Evaluating Answer Extraction for *Why*-QA using RST-annotated Wikipedia texts

SUZAN VERBERNE
Department of Linguistics, Radboud University Nijmegen
s.verberne@let.ru.nl

ABSTRACT. In this paper the research focus is on the task of answer extraction for *why*-questions. As opposed to techniques for factoid QA, finding answers to *why*-questions involves exploiting text structure. Therefore, we approach the answer extraction problem as a discourse analysis task, using Rhetorical Structure Theory (RST) as framework. We evaluated this method using a set of *why*-questions that have been asked to the online question answering system `answers.com` with a corpus of answer fragments from Wikipedia, manually annotated with RST structures. The maximum recall that can be obtained by our answer extraction procedure is about 60%. We suggest paragraph retrieval as supplementary and alternative approach to RST-based answer extraction.

1 Introduction

In my PhD research project, I aim at developing a system for answering *why*-questions (*why*-QA). More specifically, I focus on the role that linguistic information and analysis can play in the process of *why*-QA.

In this paper the research focus is on the task of answer extraction for *why*-questions. In approaches to question answering (QA) involving factoid questions, named entity recognition can make a substantial contribution to identifying potential answers in a source document. For *why*-QA on the other hand, more sophisticated techniques are needed, because most answers consist of some kind of reasoning that cannot be expressed in a noun phrase. Arguments may be distributed over several sentences, making it necessary to exploit text structure. Therefore, we decided to approach the answer extraction problem as a discourse analysis task. We aim to find out to what extent discourse structure enables *why*-QA.

We created a system that uses discourse structure for answer extraction. In [13], we evaluated our approach using a set of elicited questions to a closed corpus (the RST Treebank [2]), with a moderate degree of success.

Proceedings of the Twelfth ESSLLI Student Session
Ville Nurmi and Dmitry Sustretov (editors)

We hypothesized that part of the unsolved problems were due to the effect of the elicitation process: subjects might have been tempted to 'invent' *why*-questions that do not address the type of argumentation that one would expect for natural *why*-questions. Therefore, in the current paper, we aim to find out what the performance of discourse-based answer extraction is for *why*-questions that originate from real users' information needs. To this end, we created a corpus consisting of *why*-questions asked to the online QA system answers.com, and a set of manually selected Wikipedia fragments which we annotated with discourse structure.

2 Answer extraction using discourse structure

As a model for discourse annotation, we use Rhetorical Structure Theory (RST), originally developed by Mann and Thompson [5] and adapted by Carlson et al. [2]. In RST, the smallest units of discourse are called *elementary discourse units* (EDUs). In terms of the RST model, a rhetorical relation typically holds between two EDUs, one of which (the *nucleus*) is more essential for conveying the propositional content than the other (the *satellite*). If two related EDUs are of equal importance, there is a *multinuclear relation* between them. Two or more related EDUs can be grouped together in a larger text *span*, which in its turn can participate in another relation. By grouping and relating spans of text, a hierarchical structure of the text is created. The main reason for using RST in the variant of Carlson et al. is that their rules and guidelines for segmenting discourse units and selecting relations are largely syntax-based, which fits the linguistic perspective of the current research. Moreover, Carlson et al. created a treebank of manually annotated Wall Street Journal texts with RST structures (the RST Treebank).

We presented our discourse-based approach to answer extraction in [13]. Our method is based on the idea that the topic of a *why*-question[1] and its answer are siblings in the RST structure of the document, connected by a relation that is relevant for *why*-questions.

We performed two experiments for testing our method: (1) a manual analysis procedure and (2) and implementation of our approach.

First, we studied the theoretical upper bound of the contribution of RST to answer extraction by manually analyzing each question in our data collection and its corresponding RST structure. We apply the following manual analysis steps to each of the question-answer pairs:

1. Identify the topic of the question; in the RST tree of the source document, identify the span(s) of text that express(es) the same proposition as the question topic;

[1] The topic of a *why*-question is the proposition that is questioned. A *why*-question has the form 'WHY P?', in which the proposition P is the topic. [10]

2. Does the topic span participate in an RST relation? If it does, select the span (nucleus or satellite) that participates, and take note of the relation type;

3. Select the topic span's sibling as a potential answer;

4. Decide whether this span is satisfactory as answer to the question.

The effects of this procedure can best be demonstrated by means of an example. Consider the question *Why is the funny bone so called?* The following text fragment contains the answer:

> "The ulnar nerve comes from the medial cord of the brachial plexus, and runs inferior on the medial/posterior aspect of the humerus down the arm, going behind the medial epicondyle at the elbow. Because of the mild pain and tingling throughout the forearm associated with sudden compression of the nerve at this point, it is sometimes called the funny bone. (It may also have to do with its location relative to the humerus, as well as the fact that 'humerus' is homophonic to the word 'humorous')."

In this text, we identify the following clause representing the question topic: *it is sometimes called the funny bone.* In figure 1 below the RST annotation of the paragraph is shown. Here we see that the span representing the question topic is EDU number 6, which is the nucleus of an explanation-argumentative relation.

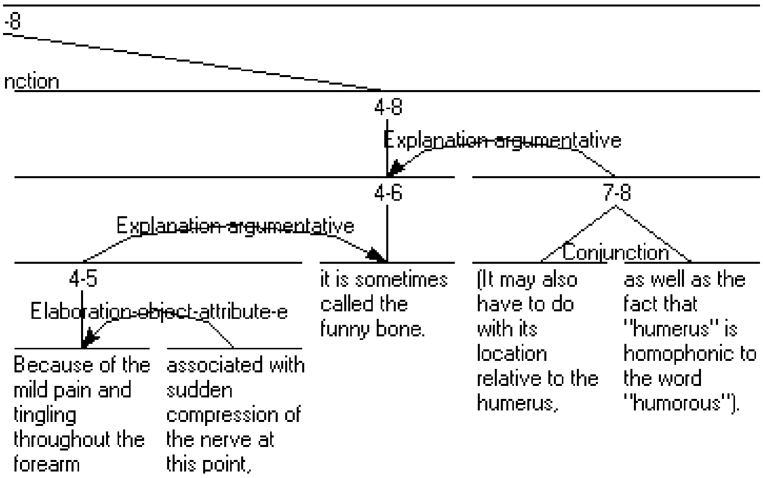
-8
nction
4-8
Explanation-argumentative
4-6
7-8
Explanation-argumentative
Conjunction
4-5
Elaboration-object-attribute-e
it is sometimes called the funny bone.
(It may also have to do with its location relative to the humerus,
as well as the fact that "humerus" is homophonic to the word "humorous").
Because of the mild pain and tingling throughout the forearm
associated with sudden compression of the nerve at this point,

Figure 1.1: Part of the RST structure for the answer paragraph on the funny bone.

The sibling of the topic span is span 4-5 in the hierarchy: *Because of the mild pain and tingling throughout the forearm associated with sudden* *compression of the nerve at this point*. We judge this span as a satisfactory answer to the question. However, we also note that the complete paragraph would have been a more complete (and therefore better) answer than this single clause, because the sentence contains one broken anaphoric reference (*this point*) and lacks background information on the *nerve* mentioned.

The second experiment is the implementation of an algorithm in Perl that reflects the manual analysis steps described above. We built an indexing script that takes as input file the RST structure of a document, and searches it for instances of potentially relevant *why*-relations. It then extracts for each relation both the participating spans and its relation type and saves the information to an index file (in plain text).

For the actual retrieval task, we wrote a second Perl script that takes as input one of the document indices, and a question related to the document. Then it performs the following steps:

1. Read the index file, normalize and lemmatize each span in the index;

2. Read the question, normalize and lemmatize it;

3. For each span in the index, calculate its likelihood using a probability model that takes into account its lexical overlap with the question and a prior on the relation type it participates in.

4. Save all spans with a likelihood greater than a predefined threshold and rank the spans according to their likelihood;

5. Retrieve as potential answers the siblings of each of these spans.

In [13], we created a test corpus consisting of seven texts from the RST Treebank and 372 *why*-questions elicited from native speakers to these documents. We performed both experiments (manual analysis and implementation) on this data collection. Following our manual analysis procedure (first experiment), we found a satisfactory answer for 58.0% of the questions. Thus, we argued that the maximum recall that can be achieved using our discourse-based answer extraction approach is 58.0%. The implementation of our algorithm (second experiment) reaches a recall of 53.3% with a mean reciprocal rank of 0.662.

We consider a recall of 53.3% (and a maximum recall of 58.0%) as mediocre at best. An in-depth analysis of the questions for which the answers could not be found suggested that a fair proportion of the questions were somewhat artificial, probably invented for the purpose of the experiment. Thus, the elicitation procedure may result in a set of questions that are not representative for users' actual information needs.

3 Real users' *why*-questions and answer fragments

In order to test the discourse-based procedure for answer extraction on a question set that is more representative for questions asked to a QA system, we created a data set from questions that have been posed to the online domain-independent QA system answers.com. Hovy et al. downloaded 17,000 questions from answers.com for their Webclopdia collection [4]. 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 our data collection.

We first study these 400 *why*-questions from the Webclopedia set independently from their answer documents. In [12], we created a typology of *why*-questions based on the classification of adverbial clauses by Quirk et al. [8]. We originally chose four classes for the semantic answer type of *why*-questions: 'motivation', 'cause', 'circumstance' and 'generic purpose'. Of these, cause (52%) and motivation (37%) were by far the most frequent answer types in our set of elicited *why*-questions pertaining to newspaper texts [12]. From other research reported on in the literature it appears that knowing the answer type helps a QA system in selecting potential answers. Some work that we did on answer type prediction is reported on in [11].

For the current set of Webclopedia *why*-questions, we find that the proportion of questions expecting a motivation as answer is much smaller than for the elicited questions (10%), and that 'circumstance' and 'generic purpose' are again negligible as question classes. The remaining category, 'cause', is too general as a class for all other questions. Therefore, we decide to split the current collection of Webclopedia *why*-questions into five classes:

- Motivation (10%), for example: *Why did NBC reject the first "Star Trek" episode, "The Cage" in 1965?*

- Physical Explanation (42%), for example: *Why can't people sneeze with their eyes open?*

- Non-physical explanation (30%), for example: *Why is the color purple associated with royalty?*

- Etymology (12%), for example: *Why are chicken wings called Buffalo wings?*

- Trivial/Nonsense (6%), for example: *Why is the word "abbreviation" so long?*

For analysis and development purposes, we created a set of answer fragments to the 400 Webclopedia *why*-questions. We manually extracted these fragments from Wikipedia using Google's domain search on en.wikipedia.org.

We chose Wikipedia as source for several reasons: it is relatively stable compared to the Internet as a whole, and its content is reliable and accurate [3]. For 54% of the questions, we can find the answer in Wikipedia. Of the other 46%, some questions had false question propositions (e.g. *Why is a computer cabinet always white?*) and other questions seem to be either too specific (e.g. *Why do cows lie down before it rains?*) or too trivial (e.g. *Why is weird spelled w-e-i-r-d and not w-i-e-r-d?*) for Wikipedia to contain the answer. In a large majority of cases (94%) the length of the answer does not exceed a single paragraph.

We let two experienced annotators create RST structures for the 216 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 did not inform the annotators about the purpose of their annotations.

For determining the consistency of our annotations, we measure inter-annotator agreement. We let both annotators annotate the first ten fragments from our data set, and we calculate κ scores for both segmentation and hierarchy (nuclearity). For the calculation of κ, we follow Marcu's [6] definition of κ_u for segmentation and κ_n for nuclearity. We get a moderate agreement for segmentation ($\kappa = 0.54$) and low agreement for nuclearity ($\kappa = 0.13$). Marcu et al. found 0.72 and 0.67 respectively for κ_u and κ_n for their RST Treebank, which is much higher. We assume that the difference in κ scores is due to the procedure used to obtain the annotations: Marcu et al. trained their annotators elaborately for the purpose of maximizing the consistency of the annotations. In our project we have to rely on annotators who received substantial training in applying RST, but they were, due to temporal and financial limitations, never put in a situation where they had to reach a common interpretation of a set of training texts. Despite the low agreement for the nuclearity annotations, we still decide to press forward and use the resulting annotations for extracting answers for *why*-questions.

We now have a set of *why*-questions and answers that differs from the first data collection in (a) the source of the questions (real user questions instead of elicited questions), (b) the source of the answer corpus (newly annotated encyclopedia fragments instead of pre-annotated newspaper texts), and (c) the collection procedure (answers extracted for existing questions instead of questions formulated for existing answer documents).

4 Results and discussion

We executed the two experiments described in section 2 on our Webclopedia/Wikipedia collection, following the same procedures as for the collection of elicited questions. We only considered the questions for which we were able to find an answer in Wikipedia (54% of all questions).

In the first experiment, involving manual analysis, we find that our answer extraction procedure leads to a satisfactory answer for 60.6% of our Webclopedia questions. The remaining 39.4% suffers from one of the following problems (in the order of the analysis steps):

1. The question topic is not represented by a text span in the answer fragment (18% of all questions);
2. The text span representing the question topic does not participate in an RST relation (2%);
3. The sibling of the span representing the question topic is not a satisfactory answer (21%).

In the last case, the correct answer is somewhere else in the tree or in the same discourse unit as the question topic. For example, the clause *Buffalo wings are named after the city of Buffalo, New York* contains both the question topic *chicken wings are called Buffalo wings* and its answer.

We find no significant differences in success rate for the fragments that were annotated by the different annotators. This suggests that the low inter-annotator agreement has no noticeable influence on the answer extraction task that we consider. This may be because the majority of the RST relations that are relevant for *why*-QA are so obvious that annotators are likely to treat these similarly, but this remains to be seen.

If we compare the success rate of the proposed answer extraction procedure for the current data collection to the success rate that we found for the elicited questions with the RST Treebank (as described in section 2), we see highly similar results: for the Webclopedia questions, 60.6% of answers can be found through manual topic matching and sibling selection. For the elicited questions, this figure was 58.0%. Thus, although the questions in both data collections came from different sources, our answer selection procedure showed highly similar results for both sets.

We also compare the set of relation types addressed by the Webclopedia questions to the set of relation types addressed by the elicited questions. Table 1 gives an overview of the relation types that leads to the correct answer for at least 6% of the questions where our discourse-based answer extraction approach succeeds in either the Webclopedia set or the elicitation data. In the second and third column are the figures for the RST Treebank and the corresponding elicited questions. In columns four and five are the percentages for the Wikipedia corpus and the Webclopedia questions. We see for example that 18.0% of relations in the RST Treebank are elaboration relations, and for 27.0% of *why*-questions where our approach succeeds, it is an elaboration relation that connects question topic and answer. For the Wikipedia corpus, these numbers are 22.4% and 20.8% respectively.

Table 1.1: Distribution of relation types in corpora and question sets

	RST Treebank 37479 relations 372 *why*-questions		Wikipedia corpus 2333 relations 400 *why*-questions	
Relation type	% of relations	% of questions	% of relations	% of questions
Elaboration	18.0%	27.0%	22.4%	20.8%
Explanation	1.4%	7.1%	3.5%	20.0%
Circumstance	1.7%	0.5%	8.1%	16.0%
Background	0.5%	0.0%	4.3%	8.8%
Purpose	1.3%	14.3%	2.6%	7.2%
Consequence	1.0%	15.3%	0.8%	2.4%
Reason	0.6%	9.7%	0.9%	4.0%
Result	0.7%	9.7%	1.2%	2.4%

Although the success rate of our discourse-based answer extraction approach is similar for the Webclopedia and elicitation data collections (around 60%), we see some interesting differences between the two data collections in table 1. First, some relations differ considerably in their relative frequencies in both corpora (columns 2 and 4): explanation-argumentative (1.4% versus 3.5%), circumstance (1.7% versus 8.1%) and background (0.5% versus 4.3%). These differences are partly due to differences in annotation styles, and partly the result of differences in text types: the RST Treebank contains newspaper texts whereas the Wikipedia corpus consists of encyclopedic items where one would expect a higher density of explanations.

Secondly, we see large differences between the relative frequencies of the relations in the set of questions (columns 3 and 5). Again, the main differences lie in the relation types explanation-argumentative, circumstance and background, but also purpose, consequence, reason and result show large differences. The last four are the most interesting since the relative frequencies of these relations are more similar for the two source corpora than for their question sets. This means that the differences for these relation types come from the question source: questions asked to a QA system are apparently more likely to expect explanations, backgrounds and circumstances as answer than elicited questions. Elicited questions on the other hand refer to purposes, consequences and reasons more often. This matches to the differences in semantic answer types that we saw in section 3.1 if we take into account that purpose and reason, as defined by Carlson et al. [1], correspond to our definition of the answer type motivation [13].

The RST relations most frequently addressed in our Webclopedia question set are elaboration, explanation-argumentative, circumstance, background and purpose. Here, we see that the very general relation type elaboration is the most frequently occurring relation type for *why*-questions. However, there is a relatively small proportion of the question topics that participate in an elaboration relation for which this relation leads to a satisfactory answer: 49%. In other words: the predictive power of elaboration relations for *why*-questions is small. The predictive power for the question topics participating in an explanation-argumentative relation is much larger: for 89% of the question topics that participate in an explanation-argumentative relation, this relation leads to a satisfactory answer. For the question topics participating in a circumstance, background and purpose relation, these relations lead to a satisfactory answer in 77%, 85% and 100% of participating question topics respectively. Thus, we can conclude that the relation types explanation-argumentative, circumstance, background and purpose are valuable for finding answers to *why*-questions, whereas elaboration relations have low relevance. Furthermore, the predictive power of some types of RST relations confirms the expected importance of answer type determination. If we can predict the answer type from the question, and we know which RST relations represent this answer type, then we can apply the knowledge on the expected answer type for answer selection and ranking.

Our manual analyses described above lead to the conclusion that the maximum recall that can be achieved using our discourse-based answer extraction approach is around 60%. The success rate that we found is similar for both data collections, but the relation types addressed are different for the two corpora.

We then performed the second experiment, implementation of our algorithm, to the Webclopedia/Wikipedia data collection. Here, we found large differences between the two data collections: our implementation obtains a recall of only 25.9% on the Webclopedia/Wikipedia data set, whereas it had scored 53.3% for the elicited questions. This difference comes from the third step of our algorithm: matching the question topic to spans in the source text using lexical overlap measures. 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 between question and source document in the system under development.

5 Conclusions and further research

We created a corpus of *why*-questions consisting of 400 questions from the Webclopedia question set and corresponding answer fragments from

Wikipedia, manually annotated with RST relations. This data collection may be of interest for other researchers in the field of question answering or discourse analysis.[2]

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 current implementation of our procedure can retrieve 25.9% of the manually selected answers to the Webclopedia questions from the corresponding Wikipedia document.

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 found that for 44% of the cases where the procedure succeeds, the complete answer paragraph would (in our judgement) be a better answer to the question than the answer span in the RST tree only. Moreover, for 30% of the questions for which the procedure does not succeed (because the question topic is not in the text or question topic and answer are no siblings), the complete paragraph gives the answer. Thus, paragraph retrieval is a good additive solution to discourse-based answer extraction. Since some types of RST relations appears to 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.

We also plan to perform user studies in order to determine what answer form users prefer for different types of *why*-questions and answers. This way, we aim to find out whether paragraph retrieval with information from (partial) RST annotations can be a good alternative to the strict procedure of topic matching and sibling selection.

We should also note that in a future application of *why*-QA using RST, the system will not have access to a manually annotated corpus—it has to deal with automatically annotated data. We assume that automatic RST annotations will be less complete and less precise than the manual annotations are. Some work has been done on automatically annotating text with discourse structure. Promising in this direction is the done work by Marcu and Echihabi [7] and Soricut and Marcu [9]. We plan to investigate to what extent we can achieve automatic partial discourse annotations that are specifically equipped to finding answers to *why*-questions.

[2] We have made both our data collections available through the project's web site `http://lands.let.ru.nl/~sverbern/`

Bibliography

[1] L. Carlson and D. Marcu. *Discourse Tagging Reference Manual.* Univ. of Southern California/Information Sciences Institute, 2001.

[2] 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.

[3] J. Giles. Special report: Internet encyclopedias go head to head. *Nature*, 438(15):900–901, 2005.

[4] 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.

[5] W. Mann and S. Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281, 1988.

[6] D. Marcu, E. Amorrortu, and M. Romera. Experiments in constructing a corpus of discourse trees. *Proceedings of the ACL99 Workshop on Standards and Tools for Discourse Tagging*, pages 48–57, 1999.

[7] D. Marcu and A. Echihabi. An unsupervised approach to recognizing discourse relations. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 368–375, Morristown, NJ, USA, 2001. Association for Computational Linguistics.

[8] R. Quirk, S. Greenbaum, G. Leech, and J. Svartvik. *A comprehensive grammar of the English language.* London, Longman, 1985.

[9] R. Soricut and D. Marcu. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology—Volume 1*, pages 149–156, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

[10] B. Van Fraassen. The pragmatic theory of explanation. In J. Pitt, editor, *Theories of Explanation*, pages 135–155. Oxford University Press, 1988.

[11] S. Verberne. Paragraph retrieval for *why*-question answering. 2007. Accepted for the Doctoral Consortium at SIGIR, to be held in Amsterdam, July 2007.

[12] S. Verberne, L. Boves, N. Oostdijk, and P. Coppen. Data for question answering: the case of *why.* In *Proceedings of the 5th edition of the International Conference on Language Resources and Evaluation (LREC 2006), Genoa, Italy*, 2006.

[13] 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.