Contrat post-doctoral : Polysemic embeddings.

Laboratoire d'accueil : LIUM, équipe LST, `https://lium.univ-lemans.fr/lium/lst/`
Partenaire industriel : SNCF Innovation Recherche
Site : Le Mans
Encadrement : Nicolas Dugué (LIUM)
Co-encadrement : Nathalie Camelin (LIUM), Luce Lefeuvre (SNCF)
Durée : Contrat post-doctoral d'un an, début en septembre si possible

1 Contexte

Le LIUM termine actuellement un projet de collaboration avec la Direction Innovation et Recherche de SNCF autour de la structuration d'un corpus de documents en thématiques. Les ressources lexicales fournies par SNCF ont permis au LIUM de prendre connaissance de la richesse et des spécificités du vocabulaire métier utilisé au sein d'entreprises telles que SNCF. Ce vocabulaire est parfois peu fréquent dans les corpus mais d'après les experts, très important pour caractériser les documents. Par ailleurs, ce vocabulaire contient des acronymes qui, pour environ 40% ne servent pas d'abréviations aux mêmes groupes de mots. Le corpus de ce projet nous a permis de mettre en lumière trois verrous scientifiques majeurs pour le traitement automatique efficace de ce type de documents en utilisant les plongements lexicaux :

- Comment apprendre des plongements de bonne qualité pour du vocabulaire spécifique parfois peu fréquent ?
- Comment apprendre des plongements pour des acronymes spécifiques ET polysémiques ?
- Comment évaluer les plongements appris ?

Le premier verrou scientifique est relatif à l'apprentissage de plongements lexicaux en langue de spécialité. C'est un problème difficile qui, à notre connaissance, peut être approché soit via des modèles capables de prendre en compte efficacement les basses fréquences (Levy et al. 2015), soit via la production de connaissances (thésaurus, ontologies) de façon à limiter la taille du vocabulaire spécifique à apprendre et à mutualiser les fréquences (Perinet, 2015), soit en utilisant des ressources capables de guider l'apprentissage (Tissier et al. 2017). Dans notre cas, SNCF dispose de ressources produites par des experts : lexiques et dictionnaires d'acronymes. Nous proposons donc d'écrire un modèle capable de tirer parti de ces ressources particulières pour guider l'apprentissage de plongements de bonne qualité pour ce vocabulaire spécialisé.

Nous pensons par ailleurs que cette approche, enrichie par une approche multi-prototypique telle que dans Tian et al. (2014) peut également permettre de résoudre le second verrou. Dans ce genre d'approches, il s'agit d'apprendre un vecteur différent pour chaque sens d'un mot, chaque vecteur étant un *prototype*. Dans notre cas, le nombre de prototypes à apprendre pour chaque acronyme correspond au nombre de définitions présentes dans le dictionnaire d'acronymes, et chaque définition pourra être utilisée pour apprendre les prototypes qui leur correspondent. Le modèle que nous souhaitons proposer aura en particulier la capacité de réaliser la désambiguïsation en même temps que l'apprentissage des prototypes.

Enfin, le dernier problème concerne l'évaluation des modèles appris sur ces corpus. Pour cela, SNCF nous permet d'accéder à des experts métier capables de réaliser des tâches d'annotation ou d'évaluation. Nous souhaitons nous baser sur des travaux préliminaires qui nous ont permis de rendre compte de la difficulté de ce travail et d'ouvrir des pistes (Dugué et al. 2019). Pour nous aider à formaliser ce problème d'évaluation, nous pourrons faire appel à Jane Wottawa, linguiste experte des tests de perception.

2 Profil idéalement recherché

Nous recherchons un.e jeune docteur.e (ou quelqu'un prêt à soutenir) en informatique, spécialisé.e dans l'apprentissage automatique via des méthodes statistiques, habitué.e à travailler avec des données textuelles. En particulier, un.e candidat.e idéal.e aurait déjà expérimenté des modèles de plongements lexicaux. Nous cherchons également un.e bon.ne programmeu.r.se Python capable de produire une librairie lisible et réutilisable. Enfin, le/la candidat.e doit très bien maîtriser la langue française, puisque les documents et le vocabulaire sont en français et qu'il s'agira de pouvoir analyser les résultats dans ce contexte.

Candidater : Contacter `nicolas.dugue@univ-lemans.fr`, `nathalie.camelin@univ-lemans.fr` avec pour sujet de mail "[PolysEmY] Candidature", et joindre un CV et une lettre de motivation. Si vous avez soutenu votre thèse de doctorat, joindre les rapports.

3 État de l'art

Firth définit l'hypothèse distributionnelle en 1954 selon laquelle le sens d'un mot est décrit par ses co-occurrences. Depuis, de nombreux travaux informatiques et linguistiques se basent sur cette assertion pour apprendre des vecteurs capables d'encoder la sémantique du vocabulaire en utilisant de grands corpus de données textuelles. Ces vecteurs, qu'on appelle plongements lexicaux ont montré leur efficacité pour de nombreuses tâches d'apprentissage automatique : Analyse de sentiments [10], traduction automatique [6], plongement de documents [5], reconnaissance de la parole [2], *etc.*

Les modèles d'apprentissage de plongements lexicaux ont usuellement pour objectif d'apprendre un espace de représentation en faible dimension, continu et dans lequel les mots similaires en sens dans la langue naturelle sont proches. Certaines méthodes se basent sur une factorisation de la matrice de co-occurrences termes-termes [11, 16]. D'autres méthodes considèrent une tâche supervisée dans laquelle il s'agit d'apprendre un espace dans lequel les mots qui co-occurrent sont proches en terme de distance cosinus [14, 13, 3]. L'une des faiblesses de ces modèles appelés word embeddings en anglais, est qu'ils aboutissent à l'apprentissage d'un vecteur par mot. Or, une grande partie du vocabulaire est polysémique [7]. Les différents sens d'un mot se trouvent ainsi encapsulés dans un unique et même vecteur. Pourtant, Li et Jurafsky [12] ont montré qu'il est important de tenir compte de la polysémie pour certaines tâches telles que la similarité sémantique entre mots, phrases ou documents. Ces tâches sont particulièrement pertinentes dans le contexte du projet, où nous souhaitons améliorer la représentation des contenus pour faciliter la recherche d'information.

Pour leur étude sur l'impact de la polysémie sur les tâches classiques en traitement automatique du langage, Li et Jurasfy [12] s'appuient sur des modèles existants. Ces modèles sont dits multi-prototype ou *multi-sense* en anglais. Il s'agit d'obtenir un vecteur de plongement par sens, ces modèles sont donc capables de tenir en partie compte de la polysémie. Ces modèles qui présentent un certain nombre de faiblesses, notamment dans notre cas, fournissent une base solide sur laquelle nous appuyer pour construire notre approche. Ainsi, les travaux de Reisinger et Mooney [17] et de Huang et al. [8] sont les premiers à notre connaissance à traiter la polysémie : ils utilisent l'algorithme des K-moyennes pour

décomposer les vecteurs en plusieurs prototypes censés représenter un sens différent pour les mots. Des travaux plus récents font une hypothèse proche de celle-ci, ils considèrent chaque mot comme un mélange sur K vecteurs de sens [1]. Neelakantan et al. [15] considèrent également un mélange de K sens, l'optimisation de chacun des vecteurs de sens est alors faite en fonction des contextes les plus probables pour chacun de ces sens. C'est également avec une modélisation probabiliste comme celle-ci que Tian et al. [18] proposent d'apprendre K vecteurs par mots. Dans chacune de ces approches, le modèle Skip-gram [13] est considéré comme une base, et K le nombre de sens est un paramètre du modèle. D'autres travaux utilisent des bases de connaissances telles que Wordnet ou Babelnet pour inférer le nombre de sens [9]. Ces modèles proposent parfois de faire la désambiguïsation des sens dans le corpus en même temps que l'apprentissage [4].

Nous souhaitons nous baser sur ces approches pour définir une tâche combinant la désambiguïsation des acronymes dans le corpus et l'apprentissage de vecteurs pour chacun des sens. En revanche, nous ne disposons que d'une base d'acronymes limitée : moins riche que Wordnet en vocabulaire, et sans organisation sémantique. De plus, nos corpus sont d'une taille bien plus limitée : moins de 10.000 documents dans notre cas d'étude réaliste. Ainsi, même s'il est envisageable d'utiliser des modèles basés sur Skip-gram, des approches adaptées aux corpus de petite taille seront considérées. L'apprentissage sera guidé par les dictionnaires d'acronymes SNCF qui fournissent une supervision distante. Enfin, l'état de l'art propose une évaluation des représentations en utilisant des tâches de classification. Nous souhaitons produire un processus d'évaluation experte adaptée à cette tâche et aux besoins d'industrialisation de ce type de solutions.

References

[1] Ben Athiwaratkun, Andrew Gordon Wilson, and Anima Anandkumar. Probabilistic fasttext for multi-sense word embeddings. *arXiv preprint arXiv:1806.02901*, 2018.

[2] Samy Bengio and Georg Heigold. Word embeddings for speech recognition. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[3] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016.

[4] Xinxiong Chen, Zhiyuan Liu, and Maosong Sun. A Unified Model for Word Sense Representation and Disambiguation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1025–1035, Doha, Qatar, 2014. Association for Computational Linguistics.

[5] Andrew M. Dai, Christopher Olah, and Quoc V. Le. Document Embedding with Paragraph Vectors. *arXiv:1507.07998 [cs]*, July 2015. arXiv: 1507.07998.

[6] Mercedes García-Martínez, Ozan Caglayan, Walid Aransa, Adrien Bardet, Fethi Bougares, and Loïc Barrault. LIUM Machine Translation Systems for WMT17 News Translation Task. *arXiv:1707.04499 [cs]*, July 2017. arXiv: 1707.04499.

[7] William L. Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. *arXiv:1605.09096 [cs]*, May 2016. arXiv: 1605.09096.

[8] Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 873–882. Association for Computational Linguistics, 2012.

[9] Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. SensEmbed: Learning Sense Embeddings for Word and Relational Similarity. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 95–105, Beijing, China, 2015. Association for Computational Linguistics.

[10] Yoon Kim. Convolutional Neural Networks for Sentence Classification. *arXiv:1408.5882 [cs]*, August 2014. arXiv: 1408.5882.

[11] Omer Levy, Yoav Goldberg, and Ido Dagan. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225, 2015.

[12] Jiwei Li and Dan Jurafsky. Do multi-sense embeddings improve natural language understanding? *arXiv preprint arXiv:1506.01070*, 2015.

[13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[14] Andriy Mnih and Geoffrey E Hinton. A scalable hierarchical distributed language model. In *Advances in neural information processing systems*, pages 1081–1088, 2009.

[15] Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. Efficient non-parametric estimation of multiple embeddings per word in vector space. *arXiv preprint arXiv:1504.06654*, 2015.

[16] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[17] Joseph Reisinger and Raymond J Mooney. Multi-prototype vector-space models of word meaning. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 109–117. Association for Computational Linguistics, 2010.

[18] Fei Tian, Hanjun Dai, Jiang Bian, Bin Gao, Rui Zhang, Enhong Chen, and Tie-Yan Liu. A probabilistic model for learning multi-prototype word embeddings. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 151–160, 2014.