

Is Context Actually Helpful? Preliminary Experiments in Contextual Question Answering

Steven Winikoff and Leila Kosseim

Department of Computer Science

Concordia University

Montreal, Canada

[smw|kosseim]@cs.concordia.ca

Abstract

In this paper, we present a preliminary experiment in contextual question-answering. The performance of the QUANTUM question answering system on a series of context questions is compared to the performance on fully specified versions of the same questions. Surprisingly, the MRR obtained for the fully specified questions is not significantly higher than that obtained for the original questions.

1 Introduction

Consider a question such as the following:

Which museum in Florence was damaged by a major bomb explosion in 1993?

The problem of answering short, factual questions such as this by searching a collection of documents has been studied widely since the eighth NIST-sponsored TREC competition in 1999 (Voorhees, 1999a; Voorhees, 1999b).

A “context task” was introduced at the tenth TREC competition, in 2001 (Voorhees, 2001b; Voorhees, 2001a). The goal of this task was to answer a sequence of questions in which only the first was fully specified. For example, following the question shown above, subsequent questions might be

On what day did this happen?

or

How many people were killed?

These subsequent questions rely on information supplied either as part of a previous question in the same sequence, or on information contained in an answer to a previous question in the same sequence.

Results from the context task at TREC 10 were surprising, in that there was no correlation between the ability to answer the first question in a series and the ability to answer subsequent questions in the same series. The following explanation is offered in (Voorhees, 2001a):

The first question in a series defined a small enough subset of documents that results were dominated by whether the system could answer the particular type of the current question, rather than by the systems’ ability to track context.

Following this experience the context task was dropped from all subsequent TREC competitions.

The obvious approach to the context task is to rewrite each context question, replacing anaphora by their antecedents so that the result is fully specified. If this can be accomplished accurately then contextual questions can be made independent, thus reducing them to the general question answering problem.

However, the TREC 10 results suggest that a simpler approach might exist, based on taking advantage of the document subset identified while answering the original questions. Perhaps anticipating this line of reasoning, Harabagiu *et al* proposed at TREC 10 that the reference resolution process for question answering can be

simplified compared to that required for discourse or dialog processing, because in the question answering domain it generally suffices to identify the question in which the antecedents of a reference occur without needing to identify the specific antecedents (Harabagiu et al., 2001). The goal of the present work was to determine whether this proposition is reasonable.

2 Experimental Design

Experiments were performed using the QUANTUM question answering system (Plamondon et al., 2001; Plamondon et al., 2002; Plamondon and Kosseim, 2002). No modifications were made to the software, which currently makes no specific provision to handle context questions.

Two sets of questions were used, each with a different document collection. The first question set consisted of the same context questions used at TREC 10 (42 questions divided into 10 sequences of 3 to 9 questions each); these were used with the same document collection provided for TREC 10 (TIPSTER, containing approximately one million documents).

The second question set was based on a collection of 225 documents obtained with permission from the web site of Bell Canada, as part of a project to develop a closed-domain question answering system to answer questions about Bell’s service offerings. This set consisted of 26 questions divided into 7 sequences of 2 to 7 questions each.

For each question set, we began by using QUANTUM to answer all of the questions in the set in their original form. Next the questions were manually reformulated so that all anaphora were replaced by their correct antecedents. For example, the two context questions presented in Section 1 were reformulated as follows:

On what day was the Uffizi gallery in Florence damaged by a major bomb explosion in 1993?

and

How many people were killed in the explosion at the Uffizi gallery in Florence in 1993?

The intent of these reformulations was to evaluate the difference in QUANTUM’s performance attributable to the lack of information in the original context questions. The goal was to establish a baseline against which to evaluate the Harabagiu approach of partial reference resolution.

3 Results

The mean reciprocal ranks (MRR) obtained using QUANTUM for the original and reformulated version of each question set are summarized in Table 1.

Question Set	Original	Reformulated
TREC 10	0.082	0.130
Bell	0.248	0.258

Table 1: MRR obtained using QUANTUM

Although the reformulated versions of the questions appeared to obtain a higher MRR, the difference between the two sets is not statistically significant using either the t-test or the sign test with $\alpha = 0.1$.

4 Conclusions and Future Work

The Bell document collection is sufficiently small that the lack of significant difference between the MRR for the original and reformulated versions is not unexpected; intuitively it seems clear that the closed domain makes it comparatively easy to find the right document for a given question even when the question itself is not fully specified. However, the lack of significance for the TREC 10 questions is much more difficult to explain; of the 42 questions in this set, QUANTUM found answers for only 7 of the original questions, and only 10 of the reformulated questions. It is unclear why this difference is so small, but three possibilities arise:

1. The manual reformulations of the context questions were badly done, and provided too little information, or were in some way misleading.
2. Although QUANTUM was unable to answer the reformulated questions, perhaps

some other question answering system might have been more successful while still being significantly less successful on the original context questions.

3. Perhaps the added context really isn't necessary.

Intuitively the last alternative seems unlikely, whereas the second is more likely to be relevant. QUANTUM's MRR on the general TREC 10 question set was 0.223 (Plamondon and Kosseim, 2002), which is nearly double its MRR on the reformulated TREC 10 context questions but still not particularly high. This suggests that even after reformulation the TREC 10 context questions are more difficult to answer than the general TREC 10 questions, but it also suggests that a better-performing question answering system might show a greater difference between the original and reformulated contextual questions.

In any case further research will be required to determine what really happened; one approach will be to study the performance of the Okapi information retrieval system (MacFarlane, 2001), which QUANTUM uses as part of its algorithm. The goal would be to determine whether QUANTUM is missing answers that are actually present in passages retrieved by Okapi, or whether the problem lies in the information retrieval stage itself.

References

- Sanda Harabagiu, Dan Moldovan, Marius Pasca, Mihai Surdeanu, Rada Mihalcea, Roxana Girju, Vasile Rus, Finley Lactusu, Paul Morarescu, and Razvan Bunescu. 2001. Answering Complex, List and Context Questions with LCC's Question-Answering Server. In *Proceedings of the Tenth Text REtrieval Conference (TREC 10)*.
- A. MacFarlane. 2001. Online Documentation for Okapi-Pack
<http://www soi.city.ac.uk/~andym/OKAPI-PACK/>.
- Luc Plamondon and Leila Kosseim. 2002. QUANTUM: A Function-Based Question Answering System. In *Proceedings of the Fifteenth Canadian Conference on Artificial Intelligence (AI'2002)*, Calgary, Canada.
- Luc Plamondon, Guy Lapalme, and Leila Kosseim. 2001. The QUANTUM Question Answering System. In *Proceedings of the Tenth Text REtrieval Conference (TREC 10)*.
- Luc Plamondon, Guy Lapalme, and Leila Kosseim. 2002. The QUANTUM Question Answering System at TREC 11. In *Proceedings of the Eleventh Text REtrieval Conference (TREC 11)*.
- Ellen M. Voorhees. 1999a. Overview of the Eighth Text REtrieval Conference (TREC-8). In *Proceedings of the Eighth Text REtrieval Conference (TREC 8)*.
- Ellen M. Voorhees. 1999b. The TREC-8 Question Answering Track Report. In *Proceedings of the Eighth Text REtrieval Conference (TREC 8)*.
- Ellen M. Voorhees. 2001a. Overview of the TREC 2001 Question Answering Track. In *Proceedings of the Tenth Text REtrieval Conference (TREC 10)*.
- Ellen M. Voorhees. 2001b. Overview of TREC 2001. In *Proceedings of the Tenth Text REtrieval Conference (TREC 10)*.