Résumés de thèses et HDR

Rubrique préparée par Sylvain Pogodalla

Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
sylvain.pogodalla@inria.fr

Rémi CARDON : remi.cardon@protonmail.com

Titre : Simplification automatique de textes spécialisés et techniques

Mots-clés : simplification automatique de textes, corpus et ressources, textes biomédicaux.

Title: *Automatic Text Simplification of Specialized and Technical Texts*

Keywords: *automatic text simplification, corpora and resources, biomedical texts.*

Thèse de doctorat en sciences du langage, Savoirs, Textes, Langages, UMR 8163, Université de Lille, sous la direction de Natalia Grabar (CR HDR, CNRS) et Anne Carlier (Pr, Sorbonne Université, Paris). Thèse soutenue le 19/04/2021.

Jury : Mme Natalia Grabar (CR HDR, CNRS, codirectrice), Mme Anne Carlier (Pr, Sorbonne Université, Paris, codirectrice), Mme Cécile Fabre (Pr, Université de Toulouse Jean Jaurès, rapporteuse), M. Thomas François (Pr, Université Catholique de Louvain, Belgique, rapporteur), Mme Emmanuelle Canut (Pr, Université de Lille, présidente), M. Pascal Denis (CR, Inria, examinateur), M. Thierry Hamon (MC, Université Sorbonne Paris Nord, examinateur), M. Horacio Saggion (MC, Universitat Pompeu Fabra, Barcelone, Espagne,).

Résumé : *La simplification automatique de textes est un domaine du traitement automatique des langues (TAL) qui vise à traiter des textes difficiles à lire pour un public donné de façon à les rendre plus accessibles. Notre objectif consiste à simplifier automatiquement les textes médicaux et de santé. Nous présentons l'ensemble de notre travail sur cette question, qui va de la collecte et analyse de corpus jusqu'aux expériences en simplification automatique.*

Nous commençons par la collecte d'un corpus comparable de textes médicaux. Ce corpus est constitué de couples de documents qui traitent du même sujet : l'un s'adressant à un public spécialiste et l'autre à un public néophyte. Le corpus contient trois types de textes : des informations sur les médicaments, des revues systématiques de littérature médicale et des articles encyclopédiques. Une fois les documents collectés, nous annotons un sous-ensemble de ces documents et analysons les transformations linguistiques qui y sont mises en œuvre lors de la simplification.

À partir du corpus comparable, nous mettons en place une méthode pour en extraire un corpus parallèle, c'est-à-dire un corpus comprenant des couples de phrases qui ont le même sens, mais diffèrent par leur degré de difficulté. Ce type de corpus représente le matériau principal pour les méthodes de simplification automatique. Notre méthode d'extraction de phrases parallèles comporte deux étapes : (1) le préfiltrage de paires de phrases candidates à l'alignement selon des heuristiques syntaxiques et (2) la classification binaire permettant de distinguer les phrases en relation de simplification. Nous évaluons différents classifieurs ainsi que l'influence du déséquilibre des données sur les performances. Afin de valoriser ce corpus parallèle, nous créons également un corpus de paires de phrases annotées selon leur similarité sémantique, avec des scores allant de 0 (sémantique indépendante) à 5 (même sémantique). Les deux corpus sont disponibles pour la recherche.

Enfin, nous présentons une série d'expériences en simplification automatique de textes médicaux en français. Ainsi, nous mettons en œuvre une méthode neuronale issue de la traduction automatique. Nous utilisons plusieurs ressources : le corpus parallèle médical construit par nous, le corpus parallèle de langue générale automatiquement traduit par nous de l'anglais vers le français ainsi qu'un lexique qui apparie des termes médicaux avec des termes ou paraphrases accessibles au grand public. Nous décrivons le protocole expérimental et menons une évaluation en deux volets, quantitatif et qualitatif. Les résultats sont comparables à l'état de l'art de la simplification en langue générale et montrent que les simplifications produites peuvent être exploitées dans le cadre d'une tâche de simplification assistée par ordinateur.

URL où le mémoire peut être téléchargé :
https://hal.archives-ouvertes.fr/tel-03343769

Hélène FLAMEIN : helene.flamein2@gmail.com

Titre : Étude de la perception d'une ville : repérage automatique, analyse et visualisation

Mots-clés : lieu, perception, ESLO, traitement automatique des langues, visualisation de l'information.

Title: *Study of the Perception of a City: Automatic Identification, Analysis and Visualization*

Keywords: *location, perception, ESLO, natural language processing, information visualization.*

Thèse de doctorat en sciences du langage, Laboratoire Ligérien de Linguistique, UMR 7270, UFR Lettres, Langues et Sciences Humaines, Université d'Orléans, sous la direction de Iris Eshkol-Taravella (Pr, Université Paris Nanterre) et Gabriel Bergounioux (Pr, Université d'Orléans). Thèse soutenue le 10/12/2019.

Jury : Mme Iris Eshkol-Taravella (Pr, Université Paris Nanterre, codirectrice), M. Gabriel Bergounioux (Pr, Université d'Orléans, codirecteur), M. Olivier Baude (Pr, Université Paris Nanterre, président), M. Thierry Poibeau (DR, CNRS, rapporteur), M. Mathieu Roche (CR HDR, CIRAD, rapporteur).

Résumé : *À l'heure où de plus en plus de corpus et de données sont accessibles, le travail initié s'interroge sur l'exploitation de données linguistiques dans un corpus oral à dimension sociolinguistique avec l'objectif d'en extraire automatiquement du contenu subjectif. À partir de l'exploitation du corpus de transcriptions de français parlé ESLO (enquête sociolinguistique à Orléans), l'objectif est de modéliser, détecter et visualiser la perception qu'ont les locuteurs de la ville d'Orléans. Pour cela, une approche pluridisciplinaire associant la linguistique, le traitement automatique des langues (TAL) et la géographie a été suivie.*

Après avoir discuté la définition du lieu, la première partie de la thèse décrit la méthodologie employée pour la détection de ce type d'information. La démarche proposée suit une approche symbolique fondée sur des lexiques et des règles élaborées grâce à une analyse approfondie des lieux et de leur nommage. Si des normes existent pour le nommage des lieux, tout individu est à même de faire varier cette norme et de se référer aux espaces de son environnement par des moyens détournés. Dans ce travail, les lieux sont identifiés tels qu'ils sont nommés, c'est-à-dire sous la forme utilisée quotidiennement par les gens. Le module développé est évalué et obtient une F-mesure de 0,91 avec un rappel de 0,90 et une précision de 0,93.

La deuxième partie de la thèse apporte un éclairage sur la notion de subjectivité et surtout sur l'articulation des émotions, des sentiments et des opinions avec la notion de perception. À partir de la détection des lieux, les transcriptions sont analysées par apprentissage automatique supervisé afin d'identifier leur caractère subjectif ou objectif ainsi que leur polarité positive ou négative. Les expériences menées ont permis d'entraîner un modèle obtenant une macro-average *de 0,77 pour la tâche de détection de la subjectivité et de 0,76 pour celle de détection de la polarité dans les segments analysés. La détection de la subjectivité et de la polarité oriente l'analyse de la perception, mais ne suffit pas à en rendre compte. Pour aller plus loin, des propositions typologiques sont réalisées au sujet de la cible de la perception et de la manière dont les locuteurs font part de leur perception.*

Afin de confronter les segments subjectifs extraits des transcriptions avec les différents éléments détectés à leur sujet, les résultats obtenus sont projetés dans un système d'information géographique (SIG). Cette visualisation cartographique permet d'avoir une

vision synthétique de l'information, mais aussi de créer de la connaissance en passant du texte à l'image. La troisième et dernière partie décrit les méthodes employées pour la visualisation de la perception de la ville d'Orléans par les locuteurs du corpus ESLO.

Finalement, la carte finale obtenue offre une nouvelle manière d'accéder au corpus ESLO qui se présente comme le portrait sonore de la ville d'Orléans. La matérialisation de ce portrait de la ville d'Orléans ancre d'une part la dimension patrimoniale et anthropologique du corpus, et d'autre part, le témoignage qu'il représente.

URL où le mémoire peut être téléchargé :
http://theses.fr/s270349

Cédric GENDROT : cedric.gendrot@sorbonne-nouvelle.fr

Titre : Traitement automatique et analyse de la variation dans la parole : des mesures phonétiques sur grands corpus aux réseaux de neurones profonds

Mots-clés : phonétique, phonologie, TAL, traitement automatique des langues.

Title: *Automatic Processing and Analysis of Variation in Speech: From Phonetic Measurements on Large Corpora to Deep Neural Networks*

Keywords: *phonetics and phonology, NLP, natural language processing.*

Habilitation à diriger des recherches en sciences du langage, Dynamique du Langage, UMR 5596, Institut des Sciences de l'Homme, Université Lumière Lyon 2, sous la direction de François Pellegrino (DR, CNRS). Habilitation soutenue le 08/07/2021.

Jury : M. François Pellegrino (DR, CNRS, directeur), M. Laurent Besacier (Pr, Université Grenoble Alpes, président), Mme Ann Bradlow (Pr, Northwestern University, Evanston, Illinois, États-Unis, rapporteuse), Mme Corinne Fredouille (MC, Avignon Université, rapporteuse), M. Kim Gerdes (Pr, Université Paris-Saclay, examinateur), Mme Christine Meunier (DR, CNRS, rapporteuse).

Résumé : *Dans ce document d'habilitation à diriger des recherches sont présentées mes activités pédagogiques et académiques, ainsi que mes activités de recherche depuis mon recrutement en tant que maître de conférences à l'Université Sorbonne Nouvelle en 2006. Ce résumé se concentre sur le dernier point en suivant le fil rouge de mes travaux : l'utilisation de grands corpus de parole non préparée pour des analyses phonétiques automatiques afin de mieux comprendre la variation présente dans la parole.*

Dans la première section, après avoir présenté des valeurs formantiques de référence pour le français, j'ai montré des phénomènes de réduction acoustique pour toutes les voyelles en fonction de leur durée phonétique, du contexte consonantique et du style de parole. Cette réduction s'observe également dans plusieurs langues avec des contraintes phonologiques différentes. Il a été démontré au cours de ces travaux que

des mesures effectuées de façon automatique sur des corpus alignés automatiquement restent cohérentes à la condition de respecter certaines précautions méthodologiques.

Dans la deuxième section, les travaux présentés ont mis en évidence l'importance de la prosodie sur la réalisation acoustique des voyelles. La position dans le mot, le syntagme accentuel et le syntagme intonatif sont trois facteurs de variation récurrents que l'on observe en français, en allemand et en espagnol. La comparaison entre trois langues aux systèmes accentuels différents m'a permis de séparer la structure accentuelle et la structure prosodique, pouvant être mises en avant respectivement soit par des informations spectrales (formants) de façon prépondérante, soit par des paramètres prosodiques (f0 et durée).

Dans la troisième section, je me suis appliqué à traiter des phénomènes linguistiques dont la variation soulève des questions sur la séparation entre phonétique et phonologie. J'ai pu montrer dans le cadre de l'analyse du schwa que la prise en compte de multiples facteurs était possible et souhaitable dans de grands corpus. La mise en évidence de variables différentes pour la réduction du schwa par rapport à son élision complète a permis de conclure à des mécanismes différents, l'un phonétique et l'autre phonologique. L'analyse du /R/ français standard d'après une combinaison de corpus de données articulatoires et de grands corpus de parole a permis de considérer la forme non voisée du /R/ comme la réalisation hyper-articulée de la forme voisée, et a montré que la variation du /R/ est grandement influencée par la position prosodique et par le style de parole, en plus du contexte consonantique. Pour finir, dans une étude postulant que /e/ et /ɛ/ sont entrés dans un processus de fusion, j'ai montré que les grands corpus avec de multiples locuteurs sont des outils appropriés pour repérer des tendances globales dans une langue malgré le maintien de variations inter-locuteurs. Ces études ont également été l'occasion de tester perceptivement les variations mesurées et ainsi valider leur pertinence dans le cadre de la communication parlée. Plusieurs aspects méthodologiques fondamentaux ainsi que des méthodes innovantes sont présentés.

Dans la quatrième et dernière section, une discussion est proposée : l'utilisation des grands corpus y est comparée à celle des petits corpus de parole lue. Une remise en question des méthodes tant pour les données que pour les analyses est également avancée et des solutions sont proposées. Mes travaux récents m'ont guidé vers la recherche de stratégies propres au locuteur et de sa caractérisation phonétique. Depuis moins de dix ans, les réseaux de neurones profonds ont bouleversé le domaine de la classification, et il paraissait indispensable d'essayer de les utiliser pour l'analyse phonétique. En ayant recours à des réseaux de neurones convolutifs (CNN) par le biais de spectrogrammes, le but était double : (1) savoir jusqu'à quel point le spectrogramme permet de caractériser le locuteur au-delà d'une analyse phonétique classique et (2) au moyen de techniques de visualisation, parvenir à localiser les zones du

spectrogramme utilisées par les CNN. Des résultats encourageants présentés dans la discussion finale donnent un aperçu de mes projets de recherche.

URL où le mémoire peut être téléchargé :
https://halshs.archives-ouvertes.fr/tel-03303801

Marine WAUQUIER : marine.wauquier@hotmail.fr

Titre : Confrontation des procédés dérivationnels et des catégories sémantiques dans les modèles distributionnels

Mots-clés : sémantique distributionnelle, morphologie, sémantique lexicale, linguistique de corpus, nominalisation.

Title: *Confrontation of Derivational Processes and Semantic Categories in Distributional Semantic Models*

Keywords: *distributional semantics, morphology, lexical semantics, corpus linguistics, nominalization.*

Thèse de doctorat en sciences du langage, CLLE-ERSS, UMR 5263, Université Toulouse 2 - Jean Jaurès, sous la direction de Nabil Hathout (DR, CNRS). Thèse soutenue le 04/12/2020.

Jury : M. Nabil Hathout (DR, CNRS, directeur), M. Olivier Bonami (Pr, Université de Paris, rapporteur), M. Ingo Plag (Pr, Université de Düsseldorf, Allemagne, rapporteur), Mme Fiammetta Namer (Pr, Université de Lorraine, présidente), M. Laurent Prévot (Pr, Université d'Aix-Marseille, examinateur), Mme Cécile Fabre (Pr, Université Toulouse 2 - Jean Jaurès, examinatrice).

Résumé : *La forme et le sens sont intimement liés en morphologie dérivationnelle, l'affixe d'un dérivé renseignant généralement sur son appartenance à une catégorie sémantique donnée. Cette relation entre affixes et catégories sémantiques n'est cependant pas exclusive, et est étudiée à partir de facteurs phonologiques, syntaxiques, ou encore sémantiques. Ces derniers sont sans doute parmi les facteurs les plus difficiles à évaluer empiriquement, et ont longtemps reposé sur une approche intuitive.*

La sémantique distributionnelle se révèle depuis quelques années comme une des alternatives les plus populaires. Il s'agit d'une approche statistique du sens basée sur les usages en corpus, qui offre une représentation vectorielle du sens des mots. La quantification de la proximité sémantique des mots et la manipulation des représentations permises par les modèles distributionnels ouvrent de nouvelles perspectives sur l'analyse sémantique de la concurrence affixale.

Nous mettons à profit dans cette thèse les modèles distributionnels pour analyser des dérivés morphologiques au regard de ces relations many-to-many*, selon quatre axes. Dans un premier temps, nous quantifions la proximité sémantique entre membres de familles dérivationnelles à l'aide de la proximité distributionnelle dans les espaces*

vectoriels, validant à grande échelle l'hypothèse d'une plus grande proximité du verbe et du nom d'action. Dans un second temps, nous étayons les différences sémantiques entre les noms en -eur, -euse *et* -rice *relatives aux propriétés axiologiques de leurs référents, en comparant les représentations globales de ces trois classes. Dans un troisième temps, nous évaluons l'hétérogénéité morphologique et sémantique de la catégorie lexicale des noms d'agent à partir de l'analyse de la représentation globale de ses représentants prototypiques. Enfin, nous explorons la différenciation sémantique des noms d'action en* -age, -ion *et* -ment, *au regard de leur degré de technicité. Nous combinons des indices distributionnels et statistiques afin de modéliser cette différence de technicité.*

Au travers de ces quatre questions, cette thèse présente différents degrés d'adaptation des modèles distributionnels pour l'analyse linguistique, en tant qu'outil de validation et d'exploration. Nous proposons à ce titre une exploration méthodologique visant à illustrer le potentiel, mais aussi les limites de l'utilisation des modèles distributionnels en linguistique.

URL où le mémoire peut être téléchargé :
http://www.theses.fr/s196917
