Résumés de thèses et HDR

Rubrique préparée par Sylvain Pogodalla

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

Ygor GALLINA : ygor.gallina@univ-nantes.fr

Titre : Indexation de bout-en-bout dans les bibliothèques numériques scientifiques

Mots-clés : indexation automatique, mots-clés, évaluation extrinsèque, recherche d'information, génération de mots-clés, méthodes de bout en bout.

Title: *End-to-End Indexation in Digital Scientific Libraries*

Keywords: *automatic indexing, keywords, extrinsic evaluation, information retrieval, keyword generation, end-to-end method.*

Thèse de doctorat en informatique, Laboratoire des Sciences du Numérique de Nantes, UMR 6004, UFR Sciences et Techniques, Université de Nantes, sous la direction de Béatrice Daille (Pr, Université de Nantes) et Florian Boudin (MC, Université de Nantes). Thèse soutenue le 28/03/2022.

Jury : Mme Béatrice Daille (Pr, Université de Nantes, codirectrice), M. Florian Boudin (MC, Université de Nantes, codirecteur), M. Richard Dufour (Pr, Université de Nantes, président), Mme Josiane Mothe (Pr, Université de Toulouse, rapporteuse), M. Patrick Paroubek (IR, CNRS, rapporteur), Mme Lorraine Goeuriot (MC, Université Grenoble Alpes, examinatrice).

Résumé : *Le nombre de documents scientifiques dans les bibliothèques numériques ne cesse d'augmenter. Les mots-clés permettant d'enrichir l'indexation de ces documents ne peuvent être annotés manuellement étant donné le volume de documents à traiter. La production automatique de mots-clés est donc un enjeu important. Le cadre évaluatif le plus utilisé pour cette tâche souffre de nombreuses faiblesses qui rendent l'évaluation des nouvelles méthodes neuronales peu fiables. Notre objectif est d'identifier précisément ces faiblesses et d'y apporter des solutions selon trois axes. Dans*

un premier temps, nous introduisons KPTimes*, un jeu de données du domaine journalistique. Il nous permet d'analyser la capacité de généralisation des méthodes neuronales. De manière surprenante, nos expériences montrent que le modèle le moins performant est celui qui généralise le mieux. Dans un deuxième temps, nous effectuons une comparaison systématique des méthodes états de l'art grâce à un cadre expérimental strict. Cette comparaison indique que les méthodes de référence comme TF#IDF sont toujours compétitives et que la qualité des mots-clés de référence a un impact fort sur la fiabilité de l'évaluation. Enfin, nous présentons un nouveau protocole d'évaluation extrinsèque basé sur la recherche d'information. Il nous permet d'évaluer l'utilité des mots-clés, une question peu abordée jusqu'à présent. Cette évaluation nous permet de mieux identifier les mots-clés importants pour la tâche de production automatique de mots-clés et d'orienter les futurs travaux.*

URL où le mémoire peut être téléchargé :
https://hal.archives-ouvertes.fr/tel-03667015

Timothee MICKUS : timothee.mickus@univ-lorraine.fr

Titre : Du statut des plongements lexicaux en tant qu'implémentations de l'hypothèse distributionnelle

Mots-clés : TAL, génération automatique de texte, sémantique distributionnelle, plongements lexicaux, génération de définition, lexicographie.

Title: *On the Status of Word Embeddings as Implementations of the Distributional Hypothesis*

Keywords: *NLP, NLG, distributional semantics, word embeddings, definition modeling, lexicography.*

Thèse de doctorat en informatique, ATILF, UMR 7118, Université de Lorraine, sous la direction de Mathieu Constant (Pr, Université de Lorraine, ATILF, UMR 7118) et Denis Paperno (universitair docent, Universiteit Utrecht, Pays-Bas). Thèse soutenue le 31/03/2022.

Jury : M. Mathieu Constant (Pr, Université de Lorraine, ATILF, UMR 7118, codirecteur), M. Denis Paperno (universitair docent, Universiteit Utrecht, Pays-Bas, codirecteur), M. Benoît Crabbé (Pr, Université de Paris, rapporteur, président), M. Nabil Hathout (DR, CNRS, rapporteur), Mme Gemma Boleda (Research Professor, Universitat Pompeu Fabra, Barcelone, Espagne, examinatrice), Mme Vera Demberg (Pr, Universität des Saarlandes, Sarrebruck, Allemagne, examinatrice), Mme Claire Gardent (DR, CNRS, examinatrice), M. Alessandro Lenci (Pr, Università di Pisa, Italie, examinateur), M. Kees van Deemter (Pr, Universiteit Utrecht, Pays-Bas, examinateur).

Résumé : *Cette thèse s'intéresse au statut des plongements lexicaux (ou «* word embeddings *»), c'est-à-dire des vecteurs de mots issus de modèles de traitement automatique des langues. Plus particulièrement, notre intérêt se porte sur leur valeur linguis-*

tique et la relation qu'ils entretiennent avec la sémantique distributionnelle, le champ d'études fondé sur l'hypothèse que le contexte est corrélé au sens. L'objet de notre recherche est d'établir si ces plongements lexicaux peuvent être considérés comme une implémentation concrète de la sémantique distributionnelle.

Notre première approche dans cette étude consiste à comparer les plongements lexicaux à d'autres représentations du sens, en particulier aux définitions telles qu'on en trouve dans des dictionnaires. Cette démarche se fonde sur l'hypothèse que des représentations sémantiques de deux formalismes distincts devraient être équivalentes, et que par conséquent l'information encodée dans les représentations sémantiques distributionnelles devrait être équivalente à celle encodée dans les définitions. Nous mettons cette idée à l'épreuve à travers deux protocoles expérimentaux distincts : le premier est basé sur la similarité globale des espaces métrisables décrits par les vecteurs de mots et les définitions, le second repose sur des réseaux de neurones profonds. Dans les deux cas, nous n'obtenons qu'un succès limité, ce qui suggère soit que la sémantique distributionnelle et les dictionnaires encodent des informations différentes, soit que les plongements lexicaux ne sont pas motivés d'un point de vue linguistique.

Le second angle que nous adoptons ici pour étudier le rapport entre sémantique distributionnelle et plongements lexicaux consiste à formellement définir ce que nous attendons des représentations sémantiques distributionnelles, puis à comparer nos attentes à ce que nous observons effectivement dans les plongements lexicaux. Nous construisons un jeu de données de jugements humains sur l'hypothèse distributionnelle. Nous utilisons ensuite ce jeu pour obtenir des prédictions sur une tâche de substituabilité distributionnelle à partir de modèles de plongements lexicaux. Bien que nous observions un certain degré de performance en utilisant les modèles en question, leur comportement se démarque très clairement de celui de nos annotateurs humains. Venant renforcer ces résultats, nous remarquons qu'une large famille de modèles de plongements qui ont rencontré un franc succès, ceux basés sur l'architecture Transformer*, présente des artéfacts directement imputables à l'architecture qu'elle emploie plutôt qu'à des facteurs d'ordre sémantique.*

Nos expériences suggèrent que la validité linguistique des plongements lexicaux n'est aujourd'hui pas un problème résolu. Trois grandes conclusions se dégagent de nos expériences. Premièrement, la diversité des approches en sémantique distributionnelle n'implique pas que ce champ d'études est voué aux approches informelles : nous avons vu que le linguiste peut s'appuyer sur la substituabilité distributionnelle. Deuxièmement, comme on ne peut pas aisément comparer la sémantique distributionnelle à une autre théorie lexicale, il devient nécessaire d'étudier si la sémantique distributionnelle s'intéresse bien au sens, ou bien si elle porte sur une série de faits entièrement distincte. Troisièmement, bien que l'on puisse souligner une différence entre la qualité des plongements lexicaux et ce qu'on attend qu'ils puissent faire, la

possibilité d'étudier cette différence sous un angle quantitatif est de très bon augure pour les travaux à venir.

URL où le mémoire peut être téléchargé :
https://tel.archives-ouvertes.fr/tel-03723503

Pedro ORTIZ SUAREZ : pedro@portizs.eu

Titre : Une approche basée sur les données pour le traitement automatique du langage naturel en français contemporain et historique

Mots-clés : modèle de langue, corpus de pré-entraînement, traitement automatique des langues, français contemporain, français historique, apprentissage par transfert.

Title: *A Data-driven Approach to Natural Language Processing for Contemporary and Historical French*

Keywords: *language model, pre-training corpora, natural language processing, contemporary French, historical French, transfer learning.*

Thèse de doctorat en informatique, ALMAnach, Centre de Recherche Inria de Paris, Sorbonne Université, sous la direction de Laurent Romary (DR, Inria) et Benoît Sagot (DR, Inria). Thèse soutenue le 27/06/2022.

Jury : M. Laurent Romary (DR, Inria, codirecteur), M. Benoît Sagot (DR, Inria, codirecteur), M. Francis Bach (DR, Inria, président), Mme Maud Ehrmann (MC, École polytechnique fédérale de Lausanne, Suisse, examinatrice), M. Alexander Geyken (DR, Berlin-Brandenburgischen Akademie der Wissenschaften, Allemagne, examinateur), Mme Anna Korhonen (DR, University of Cambridge, Royaume-Uni, rapporteuse), M. Holger (DR, Meta AI Research, rapporteur).

Résumé : *Depuis plusieurs années, les approches neuronales ont régulièrement amélioré l'état de l'art du traitement automatique des langues (TAL) sur une grande variété de tâches. L'un des principaux facteurs ayant permis ces progrès continus est l'utilisation de techniques d'apprentissage par transfert. Ces méthodes consistent à partir d'un modèle pré-entraîné et à le réutiliser, avec peu ou pas d'entraînements supplémentaires, pour traiter d'autres tâches. Même si ces modèles présentent des avantages évidents, leur principal inconvénient est la quantité de données nécessaire pour les pré-entraîner. Ainsi, le manque de données disponibles à grande échelle a freiné le développement de tels modèles pour le français contemporain et a fortiori pour ses états de langue plus anciens.*

Cette thèse met l'accent sur le développement de corpus pour le pré-entraînement de telles architectures. Cette approche s'avère extrêmement efficace, car nous sommes en mesure d'améliorer l'état de l'art pour un large éventail de tâches de TAL pour le français contemporain et historique, ainsi que pour six autres langues contemporaines. De plus, nous montrons que ces modèles sont extrêmement sensibles à la qualité, à l'hé-

térogénéité et à l'équilibre des données de pré-entraînement et montrons que ces trois caractéristiques sont de meilleurs prédicteurs de la performance des modèles que la taille des données de pré-entraînement. Nous montrons également que l'importance de la taille des données de pré-entraînement a été surestimée en démontrant à plusieurs reprises que l'on peut pré-entraîner de tels modèles avec des corpus de taille assez modeste.

URL où le mémoire peut être téléchargé :
https://tel.archives-ouvertes.fr/tel-03770337

Léon-Paul SCHAUB : lp.schaub@gmail.com

Titre : Dimensions mémorielles de l'interaction écrite humain-machine : une approche cognitive par les modèles mnémoniques pour la détection et la correction des incohérences du système dans les dialogues orientés-tâche

Mots-clés : système de dialogue orienté tâche, modèle de mémoire, réseaux de neurones, détection d'incohérences, interaction homme-machine.

Title: *Memory Dimensions of Human-Machine Written Interaction: a Cognitive Approach Using Mnemonic Models for the Detection and Correction of System Inconsistencies in Task-oriented Dialogues*

Keywords: *task-oriented dialogue system, memory model, neural networks, inconsistency detection.*

Thèse de doctorat en informatique, Laboratoire Indisciplinaire des Sciences du Numérique, Université Paris-Saclay, sous la direction de Patrick Paroubek (IR, CNRS). Thèse soutenue le 23/03/2022.

Jury : M. Patrick Paroubek (IR, CNRS, directeur), M. Frédéric Landragin (DR, CNRS, rapporteur, président), Mme Chloé Clavel (Pr, Télécom ParisTech, rapporteuse), M. Yves Lepage (Pr, Waseda Université, Tokyo, Japon, examinateur), Mme Magalie Ochs (MC, Université Aix-Marseille, examinatrice), M. Frédéric Béchet (Pr, Université Aix-Marseille, examinateur).

Résumé : *Dans ce travail, nous nous intéressons aux systèmes de dialogue orientés tâche. Nous nous concentrons sur la différence de traitement de l'information et de l'utilisation de la mémoire, d'un tour de parole à l'autre, par l'humain et la machine, pendant une conversation écrite. Après avoir étudié les mécanismes de rétention et de rappel chez l'humain en dialogue, nous émettons l'hypothèse qu'un des éléments susceptibles d'expliquer les moindres performances des machines par rapport aux humains est leur incapacité à posséder une image de l'utilisateur, mais également une image de soi, explicitement convoquée pendant les inférences liées à la poursuite du dialogue. Améliorer la machine se traduit par trois axes. Tout d'abord, par l'anticipation, à un tour de parole, de l'énoncé suivant de l'utilisateur. Ensuite, par la détection d'une incohérence dans son propre énoncé, facilitée par l'anticipation du tour suivant*

de l'utilisateur en tant qu'indice supplémentaire. Enfin, par la prévision du nombre de tours de paroles restant dans le dialogue afin d'avoir une meilleure vision de la progression du dialogue, en prenant en compte la présence d'une incohérence dans son propre énoncé. Pour les mettre en place, nous exploitons les réseaux de mémoire de bout en bout, un modèle de réseau de neurones récurrent qui possède la spécificité de créer des sauts de réflexion, permettant de filtrer l'information contenue à la fois dans l'énoncé de l'utilisateur et dans celui de l'historique de dialogue. De plus, ces trois sauts de réflexion servent de mécanisme d'attention « naturel » pour le réseau de mémoire, à la manière d'un décodeur de transformeur. Pour notre étude, nous améliorons, en y ajoutant nos trois fonctionnalités, un type de réseau de mémoire appelé WMM2Seq. *Ce modèle s'inspire des modèles cognitifs de la mémoire, en présentant les concepts de mémoire épisodique, sémantique et de travail. Il obtient des résultats performants sur des tâches de génération de réponses de dialogue sur les corpus* DSTC2 *et* MultiWOZ *qui sont les corpus que nous utilisons pour nos expériences. Les trois axes mentionnés apportent deux contributions principales à l'existant. En premier lieu, ils complexifient l'intelligence du système de dialogue en le dotant d'un garde-fou. En second lieu, ils optimisent à la fois le traitement des informations dans le dialogue et la durée de celui-ci. Les résultats obtenus avec nos différentes mesures d'évaluation montrent l'intérêt d'orienter les recherches vers des modèles de gestion de la mémoire plus cognitifs afin de réduire l'écart de performance dans un dialogue entre l'humain et la machine.*

URL où le mémoire peut être téléchargé :
https://hal.archives-ouvertes.fr/tel-03647756

Antoine SIMOULIN : antoine.simoulin@gmail.com

Titre : Plongements de phrases et leurs relations avec les structures de phrases

Mots-clés : traitement automatique des langues naturelles, plongements de phrases, apprentissage profond, réseaux de neurones structurés.

Title: *Sentence Embeddings and Their Relation with Sentence Structures*

Keywords: *natural language processing, sentence embeddings, deep learning, structured neural networks.*

Thèse de doctorat en informatique, Laboratoire de Linguistique Formelle, école doctorale Sciences Mathématiques de Paris Centre, Université Paris Cité, sous la direction de Benoît Crabbé (Pr, Université de Paris). Thèse soutenue le 07/07/2022.

Jury : M. Benoît Crabbé (Pr, Université de Paris, directeur), Mme Claire Gardent (DR, CNRS, rapporteuse), M. Éric Gaussier (Pr, Université Grenoble Alpes, rapporteur), Mme Rachel Bawden (CR, Inria, examinatrice), M. Loïc Barrault (MC, Le Mans Université, examinateur), M. Nicolas Brunel (Pr, ENSIEE, Laboratoire de Mathématiques et Modélisation d'Évry, examinateur).

Résumé : *Historiquement, la modélisation du langage humain suppose que les phrases ont une structure symbolique et que cette structure permet d'en calculer le sens par composition. Ces dernières années, les modèles d'apprentissage profond sont parvenus à traiter automatiquement des tâches sans s'appuyer sur une structure explicite du langage, remettant ainsi en question cette hypothèse fondamentale. Cette thèse cherche ainsi à mieux identifier le rôle de la structure lors de la modélisation du langage par des modèles d'apprentissage profond. Elle se place dans le cadre spécifique de la construction de plongements de phrases — des représentations sémantiques basées sur des vecteurs — par des réseaux de neurones profonds.*

Dans un premier temps, on étudie l'intégration de biais linguistiques dans les architectures de réseaux neuronaux, pour contraindre leur séquence de composition selon une structure traditionnelle, en arbres. Dans un second temps, on relâche ces contraintes pour analyser les structures latentes induites par ces réseaux neuronaux. Dans les deux cas, on analyse les propriétés de composition des modèles ainsi que les propriétés sémantiques des plongements.

La thèse s'ouvre sur un état de l'art présentant les principales méthodes de représentation du sens des phrases, qu'elles soient symboliques ou basées sur des méthodes d'apprentissage profond. La deuxième partie propose plusieurs expériences introduisant des biais linguistiques dans les architectures des réseaux de neurones pour construire des plongements de phrases. Le premier chapitre combine explicitement plusieurs structures de phrases pour construire des représentations sémantiques. Le deuxième chapitre apprend conjointement des structures symboliques et des représentations vectorielles. Le troisième chapitre introduit un cadre formel pour les transformer selon une structure de graphes. Finalement, le quatrième chapitre étudie l'impact de la structure vis-à-vis de la capacité de généralisation et de composition des modèles.

La thèse se termine par une mise en concurrence de ces approches avec des méthodes de passage à l'échelle. On cherche à y discuter les tendances actuelles qui privilégient des modèles plus gros, plus facilement parallélisables et entraînés sur plus de données, aux dépens de modélisations plus fines. Les deux chapitres de cette partie relatent l'entraînement de larges modèles de traitement automatique du langage et comparent ces approches avec celles développées dans la deuxième partie d'un point de vue qualitatif et quantitatif.

URL où le mémoire peut être téléchargé :
https://hal.archives-ouvertes.fr/tel-03791935

Chunxiao YAN : yanchunxiao5597@gmail.com

Titre : Complexité syntaxique et flux de dépendance. Études quantitatives dans les *treebanks* Universal Dependencies

Mots-clés : flux de dépendance, syntaxe de dépendance, complexité syntaxique, métrique, mémoire de travail, *Universal Dependencies treebanks*.

Title: *Syntactic Complexity and Dependency Flux. Quantitative Studies in Universal Dependencies Treebanks*

Keywords: *dependency flux, dependency syntax, syntactic complexity, metrics, working memory, Universal Dependencies treebanks.*

Thèse de doctorat en sciences du langage, MoDyCo, Université Paris Nanterre, sous la direction de Sylvain Kahane (Pr, Université Paris Nanterre). Thèse soutenue le 01/12/2021.

Jury : M. Sylvain Kahane (Pr, Université Paris Nanterre, directeur), M. François Lareau (Pr, Université de Montréal, Canada, rapporteur), M. Philippe Blache (DR, CNRS, rapporteur), Mme Marie Candito (MC, Université Paris-Diderot, examinatrice), Mme Marie-Catherine de Marneffe (Pr, The Ohio State University, Columbus, États-Unis, examinatrice), M. Kim Gerdes (Pr, Université Paris-Saclay, examinateur).

Résumé : *Nous nous intéressons à la complexité syntaxique et aux contraintes liées à la mémoire de travail chez l'humain. La mémoire de travail concerne non seulement la capacité de retenir des informations, mais aussi la capacité de les manipuler temporairement. Elle a été montrée limitée à 7 ± 2 éléments et est aujourd'hui actualisée autour de 4 selon Cowan. La limitation de la mémoire de travail peut rendre le traitement de certaines structures de phrase difficile, voire impossible. Dans cette thèse, nous nous penchons sur trois pistes d'étude : étudier et mesurer la complexité syntaxique sous différentes hypothèses cognitives, savoir s'il existe des limites à la complexité syntaxique dans les langues naturelles, et comprendre les phénomènes impliqués par les contraintes sur la complexité syntaxique.*

De ce fait, nous mesurons la complexité syntaxique en utilisant des métriques basées sur le flux de dépendance dans le corpus (les treebanks Universal Dependencies*). Ces métriques incluent non seulement des métriques devenues classiques comme la longueur de dépendance, des métriques proposées dans des travaux plus récents, mais aussi de nouvelles métriques également basées sur le flux de dépendance.*

En nous basant sur les résultats donnés par ces différentes métriques dans les plus de 100 langues appartenant à la collection des treebanks Universal Dependencies*, nous pouvons déterminer celles qui sont les plus appropriées pour étudier la complexité syntaxique. Nous montrons qu'il existe pour certaines métriques du flux des contraintes universelles, dont nous postulons qu'elles sont liées à la mémoire de tra-*

vail. Enfin, nous essayons également d'expliquer certains des phénomènes linguistiques observés dans nos données qui impliquent la complexité syntaxique.

URL où le mémoire peut être téléchargé :
https://tel.archives-ouvertes.fr/tel-03649621
