Developing an approach for *why*-question answering

Suzan Verberne
Dept. of Language and Speech
Radboud University Nijmegen
`s.verberne@let.ru.nl`

Abstract

In the current project, we aim at developing an approach for automatically answering *why*-questions. We created a data collection for research, development and evaluation of a method for automatically answering *why*-questions (*why*-QA) The resulting collection comprises 395 *why*-questions. For each question, the source document and one or two user-formulated answers are available in the data set. The resulting data set is of importance for our research and it will contribute to and stimulate other research in the field of *why*-QA. We developed a question analysis method for *why*-questions, based on syntactic categorization and answer type determination. The quality of the output of this module is promising for future development of our method for *why*-QA.

1 Introduction

Until now, research in the field of automatic question answering (QA) has focused on factoid (closed-class) questions like *who*, *what*, *where* and *when* questions. Results reported for the QA track of the Text Retrieval Conference (TREC) show that these types of *wh*-questions can be handled rather successfully (Voorhees 2003).

In the current project, we aim at developing an approach for automatically answering *why*-questions. So far, *why*-questions have largely been ignored by researchers in the QA field. One reason for this is that the frequency of *why*-questions in a QA context is lower than that of other questions like *who*- and *what*-questions (Hovy et al., 2002a). However, although *why*-questions are less frequent than some types of factoids (*who*, *what* and *where*), their frequency is not negligible: in a QA context, they comprise about 5 percent of all *wh*-questions (Hovy, 2001;

Jijkoun, 2005) and they do have relevance in QA applications (Maybury, 2002). A second reason for ignoring *why*-questions until now, is that it has been suggested 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 rather than a noun phrase (Kupiec, 1999). The current paper aims to find out whether the suggestion is true that factoid-QA techniques are not suitable for *why*-QA. We want to investigate whether principled syntactic parsing can make QA for *why*-questions feasible.

In the present paper, we report on the work that has been carried out until now. More specifically, sections 2 and 3 describe the approach taken to data collection and question analysis and the results that were obtained. Then, in section 4, we discuss the plans and goals for the work that will be carried out in the remainder of the project.

2 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, 2000). While in the context of the QA track of TREC data collections in support of factoid questions have been created, so far, no resources have been created for *why*-QA. For the purpose of the present research therefore, we have developed a data collection comprising a set of questions and corresponding answers. In doing so, we have extended the time tested procedures previously developed in the TREC context.

In this section, we describe the requirements that a data set must meet to be appropriate for development and we discuss a number of existing sources of *why*-questions. Then we describe the method employed for data collection

and the main characteristics of the resulting data set.

The first requirement for an appropriate data set concerns the nature of the questions. In the context of the current research, a *why*-question is defined as an interrogative sentence in which the interrogative adverb *why* (or one of its synonyms) occurs in (near) initial position. We consider the subset of *why*-questions that could be posed in a QA context and for which the answer is known to be present in the related document set. This means that the data set should only comprise *why*-questions for which the answer can be found in a fixed collection of documents. Secondly, the data set should not only contain questions, but also the corresponding answers and source documents. The answer to a *why*-question is a clause or sentence (or a small number of coherent sentences) that answers the question without giving supplementary context. The answer is not literally present in the source document, but can be deduced from it. For example, a possible answer to the question *Why are 4300 additional teachers required?*, based on the source snippet *The school population is due to rise by 74,000, which would require recruitment of an additional 4,300 teachers*, is *Because the school population is due to rise by a further 74,000.*

Finally, the size of the data set should be large enough to cover all relevant variation that occur in *why*-questions in a QA context.

There are a number of existing sources of *why*-questions that we may consider for use in our research. However, for various reasons, none of these appear suitable.

Why-questions from corpora like the British National Corpus (BNC, 2002), in which questions typically occur in spoken dialogues, are not suitable because the answers are not structurally available with the questions, or they are not extractable from a document that has been linked to the question. The same holds for the data collected for the Webclopedia project (Hovy et al., 2002a), in which neither the answers nor the source documents were included. One could also consider questions and answers from frequently asked questions (FAQ) pages, like the large data set collected by Valentin Jijkoun (Jijkoun, 2005). However, in FAQ lists, there is no clear distinction between the answer itself (a clause that answers the question) and the source document that contains the answer.

The questions in the test collections from the TREC-QA track do contain links to the possible answers and the corresponding source documents. However, these collections contain too few *why*-questions to qualify as a data set that is appropriate for developing *why*-QA.

Given the lack of available data that match our requirements, a new data set for QA research into *why*-questions had to be compiled. In order to meet the given requirements, it would be best to collect questions posed in an operational QA environment, like the compilers of the TREC-QA test collections did: they extracted factoid and definition questions from search logs donated by Microsoft and AOL (TREC, 2003). Since we do not have access to comparable sources, it was decided to revert to the procedure used in earlier TRECs, and imitate a QA environment in an elicitation experiment. We extended the conventional procedure by collecting user-formulated answers in order to investigate the range of possible answers to each question. We also added paraphrases of collected questions in order to extend the syntactic and lexical variation in the data collection.

In the elicitation experiment, ten native speakers of English were asked to read five texts from Reuters' *Textline Global News* (1989) and five texts from *The Guardian on CD-ROM* (1992). The texts were around 500 words each. The experiment was conducted over the Internet, using a web form and some CGI scripts. In order to have good control over the experiment, we registered all participants and gave them a code for logging in on the web site. Every time a participant logged in, the first upcoming text that he or she did not yet finish was presented. The participant was asked to formulate one to six *why*-questions for this text, and to formulate an answer to each of these questions. The participants were explicitly told that it was essential that the answers to their questions could be found in the text. After submitting the form, the participant was presented the questions posed by one of the other participants and he or she was asked to formulate an answer to these questions too. The collected data was saved in text format, grouped per participant and per source document, so that the source information is available for each question. The answers have been linked to the questions.

In this experiment, 395 questions and 769 corresponding answers were collected. The number of answers would have been twice the

number of questions if all participants would have been able to answer all questions that were posed by another participant. However, for 21 questions (5.3%), the second participant was not able to answer the first participant's question. Note that not every question in the elicitation data set has a unique topic[1]: on average, 38 questions were formulated per text, covering around twenty topics per text.

The collected questions have been formulated by people who had constant access to the source text. As a result of that, the chosen formulations often resemble the original text, both in the use of vocabulary and sentence structure. In order to expand the dataset, a second elicitation experiment was set up, in which five participants from the first experiment were asked to paraphrase some of the original *why*-questions. The 166 unique questions were randomly selected from the original data set. The participants formulated 211 paraphrases in total for these questions. This means that some questions have more than one paraphrase. The paraphrases were saved in a text file that includes the corresponding original questions and the corresponding source documents.

We studied the types of variation that occur among questions covering the same topic. First, we collected the types of variation that occur in the original data set and then we compared these to the variation types that occur in the set of paraphrases.

In the original data set, the following types of variation occur between different questions on the same topic:

Lexical variation, e.g.
    `for the second year running` vs. `again`;
Verb tense variation, e.g.
    `have risen` vs. `have been rising`;
Optional constituents variation, e.g.
    `class sizes` vs. `class sizes in England and Wales`;
Sentence structure variation, e.g.
    `would require recruitment` vs. `need to be recruited`

In the set of paraphrases, the same types of variation occur, but as expected the differences between the paraphrases and the source

---

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

sentences are slightly bigger than the differences between the original questions and the source sentences. We measured the lexical overlap between the questions and the source texts as the number of content words that are in both the question and the source text. The average relative lexical overlap (the number of overlapping words divided by the total number of words in the question) between original questions and source text is 0.35; the average relative lexical overlap between paraphrases and source text is 0.31.

The size of the resulting collection (395 original questions, 769 answers, and 211 paraphrases of questions) is large enough to initiate serious research into the development of *why*-QA.

Our collection meets the requirements that were formulated with regard to the nature of the questions and the presence of the answers and source documents for every question.

3   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 a query that contains all information about the answer that can be extracted from the question. So far, no question analysis procedures have been created for *why*-QA specifically. Therefore, we have developed an approach for proper analysis of *why*-questions. Our approach is based on existing methods of analysis of factoid questions. This will allow us to verify whether methods used in handling factoid questions are suitable for use with procedural questions. In this section, we describe the components of successful methods for the analysis of factoid questions. Then we present the method that we used for the analysis of *why*-questions and indicate the quality of our method.

The first (and most simple) component in current methods for question analysis is keyword extraction. Lexical items in the question give information on the topic of the user's information need. In keyword selection, several different approaches may be followed. Moldovan et al. (2000), for instance, select as keywords all named entities that were recognized as proper nouns. In almost all approaches to keyword extraction, syntax plays a role. Shallow parsing is used for extracting noun phrases, which are considered to be relevant key phrases in the retrieval step. Based on the query's keywords,

one or more documents or paragraphs can be retrieved that may possibly contain the answer.

A second, very important, component in question analysis is determination of the question's semantic answer type. The answer type of a question defines the type of answer that the system should look for. Often-cited work on question analysis has been done by Moldovan et al. (1999, 2000), Hovy et al. (2001), and Ferret et al. (2002). They all describe question analysis methods that classify questions with respect to their answer type. In their 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. 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.

For determining the answer type, syntactic analysis may play a role. When implementing a syntactic analysis module in a working QA system, the analysis has to be performed fully automatically. This may lead to concessions with regard to either the degree of detail or the quality of the analysis. Ferret et al. implement a syntactic analysis component based on shallow parsing. Their syntactic analysis module yields a syntactic category for each input question. In their system, a syntactic category is a specific syntactic pattern, such as 'WhatDoNP' (e.g. *What does a defibrillator do?*) or 'WhenBePNborn' (e.g. *When was Rosa Park born?*). They define 80 syntactic categories like these. Each input question is parsed by a shallow parser and hand-written rules are applied for determining the syntactic category. Ferret et al. find that the syntactic pattern helps in determining the semantic answer type (e.g. *company*, *person*, *date*). They unfortunately do not describe how they created the mapping between syntactic categories and answer types.

As explained above, determination of the semantic answer type is the most important task of existing question analysis methods. Therefore, the goal of our question analysis method is to predict the answer type of *why*-questions.

In the work of Moldovan et al. (2000), all *why*-questions share the single answer type *reason*. However, we believe that it is necessary to split this answer type into sub-types, because a more specific answer type helps the system select potential answer sentences or paragraphs. 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 (1985:15.45), we distinguish the following sub-types of *reason*: *cause*, *motivation*, *circumstance* (which combines reason with conditionality), and *purpose*.

Below, an example of each of these answer types is given.

Cause:
```
The flowers got dry because it
hadn't rained in a month.
```
Motivation:
```
I water the flowers because I
don't like to see them dry.
```
Circumstance:
```
Seeing that it is only three,
we should be able to finish
this today.
```
Purpose:
```
People have eyebrows to prevent
sweat running into their eyes.
```

The *why*-questions that correspond to the reason clauses above are respectively *Why did the flowers get dry?*, *Why do you water the flowers?*, *Why should we be able to finish this today?*, and *Why do people have eyebrows?*. It is not always possible to assign one of the four answer sub-types to a *why*-question. We will come back to this later.

Often, the question gives information on the expected answer type. For example, compare the two questions below:

```
Why did McDonald's write Mr.
Bocuse a letter?

Why have class sizes risen?
```

Someone asking the former question expects as an answer McDonald's *motivation* for writing a letter, whereas someone asking the latter question expects the *cause* for rising class sizes as answer.

The corresponding answer paragraphs do indeed contain the equivalent answer sub-types:

```
McDonald's has acknowledged
that a serious mistake was
made. "We have written to
apologise and we hope to reach
```

```
a settlement with Mr. Bocuse
this week," said Marie-Pierre
Lahaye, a spokeswoman for
McDonald's France, which
operates 193 restaurants.

Class sizes in schools in
England and Wales have risen
for the second year running,
according to figures released
today by the Council of Local
Education Authorities. The
figures indicate that although
the number of pupils in schools
has risen in the last year by
more than 46,000, the number of
teachers fell by 3,600.
```

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 the needed information on the expected answer type. In case of *why*, the question word does not give information on the answer type since all *why*-questions have *why* as 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. TOSCA's syntactic parser takes a sequence of unambiguously 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 sentences. For the purpose of evaluating the maximum contribution to a classification method that can be obtained from a principled syntactic analysis, the most plausible parse tree from the parser's output is selected manually.

For the next step of question analysis, we created a set of hand-written rules, which are applied to the parse tree in order to choose the question's syntactic category. We defined six syntactic categories for this purpose:

Action questions, e.g.
```
Why did McDonald's write Mr.
Bocuse a letter?
```
Process questions, e.g.
```
Why has Dixville grown famous
since 1964?
```

Intensive complementation questions, e.g.
```
Why is Microsoft Windows a
success?
```
Monotransitive *have* questions, e.g.
```
Why did compilers of the OED
have an easier time?
```
Existential *there* questions, e.g.
```
Why is there a debate about
class sizes?
```
Declarative layer questions, e.g.
```
Why does McDonald's spokeswoman
think the mistake was made?
```

The choice for these categories is based the information that is available from the parser, and the information that is needed for determining the answer type.

For some categories, the question analysis module only needs fairly simple cues for choosing a category. For example, a main verb with the feature *intens* leads to the category 'intensive complementation question' and the presence of the word *there* with the syntactic category *EXT* leads to the category 'existential *there* question'. For deciding on declarative layer questions, action questions and process questions, complementary lexical-semantic information is needed. In order to decide whether the question contains a declarative layer, the module checks whether the main verb is in a list that corresponds to the union of the verb classes *say* and *declare* from Verbnet (Kipper et al., 2000), and whether it has a clausal object. The distinction between action and process questions is made by looking up the main verb in a list of process verbs. This list contains the 529 verbs from the causative/inchoative alternation class (verbs like *melt* and *grow*) from the Levin verb index (Levin, 1993); in an intransitive context, these verbs are process verbs.

We have not yet developed an approach for passive questions.

Based on the syntactic category, the question analysis module tries to determine the answer type. Some of the syntactic categories lead directly to an answer type. All process questions with non-agentive subjects get the expected answer type *cause*. All action questions with agentive subjects get the answer type *motivation*. We extracted information on agentive and non-agentive nouns from WordNet: all nouns that are in the lexicographer file `noun.person` were selected as agentive.

Other syntactic categories need further analysis. Questions with a declarative layer, for example,

are ambiguous. The question *Why did they say that migration occurs?* can be interpreted in two ways: *Why did they say it?* or *Why does migration occur?*. Before deciding on the answer type, our question analysis module tries to find out which of these two questions is supposed to be answered. In other words: the module decides which of the clauses has the question focus. This decision is made on the basis of the semantics of the declarative verb. If the declarative is a factive verb – a verb that presupposes the truth of its complements – like *know*, the module decides that the main clause has the focus. The question consequently gets the answer type *motivation*. In case of a non-factive verb like *think*, the focus is expected to be on the subordinate clause. In order to predict the answer type of the question, the subordinate clause is then treated the same way as the complete question was. For example, consider the question *Why do the school councils believe that class sizes will grow even more?*. Since the declarative (*believe*) is non-factive, the question analysis module determines the answer type for the subordinate clause (*class sizes will grow even more*), which is *cause*, and assigns it to the question as a whole.

Special attention is also paid to questions with a modal auxiliary. Modal auxiliaries like *can* and *should*, have an influence on the answer type. For example, consider the questions below, in which the only difference is the presence or absence of the modal auxiliary *can*:

```
Why did McDonalds not use
actors to portray chefs in
amusing situations?

Why can McDonalds not use
actors to portray chefs in
amusing situations?
```

The former question expects a motivation as answer, whereas the latter question expects a cause. We implemented this difference in our question analysis module: CAN (can, could) and HAVE TO (have to, has to, had to) lead to the answer type *cause*. Furthermore, the modal auxiliary SHALL (shall, should) changes the expected answer type to *motivation*.

When choosing an answer type, our question analysis module follows a conservative policy: in case of doubt, no answer type is assigned.

We did not yet perform a complete evaluation of our question analysis module. For proper evaluation of the module, we need a reference set of questions and answers that is different from the data set that we collected for development of our system. Moreover, for evaluating the relevance of our question analysis module for answer retrieval, further development of our approach is needed.

However, to have a general idea of the performance of our method for answer type determination, we compared the output of the module to manual classifications. We performed these reference classifications ourselves.

First, we manually classified 130 *why*-questions from our development set with respect to their syntactic category. Evaluation of the syntactic categorization is straightforward: 95 percent of *why*-questions got assigned the correct syntactic category using 'perfect' parse trees. The erroneous classifications were due to differences in the definitions of the specific verb types. For example, *argue* is not in the list of declarative verbs, as a result of which a question with *argue* as main verb is classified as action question instead of declarative layer question. Also, *die* and *cause* are not in the list of process verbs, so questions with either of these verbs as main verb are labeled as action questions instead of process questions.

Secondly, we performed a manual classification into the four answer sub-types (*cause*, *motivation*, *circumstance* and *purpose*). For this classification, we used the same set of 130 questions as we did for the syntactic categorization, combined with the corresponding answers. Again, we performed this classification ourselves.

During the manual classification, we assigned the answer type *cause* to 23.3 percent of the questions and *motivation* to 40.3 percent. We were not able to assign an answer sub-type to the remaining pairs (36.4 percent). These questions are in the broader class *reason* and not in one of the specific sub-classes None of the question-answer pairs was classified as *circumstance* or *purpose*. Descriptions of purpose are very rare in news texts because of their generic character (e.g. *People have eyebrows to prevent sweat running into their eyes*). The answer type *circumstance*, defined by Quirk (cf. section 15.45) as a combination of reason with conditionality, is also rare as well as difficult to recognize.

For evaluation of the question analysis module, we mainly considered the questions that

did get assigned a sub-type (*motivation* or *cause*) in the manual classification. Our question analysis module succeeded in assigning the correct answer sub-type to 62.2 percent of these questions, the wrong sub-type to 2.4 percent, and no sub-type to the other 35.4 percent. The set of questions that did not get a sub-type from our question analysis module can be divided in four groups:

(a) Action questions for which the subject was incorrectly not marked as agentive (mostly because it was an agentive organization like *McDonald's*, or a proper noun that was not in WordNet's list of nouns denoting persons, like *Henk Draijen*);

(b) questions with an action verb as main verb but a non-agentive subject (e.g. *Why will restrictions on abortion damage women's health?*);

(c) passive questions, for which we have not yet developed an approach (e.g. *Why was the Supreme Court reopened?*);

(d) Monotransitive *have* questions. This category contains too few questions to formulate a general rule.

Group (a), which is by far the largest of these four (covering half of the questions without sub-type), can be reduced by expanding the list of agentive nouns, especially with names of organizations. For groups (c) and (d), general rules may possibly be created in a later stage.

With this knowledge, we are confident that we can reduce the number of questions without sub-type in the output of our question analysis module.

These first results predict that it is possible to reach a relatively high precision in answer type determination. (Only 2 percent of questions got assigned a wrong sub-type.) A high precision makes the question analysis output useful and reliable in the next steps of the question answering process. On the other hand, it seems difficult to get a high recall. In this test, only 62.2 percent of the questions that were assigned an answer type in the reference set, was assigned an answer type by the system – this is 39.6 percent of the total.

4 Conclusions and further research

We created a data collection for research into *why*-questions and for development of a method for *why*-QA. The collection comprises a sufficient amount of *why*-questions. For each question, the source document and one or two user-formulated answers are available in the data set. The resulting data set is of importance for our research as well as other research in the field of *why*-QA.

We developed a question analysis method for *why*-questions, based on syntactic categorization and answer type determination. In-depth evaluation of this module will be performed in a later stage, when the other parts of our QA approach have been developed, and a test set has been collected. We believe that the first test results, which show a high precision and low recall, are promising for future development of our method for *why*-QA.

We think that, just as for factoid-QA, answer type determination can play an important role in question analysis for *why*-questions. Therefore, Kupiec' suggestion that conventional question analysis techniques are not suitable for *why*-QA can be made more precise by saying that these methods may be useful for a (potentially small) subset of *why*-questions. The issue of recall, both for human and machine processing, needs further analysis.

In the near future, our work will focus on development of the next part of our approach for *why*-QA.

Until now we have focused on the first of four sub-tasks in QA, viz. (1) question analysis (2) retrieval of candidate paragraphs; (3) paragraph analysis and selection; and (4) answer generation. Of the remaining three sub-tasks, we will focus on paragraph analysis (3). In order to clarify the relevance of the paragraph analysis step, let us briefly discuss the QA-processes that follows question analysis.

The retrieval module, which comes directly after the question analysis module, uses the output of the question analysis module for finding candidate answer paragraphs (or documents). Paragraph retrieval can be straightforward: in existing approaches for factoid-QA, candidate paragraphs are selected based on keyword matching only. For the current research, we do not aim at creating our own paragraph selection technique.

More interesting than paragraph retrieval is the next step of QA: paragraph analysis. The paragraph analysis module tries to determine whether the candidate paragraphs contain potential answers. In case of *who*-questions, noun phrases denoting persons are potential

answers; in case of *why*-questions, *reasons* are potential answers. In the paragraph analysis stage, our answer sub-types come into play. The question analysis module determines the answer type for the input question, which is *motivation*, *cause*, *purpose*, or *circumstance*. The paragraph analysis module uses this information for searching candidate answers in a paragraph. As has been said before, the procedure for assigning the correct sub-type needs further investigation in order to increase the coverage and the contribution that answer sub-type classification can make to the performance of *why*-question answering.

Once the system has extracted potential answers from one or more paragraphs with the same topic as the question, the eventual answer has to be delimited and reformulated if necessary.

References

British National Corpus, 2002. *The BNC Sampler.* Oxford University Computing Services.

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

Ferret O., Grau B., Hurault-Plantet M., Illouz G., Monceaux L., Robba I., and Vilnat A., 2002. Finding An Answer Based on the Recognition of the Question Focus. In *Proceedings of The Tenth Text REtrieval Conference (TREC 2001)*. Gaithersburg, Maryland: NIST Special Publication SP 500-250.

Hovy, E.H., Gerber, L., Hermjakob, U., Lin, C-J, and Ravichandran, D., 2001. Toward Semantics-Based Answer Pinpointing. In *Proceedings of the DARPA Human Language Technology Conference (HLT)*. San Diego, CA

Hovy, E.H., Hermjakob, U., and Ravichandran, D., 2002a. A Question/Answer Typology with Surface Text Patterns. In *Proceedings of the Human Language Technology conference (HLT)*. San Diego, CA.

Jijkoun, V. and De Rijke, M., 2005. Retrieving Answers from Frequently Asked Questions Pages on the Web. In: *Proceedings CIKM-2005*, to appear.

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.M., 1999. MURAX: Finding and Organizing Answers from Text Search. In Strzalkowski, T. (ed.) *Natural Language Information Retrieval*. 311-332. Dordrecht, Netherlands: Kluwer Academic.

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

Litkowski, K. C., 1998. Analysis of Subordinating Conjunctions, CL Research Technical Report 98-01 (Draft). CL Research, Gaithersburg, MD.

Maybury , M., 2003. Toward a Question Answering Roadmap. In *New Directions in Question Answering 2003*: 8-11

Moldovan, D., S., Harabagiu, M., Pasça, R., Mihalcea, R., Gîrju, R., Goodrum, R., and Rus, V. 1999. Lasso: A Tool for Surfing the Answer Net. 175-184. In E. Voorhees and D. Harman (Eds.), *NIST Special Publication* 500-246. The Eight Text REtrieval Conference. Dept. of Commerce, NIST.

Moldovan, D., S., Harabagiu, M., Pasça, R., Mihalcea, R., Gîrju, R., Goodrum, R., and Rus, V. 2000. The Structure and Performance of an Open Domain Question Answering System. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000)*: 563-570.

Oostdijk, N., 1996. Using the TOSCA analysis system to analyse a software manual corpus. In: R. Sutcliffe, H. Koch and A. McElligott (eds.), *Industrial Parsing of Software Manuals*. Amsterdam: Rodopi. 179-206.

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

Text Retrieval Conference (TREC) QA track, 2003. http://trec.nist.gov/data/qamain.html

Voorhees, E. & Tice, D., 2000. Building a Question Answering Test Collection. In *Proceedings of SIGIR-2000*: 200-207

Voorhees, E., 2003. Overview of the TREC 2003 Question Answering Track. In *Overview of TREC 2003*: 1-13