Getting Answers to Natural Language Questions on the Web

Dragomir R. Radev
School of Information and Department of EECS, University of Michigan, 553 East University Avenue, Ann Arbor, MI 48109. E-mail: radev@umich.edu

Kelsey Libner
School of Information, University of Michigan, 553 East University Avenue, Ann Arbor, MI 48109. E-mail: klibner@umich.edu

Weiguo Fan
School of Business, University of Michigan, 553 East University Avenue, Ann Arbor, MI 48109. E-mail: wfan@umich.edu

Most popular search engines are not designed for answering natural language questions. However, when we asked hundreds of natural language questions of nine leading search engines, all retrieved at least one correct answer on more than three-quarters of the questions. We identified the best-performing search engines overall for factual natural language questions. We found performance differences depending on the domain of factual question asked. Other aspects of questions also predicted significantly different performance: the number of words in the question, the presence of a proper noun, and whether the question is time dependent. An additional analysis tested for differential performance by specific search engines on these four question factors. The analysis found no evidence for such interactions.

Introduction

The Web has been growing at an accelerating pace since its first appearance in the spotlight in 1993. Three years ago Lawrence and Giles (1998) estimated that the Web contains a total of 320 million documents while today the Google search engine (www.google.com) claims to index 1.347 billion Web pages. This is a colossal and unique source of information about the world. The long-range goal of our research is to tap into this source of information by extracting answers to factual natural language questions. Examples of such questions are: "What is the longest river in the United States?," "What percentage of the world's plant and animal species can be found in the Amazon forests?," and "Who was the architect of Central Park?" Correct answers to each of these questions, all of which can be found on the web, are "the Mississippi," "20 to 30 percent," and "Frederick Law Olmsted."

Until recently, natural language question understanding was beyond the reach of information retrieval (IR) systems. To surmount the barrier of question understanding, the concept of a *query language* was introduced (see Frants, Shapiro, & Voiskunskii, 1997). A query language provides an intermediate system for capturing the essence of a user's information need and matching that information need to desired items in a repository of texts or other resources. The Web—used by hundreds of millions of users, containing hundreds of millions of pages of information, offering many options but lacking widely used standards—has to some extent upended the IR research tradition. On the repository side, rather than resources being cataloged under a controlled vocabulary like thesaurus and subject headings, they are generally indexed based on full-text words while accounting for position on the page and significant tags.[1]

On the query side, different search engines have different syntaxes. Some support advanced query languages, others do not. Stop word lists differ across search engines. Some search engines have a maximum query length (e.g., Google

Received August 1, 2001; revised December 5, 2001; accepted December 5, 2001.

[1] The extensive XML Schema specification from the World Wide Web Consortium, now under review (www.w3.org/XML/Schema), opens the door for wider adoption of standards in a distributed information environment; however, for the immediate future, it falls to the user to evaluate and compare the many available options.

recently announced that only the first 10 words of a query would be considered). Even if a search engine accepts advanced search syntax in one form or another, it may also accept natural language questions. Documentation may be available, but usually does not state exactly how queries are processed. Important information about questions may be lost (for example, the presence of the word "where" at the beginning of a question indicates that the expected answer is a location, while most search engines would drop this interrogative word).

In short, in the absence of a standard query language across search engines, and with so many nonexpert searchers turning to the Web for their information needs, there is a need for a system that can maximize Web search results when a natural language question is entered by a user. Recent theoretical and computational advances make it possible to design such a system for answering factual questions. The query component of such a system works as follows: questions can be parsed and alternatives can be generated that will maximize the likelihood of retrieving documents that contain the correct answer. Words will often be added to the question to guide the search process (Agichtein, Lawrence, & Gravano, 2001). The retrieval component, relying on the principle of *predictive annotation* (Prager, Brown, Coden, & Radev, 2000), uses logistic regression based on a set of automatically extracted features to determine what phrases in a set of documents are compatible in type with a given question and are thus likely to contain the answer to that question assuming they appear near words from the question.

In this article, we address one facet of this larger research program. We have set out to determine how successful search engines are at retrieving accurate answers when *unmodified* natural language questions are asked. This is to establish a baseline for subsequent work in which the effects of question modification and predictive annotation on question-answering will be examined.

Hypotheses

We tested the following hypotheses:

1. Search engines are good at answering factual answering questions.
2. Certain characteristics of questions—we examined four including number of words and presence of proper nouns—predict the likelihood of finding a correct answer across all search engines.
3. Questions with particular characteristics are more likely to elicit the correct answer from particular search engines.

The first hypothesis will lead to a baseline measure of the likelihood of correct answers being returned. Testing the second hypothesis will provide more fine-grained insights into the types of questions that a Web-based factual question-answering system is best suited for. If the third hypothesis were proven true, question-answering performance could be enhanced by routing different types of questions to different search engines.

It is important to acknowledge that search engines, operating in a competitive commercial environment, are "black boxes." Beyond the use of stop word lists, at least some search engines included in this study implement sophisticated natural language processing techniques including bracketing of collocations and the introduction of synonyms. This "black box" issue makes it a challenge to understand a search engine's strengths and weaknesses. In this study, because of the black box problem we base our inferences only on empirical results: based on the questions we send, what do we get back. Of course, this has its limits: the search engines are constantly tweaking their algorithms. Moreover, the Web is constantly evolving, as is a search engine's representation of Web information on disk. Given this state of affairs, it is impossible to make broad pronouncements about search engine performance. Rather, we present these findings to characterize the landscape of factual information currently available on the Web, as well as to establish a baseline for the proposed Web-based question-answering system.

Method

Using the search engines' public domain search APIs, we sent 700 questions from the TREC-8 and TREC-9 lists (Voorhees & Tice, 2000) to each of nine search engines: Alltheweb, AltaVista, Excite, Google, HotBot, Lycos, MetaCrawler, NorthernLight, and WebCrawler. The experiment was done in April 2000. For each question, we downloaded and stored the top 40 documents returned by each search engine and checked whether each contained an acceptable answer (there could be multiple acceptable answers). The *score* was calculated as the sum of the reciprocal ranks of documents containing the answer. For exam-

TABLE 1. Sample questions and scores for each search engine.

	av	aw	Ex	gg	Hb	Ly	mc	nl	wc
What is molybdenum? (permissible answers: metal, metallic element, strengthening agent, alloy in steelmaking)	2.35	1.27	2.87	1.80	2.17	3.12	3.25	1.54	0.97
What cancer is commonly associated with AIDS? (permissible answer: Kaposi's sarcoma)	0.49	0.14	1.21	2.23	1.17	0.52	1.25	0.74	0.10

ple, if the answer were found in documents 7, 10, and 38, the score would be:

$$\frac{1}{7} + \frac{1}{10} + \frac{1}{38} = .269$$

Reciprocal rank is an accepted measure in Question Answering evaluations (Voorhees & Tice, 2000). It favors hits that are ranked higher, however, gives appropriate weights to lower ranked hits. The maximum possible score using Reciprocal Rank is 4.28. As an illustration, Table 1 shows scores on each search engine for two of the 700 questions.

The first hypothesis above was evaluated by analyzing mean score across all questions for each search engine. To evaluate hypotheses two and three, the 700 questions were coded on the following four factors:

1. type of answer required (e.g., date; description/requiring text understanding; name; number; person; and place)
2. proper noun presence
3. time dependency (e.g., Who is the president of the United States vs. Who was the 35th President of the United States?)
4. number of words in question.

Table 2 lists the possible values for these factors.

Corresponding to the three hypotheses, an analysis of variance (ANOVA) was planned to (1) compare the overall score of the nine search engines, (2) determine the significance of the above four factors in predicting score, and (3) test for differential performance of search engines on each of these factors.

Analysis and Results

Because of a large proportion of zero-value scores, the initial distribution of scores had a positive skew. This skew would violate the normality assumption of ANOVA. A two-step analysis was performed to correct for the problem.

TABLE 2. Factors used in the statistical analyses: see appendix for cell sizes.

Factor name	Values
Answer type[a]	DATE, DESCTEXT,[b] NAME, NUMBER, PERSON, PLACE
Number of words	Four bins in ANOVA: 2–4, 5–7, 8–10, and 11 or more words
Time dependency	Binary: nontime-dependent, time dependent
Proper noun	Binary: no proper noun, one or more proper nouns

[a] Other answer types were excluded because of small cell sizes.

[b] This category was used for (1) questions requiring a description as an answer, and (2) questions requiring text understanding. These two types of answers were combined because in both cases, the connection between question and answer is more abstract and indirect than for other categories.

TABLE 3. Main effects and interactions for logistic regression and ANOVA.

Main effects	Interactions
Search engine	
Answer type	search engine × answer type
Proper nouns	search engine × proper nouns
Time dependency	search engine × time dependency
Number of words	search engine by number of words

1. Scores were transformed to one of two values, zero or nonzero, and a binary logistic regression was performed. A nonzero score means that at least one correct answer was in the top 40 documents retrieved for a given question sent to a search engine, while a zero score means that the correct answer was not retrieved. This analysis thus tests for a relationship between the four question characteristics and whether a correct answer is retrieved at all.
2. The second part of the analysis focused on *nonzero* values—instances where at least one correct answer was retrieved. Because this restricted dataset's distribution still showed some positive skew, it was square-root transformed. An ANOVA was then performed.

This two-step analysis was carried out to evaluate the hypotheses across the entire dataset without violating the normality assumption of ANOVA.

Step 1: Logistic Regression

Scores were transformed into a binary values: 0 for a zero score and 1 for all scores greater than zero. The four question factors—answer type, proper noun presence, number of words in question, and time dependency—were entered into the equation first, followed by two-way interactions of each factor with the search engine factor. This design is shown in Table 3.

The final model was highly significant (chi-square = 509.0, $df = 17$, $p < 0.001$). Nagelkerke (corrected) R-squared = 0.143. Key results:

1. The order of mean scores across all search engines, going from highest to lowest, was: Google, Northern-Light, Hotbot, Alltheweb, Lycos, Metacrawler, Excite, AltaVista, Webcrawler.
2. All main effects were highly significant, with p-values of less than 0.001 (see source table in appendix for other regression statistics). (a) the shortest queries (two to four words) had the highest mean scores. With longer queries, there was a trend of decreasing scores. (b) queries containing one or more proper nouns showed a slight but significant advantage over those without proper nouns. To get a benchmark about the proportion of real queries that include proper nouns, we examined 100 queries[2]

[2] Sex-related queries were excluded. Some queries appeared to be routed from AskJeeves.

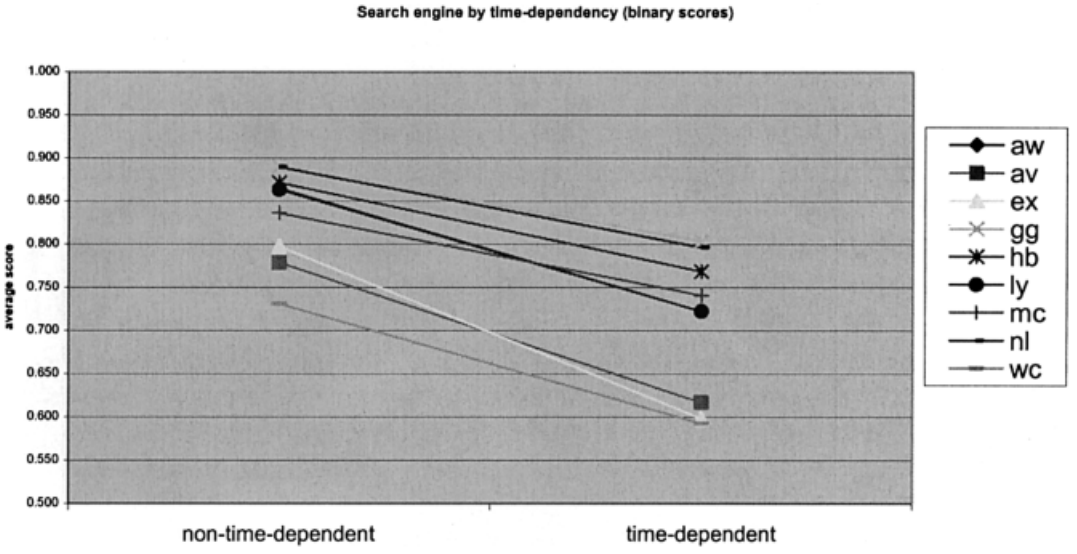

FIG. 1. Interaction between search engine and time dependency (roughly parallel lines indicate that all search engines show similar advantage on non-time-dependent questions such as "Who was the first president of the United States?").

drawn at random from a corpus of 2.4 million Excite queries and found that 51 included proper nouns. (c) Nontime-dependent queries—those where the correct answer does not change over a reasonable amount of time such as "When did Nixon visit China?"—showed an advantage over time-dependent queries (e.g., "What is the most expensive car in the world?"). (d) The following answer types all had mean scores of between 0.85 and 0.90: Description/text understanding, Place, Name, and Person. Date was somewhat lower at 0.81. Number was considerably lower at 0.59.

3. None of the interactions between search engine and the four question characteristics were found to be significant. Figure 1 illustrates the lack of an interaction between search engine and one of these factors, time dependency.

There is thus no evidence to suggest that individual search engines are differentially better at finding an answer for particular question characteristics used in this analysis (e.g., number of words in question).

Step 2: ANOVA

In the analysis above, all scores were transformed to one of two values. Subtler differences in score had to be ignored to be able to compare zero-score questions with nonzero score questions in a highly skewed distribution. In Step 2, the original scores were used for an ANOVA, but scores of zero were excluded to meet the normality assumption. (A further measure was also taken to meet this assumption: scores were square-root transformed.)

The continuous variable for number of words was placed into bins of 2–4, 5–7, 8–10, and 11 or more words.

1. There were significant differences in mean score by search engine:
2. The following results parallel those of the previous analysis: all main effects were found to be significant; questions containing proper nouns showed an advantage over those without; nontime-dependent questions showed an advantage over time-dependent questions; and the smallest bin for words in question showed the highest score.
3. Differences in score by answer type are parallel to the logistic regression but more articulated (see plot below; lines above columns are equivalence classes determined by Tukey LSD post hoc comparisons). Place answers are in the top equivalence class, followed by description/text understanding, person, and name; with date and number questions thereafter.
4. There is a significant quadratic trend in the scores for number of words in question: mean score drops for questions of five to seven words, then rises again for questions of 8–10 and 11 or more words—a finding we have no ready explanation for.
5. No evidence was found for a significant interaction between search engine and the other main effects.

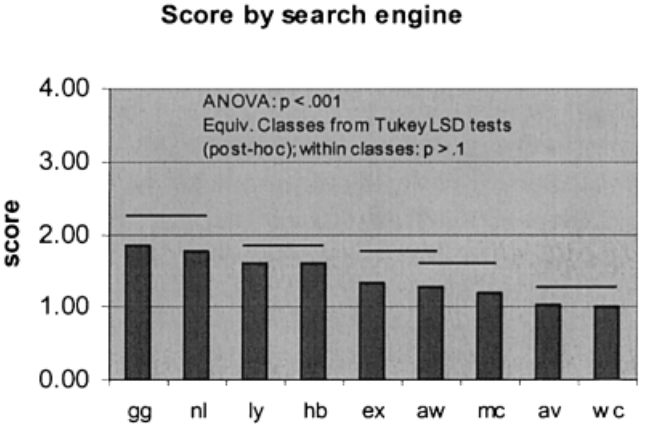

FIG. 2. Comparative performance of search engines.

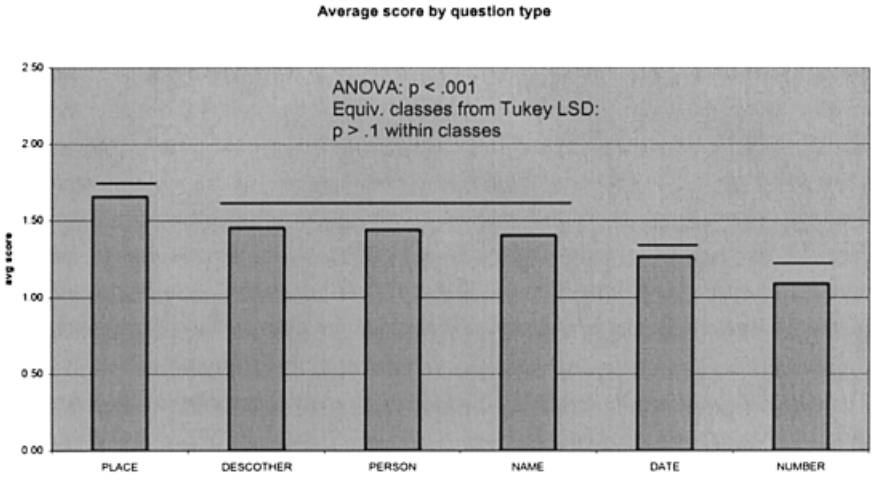

FIG. 3. Score by answer type.

Conclusions

We return to the three hypotheses stated above.

Are Search Engines Good at Answering Factual Answering Questions?

Table 4 lists the percentage of questions, for each search engine, where at least one correct answer was found in the documents returned.

All search engines retrieved the correct answer at least somewhere in the top 40 documents 75% of the time or more. As a caveat, note that mere presence of the answer anywhere in the document constituted a "hit." Of course, in a question-answering system where answers were not known in advance, an additional step, answer extraction, would be necessary and would introduce some level of noise into the process. Nevertheless, the table shows that on the whole, the correct answers are out there and are being retrieved without recourse to a query language. Moreover, the two highest performing search engines, Google and Northern Light, provide the correct answer on just short of 90% of questions. Although the additional step of answer extraction is not at all trivial, this baseline for unmodified questions suggests that Web search engines have strong potential as a component of a system for answering factual natural-language questions.

TABLE 4. Percentage of questions for which a correct answer was retrieved in the top 40 documents.

Search engine	Percentage of questions for which a correct answer was retrieved in the top 40 documents
gg	87.7
nl	87.5
hb	85.6
aw	84.3
ly	84.1
mc	82.1
ex	76.7
av	75.4

Do Certain Aspects of Questions Predict Different Levels of Question-Answering Performance Across All Search Engines? (Main Effects of Question Factors)

Yes. Highly significant main effects of answer type, question length, proper noun presence, and time dependency were found in both analyses. Date and number questions were somewhat less successful than other answer types. Nontime-dependent questions were more successful than time-dependent ones, and those containing proper nouns were higher scoring than those not containing proper nouns.

Are Different Search Engines Better at Answering Different Types of Questions?

We found no evidence for such interactions. As a result, the proposed system for question-answering on the Web should direct questions to the two or three search engines with the best overall performance.

Acknowledgments

This work was partially supported by a Research Incentive Grant at the University of Michigan. The authors would also like to thank Paul Resnick at the University of Michigan and Einat Amitay at IBM Haifa Research Center for their very useful comments on earlier versions of the article.

Appendix: Cell sizes.

Factor	Level	*N*	Factor	Level	*N*
QTYPE	DATE	58	BPROPNN	0	125
	DESCOTHER	103		1	397
	NAME	93			
	NUMBER	41			
	PERSON	132			
	PLACE	95			
TIMEDEP	0	457	WDSBINS2	2–4	103
	1	66		5–7	207
				8–10	131
				11+	82

Source table of logistic regression.

	B	S.E.	Wald	*df*	Sig.	Exp(B)
SCHENG			132.242	8	0.000	
WDSINQ	−0.039	0.010	15.772	1	0.000	0.961
BPROPNN	0.533	0.088	36.912	1	0.000	1.704
TIMEDEP	−0.453	0.101	20.114	1	0.000	0.636
QTYPENUM			252.926	6	0.000	
Constant	1.276	0.164	60.227	1	0.000	3.582

References

Agichtein, E., Lawrence, S., & Gravano, L. (2001). Proceedings of the tenth international World Wide Web conference (WWW2001), Hong Kong, 5–11 May 2001.

Frants, V.I., Shapiro, J., & Voiskunskii, V. (1997). Automated information retrieval: Theory and methods. San Diego: Academic Press.

Harabagiu, S., et al. (2000). The structure and performance of an open-domain question answering system. Proceedings of ACL 2000, Hong Kong, October 2000.

Joho, H., Liu, Y.K., Sanderson, M. (2001). Large-scale testing of a descriptive phrase finder. In Proceedings of HLT (Human Language Technologies) Conference, 2001.

Lawrence, S., & Giles, C.L. (1998). Searching the World Wide Web. Science, 280, 98.

Prager, J., Brown, E., Coden, A., & Radev, D. (2000). Question answering by predictive annotation. Proceedings of SIGIR 2000, Athens, Greece, July 2000.

Radev, D.R., et al. (2001). Mining the Web for answers to natural language questions. In Proceedings of ACM CJKM 2001, Atlanta, Georgia, November 2001.

Voorhees, E., & Tice, D. (2000). The TREC 8 question answering track evaluation. Proceedings of TREC 8, Gaithersburg, MD, 2000.