

Sujet de stage

Réalisation et évaluation d'un système d'extraction de relations sémantiques

Encadrement

Sylvie Szulman

Adeline Nazarenko

prenom.nom@lipn.univ-paris13.fr

Contexte

L'extraction de relations sémantiques est depuis longtemps reconnue comme une tâche clef du processus d'acquisition de connaissances à partir de textes. Elle a été considérée sous différents angles (pour la construction d'ontologies et la création de rôles entre concepts, pour l'extraction d'information et le remplissage de formulaires, pour l'analyse rhétorique, etc.) et a fait l'objet de nombreux travaux. Deux grandes familles de méthodes sont utilisées pour l'extraction de relations sémantiques des textes :

- La première fait partie des méthodes dites « d'extraction d'information » et repose sur des « patrons d'extraction » qui sont projetés dans les textes [1]. La qualité des résultats dépend de la précision avec lesquelles les règles ont été écrites ou de la qualité du corpus qui sert à les apprendre. La difficulté tient au fait que ces règles sont généralement très dépendantes du corpus et du domaine considérés.
- La seconde approche à base plus statistique repose sur des associations de mots repérées en corpus. Elle est généralement plus robuste et moins dépendante des corpus à analyser mais elle donne des résultats plus bruités.

Mission

L'objectif de ce stage est de développer un module d'extraction de relations sémantiques qui combine ces deux approches comme cela a été proposé dans [2] et à l'évaluer dans le cadre d'un processus de construction d'ontologies à partir de textes.

Ce travail pose une double difficulté :

- Sur le plan de l'implémentation, il s'agira de pouvoir analyser des corpus de quelques dizaines voire centaines de milliers de mots, ce qui suppose un effort d'optimisation du code et des structures de données manipulées.
- Sur le plan de l'évaluation, il faudra définir un protocole qui permette de mesurer l'apport de l'approche proposée par rapport à des approches à base de patrons ou de simples mesures d'association. Les expériences pourront être faites sur un corpus de travail de l'équipe RCLN

Plan de travail

1. Analyse de la méthode proposée dans [2] et confrontation avec l'état de l'art.
2. Conception et analyse de la méthode proposée.

3. Expérimentation sur des corpus de tests et analyse des résultats.
4. Comparaison avec des méthodes à base de patrons et des méthodes statistiques sur les mêmes corpus. Cela pourra être fait soit à partir d'outils existants soit en collaboration avec des membres de l'équipe RCLN.
5. Selon la qualité des résultats obtenus, des variantes de la méthodes initiales pourront être proposées et testées.
6. Recommandation pour l'intégration d'une méthode d'extraction de relations sémantiques dans une plate-forme de construction d'ontologies telle que Terminae [3] ou Dafoe [4].

Profil recherché

Bac + 5 en informatique

Des compétences en informatique et en TAL seront appréciées lors de l'examen des candidatures.

Conditions et cadre de travail

Le stage se déroulera au sein du LIPN (www-lipn.univ-paris13.fr).

Stage de 6 mois, financé sur projet.

En cas de succès, ce stage pourra se poursuivre par une thèse financée sur projet.

Références

- [1] Nathalie Aussenac-Gilles, Nathalie Hernandez. Du linguistique au conceptuel : étapes de l'identification de relations conceptuelles à partir de textes. Atelier "Acquisition et modélisation de relations sémantiques" associé à la conférence TIA 2009, Toulouse, nov. 2009, Sylvie Despres, Natalia Grabar (Eds.).
- [2] Rim Bentebibel, Adeline Nazarenko, Sylvie Szulman. « Un système d'aide à l'extraction de relations sémantiques pour la construction d'ontologies à partir de textes », in *Actes des 10ème journées Francophones Extraction et Gestion des Connaissances (EGC 2010)*, pp. 483-494, Hammamet, Tunisie, janv. 2010.
- [3] Brigitte Biébow, Sylvie Szulman. « TERMINAE: A Linguistic-Based Tool for the Building of a Domain Ontology ». *Proceedings of EKAW 1999*: 49-66.
- [4] Jean Charlet, Sylvie Szulman, Guy Pierra, Nadia Nadah, Henry Valéry Teguiak, Nathalie Aussenac-Gilles, Adeline Nazarenko. « DAFOE: A Multimodel and Multimethod Platform for Building Domain Ontologies », In *Actes des 2èmes Journées Francophones sur les Ontologies*, D. Benslimane, C. Roche et S. Spaccapietra (eds.), 1-2 décembre 2008, Lyon, France.