Exploration des paramètres discriminants pour les représentations vectorielles de la sémantique des mots

Frank Meyer, Vincent Dubois

France Telecom R&D
Avenue Pierre Marzin
22307 Lannion cédex
franck.meyer@francetelecom.com
vincen.dubois@francetelecom.com

Résumé : Les méthodes de représentation sémantique des mots à partir d'une analyse statistique sont basées sur des comptes de co-occurrences entre mots et unités textuelles. Ces méthodes ont des paramétrages complexes, notamment le type d'unité textuelle utilisée comme contexte. Ces paramètres déterminent fortement la qualité des résultats obtenus. Dans cet article, nous nous intéressons au paramétrage de la technique dite Hyperspace Analogue to Language (HAL). Nous proposons une nouvelle méthode pour en explorer ses paramètres discriminants. Cette méthode est basée sur l'analyse d'un graphe de voisinage d'une liste de mots de référence pré-classés. Nous expérimentons cette méthode et en donnons les premiers résultats qui renforcent et complètent des résultats issus de travaux précédents.

1 Introduction

Le but des méthodes de représentation sémantique des mots basées sur des vecteurs est d'associer à chaque mot d'un corpus un vecteur (en général dans R^N) de telle manière que la distance sémantique entre 2 mots soit reflétée par la distance entre les 2 vecteurs les représentants. On cherche donc à modéliser le sens des mots sous une forme numérique, objective, dans un espace vectoriel. Les méthodes de représentation sémantique par vecteurs les plus connues sont Latent Semantic Analysis (LSA) (Deerwester, Dumais, Furnas, Landauer et Harshman 1990) et Hyperspace Analogue to Language (HAL) (Lund et Burgess, 1996). Nous allons d'abord présenter brièvement le principe de ces deux méthodes et expliquer pourquoi, dans le cas des mots, la méthode de type HAL nous apparaît comme plus appropriée. Après avoir rapidement présenté les principaux travaux dans le domaine du choix des paramètres de HAL, nous exposerons notre méthode. Une expérimentation destinée à illustrer ses principes est ensuite décrite, ainsi que ses principaux résultats.

LSA et HAL sont des méthodes basées sur une analyse statistique d'un corpus de documents textuels. Les documents peuvent être des courts textes, des paragraphes voire des phrases. En sortie d'analyse, LSA et HAL produisent une matrice finale qui représente chaque mot i par son vecteur v_i. Pour mesurer la proximité sémantique de 2 mots, on utilise une fonction de distance entre les 2 vecteurs u et v qui les représentent. La distance utilisée est souvent la distance du cosinus. D'autres distances sont couramment utilisées : city-