Abstract

Question Answering aims at allowing users to ask a computer arbitrary questions and get correct answers back quickly and concisely. The current QA trend is moving towards answering questions from linked data resources. The objective of this thesis is to evaluate relational patterns extracted from Wikipedia and to carry out a feasibility study on the use of these relational patterns, rather than those extracted from a corpus of questions, in entailment based QA over linked data. Entailment based QA uses Recognizing Textual Entailment (RTE) paradigm. RTE relies on relational patterns that represent the various ways in which a particular relation between entities could be expressed in surface form. Linked data is based on subject, predicate and object model and the relational patterns represent the various ways in which the predicate can be expressed.

To achieve this objective, Wikipedia is used as the source of free text from where the relational patterns are extracted and DBpedia is used as an example linked data resource.This is because DBpedia was built by extracting structured information from Wikipedia, especially from the infoboxes. 10 relations, from DBpedia ontology, are used in the experiment. To acquire sentences expressing the chosen relations, the domain and range values (from the infobox) were looked up in their corresponding Wikipedia articles and only those sentences with a mention of both the domain and range were extracted as the sentences expressing that particular relation. These sentences were annotated with tags for the unit of words expressing the target relation and tags for the domain and the range values. With this, a gold-standard of annotated sentences was created. The results of the inter-annotator agreement show a high level of agreement and consistency, with respect to the built annotation guidelines, between the annotators. The average results are 0.76 for sentence agreement, 0.85 for domain tag agreement, 0.82 for range tag agreement and 0.70 for relation tag agreement.

The gold-standard sentences are used to evaluate the automatically extracted sentences and relational patterns from Wikipedia. From the gold-standard sentences, gold-standard relational patterns were extracted i.e. each relational pattern is an ordered set of domain, range and relation tags found in an annotated sentence.

Sample uses questions expressing the target relations are also acquired from the Web. The gold-standard and auto-extracted relational patterns are used (separately) to evaluate the question-pattern mapping module. Smith-Waterman string similarity metric is used in the question-pattern mapping module to measure the similarity between a user question and the set of relational patterns so as to get the relational pattern that is most similar to the user question. The results of the pattern mapping are positive, with 79% of the questions being correctly mapped to the matching pattern when using the auto-extracted patterns and 90% of the questions being correctly mapped when using the gold-standard patterns. These results show that the idea of using relational patterns extracted from text in entailment based QA over linked data is worth pursuing further.

Keywords:

Linked data, Textual Entailment, Question Answering, Relational Patterns, Relation Extraction, Inter-annotator Agreement.