

Why are People Asking these Questions? :
A Call for Bringing *Situation* into Question-Answering System Evaluation

Elizabeth D. Liddy
Center for Natural Language Processing
School of Information Studies
Syracuse University
Syracuse, New York 13210
315-443-5484 (v) 315-443-5806 (f)
liddy@syr.edu; www.cnlp.org

Introduction

I believe that in order for the field of Question-Answering (QA) to evolve to the stage where it will provide maximum utility, the environment in which a QA system is to be used should become a parameter in the evaluation of QA systems. That is, the current evaluation paradigm is becoming restrictive and may well push development in a single direction that will not produce systems that will prove useful in multiple environments. Even a quick review of the potential scenarios in which QA can be utilized suggests two key facts: 1) what is considered '*a useful answer*' in one context might not be useful in another, and; 2) currently permissible methods that systems can utilize to determine correct answers are not feasible in many real world QA environments. This paper will advance this position and suggest a range of situational dimensions that should be considered for inclusion in the QA evaluation roadmap.

QA Evaluation

While there was significant early research in Question Answering in the fields of logic and linguistics (Belnap, 1963; Belnap & Steel, 1976), automatic QA was first focused on in a large-scale evaluation framework in the TREC Conferences, beginning with TREC-8 in 1999 (Voorhees & Tice, 1999). The paradigm established in TREC-8 and continued in the next two TREC Conference QA tracks is simple fact-based, short-answer questions. Initially, answer strings were limited to either 50 or 250 bytes depending on the run type. In TREC-10, the 250 byte condition was eliminated and the list task was added. The list task consisted of 25 questions which specified the number of unique responses to be retrieved, e.g., "*What four countries are the top producers of wheat in the world?*" All other parameters of the main QA task remained the same (Voorhees, 2001).

Discussion at the TREC 2001 Workshop on QA intimated that the QA track in TREC 2002 will accept as correct only fragments which contain the minimal answer to the question. Any explanatory text, even if within the 50 byte limit, will cause the answer to be marked as incorrect. Additionally, the practice introduced in TREC 2001 of a system first determining the most frequent potential answer by searching the web, and then finding a document in the TREC collection which contained that answer fragment will continue to be allowed.

Potential Problems

The need for a more refined evaluation of answer strings was evident from some sample answers shown at the Workshop as they contained text that was non-contributory to the answer and just happened to contain the correct answer that had been provided to the relevance assessors. However, this was not always true. In some instances, the additional text can be argued to have provided useful supportive or confirmatory information. The potential problem I see in the

requirement of a minimal answer is that this evaluation paradigm, which does not permit the inclusion of supporting information that might be useful in some QA scenarios, will foster the development of systems which will be useful in only a subset of the contexts in which QA systems are truly needed.

Furthermore, the decision to allow systems to utilize redundancy on the web to select answers (Brill et al, 2001) will also foster methods that may not be useable in many QA environments. It is highly unlikely that the redundancy approach will transfer to QA systems that are developed for specialized resource environments. While the simple factoid questions for which multiple instances of responses can be found on the web have been the norm in the QA track, this is not typical in other environments for which QA systems provide great utility.

While the existing QA evaluation scenario has utilized very simple questions, has focused on a narrow definition of length of useful answer to the exclusion of other issues, and has permitted the use of a method of determining an answer which will not work in other than the simple query environment, some QA system builders have begun to call for an evaluation paradigm that considers dimensions above and beyond correctness (Breck et al, 2000). We strongly agree with this view and encourage the discussion of a broader evaluation paradigm for the QA Roadmap that will take into account the wide range of environments in which QA is already providing an essential service.

Range of Possible QA Environments

Consider the three following real-life environments for which we have developed QA systems. In each of these environments, the collection, the type of queries, how the system determines answers, and what constitutes an acceptable answer formulation for the user vary dramatically.

1. Scientific Questions from Undergraduate Students

We have developed a QA system (Liddy, 2001) with funding from NASA and AT&T for use within a collaborative learning environment for undergraduate students from two universities majoring in aeronautical engineering who are taking courses that are taught within the AIDE (Advanced Interactive Discovery Environment for Engineering Education). The students are able to ask questions and quickly get answers in the midst of their hands-on collaborations within the AIDE. The collection against which the questions are searched consists of textbooks, technical papers, and websites that have been pre-selected for their relevance and pedagogical value. We are currently working towards the addition of transcripts of class lectures and accompanying power point slides. The students' questions are not typically simple factoid questions, but tend more towards 'Why' and 'How' questions and require more than bare answers, such as:

- *How do ablating materials minimize energy conducted into a RLV?*
- *What are the changes made to the design of the Shuttle SRM since the Challenger Accident?*
- *How are malfunctions detected for the pitch and yaw gimbal actuators of the space shuttle OMS engines?*

Answers are provided in increasing window sizes, allowing the student to gradually expand the amount of text by mouse-clicking from 'answer-providing passage', to paragraph (s) containing the 'answer-providing passage' to full document(s) containing the 'answer-providing passage'. The system is currently undergoing user testing. The U S Army has funded us to create a similar capability for the students in the Army's intelligence training programs. They share NASA's

vision that work in the future will consist largely of virtual collaborative situations in which questions that arise will need to be answered electronically from selected sources.

2. Citizens' Search for Statistical Information

Naïve users need to access statistical information, but frequently do not have the sophisticated understanding required in order to translate their information needs into structured database queries using the controlled vocabulary which are currently required. However, these users can articulate quite straightforwardly in their own terms what they are looking for. One approach to satisfying the masses of citizens with needs for statistical information is to automatically map their natural language expressions of their information needs into the metadata structure and terminology that defines and describes the content of statistical tables. To accomplish this goal, under funding from NSF's Digital Government Initiative (<http://istweb.syr.edu/~tables/>), we undertook an analysis of 1,000 user email queries seeking statistical information from federal agencies which provide internet access to their statistical tables. Our goal was to understand the dimensions of interest in naïve users' typical statistical queries, as well as the linguistic regularities that could be captured in a statistical-query sublanguage grammar. We developed an ontology of query dimensions using this data-up analysis of the queries and extended the ontology where necessary with values from actual tables. We proceeded to develop an NLP statistical-query sublanguage grammar that enabled the system to semantically parse users' queries and produce a template-based internal query representation which was then mapped to the tables' metadata, in order to retrieve relevant tables which were displayed to users with the relevant cell's value highlighted (Liddy & Liddy, 2001). Typical queries were:

- *I am trying to find the percentage of women in the workforce from the years 1900 to 1998.*
- *I want to know how many people worked for small businesses last year.*
- *What was the average amount of time women spent on housework per week in 1900; 1950; 1995?*

This project made it eminently clear that the situation predicts the nature of the questions, the resources searched, and the acceptable answer formulation.

3. Speech-based Inquiries in Travel and Tourism

In an exciting project in the commercial world, we worked with a speech understanding technology company to provide answers to travelers who were planning Caribbean vacations via interaction with a voice-activated system. While the business idea was well-researched, the current status of speech-understanding technology was not, and the corporation failed to pull off the application. However, I mention it here because it introduces a third and very different set of users, answer-providing resources, and answer formulation in which appropriate supporting detail is essential.

- *We're looking for a family resort in the Caribbean with baby sitting, other activities for a family with a one and three year old. Any suggestions?*
- *My fiancée and I were wondering if there was anywhere we could go in October that would not be extremely crowded, yet more secluded?*
- *When is the best time to go on a Caribbean Cruise - and do you recommend bring our 16 year-old so? He is very bright.*

Again this situation points out that evaluation needs to reflect an environment – we do not foresee that all questions will be ones that can be satisfied with short answers which are found redundantly present on the web. Requirements in this particular situation contradict the TREC QA evaluation requirement that evidence supporting the answer should not be provided.

Conclusion

We have found that the collection of documents that will be available for querying, the nature of queries generated by real users, as well as the breadth vs. narrowness of what constitutes a useful answer in each of these instances is not the same. Therefore, it would only seem appropriate that an evaluation should fully specify the user, the purpose for which they are asking their question, and the nature of an acceptable answer. These should be parameters that can be varied in QA evaluations. It is essential that the situational aspects be known so that the criteria provided to the human relevance assessors truly reflect what users in that particular context would require. Evaluations should be designed that simulate as closely as possible the dimensions of the context in which users will be posing their questions. Clearly the use of multiple scenarios would enhance the possibility that evaluation would lead to a range of QA systems, each defined by the parameters of the situation in which they are to be used.

References

- Belnap, N. D. (1963). An analysis of questions: Preliminary report. Scientific Report TM-1287. Santa Monica, CA.
- Belnap, N. D. & Steel, T. B. (1976). The logic of questions and answers. New Haven, CT., Yale University Press.
- Breck, E.J., Burger, J.D., Ferro, L, Hirschman, L., House, D., Light, M. and Mani, I. (2000). How to evaluate your question answering system every day...and still get real work done. Proceedings of Language Resources and Evaluation (LREC).
- Brill, E., Lin, J., Banko, M., Dumais, S. & A. Ng. (2001). Data-Intensive question answering. Notebook Proceedings of the Text Retrieval Conference. Gaithersburg, MD: NIST Special Publications.
- Liddy, E.D. (2001). Breaking the Metadata Generation Bottleneck. Joint Conference on Digital Libraries. Roanoke, VA., June 25, 2001.
- Liddy, E.D. & Liddy, J.H. (2001). An NLP approach for improving access to statistical information for the masses. Proceedings of the Federal Committee on Statistical Methodology Research Conference. Arlington, VA.
- Voorhees, E. and Tice, D. (1999). The TREC-8 question answering track evaluation. In Voorhees, E. and Harman, D. Proceedings of the Eighth Text Retrieval Conference. Gaithersburg, MD: NIST Special Publications.
- Voorhees, E. (2001). Overview of the TREC 2001 question-answering track. In Voorhees, E. and Harman, D. Notebook Proceedings of the Text Retrieval Conference. Gaithersburg, MD: NIST Special Publications.