Title: Learning relations between concepts for automatic ontology enrichment

Description:
The automatic enrichment or population of ontologies is the process of properly adding concepts, instances or relations to existing ontologies. The process may be based on the analysis of texts or other semantic resources. Several works have dealt with the enrichment of concepts and instances [1]. We are interested to the automatic enrichment of ontologies with semantic relations automatically extracted from a corpora of texts. The most important issue consists in the difficulty of producing an approach which is independent from both the domain and the corpus. We do not want to impose any constraint on the types of the relations to be extracted unlike works like [5].

Context:
The candidate will work at the PRES Sorbonne Paris Cité, alternating between the LIPN (http://lipn.univ-paris13.fr/en/laboratory), RCLN team "Représentation des Connaissances et Langage Naturel" and the LATTICE (http://www.lattice.cnrs.fr/). The post will be supported by the "laboratoire d'excellence" Empirical Foundations of Linguistics (LabEx EFL, http://www.labex-efl.org/). The work will be part of a broader project on information extraction, which aims to integrate approaches based on machine learning, pattern and resource mining, a project led commonly by LIPN and LATTICE within the LabEx EFL.

Objectives:
The main objective is to enrich a given ontology by semantic relations. These relations will be extracted from a corpus, using pattern mining techniques, and link concepts that already exist in the ontology.
The process may be split into four steps: i) annotation, ii) pattern learning, iii) pattern validation, iv) relation extraction and classification.

The annotation of texts with concepts is a non-trivial task. Many approaches exist in literature: the candidate will produce a report on the existing state of the art and reuse existing annotation tools like Annotator (developed at LIPN [2]).

In order to learn the relation extraction patterns, the corpus annotated during the previous step will be used as a learning dataset. One possible choice is to exploit text mining techniques, for instance sequential pattern mining, to discover automatically morpho-syntactic patterns that are carriers of semantic relations between concepts. Recent works showed the interest of such techniques in the framework of information extraction (recognition of named entities and the relations between them) or linguistic analysis (e.g. acquiring textual characteristics patterns for the stylistic analysis [3,4]). The automatic discovery of relations between concepts by data mining is, to our knowledge, an open research topic.

The extracted patterns will be validated in order to be applied to the ontology. A possible strategy may be to use heuristics that allow to specify constraints related to the characteristics of the ontology.

The last step is to extract and classify the relationships using the selected patterns, before their integration into the target ontology. We foresee a comparison/combination with existing machine learning techniques (supervised or not) for the same task.

Bibliography

[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.

Selection Criteria:

- PhD in Computer Science
- Good writing skills
- Experience and/or interest in:
    - Knowledge Engineering and the Semantic Web
    - Natural Language Processing and Text Mining
    - Machine Learning

Duration: 12 months (bw LIPN and LATTICE)
Start: as soon as possible

The candidates should send to Isabelle Tellier (isabelle.tellier@univ-paris3.fr) and Haïfa Zargayouna (haifa.zargayouna@lipn.univ-paris13.fr) :

• a detailed CV (with a list of publications)
• a cover letter
• the names of two referents (and their e-mail address)