Linguistic Knowledge and Question Answering

Gosse Bouma
Information Science
Groningen University
{g.bouma}@rug.nl

Abstract

The availability of robust and deep syntactic parsing can improve the performance of Question Answering systems. This is illustrated using examples from Joost, a Dutch QA system which has been used for both open (CLEF) and closed domain QA.

1 Linguistically Informed IR

Information retrieval is used in most QA systems to filter out relevant passages from large document collections to narrow down the search for answer extraction modules in a QA system. Given a full syntactic analysis of the text collection, it becomes feasible to exploit linguistic information as a knowledge source for IR. Using Apache's IR system Lucene, we can index the document collection along various linguistic dimensions, such as part of speech tags, named entity classes, and dependency relations. Tiedemann (2005) uses a genetic algorithm to optimize the use of such an extended IR index, and shows that it leads to significant improvements of IR performance.

2 Acquisition of Lexical Knowledge

Syntactic similarity measures can be used for automatic acquisition of lexical knowledge required for QA, as well as for answer extraction and ranking. For instance, in van der Plas and Bouma (2005) it is shown that automatically acquired class-labels for named entities improve the accuracy of answering general WH-questions (i.e. *Which ferry sank in the Baltic Sea?*) and questions which ask for the definition of a named entity (i.e. *Who is Nelson Mandela?* or *What is MTV?*).

3 Off-line answer extraction

Off-line extraction of answers to frequent question types can be based on dependency patterns and coreference resolution (Bouma et al., 2005; Mur and van der Plas, 2006), leading to higher recall (compared to systems using surface patterns). Closed-domain (medical) QA can benefit from the fact that dependency relations allow answers to be identified for questions which are not restricted to specific named entity classes, i.e. definitions, causes, symptoms, etc. Answering definition questions, for instance, is a task which has motivated approaches that go well beyond the techniques used for answering factoid questions. In Fahmi and Bouma (2006) it is shown that syntactic patterns can be used to extract potential definition sentences from Wikipedia, and that syntactic features of these sentences (in combination with obvious clues such as the position of the sentence in the document) can be used to improve the accuracy of an automatic classifier which distinguishes definitions from non-definitions in the extracted data set.

4 Joost

Joost is a QA system for Dutch which incorporates the features mentioned above, using the Alpino parser for Dutch to parse (offline) the document collections as well as (interactively) user questions. It has been used for the open-domain monolingual QA task of CLEF 2005, as well as for closed domain medical QA. For CLEF, the full Dutch text collection (4 years of newspaper text, approximately 80 million words) has been parsed. For the medical QA system, we have been using a mixture of texts from general and medical encyclopedia, medical reference works, and web pages

dedicated to medical topics. The medical data are from mixed sources and contain a fair amount of domain specific terminology. Although the Alpino system is robust enough to deal with such material, we believe that the accuracy of linguistic analysis on this task can be further improved by incorporating domain specific terminological resources. We are currently investigating methods for acquiring such knowledge automatically from the encyclopedia sources.

References

Gosse Bouma, Jori Mur, and Gertjan van Noord. 2005. Reasoning over dependency relations for QA. In *Proceedings of the IJCAI workshop on Knowledge and Reasoning for Answering Questions (KRAQ)*, pages 15–21, Edinburgh.

Ismail Fahmi and Gosse Bouma. 2006. Learning to identify definitions using syntactic features. In Roberto Basili and Alessandro Moschitti, editors, *Proceedings of the EACL workshop on Learning Structured Information in Natural Language Applications*, Trento, Italy.

Jori Mur and Lonneke van der Plas. 2006. Anaphora resolution for off-line answer extraction using instances. submitted.

Jörg Tiedemann. 2005. Integrating linguistic knowledge in passage retrieval for question answering. In *Proceedings of EMNLP 2005*, pages 939–946, Vancouver.

Lonneke van der Plas and Gosse Bouma. 2005. Automatic acquisition of lexico-semantic knowledge for question answering. In *Proceedings of Ontolex 2005 – Ontologies and Lexical Resources*, Jeju Island, South Korea.