Exploring the use of linguistic analysis for answering *why*-questions

Suzan Verberne, Lou Boves, Nelleke Oostdijk, Peter-Arno Coppen

Dept. of Linguistics, Radboud University Nijmegen

Abstract

In the current project, we aim at developing an approach for automatically answering *why*-questions (*why*-QA). In the present paper, we investigate the relevance of linguistic analysis for *why*-QA. We focus on two tasks: the use of syntactic information for answer type determination and the use of discourse structure for the extraction of possible answers from retrieved documents.

For answer type determination, syntactic analysis appears to be of significance: we obtain 77.5% performance using a method based on syntactic parses by the TOSCA parser—compared to 58.1% using a comparable approach without syntactic analysis.

Discourse analysis appears to be very relevant for extraction of potential answers to *why*-questions. We performed a manual analysis of 336 question-answer pairs and the corresponding RST annotated texts. We found that for 58.9% of *why*-questions, the RST analysis of the source text can lead to a correct answer to the question.

1 Introduction

Up to now, *why*-questions have largely been ignored by researchers in the field of question answering (QA). One reason for this is that the frequency of *why*-questions in a QA context is lower than that of other types of question such as *who*- and *what*-questions (Hovy, Hermjakob and Ravichandran 2002). However, *why*-questions are not negligible: in a QA context, they comprise about 5 percent of all *wh*-questions (Hovy, Gerber, Hermjakob, Lin and Ravichandran 2001, Jijkoun and De Rijke 2005) and they do have relevance in QA applications (Maybury 2003). A second reason for disregarding *why*-questions until now is that the techniques that have proven to be successful in QA for closed-class questions are not suitable for questions that expect a procedural answer instead of a noun phrase (Kupiec 1999).

In the context of the current research into *why*-questions, a *why*-question is defined as an interrogative sentence in which the interrogative adverb *why* (or a synonymous word or phrase) occurs in (near) initial position. Furthermore, we only consider the subset of *why*-questions that could be posed to a QA system (as opposed to questions in a dialogue or in a list of frequently asked questions) and for which the answer is known to be present in some related document set.

The current paper aims to investigate the relevance of linguistic analysis for *why*-QA. Various types of linguistic analysis can be explored for analysis of both question and source text. For question analysis, we have researched the relevance of syntactic analysis for the determination of the answer type. For text analysis, we have been investigating the merits of Rhetorical Structure Theory (RST) (Mann and Thompson 1988) for extracting potential answers to *why*-questions.

In sections 2 and 3 we describe the work that we accomplished on question

analysis and discourse analysis for the purpose of *why*-QA. In section 2 a syntax-based method for answer type determination is presented and evaluated. Section 3 describes the results of our study into the use of RST for the purpose of answer selection. Section 4 concludes this paper with a discussion of the plans and goals for the work that will be carried out in the remainder of the project.

2 Question analysis for *why*-QA

The goal of question analysis is to create a representation of the user's information need. The result of question analysis is an answer template that contains all information about the answer that can be induced from the question. So far, no question analysis procedures have been created for *why*-QA specifically. Therefore, we have developed an approach for the analysis of *why*-questions. In this section, we first introduce the set of *why*-questions and answers that we developed for the current research into *why*-QA. We will then present the method that we used for the analysis of *why*-questions and finally indicate the quality of our method.

2.1 Data for *why*-QA

In research in the field of QA, data sources of questions and answers play an important role. Appropriate data collections are necessary for the development and evaluation of QA systems (Voorhees and Tice 2000). In the context of the QA track of TREC, data collections in support of factoid questions have been created. However, so far, no resources have been created for *why*-QA specifically.[1] For the purpose of the present research therefore, we have developed a data collection comprising a set of questions and corresponding answers and source documents. In order to meet the requirements as formulated in Verberne et al. (2006a), it would be best to collect questions posed in an operational QA environment. Since we do not have access to such an environment, we decided to revert to the procedure used in earlier TRECs, and imitate a QA environment in an elicitation experiment.

In the elicitation experiment, ten native speakers of English were asked to read texts from Reuters Textline Global News (1989) and The Guardian on CD-ROM (1992). For each text, the subjects were asked to formulate *why*-questions for which the answer can be found in the text and to formulate an answer to each of these questions. They were also asked to answer the questions of one of the other participants. The collected question-answer pairs were saved in text format, grouped per participant and per source document, so that the source information is available for each question. In this experiment, 395 questions and 769 corresponding answers were collected. For further details on the data collection, we refer to Verberne et al. (2006a).

[1] There are a few sources of *why*-questions available, but these appear to be unsuitable for the aims of the present research. See for an overview of these resources Verberne, Boves, Oostdijk and Coppen (2006a).

2.2 Syntax-based analysis of *why*-questions

As described in the introduction of section 2, no approaches for the analysis of *why*-questions have been developed until now. We decided to create a syntax-based method for the analysis of *why*-questions. We will examine the relevance of syntactic analysis for question analysis by comparing our syntax-based method to an approach without the use of syntactic parsing.

In systems for factoid-QA, the answer type is generally deduced directly from the question word (*who*, *when*, *where*, etc.): *who* leads to the answer type *person*; *where* leads to the answer type *place*, etc. This information helps the system in the search for candidate answers to the question. Hovy et al. (2001) find that, of the question analysis components used by their system, the determination of the semantic answer type makes by far the largest contribution to the performance of the entire QA system.

Since determination of the semantic answer type is the most important task of existing question analysis methods, we created a question analysis method that aims to predict the answer type of *why*-questions.

In the work of Moldovan, Harabagiu, Pasa, Mihalcea, Grju, Goodrum and Rus (2000), all *why*-questions share the single answer type *reason*. However, we believe that it is useful to split this answer type into sub-types, because a more specific answer type helps the system select potential answers from the source text. The idea behind this is that every sub-type has its own lexical and syntactic cues in a source text.

Based on the classification of adverbial clauses by Quirk, Greenbaum, Leech and Svartvik (1985), we distinguish the following sub-types of *reason*:

1. *Cause* (52% of the question-answer pairs in our data collection) which is a causal relation between two events in which no deliberate human intention is involved. For example: *Why did compilers of the OED have an easier time? – Because the OED was compiled in the 19th century when language was not developing as fast as it is today.*

2. *Motivation* (37%) which adds a human intention to a causal relation. A motivation can be either a future goal or a person's internal motivation. For example: *Why has the team of researchers been split up into two teams? – To complete the work more quickly - one team will finish "A" while the second team will start on "B".*

3. *Circumstance* (2%) which adds conditionality to the temporal relation: the first event is a strict condition for the second event. For example: *Why will people buy Windows? – Because it offers more software, it is more fun to use and it works well enough.*

4. *Generic purpose* (0%) which does not express a temporal relation between two events, but gives the physical function of an object in the real world. For example: *Why do people have eyebrows? – People have eyebrows to prevent sweat running into their eyes.*

The percentages of occurrence given above are based on a manual classification of all question-answer pairs in our data collection. To the remaining 9% of question-answer pairs, we were not able to assign one of the defined answer types. A more detailed description of the answer types, the quality of the classification and their distribution in our data collection is given in Verberne et al. (2006a).

We aim at creating a question analysis module that is able to predict the expected answer type of an input question. In the analysis of factoid questions, the question word often gives necessary information about the expected answer type. In case of *why*, the question word does not give information about the answer type since all *why*-questions have *why* as the question word. This means that other information from the question is needed for determining the answer sub-type.

We decided to use Ferret's approach, in which syntactic categorization helps in determining the expected answer type. In our question analysis module, the TOSCA (TOols for Syntactic Corpus Analysis) system (Oostdijk 1996) is explored for syntactic analysis. The TOSCA (syntactic) parser takes a sequence of unambiguously POS-tagged words and assigns function and category information to all constituents in the sentence. The parser yields one or more possible output trees for (almost) all input questions. For the purpose of evaluating the maximum contribution to a classification method that can be obtained from principled syntactic analysis, we manually selected the most plausible parse tree from the parser's output. This way, we created parse trees for the 122 *why*-questions that are linked to the first three source texts in our data collection. We decided not to create parse trees for all 395 questions because creating and correcting the parse trees is quite labour-intensive. [2]

For answer type determination, we decided to use a machine learning approach exploring automatic feature selection algorithms in Weka (Holmes, Donkin and Witten 1994). These algorithms take as input a set of feature values. In our experiment, we needed a set of question features that together can predict the answer type.

As baseline we established the majority classification: a method that classifies each question in the largest class *cause*). This baseline would lead to a correct classification of 52% of the questions.

As described above, we have manually annotated our questions with their answer type. We chose five syntactic and semantic features that, based on this manual classification, seem to be relevant to the distinction between the answer types: subject agency (agentive, non-agentive), verb type (anticausative, other), modality (*can*, *have to*, *should*, etc., none), the presence and type of declarative verb (factive, semi-factive, non-factive, none), and negation (absent, present). We added four features to the feature set that give supplementary information on the syntactic structure of the question: voice (passive, active), intensive complementation construction (absent, present), monotransitive *have* construction (absent, present), and existential *there* construction (absent, present). Thus, our feature set consists of nine syntactic and semantic question features for the purpose of answer type

[2] We are currently still working on evaluation of the parser. Part of this evaluation is measuring the difference in performance between automatic and manual parse selection.

determination.

We determined the feature values for each question by use of a Perl script that searches for patterns in the TOSCA tree. For example, the attribute *intens_compl* for the main verb of the matrix clause leads to the value 'present' for the feature *intensive complementation*. For determination of some of the features, lexical-semantic information is needed. Our script extracts this information from WordNet (Fellbaum 1998) (information on subject agency), VerbNet (Kipper, Trang Dang and Palmer 2000) (information on declaratives) and the Levin Verb Index (Levin 1993) (set of anticausatives).

The output of the script is a list of feature values for each of the 122 questions. We added the manually determined answer type for each question to complete our feature set for training. Then we used automatic feature selection algorithms to classify our questions according to their answer type.

We evaluated the classification into answer types using 10-fold cross-validation on the training set, comparing the automatically chosen answer types to the manually assigned answer types. The best-scoring algorithm (Lazy IBk) predicts 77.5% of the answer types correctly. This means our approach, classification is improved by almost 50% compared to the baseline.

To check the reliability of the feature classification, we compared the outcome of the ten individual runs of the cross-validation evaluation. We found that their standard deviation to the mean is 9%. This suggests that our results are fairly reliable, despite the small data set used.

In order to investigate the merit of syntactic parsing for answer type determination, we compared the result of our syntax-based method to an approach without the use of a syntactic parser. We created a new training set consisting of feature values for the same 122 questions as in the syntax-based method. We annotated these questions with part-of-speech tags assigned by the Brill tagger. Then we used a Perl script to extract the subject, the first auxiliary and the main verb from each question—this is feasible because of the relatively uniform syntactic structure of *why*-questions. The subject, auxiliary, main verb and the question string itself serve as input for a second Perl script for determining values for the previously defined features. Again, our script uses information from WordNet, VerbNet and the Levin Verb Index to determine subject agency, verb type and declarative type. Using this input, some of the features can relatively easily be determined: subject agency, verb type, the presence and type of declaratives, and the presence of existential *there*. For intensive complementation, modality and negation, the script can make an educated guess. On the other hand, the features *voice* (passive, active) and the presence or absence of constructions with monotransitive *have* are very difficult to determine without deep parse. Still, we are confident that our script for determining the feature values has been optimized for this set of features.

We again ran the automatic feature selection algorithm to classify our questions according to their answer type, using the non-syntax-based feature set. Now the best-scoring algorithm (Naive Bayes) classifies 58.1% of questions correctly. This is an improvement of only 12% compared to the baseline. The differences between the scores for question classification with and without syntactic parsing are due to

the fact that the set of feature values created with the Brill tagger contains more erroneous values. As a result, this feature set is less consistent than the set created with the TOSCA output. Due to this smaller consistency, it is more difficult for the classifier to induce rules that describe the data set.

These results show that adding syntactic parsing to an approach for determining answer types can improve its performance considerably. We therefore believe that syntactic analysis can play an important role for the analysis of *why*-questions.

3 Using RST for the purpose of *why*-QA

In section 2, we discussed the importance of answer type determination: knowing the answer type helps the QA system in selecting potential answers. After analysis of the input question, the QA system will retrieve a small set of documents that possibly contain the answer. Analysis of the retrieved documents is then needed for extracting potential answers. Thus, a system for *why*-QA needs a text analysis module that yields a set of potential answers to a given *why*-question. Although we now have a proper answer type determination approach, the problem of answer extraction is still very difficult. As opposed to factoid-QA, where named entity recognition can play an important role in extraction of potential answers, finding potential answers to *why*-questions is still an unsolved problem.

This means that we need to investigate how we can recognize the parts of a text that are potential answers to *why*-questions.

We decided to approach this answer extraction problem as a discourse analysis task. In this section, we aim to find out to what extent discourse analysis can help in selecting answers to *why*-questions. We also investigated the possibilities of a method based on textual cues, and used that approach as baseline for evaluating our discourse-based method.

We will first introduce RST as a model for discourse analysis. Then we shall present our method for investigating the use of RST for *why*-QA, followed by the results that we found. We will conclude this section with a discussion of the results, including a comparison to the baseline results, and the implications for future work.

3.1 Rhetorical Structure Theory (RST)

As framework for our research into discourse structure, we use the Rhetorical Structure Theory (RST), developed by Mann and Thompson (1988). In terms of the RST model, a rhetorical relation typically holds between two spans of text, of which one span (the *nucleus*) is more essential for the writer's intention than the other (the *satellite*). If two related spans are of equal importance, there is a *multi-nuclear relation* between them. Two related spans are grouped together in a larger span, which in turn can participate in a relation. The smallest units of discourse are called *elementary discourse units* (EDUs). By grouping and relating spans of text, a hierarchical structure of the document is created.

The main reasons for using RST as a model for discourse structure in the

present research are the following. First, good levels of agreement have been measured between human annotators of RST, which indicates that RST is well defined (Bosma 2005). Second, a treebank of manually annotated English texts with RST structures is available for training and testing purposes. This RST Discourse Treebank, created by Carlson, Marcu and Okurowski (2003) contains a selection of 385 Wall Street Journal articles from the Penn Treebank that have been annotated with discourse structure in the framework of RST. Carlson et al. (2003) created their own set of discourse relations for use in the treebank. The annotations by Carlson et al. (2003) are largely syntax-based. They chose clauses as EDUs, using lexical and syntactic clues to help determine the clause boundaries.

3.2 Method

Let us consider a *why*-question-answer pair and the RST structure of the corresponding source text. We hypothesize the following:

1. The question topic corresponds to a span of text in the source document and the answer corresponds to another span of text;

2. In the RST structure of the source text, an RST relation holds between the text spans representing question topic and the text span representing the answer.

If both hypotheses are true, then RST can play an important role in answering *why*-questions.

For the purpose of testing our hypotheses, we need a number of RST annotated texts and a set of question-answer pairs that are linked to these texts. Therefore, we set up an elicitation experiment using the RST treebank as data set. We followed the same elicitation method as we used for collecting data for question analysis. We selected seven texts from the RST treebank of 350–550 words each. Then we asked native speakers to read one of these texts and to formulate *why*-questions for which the answer could be found in the text. The subjects were also asked to formulate answers to each of their questions. In this experiment, they were not asked to answer one of the other participants' questions. This resulted in a set of 372 *why*-question-answer pairs, connected to seven texts from the RST treebank. On average, 53 question-answer pairs were formulated per source text. There is much overlap in the topics of the questions, as we will see later.

A risk of gathering questions following this method, is that the participants may feel forced to come up with a number of *why*-questions. This may lead to a set of questions that is not completely representative for a user's real information need. However, we believe that our elicitation method is the only way in which we can collect questions connected to a specific (closed) set of documents.

We performed a manual analysis on 336 of the collected question-answer pairs in order to check our hypotheses – we left out the other pairs for future testing purposes. We chose an approach in which we analyzed our data according to a clear step-by-step procedure, which we expect to be suitable for answer extraction

performed by a future QA system. This means that our manual analysis will give us an indication of the upper bound of the performance that can be achieved using RST.

First, we selected a number of relation types from Carlson et al.'s relation set, of which we believed that they might be relevant for *why*-QA. We started with the four answer types mentioned in section 2.2, but it soon appeared that the level of detail in the relation set made it necessary to also include relations similar to cause, purpose, motivation and circumstance. Therefore, we extended the list during the manual analysis. The final set of selected relations is shown in Table 1.

Table 1: Selected relation types

Cause	Circumstance	Condition	Elaboration
Explanation-argumentative	Interpretation	List	Problem-Solution
Purpose	Reason	Result	Sequence

Then, we used the following procedure for analyzing the questions and answers:

I. Identify the topic of the question. 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.

II. In the RST tree of the source document, identify the span(s) of text that express(es) the same proposition as the question topic.

III. Is the found span the nucleus of a relation of one of the types listed in Table 1? If it is, go to IV. If it is not, go to V.

IV. Select the satellite of the found nucleus as answer.

V. Discard the current text span.

The effects of the procedure can best be demonstrated by means of an example. Consider the following question, formulated by one of the native speakers after he had read a text about the launch of a new TV channel: *Why does Christopher Whittle think that Channel One will have no difficulties in reaching its target?* The topic of this question is *Christopher Wittle thinks that Channel One will have no difficulties in reaching its target*. According to our first hypothesis, the proposition expressed by the question topic matches a span in the RST structure of the source document. We manually selected the following text fragment which expresses the proposition of the question topic: *What we've done in eight weeks shows we won't have enormous difficulties getting to the place we want to be, said Mr. Whittle.* This sentence covers span 18–22 in the corresponding RST tree, which is shown in Figure 1 below.
In this way, we tried to identify a span of text corresponding to the question topic for each of the 336 questions. In section 3.3 we will present the results of this topic span selection step.

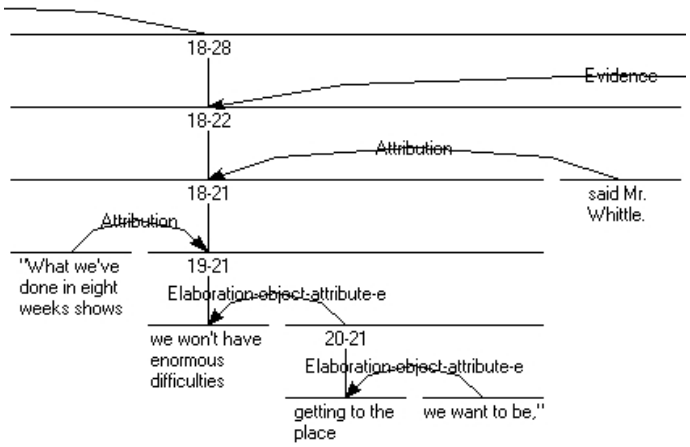

Figure 1: RST sub-tree for the text span "What we've done in eight weeks shows we won't have enormous difficulties getting to the place we want to be, said Mr. Whittle."

In cases where we succeeded in selecting a span of text in the RST tree corresponding to the question topic, we searched for potential answers following step III and IV from the analysis procedure. As we can see in Figure 1, the span *What we've done in eight weeks shows we won't have enormous difficulties getting to the place we want to be, said Mr. Whittle* is the nucleus of an evidence relation. Since we assumed that an evidence relation may lead to a potential answer, we can select the satellite of this relation, span 23–28, as an answer: *He said his sales force is signing up schools at the rate of 25 a day. In California and New York, state officials have opposed Channel One. Mr. Whittle said private and parochial schools in both states will be canvassed to see if they are interested in getting the programs.*
We analyzed all 336 *why*-questions following this procedure.

3.3 Results of the analyses

As described in section 3.1, our manual analysis procedure consists of four steps: (I) identification of the question topic, (II) matching the question topic to a span of text, (III) checking whether this span is the nucleus of an RST relation, and (IV) selecting its satellite as answer. Below, we will discuss the outcome of each of these sub-tasks.

The first step succeeds for all questions, since each *why*-question has a topic. For the second step, we were able to identify a text span in the source document that represents the question topic for 279 of the 336 questions that we analyzed (83.0%). We found that not every question corresponds to a unique text span in

the source document. For 279 questions, we identified 84 different text spans. This means that on average, each text span that represents at least one question topic is referred to by 3.3 questions. For the other 57 questions, we were not able to identify a text span in the source document that represents the topic.

For 209 of the 279 questions that have a topic in the text (62.2% of all questions), the question topic is (part of) the nucleus of a relation of one of the types in Table 1 (step III).

Evaluation of the fourth step, answer selection, needs some more explanation. For each question, we selected as answer the satellite that is connected to the nucleus corresponding to the question topic. For the purpose of evaluating the answers found using this procedure, we compared them to the user-formulated answers. If the found answer matches at least one of the answers formulated by native speakers in meaning (not necessarily in form), then we judged the found answer as correct. For example, for the question *Why did researchers analyze the changes in concentration of two forms of oxygen?*, two native speakers gave as an answer *To compare temperatures over the last 10,000 years*, which is exactly the answer that we found following our procedure. Therefore, we judged our answer as correct, even though eight subjects gave a different answer to this question. Evaluating the answer that we found to the question *Why does Christopher Whittle think that Channel One will have no difficulties in reaching its target?* (see above) is slightly more difficult, since it is longer than any of the answers formulated by the native speakers. However, since some of the user-formulated answers are part of the found answer span, and because the answer is still relatively short, we judged the found answer as correct.

We found that for 198 questions, the satellite connected to the nucleus corresponding to the topic is a correct answer. This is 58.9% of all questions.

3.4 Discussion and implications

3.4.1 Error analysis

We reported in section 3.3 that for 198 *why*-questions (58.9% of all questions), the answer could be found after matching the question topic to the nucleus of an RST relation and selecting its satellite as answer. This means that for 138 questions (41.1%), our method did not succeed. We distinguish four categories of questions for which we could not extract a correct answer using this method (percentages are given as part of the total of 336 questions):

1. Questions whose topics are not or only implicitly supported by the source text (57 questions, 17.0%). For example, the question *Why is cyclosporine dangerous?* refers to a source text that reads *They are also encouraged by the relatively mild side effects of FK-506, compared with cyclosporine, which can cause renal failure, morbidity, nausea and other problems.* We can deduce from this text fragment that cyclosporine is dangerous, but we need knowledge of the world (*renal failure, morbidity, nausea and other problems are dangerous*) to do this.

2. Questions for which the correct (i.e. user-formulated) answer is not or only implicitly supported by the text (15 questions, 4.5%). For example, the topic of the question *Why was Gerry Hogan interviewed?* corresponds to the text span *In an interview, Mr. Hogan said*. The native speaker who formulated this question gave as answer *Because he is closer to the activity of the relevant unit than the Chair, Ted Turner, since he has the operational role as President.* The source text does read that Mr. Hogan is president and that Ted Turner is chair, but the assumption that Gerry Hogan is closer to the activity than Ted Turner had been made by the reader; not by the text.

3. Questions for which both topic and answer are supported by the source text but the RST structure does not lead to the reference answer (55 questions, 16.4%). In some cases, this is because the topic and the answer refer to the same span. For example, the question *Why were firefighters hindered?* refers to the span *Broken water lines and gas leaks hindered firefighters' efforts,* which contains both question topic and answer. In other cases, question topic and answer are embedded in different, non-related spans, which are often remote from each other.

4. Questions for which the topic can be identified in the text and matched to the nucleus of a relevant RST relation, but the corresponding satellite is not suitable or incomplete as answer (11 questions, 3.3%). These are the questions that in table 2 make the difference between the last two rows (209-198). Some answers are unsuitable because they are too long. In other cases, the answer satellite is incomplete compared to the user-formulated answers. For example, the topic of the question *Why did Harold Smith chain his Sagos to iron stakes?* corresponds to the nucleus of a circumstance relation that has the satellite *After three Sagos were stolen from his home in Garden Grove*. Although this satellite gives a possible answer to the question, it is incomplete according to the reference answers, which all mention the goal *To protect his trees from thieves*.

Questions of category 1 above cannot be answered by a QA system using a closed document collection. If we are not able to identify the question topic in the text manually, then a retrieval system cannot either. A comparable problem holds for questions of category 2, where the topic is supported by the source text but the answer is not or only implicitly. If the system searches for an answer that cannot be identified in a text, the system will clearly not find it in that text. In the cases where the answer is implicitly supported by the source text, knowledge of the world is often needed for deducing the answer from the text, like in the examples of cyclosporine and Gerry Hogan above. Therefore, we consider the questions of types 1 and 2 as unsolvable by any QA system that uses a closed document collection. Together these categories cover 21.4% of all *why*-questions.

Questions of category 3 (16.4% of all questions) are the cases where both question topic and answer can be identified in the text, but where the RST structure does not lead to the reference answer. We can search for ways to extent our algorithm so

that it can handle some of the cases mentioned. For instance, we can add functionality for managing question-answer relations on sub-EDU level. For cases where question topic and answer are embedded non-related spans, we can at the moment not propose smart solutions that will increase recall without heavily lowering the MRR. The same holds for questions of category 4 (3.3%), where RST leads to an answer that is incomplete or unsuitable.

3.4.2 Comparison to baseline

In order to judge the value of this maximum recall, we compare the figure of 58.9% to the recall that can be achieved using our baseline approach. As baseline, we chose an approach that exploits textual cues in the source text. We performed a manual search on the 393 questions from our first data collection, their answers and the corresponding source documents. For each question-answer pair, we identified the item in the text that indicates the answer. For 50% of the questions, we could identify a word or group of words that in the given context is a cue for the answer. Most of these cues, however, are very frequent words that also occur in many non-cue contexts. For example, the subordinator *that* occurs 33 times in our document collection, only 3 of which are referred to by one or more *why*-questions. This means that only in 9% of the cases, the subordinator *that* is a *why*-cue. The only two words for which more than 50% of the occurrences are *why*-cues, are *because* (for 18 questions) and *since* (for 9 questions). Both are a *why*-cue in 100% of their occurrences. Almost half of the question-answer pairs that do not have an explicit cue in the source text, the answer is represented by the sentence that follows (69 cases) or precedes (11 cases) the sentence that represents the question.

Having this knowledge on the frequency of cues for *why*-questions, we defined the following baseline approach:

I. Identify the topic of the question.

II. In the source document, identify the clause(s) that express(es) the same proposition as the question topic.

III. Does the clause following the matched clause start with *because* or *since*? If it does, go to IV. If it does not, go to V.

IV. Select the clause following the matched clause as answer.

V. Select the sentence following the sentence containing the matched clause as answer.

A system that follows this baseline method can obtain a maximum recall of 24.4% ((18+9+69)/393). This means that an RST-based method can improve recall by almost 150% compared to a simple cue-based method (58.9% compared to 24.4%).

3.5 RST relations that play a role in *why*-QA

We counted the number of occurrences of the relation types from Table 1 for the 198 questions where the RST relation led to a correct answer. This distribution is

presented in Table 2. The meaning of the column *Relative frequency* in this context will be explained below.

Table 2: Addressed relation types

Relation type	# referring questions	Relative frequency
Means	4	1.000
Purpose	28	0.857
Consequence	37	0.833
Evidence	7	0.750
Reason	19	0.750
Result	19	0.667
Explanation-argumentative	14	0.571
Cause	7	0.500
Condition	1	0.333
Interpretation	6	0.333
Circumstance	1	0.143
Elaboration	50	0.098
Sequence	1	0.091
List	4	0.016
Problem-Solution	0	0.000

As shown in table 5, the relation type with most referring question-answer pairs, is the very general elaboration relation. It seems striking that *elaboration* is more frequent as relation between a *why*-question and its answer than *reason* or *cause*. However, if we look at the relative frequency of the addressed relation types, we see another pattern: in our collection of seven source texts, *elaboration* is a very frequent relation type. In the seven texts that we consider, there are 143 occurrences of an elaboration relation. Of the 143 nuclei of these occurrences, 14 were addressed by one or more *why*-questions, which gives a relative frequency of less than 1%. *Purpose*, on the other hand, has only seven occurrences in our data collection, six of which being addressed by one or more questions, which gives a relative frequency of 0.857. *Reason* and *evidence* both have only four occurrences in the collection, three of which have been addressed by one or more questions.

The table show that if we address the problem of answer selection for *why*-questions as a discourse analysis task, the range of relation types that can lead to an answer is broad and should not be implemented too rigidly.

4 Overall conclusion

We have investigated the relevance of linguistic analysis for *why*-QA. We focused on two tasks: the use of syntactic information for answer type determination and the use of discourse structure for the extraction of potential answers from retrieved documents.

For answer type determination, syntactic analysis appears to be of significance:

we obtain 77.5% performance using a method based on syntactic parses by the TOSCA parser—compared to 58.1% using a similar approach without syntactic analysis.

Discourse analysis appears to be very relevant for extraction of potential answers to *why*-questions. We performed a manual analysis of 336 question-answer pairs and the corresponding RST annotated texts. We found that for 58.9% of *why*-questions, the RST analysis of the source text can lead to a correct answer to the question. Of the remaining 41.1%, there is a subset of *why*-questions (21.4% of all questions) that cannot be answered by any QA system that uses a closed document collection since knowledge of the world is essential for answering these questions. Moreover, there is a further subset of *why*-questions (16.4% + 3.3%) that cannot be answered by a system that uses RST structure only.

We should 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. As a result of that, performance would decline if we were to use automatically created annotations. Some work has been done on automatically annotating text with discourse structure. Promising is the done work by Soricut and Marcu (2003) and Huong and Abeysinghe (2003).

At present, we are working on the implementation of a system for *why*-QA that uses the manually annotated RST treebank as document collection. For the results that we obtained until now, we refer to Verberne, Boves, Oostdijk and Coppen (2006b).

References

Bosma, W.(2005), Query-based summarization using rhetorical structure theory, *in* T. van der Wouden, M. Po, H. Reckman and C. Cremers (eds), *15th Meeting of CLIN*, LOT, Leiden, pp. 29–44. ISBN=90-76864-91-8.

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

Fellbaum, C. E. (ed.)(1998), *WordNet: An Electronic Lexical Database*, Cambridge, Mass.: MIT Press.

Holmes, G., Donkin, A. and Witten, I.(1994), Weka: a machine learning workbench, *Proceedings of the Second Australia and New Zealand Conference on Intelligent Information Systems* pp. 357–361.

Hovy, E., Gerber, L., Hermjakob, U., Lin, C.-J. and Ravichandran, D.(2001), Toward semantics-based answer pinpointing, *Proceedings of the DARPA Human Language Technology Conference (HLT)*, San Diego, CA.

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

Huong, T. L. and Abeysinghe, G.(2003), A Study to Improve the Efficiency of a Discourse Parsing System, *Proc of CICLing-03* pp. 104–117.

Jijkoun, V. and De Rijke, M.(2005), Retrieving answers from frequently asked questions pages on the web, *Proceedings CIKM-2005*.

Kipper, K., Trang Dang, H. and Palmer, M.(2000), Class-based construction of a verb lexicon, *AAAI-2000 Seventeenth National Conference on Artificial Intelligence*, Austin, TX.

Kupiec, J.(1999), Murax: Finding and organizing answers from text search, *Natural Language Information Retrieval* pp. 311–332.

Levin, B.(1993), *English Verb Classes and Alternations - A Preliminary Investigation*, The University of Chicago Press.

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

Maybury, M. (ed.)(2003), *Toward a Question Answering Roadmap*, pp. 8–11.

Moldovan, D., Harabagiu, S., Pasa, R., Mihalcea, R., Grju, R., Goodrum, R. and Rus, V.(2000), The structure and performance of an open domain question answering system, *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000)*, pp. 563–570.

Oostdijk, N.(1996), Using the tosca analysis system to analyse a software manual corpus, *Industrial Parsing of Software Manuals* pp. 179–206.

Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J.(1985), *A comprehensive grammar of the English language*, London: Longman.

Soricut, R. and Marcu, D.(2003), Sentence level discourse parsing using syntactic and lexical information, *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1* pp. 149–156.

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

Verberne, S., Boves, L., Oostdijk, N. and Coppen, P.(2006b), Discourse-based answering of *why*-questions. Submitted for *Traitement automatique des langues (TAL), Computational Approaches to Discourse and Document Processing*.

Voorhees, E. and Tice, D.(2000), Building a question answering test collection, *Proceedings of SIGIR-2000*, pp. 200–207.