# Towards Answer-Focused Summarization

Harris Wu, Dragomir R. Radev, Weiguo Fan

*Abstract*— **People query search engines to find answers to a variety of questions on the Internet. Search cost would have been greatly reduced if search engines could accept natural language questions as queries, and provide summaries that contain the answers to these questions. We introduce the notion of Answer-Focused Summarization, which is to combine summarization and question answering. We develop a set of criteria and performance metrics, to evaluate answer-Focused Summarization. We demonstrate that the summaries produced by Google, the most popular search engine nowadays, can be largely improved for question answering. We develop a proximity-based summary extraction system, and then utilize question types, i.e. whether the question is a "person" or a "place" question, to improve the performance. We suggest that there is a large application potential for Answer-Focused Summarization, such as in wireless and palm-held systems where search cost is critical.**

***Index Terms*--** **Answer-focused summarization, summarization, question-answering.**

## I. Introduction

People use search engines to find answers to a variety of questions on the Internet. Sometimes users issue queries in the original form of a natural language question, such as "Who is the president of France?". An analysis of the Excite corpus of 2,477,283 actual user queries shows that around 8.4% of the queries are in the form of natural language questions [10]. More often, users rewrite their natural language questions into keywords and try different combinations of keywords as queries on search engines. Obviously much information will be lost when a natural language question is reduced to a set of keywords. However, many users prefer to use keywords rather than natural language questions, for the search engines appear to be inefficient in accepting natural language questions as queries. In fact, search engines typically filter out stop-words such as "what" and "why" so a question would have been processed as a list of keywords even if it were submitted in a natural language question form.

Harris Wu (harriswu@umich.edu) is with the Business School, University of Michigan.

Dragomir R. Radev (radev@si.umich.edu) is with the School of Information, University of Michigan.

Weiguo Fan (wfan@vt.edu) is with Pamplin College of Business, Virginia Tech University.

After a search engine returns a set of hits related to the query, the user's next step is then to go through these documents in the hit list and find the answers to the original question. The search cost depends largely on how many document links a user has to click through, and the length of each document that a user has to skim over.

Some search engines, such as Google and NorthernLight, provide a short summary for each document with query words highlighted. If a question has a short answer and if a user can find the answer in these summaries, the search cost can be largely reduced. The quality of these summaries has a significant impact on search costs.

Search costs would have been greatly reduced if search engines could accept natural language questions as queries, and provide summaries that contain the answer to these questions. In this paper, we will focus on this specific type of summarization, namely Answer-Focused Summarization (AFS), which summarizes documents to answer questions in a natural language question form. Answer-focused summaries are different from traditional summaries in that they are tailored to provide answers to users' natural language questions.

Answer-Focused Summarization is not limited to search engines. The process of extracting answer-focused summaries from documents is a form of query-based passage retrieval [3], [11], [14]. The three most recent TREC conferences (1999 to 2001) [13] included a popular track in which Q&A systems from more than 50 institutions competed. Simultaneously, the DUC conference [4] provides a forum for the evaluation and comparison of text summarization systems. As both recent roadmap papers on Summarization [1] and Question Answering [2] indicate, the dual problems of information access, summarization and question answering, are bound to meet.

In this paper, we provide some additional motivation showing how intricately related summarization and question answering are. We describe some experiments that indicate promising directions along which we can improve Answer-Focused summarization, and present a set of methods that significantly outperforms Google's summary generator in question answering. The contributions of this paper are as follows:

1. We develop a set of criteria and performance metrics to evaluate Answer-Focused Summarization.

2. We develop a proximity-based answer-focused summarization system, which complements existing search engines and other document retrieval systems. Our

summarization system is capable of analyzing and utilizing question types to provide better answer-focused summaries.

3. Through experiments we find summarization provided by Google, one of the best search engines available, sub-optimal in question answering. Our summarization system shows significant advantage in key performance metrics.

4. We provide recommendations to search engines to improve their summarization for question answering. We also suggest some promising research directions along which better answer-focused summarizations and information retrieval systems can be made. We suggest that better answer-focused summarization may help bring Internet information systems to wireless and hand-held devices, where search cost is critical.

## II. Evaluation Answer-Focused Summaries

Not all answer-focused Summaries are created equal. Ideally, a good answer-focused summary should satisfy three criteria in decreasing order of importance:

1. Accuracy: it should contain the answer to the user's question

2. Economy: it should be as short as possible, and

3. Support: it should provide evidence that supports the answer

We should note that in above, support may contradict economy. For single-document summaries provided by search engines, users can go from a summary to the parent document to find support for the answer. In this paper we will limit ourselves to single-document summaries and focus on the first two criteria. A future paper will be devoted to the support issue for extracted answers.

The TREC Q&A evaluation (up to the latest TREC-10) uses the MRDR metric to compare participating systems. MRDR (Mean Reciprocal Document Rank) is the mean of reciprocal ranks of the highest-ranked correct answers to a set of questions. For example, if the third answer presented by the system is the first correct one, the value of Reciprocal Document Rank will be 1/3. If we treat the lists of answers in TREC evaluation as answer-focused summaries, the MRDR metrics addresses a combination of criterion 1 and criterion 2. In addition, The TREC evaluation addresses criterion 2 by stipulating a fixed number of characters (50 or 250) to be extracted by the systems. Given that most answers to the TREC evaluation corpus are rather short (less than 20 characters), a large percentage of the extracted characters do not contain answer words at all. Criterion 3, support, is not considered in the current TREC Q&A evaluation.

In this paper, we will use four different metrics to evaluate the first two criteria, accuracy and economy, on answer-focused summaries of documents returned from search engine queries:

1. Whether a question is answered
2. Summary length in characters
3. Summary rank of first answer
4. Word rank of first answer

Both summary rank and word rank depend on the order of relevant summaries returned by a system. For example, if a search engine returns 10 documents and one summary per document, each 25words long, then a hypothetical correct answer starting at the fourth word in the third summary would get a summary rank of 3, and a word rank of 54 (25+25+4). The highest possible value for word rank in this case is 250 while the lowest (best) value is 1.

Figure 1 shows the summaries produced by Google for

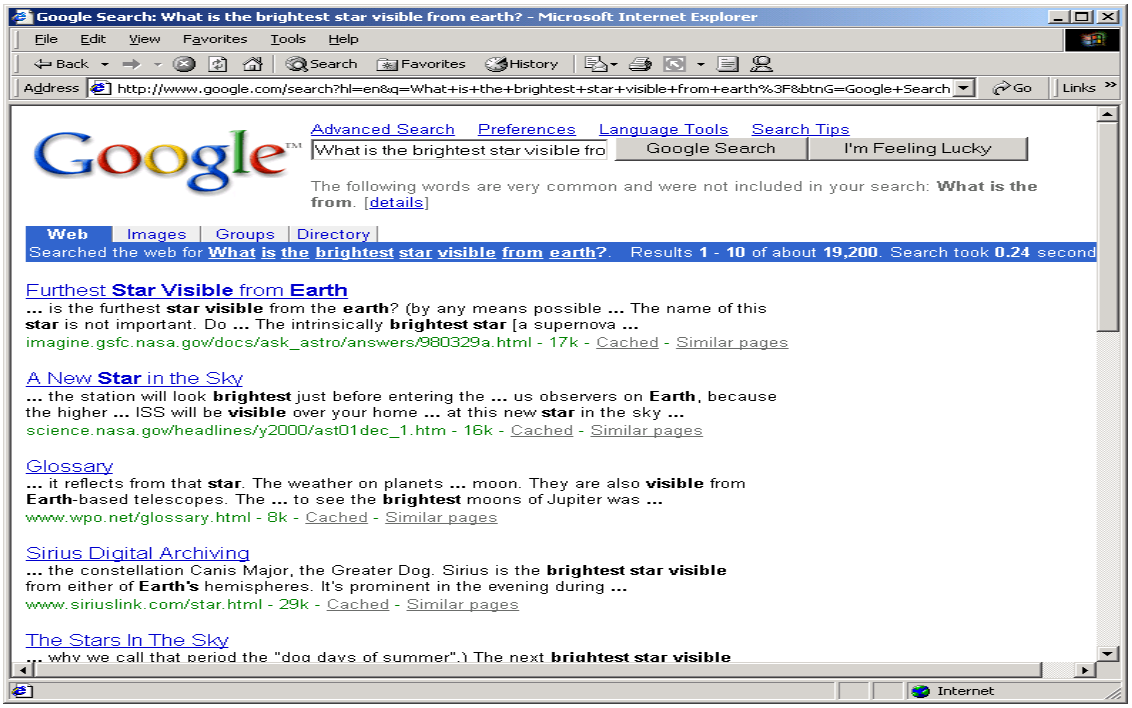


**Figure 1. Summaries provided by Google**

Question 29 of TREC-8: “What is the brightest star visible from Earth?” The correct answer, “Sirius”, also occurs in the summaries produced by Google. It appears as the 91st word from the beginning of the return page (URLs, ads, titles, and links such as Similar pages" are not counted), in the 4th hit. Therefore, the value of word rank for this question is 91, and the value for summary rank is 4. Note that although the titles and URLs may also contain answers, however they are not produced from summarization and we cannot modify them. Titles and URLs reflect the quality of document retrieval rather than the quality of summarization, thus we exclude them from our evaluation. In fact, the title of the 4th hit, “Sirius Digital Archiving", contains the answer “Sirius". But a user would not recognize Sirius as the answer by reading the title, without reading the summary in that hit, “...Sirius is the brightest star visible from either of Earth's hemispheres...”

In the example above, it is clear by reading the summaries that somehow the Google's summarizer just misses the correct answer in the first two summaries by a small number of words. Obviously, a good summarization algorithm should catch these correct answers while, at the same time, should not produce extra-long summaries.

In this paper we will show from our experiments that there is a large potential for improvement in summaries provided by search engines for question answering. Furthermore, we will show that the information contained in a question itself, as simple as whether the question is a “who” or “where” question, can significantly contribute to the performance of summarization.

## III. EXPERIMENTS

### *A. Baseline: Google*

Google is one of the most popular search engines on the Internet. It has been shown that Google has the best performance in question answering among other search engines [9], [10]. Our baseline is based on Google's summaries.

First, we sent the 200 questions from TREC-8 [12] to Google. We saved the first page of results for each question and parsed them to extract the different fields that Google produces: URL, Title, Description (rarely provided, only 203 of them were retrieved from 2000 hits), Summary, and Category. Note that Google doesn't allow access by automated agents so we had to query it manually.

Second, we removed from consideration hits originating in sites participating in TREC. Most of the time, such hits contain both the TREC question and its answer. Allowing these documents in the evaluation would have skewed the results. In total, 94 out of 2000 hits came from such places. Following is the list of the sites that appeared most frequently in the hit lists:

www-personal.umich.edu/~zzheng/answerbus/, www.isi.edu/natural-language/projects/webclopedia/, carleton.ca/~sscott2/, trec.nist.gov, www.cs.utexas.edu, www.limsi.fr/Individu/QA/, www.clairvoyancecorp.com/.

Third, we evaluate Google’s summaries using the metrics for answer-focused summaries: number of questions answered, average character length of summaries, mean reciprocal of summary ranks (MRSR) and mean reciprocal of word ranks (MRWR). To illustrate MRSR, an MRSR value of 0.2 means that on average the answer to a question appears in the 5th summary provide by the system. Even when the hit order of URLs remains the same, better quality of summaries will result in higher MRSR. An MRSR value of close to 1 means that a user on average can find the answers in the first summary they read. Similarly, an MRWR of 0.02 means that a user on average needs to read 50 words of summaries to get the answer. Of course, MRSR and MRWR are limited by the quality of document ranking provided by the search engine.

Overall, 73 of the 200 hit lists contained one or more answers. Table 1 gives the values of four metrics for Google's summaries.

**Table 1. Baseline values from Google’s summaries**

| TREC Questions | Answered | Summary Avg. Length | MRSR | MRWR |
|---|---|---|---|---|
| 1-100 | 35/100 | 159 | 0.187 | 0.0278 |
| 100-200 | 38/100 | 160 | 0.228 | 0.0474 |
| 1-200 | 73/100 | 160 | 0.207 | 0.0376 |

### *B. Experiment 1: Proximity-based summary*

This experiment is to test our proximity-based summarization system. Google's summarization algorithm is proprietary and we do not have detail of its implementation. Based on our observation of the summaries provided by Google, the algorithm seems to have used word proximity, relative word ordering, position in text, and a large indexed document database. Since we do not have a large document database, our summarization system uses a rather simple proximity-based extraction.

It has been observed that the distance from query words to answers in a document is relatively small [7]. Here we try to quantify that distance and use it for AFS extraction. We follow the steps below:

1. We identify all query words that appear in the documents on the hit list.
2. We extract short windows of size *wsize* around query words.
3. We put the windows in order so that the windows containing most number of unique query keywords are placed in the front.
4. We concatenate as many short windows as possible to produce a summary within *ssize* characters. The two variables, *wsize* and *ssize* are given as parameters. We run experiments with *wsize* from 1 to 20 at increment of 1, *ssize* from 100 to 400 at increment of 10.

**Table 2. Performance of Experiment 1, proximity-based summary extraction**

| | **Training on questions 1-100** | | | **Evaluation on questions 100-200** | | |
|---|---|---|---|---|---|---|
| *wsize* | #Answered | MRSR | MRWR | #Answered | MRSR | MRWR |
| 1 | 23 | .075 | .005 | 33 | .167 | .014 |
| 2 | 38 | .164 | .010 | 41 | .243 | .038 |
| 3 | 38 | .176 | .027 | 39 | .242 | .039 |
| **4** | **43** | **.211** | **.023** | **41** | **.237** | **.019** |
| 5 | 37 | .190 | .028 | 42 | .231 | .020 |
| 6 | 37 | .202 | .024 | 39 | .243 | .019 |
| **Google** | **35** | **.187** | **.028** | **38** | **.228** | **.047** |

The actual average passage size is below the summary size limit *ssize*. *ssize*=200 generally provides summaries of average size below 160, which is comparable to Google's. We focus on finding the optimal window size for our proximity-based summary extraction. Using the first one hundred TREC-8 questions as training data and the metrics mentioned in last section, we find the best window size given actual average passage size below Google's. It is hard to conclude on the order of importance among different metrics, however we assume the number of questions answered is the most important, then hit rank followed by word rank. The passage size is an input variable instead of a metric in our experiments. Note that for any given window size, larger *ssize* will always bring equal or better values on number of questions answered and hit ranks, but sometimes may lower the word rank. Using the optimal window size from training and keeping the average passage size below Google's, we evaluate on the second one hundred questions in the corpus, The performance results are shown in Table 2. Due to limited space, Table 2 only shows for *wsize* from 1 to 6.

As one can see in Table 2, our algorithm chooses *wsize*=4 as the best window size from training. On evaluation set of questions, 41 questions are answered by our summaries, versus 38 questions answered by Google. Our summaries have MRSR of 0.237, versus Google's 0.028. However, Google has better MRWR in both training and evaluation set. This is expected because we do not have any redundancy removal steps in our extraction procedure.

Although our simple proximity-based summary outperforms Google, note that there is not a clear, single optimal window size in Table 2. In the evaluation, the window size learned from training, *wsize*=4, is quite far way from the actual optimum, *wsize*=9. Maybe the optimal window size depends on certain characteristics of a question, such as the question type, e.g. place, date, etc.?

### *C. Experiment 2: Summarization with question type known*

This experiment is to test whether the knowledge of question types will help improve the performance of our proximity-based summary extraction. An analysis of the Excite corpus [9] shows that among the natural language question queries, 43.9% are factual questions (e.g., "What is the country code for Belgium") while the rest are either non-factual ("How do I …"), syntactically incorrect or unanswerable questions.

In this experiment, we first manually classify each of the 200 TREC-8 questions into one of the following question types based on the answers they expect: Person (56), Place (36), Name (29), Number (23), Date (22) and Other (34). We use the same proximity-based extraction described in last section, to get the summary for each question, with different window size and passage size parameters. We find the best window size for each question type, using the first one hundred TREC-8 questions as training data. Then we evaluate on the second one hundred of the TREC-8 questions. Table 3 shows the performance results. In Table 3, "ansrd" stands for "answered", and "w" stands for "*wsize*".

**Table 3. Performance of Experiment 2, proximity-based summary extraction with question type known**

| | Training optimum | | | | Eval with training *wsize* | | | Actual optimal *wsize* in eval | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **qtype** | **w** | **#ansrd** | **MRSR** | **MRWR** | **#ansrd** | **MRSR** | **MRWR** | **w** | **#ansrd** | **MRSR** | **MRWR** |
| person | 4 | 14/34 | .328 | .042 | 14/22 | .370 | .043 | 5 | 14/22 | .370 | .047 |
| place | 11 | 11/15 | .218 | .213 | 15/21 | .423 | .029 | 9 | 17/21 | .462 | .048 |
| name | 7 | 14/19 | .237 | .061 | 4/10 | .114 | .006 | 10 | 5/10 | .168 | .008 |
| number | 2 | 2/6 | .056 | .002 | 5/17 | .147 | .005 | 2 | 5/17 | .147 | .005 |
| date | 18 | 3/10 | .200 | .073 | 3/12 | .139 | .004 | 4 | 5/12 | .262 | .011 |
| other | 13 | 2/16 | .133 | .016 | 5/18 | .067 | .003 | 8 | 5/18 | .136 | .006 |
| **all types** | | **47/100** | **.234** | **.068** | **46/100** | **.247** | **.018** | | **51/100** | **.259** | **.024** |

Using question type information, the performance of summary extraction has improved notably. On the training set, 47 questions are answered compared to 41 answered in Experiment 1, which indicates that questions of different types tend to have different optimal window sizes. On the evaluation set, 46 questions are answered compared to 41 answered in Experiment 1, and 38 answered by Google. The number of questions answered is 21% more than Google's, and 12% more than that of Experiment 1.

For Person type questions, the optimal window size from training data is 4, and the optimal window size in evaluation is 5. The optimal window sizes are quite close. Actually for the training data, both *wsize*=4 and *wsize*=5 answer 14 out of 22 questions and have the exactly same MRSR, but *wsize*=5 gives a slightly higher MRWR. The training seems to be very successful in this case. Place type questions seem to have answers farther away from the query keywords, with optimal window size 11 learned from training, and actual optimal window size 9 in the evaluation.

In our experiment less than half of the Date, Number and Other questions are answered. These questions are indeed hard to answer. For example, "Why are electric cars less efficient in the north-east than in California?" or, "How much could you rent a Volkswagen bug for in 1966?" Simple proximity-based summary extraction as described here does not handle these complex questions well.

### D. Experiment 3: Relax the summary size constraint

This experiment is to investigate how different summary size constraints affect optimal window sizes.

In above experiments we have limited the average summary size to be below Google's average of 160 characters. In general, relaxing the constraint will result in longer summaries, which produce more answers and better hit ranks. How will different summary sizes affect the optimal window sizes though?

Figure 3 shows the performance for Person questions under different summary size limits from 100 to 500. Due to limited space, we only show the results for window size 1,2, 4, 5 and 9. From Figure 3, we can see *wsize*=4 is the optimal window size when the summary limit is between 130 and 300 characters. When the summary size limit is between 300 and 500 characters, *wsize*=9 becomes the optimum. As the summary size limit grows larger, the number of questions answered for a given window size first increases, then eventually becomes stagnant after the summary size limit grows over a threshold. At the right end of the chart, with summary limit at 500 characters, the numbers of questions answered by different window sizes are ordered by the window size. That is, with *ssize*=500, *wsize*=9 answers most number of questions and *wsize*=1 answers the least. Clearly the optimal window size is relative to a given summary size constraint, and as the summary size becomes larger, the optimal window size also becomes larger.

It is clear in Figure 3 that more questions are answered as summary size grows larger. In the future we plan to study the slopes of these performance curves, so we can choose the optimal summary size in a given specific situation.

## IV. Discussion and Conclusion

Our work is related with query-biased summaries in information retrieval [11], [14], where sentences ranked on the overlap with the query words are used to produce summaries for each search hit. However, answer-focused summarization is different from [11], [14] in that AFS produces summaries in favor of getting answers to natural language questions. The Webclopedia project [5], a web-based search summarization system, is more similar to what we are doing. But in our experiments we focus on a sub-problem of theirs, i.e. we assume the documents related to the question are already identified and ranked, and we only need to extract answer-focused summaries from these documents.

From our experiments we have shown that there is a large potential for search engines to improve on answer-focused summaries. A good search engine should "present the most important content to the user in a condensed form," and "in a manner sensitive to the user's need" [6]. In practice, it is not hard for search engines to recognize questions in users' queries. For example, all queries ending with a "?" can be recognized as questions. Once a search engine recognizes a query as a question, it can optimize the results in an answer-focused way. The search engine should try to display answers to the question directly on the return page, so users do not have to click through document URLs. Search engines can adopt the techniques developed in this paper to provide better answer-focused summaries. We hope to motivate practitioners as well as researchers to make one more step towards true information retrieval, rather than document retrieval. From user acceptance point of view, better answer-focused summaries will contribute to the wider acceptance of search engines and other information retrieval technologies.

We showed that the window size, a parameter frequently used in proximity based search and summarization, can be optimized for different question types in providing answer-focused summaries. By discovery of this relationship between optimal window size and question type, we largely improved the performance of our proximity-based summary extraction mechanism. We developed a question type determination algorithm, which is not discussed in the paper due to limited space. Despite its prematureness (23.5% error rate), the question type determination algorithm still made large contribution to the improvement of summary

extraction. We also observed the pattern of optimal window sizes under different summary size constraints. Studying the pattern will help us to make better decisions on summary sizes.

There are a richness of information in questions themselves besides the simple question types we devised to improve our summary extraction. Query preprocessing such as stemming, expansion, term relationships, etc. has been studied for years. We believe Web search engines can greatly benefit from query preprocessing in providing answer-focused summaries.

It will be interesting to study the human factors in answer-focused summarization. We made some assumptions when developing the performance metrics. For example, we assumed that the users will skim through the summaries when looking for answers, thus the passage size is both an important metrics and a practical constraint. In real life different users will have different expectations and behave differently when searching for answers to their questions.

One issue we did not address in this paper is the support for answers. Providing the answer is only one part of the story - a user would only appreciate if he or she can find sufficient contextual support and then accept the answer. We plan to do further study on how to provide efficient support in answer-focused summaries.

We plan to extend our study in summary size, or, the economics of summaries. Summary size directly affects accuracy, part of which we have observed in studying optimal window sizes, and support, which certainly requires extra words other than just answers. For one implication of the economy, let us think about mobilizing users from desktops to wireless devices or hand-held devices. Due to the constraints in network bandwidth or visual display, qualities of information retrieval will have a critical impact on search costs using these devices. It is important to understand the tradeoff between summary size, support and accuracy, to generate summaries for different user and application needs.

Better answer-focused summaries will not only help millions of desktop users getting answers to their questions, but also help bring Internet information systems to wireless devices and hand-held devices. This is not trivial; diffusion of information, diffusion of Internet, diffusion of IR technologies, and diffusion of next-generation devices may depend on it.

## REFERENCES


[1] B. Baldwin, R. Donaway, E. Hovy, E. Liddy, I. Mani, D. Marcu, K. McKeown, V. Mittal, M. Moens, D. Radev, K. S. Jones, B. Sundheim, S. Teufel, R. Weischedel, and M. White. An evaluation roadmap for summarization research. TIDES, July 2000.

[2] J. Burger, C. Cardie, V. Chaudhri, R. Gaizauskas, S. Harabagiu, D. Israel, C. Jacquemin, C.-Y. Lin, S. Maiorano, G. Miller, D. Moldovan, B. Ogden, J. Prager, E. Riloff, A. Singhal, R. Shrihari, T. Strzalkowski, E. Voorhees, and R. Weishedel. Issues, tasks and program structures to roadmap research in question answering. TIDES, 2000.

[3] J. P. Callan. Passage-level evidence in document retrieval. In W. B. Croft and C. van Rijsbergen, editors, Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 302-310. Spring-Verlag, 1994.

[4] D. Harman. Proceedings of the Workshop on Text Summarization 2001. NIST, 2001.

[5] E. Hovy, L. Gerber, U. Hermjakob, M. Junk, and C.-Y. Lin. Question answering in webclopedia. In NIST Special Publication 500-249:The Ninth Text REtrieval Conference (TREC 9), pages 655-664, 2000.

[6] I. Mani. Automatic Summarization. John Benjamins Publishing Company, Amsterdam/Philadephia, 2001.

[7] J. Prager, E. Brown, A. Coden, and D. Radev. Question-answering by predictive annotation. In Proceedings, 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Athens, Greece, July 2000.

[8] D. Radev, W. Fan, H. Qi, H. Wu, and A. Grewal. Probabilistic question answering from the web. Proceedings of the 11th WWW conference, 2002.

[9] D. R. Radev, K. Libner, and W. Fan. Getting answers to natural language queries on the web. Journal of the American Society for Information Science and Technology (JASIST), page to appear, 2001.

[10] D. R. Radev, H. Qi, Z. Zheng, S. Blair-Goldensohn, Z. Zhang, W. Fan, and J. Prager. Mining the web for answers to natural language questions. In the Proceedings of ACM CIKM 2001: Tenth International Conference on Information and Knowledge Management, Atlanta, GA, 2001.

[11] A. Tombros and M. Sanderson. Advantages of query biased summaries in information retrieval. In the Proceedings of SIGIR'98, pages 2-11, 1998.

[12] E. Voorhees and D. Tice. The TREC-8 question answering track evaluation. In Text Retrieval Conference TREC-8, Gaithersburg, MD, 2000.

[13] E. M. Voorhees. Overview of the trec 2001 question answering track. In Proceedings of TREC 2001 Conference, pages 157-165. NIST, 2002.

[14] R. White, I. Ruthven, and J. M. Jose. Web document summarisation: a task-oriented evaluation. In Proceedings of the First International Workshop on Digital Libraries (DLib 2001), Munich, Germany, 2001.