SuFLexQA*: an approach to Question Answering from the lexicon

Cristian A. Cardellino & Laura Alonso i Alemany

NLP group
FaMAF-UNC
Córdoba, Argentina
alemany@famaf.unc.edu.ar

Abstract. We present SuFLexQA, a system for Question Answering that integrates deep linguistic information from verbal lexica into Quepy, a generic framework for translating natural language questions into a query language. We are participating in the QALD-3 contest to assess the main achievements and shortcomings of the system.

Key words: Natural Language Processing, Question Answering, Lexicon

1 Introduction and Motivation

The amount of information that is available in RDF form, as Linked Open Data, grows day by day. For example, Freebase, the open core of Google Knowledge Graph, currently contains more than 1,000 million facts. All this information is readily available, inambiguous and interconnected to all kind of other related information. However, the typical end user does not access this kind of information because it is expressed in RDF (W3C 2004). While RDF is meant to keep some readability at human level, it is mostly intended to provide machine readability. For humans to access information stored in RDF, natural interfaces need to be provided.

Interfaces to RDF data have been mainly developed as Faceted Search (Tunkelang 2009). However, a smaller area is also developing, working on natural language interfaces to RDF data. The main goal of such interfaces is to translate questions into queries over RDF data, and to verbalize RDF graphs as answers to such questions. In this line of research, the QALD challenges within CLEF (CLEF 2013) have provided a common framework for evaluation.

We present SuFLexQA, a system for Question Answering that relies on information from a verbal lexicon to parse questions in natural language and translate them into a query language. Then, answers are provided by querying

*Part of the research in this paper was carried out within the framework of the project SKATeR: Scenario Knowledge Acquisition by Textual Reading, TIN2012-38584-C06-06, financed by the Ministry of Economy and Competition of Spain

an RDF store. Thus, SuFLexQA acts as a translation mechanism between natural language and a formal query language.

SuFLexQA builds upon Quepy, a python framework to transform natural language questions to queries in a database query language (Machinalis 2012). Quepy can be easily customized to different kinds of questions in natural language and database queries. The main development of SuFLexQA upon Quepy is that it incorporates a huge amount of regular expressions for parsing questions. These regular expressions are automatically obtained from three lexicons for English: VerbNet (Kipper, Dang, and Palmer 2000), FrameNet (Baker, Fillmore, and Lowe 1998) and PropBank (Kingsbury and Palmer 2002). They provide a description of verbs that allows to analyze the sentences where they occur and reconstruct the depicted scene. This interpretation is formally expressed and can be readily translated into the formal semantics of a knowledge base.

The rest of the paper is organized as follows. The next section provides some motivation for our work. Then, we describe the architecture of the system, and our participation in the open challenge QALD-3 to evaluate SuFLexQA. Finally, we outline some future developments.

2 Architecture of the System

2.1 The Quepy framework for translation of questions into queries

SuFLexQA is an instance of the Quepy framework. As such, its workflow is the same as for the whole framework, which is as follows:

1. A question in natural language is analyzed with NLTLK (Loper and Bird), and each word is associated to its lemma and part of speech tag.
2. The analyzed question is compared with a set of regular expressions, until a matching expression is found for the question.
3. The regular expression is associated to a semantics, determining the role that each matched element is to play in the resulting query. This semantics has to be compatible with the semantics of the RDF graph upon which the query is to be thrown.
4. A SPARQL query is generated.

Regular expressions are created manually. Quepy provides abstractions to express the different levels of analysis present in questions as a result of step **1.** However, since they are manually created, regular expressions are a bottleneck of the whole approach: in many cases, a question cannot be processed because no matching expression is found, thus the system fails in coverage. In other cases, the semantics associated to a regular expression is trivial or too shallow, thus failing in the deepness of the analysis.

We have tried to overcome both these shortcomings by creating regular expressions by automatically integrating the information in verbal lexica.

2.2 Incorporating Information from Verbal Lexica

We have integrated three verbal lexica into SuFLexQA : VerbNet (Kipper, Dang, and Palmer 2000), FrameNet (Baker, Fillmore, and Lowe 1998) and PropBank (Kingsbury and Palmer 2002). These lexicons provide a deep, rich description of verbs, based on their *subcategorization frames*: a description of how many and which arguments they take, and how they should be interpreted to reconstruct the scene depicted by each verb.

We have enriched the lexica with information from a big corpus, the OANC (OANC). From this corpus, we have obtained information about the realizations of their subcategorization frames with which each verb actually occurs, and their probabilities of occurrence. This information has been integrated with the information about the subcategorization frames of verbs.

Then, we have transformed the information in the lexica, basically the subcategorization frames of verbs, into regular expressions to parse questions. These regular expressions have been created in two phases: first, the subcategorization frame was transformed into a regular expression, then, this regular expression was transformed into a question, by transformational rules that account for the relation between enunciative and question utterances.

By doing this process, we have achieved two main objectives: first, a very broad coverage of English language and the questions that can be asked, by the automatic multiplication of regular expressions to parse questions. Second, each regular expression is associated to rich information on the semantic interpretation of each element that is matched, so that a formal representation of the utterance is readily obtained. With this formal representation, a translation into a query language is just a matter of mapping between the formal semantics of the lexica and the formal semantics of the RDF data, leaving behind the ambiguities and vagueness of natural language.

3 Evaluation

To evaluate SuFLexQA, we are participating in QALD-3 (CLEF 2013). The CLEF 2013 lab QALD-3 ist the third in a series of evaluation campaigns on question answering over linked data. We are participating in the challenge of Multilingual question answering, which consists in returning a correct answer or a SPARQL query that retrieves these answers given a RDF dataset (the DBpedia) and a natural language question or set of keywords.

4 Future Directions

SuFLexQA is currently under development. After participating in QALD-3, we will carry out a detailed error analysis to determine the most productive lines of future work. From our current perspective, we consider the following future lines:

- Incorporating automatic sense disambiguation into the analysis of questions, to determine the most adequate subcategorization frame for a given question.
- Incorporating lexicalized subcategorization frames.
- Finding mappings between the semantics of the lexica and the semantics of DBpedia by exploiting the lexical items associated to nodes in dbpedia and arguments in subcategorization frames.

References

[Baker, Fillmore, and Lowe 1998] Baker, C. F., C. J. Fillmore, and J. B. Lowe (1998). The Berkeley FrameNet project. In *COLING/ACL-98*, pp. 86–90.

[CLEF 2013] CLEF (2013). Qald, question answering over linked data. http://greententacle.techfak.uni-bielefeld.de/ cunger/qald/.

[DBpedia 2013] DBpedia (2013). `http://dbpedia.org/`.

[Kingsbury and Palmer 2002] Kingsbury, P. and M. Palmer (2002). From treebank to propbank. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002)*, Las Palmas, Spain.

[Kipper, Dang, and Palmer 2000] Kipper, K., H. T. Dang, and M. Palmer (2000). Class-based construction of a verb lexicon. In *Seventeenth National Conference on Artificial Intelligence*, Austin, Texas.

[Loper and Bird] Loper, E. and S. Bird. Natural language toolkit. `http://nltk.sf.net/`.

[Machinalis 2012] Machinalis (2012). Quepy, a python framework to transform natural language questions to queries in a database query language. `http://quepy.machinalis.com/`.

[OANC] OANC. Open american national corpus. `http://www.americannationalcorpus.org/OANC/index.html`.

[Tunkelang 2009] Tunkelang, D. (2009). Faceted search. *Synthesis Lectures on Information Concepts, Retrieval, and Services*.

[W3C 2004] W3C (2004). Resource description framework (rdf). `http://www.w3.org/RDF/`.