Evaluating paragraph retrieval for *why*-QA

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

Department of Linguistics, Radboud University Nijmegen,
s.verberne|l.boves|n.oostdijk|p.a.coppen@let.ru.nl

Abstract. We implemented a baseline approach to *why*-question answering based on paragraph retrieval. Our implementation incorporates the QAP ranking algorithm with addition of a number of surface features (cue words and XML markup). With this baseline system, we obtain an accuracy-at-10 of 57.0% with an MRR of 0.31. Both the baseline and the proposed evaluation method are good starting points for the current research and other researchers working on the problem of *why*-QA.
We also experimented with the addition of smart question analysis to our baseline system (answer type and informational value of the subject). This however did not give significant improvement to our baseline. In the near future, we will investigate what other linguistic features can facilitate re-ranking in order to increase accuracy.

1 Introduction

In the current research project, we aim at developing a system for answering *why*-questions (*why*-QA). In earlier experiments, we found that the answers to *why*-questions consist of a type of reasoning that cannot be expressed in a single clause, and that on the other hand 94% of the answers is maximally one paragraph long. Therefore, we decide to consider paragraphs as retrieval units for *why*-QA.

The goal of the present paper is to establish a baseline paragraph retrieval method for *why*-QA, including a proper evaluation method. Moreover, we aim to find out whether a system based on standard keyword based paragraph retrieval can be improved by incorporating our knowledge of the syntax and semantics of *why*-questions in query formulation.

2 Method

2.1 Data

For development and testing, we use a set of 805 *why*-questions that were submitted to the online QA system answers.com, and collected for the Webclopedia project by Hovy et al. [1].

As an answer source, we use the Wikipedia XML corpus [2], which is also used in the context of the Initiative for the Evaluation of XML Retrieval (INEX, [3]). The English part of the corpus consists of 659,388 Wikipedia articles (4.6

GB of XML data). By manual inspection we found that this corpus contains a valid answer for about one quarter of the Webclopedia *why*-questions. We randomly selected 93 questions that have an answer in the corpus and we manually extracted the answer paragraph (reference answer) from the corpus for each of them. We indexed the complete corpus using the Wumpus search engine [4] in the standard indexing modus (Wumpus version June 2007).

2.2 Baseline method

Our baseline method consists of four modules:

1. A question analysis module, which applies a list of stop words to the question and removes punctuation, returning the set of question content words;
2. A query creation module that transforms the set of question words into one or more Wumpus-style queries and sends this query to the Wumpus engine;
3. Ranking of the retrieved answers by the QAP algorithm. QAP is a scoring algorithm for passages that has specifically been developed for question answering tasks [5]. It has been implemented in Wumpus;
4. Re-ranking the results according to three answer features: (a) The presence of cue words such as *because*, *due to* and *in order to* in the paragraph; (b) the presence of one or more question terms in the title of the document in which the retrieved paragraph is embedded; (c) emphasis marking of a question term. The corpus contains XML tags for formatting information such as emphasis.
   The weights applied in the re-ranking step are variable in the configuration of our system.

After re-ranking, the system returns the top 10 results to the user.

2.3 Features for smart question analysis

In this section, we present an extension to our paragraph retrieval system incorporating smart question analysis. In previous experiments, we found that specific syntactic and semantic features of *why*-questions can play a role in retrieving relevant answers. We identified two features in particular that seem relevant in answer selection, viz. answer type and the informational value of the subject.

In factoid QA, **answer types** is known to be an important parameter for increasing system precision. The two main answer types for *why*-questions are 'cause' and 'motivation' [6]. In our question set, we encountered one other relatively frequent answer type: 'etymology'. Thus we distinguish three answer types in the current approach: 'cause' (77.4% in our question set), 'motivation' (10.2%), and 'etymology' (12.4%). We split our set of cue words in four categories: one for each of the answer types (e.g. *in order to* for motivation, *due to* for cause and *name* for etymology), and a general category of cue words that occur for all answer types (e.g. *because*). We evaluated answer type prediction for our question set using earlier defined algorithms and we found a precision of 0.806 (ranging from 0.487 for motivation to 1 for etymology) for this task.

Previous experiments have shown the relevance of a second semantic feature, the **informational value of the subject**. It appears to be a good predictor for deciding which terms from the question are likely to occur in the document title of relevant answer paragraphs. This knowledge can be used for re-ranking based on document title (step 4b in the baseline method). We defined three classes of subjects, which are automatically distinguished by our system based on their document frequency. The subjects with lowest informational value are subjects consisting of pronouns only or one of the very general noun phrases *people* and *humans*. In these cases, our re-ranking module only gives extra weight to *predicate* words occurring in the document title. The second class covers those subjects that are not semantically poor, but very common, such as *water* and *the United States*. In these cases, the baseline approach is applied, which does not distinguish between terms from subject and predicate for re-ranking. The third class consists of the subjects that have a low document frequency, and therefore have a large informational value, such as *flamingos* and *subliminal messages*. In these cases our system gives extra weight to paragraphs from documents with one or more words from the *subject* in the title.

We performed a series of experiments in order to find out what the contribution of these features is to the overall performance of our system.

2.4 Evaluation method

There are no specific evaluation procedures available for *why*-QA, but there is one evaluation forum that includes *why*-questions: the Question Answering Challenge at the Japanese NTCIR Workshop [7]. In NTCIR, all retrieved results are manually evaluated according to a four-level scale of correctness.

We propose a method for the evaluation of *why*-QA that is a combination of the procedure applied at NTCIR and the commonly-used MRR metric. We manually evaluate all retrieved answers according to the four NTCIR correctness scales. Then we count the proportion of questions that has at least one correct answer in the top 10 of the results (accuracy-at-10). For the highest ranked correct answer per question, we determine the reciprocal rank (RR). If there is no correct answer in the top 10 results, RR is 0. Over all questions, we calculate MRR.

3 Results and discussion

Table 1 shows the results (accuracy-at-10 and MRR) obtained for three configurations: (1) simple paragraph retrieval by QAP, (2) the baseline system and (3) the smart system.

Using the Wilcoxon Signed-Rank Test we find there is no significant difference between the baseline results and the results from smart question analysis (Z=-0.66, P=0.5093 for paired reciprocal ranks). The baseline is, however, sightly better than simple retrieval ($Z = 1.67$, $P = 0.0949$).

Table 1. Results per system version

Features	Version	Accuracy	MRR
QAP	Simple retrieval	47.3%	0.25
+Cue words +Title weight +Emph. weight	Baseline	57.0%	0.31
+Answer type +Subject value weight	Smart question analysis	55.9%	0.28

Apparently, the implementation of our question analysis features does not improve the ranking of the results. Since we suspected that some correct answers were missed because they are in the tail of the result list, we experimented with a larger result list (top 20 presented to user). This led to an accuracy-at-20 of 63.4% (MRR unchanged 0.31) for the baseline system and 61.3% (MRR unchanged 0.28) for the smart system.

As regards the answer type feature, we can explain its negligible contribution from the fact that answer type only affects cue word weights. Cue words apparently constitute too small a contribution to the overall performance of the system. As regards the subject value feature, we are surprised by its small influence. Our suspicion is that the ranking algorithm QAP as implemented in the baseline already gives good results with term weighting based on term frequency and inverted document frequency. Another possible explanation to the small influence of the informational value of the subject is that too many errors are still made by our question analysis module in the decision of which question part should be given the position weight.

A further error analysis shows that for 47.5% of unanswered questions, the reference answer is present in the extended result list retrieved by the algorithm (max. 450 results), but not in the top 10 of answers presented to the user. For these questions, re-ranking may be valuable. If we can define criteria that rank the reference answer for this set of questions higher than the irrelevant answers, we can increase accuracy-at-10 (and thereby MRR).

4 Conclusion and further work

We developed an approach for *why*-QA that is based on paragraph retrieval. We created a baseline system that combines paragraph ranking using the QAP algorithm with weights based cue words and the position of question terms in the answer document. We evaluated our system based on manual assessments of the answers in four categories according to two measures: accuracy-at-10 and MRR. We get 57.0% accuracy with an MRR of 0.31. We think that both the baseline and the proposed evaluation method are good starting points for the current research and other researchers working on the problem of *why*-QA.

We also implemented and evaluated a system that extends the baseline approach with two features that we obtain from linguistic question analysis: answer type and the informational value of the subject. This smart system does, however, not show significant improvement over the baseline. In section 3, we do some suggestions for explaining these results.

In the near future, we will experiment with adding a number of other linguistic features to the re-ranking module of our system. The features that we consider for re-ranking include the distinction between heads and modifiers from the question, synonym links between question and answer terms, and the presence of noun phrases from the question in the answer. We are currently preparing experiments for selecting the most relevant of these features for optimizing MRR by re-ranking.

In the more distant future, we plan to experiment with smart paragraph analysis. In [9], it is shown that rhetorical relations have relevance for answer selection in *why*-QA; the presence of (some types of) rhetorical relations can be an indication for the presence of a potential answer. Moreover, there is a connection between answer type and type of rhetorical relation; we aim to investigate whether this addition can make answer type more valuable than in the current cue-word based version of the system.

References

1. Hovy, E., Hermjakob, U., Ravichandran, D.: A question/answer typology with surface text patterns. In: Proceedings of the Human Language Technology conference (HLT), San Diego, CA (2002)
2. Denoyer, L., Gallinari, P.: The Wikipedia XML corpus. ACM SIGIR Forum **40**(1) (2006) 64–69
3. Clarke, C., Kamps, J., Lalmas, M.: Inex 2006 retrieval task and result submission specification. INEX 2006 Workshop Pre-Proceedings, Dagstuhl, Germany, December (2006) 18–20
4. Buttcher, S.: The wumpus search engine. http://www.wumpus-search.org/ (2007)
5. Buttcher, S., Clarke, C., Cormack, G.: Domain-specific synonym expansion and validation for biomedical information retrieval (multitext experiments for trec 2004). (2004)
6. Verberne, S., Boves, L., Oostdijk, N., Coppen, P.: Exploring the use of linguistic analysis for *why*-question answering. In: Proceedings of the 16th meeting of Computational Linguistics in the Netherlands (CLIN 2005), Amsterdam. (2006) 33–48
7. Fukumoto, J., Kato, T., Masui, F., Mori, T.: An overview of the 4th question answering challenge (qac-4) at ntcir workshop 6. In: Proceedings of NTCIR-6 Workshop Meeting, Tokyo, Japan (2007) 433–440
8. Itakura, K.Y., Clarke, C.L.A.: From passages into elements in xml retrieval. In: Proceedings of the SIGIR 2007 Workshop on Focused Retrieval, Amsterdam, the Netherlands (2007) 17–22
9. Verberne, S., Boves, L., Oostdijk, N., Coppen, P.: Discourse-based answering of *why*-questions. Traitement Automatique des Langues, special issue on Computational Approaches to Discourse and Document Processing (2007) 21–41