Création et validation de représentations vectorielles pour des relations lexico-sémantiques : application à l'identification/classification de relations à partir d'un corpus métier

Camille Gosset[*,**], Mokhtar Boumedyen Billami[*], Mathieu Lafourcade[**], Christophe Bortolaso[*], Mustapha Derras[*]

[*]Berger-Levrault, Labège, France
{camille.gosset , mb.billami , christophe.bortolaso, mustapha.derras}@berger-levrault.com
[**]LIRMM, Université Montpellier, Montpellier, France
{gosset,mathieu.lafourcade}@lirmm.fr

1 Introduction

Les méthodes de représentation vectorielle d'éléments lexicaux se sont avérées être importantes pour résoudre des tâches liées au traitement automatique du langage naturel. Dans cet article, nous nous intéressons à réaliser une étude comparative sur la qualité de la représentation vectorielle des relations lexico-sémantiques et cela pour deux systèmes : (1) le nôtre où des vecteurs embeddings de relations sont déduits à partir de vecteurs embeddings de termes et (2) RELATIVE (Camacho-Collados et al., 2019), un système état-de-l'art où des vecteurs embeddings de relations sont plutôt appris.

Nous proposons des vecteurs embeddings pour des termes-clés (multi-mots et expressions, p.ex. « *Projet et construction soumis à enquête publique* ») et non seulement pour des mots singuliers (p.ex. « *caravanage* »). Le corpus de données que nous utilisons provient de secteurs métiers qu'exerce la société Berger-Levrault. Les représentations vectorielles que nous créons ont été validées par le réseau lexical JeuxDeMots (Lafourcade, 2007) dont le français est la langue étudiée dans notre corpus.

2 Création de vecteurs pour des relations lexico-sémantiques

Notre corpus est édité par la société Berger-Levrault. Ce corpus a été annoté en termes-clés par un ensemble d'experts provenant de différents domaines de spécialité. Ainsi, les termes-clés constituant les relations proviennent de ces annotations. Un modèle Word2Vec (Mikolov et al., 2013) est entraîné sur notre corpus. À partir des vecteurs embeddings entraînés, nous pouvons en déduire des représentations vectorielles de relations dites typées. Pour cela, nous produisons un ensemble de paires de termes-clés liés par une relation donnée extraite depuis le réseau JeuxDeMots. Ensuite, par une utilisation d'une opération arithmétique, nous déduisons

une représentation vectorielle pour une paire de termes (typée). Pour les relations asymétriques, nous effectuons une soustraction entre les vecteurs des deux termes-clés d'une relation pour maintenir le sens de la relation tandis que pour les relations symétriques, nous prenons la valeur absolue de cette différence car le sens est bidirectionnel.

3 Évaluation

Nous nous sommes comparés à RELATIVE (*RELations as LATent dIscourse VEctors*). En prenant en considération un corpus textuel brut et un modèle de plongements de termes, RELATIVE apprend des plongements de paires de termes. Nous utilisons des classificateurs binaires pour valider la qualité des représentations vectorielles de relations typées. Cette phase de classification utilise plusieurs algorithmes (k-PPV, SVC, RFC et DT) dont chacun est employé sur deux types de relations, à savoir : (*Hyperonymie, Hyponymie*) et (*Synonymie, Antonymie*).

L'apprentissage et l'évaluation de la qualité des représentations de relations se sont nourris du réseau JeuxDeMots. En effet, nous utilisons des paires de termes-clés validés par ce réseau pour réaliser l'entraînement des classificateurs mais aussi l'évaluation. Les résultats obtenus ont montré que notre approche de déduction de vecteurs de relations surpasse la plupart du temps le système RELATIVE.Nous avons obtenu une F-mesure de 77,6 % pour (*Synonymie, Antonymie*) avec RFC et une F-mesure de 79 % pour (*Hyperonymie, Hyponymie*) avec k-PPV (k = 3).

Références

Camacho-Collados, J., L. Espinosa-Anke, J. Shoaib, et S. Schockaert (2019). A Latent Variable Model for Learning Distributional Relation Vectors. In *Proceedings of IJCAI*.

Lafourcade, M. (2007). Making people play for Lexical Acquisition with the JeuxDeMots prototype. In *SNLP'07 : 7th International Symposium on NLP*.

Mikolov, T., K. Chen, G. S. Corrado, et J. Dean (2013). Efficient Estimation of Word Representations in Vector Space. In *Proceedings of the International Conference on Learning Representations (ICLR)*, pp. 1–12.

Summary

During this last decade, word embeddings models learning continuous vector representations of words have been established and integrated in several applications of Natural Language Processing (NLP). These models have been subsequently extended to learn representations of other textual objects such as word senses/definitions, fragments of textual documents or even whole texts. In this paper, we focus on the creation of continuous vector representations for relations. The training of these representations is carried out from a business corpus referring to several specialized domains such as health, justice, urbanism or elections. The quality of these representations is evaluated on the task of identifying/classifying lexical-semantic relations from texts. The results obtained are good and surpass the performances for a recent one state-of-the-art system dedicated to the creation of relations representations.