Utilisation d'informations syntaxico-sémantiques associées à LSA

Nicolas Béchet

Utilisation d'informations syntaxico-sémantiques associées à LSA

Nicolas Béchet

LIRMM, Univ. Montpellier 2, CNRS
161 rue Ada, 34392 Montpellier, France
nicolas.bechet@lirmm.fr

MOTS-CLÉS : *analyse sémantique latente, classification conceptuelle, analyse syntaxique.*

KEYWORDS: *latent semantic analysis, conceptual classification, syntactic analysis.*

1. Introduction

L'approche présentée dans cet article s'appuie sur LSA (Latent Semantic Analysis) (Landauer *et al.*, 1997). Elle se fonde sur le fait que des mots sont jugés sémantiquement proches s'ils apparaissent dans le même contexte. S'appliquant à des corpus de grande dimension, elle permet entre autres le regroupement de termes, ce qui constitue l'objet de cet article. La méthode présentée dans celui-ci se nomme *ExpLSA* (***Exp***ansion des contextes avec ***LSA***). Elle consiste à enrichir les corpus utilisés par LSA en utilisant des informations sémantiques obtenues grâce à la syntaxe.

2. De LSA à ExpLSA

La méthode LSA décrit le corpus sous forme matricielle : les lignes représentent les mots, les colonnes les contextes choisis (document, paragraphe, etc.). Une cellule de la matrice indique le nombre d'occurrences des mots dans chacun des contextes. La matrice est ensuite normalisée (logarithme et calcul d'entropie) puis approximée grâce à une décomposition en valeurs singulières. Avoir deux mots proches sémantiquement se traduit par la proximité des vecteurs lors d'un calcul de similarité (généralement le cosinus) à partir de la matrice approximée. Notons qu'une taille réduite des contextes dégrade les résultats obtenus par LSA. Ainsi, *ExpLSA* est une méthode qui ajoute des connaissances sémantiques aux contextes en utilisant des informations syntaxiques.

Le but fixé de l'approche *ExpLSA* consiste à regrouper automatiquement des termes nominaux extraits avec des systèmes d'extraction de la terminologie. Nos travaux

s'appuient sur un corpus relatif aux Ressources Humaines issu de la société PerformanSe. Ce dernier est écrit en français et utilise un vocabulaire spécialisé. Pour effectuer ce regroupement, la méthode LSA est appliquée à partir d'un corpus dont les phrases sont enrichies. Cet enrichissement se fonde sur la régularité de certaines relations syntaxiques. Ainsi, dans un premier temps, l'analyseur syntaxique SYGMART (http ://www.lirmm.fr/∼chauche/) est appliqué. Il n'en sera conservé que les relations Verbe-Objet, permettant d'étudier la proximité sémantique entre les verbes issus des relations extraites, en utilisant la mesure d'Asium (Faure, 2000). Cette mesure considère comme proches des verbes possédant un nombre important d'objets en commun. Les objets (mots) dont les verbes ont été jugés proches sémantiquement sont ensuite regroupés. Le corpus initial est alors complété en effectuant une expansion de chaque mot avec les autres mots considérés comme sémantiquement proches. Une fois le corpus enrichi, l'approche LSA peut ensuite être appliquée sur celui-ci.

ExpLSA a été comparée à LSA en évaluant les résultats obtenus avec ceux d'un expert qui a associé les termes pertinents à 19 concepts. Cette comparaison a été menée en s'appuyant sur ces concepts choisis aléatoirement. L'évaluation des approches a été effectuée avec deux valeurs de seuil très différentes pour la mesure d'Asium (0.6 et 0.9). Ce seuil traduit le rapprochement sémantique des objets (mots) issus des relations syntaxiques utilisées. En effet, plus ce seuil est proche de 1, plus les mots ajoutés sont sémantiquement proches selon Asium. Les résultats obtenus en comparant *ExpLSA* à LSA sont améliorés avec un seuil de 0.9 contrairement à un seuil de 0.6. Ces résultats confortent le fait qu'un seuil élevé pour la mesure d'Asium enrichit beaucoup moins le corpus qu'un seuil plus faible mais de manière plus pertinente. Cela se traduit par des résultats qui sont globalement de meilleure qualité avec *ExpLSA*.

3. Conclusion

Cet article montre une approche permettant d'enrichir le corpus étudié par une méthode utilisant des informations syntaxico-sémantiques. Celle-ci offre des résultats encourageants, permettant de combler les lacunes de LSA dans la plupart des cas. Les futurs travaux envisagés consisteront à valider la qualité de l'enrichissement des corpus par la méthode d'Asium ainsi que le choix du seuil optimal. D'autres approches consistant à étudier l'ajout d'informations grammaticales avec LSA seront également testées. Par ailleurs, de manière plus générale, l'ajout de connaissances syntaxiques aux méthodes de classifications existantes sera considéré.

4. Bibliographie

Faure D., Conception de méthode d'apprentissage symbolique et automatique pour l'acquisition de cadres de sous-catégorisation de verbes et de connaissances sémantiques à partir de textes : le système ASIUM, PhD thesis, Université Paris-Sud, 20 Décembre, 2000.

Landauer T., Dumais S., « A Solution to Plato's Problem : The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge », *Psychological Review*, vol. 104, n° 2, p. 211-240, 1997.