Evaluating Discourse-based Answer Extraction for *Why*-Question Answering

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

Lou Boves
Dept. of Linguistics
University of Nijmegen
l.boves@let.ru.nl

Nelleke Oostdijk
Dept. of Linguistics
University of Nijmegen
n.oostdijk@let.ru.nl

Peter-Arno Coppen
Dept. of Linguistics
University of Nijmegen
p.a.coppen@let.ru.nl

Categories and Subject Descriptors

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

General Terms

Design

Keywords

why-questions, answer extraction, RST, discourse annotation

1. INTRODUCTION

Our research aims at developing a system for answering *why*-questions (*why*-QA). More specifically, we focus on the role that linguistic information and analysis can play in the process of *why*-QA. In the present paper, we evaluate an answer extraction method that exploits discourse relations in texts.

In approaches to factoid QA, named entity recognition can make a substantial contribution to identifying potential answers. Answers to *why*-questions on the other hand, cannot be expressed in the form of a noun phrase. Rather, they consist of propositions, and often they span multiple sentences that entertain discourse relations such as 'cause', 'motivation', 'purpose', and 'explanation'. Therefore, we decided to approach the answer extraction problem as a discourse analysis task. In order to investigate to what extent discourse structure enables *why*-QA, we created a system that uses discourse structure for answer extraction.

In the present paper, we evaluate a method for discourse-based answer extraction using two sets of *why*-questions: one obtained by elicitation of native speakers and one containing questions that are asked to the online QA system answers.com.

2. RST-BASED ANSWER EXTRACTION

As a model for discourse annotation, we use Rhetorical Structure Theory (RST) [1]. The answer extraction approach that we consider is proposed by Verberne et al. [4]. This 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 implemented an algorithm that (1) indexes all text spans from the source document that participate in a potentially relevant RST relation, (2) matches the input question to each of the text spans in the index, and (3) retrieves the sibling for each of the found spans as answer. The result is a list of potential answers, ranked using a probability model based on the general language model for Information Retrieval as defined by Croft and Lafferty [2]

3. DATA FOR WHY-QA

We created two collections of *why*-questions. For the first data collection, we manually selected seven documents from the RST Treebank [1] of 350–550 words each. The RST Treebank contains Wall Street Journal articles that have been manually annotated with RST structures by Carlson et al. We created a set of 372 *why*-questions obtained from elicitation of native speakers to these annotated texts.

Gathering questions through elicitation runs the risk that participants feel forced to come up with *why*-questions. This may lead to a set of questions that is not completely representative for a user's real information need. Therefore, we created a second data set, based on the Webclopedia question collection [3]. The complete Webclopedia collection consists of 17,000 questions downloaded from answers.com, an online domain-independent QA system. 805 questions from the Webclopedia set are *why*-questions—pragmatically defined as questions starting with the word *why*. The source of these questions guarantees that they originate from users' information needs. We randomly selected 400 of these *why*-questions. For analysis and development purposes, we created a set of answer fragments to these 400 questions, manually extracted from Wikipedia. For 54% of these questions, we were able to find the answer in Wikipedia. In a large majority of cases (94%) the length of the answer did not exceed a single paragraph. We let two experienced annotators create RST structures for the answer fragments from Wikipedia. For answer fragments shorter than one paragraph, we selected the complete paragraph for annotation. We also added the previous paragraph or the section heading to the fragment if these provided essential information for understanding the paragraph containing the answer. We

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

did not inform the annotators about the purpose of their annotations.

We believe that it is useful to categorize our questions according to their answer type, because this helps the system select potential answers from the source text. Our two question collections are very different in the types of answers that they expect. In the set of elicited questions, 38% has 'motivation' as answer type and 52% 'cause'. For the Webclopedia set, we found that the proportion of questions expecting a motivation as answer is only 10%. The remaining category, 'cause', appeared to be too general as a class for all other questions. Therefore, we decided to categorize the Webclopedia question collection into five classes: 'Motivation' (10%), 'Physical Explanation' (42%), 'Non-physical explanation' (30%), 'Etymology' (12%), and 'Nonsense'(6%).

4. EVALUATING ANSWER EXTRACTION

We use both our data collections for evaluating our approach to discourse-based answer extraction. We study the theoretically possible contribution of RST to answer extraction by manually analyzing each of the questions for which we have an answer fragment available—and its corresponding RST structure. We only considered the questions for which we were able to find an answer in Wikipedia (54% of all questions). We manually matched each question topic to a text span in the answer fragment and selected the span's sibling as answer. Following this procedure, we find a satisfactory answer for 58.0% of the question-answer pairs in our set of elicitation data, and for 60.6% of the question-answer pairs in our Webclopedia set. We see that although the questions in both data collections come from different sources, our answer selection procedure shows highly similar results for both sets.

This analysis shows that the maximum recall that can be achieved using our discourse-based answer extraction approach is around 60%. The remaining 40% suffers from one of the following problems: (1) the question topic is not represented by a text span in the answer fragment; (2) the text span representing the question topic does not participate in an RST relation; (3) the correct answer is not the sibling of the span representing the question topic but it is somewhere else in the RST structure.

If we consider the question-answer pairs where our RST-based approach succeeds, then we see that the most occurring successful RST-relations are 'explanation-argumentative', 'circumstance', 'background' and 'purpose'. The instances of each of these relation types lead to a satisfactory answer in more than 75% of question topics that participate in such a relation. Thus, these relations have the largest predictive power in answer selection using RST.

We implemented a module that automatically maps the question topic onto the correct discourse unit in the text, mainly by measuring lexical overlap between the question topic and discourse units. For the questions related to the RST Treebank documents 88.7% of the question topics could be identified automatically in the Wall Street Journal texts. However, the same procedure could only find 42.7% of the discourse units connected to the Webclopedia questions in the Wikipedia documents. This difference is due to the fact that questions elicited from subjects who are reading a text tend to use the same terms as those in the texts. This suggests that the results obtained using the Wall Street Journal texts do not generalize to any other setting. For the Webclopedia questions lexical overlap is much smaller because these questions were formulated completely independently from a specific text. Assuming that the Webclopedia/Wikipedia set is representative to an actual question answering setting, we should acknowledge the problem of small lexical overlap in the system under development.

5. CONCLUSIONS

We created two corpora of *why*-questions consisting of 372 and 400 questions respectively and corresponding answer documents annotated with discourse structure. These data collections may be of interest for other researchers in the field of question answering or discourse analysis.[1]

We evaluated an answer extraction method for *why*-questions based on the idea that question topic and answer are siblings in the RST structure. We found that our procedure is potentially successful for 60% of *why*-questions. The implementation of our procedure can match 42.7% of the question topics from the Webclopedia set to the manually chosen discourse unit in the corresponding Wikipedia fragment.

We conclude that discourse structure can be useful in solving at least a subset of *why*-questions and that some relation types have a predictive power in answer selection. However, our answer extraction approach should be combined with other methods in order to increase recall.

We consider paragraph retrieval as alternative and supplementary approach. We studied all Webclopedia questions and the corresponding Wikipedia fragments and we found that for 84.7% of questions, a complete paragraph from Wikipedia is a satisfactory answer. Thus, paragraph retrieval is a good additive solution to discourse-based answer extraction. Since some types of RST relations have a high predictive power in answer selection, we aim at developing a method for paragraph retrieval in which we incorporate knowledge about the presence of relevant RST relations.

Moreover, we plan to investigate to what extent we can achieve automatic partial discourse annotations that are specifically equipped to finding answers to *why*-questions.

6. 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] W. Croft and J. Lafferty. *Language Modeling for Information Retrieval*. Kluwer Academic Publishers, Norwell, MA, USA, 2003.

[3] E. Hovy, U. Hermjakob, and D. Ravichandran. A question/answer typology with surface text patterns. In *Proceedings of the Human Language Technology conference (HLT)*, San Diego, CA, 2002.

[4] 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.

[1] We will make the data collection available through the project's web site http://lands.let.ru.nl/~sverbern/