6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition) (JEP-TALN-RÉCITAL) [1]

6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition).

Communications des apprenti-e-s chercheur-euse-s 2020

Christophe Benzitoun, Laurine Huber (Éds.)

Nancy, France, 08-19 juin 2020

1. https://jep-taln2020.loria.fr/

Avec le soutien de

Message des présidents de l'AFCP et de l'ATALA

En ce printemps 2020, et les circonstances exceptionnelles qui l'accompagnent, c'est avec une émotion toute particulière que nous vous convions à la 6e édition conjointe des Journées d'Études sur la Parole (JEP), de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) et des Rencontres des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL). Après une première édition commune en 2002 (à Nancy, déjà !), et une expérience renouvelée avec succès en 2004, c'est désormais tous les quatre ans (Avignon 2008, Grenoble 2012) que se répète cet événement commun, attendu de pied ferme par les membres des deux communautés scientifiques voisines.

Cette édition 2020 est exceptionnelle, puisque dans le cadre des mesures sanitaires liées à la pandémie mondiale de COVID-19 (confinement strict, puis déconfinement progressif), la conférence ne peut avoir lieu à Nancy comme initialement prévu, mais se déroule à distance, sous forme virtuelle, soutenue par les technologies de l'information et de la communication. Nous remercions ici chaleureusement les organisateurs, Christophe Benzitoun, Chloé Braud, Laurine Huber, David Langlois, Slim Ouni et Sylvain Pogodalla, qui ont dû faire preuve de souplesse, d'inventivité, de détermination, de puissance de travail, et de tant d'autres qualités encore, afin de maintenir la conférence dans ces circonstances, en proposant un format inédit. Grâce aux différentes solutions mises en œuvre dans un délai court, la publication des communications scientifiques est assurée, structurée, et les échanges scientifiques sont favorisés, même à distance.

Bien entendu, nous regrettons tous que cette réunion JEP-TALN-RECITAL ne permette pas, comme ses prédécesseurs, de nouer ou renforcer les liens sociaux entre les différents membres de nos communautés respectives – chercheurs, jeunes et moins jeunes, académiques et industriels, professionnels et étudiants – autour d'une passionnante discussion scientifique ou d'un mémorable événement social... Notre conviction est qu'il est indispensable de maintenir à l'avenir de tels lieux d'échanges dans le domaine francophone, afin bien sûr de permettre aux jeunes diplômés de venir présenter leurs travaux et poser leurs questions sans la barrière de la langue, mais aussi de dynamiser nos communautés, de renforcer les échanges et les collaborations, et d'ouvrir la discussion autour des enjeux d'avenir, qui questionnent plus que jamais la place de la science et des scientifiques dans notre société.

Lors de la précédente édition, nous nous interrogions sur les phénomènes et tendances liés à l'apprentissage profond et sur leurs impacts sur les domaines de la Parole et du TAL. Force est de constater que l'engouement pour ces approches dans nos domaines a permis un retour sur le devant de la scène des domaines liés à l'Intelligence Artificielle, animant parfois un débat tant philosophique que technique sur la place de la machine dans la société, notamment à travers le questionnement sur la vie privée de l'utilisateur. Ces questionnements impactent tant la Parole que le TAL, d'une part sur la place de la gestion des données, d'autre part sur les modèles eux-mêmes. Malgré ces questionnements, nous constatons que les acquis et les expertises perdurent, et les nouvelles approches liées à l'apprentissage profond ont permis un rapprochement des domaines de la Parole et du TAL, sans les dénaturer, à la manière des conférences JEP-TALN-RECITAL qui créent un espace plus grand d'échange et d'enrichissement réciproques.

Nous terminons ces quelques mots d'ouverture en remerciant l'ensemble des personnes qui ont rendu possible cet événement qui restera, nous l'espérons, riche et passionnant, malgré les circonstances. L'ATALA et l'AFCP tiennent tout d'abord à réitérer leurs remerciements aux organisateurs des JEP, de TALN et de RECITAL, qui sont parvenus à maintenir le cap à travers vents et marées. Nos remerciements vont également à l'ensemble des membres des comités de programme, dont le travail et l'implication ont permis de garantir la qualité et la cohérence du programme finalement retenu. Un grand merci aux relecteurs pour le temps et le soin qu'ils ont dédiés à ce travail anonyme et indispensable. Ils se reflètent dans la qualité des soumissions que chacun pourra découvrir sur le site de la conférence.

En conclusion, cette 6e édition conjointe JEP-TALN-RECITAL est exceptionnelle parce qu'elle se tient dans un contexte de crise généralisée — crise sanitaire, économique, voire sociale et politique. Mais nous

formons le vœu qu'elle reste également dans les annales pour la qualité des échanges scientifiques qu'elle aura suscités, et pour le message envoyé à nos communautés scientifiques et à la société dans son ensemble, un message de détermination et de confiance en l'avenir, où la science et les nouvelles technologies restent au service de l'humain.

Véronique Delvaux, présidente de l'Association Francophone de la Communication Parlée
Christophe Servan, président de l'Association pour le Traitement Automatique des LAngues

Préface

En 2002, l'AFCP (Association Francophone pour la Communication Parlée) et l'ATALA (Association pour le Traitement Automatique des Langues) organisèrent conjointement leurs principales conférences afin de réunir en un seul lieu, à Nancy, les communautés du traitement automatique et de la description des langues écrites, parlées et signées.
En 2020, la sixième conférence commune revient à Nancy, après Fès (2004), Avignon (2008), Grenoble (2012) et Paris (2016). Elle est organisée par le LORIA (Laboratoire lorrain de recherche en informatique et ses applications, UMR 7503), l'ATILF (Analyse et traitement informatique de la langue française, UMR 7118) et l'INIST (Institut de l'information scientifique et technique) et regroupe :

— les 33es Journées d'Études sur la Parole (JEP),
— la 27e conférence sur le Traitement Automatique des Langues Naturelles (TALN),
— la 22e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL).

Les circonstances particulières liées à l'épidémie de Covid-19 en France et dans le monde ont conduit à une virtualisation de la conférence. Ainsi, malgré un rassemblement physique qui n'a pu avoir lieu, diffusions, présentations (au gré des auteurs) et discussions des articles acceptés ont lieu sur le site internet de la conférence. Les tutoriels, certains ateliers, et le salon de l'innovation qui accompagnent la conférence ont cependant dû être annulés, mais les ateliers suivants sont maintenus :

— Défi Fouille de Textes (DEFT 2020),
— Éthique et TRaitemeNt Automatique des Langues (ÉTeRNAL).

La conférence accueille également des conférencières et conférenciers invités dont les exposés sont diffusés sur le site : Dirk Hovy (université de Bocconi, Milan, Italie, invité ÉTeRNAL) ainsi que Marie-Jean Meurs (Université du Québec à Montréal, UQAM, Canada) et Hugo Cyr (Faculté de science politique et droit à l'Université du Québec à Montréaln UQAM, Canada). En raison des circonstance particulières, un exposé conjoint de Christine Meunier (Laboratoire Parole et Langage LPL, CNRS, Aix-en-Provence, France) et Christophe Stécoli (police technique et scientifique française) a dû être annulé et reporté à une journée spéciale en septembre 2020.
Ces actes regroupent les articles des conférences JEP (volume 1), TALN (volume 2), RÉCITAL (volume 3), les articles décrivant les démonstrations (volume 4), et les articles des ateliers DEFT (volume 5) et ÉTeRNAL (volume 6). Pour la première fois, un appel spécifique à résumés en français d'articles parus dans une sélection de conférences internationales en 2019 était également proposé (volume 4). Un appel spécifique apprenti·e·s chercheur·euse·s destiné aux étudiants de licence, de master, ou en première année de thèse a également été proposé, pour leur proposer des présentations courtes ou sous forme de poster de leurs projets.
Pour les JEP, 87 articles ont été soumis, parmi lesquels 74 ont été sélectionnés, soit un taux de sélection de 85%.
Pour TALN, 58 articles ont été soumis, parmi lesquels 37 ont été sélectionnés, soit un taux de sélection de 63%, dont 10 comme article longs (17% des soumissions) et 27 comme article courts dont 20 en présentation orale (34% des soumissions) et 7 en présentation poster (12% des soumissions).
Pour RÉCITAL, 22 articles ont été soumis, parmi lesquels 16 ont été sélectionnés, soit un taux de sélection de 73%.
Nous souhaitons vivement remercier toutes les personnes qui ont participé à ce travail de relecture et de sélection :

— l'ensemble des relecteurs (voir page **??**),
— le comité de programme des JEP (voir page **??**),
— le comité de programme de TALN (voir page **??**),
— le comité de programme de RÉCITAL (voir page **??**).

Nous souhaitons également remercier nos sociétés savantes : l'AFCP, assurant la continuité des éditions successives des JEP, et l'ATALA, dont le CPerm (comité permanent) assure la continuité des éditions

successives de TALN.
Nous remercions le comité d'organisation et les nombreuses personnes qui ont assuré le soutien administratif et technique pour que cette conférence se déroule dans les meilleures conditions, et en particulier Yannick Parmentier pour son travail pour la diffusion de ces actes sur HAL et les différents sites d'archives ouvertes (anthologie ACL et talnarchives.atala.org/).
Nous remercions enfin tous les partenaires institutionnels et industriels qui nous ont fait confiance, en particulier l'université de Lorraine, le CNRS, l'Inria, le LORIA, l'ATILF, l'INIST, le master TAL de l'Institut des Sciences du Digital Management & Cognition (IDMC), le projet OLKI de l'initiative Lorraine Université d'Excellence (LUE), la Région Grand Est, *The Evaluations and Language resources Distribution Agency* (ELDA), le projet ANR PARSEME-FR, la délégation générale à la langue française et aux langues de France (DGLFLF), l'Association des Professionnels des Industries de la Langue (APIL) et les entreprises Synapse, Yseop et Orange.
Bonne conférence à toutes et à tous !

Les présidentes et présidents JEP : David Langlois et Slim Ouni
TALN : Chloé Braud et Sylvain Pogodalla
RÉCITAL : Christophe Benzitoun et Laurine Huber

Table des matières

6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Évaluation des annotations par des mesures d'accord inter-annotateurs

Anaëlle Baledent
Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, 14000 Caen, France
anaelle.baledent@unicaen.fr

RÉSUMÉ

Nous présentons dans ce descriptif notre sujet de thèse portant sur l'évaluation des annotations par des mesures d'accord inter-annotateurs. Ces mesures permettent d'établir, à partir d'annotations manuelles multiples, des corpus de référence, dont leur constitution est un enjeu pour le Traitement Automatique des Langues. L'objectif de cette thèse est notamment de conseiller et d'outiller les chercheurs sur les mesures d'accord inter-annotateurs, afin d'améliorer la qualité des annotations de référence.

ABSTRACT

Annotations evaluation by inter-annotator agreement measures

In this description, we present our thesis subject on annotations evaluation by inter-annotator agreement measures. These measurements make it possible to establish, from multiple manual annotations, gold standard, whose constitution is an issue for Natural Languages Processing. The aim of this thesis is notably to advise and equip researchers on inter-annotator agreement measures, in order to improve the quality of gold standard.

MOTS-CLÉS : accord inter-annotateur, annotation de référence, évaluation d'annotations.

KEYWORDS: inter-annotator agreement, gold standard, annotation evaluation.

1 L'évaluation des annotations manuelles multiples

Notre thèse, dirigée par Yann MATHET et Antoine WIDLÖCHER au sein du GREYC à Caen, s'inscrit dans le contexte de la création des ressources langagières pour le Traitement Automatique des Langues. Nous nous intéressons aux données annotées et à l'établissement d'annotations de référence. Pour établir un corpus de référence, on a souvent recours à l'annotation manuelle multiple : on soumet les mêmes données à plusieurs annotateurs humains, puis on compare leurs annotations en s'appuyant sur des mesures d'accord inter-annotateurs. Ces mesures permettent de quantifier le degré de consensus des annotateurs. Si ce degré d'accord est jugé satisfaisant, une référence est établie à partir des annotations.

Les mesures les plus connues, comme κ (Cohen, 1968), sont adaptées pour une annotation de type catégorisation (assigner une catégorie à une occurrence prédéfinie), mais elles ne conviennent pas pour la segmentation d'un continuum, où l'annotateur doit délimiter les bornes des occurrences en plus de catégoriser ces dernières (par exemple les structures multi-échelles annotées lors du projet ANNODIS (Colléter *et al.*, 2012)). Dans ce cas-là, des mesures telles que les α dédiés à l'unitizing (Krippendorff,

1980) ou γ (Mathet *et al.*, 2015) sont préconisées pour mesurer l'accord inter-annotateurs : en plus de la catégorisation, elles prennent en compte les problèmes d'alignement des unités (positions des bornes non correspondantes, enchâssement et/ou superposition de deux unités, etc.).

Utiliser les mesures adaptées permet d'avoir des valeurs plus pertinentes concernant l'évaluation des annotations multiples et pourrait faciliter ou améliorer la constitution de corpus de référence. Or leur qualité est d'autant plus importante que de ces *gold standard* découle la fabrication d'autres outils du TAL. Par exemple, (Manning, 2011) analyse les erreurs d'un étiqueteur morpho-syntaxique et estime que plus de 40% des erreurs sont dues à une mauvaise référence (erronée ou manquante de consistance). Mais nous observons une méconnaissance des mesures d'accord et les mauvaises utilisations biaisent l'établissement d'annotations de référence, comme démontré dans (Mathet *et al.*, 2015).

2 De la mesure d'accord à la compréhension fine des désaccords

Le principal enjeu de cette thèse est d'outiller les responsables de campagnes d'annotation, et plus généralement les chercheurs, pour l'analyse des mesures d'accord. Pour ce faire, il convient dans un premier temps de dresser une vue d'ensemble des campagnes d'annotation en TAL et d'en dégager les pratiques et les méthodologies, ainsi que cerner les manques et méconnaissances, en se focalisant sur l'utilisation des mesures d'accord. Nous nous focaliserons principalement sur des campagnes où les corpus sont multi-annotés, tels que ANNODIS, ANCOR (Muzerelle *et al.*, 2014) ou le corpus émotion produit par (Le Tallec *et al.*, 2011).

Une fois cet état de l'art établi, il s'agira ensuite d'affiner, voire créer, des mesures d'accord en prenant en compte les types de données et la tâche d'annotation. S'il existe certaines métriques d'évaluation prenant en compte les relations (par exemple les mesures UAS et LAS (Kübler *et al.*, 2009) pour les analyseurs syntaxiques), nous aimerions généraliser le principe de ces mesures à d'autres types de données. Il en va de même pour des mesures prenant en compte les attributs et les valeurs. En ce sens, les chaînes de coréférence semblent être un excellent terrain d'expérimentation sur lequel nous concentrerons nos efforts. En effet, ces chaînes reposent autant sur des mécanismes de segmentation que sur des structures relationnelles. Là encore, le projet ANCOR nous permettra de mener nos investigations.

À terme, notre travail devra aussi formuler des recommandations selon les campagnes. Elles s'appuieront sur des méthodes d'évaluations et des observations liées aux types de données et aux tâches d'annotation. Elles pourront être réalisées dès les premières annotations produites et être implémentées dans un outil. Cela permettra aux responsables des campagnes de mieux cibler les causes du désaccord en identifiant les zones et les configurations dans lesquelles le désaccord émerge en priorité, et ainsi améliorer les consignes d'annotation en conséquence. Ce pan de notre travail a un double objectif : d'une part, améliorer et faciliter l'élaboration de corpus de référence, et d'autre part mieux comprendre l'accord inter-annotateurs, comme récemment préconisé par (Bregeon *et al.*, 2019).

Références

BREGEON D., ANTOINE J.-Y., VILLANEAU J. & LEFEUVRE-HALFTERMEYER A. (2019). Redonner du sens à l'accord interannotateurs : vers une interprétation des mesures d'accord en termes

de reproductibilité de l'annotation.

COHEN J. A. (1968). Weighted kappa : nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological bulletin*, **70 4**, 213–220.

COLLÉTER M., FABRE C., HO-DAC L.-M., PÉRY-WOODLEY M.-P., REBEYROLLE J. & TANGUY L. (2012). *La ressource ANNODIS multi-échelle : guide d'annotation et bonus*. Rapport interne. HAL : hal-00983076.

KRIPPENDORFF K. (1980). *Content Analysis : An Introduction to Methodology*. Beverly Hills, CA : Sage Publications, Inc.

KÜBLER S., MCDONALD R., NIVRE J. & HIRST G. (2009). *Dependency Parsing*. Morgan and Claypool Publishers.

LE TALLEC M., ANTOINE J.-Y., VILLANEAU J. & DUHAUT D. (2011). Affective Interaction with a Companion Robot for Hospitalized Children : a Linguistically based Model for Emotion Detection. In *5th Language and Technology Conference (LTC'2011)*, p. 6 pages, Poznan, Poland. 6 pages, HAL : hal-00664618.

MANNING C. D. (2011). Part-of-speech tagging from 97% to 100% : Is it time for some linguistics ? In A. F. GELBUKH, Éd., *Computational Linguistics and Intelligent Text Processing*, p. 171–189, Berlin, Heidelberg : Springer Berlin Heidelberg.

MATHET Y., WIDLÖCHER A. & MÉTIVIER J.-P. (2015). The unified and holistic method gamma (γ) for inter-annotator agreement measure and alignment. **41**(3), 437–479. DOI : 10.1162/COLI_a_00227.

MUZERELLE J., LEFEUVRE A., SCHANG E., ANTOINE J.-Y., PELLETIER A., MAUREL D., ESHKOL I. & VILLANEAU J. (2014). ANCOR_Centre, a Large Free Spoken French Coreference Corpus : description of the Resource and Reliability Measures. In ELRA, Éd., *LREC'2014, 9th Language Resources and Evaluation Conference.*, p. 843–847, Reyjavik, Iceland. HAL : hal-01075679.

6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Un projet interdisciplinaire d'initiation à la recherche par le TAL

Nicolas Ballier[1] Jean-Baptiste Yunès [2]

(1) CLILLAC-ARP, 7 rue Thomas Mann, 75013 Paris, France
(2) IRIT, 7 rue Thomas Mann, 75013 Paris, France
{nicolas.ballier, jean-baptiste.yunes}@u-paris.fr

RESUME

Le poster expliquera le déroulement du projet RELIA, qui a permis la soumission de cinq posters.

ABSTRACT

Using NLP to foster interdisciplinarity
The poster is to present the project that funded the submission of 5 undergraduate posters.

MOTS-CLES : interdisciplinarité, initiation à la recherche, python.
KEYWORDS: interdiscplinarity, learning-by-doing paradigm, python.

Le poster présente RELIA, Recherche En Licence Informatique et Études Anglophones, projet associant des étudiants de L3 de la Licence Informatique et de la Licence d'Etudes Anglophones de Paris Diderot dans le cadre d'un appel d'offre du programme IdEx Université de Paris (ANR-18-IDEX-0001), qui visait à développer l'initiation à la recherche dès la Licence. Nous expliciterons le point de vue des enseignants qui ont conduit le projet, de la réponse à l'appel d'offre à la communication institutionnelle et à la valorisation du projet. Nous expliquerons comment nous avons conduit les enseignements et conçu la ventilation du budget et l'organisation du projet. Ce cours d'initiation à la recherche à partir de l'initiation au langage de programmation Python, et notamment de la bibliothèque nltk (Bird et al. 2009), s'est étalé sur douze semaines d'1h30. Les étudiant(e)s anglicistes ont réalisé en binôme avec un(e) étudiant(e) en informatique un petit projet d'analyse automatique du langage sur des problématiques linguistiques de l'anglais. Les travaux ont été présentés sous forme de posters à des étudiants de master spécialistes de Science des données. L'idée était ensuite d'assister à une conférence de jeunes chercheurs du TAL, pour y écouter des présentations de travaux, voire de soumettre une proposition. Nous reviendrons sur les réussites et les limites de ce projet inédit pour nous.

Références

BIRD S., KLEIN E. & LOPER, E., Éds. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*, New York . O'Reilly.

6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Narrative Summarization

Claude Cunha[1] Guillaume Le Garrec [1]

(1) IRIT, Université de Paris, 7 rue Thomas Mann, 75013 Paris, France

{claude.cunha, guillaume.lgarrec}@etu.univ-paris-diderot.fr

MOTS-CLÉS : RÉSUMÉ AUTOMATIQUE DE TEXTES, RÉCITS, TRAME NARRATIVE.

RÉSUMÉ

Encore aujourd'hui, le résumé de textes narratifs reste un défi. Les différentes difficultés inhérentes à un tel procédé ont été décrites dans un grand nombre de papiers (Mani, 2004). Les aspects considérés comme les plus problématiques semblent être la construction de la ligne temporelle d'une histoire ainsi que la prise en compte des différentes intrigues et sous-intrigues se déroulant en même temps. Des travaux précédants décrivaient des representations pour les relations temporelles dans les textes narratifs (Zhou 2006, Mani and Pustejovsky, 2004), mais aucun ne présentait une procédure complète pour le faire. Ceci est principalement dû à la difficulté de traiter les informations implicites. Par exemple, un personnage se souvenant d'un évènement ne donne pas nécéssairement la date ou la période à laquelle il s'est déroulé. Il est possible que le lecteur soit supposé pouvoir situer l'évènement dans le temps en fonction du contexte dans l'histoire à résumer. Plus généralement, le problème est de réussir à créer un moyen complet pour modéliser ce qui se passe dans un texte narratif. Avec ces études en tête, nous avons essayé d'adopter une autre approche pour simplifier le processus de résumé. Ainsi, nous n'avons pas développé un algorithme pour construire le modèle d'une histoire afin d'en dériver un résumé. Au lieu de cela, nous avons essayé de créer un ensemble de critères permettant de sélectionner des phrases du texte pour constituer son résumé. Ces critères reposent sur la position des phrases dans le texte, la fréquence d'apparition des synsets représentés par les mots d'une phrase, la fréquence d'apparition des noms propres dans la phrase ainsi que la longueur d'une phrase. Nous avons conservé les phrases de moins de vingt mots, car nous avons observé que les phrases plus longues étaient des phrases de description. Nous avons également pénalisé les phrases trop courtes, qui sont peu narratives. En sélectionnant des phrases selon ces critères nous avons produit un résumé qui se voulait un rapide coup d'oeil du livre. Nous avons testé cette méthode sur les sept premiers chapitres du roman *Oliver Twist* de Charles Dickens. Le résumé obtenu était incohérent car les phrases obtenues ne formaient pas d'histoire à proprement parler. Nous avions choisi de ne pas exclure les phrases comportant des marques de dialogue car elles peuvent apporter des éléments narratifs. Cependant, les phrases retenues permettaient d'obtenir certaines informations intéressantes sur les sept premiers chapitres du livre comme le fait qu'Olivier est probablement le personnage principal de l'oeuvre ou que l'environnement dans lequel se déroule l'histoire est plutôt sombre.
Ce projet a été réalisé sous la supervision de Nicolas Ballier et Jean-Baptiste Yunès à l'Université de Paris.

REMERCIEMENTS

Nous remercions Nicolas Ballier ainsi que Jean-Baptiste Yunès qui nous ont permis d'entreprendre ce projet de recherche et qui nous ont guidé dans sa réalisation. Ce travail est l'œuvre conjointe d'étudiants de la Licence Informatique et de la Licence d'Études Anglophones de Paris Diderot. Il a été financièrement soutenu par le programme IdEx Université de Paris ANR-18-IDEX-0001.

Réferences

MANI, I. (2004). Narrative summarization. *Traitement Automatique des Langues*, 45(1):1-24.

ZHOU L., MELTON G.B., PARSONS S., HRIPCSAK G.. A temporal constraint structure for extracting temporal information from clinical narrative. *J Biomed Inform* 2006; 424-39.

MANI, I. AND J. PUSTEJOVSKY. 2004. Temporal Discourse Models for Narrative Structure. Proceedings ofthe ACL 2004 Workshop on Discourse Annotation.

6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Typologie de chaînes de référence à la lumière de corpus annotés diversifiés

Silvia Federzoni[1]
(1) Laboratoire CLLE (CNRS-UMR 5263), Université de Toulouse 2 - Jean Jaurès, 5 allées Antonio Machado, 31058 Toulouse
silvia.federzoni@univ-tlse2.fr

RÉSUMÉ

Ce projet de thèse a pour objectif la définition d'une typologie des chaînes de référence basée sur une description systématique des enchaînements des expressions référentielles dans différents corpus annotés en chaînes de référence.

ABSTRACT

Typology of reference chains in the light of diverse annotated corpora

This thesis project aims to define a typology of reference chains based on a systematic description of the sequences of referential expressions in different corpora annotated in reference chains.

MOTS-CLÉS : chaînes de référence, corpus annotés, typologie, continuité référentielle, discours .

KEYWORDS: reference chains, annotated corpora, typology, referential continuity, discourse.

Notre thèse [1] porte sur les chaînes de référence. Ce projet s'inscrit dans un contexte de recherche dynamique pour le français, comme le montrent les numéros de revue qui ont été récemment consacrés aux chaînes de référence : numéro 195 de *Langue Française* (Schnedecker *et al.*, 2017), numéro 25 de *Discours* (Sarda & Vigier, 2019) et le numéro 72 des *Cahiers de praxématique* (Gardelle *et al.*, 2019). Les *chaînes de référence* (désormais CR) sont des structures discursives qui regroupent plusieurs propositions ayant en commun un même référent – être humain, entité abstraite ou événement – signalé dans les textes par le biais d'expressions linguistiques – noms propres, syntagmes nominaux, pronoms – appelées *expressions référentielles*, *mentions* ou *maillons* (Corblin (1995) ; Charolles (1988) ; Schnedecker (1997), inter alia). Les maillons sont liés par une relation dite de *coréférence*, et leur succession dans les textes contribue à créer des liens de cohésion entre différents segments de discours. De ce fait, les CR constituent un mécanisme fondamental dans l'organisation et l'interprétation du discours. Pour cette raison, elles ont fait l'objet de nombreuses études. En TAL, les efforts ont principalement porté sur le développement et l'amélioration des systèmes de détection et résolution automatique de l'anaphore et de la coréférence (Recasens & Hovy (2010); Mitkov (2014); Oberle (2019); Désoyer *et al.* (2015); Brassier *et al.* (2018), inter alia).

En linguistique comme en TAL, les travaux portant sur les CR se fondent sur l'exploration de corpus annotés. Bien que des ressources de grande taille soient disponibles, aussi bien pour l'anglais (cf. Poesio *et al.* (2016)), comme OntoNote (Pradhan *et al.*, 2012) ou WikiCoref (Ghaddar & Langlais, 2016), que pour le français écrit, comme ANNODIS (Péry-Woodley *et al.*, 2011), Democrat (Lattice *et al.*, 2019), elles n'ont pas permis, jusqu'à présent, de mettre au jour une définition complète et systématique des CR. En effet, ces ressources ont été conçues pour répondre à des objectifs différents

1. Encadrée par Cécile Fabre et Lydia-Mai Ho-Dac.

et rassemblent donc des annotations différentes. Si les auteurs concordent sur la definition de CR comme « la suite des expressions d'un texte entre lesquelles l'interprétation construit une relation d'identité référentielle » (Corblin, 1995) ainsi que sur le nombre minimum de maillons qui doit être de trois [2] (Schnedecker, 2019), il n'y a pas de consensus sur la taille maximale d'une CR. Une autre question qui fait débat dans la littérature est de savoir quels sont les éléments aptes à constituer les maillons d'une CR et comment les délimiter, car un choix s'impose entre prendre en considération uniquement la tête lexicale ou bien inclure ses dépendants. De plus, la prise en compte des singletons (référents qui ne font pas l'objet d'une réprise référentielle) varie d'une annotation à l'autre : certains considèrent qu'il est indispensable de les annoter pour que les systèmes de résolution soient en mesure de les détecter. D'autres, ayant des objectifs différents, ne les annotent pas.
Ce manque de consensus se traduit dans les ressources existantes, conçues sur des modèles linguistiques différents, par une grande hétérogénéité en termes de choix d'annotation, ce qui rend les résultats obtenus difficilement comparables. De même, au vu de cette hétérogéneité, toutes les applications TAL ne peuvent pas exploiter n'importe quelle ressource, et les architectures des sytèmes de résolution développés sont dépendantes du type de ressource sur laquelle le modèle a été entraîné.

À cette difficulté, s'ajoute la complexité du phénomène des CR, dont l'analyse requiert la prise en compte des configuarations d'indices (Das & Taboada, 2019). Par conséquent, aucune étude à large échelle, n'a proposé une description systématique des CR dans leur complexité et leur complétude. Pour l'anglais, la plupart de travaux se focalise sur les paires coréférentielles, sans analyser les CR complètes. Pour le français, les études sur les CR ont été effectuées sur des corpus de petite taille ou échantillons de texte (Schnedecker & Longo, 2012), parfois en se focalisant sur un type de référent particulier.

Dans ce contexte, un premier objectif de notre thèse est de fournir une typologie des CR. Pour y parvenir, il s'agit préalablement d'unifier les corpus annotés afin de fournir une description, la plus exhaustive possible, de la complexité et de la variété des CR. À partir de cette typologie, la thèse proposera une étude contrastive entre différents types de textes ainsi qu'une description systématique qui puisse être exploitée pour l'amélioration d'un modèle de prédiction automatique des CR.

L'idée est de concevoir un modèle qui nous permette de considérer les enchaînements des maillons et d'aller au-délà d'un traitement par paire "antécédent-anaphorique", tout en prenant en compte les traits linguistiques et les variations qui peuvent exister entre différents genres textuels ou différents niveaux d'expertise rédactionnelle. L'application de ce modèle consentira de faire émerger les traits linguistiques ayant une influence sur la présence d'une expression référentielle donnée. L'application de ce modèle nous permettra également de confronter certaines hypothèses cognitives à la réalité des usages en corpus (en particulier les théories de l'accessibilité (Ariel, 2001) et du centrage (Walker *et al.*, 1998)). Nous serons ainsi en mesure d'évaluer systématiquement les décalages entre ce qui est théoriquement attendu en matière de chaînes de référence et ce qui est effectivement attesté dans les usages réels. L'analyse de ces décalages nous permettra de mieux caractériser les traits linguistiques que les systèmes de résolution devraient/pourraient apprendre pour améliorer leurs performances.

À ce stade, nous n'avons pas encore établi la procédure nécessaire au développement de ce modèle. Notre connaissance des modèles existants doit d'abord être approfondie. Nous explorerons ensuite l'intérêt des modèles comme le CRF (Kudo, 2005 cité par Godbert & Benoit (2017)), ou bien des modèles bayésiens, en prenant en compte une diversité de traits pour évaluer ceux qui sont les plus discriminants.

2. À default on parle d'*anaphore* ou de *coréférence* (Schnedecker, 2019)

Références

ARIEL M. (2001). Accessibility theory : An overview. *Text representation : Linguistic and psycholinguistic aspects*, **8**, 29–87.

BRASSIER M., PURET A., VOISIN-MARRAS A. & GROBOL L. (2018). Classification par paires de mention pour la résolution des coréférences en français parlé interactif. *Conférence jointe CORIA-TALN-RJC 2018*.

CHAROLLES M. (1988). Les plans d'organisation textuelle : périodes, chaînes, portées et séquences. *Pratiques*, **57**(1), 3–13.

CORBLIN F. (1995). *Les formes de reprise dans le discours. Anaphores et chaînes de référence*. Presses Universitaires de Rennes.

DAS D. & TABOADA M. (2019). Multiple signals of coherence relations. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, (24).

DÉSOYER A., LANDRAGIN F. & TELLIER I. (2015). Machine Learning for Coreference Resolution of Transcribed Oral French Data : the CROC System. *Vingt-deuxième Conférence sur le Traitement Automatique des Langues Naturelles*, p. 439–445.

GARDELLE L., ROSSI C. & VINCENT-DURROUX L. (2019). La gestion de l'anaphore en discours : complexités et enjeux. *Cahiers de praxématique*, (72).

GHADDAR A. & LANGLAIS P. (2016). WikiCoref : An English Coreference-annotated Corpus of Wikipedia Articles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia : European Language Resources Association (ELRA) European Language Resources Association (ELRA).

GODBERT E. & BENOIT F. (2017). Détection de coréférences de bout en bout en français.

LATTICE, LILPA, ICAR & IHRIM (2019). Democrat. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

MITKOV R. (2014). *Anaphora resolution*. Routledge.

OBERLE B. (2019). Détection automatique de chaînes de coréférence pour le français écrit. *Conférence sur le Traitement Automatique des Langues Naturelles (TALN-RECITAL) 2019*.

PÉRY-WOODLEY M.-P., AFANTENOS S., HO-DAC L.-M. & ASHER N. (2011). La ressource ANNODIS, un corpus enrichi d'annotations discursives. *Traitement Automatique des Langues*, **52**(3), 71–101.

POESIO M., PRADHAN S., RECASENS M., RODRIGUEZ K. & VERSLEY Y. (2016). Annotated corpora and annotation tools. In *Anaphora Resolution*, p. 97–140. Springer.

PRADHAN S., MOSCHITTI A., XUE N., URYUPINA O. & ZHANG Y. (2012). Conll-2012 shared task : Modeling multilingual unrestricted coreference in ontonotes. p. 1–40.

RECASENS M. & HOVY E. (2010). Coreference resolution across corpora : Languages, coding schemes, and preprocessing information. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, p. 1423–1432.

SARDA L. & VIGIER D. (2019). *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, (25).

SCHNEDECKER C. (1997). *Nom propre et chaînes de référence*. Recherches linguistiques. Librairie Klincksieck.

SCHNEDECKER C. (2019). De l'intérêt de la notion de chaîne de référence par rapport à celles d'anaphore et de coréférence. *Cahiers de praxématique*, (72).

SCHNEDECKER C., GLIKMAN J. & FRÉDÉRIC L. (2017). Les chaînes de référence en corpus. *Langue française*, (195).

SCHNEDECKER C. & LONGO L. (2012). Impact des genres sur la composition des chaînes de référence : le cas des faits divers. *3ème Congrès Mondial de Linguistique Française*, p. 1957–1972.

WALKER M. A., JOSHI A. K. & PRINCE E. F. (1998). *Centering theory in discourse*. Oxford University Press.

6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

From chatbot to personal assistant.

Emma JEANNETTE Sawssen HADDED
Université de Paris, 5 Rue Thomas Mann, 75013 Paris, France
emma.jeannette08@gmail.com, saw.hadded@gmail.com

RÉSUMÉ

Les chatbots (emprunt direct de l'anglais, chat = conversation amicale et informelle; bot = un programme informatique avec un fonctionnement automatique) sont des logiciels informatiques créés pour simuler des conversations à ressemblance presque humanoïde. Ces logiciels analysent les textes ou commandes vocales pour adapter leurs réponses. Eliza est un des tout premiers chatbots crée par Weizenbaum entre 1964 et 1966. Le but initial d'Eliza était de démontrer que la communication entre les humains et les machines n'était pas nécessaire. Seulement le programme a engendré de nouvelles recherches sur les machines et le langage est devenu l'ancêtre de nos chatbots actuels. Notre expérience se concentre sur les différents chatbots de renommée au fil des années de 1964 à 2018, leurs améliorations mais aussi leurs limitations. Nous avons testés chaque chatbot (ELIZA,A.L.I.C.E. 1995,Cleverbot 1997, Mitusku 2005) sur des phrases tests.

(a) How are you ?
(b) What time is it ?
(c) What's blakmail ?
(d) Can you feel pain ?
(e) Tell me a joke.
(f) Can I eat a table ?
(g) What is a family ?
(h) What is a bae ?
(I) Do you think animals have feelings ?

Phrases de protocole 1

Tout d'abord, nous nous sommes présentées sous un faux nom et nous leur avons demandé de répéter ce nom plus tard dans la conversation pour voir s'ils enregistraient les informations sur l'utilisateur. Puis une question basique (a) pour voir s'ils pouvaient répondre à des questions simples. Nous avons aussi testé des phrases avec des fautes d'orthographes et de l'argot (c et h) pour voir si les chatbots pouvaient traiter des mots dont l'orthographe n'est pas familière au système. Nous avons aussi demandé aux chatbots la definition d'un mot assez simple (g) pour voir s'ils étaient capables d'y répondre. Durant l'expérience, nous leurs avons aussi demandé s'ils pouvaient nous donner l'heure actuelle (b). Nous leur avons aussi posé des questions plus philosophiques (i) et une question qui pourrait nous montrer s'ils étaient programmés pour répondre comme un humain le ferait plutôt qu'un robot(d). Enfin, nous avions remarqué que Mitsuku pouvait assimiler des objets à des propriétés comme la taille ou l'emploi de l'objet et était donc capable d'analyser la signification derrière chaque mot et les assembler pour comprendre la signification de la phrase. Nous avons donc créé une autre question (f) pour voir si les autres chatbots étaient capables de la même chose. Le deuxième protocole a été mis en place pour voir quels chatbots étaient les plus 'performants' ou en tout cas capables de répondre à nos questions de manière cohérente. Nous avons chacune posés ses questions aux différentes chatbot puis comparés nos résultats. Sur les questions ou nous avons toute les deux jugés que la chatbot avait repondu de facon « naturel », nous lui accordions 11 points. Si la chatbot repondait de facon satisfaisante a l'une de nous pas pas a l'autre, nous lui avons attribué 5,5 points.

MOTS-CLÉS : chatbot, assistant virtuelle, Eliza,A.L.I.C.E.,Cleverbot,Mitsuku,Siri
KEYWORDS : chatbots, virtual assistant, Eliza,A.L.I.C.E.,Cleverbot,Mitsuku,Siri

State of the Art

We will discuss the various states of chatbots, from their early versions, to their di erent ameliorations and to the modern personal assistants that uses chatbots to communicate with the user. A chatbot is a computer program that simulates a human conversation whether it is by text or voice command (or both) . Eliza is known as one of the first chatbot that existed. The project was developed by Joseph Weizenbaum between 1964 and 1966 at the MIT Artificial Intelligence Laboratory. Eliza functions as a psychotherapist: designed for the user to talk about their problem. Eliza reformulates the sentences of the user into a question triggered by keywords to keep the conversation going and if it is not able to reformulate it will use pre-made questions. Because of her almost "human like" way of answering Eliza could attempt the Turing test. A.L.I.C.E., another chatbot, was very inspired by ELIZA, and became one the best performing chatbots of its category winning the Loebner prize three time in 2000, 2001 and 2004. Alice uses a heuristic pattern matching rule to the human's input. However, it could not pass the Turing test as it was deemed "too mechanical ". Clever bot is a chatter bot web application. Siri® is a virtual assistant that answers vocal human commands. We will analyze di erent chatbots, to see their limitations throughout the years, but also how they were improved to seem almost human like, even gaining new features such as vocal recognition or a humanoid body and to see if it is possible to code a chatbot to the point where you cannot distinguish the robot from the human being.

Method

The chatbots were chosen on an extended period -from 1964 to 2011- so we could analyze and document the progress of the chatbots and later the voice recognition systems. Eliza and Alice were chosen because they were both early winners of the Loebner price and the base of their system is still used to create modern chatbots. We also decided to go for a less conventional chatbot like Cleverbot which is mainly known for its humor and sometimes nonsensical responses. Mitsuku is the five-time winner of the Loebner price and was named the "best conversational chatbot". Siri was the perfect blend of a chatbot and a voice recognition device that could show an evolution towards a new type of chatbots that could process voices inputs. The bots also work differently. Eliza was pre-programmed to answer using the inputs of the user. ALICE analyze the inputs of the user to find a fitting answer. Cleverbot is not pre-programmed but it learns from all the previous inputs it had with other users, using keywords it will search in its database for a fitting answer. Mitsuku is constantly being worked on, it uses an AIML language and learns from the conversations. Siri gets its information directly from the user's phone data or the internet. To put the chatbot to the test we used similar sentences if not the same on each chatbot to see how their output could di er.

Results

Chatbots underwent an increasing evolution over the years. Eliza was developed as a proof of the superficiality of communication between humans and machines' but has definitely exceeded the expectations and is now referred and used beyond its initial purpose. Users were now creating emotional relationship with the programs. While some of the 'older' chatbots cannot compete with most recent chatbots they are still their basis and ancestors, we also obviously don't have the same expectations between a newer program and a fifty year old one. Chatbots have now evolved into personal assistants, capable of understanding and processing voices inputs. They are now capable of having full conversation with human beings that does not seem irrelevant, they can now gather information about the user and use them later in the conversation. They have access to even more data than they use to in order to help the user: they went from being pre-programmed to answer to being able to process and find the semantic meaning being sentences and words to give answers that matches the question. Users of "smart" technology are now accustomed to chatbots and personal assistant, so companies are pushing their systems to always improve and do more. Google is now trying to develop an Al system that can engage into a conversation and make appointments for the user while resembling a human voice and conversation, it will even imitate the "hum" sound as humans does when it cannot process the inputs from the other person on the phone. However, it is only trained to carry on conversation in certain domain and cannot carry out general conversation yet, 'Duplex can only carry out natural conversations after being deeply trained in such domains. It cannot carry out general conversations'. Chatbots are also taking new appearances other than on a computer screen like Sophia the humanoid robot developed by Hanson Robotics. Sophia's speech is not pre-made but comes directly from her own system where she analyzes conversations and extract data to improve her answers. This shows that chatbots are becoming even more efficient in analyzing words but also the context of the conversation, so it is now touching the pragmatic field of language.

References

BOTPRESS. (2018) *“The Whats and Why of Chatbots”*

WEIZENBAUM J. (1966). *"ELIZA - a computer program for the study of natural language communication between man and machine."* Communications of the ACM 9.1 36-45.

COPELAND B. JACK. (2000). *"The turing test."* Minds and Machines 10.4 519-539.

ABUSHAWAR, BAYAN & ERIC ATWAL. (2015). "ALICE chatbot: Trials and outputs." Computación y Sistemas 19.4: 625-632.

MAULDIN, MICHAEL L. (1994). *"Chatterbots, tinymuds, and the turing test: Entering the loebner prize competition."* AAAI. Vol. 94.

TECHJOURNEY. (2016). A.L.I.C.E Alicebot Artificial Intelligence (AI) Chatbot

CARPENTER, ROLLO. (2011). *"Cleverbot."*

APPLE. (2018). Siri

YANIV LEVIATHAN. (2018). Google Duplex: An AI System for Accomplishing Real-World Tasks Over the Phone

YUEFANG ZHOU & MARTIN H. FISCHER. (2019). AI Love You: Developments in Human-Robot Intimate Relationships Part II chapitre 3, p.106

6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Rap and Text generation : How a rap text can be generated considering the metrical, and lexical questions

E. Partouche & N. Perdriau

CLILLAC-ARP, 7 rue Thomas Mann, 75013 Paris, France
IRIT, 7 rue Thomas Mann, 75013 Paris, France
{nelson.perdriau,eitanite.partouche}@etu.univ-paris-diderot.fr

Résumé

Dans cet article, nous chercherons à montrer comment nous pouvons utiliser un langage de mémoire à long terme et à court terme (LSTM) pour générer un texte de rap ou créer une "machine à rap", y compris en prenant en compte diverses exigences récurrentes pour ce type de textes comme les syllabes, le rythme et un vocabulaire diversifié.

Abstract

In this paper, we will aim to show how we can use a Long Short-Term Memory Language (LSTM) to generate a rap text or to create a “rap machine”, including the generation of diverse recurrent requirements for this kind of texts such as syllables, rhythm and a diverse vocabulary.

Mots-clés : Apprentissage par machine, linguistique, rap, LSTM, python
Keywords: Machine Learning, linguistics, rap, LSTM, python

A large part of the previous studies on the generation of text have as basis a Neural network, especially Recurrent Neural Networks (RNN): when a Neural Network is a circuit of artificial neurons made to solve artificial intelligence (AI) problems, a RNN, still composed of Neural Networks, can remember them because they are recurrent but encounters some limits. This is why the LSTM model appears to be more effective within the scope of text generation : it has the possibility to correct the vanishing gradient problem of the RNN and thus learning what to remember and what to forget. An interesting work was made on the use of ghostwriting through LSTM : the goal is to give the impression that a rapper has produced a new song, by reproducing his style of writing.

One of the main exercices we had to perfom was writing a model in Python, using the TensorFlow library and a RNN, that would take into account our needs. In it, we have added the CMU Pronouncing Dictionnary, allowing us to read the generated text produced with the good accents, considering the word's syllable(s) and the metric of the setence, the lexical stress playing also a role in the accentuation of words. Then, we trained the model on a computer for about 113.000 iterations.

In the generated verses, we can observe a important quantity of nonsensical words, as in the iteration n°1000 : "kidsiin", "throuictifing", "griends" for examples. Then, the CMU Pronouncing Dictionary was not really useful for their pronunciation because it did not recognize these words, being non listed. The more we trained our model, the more the produced text were meaningful. The final iteration completely illustrate this fact… The poster will discuss the output productions of the LSTM. The project was supervised by Nicolas Ballier and Jean-Baptiste Yunès at Université de Paris

6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Comparaison de méthodes d'extraction de mots-clés non supervisées pour les disciplines des Sciences Humaines et Sociales

Alaric TABARIES
Master Information et Communication, Université de Toulon, France
Encadré par David REYMOND, IMSIC, 70 Avenue Roger Devoucoux, 83000 Toulon
alaric-tabaries@etud.univ-tln.fr

Accéléré par l'émergence de la voie verte, la quantité d'information scientifique disponible en ligne augmente à un rythme sans précédent. Ce phénomène rend le processus de veille documentaire, essentiel à la recherche scientifique, tant complexe que chronophage. C'est dans ce contexte que l'extraction d'information se pose en tant que service support au prétraitement de la sélection documentaire. En effet, les mots-clés, qui représentent les sujets principaux traités dans un document, sont particulièrement utiles pour distinguer les ressources intéressantes dans un ensemble de documents important. Cependant, très peu en sont pourvus. L'extraction automatique de mots-clés permet de remédier à ce problème et montre d'ores et déjà des résultats satisfaisants sur des corpus de référence. Il a cependant été établi que certaines méthodes d'extraction performent mieux que d'autres pour les productions dans les disciplines des Sciences Humaines et Sociales.

Nous proposons donc de mettre au point une expérimentation sur des jeux de données réels issus de publications identifiées sur la plateforme HAL en comparant les résultats selon les disciplines des publications afin d'identifier les méthodes d'extraction non supervisées qui performent le mieux pour servir un outil veille répondant au problème de surcharge informationnelle. Cette expérimentation consiste donc à comparer des mots-clés extraits de résumés de publications HAL à l'aide de méthodes non supervisées à des mots-clés préalablement annotés par des étudiants de Master Langues et Sociétés dans le but d'établir des mesures de pertinence ainsi qu'un classement de performance des différentes méthodes d'extraction selon diverses disciplines des Sciences Humaines et Sociales.

Le tableau 1 présente une partie des résultats obtenus.

	Sciences de l'éducation	Langues	Médias et communication	Histoire	Gestion	Psychologie
TfIdf	5.385	6	5.333	4.125	4	3.714
KPMiner	6.462	6.5	5.444	3.875	3.375	6.571
YAKE	3.462	5	5.778	2.875	**2.25**	5.286
TextRank	3.769	4	3.111	4.375	4.25	2.857
SingleRank	5	4.333	3.111	4.75	3.25	**2.714**
TopicRank	6.769	7.083	5.667	5	6.125	4.286
TopicalPageRank	**3.154**	**2.083**	**2**	**2.375**	2.875	4.143
PositionRank	3.462	3	3.222	3.625	3	3.714
MultipartiteRank	6.846	5.833	5.222	4.875	5.25	4.857

TABLE 1 – Rang moyen par méthode d'extraction pour des disciplines des SHS en anglais

Communications des apprenti-e-s chercheur-euse-s 2020, pages 15–15.

6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

TreeTagger entraîné avec le *Critical Pronouncing Dictionary* de J. Walker face aux textes modernes

Francois Huang, Blanche Miret, Preethi Srinivasan, Dao Thauvin
RELIA Recherche En Licence Informatique et études Anglophones
Encadrants : Jean-Baptiste Yunès et Nicolas Ballier
Université de Paris, 5 rue Thomas Mann, 75013, Paris
blanche.miret@etu.univ-paris-diderot.fr,
francois.huang@etu.univ-paris-diderot.fr,preethi.lfp@gmail.com,
dao.thauvin@etu.univ-paris-diderot.fr

Ce travail est l'œuvre conjointe d'étudiants de la Licence Informatique et de la Licence d'Études Anglophones de Paris Diderot. Il a été financièrement supporté par le programme IdEx Université de Paris ANR-18-IDEX-0001

RÉSUMÉ

TreeTagger (Helmut Schmid) est un outil moderne d'annotation de texte, par des lemmes et des catégories grammaticales. L'objectif de cette recherche est de déterminer si cet outil est capable d'assimiler les catégories grammaticales utilisées au 18ème siècle. Pour ce faire, nous avons utilisé le *Critical Pronouncing Dictionary* de John Walker (1791) afin de récupérer des catégories grammaticales datant du 18ème siècle des différents mots présents dans la langue anglaise pour obtenir un tagset. Ensuite nous avons créé deux fichiers .par entraînés avec des phrases du *Brown Corpus*. L'un utilise le tagset obtenu précédemment et un lexique extrait du dictionnaire de John Walker et l'autre utilise un jeu d'étiquettes et un lexique extraits du *Brown Corpus*. À partir des fichiers .par, nous avons laissé notre outil analyser certains textes modernes provenant du *Brown Corpus* de la bibliothèque NLTK et une partie du dictionnaire de John Walker. Sur des mêmes fichiers test, nous aboutissons à une précision de 33.5% en moyenne (32% de précision sur un texte provenant du dictionnaire de Walker et 35% de précision sur un texte provenant du *Brown Corpus*) avec le fichier .par utilisant les tags présents dans le dictionnaire de John Walker alors que la précision avec le fichier .par créé à partir du *Brown Corpus* est de 93.5% en moyenne (91% de précision sur un texte provenant du dictionnaire de Walker et 96% de précision sur un texte provenant du *Brown Corpus*), ce qui nous amène à penser que les tags du 18eme siècle ne sont pas adaptés à l'annotation de texte avec TreeTagger. Cependant, l'entraînement de TreeTagger et les expériences ont été effectués sur une faible quantité de données, et notre méthode pour utiliser les tags du 18ème nécessite une traduction des tags du 18ème siècle en tags de Brown Corpus. Nous perdons donc certains tags spécifiques du dictionnaire de Walker. En améliorant ces aspects, les résultats peuvent différer. Une analyse qualitative a par ailleurs montré l'incohérence de certaines étiquettes de Walker.

MOTS-CLÉS : TreeTagger, Walker, catégorie grammaticale, 18ème siècle.

KEYWORDS: TreeTagger, Walker, Part-Of-Speech Tag, 18th century.

REMERCIEMENTS

Nous remercions en premier lieu Nicolas Ballier et Jean-Baptiste Yunès qui nous ont accompagnés tout au long de ce projet et introduits aux règles de la recherche académique. Nous remercions également Nicolas Trapateau pour son travail effectué sur le *Critical Pronouncing Dictionary* de John Walker que nous avons utilisé comme jeu de données dans nos recherches, ainsi que les organisateurs de la conférence des Apprenti-e-s Chercheur-euse-s 2020 de nous donner la possibilité de partager ce travail. Enfin, merci au programme IdEx de l'Université de Paris grâce à qui ce projet fut financé.

Références

FRANCIS W. N. & KUCERA H. (1979). *Brown corpus manual*. Letters to the Editor.

HUANG F., MIRET B., SRINIVASAN P. & THAUVIN D. (2020). Github repository. https://github.com/daothauvin/TreeTaggerWithWalker.

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

TRAPATEAU N. (2015). *Placement de l'accent et voyelles inaccentuées dans la prononciation de l'anglais du XVIIIe siècle sur la base du témoignage des dictionnaires de prononciation, des vers et de la musique vocale*. Thèse de doctorat, Université de Poitiers.

WALKER J. (1791). *A Critical Pronouncing Dictionary*. British Library.

6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

TreeTagger entraîné avec des données modernes face au *Critical Pronouncing Dictionary* de J. Walker

Francois Huang, Blanche Miret, Preethi Srinivasan, Dao Thauvin
RELIA Recherche En Licence Informatique et études Anglophones
Encadrants : Jean-Baptiste Yunès et Nicolas Ballier
Université de Paris, 5 rue Thomas Mann, 75013, Paris
blanche.miret@etu.univ-paris-diderot.fr,
francois.huang@etu.univ-paris-diderot.fr,preethi.lfp@gmail.com,
dao.thauvin@etu.univ-paris-diderot.fr

Ce travail est l'œuvre conjointe d'étudiants de la Licence Informatique et de la Licence d'Études Anglophones de Paris Diderot. Il a été financièrement supporté par le programme IdEx Université de Paris ANR-18-IDEX-0001

RÉSUMÉ

Peut-on utiliser un outil d'étiquetage morpho-syntaxique pour mesurer l'évolution d'une langue à travers les siècles, et notamment reconnaître les mots devenus obsolètes ? Dans quelle mesure cet outil arrive-t-il à s'adapter à un état de langue plus ancien que celui avec lequel il a été entraîné ? C'est pour répondre à ces interrogations que nous avons appliqué TreeTagger, exercé à identifier et catégoriser les mots de l'anglais moderne, sur le *Critical Pronouncing Dictionary* de John Walker datant de 1791.

Le résultat de l'expérience permet par exemple de retrouver la différence d'évolution attendue entre les différentes catégories grammaticales de la langue : les prépositions étant sujettes à peu de transformations, la reconnaissance de celles du 18e siècle ne pose pas de problème ; celle des noms communs ou adjectifs est moins évidente. Le taux de précision mesuré sur un échantillon de 200 mots est de 93,5%, résultat inférieur aux 96,36% officiels de TreeTagger (Schmid, 1994) sur des données de la même époque que celles sur lesquelles l'outil a été entraîné. Le taux de rappel, c'est à dire la capacité de reconnaissance d'une certaine catégorie grammaticale, est de 98% pour les prépositions, 91% pour les noms, 90% pour les adjectifs.

L'approche pour la détection des mots obsolètes s'est faite en observant les termes se voyant attribuer "unknown" comme lemme, ce qui fut le cas pour 9,2% de l'ensemble des tokens, plus précisément 35 748 sur 386 172 au total. Cependant, le jeu de données étant un dictionnaire, l'analyse du résultat peut être départagée entre celle des mots vedettes hors contexte d'une part et l'ensemble des définitions d'autre part. 89,0% des mots identifiés comme potentiellement obsolètes appartiennent à l'ensemble des mots vedettes et sur un échantillon de 200 tokens, seulement 38,5% se sont révélés l'être réellement. En revanche, le taux de réussite sur les mots contenus dans les définitions est bien plus prometteur : 78,0% sont en effet obsolètes, avec une précision de marquage morpho-syntaxique de 66,7%, expression de la capacité de TreeTagger à s'adapter à un vocabulaire désuet dans le cas de tokens contextualisés. Au total, 1,2% des mots en contexte du jeu de données complet a été identifié comme éventuellement obsolète. Une suite de la recherche pourrait conduire à élargir les échantillons étudiés et mesurer le pourcentage de mots reconnus comme obsolètes parmi un jeu de token certifié comme tel.

Finalement, avec une précision relativement élevée dans l'utilisation des lemmes marqués 'unknown'

pour identifier l'obsolescence des mots, TreeTagger semble être un outil pertinent d'évolution du langage entre deux périodes.

MOTS-CLÉS : TreeTagger, catégorie grammaticale, obsolescence, évolution, prédiction , 18ème siècle.

KEYWORDS: TreeTagger, Part-of-Speech tag, obsolescence, evolution, prediction, 18th century.

REMERCIEMENTS

Nous remercions en premier lieu Nicolas Ballier et Jean-Baptiste Yunès qui nous ont accompagnés tout au long de ce projet et introduits aux règles de la recherche académique. Nous remercions également Nicolas Trapateau pour son travail effectué sur le *Critical Pronouncing Dictionary* de John Walker que nous avons utilisé comme jeu de données dans nos recherches, ainsi que les organisateurs de la conférence des Apprenti-e-s Chercheur-euse-s 2020 de nous donner la possibilité de partager ce travail. Enfin, merci au programme IdEx de l'Université de Paris grâce à qui ce projet fut financé.

Références

HUANG F., MIRET B., SRINIVASAN P. & THAUVIN D. (2020). Github repository. https://github.com/BlancheMiret/TreeTagger_on_Walker.

KHURANA D., KOLI A., KHATTER K. & SINGH S. (2017). *Natural Language Processing : State of The Art, Current Trends and Challenges*. Thèse de doctorat, Manav Rachma International University.

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

TRAPATEAU N. (2015). *Placement de l'accent et voyelles inaccentuées dans la prononciation de l'anglais du XVIIIe siècle sur la base du témoignage des dictionnaires de prononciation, des vers et de la musique vocale*. Thèse de doctorat, Université de Poitiers.

WALKER J. (1791). *A Critical Pronouncing Dictionary*. British Library.

6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Synthèse de la parole à partir du texte

Étudiante : Yongxin Zhou[1] Encadrant : Claude Montacié[2]
(1) M2 ScLan : Langue et Informatique, Sorbonne Université, Paris, France
(2) Professeur de l'UFR de Sociologie et d'Informatique pour les sciences humaines,
Sorbonne Université, 28 rue Serpente, 75006 Paris, France
yongxin.zhou@etu.sorbonne-universite.fr, Claude.Montacie@paris-sorbonne.fr

RESUME

Dans un projet précédent de « Apprentissage de Modèles de Markov cachées et détection de mots-clés », chaque étudiant a déjà choisi un fichier audio contenant 3 minutes d'interviews radiophoniques pour réaliser les tâches de projet. Il y avait plusieurs langues au choix, nous avons choisi un corpus de parole spontanée en chinois mandarin. Plus précisément, il s'agit un extrait de 3 minutes d'une interview entre deux locuteurs : une femme qui a été interviewée et un présentateur qui a posé les questions et qui a amené l'interview.
Pour ce projet de synthèse de la parole, nous avons choisi 5 tours de parole ayant une durée minimale de 7 secondes, nommés respectivement de « T1 » jusqu'à « T5 ». Les quatre premiers tours sont prononcés par la locutrice, le dernier est prononcé par le locuteur. Nous tenons aussi compte de la prosodie pour une meilleure intelligibilité.
Le projet a pour l'objectif de faire une synthèse de la parole à partir du texte (ou TTS, Text-To-Speech en Anglais). Les logiciels eSpeak et Mbrola sont utilisés pour réaliser la tâche : eSpeak sert à transformer le texte en phonèmes; à la suite, Mbrola nous permet de générer le son à partir des phonèmes obtenus après la phonétisation. La prosodie a été extraite avec la valeur moyenne de la fréquence fondamentale du phonème et les valeurs de la fréquence fondamentale au début et au milieu du phonème. De plus, les durées phonétiques sont générées à l'aide des règles de durée spécifiquement pour le mandarin. Par ailleurs, nous avons également généré des pauses et la courbe mélodique selon l'arbre syntaxique pour le premier tour, celle du deuxième tour est générée avec les structures de performance. Ensuite, en utilisant Emofilt, des émotions de joie, de colère et de tristesse sont exprimées respectivement pour les trois autres tours de parole.
Nous avons donc généré 20 fichiers d'extension pho (ce qui est compatible avec le logiciel Mbrola) et leurs fichiers sons d'extension wav correspondants.

MOTS-CLES : Synthèse de la parole à partir du texte, eSpeak, Mbrola, parole continue en mandarin.

KEYWORDS: Text-to-Speech (TTS), eSpeak, Mbrola, continuous speech in Mandarin.