A System Description of Natural Language Query over DBpedia*

Nitish Aggarwal and Paul Buitelaar

Unit for Natural Language Processing, Digital Enterprise Research Institute,
National University of Ireland, Galway
{nitish.aggarwal,paul.buitelaar}@deri.org

Abstract. This paper describes our system, which is developed as a first step towards implementing a methodology for natural language querying over semantic structured information (semantic web). This work focuses on interpretation of natural language queries (NL-Query) to facilitate querying over Linked Data. This interpretation includes query annotation with Linked Data concepts (classes and instances), a deep linguistic analysis and semantic similarity/relatedness to generate potential SPARQL queries for a given NL-Query. We evaluate our approach on QALD-2 test dataset and achieve a F1 score of 0.46, an average precision of 0.44 and an average recall of 0.48.

Introduction

The rapid growth of Linked Data offers a wealth of semantic data for facilitating a interactive way to access the Web. However, Linked Data also brings several challenges in providing a flexible access over the Web for all users. Structured query languages like SPARQL provide the capability of accessing these this data, but these languages are restricted to the vocabulary defining the data. This data should be easily searchable and consumable for casual users to query in their native language, similar as with traditional web of documents through web search engines for document search.

In order to facilitate NL-queries over Linked Data, we implemented a basic pipeline that includes entity annotation, a deep linguistic analysis and semantic similarity/relatedness. This pipeline is very similar to the system implemented

* This work is supported in part by the European Union under Grant No. 248458 for the Monnet project and by the Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Lion-2).

In: C. Unger, P. Cimiano, V. Lopez, E. Motta, P. Buitelaar, R. Cyganiak (eds.): Proceedings of Interacting with Linked Data (ILD 2012), Workshop co-located with the 9th Extended Semantic Web Conference, Heraklion, Greece, 28-05-2012, published at http://ceur-ws.org

by Freitas A. et.al. [1], which is based on the a combination of entity search, a Wikipedia-based semantic relatedness (using Explicit Semantic Analysis) measure and spreading activation. However, our work focuses additionally on a deep linguistic analysis and categorization of a NL-Query. For example, a given NL-Query, such as "who designed the Brooklyn Bridge", is first categorized as a person- type query and, then the verb "designed" is modified to "designer". We also further investigate the approaches used for computing semantic similarity and relatedness.

Query Interpretation Approach

In our system, the interpretation of NL-Query is driven by semantic matching between Linked Data vocabulary and terms appearing in the NL-Query, to construct a SPARQL query. A well- interpreted SPARQL query from a given NL-Query can overcome the semantic gap between user- described queries and Linked Data vocabularies.

This includes three components:, namely query annotation, a deep linguistic analysis and semantic similarity/relatedness as shown in Fig.1. We describe below these components by taking an example NL-Query over the DBbpedia dataset.

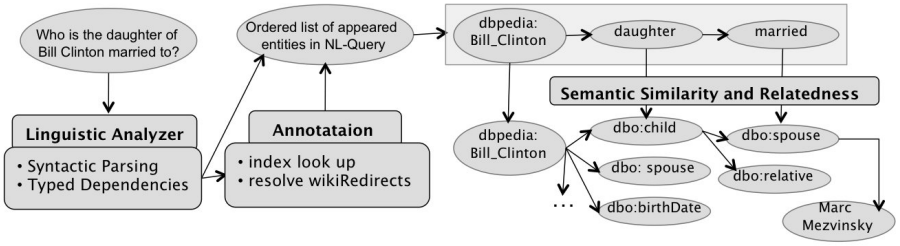

Fig. 1. Query interpretation pipeline for an example NL-Query "Who is the daughter of Bill Clinton married to?".

Query Annotation The interpretation process starts by identifying the potential entities, i.e. DBpedia instances and classes present in the NL-Query. For identifying these entities, we created two separate lucene indices, one for labels & URIs of all DBpedia instances and other one for all DBpedia classes. Annotating a NL-Query includes the extraction of keywords by removing stop words and identification of possible DBpedia classes followed by identification of DBpedia resources by performing keyword search over both lucene indices. After identifying potential resource labels, we perform disambiguation to recognize the most appropriate DBpedia resource URI, as there are multiple URIs for the same DBpedia resource label. The disambiguation is performed by retrieving wikiPageRedirects URIs, if the recognized URI redirects to any other

DBpedia resource URI, e.g. in our system "Bill Clinton" is identified as URI "http://dbpedia.org/resource/BillClinton" which redirects to the right URI of label "Bill Clinton" i.e. "http://dbpedia.org/resource/Bill_Clinton".

Linguistic Analysis A deep linguistic analysis is performed by generating a parse tree and typed dependencies by using the Stanford parser. Generated parse trees provides a phrase extraction for identifying them as potential DBpedia resources or DBpedia classes. For instance, in our example query, the phrase "Bill Clinton" is identified as a noun phrase. It suggests us to perform a lucene search over the whole phrase "Bill Clinton" rather than separate searches for "Bill" and "Clinton".

We convert the given NL-Query into an ordered list of potential terms by using typed dependencies generated by the Stanford parser. For creating this ordered list, first we select a central term among all the identified terms, where the central term is the most plausible term to start matching of a given NL-Query to the vocabulary appeared in the DBpedia graph. This selection is performed by prioritizing the DBpedia resources over DBpedia classes. Then, we retrieve the direct dependent terms of this central term following the generated typed dependencies and add them into the ordered list. Similarly, we perform the same for all the other terms in the list. For instance, in our example NL-Query, firstly, the system identifies "Bill Clinton" as a central term and then "daughter" as direct dependent of "Bill Clinton" followed by "married" as direct dependent of "daughter" shown in Fig.1.

Semantic Similarity and Relatedness A semantic similarity can be defined on the basis of taxonomic (is-a) relations of two concepts, while relatedness covers a broad range of relations, e.g. meronym and antonym. In our problem space, we want to get the best semantic match of terms appearing in the NL-Query to the vocabulary of the DBpedia dataset. We can not however rely solely on semantic similarity measures (as in our example NL-Query), as we can see relatedness can better map the term "married" on and the retrieved property "spouse" as they are semantically related terms but not semantically similar.

To find the best semantic match we are investigating two approaches for semantic relatedness, i.e. Wikipedia based Explicit Semantic Analysis (ESA) [2] and a semantic relatedness measure based on WordNet structure [3]. Due to the computational cost involved in getting the relatedness measure using ESA, currently we are experimenting with measures based on WordNet only.

Evaluation

To evaluate over approach, we calculate average precision, average recall and F1 score of the results obtained by our approach on QALD-2 test dataset, which includes 100 NL-Queries over DBpedia. The results are shown in Table 1.

Total	Answered	Right	Partially right	Avg. Precision	Avg. Recall	F1
100	80	32	7	0.44	0.48	0.46

Table 1. Evaluation on QALD-2 test dataset of 100 NL-Queries over DBpedia

Conclusion and Future Work

This paper presented a system for natural language querying over Linked Data, which includes query annotation, a deep linguistic analysis and semantic similarity/relatedness. Currently, our approach does not fully explore all the types of queries appeared in dataset as it consists more challenging complex NL-Queries such as SPARQL aggregation and ask type queries. Future work will concentrate on improving the annotation step with better handling linguistic variations and a sophisticated semantic similarity/relatedness measures by that taking contextual information into account.

References

1. Freitas, A., Oliveira, J. G., O'Riain, S., Curry, E., Da Silva, J. C. P.: Querying linked data using semantic relatedness: a vocabulary independent approach. In: Proc. of NLDB'11 (2011)
2. Gabrilovich, E., Markovitch, S.: Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In: Proc. of IJCAI 2007
3. Pirró, G.: A semantic similarity metric combining features and intrinsic information content. Data Knowl. Eng. **68**(11) (2009) 1289–1308