Acquisition automatique de relations entre concepts dans le domaine scientifique

Kata Gábor
LIPN, UMR 7030, Université Paris 13
Kata.gabor@lipn.univ-paris13.fr

Résumé : De nos jours, la production d'articles scientifiques croît à un rythme accéléré. Cette explosion d'information rend le travail des chercheurs, des experts et des relecteurs de plus en plus difficile et nécessite de nouvelles méthodes pour la compréhension, l'extraction et la structuration automatique de l'information dans les textes de spécialité. Comme la disponibilité et la couverture des bases de connaissances existantes est souvent insuffisante, nous proposons de prendre comme point de départ l'analyse sémantique du contenu afin de faire émerger un modèle de connaissances. Nous présentons deux approches non supervisées pour l'acquisition des relations sémantiques dans un corpus de spécialité. L'identification des relations ne nécessite pas des données d'apprentissage annotées et bien qu'elle soit spécifiquement dédiée à la littérature scientifique, elle reste applicable sur n'importe quel domaine pour lequel une telle littérature existe.

La présentation explorera les problématiques spécifiques à la tâche non supervisée. Deux approches complémentaires seront distinguées et explorées. La première se concentre principalement sur les relations lexicales, qui se caractérisent par une sélection sémantique des arguments, et qui ne dépendent pas du contexte. Cette approche est basée sur la représentation du sens des mots individuels par des vecteurs distributionnels (word embeddings). Les vecteurs sont créés à partir de corpus et combinés pour représenter le sens et la relation sémantique du couple d'entités. Nous proposons une nouvelle méthode de combinaison de vecteurs distributionnels qui permet de mieux estimer la similarité relationnelle entre deux couples d'entités. L'avantage de cette méthode est de pouvoir s'appliquer à des couples d'entités qui ont peu de co-occurrences dans le corpus. La deuxième approche, à son tour, s'applique aux relations contextuelles et s'appuie sur les contextes de co-occurrence des entités. Les couples d'entités sont caractérisés par leurs co-occurrences avec des motifs spécifiques à la relation, qui sont extraits automatiquement à partir du corpus. Nous montrons que cette approche peut bénéficier de la fouille de motifs séquentiels, qui crée un espace vectoriel plus adapté (moins creux) pour un clustering non supervisé.