A Guided Template-Based Question Answering System over Knowledge Graphs

Lukas Biermann[1,2], Sebastian Walter[1], and Philipp Cimiano[1,2]

[1] Knowledge Engineering Team, Semalytix GmbH
[2] Semantic Computing Group, Bielefeld University

Abstract. Question answering systems provide easy access to structured data, in particular RDF data. However, the user experience is often negatively affected by questions that are not interpreted correctly. To remedy this, we present a new guided approach to QA that ensures that all questions that can be entered into the system also return a corresponding answer. For this, a template-based approach is used to generate all possible questions from a given RDF dataset using a number of templates. The question/answer pairs can then be indexed to provide autocompletion functionality at querying time. We describe the architecture and approach and present preliminary evaluation results.

Keywords: Template based QA · QALD · Knowledge Graph

1 Introduction

Question answering systems over linked data (QALD) often suffer from their brittleness, that is the relatively high probability of not being able to parse and execute a question correctly. This is problematic as it negatively affects the user experience and trust in a system. We hypothesize that user experience would be significantly increased if all questions that can be entered into a query interface would also produce an answer. Building on this hypothesis, we have developed a QALD system that provides a controlled interface such that every question entered is interpretable and returns and appropriate answer. This is accomplished by a template-based approach that matches basic graph patterns over the data and, using a lexicon, generates different variants of asking the same question. The question and corresponding answer are stored in an index and used to provide query writing support using auto-completion to propose possible continuations of a query. In this paper we present our approach, which has been implemented in Python and relies on Hbase as database for indexing question/answer pairs. We illustrate the workings of the system using DBpedia as knowledge graph. A live demo of the system will be shown during the demonstration session, which is available at:

https://qa.semalytix.de

2 Approach

Our overall approach is visualized in figure 1. In an offline process, a set of predefined templates are used to generate natural language questions and SPARQL queries automatically. The SPARQL queries are evaluated and the resulting question/answer pairs are stored and indexed in HBase[3]. On the basis of this index, questions are retrieved while the user is typing a question and possible completions are proposed in an auto-completion functionality.

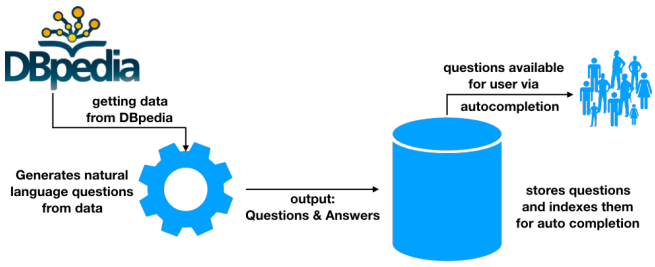

Fig. 1: System Overview

The retrieval of the data and the generation of the questions is displayed in more detail in figure 2. The templates currently supported by the system are given in the table below:

Template	Example question
Basic Triple (Noun/Verb)	Who is the wife of Barack Obama?
Verb with Prepositional Phrase	Which team Walter Payton played for?
Participle Construction	Basic In which programming language is gimp written?
Numeric Question	What is the frequency of BBC Radio Sheffield?
MaxCount	Who has the most alma maters?
Numeric Aggregation	Which Educational Institution has the lowest faculty size?
Numeric Filter	Give me all ships where the draft is less than <number>?
Ordering	Give me all american football players ordered by birth place?
Geographic	Give me all west german movies?

In order to illustrate the behaviour of the system, we focus on the *Basic Triple (Noun/Verb)* and the *Verb with Prepositional Phrase* pattern. We will assume that for the property *dbo:spouse* the following lexical variants are in the lexicon: *spouse, marry, wife, husband*. In these settings, the template-based question generation would generate the following questions for the property *dbo:spouse*:

The variance in the verbalizations for each template depends on the quality and coverage of the lemon lexicon. To extend our approach a template with the

[3] https://hbase.apache.org/

Who is the spouse of Barack Obama?	Who is married to Barack Obama?
Who is the wife of Barack Obama?	Who is the husband Barack Obama?
spouse of Barack Obama?	married to Barack Obama?
husband of Barack Obama?	wife Barack Obama?

corresponding graph pattern has to be defined. In the demo session we will show how straight forward it is to do so.

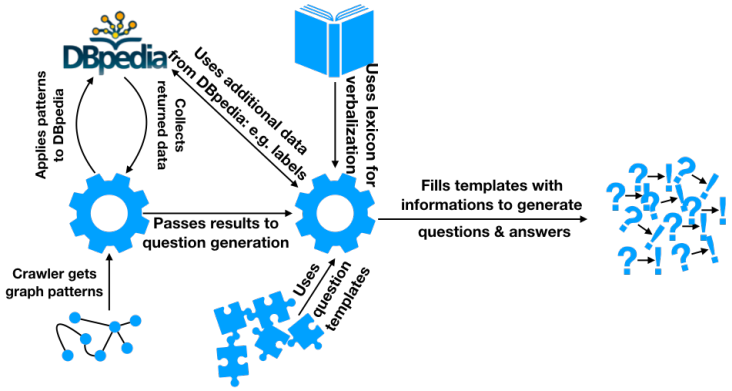

Fig. 2: Question generation workflow

3 Evaluation

For the evaluation the training data of QALD-5[3], QALD-6[4] and QALD-7[5] has been used. A total of 226 questions were used for the evaluation. Due to the limitation of the used lexicon, we were only able to answer questions for 32 out of 62 used properties, corresponding to a set 89 questions answered in total. Questions in QALD-7 were automatically matched to indexed questions using Levensthein distance, selecting the question minimizing the distance. By this, our system received a macro f-measure of 0.89%, outperforming AMAL (f-measure of 0.75% as presented in [5]). The latter was the best participant in QALD-7.

In addition to the automatic evaluation, we performed a manual evaluation in which we were able to identify types of questions that can not be answered given the current system. In particular, this analysis revealed that our system can not answer questions involving two conditions, such as *Give me all actors who were born in Paris after 1950*. One feasible solution for this example would be to generate a template which combines two kinds of questions, calculating the intersection between the answers, and using this data to generate new possible questions.

4 Conclusion

In this paper we have proposed a novel approach to guided QA that relies on a template-based approach to generate pairs of questions / answers offline and indexes them to support real-time auto-completion. Our system is very similar to a system presented earlier by Rico et al. [2], who also provides auto-completion functionality but relies on a different index. If users use the auto-completion functionality, then they are guaranteed to actually receive an answer to their question, a crucial feature from a usability point of view. We have shown that our system is competitive compared to the state-of-the-art. The main limitation is the need for a large enough lexicon covering different lexical variants for answering the same question. In the future, we plan to investigate if using an ontology lexicalization system such as MATOLL[6] can alleviate the problem. Alternatively we could use results from [1] to enrich the existing verbalization of properties.

References

1. D. Gerber and A.-C. Ngonga Ngomo. Bootstrapping the linked data web. In *Proceedings of the 1st Workshop on Web Scale Knowledge Extraction, workshop co-located with the 10th International Semantic Web Conference (ISWC 2011),Bonn, Germany, October 23-27*, 2011.
2. M. Rico, C. Unger, and P. Cimiano. Sorry, I only speak natural language: a pattern-based, data-driven and guided approach to mapping natural language queries to SPARQL. In *Proceedings of the 4th International Workshop on Intelligent Exploration of Semantic Data (IESD 2015) co-located with the 14th International Semantic Web Conference (ISWC 2015), Bethlehem, Pennsylvania , USA, October 12, 2015.*, volume 1472 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2015.
3. C. Unger, C. Forascu, V. López, A. N. Ngomo, E. Cabrio, P. Cimiano, and S. Walter. Question answering over linked data (QALD-5). In *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015.*, volume 1391 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2015.
4. C. Unger, A. N. Ngomo, and E. Cabrio. 6th open challenge on question answering over linked data (QALD-6). In *Semantic Web Challenges - Third SemWebEval Challenge at ESWC 2016, Heraklion, Crete, Greece, May 29 - June 2, 2016, Revised Selected Papers*, volume 641 of *Communications in Computer and Information Science*, pages 171–177. Springer, 2016.
5. R. Usbeck, A. N. Ngomo, B. Haarmann, A. Krithara, M. Röder, and G. Napolitano. 7th open challenge on question answering over linked data (QALD-7). In *Semantic Web Challenges - 4th SemWebEval Challenge at ESWC 2017, Portoroz, Slovenia, May 28 - June 1, 2017, Revised Selected Papers*, volume 769 of *Communications in Computer and Information Science*, pages 59–69. Springer, 2017.
6. S. Walter, C. Unger, and P. Cimiano. M-ATOLL: A framework for the lexicalization of ontologies in multiple languages. In *The Semantic Web - ISWC 2014 - 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part I*, volume 8796 of *Lecture Notes in Computer Science*, pages 472–486. Springer, 2014.