# Evaluating Web-based Question Answering Systems

**Dragomir R. Radev**[*]†, **Hong Qi**[*], **Harris Wu**[‡], **Weiguo Fan**[‡]

[*]School of Information
[†]Department of EECS
[‡]Business School
University of Michigan
Ann Arbor, MI 48109
{radev, hqi, harriswu, wfan}@umich.edu

**Abstract**

The official evaluation of TREC-style Q&A systems is done manually, which is quite expensive and not scalable to web-based Q&A systems. An automatic evaluation technique is needed for dynamic Q&A systems. This paper presents a set of metrics that have been implemented in our web-based Q&A system, namely NSIR. It also shows the correlations between the different metrics.

## 1. Introduction

Question Answering is a research area that has recently gained a lot of interest, especially in the TREC community. More than 40 research groups participated in the most recent evaluation of "static" Q&A systems, organized by NIST. We call TREC-style systems "static" because they are designed to answer factual questions from a static, 2-GB collection of newswire. In contrast to TREC-style systems, "dynamic" Q&A systems use the entire Web as a corpus, typically through the intermediary of a commercial search engine.

The official evaluation of TREC-style Q&A systems is done manually (Voorhees and Tice, 2000; Prager et al., 1999). A number of assessors judge answer strings on two criteria: how accurately they answer the question and how much justification of the answer is provided. Similarly, user-based techniques are used in similar systems on the Web (Agichtein et al., 2001; Kwok et al., 2001). However, such manual evaluation is quite expensive, and does not scale beyond a few thousand answer strings. To evaluate dynamic Q&A systems, an automatic evaluation technique is needed.

(Radev et al., 2002) compares the manual and automatic evaluation on TREC 8 questions, and gets a Pearson's correlation coefficient of 0.54. As a result, this justifies the use of automated techniques when manual evaluation is too expensive (e.g., on tens of thousands of question-document pairs). MRR (mean reciprocal rank) is the metric used in TREC Q&A evaluation. In addition, we designed a set of metrics that are more appropriate for automated evaluation.

In this paper, we will describe our NSIR system; then we will introduce our metrics for automatic evaluation of Q&A systems; correlations between different metrics will be shown, followed by a discussion of related work.

## 2. The NSIR System

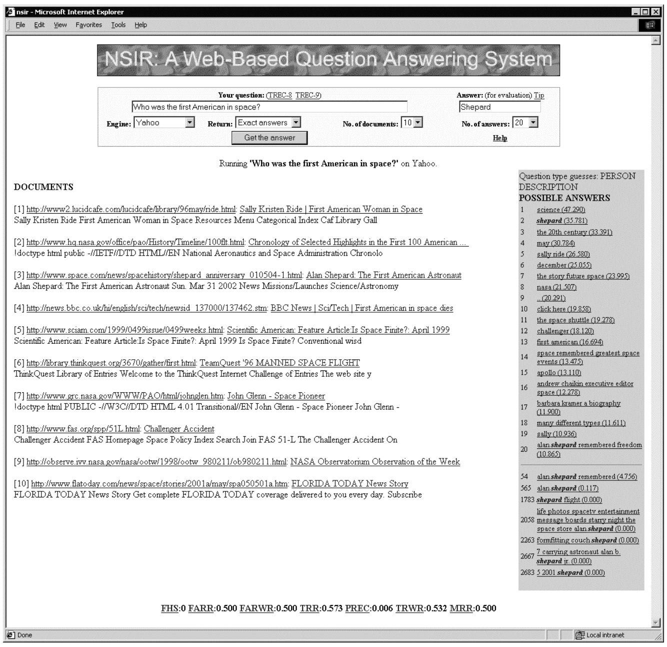
NSIR (pronounced "Answer") is a web-based question answering system under development at the University of Michigan. It utilizes existing web search engines to retrieve related documents on the web. Once NSIR gets the hit list returned by the search engine, it processes the top ranked documents and extracts a number of potential answers.

Potential answers are ranked according to a set of techniques before they are returned to NSIR users, including the proximity algorithm and probabilistic phrase ranking (Radev et al., 2002). The proximity algorithm is based on the closeness in text between the question words and the neighbors of each phrasal answer. A potential answer that is spatially close to question words gets a higher score than one that is farther away. Probabilistic phrase ranking takes expected answer type into consideration. Each phrase is assigned a probability score indicating the extent to which the phrase matches the expected answer type with respect to the part-of-speech tag sequences.

The web interface of NSIR allows users to choose from a list of search engines such as Yahoo, All the Web, Excite, etc. Users can also specify the number of documents to be processed, and the number of answers to be returned. For evaluation, NSIR allows users to specify the expected answer; after each run, NSIR uses the given answer to compute a set of evaluation metrics for current results.

Figure 1 shows the page returned by the NSIR system for the question "Who was the first American in space?". For evaluation purposes, the answer "Shepard" is specified in the answer box. The links to the top 10 documents as returned by Yahoo are displayed on the bottom left. Top 20 answers extracted by NSIR are shown on the right, each with a score of confidence. The correct answers are highlighted. Below the first 20 answers, NSIR also displays the correct ones which the system failed to rank within the first 20 positions. A set of evaluation results are displayed on the bottom of the page. These evaluation metrics will be discussed in next section.

Each answer has a link to its contexts in the original documents. Notice that each answer could be extracted from several documents. Figure 2 shows the page after the first correct answer "shepard" is clicked. The contextual information for the clicked answer is displayed on the left. Users can therefore justify the answer from the contexts. Links to full text of the original documents are also available on this page.


nsir - Microsoft Internet Explorer
File
Edit
View
Favorites
Tools
Help
NSIR: A Web-Based Question Answering System
Get the answer
Help
Done
Local intranet

**Your question:** (TREC-8 TREC-9) Who was the first American in space?

**Answer:** (for evaluation) Tip Shepard

**Engine:** Yahoo **Return:** Exact answers **No. of documents:** 10 **No. of answers:** 20

Running **'Who was the first American in space?'** on Yahoo.

DOCUMENTS

[1] http://www2.lucidcafe.com/lucidcafe/library/96may/ride.html: Sally Kristen Ride | First American Woman in Space
Sally Kristen Ride First American Woman in Space Resources Menu Categorical Index Caf Library Gall

[2] http://www.hq.nasa.gov/office/pao/History/Timeline/100flt.html: Chronology of Selected Highlights in the First 100 American ...
!doctype html public -//IETF//DTD HTML//EN National Aeronautics and Space Administration Chronolo

[3] http://www.space.com/news/spacehistory/shepard_anniversary_010504-1.html: Alan Shepard: The First American Astronaut
Alan Shepard: The First American Astronaut Sun. Mar 31 2002 News Missions/Launches Science/Astronomy

[4] http://news.bbc.co.uk/hi/english/sci/tech/newsid_137000/137462.stm: BBC News | Sci/Tech | First American in space dies

[5] http://www.sciam.com/1999/0499issue/0499weeks.html: Scientific American: Feature Article:Is Space Finite?: April 1999
Scientific American: Feature Article:Is Space Finite?: April 1999 Is Space Finite? Conventional wisd

[6] http://library.thinkquest.org/3670/gather/first.html: TeamQuest '96 MANNED SPACE FLIGHT
ThinkQuest Library of Entries Welcome to the ThinkQuest Internet Challenge of Entries The web site y

[7] http://www.grc.nasa.gov/WWW/PAO/html/johnglen.htm: John Glenn - Space Pioneer
!doctype html PUBLIC -//W3C//DTD HTML 4.01 Transitional//EN John Glenn - Space Pioneer John Glenn -

[8] http://www.fas.org/spp/51L.html: Challenger Accident
Challenger Accident FAS Homepage Space Policy Index Search Join FAS 51-L The Challenger Accident On

[9] http://observe.ivv.nasa.gov/nasa/ootw/1998/ootw_980211/ob980211.html: NASA Observatorium Observation of the Week

[10] http://www.flatoday.com/news/space/stories/2001a/may/spa050501a.htm: FLORIDA TODAY News Story
FLORIDA TODAY News Story Get complete FLORIDA TODAY coverage delivered to you every day. Subscribe

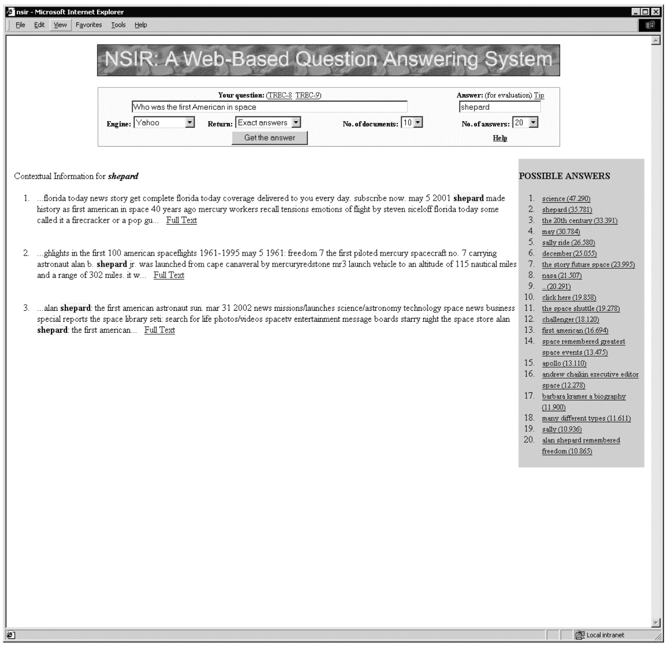
Question type guesses: PERSON DESCRIPTION
**POSSIBLE ANSWERS**

1 science (47.290)
2 ***shepard*** (35.781)
3 the 20th century (33.391)
4 may (30.784)
5 sally ride (26.580)
6 december (25.055)
7 the story future space (23.995)
8 nasa (21.507)
9 .. (20.291)
10 click here (19.858)
11 the space shuttle (19.278)
12 challenger (18.120)
13 first american (16.694)
14 space remembered greatest space events (13.475)
15 apollo (13.110)
16 andrew chaikin executive editor space (12.278)
17 barbara kramer a biography (11.900)
18 many different types (11.611)
19 sally (10.936)
20 alan ***shepard*** remembered freedom (10.865)

54 alan ***shepard*** remembered (4.756)
565 alan ***shepard*** (0.117)
1783 ***shepard*** flight (0.000)
2058 life photos spacetv entertainment message boards starry night the space store alan ***shepard*** (0.000)
2263 formfitting couch ***shepard*** (0.000)
2667 7 carrying astronaut alan b. ***shepard*** jr. (0.000)
2683 5 2001 ***shepard*** (0.000)

FHS:**0** FARR:**0.500** FARWR:**0.500** TRR:**0.573** PREC:**0.006** TRWR:**0.532** MRR:**0.500**

Figure 1: Run the question “Who was the first American in space?” on NSIR


nsir - Microsoft Internet Explorer
File
Edit
View
Favorites
Tools
Help
NSIR: A Web-Based Question Answering System
Get the answer
Help
Local intranet

**Your question:** (TREC-8 TREC-9) Who was the first American in space

**Answer:** (for evaluation) Tip shepard

**Engine:** Yahoo **Return:** Exact answers **No. of documents:** 10 **No. of answers:** 20

Contextual Information for ***shepard***

1. ...florida today news story get complete florida today coverage delivered to you every day. subscribe now. may 5 2001 **shepard** made history as first american in space 40 years ago mercury workers recall tensions emotions of flight by steven siceloff florida today some called it a firecracker or a pop gu... Full Text
2. ...ghlights in the first 100 american spaceflights 1961-1995 may 5 1961: freedom 7 the first piloted mercury spacecraft no. 7 carrying astronaut alan b. **shepard** jr. was launched from cape canaveral by mercuryredstone mr3 launch vehicle to an altitude of 115 nautical miles and a range of 302 miles. it w... Full Text
3. ...alan **shepard**: the first american astronaut sun. mar 31 2002 news missions/launches science/astronomy technology space news business special reports the space library seti: search for life photos/videos spacetv entertainment message boards starry night the space store alan **shepard**: the first american... Full Text

POSSIBLE ANSWERS

1. science (47.290)
2. shepard (35.781)
3. the 20th century (33.391)
4. may (30.784)
5. sally ride (26.580)
6. december (25.055)
7. the story future space (23.995)
8. nasa (21.507)
9. .. (20.291)
10. click here (19.858)
11. the space shuttle (19.278)
12. challenger (18.120)
13. first american (16.694)
14. space remembered greatest space events (13.475)
15. apollo (13.110)
16. andrew chaikin executive editor space (12.278)
17. barbara kramer a biography (11.900)
18. many different types (11.611)
19. sally (10.936)
20. alan shepard remembered freedom (10.865)

Figure 2: Contextual information for the first correct answer “Shepard”

## 3. Evaluation of Q&A Systems

### 3.1. Evaluation Metrics

Traditional information retrieval systems use recall and precision to measure performance. For Web-based systems, user effort should also be one of the evaluation criteria. We have developed the following metrics to address recall, precision, user effort in Web-based Q&A systems:

- FHS, First Hit Success.
  If the first answer returned by the system answers the question correctly, the FHS is 1. Otherwise the FHS is 0. For a user who relies solely on the Q&A system for answers, the user will accept the first answer returned by the system as the answer to the question. If we only consider the first answer to each question on a set of questions and assume the Web contains answers to all the questions, then the average of FHS represents the recall ratio of a Q&A system. FHS is similar to the metric used in this year’s TREC Q%A evaluation, TREC11.

- FARR, First Answer Reciprocal Rank.
  For example, if the third answer extracted by NSIR is the highest ranked correct answer, then FARR is 1/3. If no answers are correct, then FARR is 0. A user may be able to recognize the correct answer in a list

of suggested answers. A user can also find the correct answer by reading the supporting documents to each suggested answer. In both cases, the order of answers returned by the system directly affects the user's effort needed. FARR addresses the user effort criterion.

- FARWR, First Answer Reciprocal Word Rank.
  For example, for the question "In which city is Jeb Bush's office located?" if the first answer is "Florida Capital Tallahassee", then the correct answer starts from the third word, thus the FARWR is 1/3. FARWR represents the number of words a user has to read before reaching the correct answer. Humans read by saccades, which means a few words at a time. For short answers a user can read one answer in one saccade, where FARR is a fair representation of user's time-based effort. For longer answers, however, FARWR better represents a user's time-based effort.

- TRR, Total Reciprocal Rank.
  Sometimes there is more than one correct answer to a question. A user can be more certain about the correct answer, if the correct answer occurs multiple times in the list of answers provided by the system. Clearly in these cases it is insufficient to only consider the first correct answer in evaluations. TRR takes into consideration all correct answers provided by the system, and assigns a weight to each answer according to its rank in the returned list. For example, if both the 2nd and the 4th answers are correct, the TRR is $1/2 + 1/4 = 3/4$. TRR affects the likelihood for a user to retrieve the correct answer from the system. From an economic perspective, TRR reflects the diminishing returns in a user's utility function.

- TRWR, Total Reciprocal Word Rank.
  Similarly to TRR, TRWR reflects the diminishing returns in a user's utility function, and also takes a user's word-scanning effort into consideration. For example, if the first correct answer starts from the 5th word and the second correct answer starts from the 20th word, then TRWR is $1/5 + 1/20 = 0.25$.

- PREC, Precision.
  Precision is computed as the total character length of all correct answers divided by the total character length of all answers provided by the system. PREC reflects the percentage of useful content in the list of answers provided by a Q&A system.

Different Q&A systems may return different numbers of answers. A Q&A system may need to provide different numbers of answers in different situations, for example, when providing content to a browser versus a cellular phone. To ensure that we are evaluating these Q&A systems on the same ground, we have developed parameterized metrics based on some of the above metrics. For example, TRR(5) means Total Reciprocal Rank considering top 5 answers only.

### 3.2. Correlation Analysis

Each metric represents a different feature of Q&A systems. To study the consistency of different metrics, we performed a correlation analysis for some metrics. We ran 200 TREC 8 questions on the NSIR system and got the TRR, TRWR, PREC and MRR scores for each individual question. Table 1 shows the correlations between TRR, TRWR, PREC, and MRR. MRR is the metric used in TREC evaluations and will be discussed in next section.

| | TRR | TRWR | PREC | MRR |
|---|---|---|---|---|
| TRR | | | | |
| TRWR | .989** | | | |
| PREC | .367** | .332** | | |
| MRR | .974** | .981** | .342** | |

Table 1: Pearson's correlation between pairs of metrics.
**: Correlation is significant at the .01 level

The correlations given in table 1 are Pearson's correlations, which reflect the degree of linear relationship between two measures. It ranges from +1 to -1. A correlation of +1 means that there is a perfect positive linear relationship between measures.

As can be seen from table 1, the correlations within each pair are all statistically significant. This indicates consistency among different measures. Precision, though having significant correlations with other metrics, shows the weakest relationships across the table. This result suggests that precision might be a poor performance measure for web-based Q&A systems. (Kwok et al., 2001) also states that precision is an inappropriate measure in Q&A contexts.

The strongest correlation, 0.989, is found between TRR and TRWR. This is not surprising because the answers returned by NSIR are in phrasal form, normally very short. So the user effort measured in words should not be significantly different from the user effort measured in number of answers. This fact suggests that when the answers are in short phrase form, the metrics TRR and TRWR are interchangeable.

## 4. Discussion

In TREC evaluations, each question gets a score equal to the reciprocal of the rank of the first correct answer. For instance, if a question gets the first correct answer in the 2nd place, it will receive a score of $1/2 = 0.5$; a question gets 0 if none of the five returned answers are correct. The mean of the individual question's reciprocal ranks (MRR) is then computed as a measure of each submission (Voorhees and Tice, 2000). The TREC metric is one special parametric case of FARR (First Answer Reciprocal Rank) that we have implemented. The TREC metric is the same as FARR(5).

(Voorhees and Tice, 2000) points out some drawbacks of the above metric used by TREC. Q&A systems get no extra credit when they retrieve multiple correct answers. The possible scores for each question can only take values from a very limited range, namely only six values (0, .2, .25, .33, .5, 1), so it is inappropriate to do parametric statistical significance tests for this task.

(Radev et al., 2002) uses total reciprocal document rank (TRDR). For example, if the system has retrieved 10 documents, of which the second, eighth, and tenth contain the correct answer, TRDR is $1/2 + 1/8 + 1/10 = .725$. Using TRDR rather than the metric employed in TREC, they are able to make finer distinctions in performance. Our TRR, Total Reciprocal Rank, and TRWR, Total Reciprocal Word Rank, are similar to their TRDR metric.

(Kwok et al., 2001) defines the "word distance" metric to measure user effort in question answering systems. In short, the word distance measures how much work it takes a user to reach the first correct answer. They assume that answers are given in short summaries of the documents from which these summaries are extracted. They define word distance as a dependent variable of the number of snippets before the one that has correct answer and the number of words before the answer in the document. Our TRWR (Total Reciprocal Word Rank) also measures the user effort except that we do not consider the number of words that a user has to read in the original documents.

(Wu et al., 2002) discusses evaluations of answer-focused summaries. Three criteria are proposed in order of importance: Accuracy, Economy and Support. They also propose four facets to evaluate accuracy and economy, which are whether a question is answered, summary length in characters, hit rank of first answer, and word rank of first answer, respectively. Our evaluation scheme addresses these aspects. Whether a question is answered can be derived from our FARR (First Answer Reciprocal Rank) metric. Hit rank and word rank of first answer are represented by our FARR and FARWR (First Answer Reciprocal Word Rank). Instead of measuring summary lengths or answer lengths, we use PREC (Precision) to measure the percentage of key content.

## 5. Conclusion

Manual evaluations become prohibitively expensive when Q&A systems are scaled to the web. This paper proposes a set of metrics for evaluating web-based Q&A systems. In addition to MRR, the TREC evaluation metric, we introduce first hit success (FHS), first answer reciprocal rank (FARR), first answer reciprocal word rank (FARWR), total reciprocal rank (TRR), total reciprocal word rank (TRWR), and precision (PREC). The correlation analysis for TRR, TRWR, MRR and PREC suggests that precision may be an arguably inappropriate performance measure. Our metrics address the drawback of MRR, are therefore more appropriate for automatic evaluation of web-based Q&A systems.

## 6. References

Eugene Agichtein, Steve Lawrence, and Luis Gravano. 2001. Learning search engine specific query transformations for question answering. In *the Proceedings of the 10th World Wide Web Conference (WWW 2 001)*, Hong Kong.

Cody Kwok, Oren Etzioni, and Daniel S. Weld. 2001. Scaling question answering to the web. In *the Proceedings of the 10th World Wide Web Conference (WWW 2001)*, Hong Kong.

J. Prager, D. Radev, E. Brown, and A. Coden. 1999. The use of predictive annotation for question answering in trec 8. In *NIST Special Publication 500-246:The Eighth Text REtrieval Conference (TREC 8)*, pages 399–411.

Dragomir R. Radev, Weiguo Fan, Hong Qi, and Amardeep Grewal. 2002. Probabilistic question answering from the web. In *The Eleventh International World Wide Web Conference*, Honolulu, Hawaii, May.

Ellen Voorhees and Dawn Tice. 2000. The TREC-8 question answering track evaluation. In *Text Retrieval Conference TREC-8*, Gaithersburg, MD.

Harris Wu, Dragomir Radev, and Weiguo Fan. 2002. Towards better answer-focused summarization. submitted to SIGIR 2002, August.