Knowledge extraction in web media: at the frontier of NLP, Machine Learning and Semantics

Mots clés :
● Directeur de thèse : raphael TRONCY
● Co-encadrant(s) :
● Unité de recherche : Laboratoire de recherche d'EURECOM
● Ecole doctorale : École Doctorale Informatique, Télécommunications, Électronique de Paris
● Domaine scientifique principal: Divers

Résumé du projet de recherche (Langue 1)
The Web offers a vast amount of structured and unstructured content from which more and more advanced techniques are developed for extracting entities and relations between entities, one of the key elements for feeding the various knowledge graphs that are being developed by major web companies as part of their product offerings. Most of the knowledge available on the Web is present as natural language text enclosed in Web documents aimed at human consumption. A common approach for obtaining programmatic access to such a knowledge uses information extraction techniques. It reduces texts written in natural languages to machine readable structures, from which it is possible to retrieve entities and relations, for instance obtaining answers to database style queries. In the series of the WoLE workshops ([1] and [2]), we have proposed and discussed such a vision, gaining in popularity and attracting high quality papers. We contributed to the emerging idea that entities should be a first class citizen on the Web.

Résumé du projet de recherche (Langue 2)
In parallel, we have observed that entities play a pivotal role in numerous National and European projects in which we have participated, such as OpenSEM1, EventMedia2, LinkedTV3, and MediaMixer4. A common research line consists in annotating texts such as users’ posts, item descriptions, video subtitles, with entities that are uniquely identified in some knowledge bases as part of the Global Giant Graph5. The Natural Language Processing (NLP) community has been addressing this crucial task for the past few decades. As a result, the community has established gold standards and metrics to evaluate the performance of algorithms in important tasks such as coreference Resolution, Named Entity Recognition, Entity Linking and Relationship Extraction, just to mention few examples. Scientific evaluation campaigns, CoNLL (2003-2014), ACE (2005-2007), TAC (2009-2014), #Microposts (2013-2014), and ETAPE in 2012 were proposed to compare the performance of various systems in a rigorous and reproducible manner. Some of these topics overlap with research that the Database Systems and more recently the Knowledge Engineering communities have been addressing also for decades, such as Identity Resolution (Deduplication, Entity Resolution, Record Linkage), Schema Mapping (Schema Mediation, Ontology Matching) and Data Fusion (instance matching, data interlinking). Meanwhile the Semantic Web and Linked Data communities have been addressing questions related to how to model, serialize and share such information on the Web, as well as on how to use knowledge described in more expressive formalisms for a variety of integration, retrieval and discovery tasks. Finally, the Information Retrieval community has been increasingly paying attention to the intersection of structured and unstructured data, with topics encompassing Entity-Oriented Search (cf. TREC Entity, KBA), Semantic Search, etc. As a followup of the NERD6 initiative, we aim to harvest the Web for disambiguating entities providing topnotch results, leveraging on the findings of the NLP community and the expertise in the resource description from the Semantic Web community. In [3], we have presented a machine learning approach that combines the state of the art from the named entity extraction in the natural language processing domain, and named entity linking from the semantic web community. Results are encouraging, since they show improvements from the state of the art in entity recognition and classification. A lot of efforts will be dedicated in the near future for the linking task. Recently, we coorganized the #Microposts2014 NEEL Challenge [4] for which the task is to automatically link entities extracted from microposts. The TAC conference is pushing hard the competition in linking entities extracted from newswire content to a reference knowledge base. Those benchmarks constitute an ideal evaluation framework for the work to be conducted in this thesis. We foresee the following research questions: ? RQ1 What are the most performing academic and commercial tools for realizing the entity linking task? Under which assumptions are those systems working? ? RQ2 What is the role of large knowledge bases, such as DBpedia and YAGO, in structuring and linking textual documents to web resources? ? RQ3 Does the semantics of the schema influence the performance in entity linking, and, therefore, in entity recognition? ? RQ4 What is the contribution of a reference knowledge base and the associated entity linking task in improving knowledge extraction from unstructured data? ? RQ5 Given the highly domain dependent approaches proposed so far in the research communities for the entity linking, what are the best settings and conditions for better performing the linking while processing media resources? ? RQ6 How does the foreseen research can be agnostic to any domain? ? RQ7 What are the benefits of the entity linking in indexing media resources? We aim to reach an outstanding quality in publications, showing important contributions to both the NLP and Semantic Web communities. At the end, the candidate will show important expertise in the field, proven by active participations in the discussion, such paper presentations at conferences such as ISWC, ESWC, WWW, LREC or ACL and participation to various project meetings. As part of his learning phase, the candidate will be invited to contribute actively in the NERD framework, and to make in production the research modules he will be coding on.