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

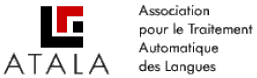
Orléans, France – 26-30 juin 2017

https://taln2017.cnrs.fr

Actes de TALN 2017, volume 3 : démonstrations

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


UNIVERSITÉ D'ORLÉANS


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 « demo »

Les TIC au service de l'enseignement : Cas de la formation et auto-formation de la langue amazighe
Boumediane Mounia 1

Wordsurf : un outil pour naviguer dans un espace de « Word Embeddings
Philippe Suignard 8

Un outil pour la manipulation de ressources arborées
Yannick Parmentier 11

Un étiqueteur en ligne du Français
Yoann Dupont, Clément Plancq 15

Apprentissage d'agents conversationnels pour la gestion de relations clients
Benoit Favre, Frederic Bechet, Géraldine Damnati, Delphine Charlet 17

Conception d'une solution de détection d'événements basée sur Twitter
Christophe Servan, Catherine Kobus, Yongchao Deng, Cyril Touffet, Jungi Kim, Inès Kapp, Djamel Mostefa, Josep Crego, Aurélien Coquard, Jean Senellart 19

Une plateforme de recommandation automatique d'emojis
Gaël Guibon, Magalie Ochs, Patrice Bellot 21

Un outil modulaire libre pour le résumé automatique
Valentin Nyzam, Aurélien Bossard 24

Uniformisation de corpus anglais annotés en sens
Loïc Vial, Benjamin Lecouteux, Didier Schwab 27

Résumer automatiquement en ligne : démonstration d'un service web de résumé multidocument
Valentin Nyzam, Nathan Gatto, Aurélien Bossard 30

Traitement automatique de la langue biomédicale au LIMSI
Christopher Norman, Cyril Grouin, Thomas Lavergne, Aurélie Névéol, Pierre Zweigenbaum . . 33

Proxem Studio : la plate-forme d'analyse sémantique qui transforme l'utilisateur métier en text scientist
Francois-Regis Chaumartin 35

Translittération automatique pour une paire de langues peu dotée
Ngoc Tan Le, Fatiha Sadat, Lucie Ménard 37

Motor, un outil de segmentation accessible en ligne
Guillaume de Malézieux, Jennifer Lewis-Wong, Vincent Berment 41

Les TIC au service de l'enseignement : Cas de la formation et auto-formation de la langue amazighe

Mounia Boumedianente[1]
(1) Institut Royal de la Culture Amazighe
Centre des Etudes Informatiques, des Systèmes d'Information et de Communication
AV. Allal El Fassi, Madinat AL Irfane,BP 2055, Hay Riad-Rabat
boumediane.mounia@gmail.com

RÉSUMÉ

Le Centre des Études Informatiques, des Systèmes d'Information et de Communication (CEISIC), issu de l'Institut Royal de la Culture Amazighe (IRCAM), fédère au sein du portail TALAM, un ensemble de ressources linguistiques informatisées et d'outils de traitement de la langue dédiées à l'amazighe. Dans ce qui suit, nous présenterons les di érentes ressources, applications et outils linguistiques développés en langue amazighe pour accompagner toute personne, de di érente tranche d'âge, assoi ée d'apprendre la langue Amazighe.

ABSTRACT

Information and Communication Technologies for Education: the case of amazigh language teaching and self-teaching.

The Center of Computer Studies, Information Systems and Communication (CEISIC), from the Royal Institute of Amazigh Culture (IRCAM), federates within the TALAM portal a set of computerized linguistic resources and processing tools Of the amazighe language dedicated to facilitate the teaching and the learning of this language to the non native speakers and as well as to native speakers, whom don't practice it in their dually life. Those tools, applications and linguistics resources are developed in Amazighe language in the center of Computer Studies, Information Systems and Communication and have as main objectives to accompany the learners whether kids or adults through the proceeding of his formation and self formation. In this article, I will present the di erent resources and applications developed in Amazighe language to facilitate to the willing learners the Amazighe language.

MOTS-CLÉS : Ressources linguistiques amazighes, applications, traitement de la langue amazighe .

KEYWORDS: Linguistic resources, Amazigh applications and tools.

1 Introduction

L'Institut Royal de la culture Amazighe est une institution publique dotée de la pleine capacité juridique et l'autonomie nancière. Elle a pour mission de « donner avis à sa Majesté sur les mesures de nature à sauvegarder et à promouvoir la culture amazighe dans toutes ses expressions »[1]. Pour accomplir ses vocations et attributions académiques et administratives, l'IRCAM, est organisée selon une structure administrative qui englobe tous les départements et services administratifs et une structure scienti que et académique qui se réfère aux 07 centres de recherche, à nommer ; le Centre d'Aménagement Linguistique, le Centre de la Traduction, l'Edition, la Documentation et la Communication, Centre des Etudes Historiques et Environnementales, le Centre des Études Anthropologiques et Sociales et le Centre des Etudes informatiques des systèmes d'information et de communication. Ce centre, en abréviation CEISIC, est dédié à la recherche scienti que qui se base sur les sciences informatiques et les technologies de l'information et de la communication pour investir les di érents aspects de recherche dans les domaines de la langue et de la culture amazighes, Le CEISIC a pour objectif de fédérer les compétences universitaires nationales pour faire aboutir les aspects de recherches béné ques à la langue et à la culture amazighes et d'encourager leur développement au sein de l'université et partant dans le tissu industriel national.

Pour encourager justement l'apprentissage et la formation et l'autoformation aux di érentes cibles non initiées à la langue amazighe, le CEISIC a pu réaliser des produits et supports en langue amazighe pour satisfaire les besoins d'une tranche de population amazighophone qui ne pratique pas sa langue maternelle et autre non amazighophone mais qui désire apprendre et connaître cette langue « comme étant une langue o cielle de l'Etat, en tant que patrimoine commun de tous les marocains sans exception. »[2]

Dans cet article, nous acharnerons à démontrer les ressources, outils et applications développés en langue amazighe pour accompagner l'apprenant de la langue amazighe d'acquérir des connaissances en cette langue et de même d'en faciliter ce processus pour les enfants en présentant des jeux ludiques visant la formation et l'auto-formation à travers des outils de TIC.

1.1 Ressources, outils et applications développés en langue amazighe

Mes premiers mots en Amazighe

Destinée à l'enfant de 2 à 12 ans, Awal inu amzwar, est une application ludique. Elle a pour vocation de motiver et d'appuyer l'apprentissage de la langue amazighe chez l'enfant. Tout en s'amusant, elle permet à l'enfant d'apprendre, à travers des jeux d'écriture et de dessin,

1 Texte du Dahir portant création de l'Institut Royal de la culture amazighe.

2 Royaume du Maroc, Secrétariat Général du Gouvernement, (2011), La constitution, série « documentation juridique marocaine».

l'alphabet ti naghe et les couleurs. A travers des jeux interactifs, cette application propice à l'enfant de découvrir et d'acquérir un vocabulaire de base sur les formes, les fruits, les légumes, les animaux, et les parties du corps humain.

Grâce à son interface tactile simple et directive, l'enfant peut apprendre à écrire l'alphabet ti naghe. In voix humaine, il peut écouter l'appellation des lettres de l'alphabet et d'apprendre la prononciation du vocabulaire. Et avec un jeu d'évaluation, il peut tester ses compétences de reconnaissance et renforcer sa mémoire lexicale.

Pour visualiser l'application, veuillez consulter l'adresse suivante : https://play.google.com/store/apps/details?id=com.ircam.vocabetlettres&hl=fr

Chi res en Amazighe

Tout comme Mes premiers mots en Amazighe, Izwiln s tamazight, est une application éducative qui cible l'enfant de 2 à 12 ans. Cette application engendre une série de jeux éducatifs visant à appuyer l'apprentissage ludique de la langue amazighe. Tout en s'amusant, elle permet à l'enfant d'apprendre, à travers des jeux d'écriture et de comptage, les chi res et les nombres en amazighe.

Pour accéder à cette application consulter l'adresse suivante : https://play.google.com/store/apps/details?id=ma.ircam.chi res

Conjugueur de la Langue Amazighe

Destiné aux enseignants et apprenants de la langue amazighe, le conjugueur Amssfti est un outil linguistique, conçu à la base de modèles de comportement exionnel augmenté de règles de régularisation morphologiques. De plus, il permet la conjugaison en ligne des formes verbales simples et dérivées dans les di érents aspects et modes de la langue amazighe.

Pour accéder au conjugueur, veuillez consulter l'adresse suivante : http://tal.ircam.ma/conjugueur/

Lexique électronique AMazighe sur mobile (LEXAM)

LEXAM est la première application mobile Android. Elle présente un lexique amazighe marocain standardisé qui couvre des domaines de la vie courante et moderne touchant les médias, l'administration, l'art, l'environnement, la civilité, le droit, la justice, l'éducation, et bien d'autres. La version actuelle contient plus de 4000 entrées amazighes. Pour une navigation plus prompte, veuillez consulter l'adresse : http://tal.ircam.ma/talam/lexam.php

Transcodeur

Cet outil de traitement automatique de la langue amazighe permet de convertir le codage des textes amazighes saisis en ti naghe à la base de l'ANSI à Unicode.

Translittérateur

Le translittérateur est un outil à accès libre. Il permet de convertir les textes amazighes transcrits en arabe ou latin à des textes écrits en ti naghe.

CONCORDANCIER

Amazigh Concord est un concordancier en ligne supportant, en outre de la langue arabe et française, les scripts latins et Unicode de la langue amazighe. Il réalise et exploite les concordances de textes écrits. Il permet d'a cher les occurrences d'un mot ou d'une expression avec son contexte d'apparition dans un texte ou un corpus. Conçu en trois langues (français, arabe, amazighe), ce qui permet à l'usager d'exploiter l'outil selon sa préférence langagière. Il est basé sur une technologie Web utilisant le PHP comme langage de développement.

Cet outil permet d'une part la recherche d'un mot ou d'une expression dans un ensemble de textes disponibles sous forme électronique, constituant un corpus, a n de répondre au besoin d'exploration suscité par l'usager. D'autre part, il permet l'a chage des fréquences des mots utilisés dans le corpus ainsi que le contenu recherché dans tous les contextes de ce corpus, dans le but d'étudier le sens et les règles d'emploi.

Pour accéder à l'application, veuillez cliquer sur ce lien : http://tal.ircam.ma/concord/application.php.

Terminologie

La terminologie est un site en ligne qui représente le référentiel de la langue amazighe standard.
Pour d'ample détail sur ce site, veuillez consulter le lien suivant : http://tal.ircam.ma/talam/ref.php

Ad nlmd ti naghe

Ad nlmd ti naɣ est une application qui a pour but l'apprentissage de l'alphabet ti naghe. Elle est destinée aux enfants âgés de 4 à 6 ans. L'application présente une approche interactive favorisant la mise en œuvre des capacités d'observation, d'écoute, d'association, du tri, de production et d'auto-évaluation. En s'exerçant, l'enfant travaille sa mémoire visuelle et auditive et mobilise son attention, et tout en s'amusant, il développe les compétences de lecture et d'écriture au moyen d'une méthode à la fois structurée et ludique.
Pour télécharger cette application, veuillez consulter cette adresse : http://tal.ircam.ma/talam/support.php

Animaux sauvages et domestiques

Imudar n lmɣrib est un double CD-ROM. Le premier CD porte sur les animaux sauvages et domestiques tandis que le deuxième sur les oiseaux et les insectes. C'est un produit entièrement en amazighe et en caractère ti naghe, ce qui permet à l'utilisateur de se familiariser avec la faune tout en apprenant la langue amazighe à son propre rythme à travers le texte, la vidéo, la voix et les images.
Pour télécharger cette application, veuillez consulter cette adresse : http://tal.ircam.ma/talam/support.php

École Amazighe

L'École Amazighe est une application multimédia qui facilite l'apprentissage de la langue amazighe et met en exergue le caractère Ti naghe. Elle présente plusieurs modules accessibles à partir d'une page principale, traitant chacun un thème particulier tels que les chi res, l'alphabet et le calendrier. L'application inclut aussi quelques exercices qui o rent une meilleure compréhension de di érents modules.
http://tal.ircam.ma/talam/support.php

Tamawalt inu tawlafant

Ce dictionnaire imagier thématique sur CD vise à servir comme support pour l'enseignement-apprentissage de la langue amazighe. Bien que cet imagier ne prétende pas à l'exhaustivité, son utilisation didactique peut concourir à la mise en œuvre des compétences lexicales chez les enfants de 5 à 11 ans, en les aidant à construire des capacités perceptives et logiques telles l'observation, le repérage et le classement des unités lexicales.
http://tal.ircam.ma/talam/support.php

Tamawalt

Destiné aux enfants, TAMAWALT est un dictionnaire trilingue mis en ligne. Il vise à assister les petits apprenants de la langue amazighe en leur introduisant les objets par image en langue arabe, française ou anglaise.
Pour le consulter veuillez saisir l'adresse suivante : http://tal.ircam.ma/talam/support.php

À noter que ces di érents outils, ressources et supports linguistiques sont accessibles et téléchargeables via le site http://tal.ircam.ma.

2 Conclusion

Les technologies d'information et de communication, comme a rme Jürgen Wagner[3] permettent certainement de faire entrer le monde extérieur dans les salles de classe et de traiter de sujets d'actualité en exploitant des documents (texte, audio, vidéo) authentiques. Pour ceux qui s'en servent, elles apportent une plus-value. Quant est il pour les langues peu dotées exemple la langue amazighe ? Les langues peu dotées ont du mal à s'imposer dans le domaine des nouvelles technologies d'information et de communication. Ceci est dû selon H. Fadili[4] aux retards et di cultés cumulés depuis plusieurs années empêchant leur intégration dans les nouveaux systèmes.

[3] http://www.bonjourdefrance.com/exercices/contenu/20/civilisation/604.html. Interview avec Jürgen Wagner,chargé de mission dans le domaine du e-learning au Landesinstitut für Pädagogik und Medien à Saarbrücken (Allemagne), professeur d'anglais et de français depuis 30 ans et intervenant de conférences en ligne, de congrès et de colloques. Administrateur de nombreux sites, auteur d'une infolettre sur les TIC et d'un guide des bonnes pratiques du Web 2.0 en classe de langue. Consulté, le 21 mars 2017, à 11 :44.

[4] (Fadili, 2012)

Les problèmes rencontrés peuvent êtes principalement d'au moins deux types : ceux liés au support des contraintes technologiques, puis ceux liés à la création et à la mise à disposition des contenus dans des formats compatibles. Pour assurer leur avenir dans le monde connecté de l'Internet et par conséquent, leur avenir tout court, les responsables et acteurs du domaine des langues peu dotées ont le devoir de s'activer pour remédier à ces problèmes a n d'assurer leur évolution et leur modernisation. Parallèlement, ils doivent prendre en charge leur développement d'un point de vue scienti que pour pouvoir y véhiculer la production, la traduction et la di usion des savoirs entant que langues savantes, aspect faisant défaut et constituant un grand handicap aujourd'hui. Surmonté les di cultés technologiques ne peuvent aboutir sans une concertation dores déjà de part les pédagogues et les andragogues qui doivent ouvrer pour l'amélioration de la qualité du processus de l'enseignement. A notre sens, ce processus, en lui-même, ne peut aboutir sans une formation continue et permanente des enseignants et chercheurs dans le domaine des technologies d'information de communication.

Références

FADILI H. (2012). « Problématiques d'usage et d'integration des langues peu dotees dans le web des donnees ouvertes (linked open data ou lod) : cas de l'amazighe ». Actes de la *5ème édition de la Conférence sur les Technologies de l'Information et de la Communication pour la langue Amazighe (TICAM'12)*, publication IRCAM.

IRCAM. (2017) Texte du Dahir portant création de l'Institut Royal de la culture amazighe, www.IRCAM.ma, consulté le, 21 mars 2017 à 10h00.

ROYAUME DU MAROC, SECRÉTARIAT GÉNÉRAL DU GOUVERNEMENT, (2011), La constitution, série « documentation juridique marocaine».

Actes de la 5ème conférence sur les technologies d'information et de communication pour la langue amazighe, (2012), Publication de l'IRCAM, Imprimerie El Maârif Al Jadida.

Wordsurf : un outil pour naviguer dans un espace de « Word Embeddings »

Philippe Suignard
EDF R&D, 7 boulevard Gaspard Monge, 91120 Palaiseau, France
philippe.suignard@edf.fr

RESUME

Dans cet article, nous présentons un outil appelé « Wordsurf » pour faciliter la phase d'exploration et de navigation dans un espace de « Word Embeddings » préalablement entrainé sur des corpus de textes avec Word2Vec.

ABSTRACT

Wordsurf : a tool to surf in a "word embeddings" space
In this article we present a tool called "Wordsurf" to facilitate the exploration and navigation phase in a "Word Embeddings" space previously trained on textual corpus with Word2Vec.

MOTS-CLÉS : Word2Vec, GloVe, word embeddings, plongement de mots
KEYWORDS: Word2Vec, GloVe, word embeddings

1 Contexte

Depuis quelques années, sont apparues des méthodes appelées « Word Embeddings » (notées WE par la suite) ou méthodes de plongement de mots en français, comme Word2Vec (Mikolov et al. 2013) ou Glove (Pennington et al. 2014), permettant de transformer des mots en vecteurs, c'est-à-dire de passer d'un espace de représentation discontinu (les mots) à un espace continu (espace vectoriel de grande dimension). Pour ce faire, Word2Vec s'appuie sur des corpus volumineux de données textuelles et utilise un réseau de neurones peu profond cherchant à prédire le mieux possible le contexte des mots. La représentation numérique de ces mots est ensuite très utile pour différentes tâches comme la classification, le clustering, la traduction, l'analyse d'opinions, etc.

Par ailleurs, EDF doit maintenant gérer des corpus textuels de plus en plus nombreux et variés : réclamations de clients, question posées à Laura l'avatar ou chatbot situé sur le site web d'« EDF Particuliers », comptes rendus d'interventions techniques, etc. En compléments des techniques d'exploration de corpus dites « classiques », comme le clustering proposé par Iramuteq (Ratinaud 2009), les technologies de type WE commencent à être utilisées à EDF pour « découvrir » ou explorer le contenu de ces différents corpus.

Pour faciliter cette phase d'exploration et de découverte, et comme les WE sont assez difficiles à interpréter, nous avons développé un outil appelé « Wordsurf » permettant de naviguer ou de « surfer » dans de telles bases ou espaces de mots.

2 L'outil Wordsurf

Wordsurf est un logiciel interactif développé en Java. Il permet de lire un modèle de mots préalablement entrainé par Word2Vec sur un corpus de textes. La démonstration sera illustrée par des résultats obtenus sur un corpus de questions posées à l'agent conversationnel ou chatbot Laura. Les fonctionnalités de Wordsurf sont les suivantes :

– Pour un mot donné, pouvoir rechercher les mots les plus « similaires » et les afficher sous la forme de nuage ou de liste de mots ;
– Visualiser l'évolution de ce nuage de mots en fonction des itérations du modèle ;
– Calculer des similarités entre phrases. Pour cela, trois mesures de similarité ont été implémentées issues de (SONG et al. 2016) ainsi que deux autres reposant sur la comparaison de la distribution des voisins des mots constituant les deux phrases ;
– Un peu plus anecdotique, voire parfois poétique, une fonction permet de calculer les interpolations entre deux mots donnés ;
– Pour un mot particulier, une liste de mots ou tous les mots du corpus, calculer un graphe où les nœuds sont les mots et les arcs traduisent la proximité entre ces mots. Le graphe peut être exporté vers le logiciel Gephi (Bastian et al. 2009) ;
– Pour un mot particulier, une liste de mots ou tous les mots du corpus, générer une visualisation des mots via l'algorithme t-SNE (Maaten et al. 2008) ;
– Pouvoir comparer qualitativement deux modèles « Word2Vec » entrainés avec des paramètres différents : pour cela, le logiciel liste les mots dont la représentation est la plus éloignés entre les deux modèles.

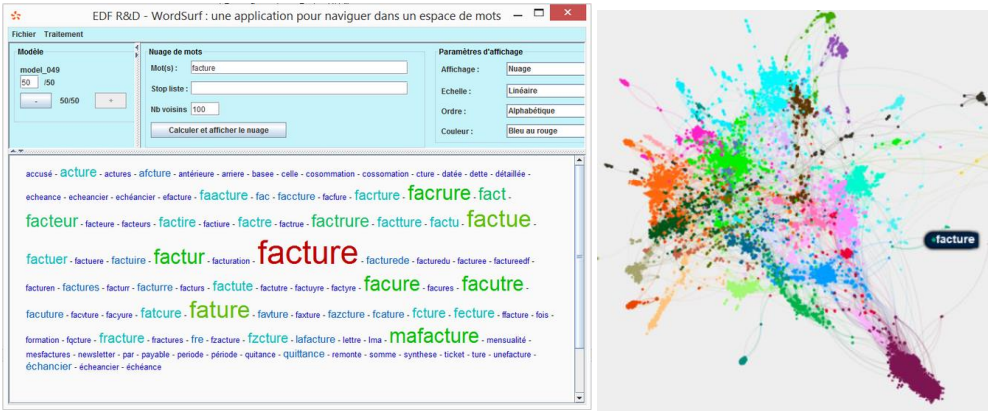


Figure 1 : Interface de l'outil Wordsurf et un exemple d'export de mots via Gephi.

La partie gauche de la Figure 1 présente l'interface de l'outil Wordsurf. On distingue un nuage de mots constitués des 100 mots les plus similaires au mot « facture ». La partie droite de la Figure 1 représente les mots du modèle ainsi que leur proximité exportés au format HTML grâce à un plug-in du logiciel Gephi. Les différentes couleurs représentent les univers lexicaux : les appareils électriques (four, radiateur, etc.), le bâti (toit, coffret, câble, poteau, etc.), les offres, les factures, les justificatifs, etc.

Références

BASTIAN M., HEYMANN S., & JACOMY M. (2009). Gephi: an open source software for exploring and manipulating networks. ICWSM, 8, 361-362.

MAATEN L. V. D., & HINTON G. (2008). Visualizing data using t-SNE. Journal of Machine Learning Research, 9(Nov), 2579-2605.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. S., & DEAN J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111-3119).

PENNINGTON J., SOCHER R., & MANNING C. D. (2014, October). Glove: Global Vectors for Word Representation. In EMNLP (Vol. 14, pp. 1532-1543).

RATINAUD P. (2009). IRAMUTEQ : Interface de R pour les Analyses Multidimensionnelles de TExtes et de Questionnaires. *Téléchargeable à l'adresse : http://www.iramuteq.org*

SONG Y., MOU L., YAN R., YI, L., ZHU Z., HU X., & ZHANG M. (2016). Dialogue session segmentation by embedding-enhanced texttiling. *arXiv preprint arXiv:1610.03955*.

Un outil pour la manipulation de ressources arborées

Yannick Parmentier[1]
(1) LIFO, Université d'Orléans, 45067 Orléans, France
yannick.parmentier@univ-orleans.fr, https://gitlab.com/parmenti/pytreeview

RÉSUMÉ

Dans cet article, nous présentons brièvement pytreeview, un outil pour la manipulation de ressources arborées (corpus annotés, grammaires électroniques). Initialement conçu pour assiter les utilisateurs linguistes dans leur tâche de développement de grammaires arborescentes, pytreeview a été étendu pour permettre de manipuler des ressources arborées variées (grammaires mais aussi corpus aux formats FTB, PTB, CoNLL, Tiger), afin d'en extraire des informations utiles (par exemple la distribution des cadres de sous-catégorisation). pytreeview est actuellement utilisé dans le cadre d'un projet visant l'extraction semi-automatique de grammaires abstraites (méta-grammaires) à partir de corpus arborés.

ABSTRACT

A tool for handling tree-based linguistic resources

In this paper, we briefly introduce pytreeview, a tool for handling tree-based linguistic resources (treebanks, electronic grammars). Initially designed for supporting linguist users in their grammar development task, pytreeview has been extended to handle various tree-based resources (grammars but also corpuses in the FTB, PTB, CoNLL, Tiger formats), in order to extract useful pieces of information from these (e.g. distributions of sub-category frames). pytreeview is currently used in a project aiming at semi-automatic abstract grammar (metagrammar) extraction from treebanks.

MOTS-CLÉS : visualisateur, grammaire d'arbres, corpus arboré.

KEYWORDS: viewer, tree grammar, treebank.

1 Introduction

De nombreuses applications de traitement automatique des langues reposent (directement ou indirectement) sur l'utilisation de ressources linguistiques sous forme arborescente. La plus courante de ces applications est probablement l'analyse syntaxique (c'est-à-dire le calcul automatique des relations entre les mots d'un énoncé, pour en construire une représentation syntaxique). Dans le cas de l'analyse syntaxique symbolique, on peut citer le projet XTAG (2001) qui a permi la création d'une grammaire noyau décrivant la syntaxe de l'anglais, comptant plus de 1000 schèmes d'arbres (c'est-à-dire, d'arbres avant lexicalisation), et utilisable en analyse. Dans le cas de l'analyse statistique, les approches courantes utilisent des ressources arborés (*treebanks*) pour la phase d'apprentissage (voir par exemple (Candito *et al.*, 2010)). Dans ce contexte, il est important de disposer d'outils permettant (au moins) l'exploration de ces ressources et leur vérification par un Humain. Les différents projets autour des grammaires et corpus arborés ont souvent été accompagnés du développement d'outils d'exploration spécifiques (voir § 3). Nous présentons ici un outil *libre*, *général* et *léger* pour la visualisation et la conversion de ressources arborescentes (voir § 2).

2 Description

Partant du constat que de nombreux formats (et outils associés) existent pour la représentation de ressources arborées (par exemple, des formats de type XML pour le French TreeBank ou le corpus Tiger, des formats textuels pour le Penn TreeBank ou le corpus CoNLL), `pytreeview` a été créé pour offrir un accès *unifié* à ces ressources, dans un but de visualisation et d'extraction d'information. `pytreeview` est donc au départ une application permettant d'*explorer* une ressource arborée (tous les formats listés ci-dessus sont supportés), et d'en afficher les arbres (plusieurs arbres peuvent être affichés simultanément, et une fonction de comparaison automatique permet de colorer les arcs et nœuds divergents entre paires d'arbres). `pytreeview` offre trois modes d'utilisation : via une interface graphique (GUI, pour une utilisation interactive, voir Figure 1 en Annexe), via une interface en ligne de commande (CLI, pour appel depuis un script par exemple), et via une interface web (pour visualiser la ressource dans un navigateur, dans le cadre d'une démo en ligne par exemple, voir Figure 2 en Annexe). En mode interactif, `pytreeview` permet également de rechercher dans la ressource arborée, l'ensemble des structures satisfaisant certains critères. Ces critères sont de 2 types : contraintes de traits et contraintes de structures (dominance ou précédence entre nœuds). Le langage utilisé pour écrire les requêtes de recherche est un sous-ensemble du langage utilisé par l'outil TigerSearch (König & Lezius, 2000, §2.2 et 2.3). En plus de pouvoir visualiser des arbres à l'écran avec zoom, `pytreeview` permet leur export sous 3 formes : image au format `png`, texte aux formats `json` (format léger permettant une interopérabilité entre ressources) et `tikz` (format permettant de documenter une ressource au moyen de l'outil d'édition LaTeX). `pytreeview` permet également d'extraire un ensemble d'information d'un corpus arborés (par exemple, le jeu d'étiquettes utilisé, la distribution des étiquettes et les différents cadres de sous-catégorisation).

`pytreeview` est développé au moyen du langage `python` (3685 lignes de codes réparties en 14 fichiers) et distribué librement sous licence libre GPLv3. La conception de `pytreeview` a été guidée par des objectifs de *légèreté* (installation et utilisation simples) et d'*extensibilité* (architecture modulaire permettant d'ajouter de nouveaux formats d'entrée / sortie, et de nouveaux traitements des arbres). `pytreeview` requiert trois librairies `python` pour la manipulation des formats d'export (`svg2tikz`, `svgwrite` et `cairosvg`), et huit autres librairies optionnelles pour l'interface graphique (par ex. `pyforms` ou `wand`), toutes sont disponibles librement et pré-compilées pour les environnements linux/windows/MacOS.

`pytreeview` est en cours de développement, et est utilisé dans le cadre d 'un projet d'extraction de (méta) grammaire. Une attention particulière a été donnée à la rapidité de traitement (utilisation de techniques de programmation parallèle pour l'export web par exemple), cependant la recherche de motif peut être longue (dizaines de sec. sur un corpus de 100000 arbres) et nécessite des optimisations.

3 Travaux liés

Le développement d'outils pour la manipulation de ressources arborées est un domaine toujours actif, comme en témoignent par exemple Gerdes (2013) pour l'annotation en dépendances et Wang *et al.* (2015) pour l'exploration de la grammaire XTAG. Les outils les plus proches de notre travail correspondent à *TigerSearch* (Brants *et al.*, 2002) et *TrEd* (Pajas & Štěpánek, 2009). Ces outils rassemblent une grande communauté d'utilisateurs et permettent non seulement la visualisation mais aussi l'édition d'arbres. La différence majeure avec notre approche réside dans le fait que ces outils ont leur propre format de représentation des données et reposent sur une architecture complexe.

Références

BRANTS S., DIPPER S., HANSEN S., LEZIUS W. & SMITH G. (2002). Tiger treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories, September 20-21 (TLT02)*, Sozopol, Bulgaria.

CANDITO M., CRABBÉ B. & DENIS P. (2010). Statistical french dependency parsing : Treebank conversion and first results. In N. C. C. CHAIR), K. CHOUKRI, B. MAEGAARD, J. MARIANI, J. ODIJK, S. PIPERIDIS, M. ROSNER & D. TAPIAS, Eds., *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, p. 1840–1847, Valletta, Malta : European Language Resources Association (ELRA).

GERDES K. (2013). Collaborative dependency annotation. In *Proceedings of the Second International Conference on Dependency Linguistics, DepLing 2013, August 27-30, 2013, Prague, Czech Republic*, p. 88–97.

KÖNIG E. & LEZIUS W. (2000). A description language for syntactically annotated corpora. In *Proceedings of the 18th Conference on Computational Linguistics - Volume 2*, COLING '00, p. 1056–1060, Stroudsburg, PA, USA : Association for Computational Linguistics.

PAJAS P. & ŠTĚPÁNEK J. (2009). System for querying syntactically annotated corpora. In *Proceedings of the ACL-IJCNLP 2009 Software Demonstrations*, p. 33–36, Suntec, Singapore : Association for Computational Linguistics.

WANG Z., ZHANG H. & SARKAR A. (2015). A python-based interface for wide coverage lexicalized tree-adjoining grammars. *The Prague Bulletin of Mathematical Linguistics*, **103**(1), 139–159.

XTAG (2001). *A Lexicalized Tree Adjoining Grammar for English*. Rapport interne IRCS-01-03, IRCS, University of Pennsylvania.

Annexes

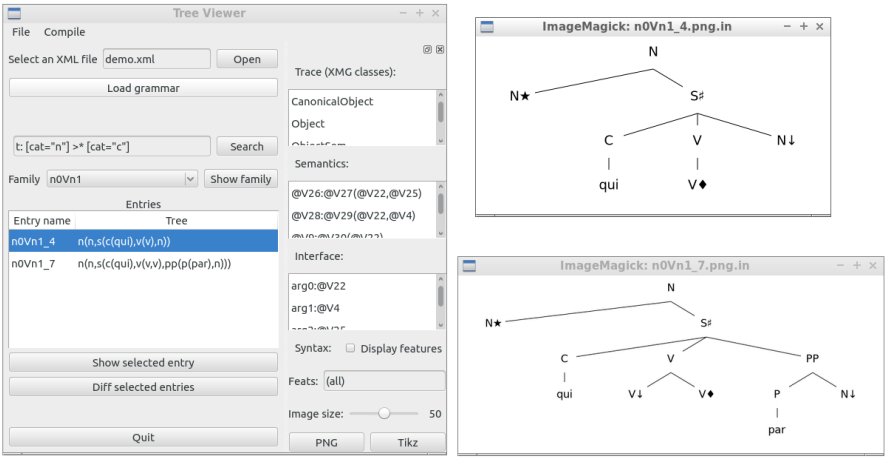


FIGURE 1 – Interface de `pytreeview` en mode *bureau*

FIGURE 2 – Interface de `pytreeview` en mode *web*

Un étiqueteur en ligne du français

Yoann Dupont[1, 2] Clément Plancq[1]
(1) Laboratoire Lattice (CNRS, ENS, Université Sorbonne Nouvelle, PSL Research University, USPC)
1 rue Maurice Arnoux, 92120 Montrouge
(2) Expert System France, 207 rue de Bercy, 75012 Paris
yoa.dupont@gmail.com, clement.plancq@ens.fr

RÉSUMÉ

Nous proposons ici une interface en ligne pour étiqueter des textes en français selon trois niveaux d'analyse : la morphosyntaxe, le chunking et la reconnaissance des entités nommées. L'interface se veut simple et les étiquetages réutilisables, ces derniers pouvant être exportés en différents formats.

ABSTRACT

An online tagger for French

We propose here an online interface for tagging French texts according to three levels of analysis : morphosyntax, chunking and named entity recognition. The interface is simple and the taggings are reusable as they can be exported in different formats.

MOTS-CLÉS : Reconnaissance d'entités nommées, French Treebank, Apprentissage automatique, CRF, IHM, en ligne.

KEYWORDS: named entity recognition, French Treebank, machine learning, CRF, GUI, online.

1 Introduction

À l'heure où les interfaces en ligne se multiplient, de plus en plus d'outils de TAL deviennent accessibles facilement, s'ouvrant alors aux non spécialistes. Dans le cadre de cette démonstration, nous nous intéresserons à la reconnaissance d'entités nommées. Il existe pour cette tâche un certain nombre d'interfaces en ligne, parmi lesquelles on peut citer celle de Cognitive Computational Group de l'Université Illinois [1] et celle d'Explosion AI [2]). De telles interfaces pour le français sont au mieux rares et sont des ressources précieuses. Notre interface [3] est une surcouche à SEM (Tellier *et al.*, 2012; Dupont & Tellier, 2014), un programme également libre [4].

Afin d'effectuer les différentes tâches d'étiquetage, nous avons entraîné un CRF (Lafferty *et al.*, 2001) sur le French Treebank (FTB) (Abeillé *et al.*, 2003). Nous avons utilisé Wapiti (Lavergne *et al.*, 2010) comme implémentation des CRF. Le FTB annoté en entités nommées (Sagot *et al.*, 2012) distingue 7 types d'entités principaux : *Company* (les entreprises), *Location* (les lieux), *Organization* (les organisations à but non lucratif), *Person* (les personnes réelles), *Product* (les produits), *FictionCharacter* (les personnages fictifs) et finalement les *PointOfInterest* (les points

1. http://cogcomp.cs.illinois.edu/page/demo_view/NERextended
2. https://demos.explosion.ai/displacy-ent/
3. accessible à l'adresse suivante : http://apps.lattice.cnrs.fr/sem/
4. disponible à l'adresse suivante : https://github.com/YoannDupont/SEM

d'intérêt). Le découpage du corpus suit le protocole entraînement–développement–test défini par Crabbé & Candito (2008). Notre étiqueteur en entités nommées atteint une précision de 87.89, un rappel de 82.34 et une f-mesure de 85.02.

2 Interface

L'interface se veut la plus simple possible, afin d'être accessible au plus grand monde. Son fonctionnement se découpe en deux passes : 1. rentrer le texte à annoter 2. récupérer le résultat dans le format attendu. Plusieurs niveaux d'analyse sont offerts à l'utilisateur : étiquetage morphosyntaxique, chunking (Abney, 1991) et entités nommées.

Une visualisation du texte en HTML avec les différents niveaux d'annotation est disponible directement afin que l'utilisateur puisse évaluer les résultats. Chaque niveau d'annotation est visible séparément et chaque élément dans un niveau spécifique est surligné, les éléments d'un même type étant surlignés avec la même couleur. La structure initiale du texte est préservée au maximum afin d'en faciliter la lecture. L'un des intérêts de notre interface est d'offrir la possibilité de récupérer les données étiquetées, en différents formats. À notre connaissance, les données ne sont que très rarement téléchargeables ou seulement en un unique format. Ces formats d'export sont les suivants : texte linéaire, tabulaire, HTML pour visualiser les annotations ainsi que json pour améliorer la réutilisation sont disponibles. Il est prévu que d'autres formats de sortie soient ajoutés, notamment le XML-TEI.

La taille maximale du texte soumis est fixée à 50 000 mots graphiques, au-delà nous conseillons aux utilisateurs d'installer le programme sur leur ordinateur. Le texte de l'utilisateur ainsi que le produit de l'annotation sont stockés dans des fichiers dont l'existence est liée à un cookie de session. La fermeture de la session web déclenche la suppression des fichiers sur le serveur.

Références

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

ABNEY S. (1991). Parsing by chunks. In *Principle-Based Parsing*, p. 257–278 : Kluwer Academic Publishers.

CRABBÉ B. & CANDITO M. H. (2008). Expériences d'analyse syntaxique statistique du français. In *Actes de TALN'08*.

DUPONT Y. & TELLIER I. (2014). Un reconnaisseur d'entités nommées du français. In *TALN 2014*, p. 40–41.

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

LAVERGNE T., CAPPÉ O. & YVON F. (2010). Practical very large scale CRFs. In *Proceedings of ACL'2010*, p. 504–513 : Association for Computational Linguistics.

SAGOT B., RICHARD M. & STERN R. (2012). Annotation référentielle du corpus arboré de paris 7 en entités nommées. In *Traitement Automatique des Langues Naturelles (TALN)*, volume 2.

TELLIER I., DUPONT Y. & COURMET A. (2012). Un segmenteur-étiqueteur et un chunker pour le français. *JEP-TALN-RECITAL 2012*, p. 7–8.

Apprentissage d'agents conversationnels pour la gestion de relations clients

Benoit Favre[1] Frédéric Béchet[1] Géraldine Damnati[2] Délphine Charlet[2]
(1) LIF/CNRS, Aix-Marseille Université, France
(2) Orange Labs, Lannion, France
prenom.nom@univ-amu.fr, prenom.nom@orange.com

RÉSUMÉ
Ce travail démontre la faisabilité d'entraîner des chatbots sur des traces de conversations dans le domaine de la relation client. Des systèmes à base de modèles de langage, de recherche d'information et de traduction sont comparés pour la tâche.

ABSTRACT
Training chatbots for customer relation management

This work demonstrates the feasability of training chatbots on customer relation conversation traces. Systems based on language models, information retrieval and machine translation are compared.

MOTS-CLÉS : Agents convesationnels, LSTM, GRU, seq2seq.

KEYWORDS: Chatbots, LSTM, GRU, seq2seq.

1 Agents conversationnels

En collectant de grandes quantités de discussions textuelles dans le cadre des services de relations client en ligne, il est aujourd'hui possible d'entraîner des agents conversationnels sur les traces de ces conversations (Shawar & Atwell, 2003). Ces "chatbots" peuvent être utilisés pour simuler le service (résoudre les problèmes des clients) lorsque des agents ne sont pas disponibles, ou simuler un client afin d'entraîner des agents débutants.

Les modèles à base d'apprentissage profond permettent de créer des agents conversationnels, en considérant le problème comme de la traduction (traduire l'historique du dialogue en l'intervention suivante), des modèles de langages (générer du texte étant donné un historique), ou de la recherche d'information (trouver l'intervention la plus propice dans la base d'apprentissage étant donné l'historique du dialogue).

La plupart des approches testées jusqu'ici s'intéressent à des échanges courts (Vinyals & Le, 2015), de type question / réponse, sans considérer les longs contextes des dialogues à finalité précise. Ces approches sont entraînées sur des traces issues de conversations IRC, de sous-titres de films ou de tweets (Lowe *et al.*, 2015). La contribution de ce travail est d'entraîner des agents conversationnels sur de grandes quantités de conversations liées à une même tâche pour explorer l'utilité de tels systèmes dans un contexte beaucoup plus réaliste, et ainsi d'en voir les limites.

2 Description des systèmes

Nous avons construit trois systèmes qui peuvent être comparés. Le premier est un modèle de langage dit "alternant". Ce modèle est entraîné à prédire la séquence de mots de la conversation, en séparant les tours des participants par un symbole spécial. Ce modèle génère une distribution de probabilité pour le mot suivant étant donné le mot courant et un état caché. Après l'entraînement, l'état caché est mis à jour sur les mots de l'humain pour ensuite générer les mots de la machine. Ces derniers sont tirés aléatoirement dans la distribution prédite jusqu'à génération d'un changement de participant. Ce modèle est implémenté dans TensorFlow comme un réseau de neurones récurrent LSTM à deux couches (Sundermeyer *et al.*, 2012), avec des couches cachées de taille 650, un vocabulaire de taille 30 000.

Le second modèle est un système à base de recherche d'information. Il créé des représentations de taille fixe pour l'historique et la prochaine intervention (appelée réponse) et est entraîné de manière à ce que les paires (historique, réponse) aient une représentation proche alors qu'elles doivent être éloignées lorsque la paire n'a pas été observée en entraînement. De cette manière, on peut rechercher la réponse la plus proche d'un historique donné pour continuer la conversation. Le désavantage de cette méthode est qu'elle ne permet pas de générer du texte nouveau, mais le concepteur de la base de réponses à un contrôle total des réponses possibles, ce qui est une contrainte pour de nombreuses applications industrielles. Ce système est implémenté dans Keras et entraîné par maximisation de la marge dans un triplet (historique, réponse, bruit). Les représentations sont extraites avec des RNN de type GRU, et ont une taille de 128. L'historique est limité aux 64 mots précédents.

Le troisième système est un système de traduction avec un mécanisme d'attention qui lui permet d'apprendre à localiser dans l'historique les mots les plus pertinents. Il fonctionne essentiellement comme le premier modèle sauf que les poids du modèle de l'historique ne sont pas partagés avec ceux du modèle pour la génération, et que le mécanisme d'attention permet de mieux tirer parti de l'historique lointain. Le système de traduction est fondé sur OpenNMT (Klein *et al.*, 2017) dans son paramétrage par défaut. L'historique est limité aux 64 mots précédents.

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

KLEIN G., KIM Y., DENG Y., SENELLART J. & RUSH A. M. (2017). Opennmt : Open-source toolkit for neural machine translation. *arXiv preprint arXiv :1701.02810*.

LOWE R., POW N., SERBAN I. & PINEAU J. (2015). The ubuntu dialogue corpus : A large dataset for research in unstructured multi-turn dialogue systems. *arXiv preprint arXiv :1506.08909*.

SHAWAR B. A. & ATWELL E. (2003). Using dialogue corpora to train a chatbot. In *Proceedings of the Corpus Linguistics 2003 conference*, p. 681–690.

SUNDERMEYER M., SCHLÜTER R. & NEY H. (2012). Lstm neural networks for language modeling. In *Interspeech*, p. 194–197.

VINYALS O. & LE Q. (2015). A neural conversational model. *arXiv preprint arXiv :1506.05869*.

Conception d'une solution de détection d'événements basée sur Twitter

Christophe Servan, Catherine Kobus, Yongchao Deng, Cyril Touffet,
Jungi Kim, Inès Kapp, Djamel Mostefa, Josep Crego et Jean Senellart

SYSTRAN, 5 rue Feydeau, 75002 Paris, France
firstname.familyname@systrangroup.fr

RÉSUMÉ

Cet article présente un système d'alertes fondé sur la masse de données issues de Tweeter. L'objectif de l'outil est de surveiller l'actualité, autour de différents domaines témoin incluant les événements sportifs ou les catastrophes naturelles. Cette surveillance est transmise à l'utilisateur sous forme d'une interface web contenant la liste d'événements localisés sur une carte.

ABSTRACT

Design of a solution for event detection from Tweeter

This article presents a detection system based on Big Data from Tweeter. The goal of the tool is to supervise news from various domains such as sport events or natural disasters. This monitoring is transmitted to the user as a website, which contains a list of events located on a map.

MOTS-CLÉS : Détection d'événements, Masse de Données, Tweeter.

KEYWORDS: Event Detection, Big Data, Tweeter.

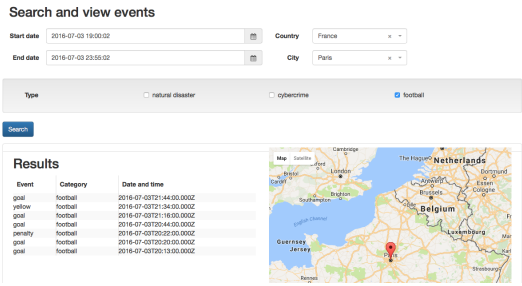


FIGURE 1 – Example de l'interface utilisateur, permettant la visualisation des données.

1 Introduction

Lors de la dernière décennie, la quantité d'information disponible sur Internet a radicalement explosé d'une part par le passage au numérique de la quasi-totalité des médias traditionnels, mais aussi par l'explosion de l'usage des réseaux sociaux et des sites d'informations non officiels de type blogs. Cette explosion à l'échelle mondiale donne théoriquement accès à une palette gigantesque d'informations multilingues mais noie pratiquement l'utilisateur final dans un excès d'information. Face à l'explosion

des flux et des sources d'information, de nouveaux outils et services voient le jour pour les exploiter. L'outil présenté dans cet article en fait partie.

L'objectif de l'outil est de surveiller l'actualité, en particulier autour de différents domaines témoin incluant les événements sportifs ou les catastrophes naturelles. L'outil utilise une approche polylingue, c'est-à-dire qu'il traite de manière native dans les différentes langues concernées (soit le français, l'anglais et l'arabe). Cet outil est l'aboutissement du projet PAPYRUS (Plateforme Adaptative PolYlingue de suRveillance de FlUx Spécialisés). Ce dernier s'intéresse à la recherche d'approche polylingues pour améliorer la précision des modules de fouilles documentaires.

2 Système de détection

Le système effectue un premier filtrage en utilisant des mots-clés relativement courants. Puis, les tweets ainsi filtrés sont classés suivant les différents événements auxquels ils appartiennent. Ces derniers peuvent n'appartenir à aucun événement et pourront être rejetés par le système de classification. L'approche de classification est fondée sur les réseaux de neurones convolutionnels tels que décrits par (Kim, 2014). Une fois la classification faite, nous étudions l'évolution de la fréquence des tweets répondant à chacun des événements, telle que décrite dans (Earle *et al.*, 2012). Les événements ainsi collectés sont stockés dans une base de données, de même que les identifiants des tweets associés à ces mêmes événements.

2.1 Géolocalisation

Afin de géolocaliser les événements, nous utilisons d'une part les informations contenues dans les tweets, lorsque ceux-ci sont géolocalisés et, d'autre part, nous utilisons une détection d'entités nommées, lorsque la géolocalisation du tweet n'est pas présente ou non-pertinente. En effet, lorsqu'un utilisateur tweete un message à propos d'un match de foot depuis son lieu d'habitation, la géolocalisation associée à ce tweet n'est pas pertinente. Par contre, la géolocalisation d'un tweet émis par un utilisateur proche d'un épicentre d'un tremblement de terre est pertinente.

2.2 Visualisation des données

La figure 1 présente un exemple de visualisation des données collectées. Ainsi, l'événement et ses caractéristiques (date, lieu, contexte, *etc.*) peuvent être affichés et visualisés sur une carte. Les tweets correspondant aux événements peuvent également être affichés dans une autre partie de l'interface, permettant une validation humaine de la détection automatique.

Remerciements

Cette réalisation a été financée à travers le projet DGA-RAPID 2014 N°1429060465 PAPYRUS

Références

EARLE P. S., BOWDEN D. C. & GUY M. (2012). Twitter earthquake detection : earthquake monitoring in a social world. *Annals of Geophysics*, **54**(6).

KIM Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 1746–1751, Doha, Qatar.

Une plateforme de recommandation automatique d'emojis

Gaël Guibon[1, 2] Magalie Ochs[1] Patrice Bellot[1]
(1) Aix Marseille Université, Université de Toulon, CNRS, ENSAM, LSIS, Marseille, France
(2) Caléa Solutions, 1 place Francis Chirat, 13002 Marseille, France
prenom.nom@lsis.org

RÉSUMÉ

Nous présentons une interface de recommandation d'emojis porteurs de sentiments qui utilise un modèle de prédiction appris sur des messages informels privés. Chacun étant associé à deux scores de polarité prédits. Cette interface permet permet également d'enregistrer les choix de l'utilisateur pour confirmer ou infirmer la recommandation.

ABSTRACT

An emoji recommandation platform

We show an emoji recommendation web interface dedicated to sentiment-related emojis. This application uses a model learnt on private informal short text messages associated with two predicted polarity scores. The application also saves the user's choices to validate or invalidate the recommendation.

MOTS-CLÉS : emoji, recommandation, apprentissage automatique, medias sociaux, messagerie.

KEYWORDS: emoji, recommendation, machine learning, social media, messaging application.

1 Introduction

Les emojis sont l'un des principaux vecteurs d'émotions et de sentiments. Depuis leur création dans les années 90 et leur instauration dans le clavier de l'iPhone en 2011, les emojis sont de plus en plus présents dans le paysage actuel de la communication écrite sous presque toutes ses formes : SMS, message instantané, chat, forum, email, etc. Avec 2389 emojis "standards" fin 2016, et 2683 à ce jour [1], leur nombre ne cesse d'augmenter. Il convient donc de pouvoir recommander les emojis non plus uniquement à l'aide d'un lexique, comme c'est le cas dans Mood Messenger [2] et iMessage sous iOS 10 d'Apple, mais également de manière plus intelligente en prenant en compte la phrase entière. L'application proposée montre le résultat de travaux en cours sur la recommandation d'emojis porteurs de sentiments dans un contexte phrastique.

2 Architecture de l'application

Le système de recommandation de l'application propose une prédiction d'emojis porteurs de sentiments sélectionnés selon l'Emoji Sentiment Ranking (Kralj Novak *et al.*, 2015), un lexique d'emojis

1. Selon les emojis Unicode de mai 2017 : `http://unicode.org/emoji/charts/full-emoji-list.html`
2. `http://moodmessenger.com/`

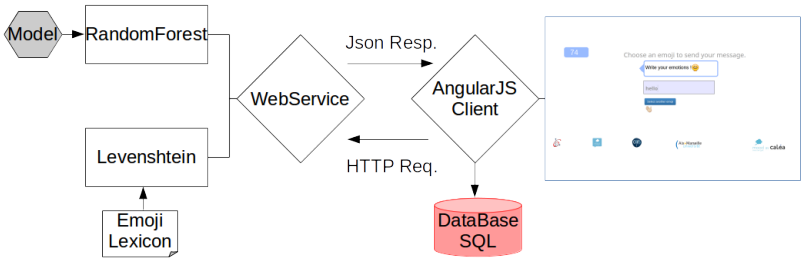


FIGURE 1 – Architecture globale de l'application

avec polarités associées issu de 83 annotateurs humains dans 13 langues. Cette recommandation se limite à la phrase en cours et non au message (Barbieri *et al.*, 2017) ni à l'ensemble de la conversation (Xie *et al.*, 2016). La séparation des phrases, le découpage en unités lexicales élémentaires et la lemmatisation sont effectués à l'aide de modèles de nltk [3].

Le système de recommandation proposé combine deux approches (Figure 1) : une prédiction automatique d'emojis par apprentissage automatique et une distance d'édition.

Prédiction automatique. La prédiction automatique d'emojis est effectuée à l'aide de *Random Forest* (Breiman, 2001) pour classification multi-étiquettes appris sur 9700 phrases contenant des emojis. L'étendue de l'apprentissage est volontairement limitée à 169 emojis porteurs de sentiments. Lors de l'apprentissage et lors de la prédiction chaque phrase est traduite en une représentation vectorielle tf-idf qui est ensuite enrichie avec des n-grammes de 1 à 5 mots, et 2 scores de polarité prédits à l'aide d'Echo(Hamdan *et al.*, 2015) et de SentiStrength(Thelwall *et al.*, 2012). Le tout est ensuite utilisé comme caractéristiques pour le classifieur (Guibon *et al.*, 2017).

Distance d'édition. La distance d'édition utilisée est celle de Levenshtein (Levenshtein, 1966). Elle est associée à un lexique propriétaire de correspondance mots-clés - emojis. Plus précisément, l'insertion et la suppression ont une valeur de 1, tandis que la substitution a une valeur de 2 correspondant aux deux opérations de suppression puis d'insertion nécessaires pour les lettres déjà tapées.

Interface. L'interface web consiste en une SPA (*Single Page Application*) effectuée en AngularJS [4] et dont le code source est disponible sur github [5]. Il s'agit d'un client qui consomme une API par requêtes HTTP pour récupérer du json. L'interface et le moteur de prédiction sont indépendants et communiquent via 2 ports. Enfin, à des fins de validation du système par l'utilisateur, l'interface est également liée à une base de données MySql qui permet d'enregistrer les phrases tapées, les emojis sélectionnés et s'il s'agit d'un emoji recommandé ou d'un autre disponible parmi ceux représentant les 7 émotions basiques d'Ekman (Ekman, 1993).

L'application est hébergée sur le site du laboratoire à l'adresse suivante : `http://lsis-mood-emoji.lsis.org/`. Quant au système de recommandation, en voici un exemple de requête : `http://lsis-mood-emoji.lsis.org:8080/?query=hello`

3. `http://www.nltk.org/`
4. `https://angularjs.org/`
5. `https://github.com/gguibon/lsis-mood-emoji`

Références

BARBIERI F., BALLESTEROS M. & SAGGION H. (2017). Are emojis predictable ? *arXiv preprint arXiv :1702.07285*.

BREIMAN L. (2001). Random forests. *Machine learning*, **45**(1), 5–32.

EKMAN P. (1993). Facial expression and emotion. *American psychologist*, **48**(4), 384.

GUIBON G., OCHS M. & BELLOT P. (2017). Prédiction automatique d'emojis sentimentaux. In *COnférence en Recherche d'Information et Applications (CORIA) 2017*.

HAMDAN H., BELLOT P. & BECHET F. (2015). Sentiment lexicon-based features for sentiment analysis in short text. In *In Proceeding of the 16th International Conference on Intelligent Text Processing and Computational Linguistics*.

KRALJ NOVAK P., SMAILOVIC J., SLUBAN B. & MOZETIC I. (2015). Sentiment of emojis. *PLOS ONE*, **10**(12).

LEVENSHTEIN V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, p. 707–710.

THELWALL M., BUCKLEY K. & PALTOGLOU G. (2012). Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, **63**(1), 163–173.

XIE R., LIU Z., YAN R. & SUN M. (2016). Neural emoji recommendation in dialogue systems. *arXiv preprint arXiv :1612.04609*.

Un outil modulaire libre pour le résumé automatique

Valentin Nyzam Aurélien Bossard
LIASD, Université Paris 8 - IUT de Montreuil, 140 rue de la Nouvelle France,
93100 Montreuil, France
valentin.nyzam@iut.univ-paris8.fr,
aurelien.bossard@iut.univ-paris8.fr

RÉSUMÉ

Nous proposons une démonstration d'un outil modulaire et évolutif de résumé automatique qui implémente trois méthodes d'extraction de phrases de l'état de l'art ainsi que sept méthodes d'évaluation des phrases. L'outil est développé en Java et est d'ores-et-déjà disponible sur la plateforme Github.

ABSTRACT

A Modular Open Source Tool for Automatic Summarization

We propose a demonstration of an evolutive and modular open source tool for automatic summarization. The tool is developed in Java, implements three sentence extraction methods and seven sentence scoring methods, and is available on Github platform.

MOTS-CLÉS : résumé automatique, open source.

KEYWORDS: automatic summarization, open source.

Le résumé automatique est une tâche explorée depuis les années 1950 (Luhn, 1958). Depuis, de nombreuses méthodes ont vu le jour, qui s'appuient sur des domaines aussi variés que l'analyse de graphes (Erkan & Radev, 2004; Mihalcea & Tarau, 2004), les modèles statistiques génératifs (Blei *et al.*, 2003), la programmation linéaire en nombres entiers (Gillick & Favre, 2009b) ou encore les algorithmes évolutionnaires (Bossard & Rodrigues, 2015). On peut être amené, pour des besoins d'évaluation ou d'expérimentation sur de nouveaux corpus, à tester l'efficacité de différentes méthodes reconnues. Cependant, l'implémentation de ces méthodes n'est pas toujours disponible. De plus, quand elle l'est, il est souvent compliqué de la comparer à d'autres, car elles mettent en œuvre des pré et post-traitements différents qui influent sur la qualité des résumés produits.

Certains systèmes existent déjà, par exemple MEAD [1], News In Essence ou encore ICSISUMM [2], mais ceux-ci implémentent uniquement la ou les méthodes de leurs auteurs. Sumy [3], un système de résumé automatique multidocument, implémente plusieurs méthodes de résumé, mais ne propose pas une approche modulaire qui permet de modifier chaque composant de la chaîne de traitement du résumé automatique. Au contraire, l'outil que nous proposons est le plus modulaire possible afin de permettre par exemple de tester l'apport de nouvelles approches du TAL en changeant la représentation des mots ou des phrases dans des méthodes de résumé déjà existantes.

L'objectif de la démonstration est de montrer l'utilisation et la modularité de ce nouvel outil open

1. http://www.summarization.com/mead/
2. https://code.google.com/archive/p/icsisumm/
3. https ://github.com/miso-belica/sumy

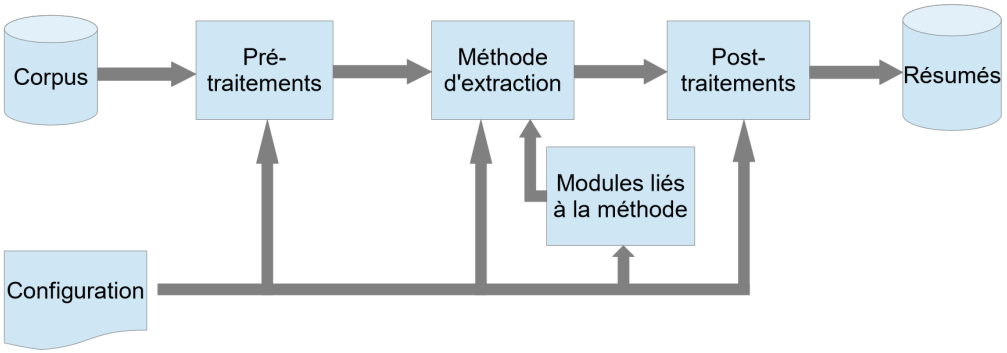


FIGURE 1 – Architecture globale de l'outil de résumé automatique

source évolutif pour le résumé automatique. Développé en Java, nous y avons implémenté des méthodes bien connues d'extraction de phrases pour le résumé automatique :

— MMR (Carbonell & Goldstein, 1998) ;
— ILP (Gillick & Favre, 2009b) ;
— Algorithme évolutionnaire (Bossard & Rodrigues, 2015).

Toutes ces méthodes d'extraction comportent des composants modulables. Par exemple, la méthode MMR nécessite des méthodes d'évaluation de la pertinence des phrases. Ainsi, nous avons implémenté les méthodes d'évaluation suivantes :

— Variantes fondées sur le tfidf ;
— Centroid (Radev *et al.*, 2004) ;
— LexRank (Erkan & Radev, 2004) ;
— LDA (Blei *et al.*, 2003) ;
— *Word embeddings* (Zhang *et al.*, 2015)

L'outil permet de choisir des pré-traitements, une méthode d'indexation, une méthode d'extraction ainsi que différentes options propres à cette méthode, et des post-traitements. Les méthodes communiquent entre elles leurs résultats à l'aide d'interfaces génériques. Cela permet ainsi d'ajouter une nouvelle méthode sans toucher au cœur de l'outil mais simplement en utilisant les interfaces proposées. L'outil dispose également d'un algorithme génétique (à bien différencier de l'algorithme évolutionnaire dédié au résumé) qui permet d'optimiser les paramètres de l'outil d'une méthode donnée sur un corpus pour lequel on dispose de résumés de référence.

Sans rentrer ici dans les détails de l'architecture de chacune de ces méthodes, l'architecture globale de l'outil est présentée en Figure 1.

Cet outil permet d'évaluer différentes méthodes dans un cadre unifié (pré et post traitements identiques) et nous espérons qu'il sera rapidement adopté par la communauté du résumé automatique afin qu'y soient implémentées de nouvelles méthodes de résumé automatique.

Un *web service* qui utilise l'outil de résumé dont nous ferons la démonstration est en cours de développement pour permettre d'obtenir rapidement, et sans avoir à l'installer, des résumés multidocuments.

1 Remerciements

Les travaux décrits dans cet article ont bénéficié du soutien de l'ANR (projet Jeunes Chercheurs/Jeunes Chercheuses ASADERA - convention ANR-16-CE38-0008)

Références

BLEI D. M., NG A. Y. & JORDAN M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, **3**, 993–1022.

BOSSARD A. & RODRIGUES C. (2015). Une approche évolutionnaire pour le résumé automatique. In *TALN 2015 - 22ème Conférence sur le Traitement Automatique des Langues Naturelles*.

CARBONELL J. & GOLDSTEIN J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR'98 : Proceedings of the 21st ACM SIGIR Conference*, p. 335–336.

ERKAN G. & RADEV D. R. (2004). Lexrank : Graph-based centrality as salience in text summarization. *Journal of Artificial Intelligence Research (JAIR)*.

GILLICK D. & FAVRE B. (2009b). A scalable global model for summarization. In *Proceedings of the Workshop on Integer Linear Programming for Natural Language Processing*, p. 10–18 : Association for Computational Linguistics.

LUHN H. (1958). The automatic creation of literature abstracts. *IBM Journal*, **2**(2), 159–165.

MIHALCEA R. & TARAU P. (2004). TextRank : Bringing order into texts. In *Proceedings of EMNLP-04and the 2004 Conference on Empirical Methods in Natural Language Processing*.

RADEV D. R., JING H., STY M. & TAM D. (2004). Centroid-based summarization of multiple documents. *Information Processing Management*, **40**, 919–938.

ZHANG Y., ER M. J. & ZHAO R. (2015). Multi-document extractive summarization using window-based sentence representation. In *2015 IEEE Symposium Series on Computational Intelligence*, p. 404–410.

Uniformisation de corpus anglais annotés en sens

Loïc Vial, Benjamin Lecouteux, Didier Schwab
GETALP – LIG – Univ. Grenoble Alpes
{loic.vial, benjamin.lecouteux, didier.schwab}
@univ-grenoble-alpes.fr

RÉSUMÉ

Pour la désambiguïsation lexicale en anglais, on compte aujourd'hui une quinzaine de corpus annotés en sens dans des formats souvent différents et provenant de différentes versions du *Princeton WordNet*. Nous présentons un format pour uniformiser ces corpus, et nous fournissons à la communauté l'ensemble des corpus annotés en anglais portés à notre connaissance avec des sens uniformisés du *Princeton WordNet* 3.0, lorsque les droits le permettent et le code source pour construire l'ensemble des corpus à partir des données originales.

ABSTRACT

Unification of sense annotated English corpora for word sense disambiguation

In word sense disambiguation, there are today about almost fifteen sense annotated English corpora, in various formats and using different versions of *Princeton WordNet*. We present a format that unifies these corpora, and we give to the community the whole set of corpora sense annotated in English that we know, with senses from *Princeton WordNet* 3.0, in this unified format, when the copyright allows it, and the source code for building these corpora from the original data.

MOTS-CLÉS : désambiguïsation lexicale, corpus annotés en sens, ressource uniformisée.

KEYWORDS: word sense disambiguation, sense annotated corpora, unified resource.

1 Introduction

Que ce soit pour l'évaluation ou l'apprentissage d'un système de désambiguïsation lexicale (DL), les corpus annotés en sens sont essentiels. En effet, les systèmes de DL exploitant les exemples issus de corpus annotés en sens sont généralement bien meilleurs que ceux qui n'en exploitent pas (Navigli *et al.*, 2007; Moro & Navigli, 2015).

En anglais, le *Princeton WordNet* (Miller, 1995) est aujourd'hui la base lexicale standard *de facto*. La plupart des corpus annotés en sens sont ainsi soit annotés directement grâce à WordNet soit annotés avec un inventaire de sens lié aux sens de WordNet comme BabelNet (Navigli & Ponzetto, 2010).

Il n'est toutefois pas aisé d'utiliser ces corpus car la plupart diffèrent grandement par leur format et par la version du *Princeton WordNet* utilisée. De plus, les systèmes sont systématiquement évalués sur les corpus destinés à l'origine à l'évaluation et jamais sur les corpus destinés à l'origine à un autre usage sans qu'il n'y ait de raison scientifique pour cela.

Nous présentons ainsi un travail d'unification de tous les corpus anglais annotés avec WordNet portés à notre connaissance, dans un format unique, simple à comprendre et rapide à utiliser en pratique. Nous mettons au même plan les corpus destinés à l'origine à l'évaluation et ceux destinés à l'apprentissage, pour faciliter la construction de systèmes de DL qui pourraient ainsi réaliser une

évaluation à plus large échelle en procédant, par exemple, à une validation croisée par rotation dans laquelle on utilise tour à tour chacun des corpus pour l'évaluation d'un système et l'ensemble des autres pour sa construction.

Nous avons aussi effectué la conversion de toutes les annotations en sens depuis leur version de WordNet d'origine à la dernière version (3.0) grâce à des tables de conversion dont la méthode de fabrication est issue de Daudé *et al.* (2000) [1].

Notre travail est proche de celui de Raganato *et al.* (2017) mais il diffère en deux points : premièrement, ils séparent les corpus en corpus d'évaluation et corpus d'apprentissage, et deuxièmement, ils n'intègrent que 7 corpus contre 12 pour nous.

Nous fournissons du code Java permettant de lire et écrire facilement ce format, ainsi que tous les corpus utilisés, à la fois dans leur format original ainsi que dans notre format. Le code qui nous a permis de faire la conversion est lui aussi fourni. Le tout est disponible à l'adresse suivante : `https://github.com/getalp/WSD-TALN2017-Corpus-Vialetal`

2 Corpus anglais annotés en sens

Notre ressource contient tous les corpus anglais annotés en sens *Princeton WordNet* à notre connaissance, c'est à dire :

- Le SemCor (Miller *et al.*, 1993), annoté originellement avec WordNet 1.6 ;
- Le DSO (Ng & Lee, 1996), annoté avec WordNet 1.5 ;
- Le corpus des définitions de WordNet [2] , annotées en sens depuis la version 3.0 ;
- L'OMSTI (Taghipour & Ng, 2015), annoté avec WordNet 3.0 ;
- Le MASC (Nancy Ide & Passonneau, 2008), annoté avec WordNet 3.0 ;
- L'Ontonotes (`https://catalog.ldc.upenn.edu/ldc2013t19`), annoté avec WordNet 3.0 ;
- Les 6 corpus des campagnes d'évaluation de DL pour l'anglais SemEval-SensEval.

3 Format de corpus unifié

Notre format de corpus est voulu pour être clair et facile à comprendre, tout en étant à la fois efficace à traiter, et contenant toutes les informations données par les différents corpus originaux.

Ainsi, nous avons opté pour un format descriptif XML, composé des 5 noeuds suivants : `corpus`, `document`, `paragraph`, `sentence` et `word`. À n'importe quel noeud, on peut y attacher un ou plusieurs attributs quelconques. Tous les attributs sont optionnels, sauf pour l'attribut `surface_form` d'un mot qui correspond à sa forme de surface. Les parties du discours sont annotés avec l'attribut `pos`, les lemmes avec `lemma`, les clés de sens pour une version spécifique de *Princeton WordNet* avec `wn{version}_key`.

L'extrait suivant est un exemple de XML résultant :

```
<corpus>
 <document id="d001" >
  <paragraph>
   <sentence id="d001.s001" >
    <word surface_form="exemple" lemma="exemple" wn30_key="example%1:09:00::" />
   </sentence>
  </paragraph>
 </document>
</corpus>
```

1. `http://www.talp.upc.edu/index.php/technology/tools/45-textual-processing-tools/98-wordnet-mappings/`
2. `http://wordnet.princeton.edu/glosstag.shtml`

Références

DAUDÉ J., PADRÓ L. & RIGAU G. (2000). Mapping wordnets using structural information. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, ACL '00, p. 504–511, Stroudsburg, PA, USA : Association for Computational Linguistics.

MILLER G. A. (1995). Wordnet : A lexical database. *ACM*, **Vol. 38**(No. 11), p. 1–41.

MILLER G. A., LEACOCK C., TENGI R. & BUNKER R. T. (1993). A semantic concordance. In *Proceedings of the workshop on Human Language Technology*, HLT '93, p. 303–308, Stroudsburg, PA, USA : Association for Computational Linguistics.

MORO A. & NAVIGLI R. (2015). Semeval-2015 task 13 : Multilingual all-words sense disambiguation and entity linking. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, p. 288–297, Denver, Colorado : Association for Computational Linguistics.

NANCY IDE, COLLIN BAKER C. F. C. F. & PASSONNEAU R. (2008). Masc : the manually annotated sub-corpus of american english. In B. M. J. M. J. O. S. P. D. T. NICOLETTA CALZOLARI (CONFERENCE CHAIR), KHALID CHOUKRI, Ed., *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco : European Language Resources Association (ELRA). http ://www.lrec-conf.org/proceedings/lrec2008/.

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.

NAVIGLI R. & PONZETTO S. P. (2010). Babelnet : Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, p. 216–225 : Association for Computational Linguistics.

NG H. T. & LEE H. B. (1996). Integrating multiple knowledge sources to disambiguate word sense : an exemplar-based approach. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, ACL '96, p. 40–47, Stroudsburg, PA, USA : Association for Computational Linguistics.

RAGANATO A., CAMACHO-COLLADOS J. & NAVIGLI R. (2017). Word sense disambiguation : A unified evaluation framework and empirical comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics : Volume 1, Long Papers*, p. 99–110, Valencia, Spain : Association for Computational Linguistics.

TAGHIPOUR K. & NG H. T. (2015). One million sense-tagged instances for word sense disambiguation and induction. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, p. 338–344, Beijing, China : Association for Computational Linguistics.

Résumer automatiquement en ligne : démonstration d'un service web de résumé multidocument

Valentin Nyzam Nathan Gatto Aurélien Bossard
LIASD, Université Paris 8 - IUT de Montreuil, 140 rue de la Nouvelle France,
93100 Montreuil, France
valentin.nyzam@iut.univ-paris8.fr,
nathan.gatto@free.fr,
aurelien.bossard@iut.univ-paris8.fr

RÉSUMÉ

Nous proposons une démonstration d'un webservice de résumé automatique multidocument. Ce webservice s'appuie sur un outil ouvert qui implémente plusieurs algorithmes reconnus de résumé automatique, et permet de résumer des documents en utilisant des configurations différentes.

ABSTRACT

Summarizing Automatically Online :

We propose a demonstration of an automatic multidocument summarization web service. This web service relies on an open source summarization tool that implements several known summarization algorithms, and allows to summarize documents using different configurations.

MOTS-CLÉS : résumé automatique, service web.

KEYWORDS: automatic summarization, web service.

1 Introduction

Le résumé automatique multidocument par extraction, comme beaucoup d'autres tâches du traitement automatique du langage, nécessite au-delà des algorithmes d'extraction des phrases, des ressources externes. Les systèmes développés par les chercheurs de la communauté s'appuient parfois sur des librairies externes.

Ainsi, effectuer des expérimentations sur des corpus en utilisant des algorithmes reconnus peut se révéler fastidieux. Un service web de résumé automatique proposant des méthodes de l'état de l'art peut être une solution efficace à ce problème. Aujourd'hui, quelques services web existent en ligne. Certains sont payants (par exemple `https://resoomer.com`), d'autres gratuits (SweSum [1], Text Compactor [2], ou encore Open Text Summarizer [3]) mais ceux-ci sont parfois datés, voire n'annoncent pas les méthodes qu'ils utilisent.

1. http ://swesum.nada.kth.se/index-eng.html
2. http ://textcompactor.com/
3. https ://www.splitbrain.org/services/ots

2 Présentation du service web

Nous présentons un service web de résumé automatique qui s'appuie sur un outil libre qui implémente plusieurs méthodes reconnues de résumé automatique multidocument. Ce service web permet d'une part d'envoyer le corpus que l'on souhaite résumer, d'autre part de choisir la configuration complète de lancement du résumé automatique. Le service web donne également au client le choix entre plusieurs méthodes, et permet pour chacune des méthodes, de choisir des options propres à celle-ci.

A titre d'exemple, les méthodes ou algorithmes suivants peuvent être lancés par le service web :

— MMR (Carbonell & Goldstein, 1998) ;
— ICSISUMM : résumé par programmation linéaire en nombres entiers (Gillick & Favre, 2009b) ;
— Algorithme évolutionnaire (Bossard & Rodrigues, 2015).
— Centroid (Radev *et al.*, 2004) ;
— LexRank (Erkan & Radev, 2004) ;
— LDA (Blei *et al.*, 2003) ;
— LSA ;
— *Word embeddings* (Zhang *et al.*, 2015)

Ainsi, la constitution du corpus à envoyer ainsi que la configuration pas à pas nécessitent une prise en main et n'est sûrement pas optimale malgré nos efforts.

3 Objectifs de la démonstration

La démonstration a pour objectif de présenter à la communauté notre service web. Nous proposerons également un tutoriel pour apprendre à constituer les corpus à envoyer et pour comprendre les configurations de lancement de l'outil proposées par le service web.

4 Remerciements

Les travaux décrits dans cet article ont bénéficié du soutien de l'ANR (projet Jeunes Chercheurs/Jeunes Chercheuses ASADERA - convention ANR-16-CE38-0008)

Références

BLEI D. M., NG A. Y. & JORDAN M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, **3**, 993–1022.

BOSSARD A. & RODRIGUES C. (2015). Une approche évolutionnaire pour le résumé automatique. In *TALN 2015 - 22ème Conférence sur le Traitement Automatique des Langues Naturelles*.

CARBONELL J. & GOLDSTEIN J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR'98 : Proceedings of the 21st ACM SIGIR Conference*, p. 335–336.

ERKAN G. & RADEV D. R. (2004). Lexrank : Graph-based centrality as salience in text summarization. *Journal of Artificial Intelligence Research (JAIR)*.

GILLICK D. & FAVRE B. (2009b). A scalable global model for summarization. In *Proceedings of the Workshop on Integer Linear Programming for Natural Language Processing*, p. 10–18 : Association for Computational Linguistics.

RADEV D. R., JING H., STY M. & TAM D. (2004). Centroid-based summarization of multiple documents. *Information Processing Management*, **40**, 919–938.

ZHANG Y., ER M. J. & ZHAO R. (2015). Multi-document extractive summarization using window-based sentence representation. In *2015 IEEE Symposium Series on Computational Intelligence*, p. 404–410.

Traitement de la langue biomédicale au LIMSI

Christopher Norman[1] Cyril Grouin[1] Thomas Lavergne[2]
Aurélie Névéol[1] Pierre Zweigenbaum[1]
(1) LIMSI, CNRS, Université Paris-Saclay, 91405 Orsay, France
(2) LIMSI, CNRS, Univ. Paris-Sud, Université Paris-Saclay, 91405 Orsay, France
prenom.nom@limsi.fr

RÉSUMÉ

Nous proposons des démonstrations de trois outils développés par le LIMSI en traitement automatique des langues appliqué au domaine biomédical : la détection de concepts médicaux dans des textes courts, la catégorisation d'articles scientifiques pour l'assistance à l'écriture de revues systématiques, et l'anonymisation de textes cliniques.

ABSTRACT

Biomedical language processing at LIMSI

We propose demonstrations of three natural language processing tools developed at LIMSI for applications in the biomedical domain : medical concept detection in short texts, scientific paper categorization to assist systematic review authors, clinical text de-identification.

MOTS-CLÉS : domaine médical ; classification de textes ; extraction d'information ; désidentification ; revues systématiques.

KEYWORDS: medical domain; text classification; information extraction; de-identification; systematic reviews.

1 Détection de concepts dans des textes courts

Nous présentons un système qui effectue de la détection de concepts dans des textes courts (Zweigenbaum & Lavergne, 2017). Il est spécialisé dans la détermination de codes diagnostiques de la Classification internationale des maladies (CIM-10) de l'OMS pour des certificats de décès. Il combine deux techniques classiques : l'application d'un dictionnaire spécialisé et un apprentissage supervisé, entraîné sur plus de 300 000 exemples de diagnostics. La démonstration montre la détection de diagnostics dans des textes courts, ainsi que l'apport comparé de différentes méthodes d'hybridation entre dictionnaire et apprentissage supervisé :

— le calibrage du dictionnaire par apprentissage supervisé sur le corpus d'entraînement ;
— la prise en compte du dictionnaire comme attributs dans la classification ;
— la fusion des résultats du dictionnaire et de l'apprentissage supervisé.

Le système est démontré en français et en anglais. Sur les données de la campagne d'évaluation CLEF eHealth 2017 [1], la version française produit actuellement des résultats meilleurs que le meilleur système participant et la version anglaise est juste derrière les résultats du meilleur système participant.

1. https://sites.google.com/site/clefehealth2017/task-1

2 Assistance à l'écriture de revues systématiques

Les revues systématiques de la littérature dans le domaine biomédical reposent essentiellement sur le travail bibliographique manuel d'experts. Nous présentons un système de classification automatique d'articles présélectionnés à l'aide d'une requête soumise à un moteur de recherche (Norman *et al.*, 2017). Le système propose un ordonnancement des articles par ordre de pertinence par rapport aux critères d'inclusion définis par un corpus d'entraînement. La mise en œuvre ainsi que les résultats de cet outil de classification peuvent être exploités au travers de diverses interfaces graphiques. À titre d'exemple, nous présentons l'interface BibReview, utilisée dans le cadre du Yearbook of Medical Informatics (Névéol & Zweigenbaum, 2016) qui rapporte chaque année les résultats de revues de la littérature dans seize sous-domaines de l'informatique biomédicale. Cet outil permet de visualiser les articles de la préselection dans l'ordre proposé par le classifieur, de valider et de visualiser la sélection d'articles pour inclusion dans une revue de la littérature.

3 Désidentification de comptes rendus cliniques

MEDINA (MEDical INformation Anonymization) (Grouin, 2013) est un outil permettant de désidentifier (anonymiser) les informations potentiellement identifiantes contenues dans des comptes rendus cliniques. L'outil identifie les informations par type, puis remplace les informations identifiées par des pseudonymes (noms, prénoms, adresses, téléphones, etc.) et par un décalage des dates dans le passé (permettant de conserver les écarts temporels entre dates). Le résultat produit permet de bénéficier de textes représentatifs du domaine clinique sur lesquels réaliser des études, sans que la vie privée des patients ne soit mise en cause.

La démonstration proposée consiste à présenter les différentes étapes du système à base de règles et de lexiques pour anonymiser des documents cliniques.

Références

GROUIN C. (2013). *Anonymisation de documents cliniques : performances et limites des méthodes symboliques et par apprentissage statistique*. Thèse de doctorat, Université Pierre et Marie Curie, Paris, France.

GROUIN C., DELÉGER L., MINARD A.-L., LIGOZAT A.-L., BEN ABACHA A., BERNHARD D., CARTONI B., GRAU B., ROSSET S. & ZWEIGENBAUM P. (2011). Extraction d'informations médicales au LIMSI. In *Démonstrations, TALN 2011*, Montpellier.

NORMAN C., LEEFLANG M., ZWEIGENBAUM P. & NÉVÉOL A. (2017). Tri automatique de la littérature pour les revues systématiques. In *TALN 2017*, Orléans, France : Association pour le Traitement Automatique des Langues.

NÉVÉOL A. & ZWEIGENBAUM P. (2016). Clinical natural language processing in 2015 : Leveraging the variety of texts of clinical interest. *Yearb Med Inform*, p. 234–239.

ZWEIGENBAUM P. & LAVERGNE T. (2017). Détection de concepts et granularité de l'annotation. In *TALN 2017*, Orléans, France : Association pour le Traitement Automatique des Langues.

Proxem Studio : la plate-forme d'analyse sémantique qui transforme l'utilisateur métier en *text scientist*

François-Régis Chaumartin [1]
(1) Proxem SAS, 105 rue La Fayette, 75010 Paris, France
frc@proxem.com

RESUME

Proxem édite depuis 2011 une plate-forme d'analyse sémantique multilingue utilisé en entreprise pour de multiples usages : relation clients, ressources humaines, veille stratégique… La version la plus récente du logiciel, lancée en mars 2017, lève le principal goulet d'étranglement des outils classiques de text mining : un utilisateur métier devient enfin autonome pour définir lui-même les ressources linguistiques nécessaires à l'analyse sémantique d'un corpus donné. Une fois le corpus chargé, la plate-forme en extrait une terminologie et organise les termes en regroupements hiérarchisés de proto-concepts ; l'utilisateur n'a plus qu'à valider ces concepts au niveau de granularité qui lui semble pertinent pour constituer un extracteur d'entités nommées de granularité fine, adapté au corpus à traiter, avec un rappel élevé grâce à l'identification automatique des quasi-synonymes. La plate-forme détecte aussi dans ces termes les homonymes potentiels et propose à l'utilisateur des contextes de désambiguïsation, fournissant ainsi une bonne précision.

ABSTRACT

ProxemStudio: the semantic analysis platform that turns the business user into a text scientist. Proxem publishes since 2011 a multilingual semantic analysis platform used by companies for multiple use cases: customer feedback management, human resources, business intelligence. The latest version, launched in March 2017, removes the major bottleneck of conventional text mining tools: a business user finally becomes autonomous to define himself or herself the linguistic resources necessary for the semantic analysis of a given corpus. Once the corpus is uploaded, the platform extracts a terminology and organizes the terms into hierarchical clusters of proto-concepts. The user only has to validate these concepts at the granularity level that seems pertinent to constitute a fine-grained named entities recognizer, perfectly adapted to the corpus, with a high recall thanks to the automatic identification of synonyms. The platform also detects potential homonyms among these terms and provides the user with disambiguation contexts, thus providing a good precision.

MOTS-CLES : entités nommées, catégorisation, désambiguïsation, apprentissage profond.
KEYWORDS: named entities, categorization, disambiguation, deep learning.

1 Evolutions de la plate-forme Proxem en 10 ans

Développée à partir de 2007, la plate-forme Antelope (Chaumartin, 2012) avait pour objectif d'accélérer le travail de l'infolinguiste en lui proposant un cadre de travail simplifié, fourni avec des modules d'analyse prêts à l'emploi. Des algorithmes d'apprentissage ont été intégrés dès 2009 pour la reconnaissance des entités nommées (par CRF), puis en 2010 pour la classification automatique.

Ils ont été ensuite enrichi par un module de catégorisation générique (Chaumartin, 2013) permettant d'associer finement à un texte tout-venant écrit dans une langue donnée, un graphe de catégories de la Wikipédia dans cette langue. D'autres évolutions ont permis de traiter des corpus multilingues.

A partir de 2013, Proxem a commencé à s'approprier les techniques à base de réseaux de neurones, et a généralisé l'approche de type *word embedding* à l'apprentissage simultané de plusieurs langues (Coulmance *et al.*, 2015).

Une réflexion ergonomique a été menée courant 2016 pour déterminer comment proposer à un utilisateur métier une application web simple d'utilisation, intégrant tous ces modules. Cette application, le Proxem Studio, a été lancée en mars 2017. Elle permet à présent à un utilisateur qui n'a pas de compétences particulières en linguistique d'être autonome pour réaliser, de bout en bout, un projet comportant des tâches de web mining, text mining et data mining. Cette application permet de transformer facilement une source web ou un corpus d'entreprise en un nouveau canal de données originales, de procéder à une visualisation riche des données extraites du texte, et d'en tirer une connaissance permettant de prendre des meilleures décisions sur les enjeux métiers.

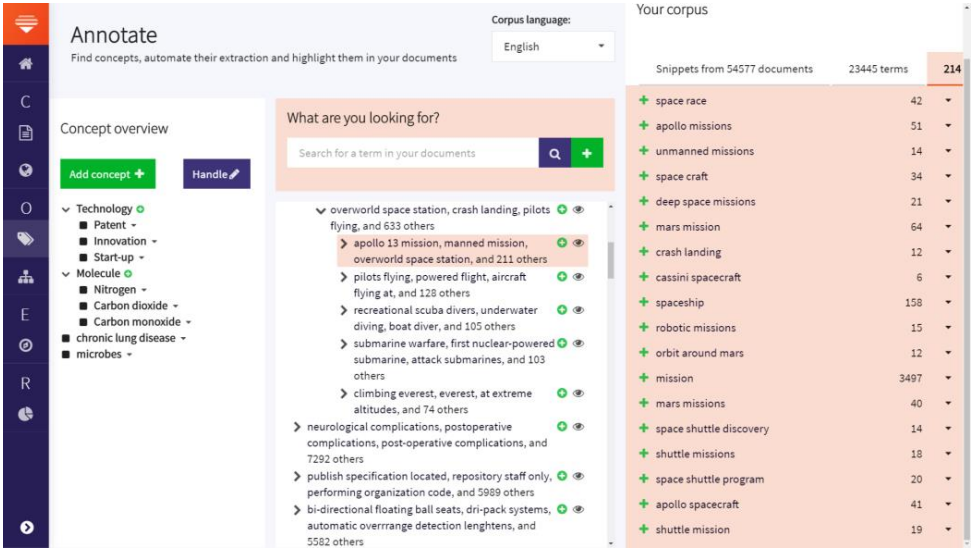


FIGURE 1: Module qui permet à l'utilisateur métier de définir les entités nommées à extraire.

Références

CHAUMARTIN F.-R. (2012). Antelope, une plate-forme de TAL permettant d'extraire les sens du texte : théorie et applications de l'ISS. Thèse de doctorat, Université Paris Diderot.

CHAUMARTIN F.-R. (2013). Apprentissage d'une classification thématique générique et cross-langue à partir des catégories de la Wikipédia. Actes de TALN, 659–666.

COULMANCE J., MARTY J.-M., WENZEK G., BENHALLOUM A. (2015). Trans-gram, Fast Cross-lingual Word-embeddings. Actes de EMNLP.

Translittération automatique pour une paire de langues peu dotée

Ngoc Tan Le[1] Fatiha Sadat[1] Lucie Ménard[2]
(1) Université du Québec à Montréal, 201 avenue du Président-Kennedy, H2X 3Y7, Montréal, Canada
(2) UQÀM, Laboratoire de phonétique, 320 Sainte-Catherine Est, H2X 1L7, Montréal, Canada
le.ngoc_tan@courrier.uqam.ca, sadat.fatiha@uqam.ca, menard.lucie@uqam.ca

RÉSUMÉ
La translittération convertit phonétiquement les mots dans une langue source (*i.e. français*) en mots équivalents dans une langue cible (*i.e. vietnamien*). Cette conversion nécessite un nombre considérable de règles définies par les experts linguistes pour déterminer comment les phonèmes sont alignés ainsi que prendre en compte le système de phonologie de la langue cible. La problématique pour les paires de langues peu dotées lie à la pénurie des ressources linguistiques. Dans ce travail de recherche, nous présentons une démonstration de conversion de graphème en phonème pour pallier au problème de translittération pour une paire de langues peu dotée, avec une application sur français-vietnamien. Notre système nécessite un petit corpus d'apprentissage phonétique bilingue. Nous avons obtenu des résultats prometteurs, avec un gain de +4,40% de score BLEU, par rapport au système de base utilisant l'approche de traduction automatique statistique.

MOTS-CLÉS : Translittération, graphème, phonème, traduction automatique, langue peu dotée, français-vietnamien.

KEYWORDS: Transliteration, grapheme, phoneme, machine translation, under-resourced language, French-Vietnamese.

1 Introduction

La translittération consiste en un processus de transformation d'un mot d'un système d'écriture (appelé mot source) vers un mot, phonétiquement équivalent, d'un autre système d'écriture (appelé mot cible) (Knight & Graehl, 1998). Beaucoup d'entités nommées (*i.e. les noms de personne, de location, d'organisation, les termes techniques, etc.*) sont souvent translittérées d'une langue source vers une langue cible quand la traduction est difficile ou impossible. La translittération peut être considérée comme une sous-tâche de la traduction automatique (TA).

Par ailleurs, avec l'évolution de hautes technologies, les gens ont tendance à inventer de nouveaux mots. Il est très difficile de définir toutes les règles possibles de conversion phonétique entre la langue source et la langue cible. Nous nous intéressons à résoudre les mots hors vocabulaire (MHV) considérés comme noms propres ou termes techniques issus d'un système de traduction automatique (TA) pour une paire de langues peu dotée, français-vietnamien.

2 Approche proposée

Notre approche se déroule en trois étapes principales : *(1) prétraitement, (2) classification et (3) re-classement de la liste des k-meilleurs résultats*. Tout le processus est illustré dans la Figure 2.

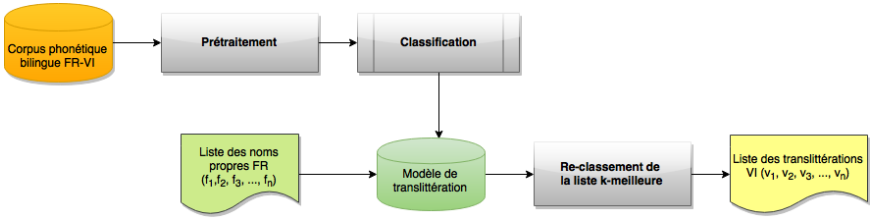


FIGURE 1 – Architecture de translittération des entités nommées bilingues pour une paire de langues peu dotée

3 Expérimentation

Nous utilisons un corpus phonétique bilingue qui a été collecté depuis les sites Web d'actualités comme présentés dans (Cao *et al.*, 2010). Ce corpus d'apprentissage possède 4 259 paires de noms propres bilingues français-vietnamien. Le corpus phonétique bilingue d'apprentissage est découpé en deux ensembles d'apprentissage et de test avec un ratio de 90% et 10% respectivement. Pour évaluer notre approche proposée, nous implémentons trois systèmes, notamment le système de base (*pbSMT sans distorsion*), le système 1 (*pbSMT avec distorsion*) et le système 2 (*notre approche proposée*) (Table 1).

Métrique	**Système**	**Moyenne**	$\overline{s}_{sel}$	s_{Test}	p-**valeur**
BLEU ↑	Système de base (*pbSMT sans distorsion*)	61,3	1,7	-	-
	Système 1 (*pbSMT avec distorsion*)	61,6	1,7	-	0,79
	Système 2 (*Notre approche*)	**65,7**	1,5	-	0,01
TER ↓	Système de base (*pbSMT sans distorsion*)	24,8	1,2	-	-
	Système 1 (*pbSMT avec distorsion*)	24,5	1,2	-	0,13
	Système 2 (*Notre approche*)	**20,5**	1,0	-	0,00
PER ↓	Système de base (*pbSMT sans distorsion*)	4,42	-	-	-
	Système 1 (*pbSMT avec distorsion*)	4.05	-	-	-
	Système 2 (*Notre approche*)	**3,80**	-	-	-

TABLE 1 – Évaluation des scores pour tous les systèmes : **BLEU**, **TER** et **PER**.
p-**valeurs** sont relatives au système de base et indiquent si une différence de cette magnitude entre le système de base et le système à comparer. $\overline{s}_{sel}$ indique la variance due à la sélection du test.

4 Conclusion et perspective

Dans cet article, nous avons présenté une méthode originale supervisée pour pallier au problème de translittération pour une paire de langues peu dotée, avec une application sur la paire de langues français-vietnamien. Nous avons obtenu des résultats prometteurs, avec un gain de +4,40% de score BLEU, en comparant notre approche au système de base. Ce système de translittération des entités nommées bilingues possède la capacité d'apprendre, de manière automatique, les régularités linguistiques à partir du corpus phonétique bilingue.

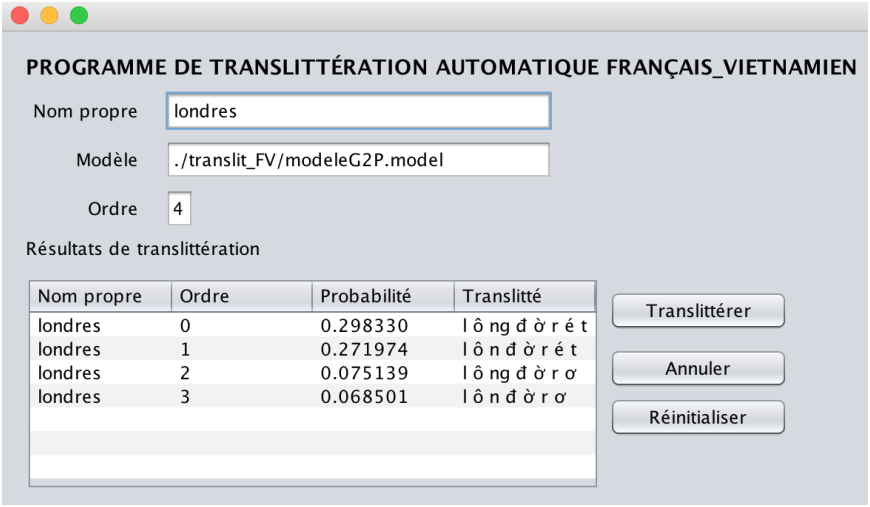


FIGURE 2 – Interface de translittération des entités nommées bilingues pour une paire de langues peu dotée

Références

BISANI M. & NEY H. (2008). Joint-sequence models for grapheme-to-phoneme conversion. *Speech communication*, **50**(5), 434–451.

CAO N. X., PHAM N. M. & VU Q. H. (2010). Comparative analysis of transliteration techniques based on statistical machine translation and joint-sequence model. In *Proceedings of the 2010 Symposium on Information and Communication Technology*, p. 59–63 : Association for Computing Machinery.

KNIGHT K. & GRAEHL J. (1998). Machine transliteration. *Computational Linguistics*, **24**(4), 599–612.

KOEHN P., HOANG H., BIRCH A., CALLISON-BURCH C., FEDERICO M., BERTOLDI N., COWAN B., SHEN W., MORAN C., ZENS R. *et al.* (2007). Moses : Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, p. 177–180 : Association for Computational Linguistics.

NGO H. G., CHEN N. F., NGUYEN B. M., MA B. & LI H. (2015). Phonology-augmented statistical transliteration for low-resource languages. In *Interspeech*, p. 3670–3674.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). Bleu : a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, p. 311–318 : Association for Computational Linguistics.

THU Y. K., PA W. P., SAGISAKA Y. & IWAHASHI N. (2016). Comparison of grapheme–to–phoneme conversion methods on a myanmar pronunciation dictionary. *Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing 2016*, p. 11–22.

Motor, un outil de segmentation accessible en ligne

Guillaume de Malézieux[1] Jennifer Lewis-Wong[1,2] Vincent Berment[1, 3]

(1) Équipe de Recherche Textes, Informatique, Multilinguisme (ERTIM - EA 2520), INALCO, 2 rue de Lille, 75343 Paris Cedex 07, France
(2) Langues et Civilisations à Tradition Orale - CNRS / Paris III / INALCO (LACITO - UMR 7107), 7 rue Guy Môquet (bât. D), 94801 Villejuif Cedex, France
(3) GETALP, LIG-campus, CS 40700 - 38058 Grenoble cedex 9, France

guillaume.de-malezieux@protonmail.com, jennifer.wong@inalco.fr, vincent.berment@inalco.fr

RÉSUMÉ

Dans cette démonstration, nous montrons le fonctionnement des segmenteurs disponibles en ligne pour diverses langues (birman, khmer, lao, thaï et tibétain) et réalisés avec l'outil Motor.

ABSTRACT

Motor, a segmentation tool accessible online.
In this demonstration, we present the use of segmenters available online for several languages (Burmese, Khmer, Lao, Thai, Tibetan), and developed with a tool called Motor.

MOTS-CLÉS : Segmentation, tokenization, langues peu dotées.

KEYWORDS: Segmentation, tokenization, under-resourced languages.

La plupart des langues d'Asie utilisent des systèmes d'écriture dits non segmentés, car ne séparant pas les mots par des espaces. La segmentation en mots est une question centrale pour ces langues, et la performance des segmenteurs est déterminante pour tous les traitements avals. Dans cet article, nous présentons les segmenteurs réalisés avec Motor, qui est un outil permettant de développer des segmenteurs à partir d'une liste des mots de la langue. L'algorithme mis en œuvre dans Motor est un algorithme de « plus petit nombre de mots ». Il a été utilisé pour plusieurs langues d'Asie du Sud-Est : birman, khmer, lao, thaï (siamois) et tibétain (avec XXX et son équipe à XXX), ainsi que pour le japonais. L'état courant de ces segmenteurs est sur www.lingwarium.org/motor/Segmentation.

Le développement des segmenteurs à partir de Motor se fait a partir d'une liste de mots fournie a Motor sous forme d'une table dans une base de données sqlite3. Cette base est dotée d'une unique table avec une colonne ≪ Cle ≫ (simple numéro d'ordre) et une colonne « Article » pour les mots. Pour optimiser la vitesse de segmentation, un index doit être mis sur la colonne des mots.

L'environnement de développement est constitué de l'interface de test de la page publique citée plus haut, et d'une zone permettant de mettre à jour la base de données (voir figures 1 et 2).

Les segmenteurs réalisés ont été utilisés comme première étape dans des systèmes de traduction automatique (cf. projet « Petit Prince » : lingwarium.org/heloise/index.php?Ref=&ws=LittlePrince). Dans ce cadre, des informations ont été ajoutées à la base lexicale, la même base ayant ainsi pu produire la segmentation, mais aussi l'analyse morphologique et le transfert lexical.

Motor, comme les autres outils disponibles sur lingwarium.org, est disponible en ligne, de manière à permettre des développements collaboratifs. Un service d'API est aussi disponible pour permettre l'utilisation des segmenteurs par des sites ou outils externes (appels en cURL...).

Figure 1: Interface de test

Figure 2: Mise à jour de la base de données