Bringing Why-QA to Web Search

Suzan Verberne, Lou Boves, Wessel Kraaij

Centre for Language Studies / Institute for Computing and Information Sciences
Radboud University Nijmegen
s.verberne@let.ru.nl

Abstract. We investigated to what extent users could be satisfied by a web search engine for answering causal questions. We used an assessment environment in which a web search interface was simulated. For 1 401 *why*-queries from a search engine log we pre-retrieved the first 10 results using Bing. 311 queries were assessed by human judges. We found that even without clicking a result, 25.2% of the *why*-questions is answered on the first result page. If we count an intended click on a result as a vote for relevance, then 74.4% of the *why*-questions gets at least one relevant answer in the top-10. 10% of *why*-queries asked to web search engines are not answerable according to human assessors.

1 Introduction and background

The problem of automatically answering open-domain questions by pinpointing exact answers in a large text (web) corpus has been studied since the mid 1990s. Already in 2001, Kwok et al. [1] argued that if developers of open-domain Question Answering (QA) systems want their system to be useful for a large audience, QA should be web-based and integrated in existing search interfaces.

While approaches to QA for factoid, list and definition questions that might scale up to the web have been implemented and evaluated in the TREC (and CLEF) evaluation forums, QA for more complex questions such as *why*- and *how*-questions is considered a complex NLP task that requires advanced linguistic processing [2]. Accordingly, most research in *why*-QA (also called causal QA) is directed at extracting causal relations from text [3–6]. Some papers describe IR-based approaches to *why*-QA [7–9], but these use a restricted corpus of candidate answer documents (either newspaper texts or a static version of Wikipedia). Until now, no attempts have been made to investigate causal QA using web data and web search engines.

There is a huge amount of user-generated QA data available on the web [10], also for *why*-questions [11]. We think that *why*-QA, just as any other open-domain retrieval task, should be studied in the context and scale of the web. Ideally, a search engine should be able to recognize a query as a causal question, and provide the answer directly. The implementation of such a service is only possible if the answers to *why*-questions can be found by web search engines.

In the current paper, we investigate to what extent users could be satisfied by a web search engine for answering *why*-questions. We approximate user satisfaction using an assessment environment in which a web search interface is simulated. Participants in the experiment are presented with *why*-queries from a search engine log and a ranked list of results that have been pre-retrieved with the Bing search engine[1]. The judges are asked to assess the relevance of the first ten results for each query.

Using this assessment environment we address the question "What proportion of *why*-questions can be answered satisfactorily by a state-of-the-art web search engine?" Other questions that we will address are: "To what extent do judges agree in their assessment of answers to causal questions?", and "What proportion of *why*-queries that are posed to search engines are not answerable according to human assessors?".

2 Web data for evaluating why-QA

We obtained user-generated *why*-questions from the Microsoft RFP data set[2]. This collection consists of approximately 14 million queries from US users entered into the Microsoft Live search engine in the spring of 2006. In this data set, 2 879 queries (1 401 unique) start with the word 'why'. There are a number of possible paraphrases of the question word 'why', but their frequency in the data is very low: 'for what reason' does not occur at all and 'how come' occurs in 11 queries. Therefore, we decided to only extract queries that start with the word *why*, assuming that they are representative for all causal queries. Since the majority of queries in our set are questions, we will use the words query and question interchangeably in the remainder of this paper.

Not all queries are syntactically complete sentences (e.g. "why so many religions") and some are not even questions: "why you wanna lyrics". We kept ungrammatical questions and queries with spelling errors in the set in order to reflect the noisiness of user queries. Moreover, we did not filter out questions with subjective answers ("why marriages fail") or queries that are probably no questions ("why do fools fall in love"); we left the decision whether a query is answerable or not to the assessors. We only filtered out the queries that contain the word 'lyrics'. The result is a set of 1 382 unique *why*-queries.

Using the Perl module LWP::Simple[3], we sent all queries to the Bing search engine and extracted the 10 results from the first result page. For each of the results, we saved the title, URL and snippet. We refer to the combination of title, URL and snippet as an 'answer'.

3 Relevance assessment set-up

For the manual assessment of answers, we recruited subjects among colleagues and friends. We promised them a treat if they assessed at least twenty questions.

[1] www.bing.com

[2] This set was distributed for the WSCD 2009 workshop.

[3] The module is available at `http://www.cpan.org/`.

Table 1. General statistics of the collected data

Number of assessors	22
Number of questions assessed (including skipped)	311
Number of unique questions assessed (including skipped)	271
Number of unique questions skipped	33
Number of unique questions with at least one assessed answer	238
Number of answers with a judgment (other than 'don't know')	2 105

For the levels of answer relevance in our assessment task we adopt the distinction between "the passage answers the query" and the "passage does not answer the query" from [9]. In addition to this binary choice, we add the alternatives "don't know" and "I expect to find the answer if I would click the link. The latter option was added because the snippets returned by Bing are short and users of web search engines are used to click on results they expect them to be relevant. In evaluation studies using click data, a click on a document is often considered a vote for the document's relevance.[4]

The participants were allowed to skip a question, with or without providing a reason for skipping. Three different reasons are predefined: *There is no answer possible (the query is a joke or title)*', '*The answer is subjective/depends on the person*' and '*I don't understand the query*'. If a question is skipped, all the corresponding snippets in the result list receive the label "don't know". In the user interface of the experiment, we imitate a web search engine's formatting by using a larger font and blue for titles and green for the URL under the snippet. In the title and the snippet, query words are printed in bold face. Next to each answer are the radio buttons for the four assessment options.

Each time a participant logged into the assessment environment, a random query from the total set of *why*-queries is presented to him/her together with the first ten results that were returned by Bing. For the purpose of calculating agreement between the assessors, each assessor was presented at least one query that was already assessed by another assessor.

4 Results

Table 1 shows general statistics of the collected data. We collected answer assessments for 238 unique queries, 42 of which have been assessed by two assessors. This set may seem small but is big compared to sets of *why*-questions collected in previous work (see Section 5). When calculating the inter-judge agreement for the answers to these queries, we disregarded the answers that were labeled as "don't know". We measured strict agreement, where each of the three categories 'yes', 'click' and 'no' is considered independently, and lenient agreement, where 'yes' and 'click' are taken together as one category because they both

[4] Although it was shown by [12] that click behavior is biased by trust in the search engine and the quality of the ranking.

Table 2. Evaluation of Bing (Summer 2010) for *why*-queries on the basis of collected user assessments. In addition to success@10 and precision@10, we present MRR (based on the highest ranked answer per question) because it was used in previous research on *why*-QA.

	strict	lenient
Success@10	25.2%	74.4%
Precision@10	8.1%	34.1%
Mean Reciprocal Rank (MRR)	0.163	0.500

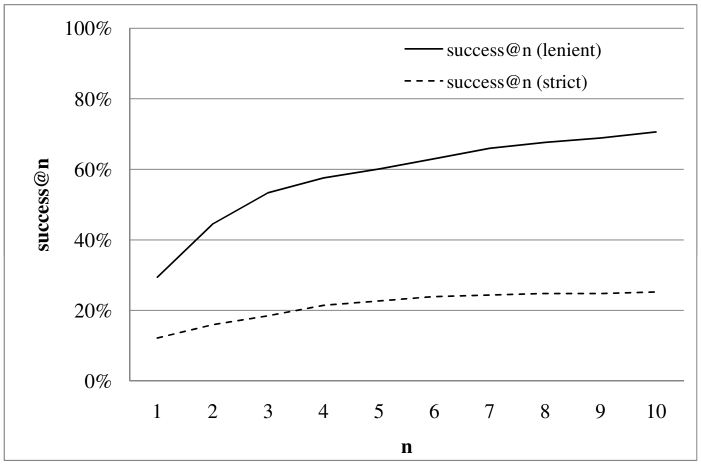

Fig. 1. Success@n as a function of n (the length of the result list), over all *why*-queries.

represent relevant answers. Measured strictly, we found a fair agreement (Cohen's $\kappa = 0.34$); measured leniently, there was a moderate agreement (Cohen's $\kappa = 0.47$).

In the evaluation of the retrieval results on the basis of the user assessments, we also distinguish between strict and lenient evaluation. In Table 2, we present the results in terms of Success@10 (the percentage of questions with at least one relevant answer) and Precision@10 (the percentage of answers that is judged relevant). Figure 1 shows Success@n as a function of n.

We investigated the influence of query frequency and query length on the query success. We hypothesized that web search engines are more successful for more frequent queries and for longer queries. For both predictors however, we found no correlation with average precision (Pearson's $\rho = 0.08$ for query frequency and $\rho = 0.06$ for query length; $N = 238$). Query length is approximately normally distributed with $\mu = 6.2$ and $\sigma = 2.6$ but since query frequency is relatively sparse (55% of the queries has a frequency of 1), these results may be falsified when using a larger click data set.

5 Comparison to other approaches

It is difficult to compare our current results to other work in *why*-QA because previous approaches either disregard the retrieval step of the QA task [3, 4] or address a different language than English [7]. The work described in [8, 9] is the most comparable to our work because it evaluates an approach to *why*-QA for English that is based on passage retrieval, using a set of real users' questions (questions asked to answers.com). The answer corpus used is a Wikipedia XML corpus and the evaluation set only contains *why*-questions for which the answer is present in this corpus (only 186 of the 700 Webclopedia *why*-questions). Moreover, grammatically incomplete questions and questions containing spelling errors were removed. Using TF-IDF only for ranking, MRR was 0.24 with a success@10 of 45% (precision@10 was not measured).

Since the web contains much more redundant information than Wikipedia proper, our evaluation shows a different pattern. Dependent on strict or lenient evaluation, success@10 is lower (25.2%) or much higher (74.4%) than the results in [9]. More strikingly, precision@10 is 34%, which means that one in three retrieved answers is relevant. Although we did not filter out questions for which no answer is available on the web (clearly, there is no good way to do this), both Success@10 and MRR (measured leniently) are much higher than the results reported in [9]. Apparently, the abundance of information on the web compensates heavily for the careful filtering of questions in previous work. This confirms previous findings in redundancy-based QA [13].

6 Discussion and Conclusion

We found that even without clicking a result, 25.2% of *why*-questions is answered by a state-of-the-art web search engine on the first result page. If we count an intended click on a result as a vote for relevance, then 74.4% of the *why*-questions gets at least one relevant answer in the top-10. The large difference between strict and lenient evaluation suggests that for *why*-queries, presenting longer snippets in the search engine output may increase user satisfaction.

We measured the difficulty of answer assessment for causal questions in terms of inter-judge agreement. Our κ-values suggest that if assessors are asked to judge other persons' causal questions, they quite often disagree on the relevance of the answers proposed by a search engine.

The assessors had the possibility of skipping questions, with or without ticking one of the reasons provided in the interface. 33 of the 311 *why*-queries were skipped. For most of these, the assessor provided a reason. The reason '*There is no answer possible (the query is a joke or title)*' was chosen for 5 queries such as "why did the chicken cross the road". The reason '*The answer is subjective/depends on the person*' was selected for 10 questions containing subjective judgements such as "why Disney Is Bad" and personal questions such as "why am I the best hockey coach for the position". The last reason '*I don't understand the query*' was selected 13 times, for complex queries such as "why circumscribing the role and behavior in the biblical community according to ephesian 6 text" but also for underspecified queries such as "why photography". Overall, we can conclude that 10% of *why*-queries asked to web search engines are not answerable according to human assessors.

In future work, we aim to investigate how a dedicated approach to *why*-QA can be implemented for web search, combining the output of Bing with knowledge from previous research in the computational linguistics community.

References

1. Kwok, C., Etzioni, O., Weld, D.: Scaling question answering to the Web. ACM Transactions on Information Systems (TOIS) **19**(3) (2001) 242–262
2. Maybury, M.: Toward a Question Answering Roadmap. In: New Directions in Question Answering. (2003) 8–11
3. Girju, R.: Automatic detection of causal relations for question answering. In: Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering-Volume 12, Association for Computational Linguistics (2003) 83
4. Pechsiri, C., Kawtrakul, A.: Mining Causality from Texts for Question Answering System. IEICE Transactions on Information and Systems **90**(10) (2007) 1523–1533
5. Verberne, S., Boves, L., Oostdijk, N., Coppen, P.: Discourse-based Answering of Why-Questions. Traitement Automatique des Langues (TAL), special issue on "Discours et document: traitements automatiques" **47**(2) (2007) 21–41
6. Vazquez-Reyes, S., Black, W.: Evaluating Causal Questions for Question Answering. In: Mexican International Conference on Computer Science, 2008. ENC'08. (2008) 132–142
7. Higashinaka, R., Isozaki, H.: Corpus-based Question Answering for Why-Questions. In: Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP). (2008) 418–425
8. Verberne, S., Boves, L., Oostdijk, N., Coppen, P.: What is not in the Bag of Words for Why-QA? Computational Linguistics **32**(2) (2010) 229–245
9. Verberne, S., Van Halteren, H., Theijssen, D., Raaijmakers, S., Boves, L.: Learning to Rank for Why-Question Answering. Information Retrieval (2010) Published online: DOI 10.1007/s10791-010-9136-6.
10. Adamic, L., Zhang, J., Bakshy, E., Ackerman, M.: Knowledge sharing and yahoo answers: everyone knows something. In: Proceeding of the 17th international conference on World Wide Web, ACM (2008) 665–674
11. Ignatova, K., Toprak, C., Bernhard, D., Gurevych, I.: Annotating Question Types in Social Q&A Sites. In: GSCL Symposium" Language Technology and eHumanities, Citeseer (2009)
12. Joachims, T., Granka, L., Pan, B., Hembrooke, H., Gay, G.: Accurately interpreting clickthrough data as implicit feedback. In: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, ACM (2005) 154–161
13. Dumais, S., Banko, M., Brill, E., Lin, J., Ng, A.: Web question answering: Is more always better? In: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, ACM (2002) 291–298