Complémentarités de représentations vectorielles pour la similarité sémantique

Julien Hay[*,**], Tim Van de Cruys[***]
Philippe Muller[***] Bich-liên Doan[**,****]
Fabrice Popineau[**,****] Lyes Benamsili[*]

[*]Octopeek
22 Rue du Général de Gaulle, 95880 Enghien-les-Bains, France
https://www.octopeek.com
[**]LRI
Bat 650 / 660, Rue Noetzlin, 91190 Gif-sur-Yvette, France
[***]IRIT
Université Toulouse III Paul Sabatier
118 Route de Narbonne, 31062 Toulouse, France
[****]CentraleSupélec
3 rue Joliot-Curie, 91192 Gif-sur-Yvette, France

Résumé. La tâche de similarité sémantique textuelle consiste à exprimer automatiquement un nombre reflétant la similarité sémantique de deux fragments de texte. Chaque année depuis 2012, les campagnes de *SemEval* déroulent cette tâche de similarité sémantique textuelle. Cet article présente une méthode associant différentes représentations vectorielles de phrases dans l'objectif d'améliorer les résultats obtenus en similarité sémantique. Notre hypothèse est que différentes représentations permettraient de représenter différents aspects sémantiques, et par extension, d'améliorer les similarités calculées, la principale difficulté étant de sélectionner les représentations les plus complémentaires pour cette tâche. Notre système se base sur le système vainqueur de la campagne de 2015 ainsi que sur notre méthode de sélection par complémentarité. Les résultats obtenus viennent confirmer l'intérêt de cette méthode lorsqu'ils sont comparés aux résultats de la campagne de 2016.

1 Introduction

De nombreux travaux récents s'intéressent à la similarité sémantique, soit entre mots, soit entre groupes de mots, depuis les syntagmes jusqu'à des documents complets, en passant par la similarité entre phrases. Rapprocher des mots ou des phrases par leur sens permet d'utiliser des traits sémantiques dans des modèles en évitant la dispersion inhérente liée à la taille du vocabulaire, ou à l'espace des phrases possibles. La plupart des travaux dans ce sens calculent des similarités entre des représentations construites sur des bases distributionnelles, c'est à dire où la similarité de sens dérive d'une similarité des contextes d'apparition des mots, une hypothèse énoncée par Harris (1954).