The Task First, Please

Valentin Jijkoun
ISLA, University of Amsterdam
Kruislaan 403, 1098 SJ Amsterdam
jijkoun@science.uva.nl

Maarten de Rijke
ISLA, University of Amsterdam
Kruislaan 403, 1098 SJ Amsterdam
mdr@science.uva.nl

ABSTRACT

We examine the current state of evaluation exercises for automatic Question Answering (QA) systems, specifically targeting the QA task (QA@CLEF) as it is being evaluating with the setting of the Cross-Language Evaluation Forum (CLEF). We describe several key issues for the evaluation of QA systems and show how they are problematic in the current setup of the tasks at QA@CLEF. We argue that many of the problems are caused by the lack of a clear understanding of the QA task that should include potential users, types of information needs, types of available information resources. Finally, we propose several scenarios for QA and focused retrieval tasks that address these problematic issues. Our main conclusion is simple but important: a clear task definition is paramount for a meaningful evaluation of automatic systems, as evidenced by the overview of the QA evaluation setups.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.4 [**Information Systems Applications**]: H.4.2 Types of Systems; H.4.m Miscellaneous

General Terms

Algorithms, Measurement, Performance, Experimentation

Keywords

Focused retrieval, Question answering, Evaluation

1. INTRODUCTION

Question answering (QA) provides an important example of focused retrieval. In response to a user's question, QA systems are supposed to return an answer instead of a ranked list of documents from which the user has to extract the answer herself. Situated at the interface between computational linguistics and information retrieval, the task has attracted a great deal of attention over the past few years.

The launch of a dedicated question answering track at TREC 1999 has proved to be an especially important stimulus to research in the area. Most of the questions considered at the TREC QA track (and its descendants, the NTCIR and CLEF QA tracks) are fact-based, whose answers are typically named entities such as people, organizations, locations, dates (*"Who killed Lee Harvey Oswald? When was Mozart born?"*). Variations that have also been receiving attention include "list" questions such as *"Name all airports north of the polar circle*", definition questions such as "What is a sequencer?", as well as more complex questions that ask for "information nuggets" to be gathered from multiple documents.

Since focused retrieval and QA tasks are relatively new for the Information Retrieval community, there is an ongoing discussion about the nature of the tasks and approariate evaluation environments. Following up on analysis on the QA task in the literature [12], in this paper we identify a number of issues with the task scenarios being used today at one of the evaluation platforms for QA: QA@CLEF. We argue

(1) that a task model is important for informing the key ingredients of a retrieval task, including QA;

(2) that the QA task definitions used at CLEF leave a number of things to be desired, as a result of which key notions such as "answerhood" and "exactness" are seriously underspecified.

To remedy these shortcomings we argue that explicit *task definitions* should come first and that key ingredients (such as the definition of answerhood, the metrics to be used, etc) should be provided as part of the task definition. We propose a few QA task models that come with natural definitions of answerhood and metrics. One of these tasks (WiQA) was run as a pilot at CLEF 2006.

The aim of this note is ask questions and to stimulate discussion about current QA evaluation practices. In our analysis of current QA evaluation practices we use QA@CLEF as a main vehicle for discussion—this should not be interpreted as "CLEF-bashing." On the contrary, QA@CLEF has proved to be tremendously useful as a platform for fostering QA research in Europe, especially in languages other than English. Our comments and suggestions should be interpreted as suggestions for making the task even more valuable.

The remainder of this paper is organized as follows. In Section 2 we review the assumptions underlying the QA@CLEF track, relating it to the TREC QA on which it builds. Then, in Section 3 we indentify some of the key issues problematic for current QA evaluation exercises at CLEF. We proposed a few alternative scenarios in Section 4 and conclude in Section 5.

SIGIR 2007, Workshop on Focused Retrieval July 27, 2007, Amsterdam, The Netherlands

2. QUESTION ANSWERING AT CLEF

At TREC, the QA scenario that is being used as a model that informs decisions about what constitutes a correct answer and about what suitable metrics are, is that of an information analyst. Actually, very little of the analyst's rich context is included in the scenario used at TREC—no background knowledge, no factbooks or definition of an overarching task is included the task definition at the TREC QA track.

In 2003, a QA track was launched at CLEF, the Cross-Language Evaluation Forum [8]. Traditionally, retrieval tasks that are being evaluated at CLEF have involved multiple languages, either in the form of multiple monolingual tasks, bilingual tasks, or crosslingual tasks. The QA@CLEF track was no different: initially, it "copied" its task definition from the TREC QA track into three monolingual tasks (Dutch, Italian, Spanish), without, however, taking over TREC's assumption of the analyst's user scenario. In later years several languages were added, as were a cross-language variation of the task (with questions in one language and answer bearing documents in another). Originally, the conditions copied from the TREC QA track were more or less those from its 2001 edition: answers were 50 bytes long or exact, and participants could return up to three answers per question. The corpus used consisted of 1994/1995 newspapers. The questions were factoids only, and they were back-generated from the corpus.

In 2004, nine source languages and seven target languages were considered at QA@CLEF [9]. In addition to factoids, about 10% of the questions were definition questions, and another 10% did not have any answer in the corpora. To reduce the asssessment effort, the systems' output was reduced to a single, exact answer-string per question.

In 2005 the number of languages considered grew again (yielding 8 monolingual and 73 cross-lingual tasks) [13]. There was little or no innovation in the main task being assessed. So-called *temporally restricted* questions were added, which contain either an event that constrains the answer (e.g., *Who was Uganda's president during Rwanda's war?*), or a date (e.g., *Which Formula 1 team won the Hungarian Grand Prix in 2004?*), or a period of time (e.g., *Who was the president of the European Commission from 1985 to 1995?*). As in the previous year, a single exact answer per question was required.

In 2006, a number of changes were implemented for the main task at QA@CLEF [10]. For a start, list questions were included for the first time, and systems had to return short snippets containing answers to the test questions; the snippets were required to be short but sufficiently informative so as to allow the assessors to determine the correctness of the answer (without additional means). In addition, three pilots were run: the Answer Validation Exercise (AVE), the Real-Time QA Exercise (RTE), and Question Answering Using Wikipedia (WiQA). For AVE, systems were given triples of the form (Question, Answer, Supporting Text) and were asked to decide whether the Answer to the Question is correct and supported or not according to the given Supporting Text. At RTE, QA systems had to answer as many questions as possible in as little time as possible. WiQA was based on a scenario of a user exploring Wikipedia, and wanting to harvest additional bits of important information relevant to a Wikipedia article she is currently reading [3].

At the time of writing, the 2007 edition of the QA@CLEF task is in progresss. The main task has changed somewhat. It now uses a heterogeneous corpus, consisting of newspapers and Wikipedia—from non-overlapping periods of time: for some languages the newspaper collection dates back to 1994/1995, while the Wikipedia dump used is from late 2006. The Real-Time Evaluation pilot will run for a second time, and the WiQA pilot has merged with the WebCLEF web retrieval task at CLEF. A new pilot is being run which aims to assess the performance of QA systems when working on speech transcripts.

The number of participating teams at QA@CLEF has risen from 8 in 2003 to 37 in 2006—that's a tremendous success, but it also underlines the importance of a solid and well-defined task definition.

3. WHAT'S WRONG?

Increasingly, the QA@CLEF track has moved away from the TREC QA information analyst scenario. In principle, this is a laudable development, as we, as a community, will not be pushing the state of the art in QA in case we are merely repeating the same task at different venues and in different languages. However, this move has not been a move toward an alternative task scenario—no explicit task scenario has now been adopted at QA@CLEF, leaving many key dimensions of the task underspecified. Below, we review some of these dimensions and identify ones where, in our view, the current practice is not sufficiently explicitly specified as well as ones where clear choices have been made.

3.1 Exact Answers

Despite the drive of many researchers (and the TREC track) to focus on *exact* answers, users might not actually like or want simply exact answers: Lin et al. [7] show that users generally prefer answers *embedded in context*. The QA@CLEF task has already made an important step in this direction: in the 2006 setup systems were required to provide short document *passages* that justify the "answerhood" of the returned answers. The maximum length of supporting snippets, though, was set somewhat arbitrarily to 700 characters.

Another important issue is the "*exactness*" of answers. Assessors are typically asked to check whether answers are exact, both syntactically (i.e., they do not contain any "noise") and semantically. As was noted by participants in the TREC QA task,[1] this decision highly depends on assessors' background and expectations, and on the context in which a question arises. E.g., for the TREC question *Q160.7 "Where is the IMF headquartered?"*, the answer *"Washington"* was judged as exact, but for the question *Q152.1 "Where was Mozart born?"* the answer *"Salzburg"* was judged as inexact because assessors had the answer *Salzburg, Austria* in mind.

3.2 How Many Answers?

Whereas in the 2003 edition of the QA@CLEF tasks systems could return up to three answers for one questions, in the latest evaluation campaigns (both CLEF and TREC) a single answer is required. The decision to allow only answer might be a compromise between the amount of manual assessment of the submitted runs and the potential usability of QA systems. Since in the "information analyst" setting for document retrieval systems (at TREC and CLEF), as many as 1000 document are typically examined for relevancy, QA's focus on the top-1 answer in the simiar setting is hard to justify. At the same time, in the context of, e.g., Mobile QA [16] or real-time quizzes such a restriction would seem natural and even essential.

[1]Discussion on the TREC QA mailing list on October 6, 2006 started by Mark Greenwood.

The TREC complex interactive QA (ciQA) task [6] partially addresses this issue by allowing assessors to *interact* with a QA system for 15 minutes for one topic.

List questions, such as *"Name all airports in London"* present a particular challenge for the evaluation. E.g., QA@CLEF task switched from precision/recall-based evaluation for list questions to a "one complete answer" evaluation, and distinguishes closed list questions (e.g., *"What are the names of all of Bach's children?"*) and open list questions (e.g., *Name several most famous Bach's works.*, with the idea that they require different evaluation measures.

3.3 NIL Questions

What should a system do if it is not capable to locate an answer in a given document collection? Should a system back off and explicitly indicate with a *NIL* response, or try to use other available resources (e.g., Web, encyclopedias, newspapers) to find answers? Would a real-world user be interested in knowing that the system cannot find an answer, or would she prefer to at least receive "the best guess" so as to get started [12]?

3.4 Types of Questions

What types of questions should a QA system deal with? What questions should a QA evaluation exercise deal with? A small study of questions extracted from search engine logs [4] indicates that most users ask *procedural* questions (38%) such as *"How to cook a ham?"*, but factoids (e.g., *"What did caribs eat?"*) also constitute a substantial portion of questions (10%). Other common question types include *description* (13%, e.g., *"Who is victoria gott?"*) and *explanation* (10%, e.g., *"Why people do good deeds?"*).

Although QA evaluation exercises traditionally focus on factoid questions, with TREC's "OTHER" questions, CLEF's "definition" question and NTCIR's "why" questions [2] the attention is moving beyond factoids. Still, the reasons why specific types of questions are included in evaluations in specific proportions are not motivated by the requirements of potential users of QA systems.

3.5 Question Generation

Generating questions for a QA evaluation exercise is a laborious process. In its first year, the TREC QA track used questions back-generated from a corpus of newspaper/newswire documents, which made the questions somewhat unnatural and the task somewhat easier since the target document contained most of the question words [14]. In later years, questions at TREC's QA track were created by assessors, informed by query logs and based on their own interests. In contrast, for lack of a clear scenario, QA@CLEF has only dealt with question back-generated from the corpus of newspaper documents used. We believe that this is problematic (for the reasons described above)

3.6 Matching Needs and Sources

Librarians are good at selecting appropriate sources for addressing a specific user's information need. For questions like *"When was Mozart born?"* or *"What is a sequencer?"* they would probably consult an encyclopedia, while for a question like *"Which countries did Bush visit in 2005?"* newspapers seem a more appropriate source. A QA system intended for real world use should also match different available information resources to user's information needs. Why would we want to find Mozart's date of birth in a newspaper collection (at WebCLEF 2007) or Marlon Brando's age in a blog (as in the 2007 of the TREC QA track), if more natural and even more reliable sources are available?

The QA evaluation exercises are moving in the direction of diversifying data sources: Wikipedia is used at CLEF QA, a blog corpus is used at TREC (although the TREC questions are still mostly factoid), Google's view of the Web is used at WebCLEF 2007. Still, there is a long way to working with types of questions that match the types of collections used in the collection-based QA.

3.7 Multilinguality

At CLEF and NTCIR, multilinguality has been one of the key starting points. However, for the cross-lingual tasks (i.e., questions in language X are supposed to be answered using a document collection in language Y, a so called "X to Y" task), the evaluation questions are typically constructed by translating questions of a monolingual QA task into a different language (e.g., translating questions from Y into X, and thus creating an evaluation set for the "X to Y" QA task). This simplifies evaluation, but unfortunately creates many questions that are highly unnatural regarding the information sources. Why would a Dutch-speaking user be interested in answering the questions *"Who was Flaubert?"* from a collection of Spanish newspaper articles?

4. NEW SCENARIOS

Given the many dimensions outlined above, how should we go about evaluating QA? We see two possible options here:

1. Evaluating QA as a user-driven information access task: we first define who our users are. This will imply determining what kind of information needs they have, what resources they allow, and what constitutes proper result presentation(s), and evaluation measures.
2. Answering questions as a means for evaluating certain NLP tools or techniques: "I have a parser/tagger/analyzer/... and I want questions for which I can use the parser/tagger/analyzer/... to demonstrate its usability." Usually, this strategy leads to a clear but narrow definition of QA, not driven by information needs but by expected applicability of a specific tool or technique. E.g., the IR step can be dropped, questions can be pre-categorized, e.g., as "targeting synonymy and paraphrasing," "requiring basic world knowledge"—creators of different NLP tools may be interested in different categories of questions.

We believe that much confusion results from mixing options 1 and 2, and this is what has happened at QA@CLEF. The result is that many things are dealt with in a very ad hoc way: types of questions, evaluation measures, result presentation, choice of collection, etc.

If we are right, and the lack of an explicit task scenario at QA@CLEF is problematic, how should we move forward? Below we list a number of possibilities of task scenarios that we believe address the isssues identified in the previous section *and* that are worth pursuing.

Before we list our suggestions, we specify what we believe are natural criteria on scenarios to be considered for retrieval experiments at CLEF:

- The task should correspond as close as possible some real-world information need with a clear definition of a user;
- Multi- and cross-linguality should be natural (or even essential) for the task;

- The collection(s) used in the task should be the source of choice for the user's information need;
- Test questions should be generated by people having a genuine interest in the topic at hand;
- Collections, topics and assessors' judgements, resulting from the task should be re-usable in future; and finally,
- The task should be challenging for the state-of-the-art technology.

Against this background, then, we list a number of alternative task scenarios that we believe would make a meaningful QA evaluation effort.

4.1 Intelligence gathering

In analytical question answering [11], the users are information analysts and questions are not factoids for which answers come in a fairly limited number of "formats," but they are exploratory in nature, seeking to find out what is generally available on the topic of the question. E.g., *"What has been Russia's reaction to the U.S. bombing of Kosovo?* Here, appropriate responses can be taken to be frames, consisting of bags of atributes associated with a (news) event. Newspapers form a natural corpora to use in this scenario. The TREC QA task and especially the TREC ciQA task target at this type of scenarios: questions are assumed to follow one of the pre-defined templates (reflecting recurring interests of analysts) and assessors (users) may interact with the QA system within a specified time interval. The final decision about the correctness and the number of answers found with the help of the system is up to the assessor.

4.2 Event-targeted QA

In a different scenario, a user (e.g., a journalist or a history student) needs to collect background information around a specific event: e.g., persons involved, their occupations at the time of the event or later, ages, relations, places (populations, exact locations, distances), other details mentioned in connection with the event, possibly other related events, or even different perspectives on the event, etc.—whatever she might find important. The scenario is that the user starts with an article mentioning or describing event and has further questions about it, "stemming" from this initial information and her own knowledge.

In this setting the use of heterogeneous collections (newswire, blogs, encyclopedias, etc.), is much justified: more general questions ("*Where exactly in Iraq is Basra?*") are naturally answered using an encyclopedia, but for more specific questions (e.g., *"Which countries did Hussein visit in 1991?*") newspaper texts are a good (and maybe the most appropriate) source. Possible question types would include temporally and geographically restricted questions, as well as definition, relationship, list questions, and questions about subjective aspects and opinion questions (for which, e.g., blogs would be a natural source). Questions can be of any type in this scenario, and a ranked list of answers would seem most appropriate here, while a limited form of multilinguality seems natural, especially when the event at handtook place acros the border, or if the user is interested in the international perspective on the event being considered.

4.3 Trivia game show

Trivia are a source of entertainment for many, as is witnessed by game shows, trivia board games, as well as a large number online resources, where users both ask and answer such questions.[2] It is usually clear what the answer to a trivia question is, which makes the evaluation of trivia-based QA easier. The unique correct answer is known in advance, as defined by the game organizers. Answers to questions are always short strings (entities, actions, events). No specific information source is enforced, which means that a system may use any sources available (encyclopedias, the web, thematic corpora, etc.).

While a QA system that is intended for answering specifically trivia questions is not necessarily a useful real-world application (other than for entertainment purposes), if provides a clear definition of the task and straightforward evaluation measures, that take into account both answer correctness and the time spent by the system. Using such scenario would be a natural option for, e.g., the Real-Time QA Exercise (RTE) held at CLEF [1]. Multilinguality does not seem appropriate here, while the requirement that the answers are sufficiently exact seems reasonably natural.

4.4 WiQA 2006 and WebCLEF 2007

The CLEF 2006 WiQA task [3, 5] and CLEF 2007 WebCLEF task [15] take, as the scenario, a user collecting important information on a specific topic. E.g., the user might be writing an essay or updating a encyclopedia article on the topic, and is gathering "important" information nuggets that are worth including in her report. An automatic system is supposed to help the user to locate new important bits of information in Wikipedia (for the WiQA task) or on the Web (for the WebCLEF task). While not instantiating a traditional QA scenario—it really only asks a single question about the topic at hand: what should I know about it?—, these two tasks provide two different frameworks for evaluating focused retrieval systems, in which, moreover, multilinguality comes natural, as important information may be expressed in a language other than the language of the topic statement. Finally, the task suggets natural document sources—Wikipedia and/or the web.

5. CONCLUSIONS

We examined the current state of evaluation exercises for automatic Question Answering systems, specifically targeting the QA@CLEF task. We have described several key issues for the evaluation of QA systems and showed that they are problematic in the current setup of the tasks.

We argue that many of these problems are caused by the lack of a clear understanding of the QA task that should include potential users, types of information needs, types of available information resources. The lack of clarity on these dimensions makes it difficult to justify the setup and evaluation decisions for a QA task. Finally, we proposed several scenarios for QA and focused retrieval tasks that address the problematic issues.

Our main conclusion is simple but important: a clear task definition is paramount for meaningful evaluation of automatic systems, as evidenced by our overview of the QA evaluation setups.

6. ACKNOWLEDGEMENTS

We are grateful to Diana Santos and other members of the organizing group responsible for QA@CLEF. This research was supported by the Netherlands Organisation for Scientific Research (NWO) under project number 220-80-001.

[2] E.g., `http://www.funtrivia.com`

7. REFERENCES

[1] C. L. A. Clarke, G. V. Cormack, and T. R. Lynam. Exploiting redundancy in question answering. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 358–365, New York, NY, USA, 2001. ACM Press.

[2] J. Fukumoto, T. Kato, F. Masui, and T. Mori. An overview of the 4th question answering challenge (QAC-4). In *Proceedings of the NTCIR Workshop 6*, 2006.

[3] V. Jijkoun and M. de Rijke. Overview of the WiQA task at CLEF 2006. In *Proceedings CLEF 2006*, to appear.

[4] V. Jijkoun and M. de Rijke. Retrieving answers from frequently asked questions pages on the web. In *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 76–83, New York, NY, USA, 2005. ACM Press.

[5] V. Jijkoun and M. de Rijke. WiQA: Evaluating Multi-lingual Focused Access to Wikipedia. In T. Sakai, M. Sanderson, and D. Evans, editors, *Proceedings EVIA 2007*, pages 54–61, 2007.

[6] D. Kelly and J. Lin. Overview of the TREC 2006 ciQA Task. *SIGIR Forum*, 41(1):107–116, June 2007.

[7] J. Lin, D. Quan, V. Sinha, K. Bakshi, D. Huynh, B. Katz, and D. Karger. What makes a good answer? The role of context in question answering. In *Proceedings of INTERACT 2003*, 2003.

[8] B. Magnini, S. Romagnoli, A. Vallin, J. Herrera, A. Penas, V. Peinado, F. Verdejo, and M. de Rijke. The multiple language question answering track at CLEF 2003. In C. Peters, J. Gonzalo, M. Braschler, and M. Kluck, editors, *Comparative Evaluation of Multilingual Information Access Systems, CLEF 2003*, volume 3237 of *Lecture Notes in Computer Science*, pages 471–486. Springer, 2004.

[9] B. Magnini, A. Vallin, C. Ayache, G. Erbach, A. Penas, M. de Rijke, P. Rocha, K. Simov, and R. Sutcliffe. Overview of the CLEF 2004 multilingual question answering track. In C. Peters, P. Clough, G. Jones, J. Gonzalo, M. Kluck, and B. Magnini, editors, *Multilingual Information Access for Text, Speech and Images: Results of the Fifth CLEF Evaluation Campaign*, volume 3491 of *Lecture Notes in Computer Science*, pages 371–391, 2005.

[10] B. Magnini, D. Giampiccolo, P. Forner, C. Ayache, P. Osenova, A. Penas, V. Jijkoun, B. Sacaleanu, P. Rocha, and R. Sutcliffe. Overview of the CLEF 2006 Multilingual Question Answering Track. In *Proceedings CLEF 2006*, to appear.

[11] S. Small, T. Strzalkowski, T. Liu, S. Ryan, R. Salkin, N. Shimizu, P. Kantor, D. Kelly, R. Rittman, and N. Wacholder. HITIQA: towards analytical question answering. In *Proc. of COLING 2004*, 2004.

[12] K. Sparck Jones. Is question answering a rational task? In *Proceedings 2nd CoLogNET-ELSNET Symposium on Questions and Answers: Theoretical and Applied Perspectives*, 20003.

[13] A. Vallin, B. Magnini, D. Giampiccolo, L. Aunimo, C. Ayache, P. Osenova, A. Penas, M. de Rijke, B. Sacaleanu, D. Santos, and R. Sutcliffe. Overview of the CLEF 2005 Multilingual Question Answering Track. In C. Peters, F. Gey, J. Gonzalo, H. Müller, G. Jones, M. Kluck, B. Magnini, and M. D. Rijke, editors, *Accessing Multilingual Information Repositories*, volume 4022 of *Lecture Notes in Computer Science*, pages 307–331. Springer, September 2006.

[14] E. Voorhees. Overview of the TREC-9 Question Answering Track. In *The Ninth Text REtrieval Conference (TREC 9)*, 2001.

[15] WebCLEF. Definition of the clef 2007 webclef task, 2007. `http://ilps.science.uva.nl/WebCLEF/WebCLEF2007`.

[16] E. Whittaker, J. Mrozinski, and S. Furui. Factoid question answering with web, mobile and speech interfaces. In *HLT-NAACL*, pages 288–291, 2006.