

Sujet de stage

Réalisation d'un système d'annotation de texte à partir d'ontologies

Encadrement

Laurent Audibert

Adeline Nazarenko

prénom.nom@lipn.univ-paris13.fr

Contexte

Au cours des dernières années, de nouveaux systèmes d'analyse de textes reposant sur une approche d'annotation sémantique ont vu le jour. Ces systèmes décrivent le contenu d'un texte en liant ce dernier à une ontologie. Ils proposent donc une interprétation du texte au regard de l'ontologie considérée. Cette approche prend tout son sens dans le cadre du Web Sémantique. En tant que système sémantique formel, l'ontologie définit une manière standardisée de coder les connaissances (sous la forme de concepts, d'instances et de rôles conceptuels) qui donne une vue nécessairement partielle du contenu du texte mais qui supporte des raisonnements qui ne peuvent être faits sur le texte de départ.

Ce stage s'inscrit dans le cadre du programme Quaero dont l'un des objectifs consiste à annoter sémantiquement des documents, les annotations produites étant utilisées par différents systèmes documentaires (recherche d'information sémantique, catégorisation de documents, etc.).

Mission

Le but de ce stage est de développer un module d'annotation sémantique qui prend en entrée un document – éventuellement déjà partiellement analysé –, une ontologie ainsi que des règles d'annotation attachées à l'ontologie et qui produit en sortie un texte annoté au regard de l'ontologie de départ. Toute la difficulté consiste à prendre en compte une grande diversité de règles d'annotation : certaines peuvent être très simples – comme l'association d'un terme non ambigu à un concept – mais d'autres, plus complexes, s'expriment sous la forme d'expressions régulières ou nécessitent des calculs probabilistes. Il faut également s'appuyer sur un premier étiquetage du texte lorsque celui-ci est fourni par des outils d'analyse linguistique (étiquetage morpho-syntaxique, terminologique, etc.).

Ce module d'annotation devra pouvoir être interfacé avec différents outils développés au sein du LIPN, ce qui nécessitera un travail de conception initial très rigoureux. La complexité des problèmes d'annotation et le volume de données à traiter supposera également une analyse approfondie des différentes solutions techniques permettant d'appliquer un ensemble de règles d'annotation sur un corpus textuel et du travail de développement.

Une fois développé le système d'annotation initial, deux pistes de travail pourront être explorées selon le profil et les goûts du candidat :

- L'annotation en mode interactif. Si le premier système d'annotation doit fonctionner en mode automatique, un mode de fonctionnement interactif est également envisagé. Il s'agira, pour de petits textes, de proposer des annotations à l'utilisateur mais de lui laisser la possibilité de choisir entre différentes annotations, de réviser les

annotations faites et de modifier les règles d'annotation quand elles paraissent inappropriées.

- L'analyse de la qualité de l'annotation (orientation plutôt recherche). Il est essentiel de pouvoir apprécier la qualité d'une annotation (degré d'ambiguïté, couverture, granularité, cohérence, complétude, etc.). Il faudra donc proposer des méthodes et mesures permettant de le faire. Différentes approches complémentaires sont envisageables : mesures prédictives calculées sur la seule base de règles d'annotations, mesures d'adéquation *a priori* de l'ontologie à un corpus (avant annotation), mesure de la qualité *a posteriori* de l'annotation.

Profil recherché

Bac + 5 en informatique

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

Conditions

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.

Candidature

Envoyer une lettre de motivation et un CV aux encadrants.

Références

1. Amardeilh F. (2007). *Web sémantique et informatique linguistique : propositions méthodologiques et réalisation d'une plateforme logicielle*. In Thèse de doctorat, Univ. Paris X, p. 223–253.
2. Abdoulaye Guissé, François Lévy, Adeline Nazarenko, Sylvie Szulman. « Annotation sémantique pour l'indexation de règles métiers », in *Actes de la Conférence Internationale sur la Terminologie et l'Intelligence Artificielle (TIA 2009)*, Marie-Claude L'Homme, Sylvie Szulman (Eds.), Toulouse (France), (version électronique, 11 pages), nov. 2009.
3. Kalyanpur A., Hendler J., Parsia B. & Golbeck J. (2003). Smore - semantic markup, ontology, and rdf editor. In <http://www.mindswap.org/papers/SMORE.pdf>.
4. Ma Y., Audibert L. & Nazarenko A. (2009). Ontologies étendues pour l'annotation sémantique. In F. L. Gandon, Ed., *Actes des 20es Journées Francophones d'Ingénierie des Connaissances (IC 2009)*, p. 205–216, Hammamet, Tunisie : PUG.
5. Uren V., Cimiano P., Iria J., Handschuh S., Vargas-Vera M., Motta E. & Ciravegna F. (2006). Semantic annotation for knowledge management : Requirements and a survey of the state of the art. *Journal of Web Semantics*, 4.