Paragraph Retrieval for *Why*-Question Answering

Suzan Verberne
Department of Linguistics
University of Nijmegen
s.verberne@let.ru.nl

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software

General Terms

Design

Keywords

why-questions, paragraph retrieval, discourse annotation

1. INTRODUCTION

In the current research project, we aim at developing a system for answering *why*-questions (*why*-QA). In previous research, we investigated the possibilities of answer extraction for *why*-QA exploiting discourse structure in the source text. One conclusion was that many *why*-questions require a complete paragraph as an answer. In the present paper, we will first discuss the results and the main conclusions that we obtained from these experiments. Then, we will present our research plans concerning paragraph retrieval for *why*-QA.

2. ANSWER EXTRACTION

In [3], we present and discuss an approach to answer extraction for *why*-questions. As a model for discourse annotation, we use Rhetorical Structure Theory (RST), in the variant developed by Carlson et al. [1]. Our answer extraction method is based on the idea that the topic of a *why*-question and its answer are siblings in the RST structure of the document, connected by a relation that is relevant for *why*-questions.

We evaluated our discourse-based method for answer extraction on two sets of *why*-questions. The first data collection consists of 372 *why*-questions obtained from elicitation of native speakers to seven annotated texts from the RST Treebank [1]. The second data collection contains 400 questions that have been submitted to the online QA system `answers.com` with a set of answer fragments that we manually extracted from Wikipedia and annotated with RST structures.

A manual analysis of the questions and the corresponding RST structures shows that the maximum recall that can be achieved using our discourse-based answer extraction approach is around 60%. We find that some ('explanation-type') RST-relations have a large predictive power in answer selection but we conclude that our answer extraction approach should be combined with other methods in order to increase recall.

We consider paragraph retrieval as an alternative and supplementary approach. We studied all questions in our second data collection and we found that for 84.7% of questions, a complete paragraph from Wikipedia is a satisfactory answer.

SIGIR'07, July 23–27, 2007, Amsterdam, The Netherlands.
ACM 978-1-59593-597-7/07/0007.

3. PROPOSED RESEARCH

We aim at developing a method for paragraph retrieval in which we incorporate knowledge about the presence of relevant RST relations. We formulate the following research question: "How can we realize intelligent paragraph retrieval and paragraph ranking for *why*-QA, incorporating knowledge on discourse relations?"

We suggest the following approach for answering *why*-questions: (1) question analysis and query creation; (2) document retrieval; (3) paragraph retrieval and ranking.

One problem that we face in developing a paragraph retrieval method is that general language models are not well geared for retrieving short text units such as paragraphs. This means that we need to develop a language model that is suitable for possibly very short text fragments. Moreover, we aim to incorporate quantifiable information on the presence of RST relations in the language model. For automatically detecting RST relations, we plan to extend the approach as implemented in Soricut and Marcu's Spade tool [2] so that it is specifically geared for detecting answers to *why*-questions.

4. REFERENCES

[1] L. Carlson, D. Marcu, and M. E. Okurowski. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In J. van Kuppevelt and R. Smith, editors, *Current Directions in Discourse and Dialogue*, pages 85–112. Kluwer Academic Publishers, 2003.

[2] R. Soricut and D. Marcu. Sentence level discourse parsing using syntactic and lexical information. *Proceedings of NAACL-HLT 2003 Volume 1*, pages 149–156, 2003.

[3] S. Verberne, L. Boves, N. Oostdijk, and P. Coppen. Discourse-based answering of *why*-questions. 2007. Accepted for *Traitement Automatique des Langues*, special issue on Computational Approaches to Discourse and Document Processing.