The Impact of Corpus Size on Question Answering Performance

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

University of Waterloo

Overview

We evaluated our question answering system over 1323 TREC QA track questions[1] using a series of increasingly larger Web data collections. The sizes of the collections varied from 25 gigabytes up to nearly a terabyte. Our goal was to determine the effect of corpus size on question answering performance.

Experiment

A terabyte of HTML, crawled from the general Web, forms the basis for the experiment.

An archived version of our TREC 2001 QA system was used to generate the responses.

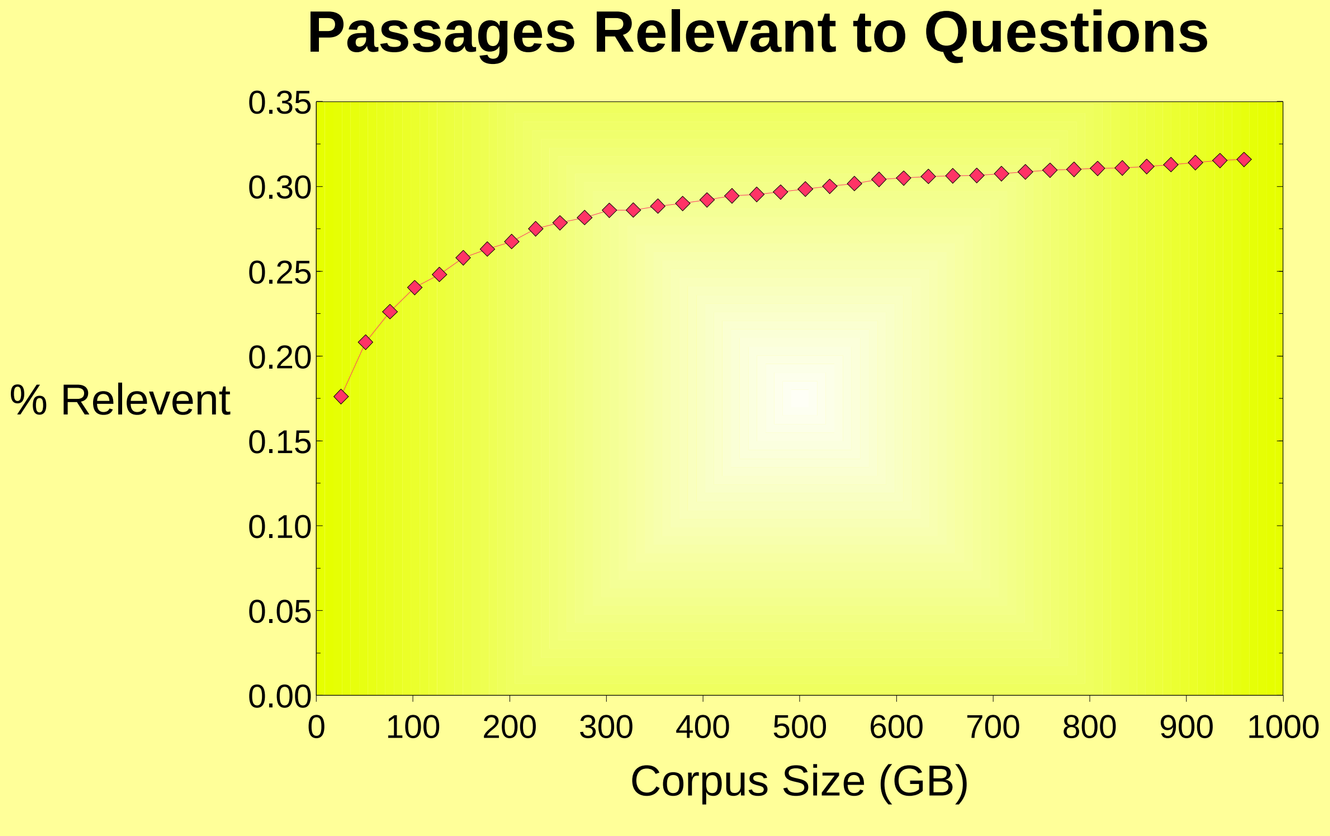
For each question, the top 40 passages were retrieved from each collection, and five 50−byte response strings were selected from these passages.

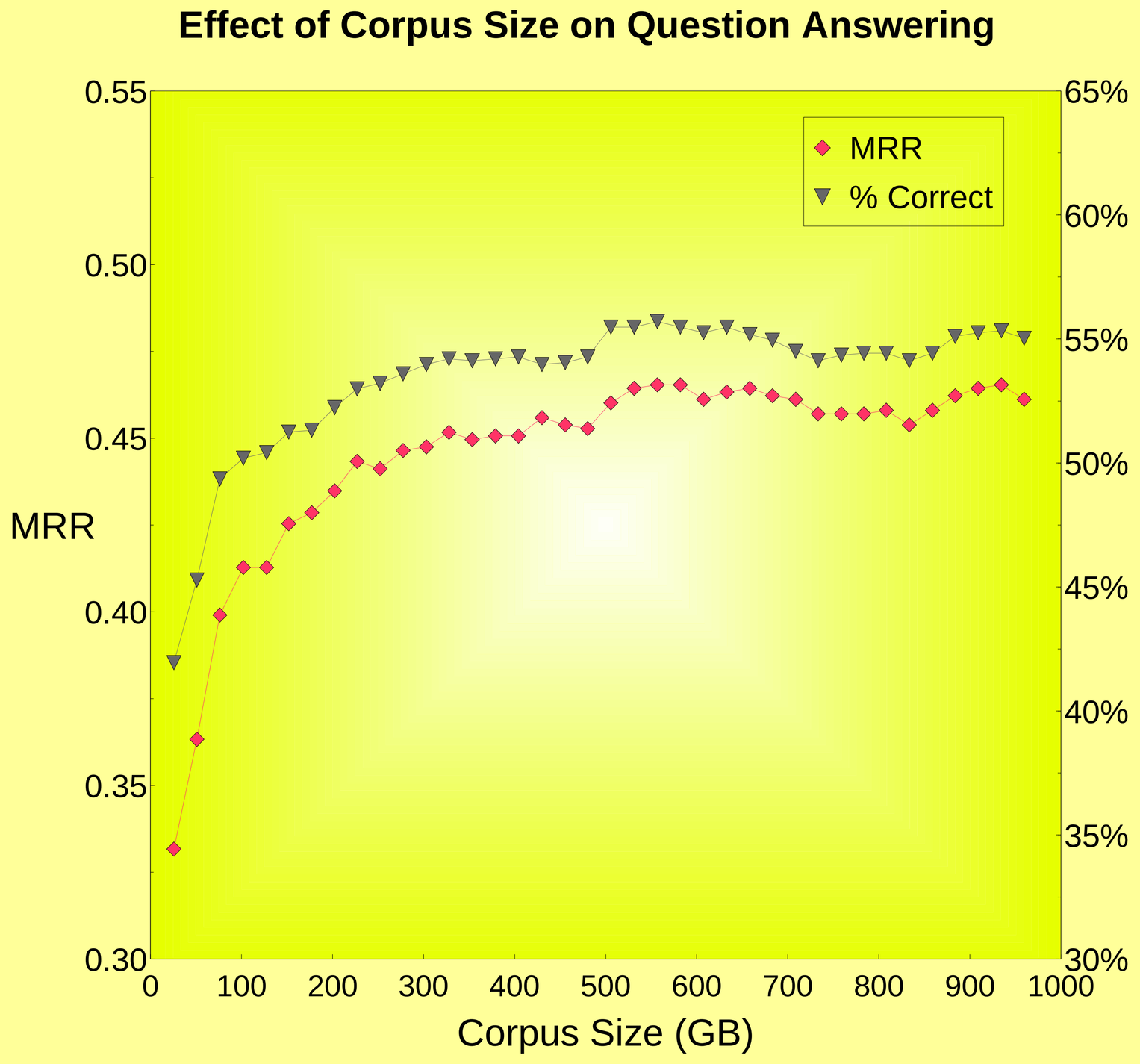
An automatic evaluation script supplied on the TREC web site was used to judge correct responses. The script executes a series of question−specific regular expressions over the responses returned for each question. Whenever a match occurs, the response is marked as correct. System performance was measured using Mean Reciprocal Rank (MRR)[2].

Relevant Passages

The number of relevant passages increases as the corpus size increases.

A passage is considered relevant if it contains a match to a regular expression.





Performance

Both mean reciprocal rank and percentage of questions answered correctly increase up to approximately 500GB where an asymptote appears to be reached.

1. Only 1323 of the 1393 Text REtrieval Conference (TREC) QA track questions were used in this evaluation. Questions with no answer patterns were excluded.

2. 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.