UNIVERSITÀ
DEGLI STUDI DI BARI
ALDO MORO

DIPARTIMENTO DI
INFORMATICA

PhD Program in Computer Science and Mathematics
XXXIII cycle

Research Project

PhD Student: Lucia Siciliani

Supervisor: Pasquale Lops

Coordinator: Prof. Maria F. Costabile

PhD student signature Lucia Siciliani

Supervisor signature Pasquale Lops

1. Research title

Question Answering over Knowledge Bases

2. Research area

Machine Learning and Data Mining: Text Mining, Natural Language Processing, Information Retrieval, Question Answering

3. Research motivation and objectives

Question Answering has been an important research topic in the field of Artificial Intelligence for many years.

Due to its nature, the developments in this area are strictly connected with the paradigms for Knowledge Representation.

The research about this topic began in the late sixties and early seventies, when the first Natural Language Interfaces were developed as a way to access data contained into databases [1].

Later on the attention focused over the extraction of relevant information from free text. Compared to traditional Information Retrieval systems, Question Answering systems allow the users to retrieve the information they are looking for in a easier way since the query is expressed through natural language instead of a set of keywords and the answer is presented as a concise answer rather than a list of documents.

Another significant milestone was the birth of the Semantic Web [2] which led to the spread of Knowledge Bases that encode an huge amount of information in the form of structured data.

Some examples of Knowledge Bases are DBpedia [3], YAGO [4] and Freebase [5] but there are more than a thousand datasets composing the so called Linked Open Data Cloud[1].

Although being an innovation, as more semantic data was published on the web, the challenges connected to how effectively exploit this huge amount of information arised.

The first problem to be dealt with regards the lexical gap between users and Linked Data.

Much of the spread of the Knowledge Bases was in fact due to the publication by the W3C[2] of the two standards for the publication and interrogation of semantic data: RDF[3] (Resource Description Framework) and SPARQL[4] (SPARQL Protocol and RDF Query Language). Nevertheless SPARQL, to be used effectively, requires a specific expertise difficult to acquire for common users.

In this context, the task that Question Answering systems try to accomplish is to construct a SPARQL query starting from a natural language question and this involves two main steps: mapping natural language expressions to the vocabulary elements used in the dataset and handling meaning variations introduced by ambiguous expressions.

Another challenge regards the performances and the scalability of the system. Since a dataset can be composed by billions of RDF triples, it is clear that in order to guarantee a real time answer, it necessary to use appropriate index structures, search heuristics and even distributed computing

[1] http://lod-cloud.net/

[2] https://www.w3.org/

[3] https://www.w3.org/RDF/

[4] https://www.w3.org/TR/sparql11-query/

principles. This issue is even more critical if we consider that some answers can be given only consulting more knowledges at the same time.

Despite the efforts made through all these years, a solution for all these problems is obviously really hard to get and this makes Question Answering one of the main research topics in the field of Natural Language Processing.

The aim of this project concerns the development of a novel model for Question Answering over Knowledge Bases. Through an in-depth study of the state of the art, the goal is to try to overcome the shortcomings that affect current systems. The proposed methods will be evaluated against the gold standard benchmarks commonly adopted by the research community in this field in order to provide a fair comparison.

4. State of the art

The problem of Question Answering over knowledge bases has been addressed in a wide variety of ways. Question answering over Knowledge Bases requires to fulfil different subtasks.

Following the partition proposed by [6] and [7], there are six steps that usually characterize a Question Answering system: *data preprocessing*, *question analysis*, *phrase mapping*, *disambiguation*, *query construction* and *distributed knowledge*.

The *data preprocessing* phase can be performed in order to help to reduce the overall running time of the system. Usually it consist in the indexing of the dataset information into specific structures, specifically designed to facilitate the retrieval performed on-line when answering a question.

The *question analysis* step is the first one where the user question is analyzed. This analysis ranges from the syntactic and semantic analysis, which includes part-of-speech tagging, parsing, named entity recognition and expected answer type identification among the others.

The next three steps, phrase mapping, disambiguation and query construction are the most important for question answering because it is where the system has to try to bridge the lexical gap (see section 3).

Given a question, the goal of the *phrase mapping* task is to search if there is any phrase within the question that actually matches the dataset terminology. This can be done in several ways, from a simple string matching to more sophisticated semantic matching approaches.

However, a single phrase can match with more than one entry in the knowledge base so a *disambiguation step* could be necessary in order to identify the right one. The context of a phrase is used as a hint to identify which one represents the right relationship.

Once all the mappings are performed, the question in natural language can finally be translated into its equivalent in SPARQL or any data query language required to retrieve the answer in the knowledge base. The *query construction* task is particularly tricky even when the question is quite simple and gets worse when it involves special operators such as quantifiers, comparatives, cardinals and superlatives.

Finally, the *distributed knowledge* task is performed when the answer can be found over distributed but linked datasets. Actually this problem has been addressed only by few systems but it is something that has to be faced in the future considering the evolution of the semantic web.

Since so many different steps have to be accomplished, every system can differ from the other on many aspects. The most distinctive one regards the way in which the query construction is performed. We can thus identify the following categories:

- *template based approaches;*
- *approaches based on the informations coming from the question analysis;*
- *approaches based on the Semantic Parsing;*
- *machine learning approaches;*
- *approaches based on semantic information;*
- *approaches not using SPARQL.*

One way to cope with the query construction issue is to make use of *templates*. Since most questions can be led back to the same semantic syntactic structure, it makes sense to use a set of templates defined by default. Each template reflects a semantic structure where entities and relation are substituted by empty slots which have to be filled online. For example the questions “What is the capital of Italy?” and “What is the currency of Japan?” both can be associated with the template “What is the ___ of ___?” where the first empty slot must be filled with a relation and the second one with an entity. Each template is correlated to a specific query thus, once the slots have been correctly filled, the translation process becomes straightforward. Methods that use this approach are QAKiS [8] and ISOFT [9]. The clear disadvantage of this method is that if a query doesn’t match one of the templates, then the system is infeasible to retrieve an answer.

Another approach is to base the query construction step over the information obtained in the question analysis part. This category includes several system that use different strategies, one of them is QAnswer [10]. In QAnswer, in order to generate the query, it is required that all the DBentities are correctly identified in the question and that they are properly linked. For this purpose, first a directed graph is created where the vertices correspond to tokens in the questions (that have been annotated with lemma and part-of-speech tags) while the edges correspond to the collapsed dependencies generated by Stanford Core NLP. Secondly, the DBpedia entities are detected with a different method depending on the type of entity. During these two phases, the system creates different graphs, one for every possible match, each one having a score. At the end of the process, the graph with the best score is selected and then used to build the query. The fact that this approach is based on the hypothesis that the from the structure of the question we can deduce the structure of the SPARQL query leds also to an important drawback since it does not take account for the way in which knowledge is encoded into the Knowledge Base.

Some systems adopt methods based on the use of a *semantic parser*. Through semantic parsing a sentence, written in natural language, is mapped to a formal representation for its meaning. There are many different kind of grammars that can be used for semantic parsers, such as: Grammatical Framework grammars, Feature-Based Context-Free grammars, Combinatory Categorial Grammar and Lexical Tree-Adjoint grammars. Without going into details, each phrase in the question can be associated with a syntactic category and a semantic representation. This semantic representation can vary for each system, for example in [11] it is expressed with the lambda calculus. An advantage of this kind of approach is that it can allow to handle even complex questions, such as those containing superlatives and comparatives. The disadvantages, on the other hand, are that the question has to be well formed and that the semantic representation require a corpus to be generated. Thus, if many lexical items don’t appear in the corpus this can lead to low recall.

Another way to address the query construction phase is to use *machine learning* techniques. In CASIA [12] a machine learning approach is used throughout the overall process of question answering. First of all, the phrase detection step is performed without a Named Entity Recognizer but retaining all the n-grams as candidate entities and then using rules to select only the right ones. Next they construct a candidate space for the mapping phrases and entities in the knowledge base. In order to select the right candidate for each phrase, the question is analyzed to extract some feature which are then used in a Markov Logic Network that combines Markov networks with first-order logic formulas. The inference results are then used to generate a SPARQL query.

Approaches using *semantic information* avoid the use syntactic features to construct the query. The Question Answering system SINA [13] first finds all the resources that appear in the user question and disambiguates them using an Hidden Markov Model, then constructs all the possible graphs which can represent the question. This is done by creating a vertex for each instance or class and an edge for each property. Then if the types of the range and domain are compatible the edges can be directly connected to the vertices, otherwise one or more vertices, representing the variables, are added to the graph. It is clear that this procedure can create more than one graph and these are all considered by the system. Ignoring the syntactic structure of the query can lead to heavy misinterpretations of the user intent.

Finally other system try to retrieve the answer from the Knowledge Base of choice without using SPARQL at all. The approach proposed in Treo [14] sees the process of retrieving the answer in the Knowledge Base as a graph exploration problem. The first step consists in the extraction of all the key entities in the questions, i.e. phrases that can be mapped to a resource in the knowledge base. Among all these key entities, exploiting a dependency graph, a pivot entity is identified thus the exploration of the Knowledge Base graph start from that point. While exploring the graph, all the key entities previously found in the question are progressively connected, leading to the resource representing the final answer. The main bottleneck is that an exhaustive search on the graph is obviously impossible, so some heuristics must be used. Another issues arises when the question does not have a one-to-one correspondence with a path in the graph: in this case the search stops and an answer can not be given.

5. Approach

For my Master Thesis I have developed a Question Answering system over Linked Open Data extending the approach proposed in CANaLI [15] applying some distributional semantics techniques. The method proposed in CANaLI is based on the use of controlled languages. Given a language, we obtain a controlled language by considering only a subset of its vocabulary and its grammatical rules. In this way it is possible to build a finite automata that is capable to recognize any sentence written in that controlled natural language.

This work allowed me to grasp the problem of question answering over knowledge bases and have a sample of the main issues that must be faced when developing a Question Answering system. This kind of approach in fact has a great drawback: if a question does not follow the syntax allowed by the controlled grammar, then is impossible to generate the SPARQL query and retrieve the answer. Therefore the aim of this project is to propose a new model for Question Answering over Knowledge Bases capable to mitigate the major drawbacks observed in the literature.

For this purpose it is necessary to carry out not only a deeper study of the state-of-the-art systems, but also of the single techniques that can be used in each phase of the Question Answering process. Another aspect to care about regards Knowledge Bases themself. It is important to understand how the most used Knowledge Bases are structured in order to take full advantage of the information that they contain. Moreover evaluations will be conducted to make a comparison with works already present in literature.

6. Expected results

The field of application for Question Answering systems over Knowledge Bases is very wide. As explained in section 3, the research in this field began for pragmatic reasons, mainly bound to the spread of Data Bases.

Through Question Answering the user is capable of finding the information he is looking for in a easier way since there is no need to use a keyword search or a specific data query language which can be difficult to use for an inexpert user. From this point of view a Question Answering system gives more support to the users than a classical Information Retrieval one.

Thus, a Question Answering module can be integrated in every system that needs a user interface for its data, for example an ontology or a database. For this purpose it is important to focus the research on methods that are open-domain: this certainly makes the task more difficult but brings the important advantage that in this way the system becomes applicable in different contexts.

One of the major challenges for Question Answering regards multilinguality: even if the Semantic Web is suited for this, since the URIs used to identify each resource are language independent, it is not the same for their labels. Moreover the number of users who do not speak English as native language is constantly growing and this makes this issue even more compelling.

During the Ph.D, the proposed model will be applied to two European projects.

The first one is SEO-DWARF (Semantic EO Data Web Alert and Retrieval Framework) project that is related to the Marie Sklodowska-Curie Research and Innovation Staff Exchange (MSCA-RISE) funding of the Horizon 2020 Programme. The main goal of this project is to create a system capable of exploiting the information coming from remote sensing applications for the marine domain. All the data fetched by the satellites will be organized in an ontology, so a model of Question Answering will be used in order to allow the retrieval of relevant informations through the use of queries formulated in natural language.

The second project is called TALIA (Territorial Appropriation of Leading-edge Innovation Action), which is part of Interreg-Mediterranean Programme. The aim of this project is to improve the policy impact of different modular projects and allow for individual projects to work together as a system. Within this project, the main contribution will be given by exploiting the given ontology to retrieve information about the different targets and stakeholders.

7. Phases of the project

The activities that will be carried out over the three years of Ph.D can be scheduled as follows:

- First year
    - Study of the approaches proposed in the literature for Question Answering over Linked Data;
    - In-depth study of the main advantages and drawbacks of each approach;
    - Study of deep learning approaches applied to Natural Language Processing;
    - Analysis of the most relevant Knowledge Bases used for Question Answering;

    - Gathering and analysis of the datasets used for the evaluation of Question Answering Systems;
    - Participation to international schools, conferences, workshop and doctoral consortium.
- Second year:
    - Definition of a new proposal to address the problem;
    - Implementation of the proposed model;
    - Definition of a test plan to evaluate the model and compare its results with the state of the art systems;
    - Submission of the obtained results in national and international journals and conferences.
- Third year
    - Identification of the possible improvements that can be applied to the model;
    - Development of a new version of the model based on the previous results;
    - Integration and evaluation of the model in the industrial field.
    - Internship period to international research groups relevant for the research topic;
    - Drafting of the Ph.D thesis.

8. Result evaluation

Until 2011 there was still no official benchmark to evaluate a Question Answering system with and so a comparison was almost infeasible however, in the last years, the growing interest towards this topic is leading to a solution of this problem.

Three major benchmarks for Question Answering system over Knowledge Bases are: QALD [16, 17, 18, 19, 20], WebQuestions [21] and SimpleQuestions [22].

In particular, QALD is not a single benchmark but a series of evaluation campaigns for Question Answering over Knowledge Bases.

So far there have been held eight editions of this challenge and the last one took place during the International Semantic Web Conference in October.

The main difference between WebQuestions, SimpleQuestions and QALD is that the first two contain questions that can be answered using Freebase while the last involves the use of DBpedia besides some other Knowledge Bases.

Other discrepancies regard the way the question are generated (manually for QALD, with a crowd-sourcing method for WebQuestions and SimpleQuestions) and obviously the dimension of the datasets.

Regardless of the benchmark that is used, the evaluation of a Question Answering system is performed in accordance with three parameters: precision, recall and F-measure.

Precision assesses the number of correct answers for a given question. Given a question q, it is computed as follows:

$$precision(q) = \frac{number\ of\ correct\ system\ answers\ for\ q}{number\ of\ system\ answers\ for\ q}$$

Recall must be computed taking account of the set of expected correct answers which, in literature, are called the gold standard answers. The recall indicates how many of the gold standard answers are actually returned by the system. Given a question q, it is computed as follows:

$$recall(q) = \frac{number\ of\ correct\ system\ answers\ for\ q}{number\ of\ gold\ standard\ answers\ for\ q}$$

The aforementioned measures are then used to compute the global precision and recall of the system.

This can be done in two different ways that lead to the so called micro/macro precision and recall. The micro precision and recall are computed by making the average of the precision or the recall without taking into account the questions not answered by the system, otherwise, if we consider all the questions that compose the dataset, we obtain the macro precision and recall.

Lastly the micro/macro F-Measure is the weighted average between the micro/macro precision and recall. It is computed as follows:

$$F - measure = 2 \times \frac{precision \times recall}{precision + recall}$$

9. Possible reference persons external to the department

Possible external scientific referents will be identified during the three years of doctorate, during participation in summer schools, conferences, workshop and doctoral consortium. However, it is possible to provide a preliminary list of researches in the field of Question Answering over Knowledge Bases that will be taken into account.

- *André Freitas,* lecturer at the School of Computer Science at the University of Manchester;
- *Dennis Diefenbach*, Researcher and Doctorand at Laboratoire Hubert Curien;
- *Philipp Cimiano,* head of the Semantic Computing Group at Bielefeld University;
- *Christina Unger*, Research Associate at CITEC at Bielefeld University.

10. References

[1] Androutsopoulos, I., Ritchie, G. D., & Thanisch, P. (1995). Natural language interfaces to databases–an introduction. Natural language engineering, 1(1), 29-81.

[2] Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. Scientific american, 284(5), 28-37.

[3] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. The semantic web, 722-735.

[4] Suchanek, F. M., Kasneci, G., & Weikum, G. (2007, May). Yago: a core of semantic knowledge. In Proceedings of the 16th international conference on World Wide Web (pp. 697-706). ACM.

[5] Bollacker, K., Evans, C., Paritosh, P., Sturge, T., & Taylor, J. (2008, June). Freebase: a collaboratively created graph database for structuring human knowledge. In Proceedings of the 2008 ACM SIGMOD international conference on Management of data (pp. 1247-1250). AcM.

[6] Diefenbach, D., Lopez, V., Singh, K., Pierre, M.: Core Techniques of Question Answering Systems over Knowledge Bases: a Survey. Knowledge and Information systems (2017 (to appear))

[7] Unger, C., Freitas, A., & Cimiano, P. (2014). An introduction to question answering over linked data. In Reasoning Web. Reasoning on the Web in the Big Data Era (pp. 100-140). Springer International Publishing.

[8] Cabrio, E., Cojan, J., Aprosio, A. P., Magnini, B., Lavelli, A., & Gandon, F. (2012, November). QAKiS: an open domain QA system based on relational patterns. In Proceedings of the 2012th International Conference on Posters & Demonstrations Track-Volume 914 (pp. 9-12). CEUR-WS. org.

[9] Kwon, S., Park, S., Nam, D., Lee, K., Yu, H., & Lee, G. G. (2016). ISOFT-Team at NTCIR-12 QALab-2: Using Choice Verification. In NTCIR.

[10] Ruseti, S., Mirea, A., Rebedea, T., & Trausan-Matu, S. (2015). QAnswer-Enhanced Entity Matching for Question Answering over Linked Data. In CLEF (Working Notes).

[11] Hakimov, S., Unger, C., Walter, S., & Cimiano, P. (2015, June). Applying semantic parsing to question answering over linked data: Addressing the lexical gap. In International Conference on Applications of Natural Language to Information Systems(pp. 103-109). Springer, Cham.

[12] He, S., Zhang, Y., Liu, K., & Zhao, J. (2014, September). CASIA@ V2: A MLN-based Question Answering System over Linked Data. In CLEF (Working Notes) (pp. 1249-1259).

[13] Shekarpour, S., Marx, E., Ngomo, A. C. N., & Auer, S. (2015). Sina: Semantic interpretation of user queries for question answering on interlinked data. Web Semantics: Science, Services and Agents on the World Wide Web, 30, 39-51.

[14] Freitas, A., Oliveira, J. G., Curry, E., O'Riain, S., & da Silva, J. C. P. (2011, June). Treo: combining entity-search, spreading activation and semantic relatedness for querying linked data. In Proc. of 1st Workshop on Question Answering over Linked Data (QALD-1) at the 8th Extended Semantic Web Conference (ESWC 2011).

[15] Mazzeo, G. M., & Zaniolo, C. (2016). Answering Controlled Natural Language Questions on RDF Knowledge Bases. In EDBT (pp. 608-611).

[16] Lopez, V., Unger, C., Cimiano, P., & Motta, E. (2013). Evaluating question answering over linked data. Web Semantics: Science, Services and Agents on the World Wide Web, 21, 3-13.

[17] Cimiano, P., Lopez, V., Unger, C., Cabrio, E., Ngomo, A. C. N., & Walter, S. (2013, September). Multilingual question answering over linked data (qald-3): Lab overview. In International Conference of the Cross-Language Evaluation Forum for European Languages (pp. 321-332). Springer, Berlin, Heidelberg.

[18] Unger, C., Forascu, C., Lopez, V., Ngomo, A. C. N., Cabrio, E., Cimiano, P., & Walter, S. (2014, September). Question answering over linked data (QALD-4). In Working Notes for CLEF 2014 Conference.

[19] Unger C, Forascu C, Lopez V, et al. Question Answering over Linked Data (QALD-5). In: Cappellato L, Ferro N, Jones G, San Juan E, eds. Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum. Vol 1391. Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum; 2015. **

[20] Unger C., Ngomo AC.N., Cabrio E. (2016) 6th Open Challenge on Question Answering over Linked Data (QALD-6). In: Sack H., Dietze S., Tordai A., Lange C. (eds) Semantic Web Challenges. SemWebEval 2016. Communications in Computer and Information Science, vol 641. Springer, Cham

[21] Berant, J., Chou, A., Frostig, R., & Liang, P. (2013, October). Semantic Parsing on Freebase from Question-Answer Pairs. In EMNLP (Vol. 2, No. 5, p. 6).

[22] Bordes, A., Usunier, N., Chopra, S., & Weston, J. (2015). Large-scale simple question answering with memory networks. arXiv preprint arXiv:1506.02075.