à l'intérieur d'un énoncé, comme c'est le cas des entités nommées. En pratique, cette condition n'est pas entièrement satisfaite ici, car le choix d'un concept pour certains diagnostics peut dépendre du contexte. Ainsi, dans un certificat contenant le terme ambigu *choc hémorragique* ainsi que *polytraumatisme*, le classifieur entraîné et appliqué sur les certificats propose bien le concept T79.4 (*choc traumatique*, choc en présence de traumatisme) alors qu'appliqué sur les lignes il propose R57.1 (*choc hypovolémique*), qui est incorrect ici.

Cette étude possède diverses limitations. Ces expériences ont été réalisées sur un seul jeu de données. Leur généralisation à d'autres données reste à tester. Pour des questions de temps, nous n'avons pas

appliqué de validation croisée pour construire les tableaux 3 et 4 : les différences observées sont susceptibles de varier avec des découpages différents en entraînement et test. Notons cependant qu'un certain nombre des différences constatées dans le tableau 4 sont larges.

Les jeux de données employés correspondent aux corpus d'entraînement de CLEF eHealth 2016 et 2017, mais pas à un corpus de test existant. Les résultats présentés ne sont de ce fait pas directement comparables à ceux des participants de la tâche de 2016 ou 2017. On peut noter cependant qu'ils sont du même ordre de grandeur que le résultat du meilleur participant sur le corpus de test de 2016 (F1=84,8 % (Van Mulligen *et al.*, 2016)), qui était évalué au niveau ligne, ou que nos résultats sur ce même corpus de test pour l'union du dictionnaire calibré et du classifieur mono-étiquette (F1=85.9 % (Zweigenbaum & Lavergne, 2016)). Ils sont aussi du même ordre que nos résultats avec le classifieur multi-étiquette sur le corpus de test français de CLEF eHealth 2017 (ligne : F1=86.5 %, certificat : F1=81.7 %, (Zweigenbaum & Lavergne, 2017)), qui surpassent de 5 à 6 points ceux du meilleur participant.

4 Conclusion

Dans le cas des certificats de décès, nous confirmons l'intuition selon laquelle la détection supervisée de concepts au niveau des énoncés (lignes) fonctionne mieux qu'au niveau du texte entier (certificat). Ce constat était bien sûr attendu lorsque l'on dispose d'énoncés annotés ; nous avons observé qu'il s'applique aussi lorsque l'on dispose uniquement de textes annotés, un classifieur entraîné sur les textes entiers donnant en moyenne de meilleurs résultats si on l'applique séparément à chaque énoncé que si on l'applique globalement au texte entier. Nous avons vu également qu'un classifieur multiétiquette était plus robuste mais cependant sensible aussi à cette différence de granularité.

Ces observations ouvrent plusieurs perspectives. Tout d'abord, elles sont directement applicables à des données de nature similaire mais provenant d'une autre source et écrites dans une autre langue : les certificats de décès américains fournis lors de la campagne CLEF eHealth 2017. Contrairement aux certificats français, ces données ne sont annotées qu'au niveau de chaque certificat. Nous avons réalisé des expériences préliminaires qui confirment l'intérêt d'entraîner un classifieur sur ces textes puis de l'appliquer au niveau des lignes individuelles plutôt que des textes entiers. De plus, ces observations incitent à mettre en place des méthodes pour transférer automatiquement au niveau des énoncés les annotations disponibles au niveau des textes. Là aussi, nos premiers essais par annotation automatique d'une partie des énoncés puis ré-entraînement sur le corpus ainsi étendu sont encourageants.

Remerciements

Ce travail a bénéficié d'un soutien de l'action Horizon 2020 Marie Sklodowska-Curie Innovative Training Networks — European Joint doctorate (ITN-EJD) de l'Union européenne, projet No :676207 (MiRoR).

Références

ARONSON A. R. & LANG F.-M. (2010). An overview of MetaMap : historical perspective and

recent advances. *J Am Med Inform Assoc*, **17**(3), 229–36.

LAVERGNE T., NÉVÉOL A., ROBERT A., GROUIN C., REY G. & ZWEIGENBAUM P. (2016). A dataset for ICD-10 coding of death certificates : Creation and usage. In *Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM2016)*, p. 60–69, Osaka, Japan : The COLING 2016 Organizing Committee.

LIN J. & WILBUR W. J. (2007). PubMed related articles : a probabilistic topic-based model for content similarity. *BMC Bioinformatics*, **8**, 423.

NOUVEL D., ERHMANN M. & ROSSET S. (2015). *Les entités nommées pour le traitement automatique des langues*. Iste Editions.

NÉVÉOL A., COHEN K. B., GROUIN C., HAMON T., LAVERGNE T., KELLY L., GOEURIOT L., REY G., ROBERT A., TANNIER X. & ZWEIGENBAUM P. (2016). Clinical information extraction at the CLEF eHealth evaluation lab 2016. In *CLEF eHealth Evaluation Lab*.

OMS (1993). *Classification statistique internationale des maladies et des problèmes de santé connexes — Dixième révision*. Organisation mondiale de la Santé, Genève.

SEBASTIANI F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, **34**(1), 1–47.

TSATSARONIS G., BALIKAS G., MALAKASIOTIS P., PARTALAS I., ZSCHUNKE M., ALVERS M. R., WEISSENBORN D., KRITHARA A., PETRIDIS S., POLYCHRONOPOULOS D., ALMIRANTIS Y., PAVLOPOULOS J., BASKIOTIS N., GALLINARI P., ARTIÈRES T., NGONGA A., HEINO N., ÉRIC GAUSSIER, BARRIO-ALVERS L., SCHROEDER M., ANDROUTSOPOULOS I. & PALIOURAS G. (2015). An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, **16**, 138 :1–138 :28.

VAN MULLIGEN E., AFZAL Z., AKHONDI S. A., VO D. & KORS J. A. (2016). Erasmus MC at CLEF eHealth 2016 : Concept recognition and coding in French texts. In *CLEF 2016 Online Working Notes* : CEUR-WS.

ZHENG J. G., HOWSMON D., ZHANG B., HAHN J., MCGUINNESS D., HENDLER J. & JI H. (2015). Entity linking for biomedical literature. *BMC Medical Informatics and Decision Making*, **15**(1), S4.

ZWEIGENBAUM P. (2017). Le traitement des langues naturelles dans le contexte de la eSanté. In P. DEGOULET, M. FIESCHI & J. MÉNARD, Eds., *e-santé en perspective*, volume 20 of *Informatique et santé*. Paris : Lavoisier.

ZWEIGENBAUM P. & LAVERGNE T. (2016). Hybrid methods for ICD-10 coding of death certificates. In *Seventh International Workshop on Health Text Mining and Information Analysis*, p. 96–105, Austin, Texas, USA : EMNLP 2016.

ZWEIGENBAUM P. & LAVERGNE T. (2017). Multiple methods for multi-class, multi-label ICD-10 coding of multi-granularity, multilingual death certificates. In *CLEF 2017 Online Working Notes* : CEUR-WS. Sous presse.

Tri Automatique de la Littérature pour les Revues Systématiques

Christopher Norman,[1, 2] Mariska Leeflang,[2] Pierre Zweigenbaum,[1] Aurélie Névéol[1]
(1) LIMSI, CNRS, Université Paris Saclay, 91405 Orsay, France
(2) Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands
prenom.nom@limsi.fr, m.m.leeflang@amc.uva.nl

RÉSUMÉ

Les revues systématiques de la littérature dans le domaine biomédical reposent essentiellement sur le travail bibliographique manuel d'experts. Nous évaluons les performances de la classification supervisée pour la découverte automatique d'articles à l'aide de plusieurs définitions des critères d'inclusion. Nous appliquons un modèle de regression logistique sur deux corpus issus de revues systématiques conduites dans le domaine du traitement automatique de la langue et de l'efficacité des médicaments. La classification offre une aire sous la courbe moyenne (AUC) de 0.769 si le classifieur est contruit à partir des jugements experts portés sur les titres et résumés des articles, et de 0.835 si on utilise les jugements portés sur le texte intégral. Ces résultats indiquent l'importance des jugements portés dès le début du processus de sélection pour développer un classifieur efficace pour accélérer l'élaboration des revues systématiques à l'aide d'un algorithme de classification standard.

ABSTRACT

Automatically Ranking the Literature in Support of Systematic Reviews.

Current approaches to document discovery for systematic reviews in biomedicine rely on exhaustive manual screening. We evaluate the performance of classifier based article discovery using different definitions of inclusion criteria. We test a logistic regressor on two datasets created from existing systematic reviews on clinical NLP and drug efficacy, using different criteria to generate positive and negative examples. The classification and ranking achieves an average AUC of 0.769 when relying on gold standard decisions based on title and abstracts of articles, and an AUC of 0.835 when relying on decisions based on full text. Results suggest that inclusion based on title and abstract generalizes to inclusion based on full text, so that references excluded in earlier stages are important for classification, and that common-off-the-shelves algorithms can partially automate the process.

MOTS-CLÉS : Recherche d'Information, Classification Supervisée, Revues Systématiques.

KEYWORDS: Information Retrieval, Supervised Classification, Systematic Reviews.

1 Revues systématiques de la littérature

1.1 Processus de sélection des articles analysés

Les revues systématiques de la littérature ont pour objectif de présenter une synthèse complète et objective de l'ensemble des informations publiées sur une question donnée. Elément essentiel de la médecine factuelle, ces études constituent le niveau de preuve scientifique le plus élevé. Elles contribuent également à l'information du grand public, à l'élaboration de politiques de santé publique.

Cependant, la réalisation d'une revue systématique repose sur des experts et s'avère extrèmement couteuse en temps. Le nombre d'articles à consulter pour effectuer une revue **complète** de la littérature peut se chiffrer à plusieurs milliers, alors que seules quelques dizaines de références passeront le filtre de la méta-analyse finale. Ce processus de sélection, essentiellement manuel, demande plusieurs mois de travail à un collège d'experts.

Le processus de sélection se déroule en plusieurs étapes, comme illustré en figure 1. Un large ensemble d'articles candidats à l'analyse est défini, typiquement grâce à une requête booléenne établie par les experts et soumise à des moteurs recherche spécialisés. Une première sélection est effectuée dans cet ensemble en consultant le titre et le résumé des articles retournés par la requête. Cette étape permet d'éliminer les articles qui ne portent pas sur la question étudiée et les articles qui ne sont pas conformes aux critères d'inclusion de l'étude. Nous désignons les articles éliminés par ce premier filtre comme la catégorie "**N**". Une deuxième sélection est ensuite effectuée sur consultation du texte intégral des articles afin de confirmer que l'ensemble des critères d'inclusion sont bien réunis. Nous désignons les articles éliminés par ce deuxième filtre comme la catégorie "**M**", et les articles finalement retenus comme la catégorie "**Y**".

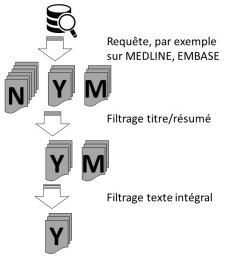


FIGURE 1 – Processus de sélection des articles pour une revue systématique.

Ces deux étapes sont nécessaires, car si un grand nombre d'articles peut être rejetté suite à l'examen du titre et du résumé, la décision finale d'inclusion des articles dans l'analyse ne peut être prise que sur la base du texte intégral car le niveau de détail requis ne se trouve pas dans les résumés. Les décisions d'inclusion réputées difficiles sont prises lors de la deuxième étape, et les articles des catégories Y et M sont listés dans la revue finale, alors que les articles de la catégorie N sont simplement écartés.

Bien que la méthode experte de sélection des articles à inclure dans une revue systématique repose sur plusieurs étapes, les travaux cherchant à automatiser ce processus ont abordé le problème comme une seule étape, celle de la sélection finale des articles à retenir. En effet, le but final de la sélection de documents dans l'élaboration d'une revue systématique est d'identifier les documents satisfaisant les critères d'inclusion. Lors de la construction de jeux de données il n'est donc pas toujours jugé pertinent de distinguer les documents appartenant aux catégories M et N. Nous faisons l'hypothèse que les documents de la catégorie M peuvent être utiles pour entrainer un classifieur et nous étudions différentes configurations des jeux d'entrainement afin définir la configuration la plus efficace.

Dans cet article, nous étudions l'intérêt de distinguer les deux étapes du processus de sélection pour automatiser le tri d'article à inclure dans les revues systématiques. Notre contribution est double : tout d'abord, nous présentons des expériences permettant de choisir une définition adaptée du problème de classification des articles considérés pour inclusion dans les revues systématiques et ensuite, nous nous appuyons sur un corpus existant dans le domaine de l'éfficacité des médicaments et confirmons

les résultats obtenus sur un nouveau corpus dans le domaine orthogonal du TAL.

1.2 Sélection automatique d'articles pour les revues systématiques

De nombreux travaux ont abordé la question de l'automatisation du processus d'élaboration des revues systématiques, avec un succès variable comme l'indique une revue approfondie de la recherche récente sur ce sujet dans le domaine biomédical (O'Mara-Eves *et al.*, 2015). D'autres domaines ont proposé des méthodes pour améliorer la recherche exhaustive de documents, par exemple pour les brevets (Stein *et al.*, 2012) ou les procédures civiles en droit américain (Grossman & Cormack, 2011). Le problème est abordé comme une tâche de classification ou de tri ordonné d'articles. Les méthodes de classification supervisées nécessitent la disponibilité de données d'entrainement, ce qui n'est pas toujours le cas pour des revues dans un nouveau domaine. Cependant, pour des revues de mise à jour ou des revues abordant une nouvelle question dans un domaine déjà exploré, le résultat des revues précédentes peut être exploité.

Ainsi, des classifieurs Bayésiens (Matwin *et al.*, 2010; Matwin & Sazonova, 2012), des SVMs (Support Vector Machines) et des arbres de décision (Bekhuis & Demner-Fushman, 2010) ont permis d'obtenir de bonnes performances. Parmi les méthodes utilisées on dénombre également : des perceptrons (Cohen *et al.*, 2006), kNN (García Adeva *et al.*, 2014), des relations sémantiques (Fiszman *et al.*, 2010), des ontologies (Sun *et al.*, 2012), des modèles linéaires (Shekelle *et al.*, 2012), des classifieurs ensemblistes (Shekelle *et al.*, 2012), l'indexation aléatoire (Jonnalagadda & Petitti, 2014) ou les forêts d'arbres décisionnels (Khabsa *et al.*, 2016). Cependant, peu de ces méthodes ont été évaluées sur des jeux de données communs hormis le corpus Cohen. Nous décrivons ce jeu de données ci-dessous et l'utilisons pour permettre une comparaison directe de nos résultats avec l'état de l'art.

2 Régression logistique pour le tri automatique de documents

2.1 Corpus de Travail

Chaque corpus comporte des références d'articles sous forme de numéro identifiant associé à des métadonnées ainsi qu'à une classification selon les catégories N, M et Y. La distribution des articles issus de chaque étape de sélection est présentée dans le Tableau 1.

Le **corpus du** *Yearbook* est issu de la revue de la littérature en Traitement Automatique de la Langue Clinique effectuée pour le *Yearbook of Medical Informatics* (Névéol & Zweigenbaum, 2016). Ce corpus illustre le cas d'une revue systématique récurrente, proposant des mises à jour successives de l'état de l'art. Pour chaque nouvelle version de la revue, les données issues des revues précédentes peuvent être exploitées pour entraîner un classifieur *intra-topic*.

Le **corpus Cohen**, disponible librement [1], est issu d'un travail précurseur sur la classification automatique d'articles pour les revues systématiques exploitant les données de 15 revues sur l'efficacité des médicaments (Cohen *et al.*, 2006). Ce corpus illustre le cas d'une série de revues sur le même thème général, mais déclinant plusieurs questions. Pour chaque nouvelle question, les données issues des revues précédentes peuvent être exploitées pour entraîner un classifieur *inter-topic*.

1. `http://skynet.ohsu.edu/~cohenaa/systematic-drug-class-review-data.html`

Topic	Y	M	N	Topic	Y	M	N
Yearbook							
ClinicalNLP (2017)	11	70	244				
ClinicalNLP (2016)	23	60	267				
Cohen							
CalciumChannelBlockers	100	180	938	ProtonPumpInhibitors	51	187	1095
ACEInhibitors	41	142	2361	Triptans	24	194	453
BetaBlockers	42	260	1770	NSAIDS	41	47	305
Opiods	15	33	1867	ADHD	20	64	767
OralHypoglycemics	136	3	364	AtypicalAntipsychotics	146	218	756
Statins	85	88	3292	UrinaryIncontinence	40	38	249
SkeletalMuscleRelaxants	9	25	1609	Estrogens	80	0	288
Antihistamines	16	76	218				

TABLE 1 – Distribution des catégories d'article dans les corpus de travail.

2.2 Méthode de classification

Nous utilisons l'implémentation de regression logistique de sklearn(Pedregosa *et al.*, 2011) entrainé à l'aide d'un gradient stochastique conjugué. Nous avons choisi cette méthode pour sa capacité à traiter rapidement des représentations de données creuses et à effectuer des mises à jour partielles du modèle. Par ailleurs, cette méthode produit un score de confiance associé aux prédictions, ce qui est essentiel pour notre tâche de tri de documents.

Les traits du classifieurs reposent sur les métadonnées extraites de la base MEDLINE pour chacun des articles du corpus : Nous créons des traits avec les sacs-de-n-grammes ($n \leq 3$) issus des titres des résumés, les scores tf-idf ou binaire associés aux mots, les formes racinisées ou originale des mots. Nous considérons séparéments les n-grammes issus des sections "background", "method", "results", et "conclusion" des résumés. Nous considérons également les mots-clés attribués aux articles par les auteurs, les noms des journaux et les types de publication. Pour le corpus Cohen nous extrayons les termes d'indexation MeSH. Pour le corpus Yearbook, ces termes ne sont pas disponibles lors de la préparation de la revue systématique.

Nous abordons deux types de classification : **la classification *intra-topic*** pour laquelle des données d'entrainement et les données de test appartiennent au même thème. Dans le cas du corpus Yearbook, nous entrainons le classifieur sur les données 2016 et nous testons sur les données 2017. Dans le cas du corpus Cohen, nous séparons aléatoirement les données en deux ensembles. Cela présete l'inconvénient de produire des corpus d'entrainement de petite taille, en particulier pour certains thèmes. Afin de pallier cet inconvénénient, nous abordons également **la classification *inter-topic*** pour laquelle des données d'entrainement et les données de test appartiennent à des thèmes différents, mais proches. C'est le cas pour les différents thèmes du corpus Cohen. Dans ce cas, le classifieur est entrainé sur 14 thèmes et testé sur le 15ème thème.

Nous suivons la méthodologie expérimentale utilisée dans la littérature (Cohen *et al.*, 2006; Khabsa *et al.*, 2016). Ainsi, pour la validation croisée inter-topic nous répétons le processus de classification 10 fois. Pour la validation croisée intra-topic nous répétons le processus de classification 5 fois, sur 2 plis. Nous rapportons les résultats moyens (+/- SD) pour chaque expérience en termes d'aire moyenne sous la courbe (AUC) et de score WSS@95 ("work saved over sampling at 95% recall"). Le WSS@95

représente la quantité de travail économisée par un expert qui parcourerait la liste des articles triés par le classifieur jusqu'au seuil de 95% de rappel pour les articles à inclure dans la revue (catégorie Y).

Nous utilisons les paramètres par défaut, sauf dans les cas indiqués ci-dessous. Nous augmentons le poids des exemples positifs en leur affectant un coefficient de 80 afin d'amplifier le coût des erreurs de classification pour les articles pertinents. Nous utilisons la valeur $\alpha = 10^{-4}$ pour la regularisation sur le corpus Cohen, et $\alpha = 0.05$ pour le corpus du Yearbook. Ces valeurs ont été sélectionnées sur la base du résultat d'expériences préliminaires sur le thème (CalciumChannelBlockers) pour le corpus Cohen, et sur la première itération pour le corpus du Yearbook 2016.

3 Résultats

La table 2 présente une comparaison de nos résultats avec l'état de l'art afin de valider la pertinence de notre méthode. Notre classifieur offre des performances supérieures à l'état de l'art pour la mesure WSS@95, et légèrement en dessous pour l'AUC. Notre implémentation semble notament faible lorsque peu d'articles de la catégorie M sont présents (OralHypoGlycemics, Estrogens). On constate cependant qu'aucune méthode ne semble s'imposer sur l'ensemble des thèmes.

	Intertopic			Intratopic			
Thème \ Mesure	WSS@95	AUC		WSS@95		AUC	
	NC	NC	(Cohen)	NC	(Khabsa)	NC	(Khabsa)
CalciumChannelBlockers	.13	**.76**	.71	**.40**	.29 (RF)	.83	**.87** (SVM)
ACEInhibitors	.57	**.82**	.81	**.63**	.52 (CNB)	.92	**.95** (RF)
BetaBlockers	.40	**.84**	.80	**.51**	.37 (CNB)	.86	**.89** (RF)
Opiods	.30	**.89**	.86	**.59**	.55 (CNB)	.91	**.91** (RF)
OralHypoglycemics	.07	**.66**	.57	**.11**	.08 (CNB)	.57	**.78** (SVM)
Statins	.27	**.83**	.77	**.44**	.40 (RF)	.87	**.92** (RF)
SkeletalMuscleRelaxants	.24	.83	**.84**	**.43**	.37 (RF)	.74	**.79** (RF)
Antihistamines	.07	**.65**	.62	**.15**	.15 (CNB)	.65	**.72** (SVM)
ProtonPumpInhibitors	.38	**.82**	.79	**.31**	.29 (RF)	.83	**.88** (RF)
Triptans	.46	.82	**.82**	.30	**.31** (RF)	.79	**.91** (SVM)
NSAIDS	.671	**.91**	.90	**.54**	.53 (CNB)	.86	**.95** (SVM)
ADHD	.13	**.59**	.47	.62	**.67** (VP)	.91	**.95** (RF)
AtypicalAntipsychotics	.16	**.76**	.65	**.21**	.21 (CNB)	.78	**.84** (RF)
UrinaryIncontinence	.37	**.89**	.85	**.42**	.41 (RF)	.78	**.89** (SVM)
Estrogens	.18	**.69**	.59	.29	**.38** (CNB)	.69	**.89** (SVM)

TABLE 2 – Comparaison de notre classifieur (NC) sur le corpus Cohen avec (Cohen, 2008) et (Khabsa *et al.*, 2016) pour la classification (Y||MN).

Les Tables 3 et 4 présentent les resultats de nos expériences utilisant différentes combinaisons des classes Y, M et N. Nous indiquons entre parenthèses les classes considérées : par exemple, (Y||MN) dénote l'utilisation de la catégorie Y comme classe positive, les catégories M et N étant considérées comme classe négative. Dans l'expérience (Y|M|N) les catégories Y et M sont utilisées comme classe positive à l'entrainement mais seule la catégorie Y est considérés comme classe positive lors du test.

	(Y||MN)				(YM||N)				(Y|M|N)			
	WSS@95		AUC		WSS@95		AUC		WSS@95		AUC	
Topic	avg	std	avg	std	avg	std	avg	std	avg	std	avg	std
ClinicalNLP	.00	.00	.63	.01	.23	.01	.80	.00	.19	.01	.81	.00
CChannelBlockers	.40	.10	.83	.02	.22	.06	.76	.03	.34	.07	.79	.01
ACEInhibitors	.63	.16	.92	.02	.28	.05	.80	.02	.60	.13	.88	.03
BetaBlockers	.51	.16	.86	.03	.19	.05	.73	.03	.48	.21	.83	.02
Opiods	.59	.19	.91	.05	.37	.10	.82	.03	.71	.06	.88	.04
OHypoglycemics	.11	.05	.57	.03	.14	.07	.58	.04	.09	.02	.58	.03
Statins	.44	.18	.87	.02	.25	.09	.78	.03	.42	.10	.86	.02
SkMuscleRelaxants	.43	.22	.74	.11	.26	.18	.83	.06	.45	.12	.75	.06
Antihistamines	.15	.09	.65	.09	.13	.04	.57	.03	.24	.09	.60	.01
PPumpInhibitors	.31	.19	.83	.04	.17	.04	.73	.02	.38	.06	.77	.04
Triptans	.30	.24	.79	.08	.30	.04	.75	.03	.41	.07	.69	.03
NSAIDS	.54	.18	.86	.02	.40	.07	.76	.04	.46	.06	.73	.02
ADHD	.62	.15	.91	.03	.70	.10	.91	.02	.83	.06	.91	.01
AAntipsychotics	.21	.04	.78	.01	.12	.02	.71	.03	.28	.06	.80	.02
UIncontinence	.42	.14	.78	.03	.21	.09	.66	.04	.48	.07	.75	.04
Estrogens	.29	.09	.69	.03	.27	.09	.72	.04	.32	.06	.69	.03

TABLE 3 – Classification Intra-topic à trois classes. Les résultats présentent la moyenne (avg) et l'éccart type (std) obtenus sur 10 itérations (5×2 validation croisée) pour différentes compositions des jeux d'entrainement.

	(Y||M)				(Y||N)				(M||N)			
	WSS@95		AUC		WSS@95		AUC		WSS@95		AUC	
Topic	avg	std	avg	std	avg	std	avg	std	avg	std	avg	std
ClinicalNLP	.01	.00	.48	.01	.02	.00	.74	.00	.26	.00	.79	.00
CChannelBlockers	.14	.04	.59	.03	.42	.11	.85	.02	.21	.07	.74	.03
ACEInhibitors	.17	.08	.63	.06	.41	.37	.92	.03	.26	.06	.77	.02
BetaBlockers	.38	.10	.74	.02	.52	.14	.87	.03	.19	.03	.71	.02
Opiods	.13	.10	.53	.01	.59	.21	.91	.06	.25	.18	.76	.05
OHypoglycemics	.06	.00	.39	.17	.11	.04	.58	.03	.75	.19	.83	.11
Statins	.13	.05	.56	.04	.44	.18	.88	.05	.24	.09	.71	.03
SkMuscleRelaxants	.24	.14	.55	.02	.30	.15	.67	.08	.23	.16	.80	.07
Antihistamines	.20	.17	.55	.06	.16	.09	.70	.03	.13	.07	.58	.04
PPumpInhibitors	.16	.05	.58	.02	.42	.17	.85	.03	.12	.05	.69	.03
Triptans	.20	.13	.70	.07	.44	.24	.88	.04	.27	.06	.75	.03
NSAIDS	.13	.05	.58	.06	.48	.19	.85	.02	.32	.09	.72	.03
ADHD	.19	.14	.59	.09	.71	.17	.94	.02	.64	.17	.92	.01
AAntipsychotics	.11	.02	.55	.02	.26	.11	.79	.03	.11	03	.63	.03
UIncontinence	.09	.04	.55	.02	.43	.16	.79	.03	.12	.10	.59	.05
Estrogens	-	-	-	-	.23	.03	.69	.04	-	-	-	-

TABLE 4 – Classification Intratopic à deux classes. Les résultats présentent la moyenne (avg) et l'éccart type (std) obtenus sur 10 itérations (5×2 validation croisée) pour différentes compositions des jeux d'entrainement.

Globalement, nos résultats montrent l'intérêt de disposer des documents de la catégorie N pour entraîner nos classifieurs. En effet, les performances les plus médiocres sont obtenues avec la catégorisation (Y||M), qui correspond pourtant à la deuxième étape de filtrage, considérée plus importante et mieux documentée.

Dans la Table 3, on observe des résultats similaires pour les classifications (Y|M|N) et (YM||N), ce qui indique que les exemples de la catégorie M peuvent être utilement utilisés comme substituts aux exemples de la catégorie Y lorsque ceux-ci sont rares. Par ailleurs, les résultats de la classification (Y||MN) sont largement inférieurs, ce qui confirme notre hypothèse et montre l'importance de bien distinguer la catégorie M de la catégorie N.

4 Conclusion et perspectives

En accord avec la littérature, nos résultats indiquent qu'un classifieur entrainé sur le titre et le résumé d'articles offre de bonnes performances pour le tri d'articles sur des critères d'inclusion majoritairement présent dans le texte intégral. Cependant, pour atteindre ce résultat, il est important de disposer d'exemples d'articles exclus sur la base du résumé seul.

Ainsi, pour construire un corpus de référence qui permette d'évaluer ou d'entrainer des outils automatiques de tri de la littérature pour les revues systématiques, il est nécessaire de disposer d'une catégorisation des articles qui reflète les deux étapes du processus manuel de filtrage.

Dans la suite de ce travail, nous envisageons plusieurs directions : tout d'abord, nous souhaitons évaluer la généralisabilité de la classification aux revues systématiques des études sur l'évaluation des tests diagnostiques, réputées fastidieuses de par la difficulté de construire des requêtes permettant de réduire l'ensemble des documents à considérer. Ensuite, nous prévoyons d'étudier l'apport du retour de pertinence (*relevance feedback*) en particulier pour les thèmes pour lesquels il n'est pas possible de construire d'emblée un classifieur intra-topic. L'intégration du retour de pertinence correspondrait en outre à un cas pratique d'utilisation de méthodes de classification automatique pour la sélection d'articles en vue d'une revue systématique où les experts valideraient la classification des articles dans l'ordre proposé par l'outil - produisant ainsi des jugements intégrables dynamiquement dans le classifieur.

Remerciements

Ce travail a bénéficié d'un financement du programme européen Horizon 2020 Marie Sklodowska Curie Innovative Training Networks—European Joint doctorate (ITN-EJD) No :676207, Methods in Research on Research (MiRoR).

Références

BEKHUIS T. & DEMNER-FUSHMAN D. (2010). Towards automating the initial screening phase of a systematic review. *Studies in Health Technology and Informatics*, **160**(PART 1), 146–150.

COHEN A. M. (2008). Optimizing feature representation for automated systematic review work prioritization. *AMIA Annual Symposium proceedings*, p. 121–5.

COHEN A. M., HERSH W. R., PETERSON K. & YEN P. (2006). Reducing Workload in Systematic Review Preparation Using Automated Citation Classification. p. 206–219.

FISZMAN M., BRAY B., SHIN D., KILICOGLU H., BENNETT G., BODENREIDER O. & RINDFLESCH T. (2010). Combining Relevance Assignment with Quality of the Evidence to Support Guideline Development. *Stud Health Technol Inform*, **160**(1), 709—-713.

GARCÍA ADEVA J. J., PIKATZA ATXA J. M., UBEDA CARRILLO M. & ANSUATEGI ZENGOTITABENGOA E. (2014). Automatic text classification to support systematic reviews in medicine. *Expert Systems with Applications*, **41**(4 PART 1), 1498–1508.

GROSSMAN M. & CORMACK G. (2011). Technology-assisted review in e-discovery can be more effective and more efficient than exhaustive manual review. *Richmond Journal of Law and Technology*, **17**(3).

JONNALAGADDA S. R. & PETITTI D. (2014). A new iterative method to reduce workload in the systematic review process. *Int J Comput Biol Drug Des*, **6**(0), 5–17.

KHABSA M., ELMAGARMID A., ILYAS I., HAMMADY H. & OUZZANI M. (2016). Learning to identify relevant studies for systematic reviews using random forest and external information. *Machine Learning*, **102**(3), 465–482.

MATWIN S., KOUZNETSOV A., INKPEN D., FRUNZA O. & O'BLENIS P. (2010). A new algorithm for reducing the workload of experts in performing systematic reviews. *Journal of the American Medical Informatics Association : JAMIA*, **17**(4), 446–53.

MATWIN S. & SAZONOVA V. (2012). Direct comparison between support vector machine and multinomial naive Bayes algorithms for medical abstract classification. *Journal of the American Medical Informatics Association*, **19**(5), 917–917.

NÉVÉOL A. & ZWEIGENBAUM P. (2016). Clinical natural language processing in 2015 : Leveraging the variety of texts of clinical interest. *IMIA Yearbook*, p. 234–239.

O'MARA-EVES A., THOMAS J., MCNAUGHT J., MIWA M. & ANANIADOU S. (2015). Using text mining for study identification in systematic reviews : a systematic review of current approaches. *Systematic reviews*, **4**(1), 5.

PEDREGOSA F., VAROQUAUX G., GRAMFORT A., MICHEL V., THIRION B., GRISEL O. *et al.* (2011). Scikit-learn : Machine learning in python. *Journal of Machine Learning Research*, **12**(Oct), 2825–2830.

SHEKELLE P. G., DALAL S. R. & SHETTY K. D. (2012). A Pilot Study Using Machine Learning and Domain Knowledge To Facilitate Comparative Effectiveness Review Updating. *AHRQ*.

STEIN B., HOPPE D. & GOLLUB T. (2012). The impact of spelling errors on patent search. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, p. 570–579 : Association for Computational Linguistics.

SUN Y. B., YANG Y., ZHANG H., ZHANG W. & WANG Q. (2012). Towards evidence-based ontology for supporting systematic literature review. *Evaluation and Assessment in Software Engineering*, **2012**(1), 171–175.

Une approche linguistique pour la détection des dialectes arabes

Houda Saâdane[1] Damien Nouvel[2] Hosni Seffih[1, 3] Christian Fluhr[1]
(1) GEOLSemantics, 12 Avenue Raspail, 94250 Gentilly, France
(2) ERTIM, INALCO, 2 rue de Lille, 75007 Paris, France
(3) LIASD (EA4383), Université Paris8, 2 rue de la Liberté, 93526 Saint-Denis cedex, France
houda.saadane,hosni.seffih,christian.fluhr@geolsemantics.com,
damien.nouvel@inalco.fr

RÉSUMÉ

Dans cet article, nous présentons un processus d'identification automatique de l'origine dialectale pour la langue arabe de textes écrits en caractères arabes ou en écriture latine (arabizi). Nous décrivons le processus d'annotation des ressources construites et du système de translittération adopté. Deux approches d'identification de la langue sont comparées : la première est linguistique et exploite des dictionnaires, la seconde est statistique et repose sur des méthodes traditionnelles d'apprentissage automatique (n-grammes). L'évaluation de ces approches montre que la méthode linguistique donne des résultats satisfaisants, sans être dépendante des corpus d'apprentissage.

ABSTRACT

A linguistic approach for the detection of Arabic dialects.

In this work, we present the automatic identification process of the dialectal origin of the Arabic language in text written in Arabic characters and in Latin characters (Arabizi). We describe the annotation process of the constructed resources and the transliteration system. We compare two systems : the linguistic one is based on dictionary lookup, the statistical one is based on machine learning (ngrams). the evaluation of those systems shows that the linguistic approach obtains good results, without relying on a training corpus.

MOTS-CLÉS : dialectes arabes, arabizi, alternance codique, translittération, identification des dialectes, analyse morphologique.

KEYWORDS: Arabic Dialects, Arabizi, Code switching, transliteration, Identification of Dialects, Morphological analyzes.

1 Introduction

Le traitement du texte informel est devenu un domaine d'investigation extrêmement populaire (Yang & Eisenstein, 2013). Pour l'arabe, l'identification du dialecte est une tâche devenue centrale pour la plupart des applications de traitement de l'arabe, comme par exemple la traduction automatique ou l'analyse des médias sociaux. Selon (Zaidan & Callison-Burch, 2011), l'identification des dialectes peut être vue comme une identification des langues appliquée à un groupe de langues étroitement apparentées. Pour l'arabe, cette identification est rendue complexe par l'absence de conventions orthographiques et par la translittération en écriture latine «***arabizi***». Des travaux récents ont proposé des approches statistiques pour l'identification de langue. Cependant, les méthodes actuelles reposent

sur l'hypothèse que des ressources existent (corpus et dictionnaires), ce qui n'est pas toujours le cas, notamment pour l'arabe maghrébin. L'objectif de notre travail est de construire des ressources adéquates pour l'algérien, le tunisien et le marocain, ainsi que pour le dialecte égyptien. Nous présentons une méthode linguistique pour l'identification de l'origine dialectale et discutons les résultats, en les comparant à ceux obtenus par une méthode statistique.

2 État de l'art

La constitution de ressources et de méthodes pour traiter les dialectes de l'arabe est devenue un enjeu de taille, et nécessite des ressources spécialisées. Le travail de (Zaidan & Callison-Burch, 2014) s'inscrit dans cette optique : les auteurs utilisent les sites de trois journaux arabes des dialectes levantin, golfe et égyptien. L'annotation, réalisée avec Amazon Mechanical Turk, permet une approche de modélisation linguistique et est exploitée afin de prédire le dialecte pour une phrase donnée. D'autres initiatives ont visé à créer des données dialectales afin de remédier au manque de ressources dédiées, comme par exemple l'extraction de commentaires, de journaux en ligne et de Twitter pour les dialectes égyptien, golfe, levantin, algérien et irakien (Cotterell & Callison-Burch, 2014).

Le travail de (Elfardy & Diab, 2012) propose des directives pour la formation de larges corpus de ressources arabes mixtes à code alterné pour le dialecte égyptien et a abouti au système *AIDA "Automatic Identification and glossing of Dialectal Arabic"* (Elfardy *et al.*, 2014), pour l'identification, la classification et l'interprétation des dialectes. Les auteurs de (Elfardy & Diab, 2013) ont ensuite présenté une approche supervisée pour l'identification des phrases en dialecte égyptien et ont étudié les effets des techniques de prétraitement sur la performance de classifieurs.

L'absence de convention orthographique standardisée est une difficulté, objet de travaux proposant une convention de transcription nommée *CODA "Conventional Orthography for Dialectal Arabic"*. Elle a été proposée dans un premier temps pour le dialecte égyptien (Habash *et al.*, 2012), puis étendue à d'autres dialectes comme le tunisien (Zribi *et al.*, 2014), l'algérien (Saâdane & Habash, 2015) et le palestinien (Jarrar *et al.*, 2014).

Ces approches traitent des dialectes écrits en caractères arabes, or ces dernières années ont vu l'apparition de l'arabizi : l'écriture en langue arabe avec des caractères latins. L'arabizi est souvent utilisé dans des contextes informels comme les réseaux sociaux, et alterné avec d'autres langues étrangères, comme l'anglais ou le français. Les quelques outils TAL pour l'arabizi (Saâdane *et al.*, 2013; Darwish, 2013; Eskander *et al.*, 2014) visent au préalable à le détecter dans les textes, à translittérer les textes vers l'écriture arabe, ce qui permet de les traiter avec des systèmes TAL dédiés à l'ASM [1].

L'ensemble des travaux référencés ci-dessus sont focalisés sur l'égyptien et le levantin (jordanien, libanais, palestinien, syrien) et beaucoup moins de travaux ont été réalisés pour les langues maghrébines. Notons également que les ressources présentées ci-dessus ne sont généralement pas disponibles, ce qui rend difficile la reproduction des expériences et la comparaison des résultats. Dans notre travail, nous nous intéressons prioritairement à la constitution de ressources pour les dialectes maghrébins et plus spécifiquement à l'identification automatique du dialecte écrit en caractères arabes et latins : ce sont des problèmes bien réels qui appellent à des solutions concrètes.

1. ASM : Arabe standard moderne

3 Constitution des corpus

Nos ressources linguistiques ont été obtenues à partir d'une collecte de données faite sur les commentaires des lecteurs extraits des journaux arabes ainsi que ceux du réseau social Facebook. Nous avons choisi d'utiliser des sites et journaux spécifiques pour chaque dialecte étudié, comme exposé dans le tableau 2. Cette collecte nous a permis de constituer un corpus segmenté selon le dialecte et le mode d'écriture (arabe ou arabizi). Nous signalons que les commentaires rédigés en français ou en anglais ne contenant pas d'arabizi ont été exclus de nos corpus. Le processus suivi pour la construction de nos ressources est décrit dans (Saâdane *et al.*, 2013).

	ALGERIEN		**MAROCAIN**		**TUNISIEN**		**EGYPTIEN**	
Source	**Echorouk**	**Dzfoot**	**Yabiladi**	**Kifache**	**Aljarida**	**F.B**	**Misr5**	**F.B**
#messages	10.5K	67.5K	15.6K	103.1K	5.5K	13.6K	30.9K	4.3K
#phrases	20.3K	182.3K	100.7K	131.1K	13.7K	20.2K	36.2K	8.1K
#mots	400.9K	1.5M	2.4M	2.5M	70K	263K	31.8K	348K
#msg arabe	5.2K	60K	8.6K	3.3K	4K	5.1K	29.1K	2.1K
#msg latin	5.3K	7.4K	7.0K	99.8K	1.5K	8.4K	1.8K	2.2K

TABLE 1 – Volumes des corpus constitués

L'extraction de ces données, tirant parti de la localisation dialectale des sources, permet de constituer un corpus volumineux et relativement fiable de textes, pour lesquels le dialecte des messages est connu *a priori*. Ils seront ainsi exploités à la fois pour constituer un corpus de test pour les deux méthodes présentées ci-après, et un corpus d'apprentissage pour la méthode statistique.

4 Système de détection des dialectes à base de dictionnaires

4.1 Format d'annotation dialectale des données

Une annotation automatique détermine, pour chaque mot du texte, les étiquettes suivantes :

- **lang1 :** mot en ASM dans un texte arabe, par ex. *AlĀn* [2] «maintenant». Elle annote également dans les textes arabizi les mots étrangers qui conservent leurs formes orthographique comme les mots «belle» du français ou «good» de l'anglais.
- **lang2 :** mot dialectal <AD> dans un texte arabe ou arabizi et information pour un texte arabe : <DZ> (*algérien*), <TN> (*tunisien*), <MA> (*marocain*) ou <EG> (*égyptien*). Cette information est rajoutée dans les textes arabizi après leur translittération.
- **entité Nommée :** une entité nommée, comme *AljazaAŷir* «l'Algérie».
- **autres :** ponctuations, chiffres, sons et émoticônes, URL, etc.
- **ambiguïté :** mot où la classe (lang1 ou lang2) ne peut être déterminée étant donné le contexte courant (par ex. *Tyb* peut être utilisé en ASM «bon» et en dialecte égyptien «ok»), observée aussi dans l'annotation de mots communs à plusieurs dialectes (comme la particule *wqtAš* «quand» qui est utilisée dans les dialectes maghrébins).

2. Translittération arabe présentée dans scheme Habash-Soudi-Buckwalter (HSB) (Habash *et al.*, 2007)

4.2 Translittération de l'arabizi

Après avoir annoté les textes dialectaux écrits en latin, les messages en arabizi sont convertis en écriture arabe, en suivant la convention CODA. Cette translittération arabizi-arabe se focalise d'abord sur les entités nommées en utilisant un système (Saâdane *et al.*, 2012) qui convertit l'arabizi en texte arabe, et inversement, par utilisation de transducteurs à états finis. Les translittérations possibles sont filtrées à l'aide d'un analyseur morphologique de l'arabe. Pour les mots qui ne sont pas des entités nommées, un translittérateur basé sur une approche à base de règles et de dictionnaires spécifiques pour chaque dialecte est utilisé. Les dictionnaires regroupent 24442 paires de correspondances (mots, mais aussi bigrames ou trigrammes) arabizi-arabe répartis entre les dialectes étudiés. Ces correspondances ont été établies par des natifs arabes et qui ont translitteré manuellement les motifs ayant le plus grand nombre d'occurrences dans des corpus. Les règles proposées permettent de définir les frontières des lettres dans un mot arabe et d'appliquer des règles d'exception pour chaque mot afin de supprimer les variantes orthographiques qui ne sont pas attestées dans l'usage. La liste générée est ensuite filtrée à l'aide de l'analyseur morphologique pour prédire l'appartenance d'un mot à un des dialectes étudiés.

Si la conversion de l'arabizi en écriture arabe est une étape importante du processus, cet article se focalise sur l'identification des dialectes et nous ne pouvons, faute de place, détailler plus avant cette partie du traitement. Notons cependant que ce processus conserve une information importante pour identifier le dialecte : la présence de voyelles en arabizi, permet d'indiquer les diacritiques en arabes (ils sont généralement omis), ce qui apportera un indice décisif pour déterminer le dialecte d'un texte.

4.3 Système ambigu d'annotation des mots à base de dictionnaire

L'originalité de notre approche d'annotation réside dans la production d'analyses morphologiques diacritiques. Les dialectes maghrébins produisent effectivement de nombreux morphèmes et éléments lexicaux qui sont tout à fait semblables, et ne diffèrent souvent que par les voyelles courtes. Si cette information n'est pas écrite en ASM, elle est en revanche présente par translittération de l'aribizi, ce qui va aider à la reconnaissance du dialecte.

Notre approche d'annotation est résumée comme suit :

- **Prétraitements :** séparation de la ponctuation et des nombres, normalisation des effets d'allongement de lettres, détection des URLs et des nombres, étiquetetage de la ponctuation, des émoticônes et des sons.
- **Analyseurs morphologiques :** des analyseurs morphologiques réalisent la segmentation, la lemmatisation et l'étiquetage pour déterminer des traits morphologiques et l'appartenance d'un mot à l'ASM, à un dialecte (DZ,TN, MA ou EG), au français ou à l'anglais. Nous utilisons par ailleurs le système présenté dans (Fluhr *et al.*, 2012) pour l'identification des mots étrangers, des entités nommées et des mots dialectaux (les mots hors vocabulaires) dans des textes arabizi. Après une phase de translitération et d'étiquetage des mots en arabizi et des noms propres, nous appliquons une analyse morphologique sur le texte arabe obtenu. Enfin, nous annotons les éléments obtenus et les textes en arabe via le système présenté dans (Saâdane, 2013; Saâdane, 2015).
- **Liste des entités nommées :** ressources proposées par ANERGazet (Benajiba & Rosso, 2007) et GeoNames [3] pour identifier les entités nommées arabes, français et anglais, dont :

3. http ://download.geonames.org/export/dump/

— **personnes :** 7387 entrées,
— **locations :** 73892 entrées (pays, villes, continents, etc.),
— **organisation :** 22772 entrées (entreprises, équipes de football, etc.).

Une étape de combinaison permet d'agréger plusieurs composants, (analyseurs morphologiques, dictionnaires des entités nommées) afin d'effectuer l'identification de la langue du texte en entrée. Chaque mot de la phrase peut obtenir différentes étiquettes selon chaque composant. L'étape de combinaison, en se basant sur les étiquettes générées, utilise de plus un ensemble de règles de décision qui affectent un tag final à chaque mot de la phrase d'entrée, en donnant la priorité aux règles les plus sûres (ponctuations, dictionnaires) tout en conservant de l'ambiguïté, en particulier si les mots sont reconnus comme pouvant appartenir à plusieurs dialectes. Les règles de décision utilisées sont présentées comme suit :

— Si le mot contient des numéros ou des signes de ponctuation, alors il est associé à la balise *Autre* (Ponct, NUM, etc) ;
— Sinon si le mot est présent dans l'un des dictionnaires ou si les analyseurs morphologiques assignent la balise entité nommée, alors le mot est étiqueté comme *Entité nommées* <*NE*> ;
— Sinon si le mot est identifié par les analyseurs que ce soit *Lang1* ou *Lang2*, le mot est alors associé à l'étiquette correspondante ;
— Sinon si le mot identifié est associé à la fois à *Lang1* et *Lang2*, alors nous attribuons au mot les balises *Lang1* et *Lang2*. Toutefois ce cas introduit une certaine ambiguïté.
— Sinon si les analyseurs n'étiquettent pas le mot, cela veut dire que le mot est considéré comme un mot hors vocabulaire, alors nous associons la balise <*UNK*> au mot analysé.

4.4 Désambiguïsation des messages

Après avoir annoté les mots des messages, le système de détection de dialecte a pour objectif de déterminer le dialecte du corpus analysé, en général ou texte par texte. Pour cela, les seuls indices disponibles sont les annotations dialectales présentées dans l'étape précédente. Deux méthodes ont été implémentée pour déterminer le dialecte.

La première est basée sur le nombre de mots discriminants, soit ceux qui n'appartiennent qu'à un seul des dialectes candidats. L'hypothèse est que la présence d'un mot discriminant suffit à donner une bonne idée du dialecte. Cette approche permet de renvoyer le nombre d'occurrences de chaque label ainsi que le dialecte détecté. L'inconvénient de cette méthode est que la détection est impossible si le message ne contient aucun mot propre à un dialecte donné, auquel cas le système fait appel à la deuxième méthode.

La seconde méthode est basée sur un système de notation pondéré. Pour un mot donné, le poids d'un dialecte est inversement proportionnel au nombre de dialectes détectés. Par sommation sur les mots d'un texte, cette méthode renvoie le dialecte qui a le poids le plus important.

5 Systèmes statistique de référence

Pour la méthode statistique, deux approches ont été mises en œuvre et évaluées. Elles diffèrent par le niveau d'analyse : la première s'appuie sur les mots (approche lexicale), la seconde sur les caractères (approche graphémique). Des modèles de langage N-grammes de mots ou de caractères sont construits

pour chaque dialecte, et un modèle probabiliste détermine le langage le plus vraisemblable pour un texte donné. La qualité du modèle dépend donc entièrement du corpus d'apprentissage et de sa représentativité pour chaque dialecte, directement liée à la quantité de données disponibles. Lors de la phase de test, le système calcule le score de chaque message de test avec les modèles n-grammes de chaque dialecte, et renvoie celui qui obtient le meilleur score.

6 Résultats expérimentaux

6.1 Corpus de test

Nous avons sélectionné 820 messages issus des corpus dialectaux, écrits en caractères arabes et latins. Les deux systèmes linguistiques et statistiques sont comparés en calculant le taux d'erreur d'identification des dialectes (LER), c'est-à-dire la proportion de messages pour lesquels le dialecte est incorrectement détecté. Deux méthodes sont évaluées sur ces corpus : la première est purement statistique, la seconde exploite les dictionnaires comme décrits en section 4.1. Les résultats des deux méthodes sont présentés dans le tableau 2.

Dial. (#nb-ar/#nb-lat)	Approche lexicale	Approche graphémique	Approche dictionnaire
ALG (102/100)	38,6% (38/40)	39,1% (37/42)	20,3% (27/14)
EGY (105/103)	03,3% (0/7)	03,3% (0/7)	03,3% (2/5)
MAR (101/102)	15,3% (13/18)	11,3% (11/12)	14,8% (15/14)
TUN (100/107)	08,7% (10/8)	15,0% (22/9)	20,3% (26/16)
TOUS (408/412)	16,3% (61/73)	17,1% (70/70)	14,5% (70/49)

TABLE 2 – Taux d'erreurs de l'identification avec les méthodes linguistique et statistique

6.2 Résultats pour la méthode statistique

Pour la méthode statistique, le taux d'erreurs pour l'identification du dialecte (LER) obtenu sur les corpus de test est de 16.3% avec l'approche lexicale et de 17.1% avec l'approche graphémique. D'un point de vue global, les résultats selon que les algorithmes soient testés sur des textes écrits en arabe ou en latin sont assez similaire : 17,2% en arabe versus 17,0% pour le latin avec l'approche graphémique, et respectivement 15,0% et 17,7% pour l'approche lexicale. Pour ce qui concerne les dialectes, nous constatons d'emblée la difficulté de distinguer les textes algériens et marocains. L'égyptien est particulièrement bien détecté, ce qui s'explique de par la distance linguistique de l'arabe égyptien avec les autres dialectes. Les deux approches donnent des résultats assez proches, à l'exception du marocain (l'approche graphémique fonctionne mieux) et du tunisien (c'est l'approche lexicale qui l'emporte).

6.3 Résultats pour la méthode à base de dictionnaire

L'approche à base de dictionnaire donne un taux d'erreur de 14,5%. Sur le corpus en écriture latine, ce taux d'erreur descend même jusqu'à 11,8%, ce qui pourrait être lié à la présence de voyelles comme indice discriminant supplémentaire par rapport à l'écriture arabe. Il y a peu d'erreurs pour

l'égyptien. L'algérien obtient un LER de 26,4% , important mais bien moins élevé que celui obtenu avec les approches statistiques (plus de 35%). Le marocain obtient des résultats proches de ceux des approches statistiques. Enfin, seul le tunisien présente de moins bons résultats, avec 20,3% (contre 11% en moyenne avec les approches statistiques).

En complément, nous avons étudié les résultats ambigus de la méthode à base de dictionnaire. Il est en effet possible que cette méthode ait un choix entre plusieurs dialectes avant de prendre une décision. Un oracle permettrait alors de choisir le bon dialecte parmi ceux envisagés. Tel est le cas pour 29/24 messages algériens, 5/4 égyptiens, 31/12 marocains et 30/21 tunisiens. Le même constat est dressé : la confusion porte sur l'algérien, le marocain et le tunisien, et une méthode plus fine serait bienvenue pour distinguer ces dialectes.

6.4 Discussion

Sans surprise, cette étude montre que la détection des dialectes est une tâche difficile lorsque l'on confronte des dialectes proches, comme cela a été constaté pour l'algérien, le marocain et le tunisien, et ceci quelle que soit la méthode choisie. Les résultats rapportés montrent que, dans les conditions d'expérimentation réalisées, la méthode statistique basée sur des n-grammes de mots ou de caractères est confrontée aux même difficultés qu'une approche basée sur des ressources lexicales (dictionnaires) et morphologiques (règles de flexion). Il semble pourtant que cette dernière donne de meilleurs résultats, ce qui pourrait être lié à l'insuffisance des corpus d'apprentissage par rapport aux lexiques constitués sur lesquels s'appuient l'approche à base de dictionnaire. Comme cela a déjà été constaté pour d'autres tâches, un meilleur contrôle de la qualité des ressources permet de garantir de meilleurs résultats en sortie. Une analyse plus détaillée des résultats nous a permis de constater quelques difficultés pour constituer un corpus de qualité : messages écrits dans un dialecte mais collectés dans un corpus d'un autre dialecte, présence de mots anglais ou français identiques à des translittération de mots dialectaux. Même si ces phénomènes restent marginaux, il reste impératif, pour évaluer correctement les résultats, de porter grande attention à la qualité des corpus.

Une avancée importante de notre travail porte sur le traitement de l'arabizi. Non seulement cette écriture peut être translittérée en arabe, mais nous constatons de plus que l'ajout des voyelles courtes par rapport à l'écriture arabe, aide à mieux distinguer des dialectes.

7 Conclusion

Dans cet article, nous avons décrit un système de détection de l'origine dialectale de textes écrits en caractères arabes ou latins (arabizi). Nous avons constitué des corpus pour quatre dialectes auxquels nous nous sommes intéressés : l'algérien, le tunisien, le marocain et l'égyptien. Un translittérateur de l'arabizi vers l'arabe et des analyseurs base de dictionnaires ont été développés. Les expériences montrent qu'une approche contrôlée à base de dictionnaires obtient de meilleurs résultats qu'une approche statistique, même si le système reste à améliorer : l'algérien, le tunisien et le marocain restent difficiles à distinguer. Nos futurs travaux s'orientent, d'une part, vers une évaluation à large échelle de notre outil d'identification des dialectes en vue de consolider les résultats déjà obtenus, et d'autre part, vers la constitution d'outils et de ressources pour d'autres dialectes arabes.

Remerciements

Le projet a été soutenu par la DGE (Ministère de l'Industrie) et par la DGA (Ministère de la Défense) : projet RAPID "ORELO", référencé par le N :142906001.

Références

BENAJIBA Y. & ROSSO P. (2007). Conquering the ner task for the arabic language by combining the maximum entropy with pos-tag information. In *Proceedings of Workshop on Natural Language-Independant Engineering*.

COTTERELL R. & CALLISON-BURCH C. (2014). A multi-dialect, multi-genre corpus of informal written arabic. In *LREC*, p. 241–245.

DARWISH K. (2013). Arabizi detection and conversion to arabic. p. 217—-224, Doha, Qatar. Association for Computational Linguistics.

ELFARDY H., AL-BADRASHINY M. & DIAB M. (2014). Aida : Identifying code switching in informal arabic text. p.9̃4 : Citeseer.

ELFARDY H. & DIAB M. T. (2012). Simplified guidelines for the creation of large scale dialectal arabic annotations. In *LREC*, p. 371–378 : Citeseer.

ELFARDY H. & DIAB M. T. (2013). Sentence level dialect identification in arabic. In *ACL (2)*, p. 456–461.

ESKANDER R., AL-BADRASHINY M., HABASH N. & RAMBOW O. (2014). Foreign words and the automatic processing of arabic social media text written in roman script. In *Proceedings of The First Workshop on Computational Approaches to Code Switching*, p. 1—12, Doha, Qatar. Association for Computational Linguistics.

FLUHR C., ROSSI A., BOUCHESECHE L. & KERDJOUDJ F. (2012). Extraction of information on activities of persons suspected of illegal activities from web open sources. p.1̃9.

HABASH N., DIAB M. T. & RAMBOW O. (2012). Conventional orthography for dialectal arabic. In *LREC*, p. 711—-718, Istanbul, Turkey.

HABASH N., SOUDI A. & BUCKWALTER T. (2007). On arabic transliteration. In *Arabic computational morphology*, p. 15–22 : Springer.

JARRAR M., HABASH N., AKRA D. & ZALMOUT N. (2014). Building a corpus for palestinian arabic : a preliminary study. In *Proceedings of the EMNLP 2014 Workshop on Ar-abic Natural Langauge Processing (ANLP)*, p. 18—-27, Doha, Qatar. Association for Computational Linguistics.

SAÂDANE H. (2015). Traitement automatique de l'arabe dialectalisé : aspects méthodologiques et algorithmiques. In *PhD thesis, Université Grenoble Alpes*.

SAÂDANE H. (2013). Une approche linguistique pour l'extraction des connaissances dans un texte arabe. Les Sables d'Olonne, France : TALN-Récital.

SAÂDANE H., GUIDERE M. & FLUHR C. (2013). La reconnaissance automatique des dialectes arabes à l'écrit. In *colloque international «Quelle place pour la langue arabe aujourd'hui*, p. 18–20.

SAÂDANE H. & HABASH N. (2015). A conventional orthography for algerian arabic. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, p. 69–79, Beijing, China : Association for Computational Linguistics.

SAÂDANE H., ROSSI A., FLUHR C. & GUIDERE M. (2012). Transcription of arabic names into latin. In *2012 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT)*, p. 857–866.

YANG Y. & EISENSTEIN J. (2013). A log-linear model for unsupervised text normalization. In *EMNLP*, p. 61–72.

ZAIDAN O. & CALLISON-BURCH C. (2014). Arabic dialect identification. p. 171–202.

ZAIDAN O. F. & CALLISON-BURCH C. (2011). The arabic online commentary dataset : An annotated dataset of informal arabic with high dialectal content. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies : Short Papers - Volume 2*, HLT '11, p. 37–41, Stroudsburg, PA, USA : Association for Computational Linguistics.

ZRIBI I., BOUJELBANE R., MASMOUDI A., ELLOUZE M., BELGUITH L. H. & HABASH N. (2014). A conventional orthography for tunisian arabic. In *LREC*, p. 2355—-2361, Reykjavik, Iceland.