The Impact of Corpus Size on Question Answering Performance

C. L. A. Clarke G. V. Cormack M. Laszlo T. R. Lynam E. L. Terra

Department of Computer Science, University of Waterloo, Canada
mt@plg.uwaterloo.ca

ABSTRACT

Using our question answering system, questions from the TREC 2001 evaluation were executed over a series of Web data collections, with the sizes of the collections increasing from 25 gigabytes up to nearly a terabyte.

Categories and Subject Descriptors

H.3.4 [**Information Systems**]: Information Storage and Retrieval—*Systems and Software*

General Terms

Measurement, Performance, Experimentation

Keywords

Question answering

1. QUESTION ANSWERING

The goal of an open-domain question answering (QA) system is to provide concise answers to short, factual questions. Questions might range from those eliciting straightforward facts, such as names or dates, to those requiring brief explanations. Examples of the former include "What is the speed of light?" and "Who won the world series in 1993?". Examples of the latter include "Who is Harry Houdini?" and "Why is the sky blue?". Often, the answers are drawn from a target corpus consisting of documents created for other purposes, such newspaper articles and Web pages.

Since 1999 a QA track has been part of the annual TREC evaluation of information retrieval systems conducted by the U.S. National Institute of Standards and Technology (NIST) [8]. In TREC 2001, the main QA task used a set of 500 questions and a 3GB target corpus [7]. Each system was requested to return five ranked responses to each question, where each response could have a maximum length of 50 bytes. A response was judged correct only if an answer was contained in the 50 bytes.

The main evaluation measure for the task was *mean reciprocal rank* (MRR). To compute an MRR, each question is assigned a score that is the inverse of the rank of the first response that is judged correct. If no response contains an answer, the question is assigned a score of zero. The scores of the individual questions are then averaged to produce an MRR value for the question set as a whole.

Our QA system was one of the better-performing systems in both the 2000 and 2001 TREC QA experiments [3]. The system takes a simple statistical approach to QA [2]. A lightweight parser first performs question analysis and generates a query and a set of question-specific selection rules. The query is used by a passage retrieval component to locate short text fragments within the target corpus that are related to the topic of the question. The passages and the selection rules are then passed to an answer selection component that identifies 50-byte text snippets within the retrieved passages that are likely to include the answer.

A new feature of our system for TREC 2001 was the use of Web data for candidate reinforcement during answer selection [3]. In the experiment reported here, we continue this work, using Web data as the target corpus and examining the impact of target corpus size on system performance.

2. EXPERIMENT

A terabyte of HTML, crawled from the general Web in mid-2001, forms the basis for the experiment. Starting with a seed set of URLs representing the home pages of 2392 universities and other educational organizations, pages were gathered in breadth-first order with one exception. If a breadth-first ordering would place an excessive load on a single host, defined as more than 0.2% of total crawler activity over a time period of approximately one hour, URLs associated with that host were removed and requeued until the anticipated load dropped to an acceptable level. Pages at a given depth from the seed set were crawled in random order. During the crawl, duplicate pages were detected and eliminated, and do not form part of the final collection. A breadth-first ordering is known to produce high-quality pages [6], and we expected the crawl to contain the answer to many simple factual questions.

To generate responses, we used an archived version of our QA system as it existed at the time our TREC 2001 experimental runs were generated. Thus, the system has not been trained or influenced by the TREC 2001 questions. For the experiments reported here, the Web reinforcement component was disabled.

The TREC 2001 questions were executed over a range of target corpus sizes, representing from 3% to 100% of the available Web collection. During the experiment a small portion of the full terabyte collection was off-line, and the experiment was run over an actual Web collection consisting of 960GB of HTML.

Copyright is held by the author/owner.
SIGIR'02, August 11-15, 2002, Tampere, Finland.
ACM 1-58113-561-0/02/0008.

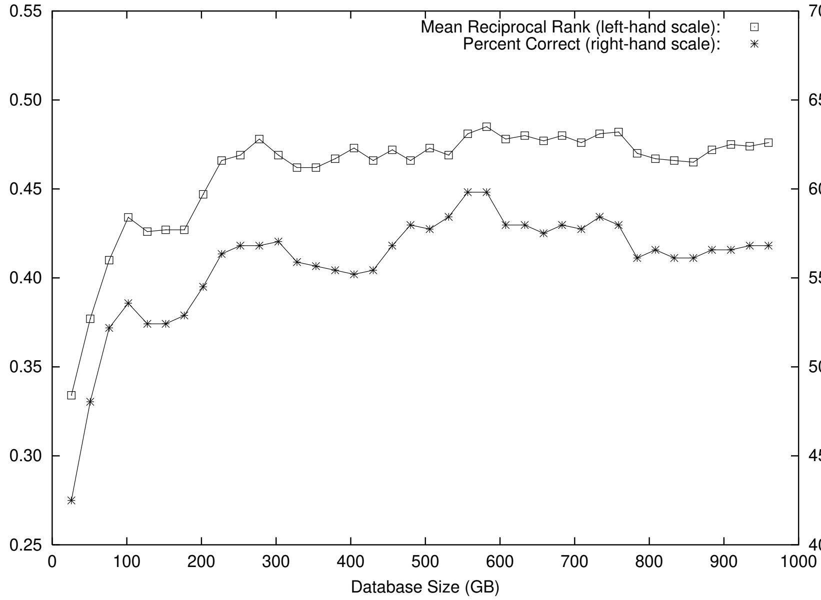

Figure 1: The Impact of Corpus Size on Question Answering Performance

Responses are judged using an automatic evaluation script created by Ken Litkowski of CL Research, which is similar to earlier scripts created by Ellen Voorhees for the 1999 and 2000 TREC questions [8]. The script executes a series of question-specific regular expressions over the responses returned for each question. Whenever a match occurs, the response is marked as correct. The script contains patterns for only 433 of the original 500 TREC 2001 questions. Most of the remaining questions either did not have an answer in the TREC 2001 target corpus or were discarded by NIST due to typos and similar errors. The 67 missing questions are not considered in our calculation of MRR.

The results are presented in Figure 1. The figure reports both the MRR and the percentage of questions with a correct answer among the top five responses. For comparison, the same system and questions executed over the official 3GB TREC 2001 QA corpus and evaluated with the automatic script gave a MRR of 0.408.

3. DISCUSSION

We expected performance to improve with corpus size. While performance does improve up to 400-500GB, it then appears to reach an asymptote and actually declines slightly after that. Experience with traditional IR systems indicates that system performance is directly related to corpus size [4] and we have observed a similar relationship in QA systems with corpora in the range 10-100GB [2]. Further work is needed to determine if the observed asymtote represents a real effect, or if it is an artifact of our system or of the evaluation methodology.

An examination of the answers returned by the system suggests a possible weakness in the evaluation methodology. Generally, automatic evaluation by a script is not as accurate as manual evaluation by a human judge. Automatic scripts err both by marking responses as correct when the surrounding context does not support the answer and by missing correct answers that do not match the expected syntactic form. In this case, the script may be "overfitted" to the syntactic forms that appear in the TREC corpus. Certainly many correct answers are missed.

Other groups have used Web data to support open-domain question answering, generally by depending on a commerical search engine to supply the target corpus and information retrieval facilities [1, 5]. While this approach can produce practical systems, it does not allow the information retrieval strategies for question answering to be studied in detail.

4. REFERENCES

[1] E. Brill, J. Lin, M. Banko, S. Dumais, and A. Ng. Data-intensive question answering. In *2001 Text REtrieval Conference*, 2001.

[2] C. L. A. Clarke, G. V. Cormack, and T. R. Lynam. Exploiting redundancy in question answering. In *24th ACM SIGIR Conference*, pages 358–365, 2001.

[3] C. L. A. Clarke, G. V. Cormack, T. R. Lynam, C. M. Li, and G. L. McLearn. Web reinforced question answering. In *2001 Text REtrieval Conference*, 2001.

[4] G. V. Cormack, O. Lhotak, and C. R. Palmer. Estimating precision by random sampling. In *22rd ACM SIGIR Conference*, pages 273–274, 1999.

[5] C. C. T. Kwok, O. Etzioni, and D. S. Weld. Scaling question answering to the Web. In *Tenth International World Wide Web Conference*, pages 150–161, 2001.

[6] M. Najork and J. L Wiener. Breadth-first search crawling yields high-quality pages. In *10th International World Wide Web Conference*, pages 114–118, 2001.

[7] E. M. Voorhees. Overview of the TREC 2001 Question Answering Track. In *2001 Text REtrieval Conference*, 2001.

[8] E. M. Voorhees and D. Tice. Building a question answering test collection. In *23rd ACM SIGIR Conference*, pages 200–207, 2000.