Une mesure de similarité entre phrases basée sur des noyaux sémantiques

Samir Amir*, Adrian Tanasescu*, Djamel A. Zighed*

*Institut des Sciences de l'Homme, 14 avenue Berthelot,
69007 Lyon, France, prenom.nom@ish-lyon.cnrs.fr,

Résumé. Nous proposons une nouvelle approche pour le calcul de similarité sémantique entre phrases en utilisant les noyaux sémantiques qui les composent. Ces noyaux, sous la forme de triplets (sujet, verbe et objet) sont supposés porteurs de l'information des phrases dont ils sont extraits. Sur la base de la comparaison sémantique de noyaux, on extrait un ensemble d'indicateurs descriptifs. Nous utilisons ensuite un apprentissage automatique, sur un benchmark contenant des phrases dont la similarité sémantique a été évaluée par des experts humains, afin de déterminer l'importance de chaque indicateur et de construire ainsi un modèle capable de fournir une mesure de similarité sémantique entre phrases. Les expérimentations et les études comparatives, effectuées avec d'autres approches permettant l'estimation des similarités sémantiques entre phrases, montrent les bonnes performances de notre approche. En se basant sur cette dernière, un outil de navigation sémantique est en cours de développement.

1 Introduction

Avec l'émergence de l'internet et les différents moyens de son utilisation, la capacité de déterminer la similarité sémantique entre textes s'est avérée utile pour un large nombre d'applications. Le calcul de similarité entre textes en utilisant les méthodes classiques issues du domaine de la recherche d'information et du traitement automatique de la langue présente certaines limitations, notamment pour les textes courts (phrases). Cela est du au fait que le nombre de mots communs ou leur cooccurrence n'est pas toujours une information pertinente sur le plan statistique. Dans ce contexte, plusieurs mesures de similarité entre phrases ont été proposées. Cependant, due à la complexité du langage humain, des améliorations sont toujours requises.

Deux phrases peuvent avoir une sémantique équivalente (ou proche) avec une structure syntaxique et lexicale complètement hétérogène en terme de nombre et de position de mots, ainsi qu'en termes de relations grammaticales. Par exemple, selon le benchmarks décrits dans (O'shea et al. (2014)), les deux phrases "*I am so hungry I could eat a whole horse plus dessert*." et "*I could have eaten another meal, I'm still starving*." Ont une forte similarité évaluée dans ce benchmark à 0.77 alors qu'on peut clairement voir que leurs structures présentent des hétérogénéités sur plusieurs niveaux.

Afin de faire face au défi décrit précédemment, nous proposons une approche basée sur des noyaux sémantiques. Dans la prochaine section, nous décrivons cette approche plus en détails