AN ONTOLOGY-BASED QUESTION ANSWERING SYSTEM EXPLOITING SEARCH ENGINES' RESULTS

Plegas Yannis and Kafeza Evanthia
Computer Engineering and Informatics Department, University of Patras, Rio 26500, Patras, Greece

Keywords: Question Answering, Semantic Web, Ontologies, Queries, Application, Natural Language, Syntactic Analysis, Semantics, Search Engines.

Abstract: This paper proposes a Semantic Web Application which extends search engines giving them the ability to answer natural language queries. The application builds an automatic ontology-based question-answering system from the texts of the search engine's results. The automatic question- answering system converts the results of a traditional search engine to ontology, integrating syntactic analysis as well as the respective semantic rules. The main idea of this paper is the use of OWL Description Logic rules to model the human logic for the process of an answer in a question through the syntactic structure of the texts which contain the answer. Specifically, it is described the process for the creation of the ontology and the construction of the proper queries to the ontology for each type of question through the OWL semantic web language.

1 INTRODUCTION

For many years, search engines did not focus on natural language questions. Answering questions in natural language has obtained much attention during the last years, incorporating semantic information in their search results. Many of them try to face this disadvantage by adding to Web pages rich semantic snippets (like microdata, microformats, and RDFa) in order to be able to accept semantically enhanced queries as cited in Heinrich and Gaedke(2011); Khare (2006), and Delmonte and Tripodi (2011).

This paper presents an attempt of combating this shortcoming of search engines, creating a prototype question answering system that exploits the search engines' results. The application handles the queries, which are submitted to the search engine as natural language questions (syntactically structured sentences), in contrast to search engines that handle them as sets of keywords. Consequently, and in order to avoid any misunderstanding, the natural language questions are called natural language queries when they are submitted to the search engines for the rest of the work.

Our aim is to take advantage of the speed and accuracy of general-purpose search engines in order to create a small set of texts very quickly, before applying our methods. The application achieves two main goals: a) enables answers to natural language queries very fast, and b) performs semantic classification of the obtained results based on their relevancy to the query.

The implemented Semantic Web model combines two main axes. One of the two axes is the syntactic parsing of the texts, and the other is the language for authoring ontologies, the OWL semantic web language. The OWL language converts the information of syntactic parsing and the syntactic rules in a format recognizable by the computer, using an ontology. The aim of the ontology is the representation of the human logic for the answer of a question.

The rest of the paper is organized as follows. Section 2 provides the main characteristics of previous and related work. Section 3 contains the Description Logic Rules for the creation of the ontology. Section 4 describes the main part of te application, the question answering system. Section 5 discusses the experimental process and results. Finally, Section 6 reports some conclusions and discusses the future work.

2 PREVIOUS AND RELATED WORK

There are plenty of published works, but the paper will focus in the most related approaches towards our application. The following papers are trying to

Yannis P. and Evanthia K..
AN ONTOLOGY-BASED QUESTION ANSWERING SYSTEM EXPLOITING SEARCH ENGINES' RESULTS.
DOI: 10.5220/0003929404260429
In *Proceedings of the 8th International Conference on Web Information Systems and Technologies* (WEBIST-2012), pages 426-429
ISBN: 978-989-8565-08-2

incorporate semantic information, or make use of ontologies, to answer natural language questions.

A related approach to our work, using triplets (subject, verb, and object) to answer questions is that by Lorand, Rusu, Fortuna, Mladenic and Grobelnik (2009). The main difference with our work is that the questions are directly applicable to the texts. In the present study, the data carrying semantic meaning creates an ontology, from which the answers are exported. Saias and Quaresma (2003) allow users to query the semantic content of the documents using ontologies. Kotov and Zhai (2010) propose a new framework for question-guided search, in which a retrieval system would automatically generate potentially interesting questions to users based on the search results of a query. Moreover, Moise and Gheorghe (2010) answer a predefined set of questions using text patterns. Additionally, there are question answering systems like the QuestIO from Damljanovic et al. (2008), and TextRunner from Yates, Cafarella, Banko, Etzioni, Broadhead and Soderland (2007).

3 DESCRIPTION LOGIC RULES AND ONTOLOGY

This section describes the logic for the creation of the ontology through the texts of the search engines' results. After the definition of the ontology, the Hermit reasoner is used to extract the knowledge base model as cited in Shearer, Motik and Horrocks (2008).

Initially the morphosyntactic structure of the texts is extracted. Specifically, the texts are organized in different levels. In the first level, the texts are divided into sentences. Then every sentence is divided into two parts: the nominal part NP which is the subject of the sentence and its determinations, and the verbal part VP which is the rest of the sentence; the verb, the object and subordinate clauses that may exist. The nominal part and the verbal part are divided in different depth syntactic levels, until the parser reaches the level of words.

For the representation of the information contained in a sentence, a data structure which is called triplet is used. A triplet consists of the subject, the verb and the object of the sentence as cited in Lorand et al. (2009). The triplets incorporate the syntactic rules in the ontology after the insertion of the syntactic structure of the text.

The Ontology consists of a set of the following concepts: Named Classes (A), Individuals (o) and Named Properties (P). Each level of the morphosyntactic analysis marks a new level in the ontology's structure. The root of the text's subclass is the title of the text. For each sentence a new subclass is imported under the root in the ontology representing its structure. Then, for each pos tag assigned from the parser, a corresponding *named class* is created in the ontology such as NPsub for the nominal part of the subject or NPobj for the nominal part of the object of each sentence. For each word of the text, an *individual* of the class corresponding to the pos-tag of the word is imported into the ontology.

The rules for the creation of most appropriate triplets are defined below.

- If the subject of the sentence is a personal pronoun, there are two different cases in order to determine it. The subject can be found either in a subordinate clause of the same sentence which should be preceded by the main clause or in the previous sentence.
- In the main sentence, if the verb has the pos tag VBZ or VBP then the creation of the triplet is simple. The subject of the sentence is the nominal total which is in the same level with the superclass of the class-verb. The object is the nominal total which is located to the next syntactic level of the verb's class. If another verb exists with pos tag VBN then this is a verb in indefinite tense. The combination of these two verbs should be one tense in the sentence's triplet. The extraction of the sentence's object is complicated in this case, because the classes located in the same level of the verb with the tag VBN must be checked and can be more than one. Into the ontology, the class VBZ or VB, which corresponds to the auxiliary verb, must be equivalent with the class VBN. The previous rule also applies to subordinate clauses.
- When a relative pronoun exists in the sentence, additional triplets should be created to determine where the pronoun refers. The nominal total, in which the pronoun refers, would be at the classes located in the same level of the relative pronoun. So the right triplet is created defining a new rule which presupposes that the class WDT must be equivalent class with the resulting class NPobj.

The *named properties* are presenting the relationships between the classes or the individuals. It is necessary to create a property for the definition of the triplet's relations with domain class the NPsub and range class the NPobj. Some sentences contain

important determinations for the proper construction of the ontology. For the above reason additional properties must be created as sub-properties of the above property. Moreover, the relation between the verb and the property is necessary and is created by setting the property which indicates the existence of the verb, as disjoint property of the above property. Also, for each verb an assertion is created which shows the relation between the individual and its class.

4 AUTOMATIC QUESTION ANSWERING SYSTEM

In this section we describe the Question Answering System which is the main part of our application and its architecture is shown in Figure 1. We begin with the conversion of the search engine's results into an ontology, using the described rules in Section 3. The next action after the creation of the ontology, is the construction of the query which is applied to the ontology in order to retrieve the answers.

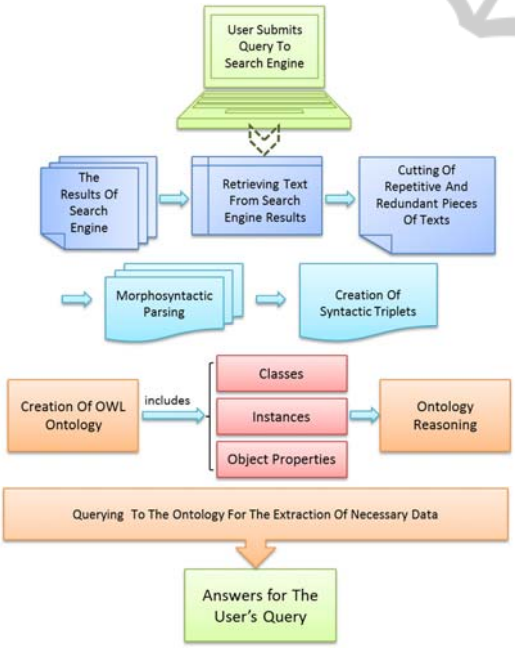

Figure 1: From top to bottom question answering architecture.

Initially, the user's query must be analyzed syntactically as a question. The questions can be classified into different categories. The syntactic parsing of the questions determines the object properties which must be searched in the ontology. The following types of questions are supported by the system:

- Yes/No questions (Does science have future?),
- List questions (What is computer science?),
- Reason questions (Why is science important?)
- Quantity questions (How many questions do computer engineers ask?),
- Location questions (Where can I buy a coffee?)
- Time questions (When the train was arrived?)

The second step is constituted by the development of the appropriate query based on the user's question in order to extract from the ontology the possible answers. To be successful the extraction of the proper answers, the reasoner infers the logical consequences of a set of attested facts or axioms in the ontology.

The set of sentences S is created seeking all the object properties which are referred in verbs that have as instances the verb of the question. Next, depending on the type of question, the classes of the above object properties, are used for the extraction of the proper individuals. These individuals must be identical to the nominal set of the user's question. The composition of the above individuals are creating the proper sentences as answers to the user's question. These sentences are generating the set of sentences SR.

The overall process is shown below, written in three basic steps:

Step 1: Submission of a natural language query in the search engine.
Step 2: Morphosyntactic parsing of the search engine's results and creation of the Ontology.
Step 3: Execution of the appropriate query in the ontology and return of the answers back to the user.

5 EXPERIMENTS

In order to evaluate the proposed systems, experiments were carried out to ensure, that the questions are answered correctly. The dataset used in the experiments is the Category B part of the ClueWeb09 Dataset of Lemur Project (lemurproject.org, 2009).

The proposed application is applied to two different systems. The first is a standalone application, which uses as input data, the results of a locally installed instance of the Indri search engine (Strohman, Metzler, Turtle and Croft, 2000). The Indri search engine has the ability to search for information in two parts of the ClueWeb09 Dataset, the English Wikipedia and the Category B Dataset

(which includes the English Wikipedia). Moreover the application has been developed as an add-on tool for Internet Explorer. The created tool is taking as input data the Google's results and returns the answers together with the initial results.

The evaluation process includes the execution of one hundred questions in the two systems. The questions were constructed from the queries of the Web Tracks 2009 and 2010(TREC Collections). One question deemed to have been answered correctly, when all the correct answers with their respective texts are returned.

Tables 1 and 2 contain the percentage of the correct answers in the two systems for our dataset and show clearly that our application answers satisfactory the questions.

Table 1: Percentage of correct answers for Indri.

	English Wikipedia	Category B
Percentage	75%	94%

Table 2: Percentage of correct answers for Google.

	Google
Percentage	66%

6 CONCLUSIONS AND FUTURE WORK

This paper presents a novel idea which allows search engines to quickly answer to natural language queries locally. We have also presented a prototype system based on the integration in a unified ontology of the texts of the search engine's results, together with their syntactic structure. Then applying reasoning tools in the ontology, the answers are extracted by executing specific queries.

Future improvements to the question answering system could be the development of a module for automatic adjustment of questions submitted in the wrong way by the user. Moreover we plan to speed up even further the whole computation by employing various optimization techniques.

ACKNOWLEDGEMENTS

This research has been co-financed by the European Union (European Social Fund-ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF)-Research Funding Program: Heracleitus II. Investing in knowledge society through the European Social Fund.

REFERENCES

Croft, W., Callan, J., Allan, J., Zhai, C., Fisher, D., Avrahami, T., Strohman, T., Metzler, D., Ogilvie, P., Hoy, M., Lafferty, J., Brown, J., Si, L., Collins-Thompson, K., Bilotti, M., Feng, F., and Larkey, L., 2006. *The Lemur Project.* Available at: <http://www.lemurproject.org/>, <http://lemurproject.org/clueweb09.php/>

Damljanovic, D., Tablan, V. and Bontcheva, K., 2008. *A text-based query interface to owl ontologies.* The 6th Language Resources and Evaluation Conference (LREC).

Delmonte, R. and Tripodi, R., 2011. *Linguistically-Based Reranking of Google's Snippets with GreG.* Studies in Computational Intelligence,Vol. 361, p. 59-79.

Heinrich, M., Gaedke, M., 2011. *WebSoDa: A Tailored Data Binding Framework for Web Programmers Leveraging the WebSocket Protocol and HTML5 Microdata.* Lecture Notes in Computer Science.

Khare, R., 2006. *Microformats: the next (small) thing on the semantic Web?.* Internet Computing, IEEE 2006.

Kotov, A., Zhai, C., 2010. *Towards natural question guided search.* WWW '10 Proceedings of the 19th international conference on World wide web ACM NewYork. Available at: <http://portal.acm.org/citation.cfm?id=1772690&picked=prox&cfid=27961346&cftoken=36016737>.

Lorand, D., Rusu, D., Fortuna, B., Mladenic, D. and Grobelnik, M., 2009. *Question answering based on semantic graphs.* Proc. of the Workshop on Semantic Search.

Moise, M., Gheorghe, C., 2010. *Developing question answering (QA) systems using the patterns*. WSEAS Transactions on Computers.

Saias, J. and Quaresma, P., 2003. *A methodology to create ontology-based information retrieval systems.* Lecture Notes in Computer Science, Vol. 2902, p.424-434.

Shearer, R., Motik, B. and Horrocks, I., 2008. *HermiT: a Highly-Efficient OWL Reasoner*. In Proceedings of OWLED'2008.

Strohman, T., Metzler, D., Turtle, H. and Croft, W., 2000. *Indri: A language-model based search engine for complex queries.* Center for Intelligence Information Retrieval University of Massachusetts Amherst, Available at: <http://lemurproject.org/indri/>.

Yates A., Cafarella M., Banko M., Etzioni O., Broadhead M., Soderland S., 2007. *TextRunner: open information extraction on the web.* Proceedings of Human Language Technologies.