1 as a "team" vs. the keyword searching technique (Google and MSN as another "team"). We also wanted to know whether QA might decrease the cognitive load during the answer-seeking process. We therefore claim only that the idea of "going beyond keywords" is possible while searching for answers to questions, not that a particular system is better than another particular system.

No researcher has yet claimed to have produced a representative set of questions for evaluating QA systems. Indeed, such a set might have to include thousands of questions to adequately represent each possible type of question. We built a data set based on our experience with IT practitioners. We merged all the TREC questions with a set of 2,477,283 questions extracted earlier by Radev et al.[5] from the Excite search engine log of real search sessions.[5] We then distributed nonoverlapping sets of 100 randomly drawn questions to each of the 16 students in a technology-related MBA class at Arizona State University. The survey was followed by interviews and resulted in the selection of 28 test questions guided by participant choices and comments. In order to avoid researcher bias, it was crucial that we not enter any of the questions into an online system—search engine or QA—before deciding whether to select that particular question for the test.

We used the mean reciprocal rank (MRR) of the first correct answer, a metric also used during the 2001 and 2002 TREC competitions and in several follow-up studies. It assigns a score of 1 to the question if the first answer is correct. If only the second answer is correct, the score is ½. The third correct answer results in a score of ⅓. The intuition that went into devising this metric is that a reader of online question-answering results typically scans answers sequentially, and "eyeballing" time is approximately proportional to the number of wrong answers before the correct one pops up. However, this computation is known to "misbehave" statistically, being overly sensitive to the cut-off position, the lowest-ranked answer considered,[5] thus its reciprocals are typically reported and used for averaging and statistical testing. Results are outlined in Table 2.

By rerunning our analysis with each of the members excluded from the QA team, we verified that no weak players would pull down the QA team's performance. Because our intention was not to compare individual QA systems, we did not include the data for each individual QA system. The average results support the following observations:

- The QA team performed much better than the keyword-search-engine team, an MRR of 0.42 vs. 0.27; a remarkable 50% improvement was statistically significant, with the p value of the t-test at 0.002;
- The average performance of the QA team is better than the performance of each search engine individually; moreover, each QA system performed better than each keyword search engine;
- For each question to which an answer was found by a keyword search engine, at least one QA system also found an answer; the reverse was not always the case; and
- If a QA system found the correct answer, it was typically second or third in the ranked list; only the fourth or fifth snippet from Google or MSN typically provided the correct answer.

To verify the stability of these observations, we re-ran our tests in spring 2006. Although most of the measurements of the specific systems with respect to the specific questions had changed, their overall performance did not change significantly, and our observations were further reinforced.

Conclusion

Based on our interaction with business IT practitioners and an informal evaluation, we conclude that open-domain QA has emerged as a technology that complements or even rivals keyword-based search engines. It allows information seekers to go beyond keywords to quickly answer their questions. Users with limited communication bandwidth (as a result of small-screen devices or having some visual handicap) will benefit most. And users under some time constraint (such as first responders at a natural disaster) will likely find it more suitable compared to the keywords-to-snippets approach offered by popular search portals like Google and MSN.

However, to compete with established keyword-based search engines, QA systems still must address several technical challenges:

Scalability. Web QA system response time lags the one-to-two-second performance provided by today's search engines; more research needs to be done as to how to make Web QA systems more scalable in order to process the comparable loads simultaneously;

Credibility. Information on the Web, though rich, is less factually reliable than counterpart material published on paper; how can QA system developers, as well as search users, factor source credibility into answer ranking?; and

Usability. Designers of online QA interfaces must address whether QA systems should display precise answers, sentences, or snippets.

We look forward to the next five to 10 years for advances in all of them. C

References

1. Berwick, R.C. Principles of principle-based parsing. In *Principle-Based Parsing Computation and Psycholinguistics*, R.C. Berwick, S.P. Abney, and C. Tinny, Eds. Kluwer Academic Publishers, Norwell, MA, 1991, 1–38.
2. Dumais, S., Banka, M., Brill, E., Lin, J., and 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* (Tampere, Finland, Aug. 11–15). ACM Press, New York 2002, 291-298.
3. Kwok, C., Etienne, O., and Weld, D.S. Scaling question answering to the Web. *ACM Transactions on Information Systems 19*, 3 (2001), 242–262.
4. Lempert, R.J., Popper, S.W., and Bankes, S.C. *Shaping the Next One Hundred Years: New Methods for Quantitative, Long-Term Policy Analysis.* RAND Corp., Santa Monica, CA, 2003; direct.bl.uk/bld/PlaceOrder.do?UIN=138854587&ETOC=RN&from=searchengine.
5. Radev, D., Fan, W., Qi, H., Wu, H., and Grewal, A. Probabilistic question answering on the Web. *Journal of the American Society for Information Science and Technology 56*, 6 (Apr. 2005), 571–583.
6. Roussinov, D. and Robles, J. Applying question answering technology to locating malevolent online content. *Decision Support Systems 43*, 4 (Aug. 2005), 1404–1418.
7. Surdeanu, M., Moldovan, D.I, and Harabagiu, S.M. Performance analysis of a distributed question/answering system. *IEEE Transactions on Parallel and Distributed Systems 13*, 6 (2002), 579–596.
8. Voorhees, E. and Buckland, L.P., Eds. *Proceedings of the 13th Text Retrieval Conference TREC 2004* (Gaithersburg, MD, Nov. 16–19). National Institute of Standards and Technology, Gaithersburg, MD, 2004; trec.nist.gov/pubs/trec13/t13_proceedings.html.

Dmitri Roussinov (Dmitri.Roussinov@cis.strath.ac.uk) is a senior lecturer in the Department of Computer and Information Sciences at the University of Strathclyde, Glasgow, Scotland. The study described here was performed when he was an assistant professor in the Department of Information Systems in the W.P. Carey School of Business at Arizona State University, Tempe, AZ.

Weiguo Fan (wfan@vt.edu) is an associate professor of information systems and of computer science at Virginia Polytechnic Institute and State University, Blacksburg, VA.

José Robles-Flores (jrobles@esan.edu.pe) is an assistant professor in the School of Business at ESAN University, Lima, Perú. The study described here was performed while he was working on his doctoral dissertation in the Department of Information Systems in the W.P. Carey School of Business at Arizona State University, Tempe, AZ.