Plongement de métrique pour le calcul de similarité sémantique à l'échelle

Julien Subercaze*, Christophe Gravier*, Frédérique Laforest*

* Université de Lyon, F-42023, Saint-Etienne, France
CNRS, UMR5516, Laboratoire Hubert Curien, F-42000, Saint-Etienne, France,
Université de Saint-Etienne, Jean Monnet, F-42000, Saint-Etienne, France

Résumé. Nous explorons le plongement de la métrique de plus court chemin dans l'hypercube de Hamming, dans l'objectif d'améliorer les performances de similarité sémantique dans Wordnet (Subercaze et al. (2015)). Nous montrons que bien qu'un plongement isométrique est impossible en pratique, nous obtenons de très bons plongements non isométriques. Nous obtenons une amélioration des performances de trois ordres de grandeur pour le calcul de la similarité de Leacock et Chodorow (LCH).

1 Introduction

Le concept de similarité sémantique encode la distance conceptuelle entre deux unités de langage. Quand les mots sont les unités de discours, cette similarité est au cœur de nombreuses tâches de TALN. Elle est notamment essentielle pour la désambiguisation (Basile et al. (2014)). Deux approches dominent la mise en œuvre du calcul de la similarité sémantique.

Les approches basées sur les bases de connaissance exploitent aussi bien la structure de la taxonomie Leacock et Chodorow (1998), que son contenu Banerjee et Pedersen (2002) ou les deux Jiang et Conrath (1997). Ces approches ignorent les informations contextuelles et utilisent des bases décrites manuellement. A l'opposé, la sémantique statistique encode la similarité sémantique en se basant sur des observations statistiques, par exemple sur les occurrences dans un corpus. Cependant cette approche est largement limité quant au volume des données qu'elle peut traiter. Les nouvelles approches de plongement de sémantique statistique dans des espaces vectoriels compact (*word embedding*, Collobert et Weston (2008)) apportent une réponse efficace à ce problème. Les architectures neuronales permettent un traitement de larges volumes de données au prix d'un temps d'entrainement de l'ordre de plusieurs jours. La popularité des approches neuronales montre un enthousiasme certain pour des approches efficaces de calcul de similarité entre des paires de mots.

Dans cet article, nous proposons un plongement de la similarité sémantique de Leacock et Chodorow (1998) dans un hypercube de Hamming de dimensions alignées sur la taille des mots processeurs. La similarité de Leacock et Chodorow, basée sur la métrique de plus court chemin dans la relation d'hypernymie de Wordnet, est une des mesures les plus précises. Dans l'évaluation de Miller et Charles (1991) elle atteint le deuxième rang. Ses performances sont légèrement dépassées par l'approche de Jiang et Conrath (1997), basée sur le contenu.