AnswerArt - Contextualized Question Answering

Lorand Dali, Delia Rusu, Blaž Fortuna, Dunja Mladenić, and Marko Grobelnik

Jožef Stefan Institute, Department of Knowledge Technologies, Jamova 39,
1000 Ljubljana, Slovenia
{Lorand.Dali,Delia.Rusu,Blaz.Fortuna,
Dunja.Mladenic,Marko.Grobelnik}@ijs.si

Abstract. The focus of this paper is a question answering system, where the answers are retrieved from a collection of textual documents. The system also includes automatic document summarization and document visualization by means of a semantic graph. The information extracted from the documents is stored as subject-predicate-object triplets, and the indexed terms are expanded using Cyc, a large common sense ontology.

Keywords: question answering, summarization, ontology.

1 Introduction

We describe an enhanced question answering system that integrates two important functionalities: providing answers to questions and browsing through the document that supports the answer. The documents have to undergo several natural language processing steps before they are indexed. Moreover, the indexed terms are semantically enhanced by inference using WordNet[1] and the Cyc [1] ontology. Section 3 will explain the preprocessing steps in more detail. Another feature of the system is that the user can choose at query time in which document collection the answer should be searched. Also, the user can upload his own document collection. To our knowledge such flexibility is not provided by other similar systems.

Previous work has typically focused on a single topic (question answering, summarization, semantic representation and visualization of documents) and we see the advantage of the proposed system in combining these topics together. Many of the previous approaches, like Aqualog [2] and QuestIO [3], query structured data stored in ontologies. Aqualog has a restricted grammar and restricted vocabulary to which the query has to be compatible. QuestIO does not require a fixed grammatical structure of the question, but the words which it can handle are limited because of the dependency on an underlying ontology. Our system derives the answers only from unstructured text, which means that the range of questions is not limited or domain specific. However the questions must be in fixed grammatical forms for our system to "understand" them. TextRunner [4] is similar to our system in the way that it also consists of structured queries on unstructured text but the difference is that we also

[1] http://wordnet.princeton.edu/

J.L. Balcázar et al. (Eds.): ECML PKDD 2010, Part III, LNAI 6323, pp. 579–582, 2010.

provide a natural language interface to the search. Powerset[2] enables search and discovery in Wikipedia and Freebase, by entering keywords, phrases or simple questions. What distinguishes our system from Powerset is the way we describe the answer: by a visual representation of the document in the form of a semantic graph and by the document summary, which is automatically extracted based on the semantic graph of the document.

2 System Overview

AnswerArt [5] combines question answering, summarization and document visualization. Firstly, in a step performed offline, facts (consisting of subject - predicate - object triplets) are extracted from text and then stored in a triplet store. The user queries these triplets by asking a natural language question which is transformed into a structured query for the triplet store. The result consists of a list of matching triplets and the list of documents in which they occurred. In a detailed overview of the document the user can also see the list of all triplets from that document, the semantic graph (made out of triplets) and an automatically generated summary. Fig 1 shows a possible use case, while Fig 2 is a screenshot showing as an example the results we get for the question *What could pollution have affected?*

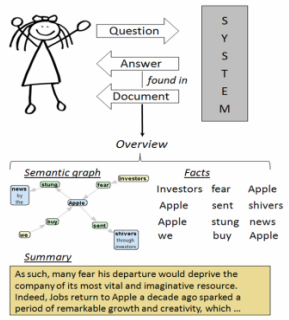

Fig. 1. System Overview

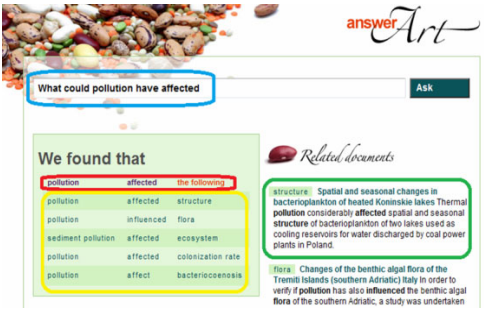

Fig. 2. Example Screenshot

3 Document Preprocessing

We shall now describe the preprocessing steps which are necessary to make the functionality described in Section 2 possible. The starting point is a document collection, containing unstructured text which needs to be preprocessed. The central part of preprocessing is triplet extraction [6]. Triplets, which are made of *subject*, *predicate* and *object*, are extracted from each sentence. It is in this form that the knowledge contained in the text is stored and made available for searching. Before triplets can be extracted, certain other preprocessing steps have to be done: Part of speech tagging, noun phrase chunking, parsing and named entity extraction. Thus

[2] `http://www.powerset.com/`

having obtained the triplets we construct a semantic graph by merging equivalent terms in the separate triplets. The semantic graph is used in document visualization [7] and summarization [8].To make the semantic graph more connected, co-reference resolution, anaphora resolution and normalization has to be done on the triplet terms.

To make the search efficient, the terms contained in the triplets are indexed. Before indexing, the terms are expanded with related terms which are found from WordNet and the Cyc ontology. The related terms are found from the synonyms and related concepts in the ontology. For example the term *water* would be expanded with: *lake*, *sea*, *ocean*, *stream*, *freshwater* etc. Indexing related terms results in a semantic enhancement of the stored knowledge and has the goal of increasing the recall of the system.

4 A Machine Learning Approach to Document Summarization

In order to automatically generate document summaries, we consider a machine learning approach, where we aim at learning which sentences belong to the summary. More exactly, we describe a set of features for each triplet extracted from the document sentences, and train an SVM model for binary classification of triplets as belonging or not to the summary. Further we identify the corresponding sentences which yielded the triplets, and, relying on them, construct the document summary. The features used are of three kinds: *document features* (for e.g. position of the sentence in the document, position of the triplet in the sentence, words in the triplet elements), *linguistic features* (for e.g. part of speech tags, location of the triplet in the parse tree) and *graph features* (for e.g. hub and authority weights, page rank, node degrees, connected components).

For training the linear SVM model and for evaluating the triplet ranking, we use the *DUC (Document Understanding Conferences)*[3] datasets from 2002 and 2007, respectively. The 2002 dataset comprises 300 newspaper articles on 30 different topics and for each article we have a 100 word human written abstract. The DUC 2007 dataset comprises 250 articles for the update task and 1125 articles for the main task.

We evaluated our summarization system, by comparing our results to the ones obtained by other systems participating in the DUC 2007 update task; we refer to [8] for more details regarding the evaluation outcome.

5 Evaluation

To evaluate the contribution of the triplet enhancement with ontologies to the performance of the question answering, we have conducted the following experiment. We have asked 27 questions to which the system responded both with and without using inference from ontologies. To each question the system gave a number of answers. The correctness or relevance of the given answers was determined according to the judgement of the authors. On average a question was answered with 8 answers out of which on average 3 were due to using ontologies. Hence the usage of ontologies

[3] `http://duc.nist.gov/`

increases the number of answers retrieved by about 60%. However the number of answers that are actually correct increases by only 40% when using ontologies. This shows that the precision of answers obtained using ontologies is lower and that trying to obtain more answers by inference has a negative effect on the precision. Indeed, the precision of the system drops from 84.17% to 76.61% when adding answers obtained from ontologies, because the answers using ontologies have a precision of only 63.29% Although the size of the experiment is too small to base any solid conclusions on it, we can argue that the AnswerArt system cannot find an important number of correct answers unless it uses ontologies. On the negative side however, ontologies introduce more mistakes and decrease the precision of the system.

6 Conclusions

We have presented a question answering system enhanced with summarization and document visualization functionalities. The information from which the answers are retrieved is stored as subject-predicate-object triplets. The indexed terms are expanded using inference from the Cyc ontology and WordNet. Evaluation shows that the use of ontologies increases recall but decreases precision.

References

1. Cyc, L.D.B.: A Large-Scale Investment in Knowledge Infrastructure. Comm. of the ACM 38(11) (November 1995)
2. Lopez, V., Uren, V., Motta, E., Pasin, M.: AquaLog: An ontology-driven question answering system for organizational semantic intranets. Journal of Web Semantics, 72–105 (2007)
3. Damljanovic, D., Tablan, V., Bontcheva, K.: A text-based query interface to owl ontologies. In: The 6th Language Resources and Evaluation Conference (LREC), Marrakech, Morocco, ELRA (May 2008)
4. Banko, M., Etzioni, O.: The Tradeoffs Between Open and Traditional Relation Extraction. In: Proceedings of the Association for Computational Linguistics, Columbus, Ohio (2008)
5. Dali, L., Rusu, D., Fortuna, B., Mladenic, D., Grobelnik, M.: Question Answering Based on Semantic Graphs. In: LTC 2009, Poznan, Poland (November 6, 2009)
6. Rusu, D., Dali, L., Fortuna, B., Grobelnik, M., Mladenic, D.: Triplet extraction from sentences. In: SiKDD 2007, Ljubljana, Slovenia (October 2009)
7. Rusu, D., Fortuna, B., Mladenic, D., Grobelnik, M., Sipos, R.: Document Visualization Based on Semantic Graphs. In: Proceedings of the 13th International Conference Information Visualisation (IV'09), Barcelona, Spain, pp. 292–297 (2009)
8. Rusu, D., Fortuna, B., Grobelnik, M., Mladenic, D.: Semantic Graphs Derived From Triplets with Application in Document Summarization. In: Proceedings of the 11th International Multiconference Information Society - IS 2008, Ljubljana, Slovenia, pp. 198–201 (2008)