Title: Learning RDF pattern extractors for a language from dual bases Wikipedia/LOD

Research axis of the 3IA: Core elements of AI

Supervisor: Fabien Gandon (Fabien.gandon@inria.fr), Inria University Cote d'Azur (UCA)

The laboratory and/or research group: Inria University Cote d'Azur (Wimmics team, https://team.inria.fr/wimmics/), I3S Laboratory I3S - Inria Sophia Antipolis

Description of the topic:

Whether automatically extracted from structured elements of articles or centrally populated and crowdsourced, the open and linked data published by DBpedia and Wikidata now offer rich and complementary views of the textual descriptions found in Wikipedia.

However, the text of Wikipedia articles still contains a lot of information that would be interesting to extract in order to improve structured databases in terms of coverage and quality, for example by relying on natural language processing techniques.

This thesis proposes to exploit the dual bases that can be formed from Wikipedia pages and Linked Open Data (LOD) bases to produce RDF pattern extractors for a language. In particular the initial research question will be can we learn customized RDF graph extractors from the dual base Wikipedia/DBpedia+Wikidata.

In a first stage the candidate will consider the question: how to generate a training set for a specific RDF pattern? The problem is to select an adequate training set to learn an extractor, for instance a set of graphs corresponding to a SPARQL pattern and the union of the sets of the pages mentioning in their plain text all the entities referenced in each one of these graphs.

In a second stage the candidate will address the question: how to learn an extractor from a training set? The candidate will consider using the latest NLP techniques for instance to "translate" from English to Turtle. The task includes learning to select and learning transform the relevant parts of the page to RDF.

Incremental steps can be introduced to achieve this goal: starting from a single triple pattern before generalizing to arbitrary basic graph patterns; starting from a subdomain of Wikipedia before generalizing do any domain; focusing on one natural language before generalizing to more languages.

The subject also supports extensions such as the application of the learned extractors to augment the dataset from Wikipedia or other corpora and to improve the quality of the data detecting silences or errors. The inverse problem of suggesting missing texts from data could also be considered. Finally, the perspectives include approaches evaluating transfer learning and active learning.

Internal references:

Fabien Gandon, Raphael Boyer, Olivier Corby, Alexandre Monnin. Materializing the editing history of Wikipedia as linked data in DBpedia, ISWC 2016 - 15th International Semantic Web Conference, Oct 2016, Kobe, Japan

Fabien Gandon, Raphaël Boyer, Alexandre Monnin. DBpédia.fr : retour sur la publication de données de la culture française, I2D – Information, données & documents, A.D.B.S., 2016, Web de données et création de valeurs : le champ des possibles, 53 (2016/2), pp.84

Elena Cabrio, Serena Villata, Fabien Gandon. Classifying Inconsistencies in DBpedia Language Specific Chapters, Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), May 2014, Reykjavik, Iceland. pp.1443-1450

External references:

Nayak, Tapas, et al. "Deep neural approaches to relation triplets extraction: A comprehensive survey." Cognitive Computation 13.5 (2021): 1215-1232.

Martinez-Rodriguez, Jose L., Aidan Hogan, and Ivan Lopez-Arevalo. "Information extraction meets the semantic web: a survey." Semantic Web 11.2 (2020): 255-335.