Context-Based Question-Answering Evaluation

Elizabeth D. Liddy  
*Syracuse University*, liddy@syr.edu

Anne R. Diekema  
*Utah State University*, diekemar@syr.edu

Ozgur Yilmazel  
*Syracuse University*, oyilmaz@syr.edu

Follow this and additional works at: [http://surface.syr.edu/istpub](http://surface.syr.edu/istpub)

Part of the [Library and Information Science Commons](http://surface.syr.edu/istpub) and the [Linguistics Commons](http://surface.syr.edu/istpub)

**Recommended Citation**


This Conference Document is brought to you for free and open access by the School of Information Studies (iSchool) at SURFACE. It has been accepted for inclusion in The School of Information Studies Faculty Scholarship by an authorized administrator of SURFACE. For more information, please contact surface@syr.edu.
ABSTRACT
In this poster, we will present the results of efforts we have undertaken to conduct evaluations of a QA system in a real world environment and to understand the nature of the dimensions on which users evaluate QA systems when given full reign to comment on whatever dimensions they deem important.

Categories and Subject Descriptors
H.3.4 Systems and Software Performance Evaluation

General Terms
Measurement

Keywords
Question-answering systems, question taxonomies, question understanding, real-time systems.

1. INTRODUCTION
While research on question-answering (QA) systems has continuously advanced the quality of such systems (Voorhees, 2000), the evaluation of QA systems has not made similar advances. The standard evaluation paradigm is based on the well-known test collection paradigm developed in years of information retrieval research. And while the issue of whether this paradigm is appropriate for question-answering systems was addressed in a report on the TREC QA track (Voorhees & Tice, 2000) the perspective was from that of the controlled TREC environment, where assessors are hired to make the relevance decisions, rather than from the perspective where actual users are the ones who ask the questions based on real information needs.

In recent years, we have developed (or specialized) QA systems for a range of environments and have recognized that the basis on which individuals evaluate such systems differs quite dramatically from the test-collection based evaluation with which we are all most familiar. And while we are not saying that what we have learned in these environments will necessarily hold in all QA environments, we believe that our findings will provide informative discussion points and serve to advance all of our understandings of evaluation of QA.

2. QA SYSTEM ENVIRONMENT
The focus of this poster is our eQuery capability as adapted for use in the Knowledge Acquisition and Access System (KAAS). It was developed for use in a NASA, New York State, and AT & T funded collaborative learning environment, the Advanced Interactive Discovery Environment for Engineering Education (AIDE) for undergraduate students from two universities majoring in aeronautical engineering. While students are working within the AIDE, either in a group or alone, they can ask questions on any topic related to the course. The collection against which the questions are asked consists of textbooks, technical papers, and websites that have been pre-selected by the team of engineering professors teaching the course for their relevance and pedagogical value. This system has been explained in detail elsewhere (Diekema et al, In Press), and can be considered a fairly standard QA system architecture in which rather sophisticated NLP techniques are used, and effort has been focused on the user’s experience as well as the standard issues of precision and recall.

Since the environment in which the QA system is used is itself an experimental learning environment, it provided us the opportunity and permission to seek and obtain extensive user feedback. Our evaluations consisted both of logged questions asked by the student users of KAAS and end-of-semester student surveys for two different semesters. Not surprisingly, the logged questions of this real user group did not closely resemble questions from the more standard QA evaluation test collections. Rather, we found the students to utilize both a broader range of question types and to ask more complex, multi-faceted questions, including the following question types: quantification, conditional, yes/no, alternative, why, how, and definition questions (Liddy et al, 2003) Example questions include, “Are aerogels rigid enough to sustain the compression inflicted on it by the shell of a sandwich panel-type Thermal Protection System when under the influence of an applied load?” (yes/no). “How difficult is it to mold and shape graphite-epoxies compared with alloys or ceramics that may be used for thermal protective applications?” (alternative) and “In
preliminary stages of product fabrication, is it common practice to
first test highly simplified scenarios?" (conditional).

What we will focus on in this poster is the results of the open­
ended student surveys which were conducted at each site at the
end of two different semesters. We believe that these open-ended
surveys enabled us to learn first-hand about the dimensions of
QA system performance that the users themselves found worthy of
note. And while answer correctness does matter in a QA system,
we believe that these findings indicate that in the context of the
information need that brought about the question in the first place,
there are other dimensions of importance to the user.

3. DIMENSIONS OF USER EVALUATION

The KAAS survey was part of a larger scale, cross-university
course evaluation which looked at the students’ perceptions of
distance learning, collaboration at a distance, the collabora­tive
software package, the KAAS, and each participating faculty
member. While there was some structure and guidance in the user
survey of the QA system, it was minimal and the survey is mainly
characterized by the open nature of the responses. There were 25
to 30 students participating in each full course survey, but since
we do not have the actual surveys that were turned in, we are not
certain as to exactly how many students completed the survey
section on the KAAS. However, it appears that most, if not all of
the students provided feedback.

Given the free text nature of the responses, it was decided that the
three researchers would do a content analysis of the responses and
independently derive a set of evaluation dimensions that they
detected in the students’ responses. In follow-up discussion, we
shared our dimensions, removed duplicates, selected the most
appropriate phrasing for each distinct dimension, and produced a
hierarchical classification structure which covered the content of
the survey comments. This schema is presented in a hierarchy
below.

1 System Performance
   1.1 Speed
   1.2 Availability / reliability / upness
2 Answers
   2.1 Completeness
   2.2 Accuracy
   2.3 Relevance
   2.4 Applicability to task / utility / usefulness
3 Database Content
   3.1 Authority / provenance / Source quality
   3.2 Scope / extensiveness / coverage
   3.3 Size
   3.4 Updatedness
4 Display (UI)
   4.1 Input
      4.1.1 Question understanding / info need understanding
      4.1.2 Querying style
      4.1.2.1 Question
   4.1.2 Keywords
   4.1.3 Question formulation assistance
      4.1.3.1 Spell Checker
      4.1.3.2 Abbreviation recognition
   4.2 Output
      4.2.1 Organization
      4.2.2 Feedback Solicitation
5 Expectations
   5.1 Googleness

As is evident from the different dimensions, a QA system needs to
be evaluated in context. A meaningful and successful system can
only be created if it is situated in the context in which it is used.
Hence, a QA evaluation has to be situated in the task, domain, and
user community for which the system is developed. We believe
that the evaluation should be driven by the dimensions identified
by the users as important: system performance, answers, database
content, display, and expectations. How many seconds does it take
to answer a question? Is the system available at all times? How
relevant are the answers to the task at hand? How complete is the
domain coverage of the database? How easy is the system to use?

4. FUTURE WORK

Having extracted these dimensions from an examination of the
responses of users who evaluated their interactions with and
output from a QA systems, we plan to have un-involved
individuals utilize this schema to code the nature of the evaluation
dimensions of a new set of comments from users.

5. ACKNOWLEDGMENTS

Funding for this research has been jointly provided by NASA,
NY State, and AT&T.

6. REFERENCES

Gaithersburg, MD, NIST, 2000.