Linguistic-based SPARQL translation model for semantic question answering system

ABSTRACT

Semantic Question Answering (SQA) aims to translate natural language (NL) questions to Simple Protocol and RDF Query Language (SPARQL) queries to retrieve answer from linked data. SQA deals with the complexity of NL questions because of the usersô styles of writing. Furthermore, the process to construct the SPARQL query to retrieve answer from linked data is complex due to the different merging scenarios depending on the six meta-mapping aspects: (1) the question type; (2) the sequence of important POS tags; (3) the preposition occurrence (4) the datatype of the matched RDF triples; (5) the resource heterogeinity; (6) the structure of the matched RDF triples. To date, most existing researchers on SQA system have treated the focus for SQA system to accept complex NL question separately from the focus to address meta-mapping scenarios. The motivation of this study is to design and develop an SQA system that accepts complex NL questions while addressing the meta-mapping scenarios. This is vital because each user has their own idiosyncrasy in composing NL question which needs to be translated to SPARQL query that involve different merging meta-mapping scenarios. We designed the selective POS tag extraction technique and the semantic representation composition technique to handle the complex NL questions. Meanwhile, we formulated a new linguistic-based SPARQL translation model to address the meta-mapping scenarios. The model is formulated using our proposed QALD dataset analysis methodology which can also be used by other researchers to implement on any QALD dataset. Model-Driven Semantic Question Answering (MDSQA) system that is integrated with the two techniques and formulated model is developed to automate the translation of the NL questions to SPARQL queries. MDSQA is evaluated using the QALD-3 test dataset that consists of 100 NL questions as input. The output of the MDSQA are the constructed SPARQL queries. The evaluation results are derived by comparing the constructed SPARQL queries against the actual SPARQL queries provided by the QALD-3 test dataset. MDSQA is able to process all complex NL questions in QALD-3 which consist of simple and complex NL questions without any manual modification of the question. Based on precision and recall of answer type, SPARQL query form, number of triples, placement of triples and SPARQL condition, MDSQA is capable of addressing meta-mapping scenario. Further enhancement is needed to address the drawbacks of this approach.

Keyword: Complex question; Linked data; Natural language question; SPARQL; Semantic question answering