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