Titre Apprentissage et fouille de relations entre concepts pour l'enrichissement d'ontologies

Problématique :

L'enrichissement ou peuplement d'ontologies consiste à ajouter des concepts, instances ou relations à des ontologies existantes. Cet enrichissement peut se fonder sur des textes ou d'autres ressources sémantiques. Plusieurs travaux se sont intéressés à l'enrichissement de concepts et instances [1]. Nous nous intéressons à l'enrichissement des ontologies par la découverte des relations sémantiques à partir de textes. La principale difficulté consiste à proposer une approche générique et indépendante du type de corpus et du domaine. Nous ne voulons imposer aucune contrainte sur les types de relations à extraire contrairement à des travaux comme [5].

Contexte

Le travail de postdoc se déroulera au sein du PRES Sorbonne Paris Cité entre le LIPN (http://www-lipn.univ-paris13.fr/), dans l'équipe «Représentation des Connaissances et Langage Naturel» (RCLN) et le LATTICE (http://www.lattice.cnrs.fr/). Ce poste est rémunéré grâce au soutien du laboratoire d'excellence "Empirical Foundations of Linguistics" (labex EFL, http://www.labex-efl.org/). Il fait partie d'un projet plus large sur l'extraction d'information, visant à intégrer des approches à base d'apprentissage automatique, de fouille de données séquentielles et de ressources, mené en commun entre le LIPN et le LATTICE dans le cadre du labex EFL.

Objectifs de la recherche

L'objectif est d'enrichir une ontologie par des relations sémantiques entre concepts extraites de corpus. Le processus peut se décomposer en quatre étapes : i) étape d'annotation, ii) étape d'acquisition de patrons, iii) étape de validation des patrons et enfin iv) une étape d'extraction et de typage des relations.

L'annotation des textes à partir de concepts d'une ontologie est une tâche non triviale. Plusieurs approches sont déjà proposés dans la littérature, il s'agit dans un premier temps de faire un état de l'art et de réutiliser des outils d'annotation existants tels que l'outil Annotator développé au LIPN [2].

Pour acquérir des patrons d'extraction, le corpus annoté dans l'étape précédente fournira un jeu de données d'apprentissage. Une possibilité est de tirer parti de la fouille de textes, par exemple de la fouille de motifs séquentiels, pour découvrir automatiquement des patrons morpho-syntaxiques porteurs de relations sémantiques entre concepts. De récents travaux ont montré l'intérêt de ces techniques dans le cadre spécifique de l'extraction d'information (reconnaissance d'entités nommées, relations entre entités nommées) ou de l'analyse linguistique (e.g. acquisition de motifs caractéristiques d'un genre textuel pour l'analyse stylistique [3,4]). La découverte automatique de relations entre concepts par des techniques de fouille de données reste un défi encore peu exploré à notre connaissance.

Les patrons extraits devraient être filtrés et validés pour s'appliquer à l'ontologie utilisée. Une piste possible consiste à proposer des heuristiques qui permettent de spécifier des contraintes liées aux caractéristiques de l'ontologie.

La dernière étape consiste à extraire et typer des relations à l'aide des patrons sélectionnés, en vue de leur intégration dans l'ontologie cible. Une comparaison / combinaison avec des méthodes d'apprentissage automatique (supervisées ou non) pour la même tâche sera aussi considérée.

Références

[1] Agirre E., Olatz A., Hovy E.H., Martinez D. (2000) Enriching very large ontologies using the WWW. In ECAI Workshop on Ontology Learning.

[2] http://lipn.univ-paris13.fr/~szulman/Annotator/annotator.html

[3] Auger, A., & Barrière, C. (2008). Pattern-based approaches to semantic relation extraction: A state-of-the-art. *In* Terminology, 14(1), pp. 1-19.

[4] Nicolas Béchet, Peggy Cellier, Thierry Charnois, Bruno Crémilleux (2012). Discovering Linguistic Patterns Using Sequence Mining. In CICLing 2012. pp. 154-165

[5] Fabian M. Suchanek, Mauro Sozio,Gerhard Weikum (2009). Sofie: A self-organizing framework for information extraction. In WWW conference, pp. 631– 640.

Profil recherché

- Thèse en Informatique
- Qualité de rédaction en français et en anglais
- Expertise et/ou Intérêt pour
    - L'Ingénierie des Connaissances et le web sémantique
    - Le Traitement Automatique des Langues et la Fouille de Textes
    - L'Apprentissage Automatique

Durée : 12 mois (entre le LIPN et le LATTICE)
Début souhaité : Dès que possible

Pour candidater, envoyer à Isabelle Tellier (isabelle.tellier@univ-paris3.fr) et Haïfa Zargayouna (haifa.zargayouna@lipn.univ-paris13.fr) :

- un CV (avec liste de publications)
- une lettre de motivation
- le nom de deux référents (avec adresse mail)