

Chapter 11

Finding Answers to Complex Questions

*Anne R. Diekema, Ozgur Yilmazel, Jiangping Chen,
Sarah Harwell, Lan He, & Elizabeth D. Liddy*

11.1 Introduction

While there was significant early research in question-answering in the fields of logic and linguistics (Belnap 1963, Belnap and Steel 1976), automatic question-answering research has been largely driven by the Text Retrieval Conference (TREC), cosponsored by the National Institute of Standards and Technology (NIST) and the Defense Advanced Research Projects Agency (DARPA). The purpose of TREC is to support research in the area of information retrieval by organizing yearly large-scale system evaluations of a variety of retrieval related tasks (tracks). Although progress has been made since question answering was first added as a track at TREC in 1999 (Voorhees 2000), the research has largely converged on shorter, fact-based, general domain questions, many of whose answers can be found using a redundancy of potential answers on the Web. While these simpler approaches may work well for factoid questions, having a relatively successful QA system tailored to the TREC question-answering task (Diekema et al. 2001) does not necessarily ensure success in question-answering applications outside TREC.

As question answering fully emerges as a field in its own right, a broader definition of QA is developing which encompasses a wider range of question types and answer types than those represented in the TREC QA paradigm. Important in this definition is the notion of context, i.e., who is the user?...what is the task they are attempting to accomplish when they ask a question?...what constitutes a useful answer to them?...what format should the answer be presented in?...might the most useful answer be an unsupported factoid?...or an answer-providing passage?...or the full document?...or a cell value highlighted in a table? As the field advances from the somewhat homogenous approaches

initially adopted across QA systems (Hirschman and Gaizauskas 2001), the lessons learned from deployment of QA systems into real user environments are providing insights into requirements for next generation QA systems.

In this chapter, we motivate one potential type of future QA system that deals with questions more complex than simple factoid questions and which provides answers with their supporting context. Our approach is based on the issues we faced when developing and delivering a QA system to deal with real time questions in the domain of RLVs within the larger field of aerospace engineering. This particular domain, the actual users of the system, and the questions asked, all demanded a change in our question-answering strategy. First, the chapter will present background on the project that provided the context and a description of the system that was deployed. Next, the chapter analyzes the questions put to the system by the users and discusses the implications that this analysis and the user evaluation study had on our design of a QA system of the future.

11.2 Background

We have developed a QA system (Liddy 2002) with funding from NASA and AT&T for use within a collaborative learning environment for undergraduate students majoring in aeronautical engineering at Cornell University and Syracuse University. The students are taking courses that are taught within the advanced interactive discovery environment for engineering education (AIDE) environment. The students are able to ask questions and quickly get answers in the midst of their hands-on collaborations within the AIDE or while working independently. The collection against which the questions are asked is comprised of textbooks, technical papers, and websites that have been pre-selected by the engineering faculty of both universities for their relevance and pedagogical value.

The students' questions are not typically simple factoid questions, but tend to be more complex and require more than bare answers, such as: What are the changes made to the design of the Shuttle SRM since the *Challenger* accident?

The system provides up to 20 short answers on the answer page. The student can then click on a link that provides access to the full document. In case the link is dead, or the student is otherwise having trouble accessing the page, a cached version is also provided. The system has undergone a first round of user testing, the results of which are reported later in this chapter.

11.3 System Overview

The architecture of the current CNLP AIDE QA system (see unshaded portions

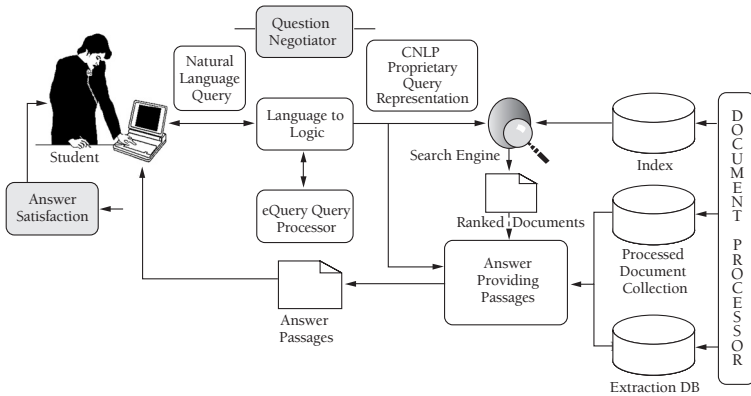


Figure 11.1. System Architecture

in figure 11.1) consists of four modules: (1) document processing; (2) language-to-logic (L2L); (3) search engine, and (4) answer providing passages. Given that the instructors want to vet the documents that will be used to answer the students' questions, it is possible to do document processing offline. When a user submits a question to the system, the question is first sent to the L2L module, which generates the L2L query representation and identifies the question focus. The search engine module then searches the index and returns the top 50 relevant passages. At last, the L2L query representation, question focus, and the retrieved passages are passed to the answer providing passages module, which returns the top 20 most relevant answer passages.

Three months into the project, we conducted an informal evaluation of the system. Twenty-seven students responded. While many students expressed satisfaction with the system and the answers to their questions, half of the negative comments mentioned the need to improve the accuracy of our answers. Although our system contained vetted documents that were highly relevant to their questions, the students perceived Google as their preferred system. While preference for the system that has been in use longer than the system being introduced is a typical finding, we still wanted to better understand the causes of this evaluation result. We began to analyze: (1) the nature of the students' questions, (2) the kinds of answers the students expected and wanted, and (3) how their questions and expected answers differed from the TREC questions of our past experience.

11.4 Analysis of NASA Questions

Characteristics of NASA Questions. For our analysis, we closely examined 342 questions that were asked of the system by students in the aeronautical engi-

neering program. This analysis found these questions to be similar in language usage to scientific writing generally. They are: Objective-personal pronouns seldom appear in the questions, and even if they do, they are not very useful in representing the semantics of the questions; Plain-the adjectives and adverbs used are necessary modifiers and are used either to convey a certain feature or to specify a level; Precise-the questions require certain prepositional phrases to convey the temporal, spatial, or conceptual domain of an occurrence.

These questions were observed to present the following linguistic features:

1. A large number of domain-specific phrases, including Proper Noun phrases, common-noun phrases, and verb phrases.

The domain-specific noun phrases and verb phrases are essential for complete understanding of the meaning of the questions. A preliminary analysis shows that these noun phrases can be classified into 49 categories and the verb phrases can be classified into 41 categories. Examples of the classes for nouns phrases are system, application, reason, alternative, weight, part, structure, and risk. The classes for verb phrases include compositional, dependency, reference, tendency, examination, and alteration. These classifications of the noun and verb phrases are important for proper representation and identification of the key components of a question together with the relations among components of a question. They are also used for question focus identification, which will be discussed later.

2. There are clear linguistic patterns that can be used to categorize questions into classes. A question type classification is presented in table 11.1.
3. These questions are comparatively longer, with complex syntax containing several prepositional phrases as modifiers (see sample questions in table 11.1).

Question Type and Question Focus. To identify the focus of a question, the L2L Module first determines the question type (Chen et al. 2002). As shown in table 11.1, eight different question types emerged from the analysis of questions in this application. Some of the question types are the same as the TREC questions, but some are new types, such as yes/no questions and alternative questions. While there are a large number of Wh-questions, only a portion of them are simple factoid questions. To capture this, we divided the Wh-questions into simple and complex types. Table 11.1 also shows the distribution of the 342 questions. Each question type was identified based on lexical and/or syntactic information. As discussed above, noun phrases, verb phrases, and prepositional phrases in the questions were categorized into classes with attached semantic relations. For this classification, a domain expert was consulted for the definition of those phrases in students' questions with which our team was unfamiliar. A list of sample question focuses is shown in the third column of table 11.1. "Unknown" is assigned to questions for which the system cannot identify a proper question focus. These questions may need special processing or interaction with the user. The system approach to focus identification and phrase analysis is detailed in Diekema et al. (2001).

Type/Cat.	Definition / Frequency	Example / Sample Focus
Wh-	Questions starting with What, When, and Where. Some are simple factoid questions, others are quite complex.	<i>What are the downselect criteria for design of the thermal protection system for the second generation RLV?</i>
Wh-simple	80 (23.4%)	time, material, cost, part, position
Wh-complex	49 (14.3%)	process, comparison, feature, goal, application, method, mission, danger, design, unknown
Yes / No	Require a yes or no response, but may mask a complex inquiry	<i>Doesn't the simplification of the complex honeycomb design for the thermal protection system of a reusable launch Vehicle jeopardize the accuracy of results?</i>
Yes/No	51 (14.9%)	unknown
How	Require an explanation.	How difficult is it to mold and shape graphite-epoxies compared with alloys or ceramics that may be used for thermal protective applications?
How	48 (14.0%)	level, manner,
Quantification	Looking for a specific amount, such as cost, weight, number, maximum, volume, etc.	<i>What is the highest temperature the space shuttle undersurface experiences during its mission?</i>
Quantification	41 (12.0%)	number, money, time, weight, temperature, volume
Conditional	Inquiry indicates a condition that the answer needs to take into account. Indicated by phrases such as: <i>in addition to, aside from, other than, etc</i>	<i>Aside from contact of two tiles that can be damaging, are there any other reasons why insulating tiles on Reusable Launch Vehicles must be isolated from one another?</i>
Conditional	12 (3.5%)	Manner, unknown
Alternative	User provides several alternatives, one of which needs to be proven true, e.g. <i>A or B or C</i>	<i>Are Thermal Protection systems of spacecrafts commonly composed of one panel or a collection of smaller tiles?</i>
Alternative	11 (3.2%)	unknown
Why	Require an explanation	Why are all shear loads and twisting moments set to zero for the preliminary design phase of TPS?
Why	16 (4.7%)	reason
Definition	Looking for a formal or semi-formal definition of an element, process, material, etc.	What is a liquid metal?
Definition	34 (10.0%)	definition

Table 11.1. NASA AIDE Question Types, Distributions, and Focuses

11.4.1 Complexity of NASA questions

NASA questions differ from TREC questions in several respects. First, a NASA question is written in real time by a student whose question can be ambiguous, or dependent upon implicit knowledge that isn't explicitly stated in the question. Real-time questions are often hurried and rife with malformed syntax and spelling errors. Due to the nature of the subject area and the fact that the QA system supports an educational endeavor, the NASA questions are complex, needing complex answers or sometimes returning information from which the answer needs to be inferred by the student once the answer-providing passages are read.

For example the simple question: "How dsoe the shuttle fly?" (leaving aside its obvious typo) is so broad as to thoroughly confound even a reference librarian, let alone an automatic QA system. Does the student wish to know that the shuttle flies upside down, i.e., the physical orientation of the space shuttle as it flies? Or is the student looking for specifications related to the way the space

shuttle flies during its launch, the way it orbits when it arrives in space, or its re-entry into our atmosphere? Or does the student need information about the way the shuttle navigates?

The question “Do welding sites yield any structural weaknesses that could be a threat for failure?” is not specific, i.e., it doesn’t specify where or on what the welding sites are located. We can assume (as humans who know what course the students are taking) that the welding sites are probably located on the space shuttle, but QA systems of today do not typically make this type of assumption.

For another type of question that appears simple, e.g., “At what temperatures do liquid metals typically exist?” the QA system would typically look for “liquid metals,” plus a particular semantic class of verb, and a temperature (determined by the L2L focus analyzer). However, the actual answer is much more complex. Melting points depend on the type of liquid metal, with binary liquids having a narrow melting point (e.g., mercury -39 C), liquid metals made of heavier elements having a lower melting point (unspecified), and alkali metals having melting points below 200 C. This answer can be found in one document, but over several paragraphs, and it is still not the complete answer because it fails to specify the exact temperatures of liquid metals made of heavier melting points.

A fourth type of complex question requires comparison of two different elements from two different documents where the answer has to be synthesized by the actual questioner. For example: What advantages/disadvantages does an aluminum alloy have over Ti alloy as the core for a honeycomb design? It is unlikely that the system will find a particular sentence or paragraph that will answer this question thoroughly. This type of question requires higher order thought processes that utilize synthesis and analysis of existing information within the document collection. To assist the questioner, the system must be able to parse the question into different parts, e.g., return a passage on the strengths and weaknesses of “aluminum alloy” for honeycomb design, as well as return a passage that talks about the advantages and disadvantages of “Ti alloy” for honeycomb design. It will then be necessary for the system or the questioner to deduce an answer from the pieces of returned information.

Research has shown that “ill-formed queries” can be expected when dealing with users of any system. Reference librarians are familiar with the general patterns that can be expected when dealing with real-time questions. These problems fall into the following categories (Ross et al 2002): (1) Too broad a query; (2) Queries whose answer, even though correct, does not meet the unstated needs of the user; (3) A question that is based on a misunderstanding, either of the system or of the subject; (4) Ambiguous keywords; (5) A question based on erroneous details or memory that is wrong; and (6) A question containing a faulty assumption about the world or subject.

Any QA system that deals with real-time questions from users will need to develop strategies to deal with these types of questions successfully.

11.5 Human Query Negotiation

Reference librarians have successfully fielded ambiguous and open-ended questions for years using the reference interview to narrow a broad question and clarify an apparently unfocused question. The reference interview tries to “elicit from the user sufficient information about the real need to enable the librarian to understand it enough to being searching” (Ross et al. 2002). The question is “clarified, narrowed down, made more detailed and contextualized” (Ross et al. 2002). Real questions from real users are often “ill-formed” with respect to the information system; i.e., they do not match the structure of ‘expectations’ of the system.” (Ross et al. 2002). A reference interview translates the user’s query into a representation the library system can interpret correctly. Straightforward questions might be well served by the “focus” approach, but we believe the more complex questions need an alternative process to help the system interpret the question correctly. One possible solution for a QA system faced with broad or ambiguous questions is query clarification, where the system asks the questioner for more information in order to return better results. We posit that utilizing reference interview theory to provide a framework for automatic query negotiation between the system and the questioner will make system question answering more accurate and satisfying.

A reference interview has three parts: questioning the questioner, locating the answer, and returning the answer to the user (Bopp and Smith 2001). The reference interview typically begins by restating the question to the user in order to allow the patron to refine his or her thoughts and to ensure that the librarian has understood the query accurately. Based on the initial question, the librarian might respond with an open-ended question. For example, when the system is faced with a question as ambiguous as “How does a shuttle fly?” it might best respond, “What part of flying a shuttle would you like to explore?” thus allowing the patron to rephrase his or her question and make it more specific.

Since we are basing the automatic query negotiation on the model of the reference interview, which is an actual conversation between librarian and user, certain modifications of the question negotiation process are in order. It is important to verify first that the question has been understood correctly. After the information need has been established, follow-up interview questions by the system might be in order to further clarify what the user is looking for. Finally, once the answer has been provided, the user should be asked whether the answer is what he or she was looking for. If not, the process begins again. The implementation of each of these stages in the interactive QA process is further detailed in section 11.6 (see figure 11.3 for a sample interaction).

User: How are layers in TABI bonded together?
Computer: Do you mean How are layers in TABI bonded together?
User: Yes
Computer: Our databases expand *TABI to tailorable advanced blanket insulation*. Please press enter if this is correct.
User: Enter
Computer: Are you looking for an explanation? (Computer is alerted to this possibility using the classification schema shown in table 1).
User: Yes
Computer: The subject of your question is *layers in tailorable advanced blanket insulation*. The action is *bonded*. Is this correct?
User: Yes
Computer: We show the following synonyms for bonded:
 1 SHACKLE, chains, fetter(s), gyve(s), iron(s)
 2 CONTRACT, agreement, bargain, compact, convention, covenant, pact, transaction
 3 ADHERENCE 1, adhesion, cling, clinging, coherence, cohesion, stickage, sticking
 4 GUARANTEE 1, bail, guaranty, security, surety, warranty
 Please choose the appropriate sense.
User: 4
Computer then returns appropriate passages.
Computer: Does this answer your question?
User: No.
Computer: Would you like to try again? Use the following hints in rephrasing your question (e.g., be as specific as possible, make sure the spelling is correct, think about what kind of answer you want).

Figure 11.2. Sample Interaction

11.6 Architecture of an Interactive Question Answering System

From a technical perspective, the system can be described as follows: domain specific documents are collected, preprocessed, split into sections, indexed, and tagged offline at multiple levels of linguistic analysis the same as before. A user who comes to the system can ask a natural language question immediately, or go to the pages containing help information explaining the system's processing and then submit a question. The system will conduct spell checking, acronym expansion, and focus identification after receiving the question. Next, the query negotiation process will be activated to refine the question.

The user will have total control of the negotiation process and can stop at any point. The finalized question will then go through passage retrieval and answer finding. The system will return 20 possible answer-providing passages and display them to the user. The user will be given the opportunity to evaluate these answers. The system will provide several options if the user is not satisfied, including rephrasing the question for another search attempt, asking an expert for the answer, or returning answers from other resources.

The QA system will have the following modules: (1) document processing; (2) question interpretation and negotiation which has three components: question entry/revision page, question interpreter, and question negotiator; (3) search engine; (4) answer providing passage provision; and (5) answer satisfaction. The document processing, search engine, and answer providing passage provision modules stay the same as in the current system model. Question interpretation and negotiation and answer satisfaction modules are new and are described in detail below and can be seen in figure 11.1 as the shaded modules.

11.6.1 Question Interpretation and Negotiation Module

Question Entry/Revision Page. On this page, the user may type in the question, review the rephrased question returned by the system, or go to the help page for more information. This page will be brought up at each interaction between the system and the user. The user can ask a natural-language question immediately, or go to the help pages containing a description of question techniques, an explanation of what kinds of questions can be best answered by the system, sample questions, and a collection of domain specific terms and their definitions or de-abbreviations.

Question Interpreter. This module includes spell-checking and acronym expansion. If unusual spellings are detected, the system will ask the user whether they are intended and suggest alternative spellings. At this stage the system will ask the user to pick the full form of any acronym used in the question. The original L2L module, will be extended to facilitate interactive question answering. In addition to identifying the question's focus, the L2L module combines linguistic features such as phrases, categorizations, and extractions to construct what we call "semantic expected-answer frames" based on an understanding of users' queries.

For those questions where the system automatically creates a semantic expected-answer frame, the goal is to establish one or more of the following extraction segments: type, frame type, and content.

The frame type indicates whether the extraction is about a named entity, an entity, or an event. This information is not only important for finding the answer, but also in the dialog between the system and the user. By showing the user what the system understands is expected as an answer and soliciting feed-

Answer Frame

event (frame type): composed of
 agent: Thermal Protection systems
 (two possible answers expected)
 object: panel
 object: tile
 entity (frame type) : panel
 type=part
 entity (frame type): tile
 type=part

Figure 11.3. Answer Frame

back, a much more exact picture of the expected answer is likely to emerge. Extraction content is textual information from the document that can be found in the answer frame that identifies logical relations between the question's terms, and term importance.

For example, the question "Are thermal protection systems of spacecrafts commonly composed of one panel or a collection of smaller tiles?" would produce an answer frame as in figure 11.3.

The semantic expected answer frames will be fed into the question negotiator component to be shown to the user and to receive feedback on the system's understanding of the question.

Question Negotiator. The question negotiator begins the conversation between the system and the user. Complex questions such as How, Why, Alternative, and Conditional questions (and to some extent Yes/No questions) require complex answers and may not be satisfied by the traditional "focus" approach. Each time the system does not understand the focus for a question, it turns to the question negotiator for help in refining the question. In each interaction, the system will carry out one of the following actions: rephrase the question, ask for more information from the user, or end the interaction to start the answer-finding process.

For some questions, the information need is easily determined by the system. The question "What is the weight of the space shuttle?" clearly asks for a weight of a certain item. This type of question is currently recognized by our system as it is familiar with weight measures (i.e., tons, kilos, pounds) and can provide a short factual answer. It is therefore fairly straightforward to add an extra step where the system paraphrases a weight question: "Do you want to know how much the <OBJECT> weighs?"

Paraphrasing the question becomes more complicated when questions are open-ended ("Why must there be a buffer between tiles on the thermal protection system surface?"), or ambiguous ("How does an X-Ray spectrometer lo-

cate stress fields?”). For those questions that cannot be paraphrased as easily, the question needs to be processed in more detail. During this process, information about the entities, events, and relations is extracted and presented in human-readable form. The user can quickly see whether the system has understood the question. If so, the system proceeds to the next step in the reference interview.

Each type of question would require a separate query negotiation process based on its classification. For example, an alternative question would be split into two or more questions. The question “Are Thermal Protection systems of spacecrafts commonly composed of one panel or a collection of smaller tiles?” would necessitate splitting the later part of the question into the two different parts, so that the answer documents could be retrieved using the questions, “Are TPS of spacecrafts commonly composed of one panel” and “Are TPS or spacecrafts commonly composed of a collection of smaller tiles?” The returned answer passages would be ranked for each separate question allowing the user to compare and synthesize the question alternatives.

If the L2L conversion process returns a question missing key information (e.g., no focus), the system will then ask the user to supply additional information pertaining to the question. The user will be able to alter the question as understood by the system. For example if the system returned the wrong subject, the user can type over the incorrect subject and resubmit the question with corrections.

11.6.2 Answer Satisfaction

The user is then shown a page that allows feedback about the answers provided. If the answer is unsatisfactory, the user will be provided with three options: (1) the user can choose to return to the question entry/revision page to begin the process with an alternative question; (2) the user will be able to ask the question using the resources from the web, or; (3) the user can choose to get help from a subject specialist.

11.7 Conclusion

This chapter is based on our findings in a real-world environment where we are providing a QA system to a real set of users with particular tasks required of them. We have found that the nature of queries (examples of which are included throughout this chapter) generated by real users, as well as the breadth vs. narrowness of what constitutes a useful answer, diverges substantially from the TREC experimental setup. While the TREC approach may be required for comparative testing of multiple systems, we have found that answering complex questions in a real-time environment requires a quite different approach. We propose to provide this type of QA system capability by incorporating the

well-known reference interview process from library science into the automatic question-answering process. By helping the user to formulate questions that the system can understand and to reformulate complex questions based on the system's feedback, we believe we can improve the question-answering process for such real world environments.

Acknowledgments

This research was supported by funding from National Aeronautics and Space Administration under NCC1-01-004, the State of New York, and AT&T.

Elizabeth D. Liddy is a professor in the School of Information Studies at Syracuse University and director of its Center for Natural Language Processing. Her research interests include natural language processing, question-answering, and text mining. For further information and contact, see www.cnlp.org.

Jiangping Chen is an assistant professor in the School of Library and Information Sciences at University of North Texas. Her research interests include cross language information retrieval and Chinese information processing. She can be reached at jpchen@unt.edu

Ozgur Yilmazel is a senior software engineer at the Center for Natural Language Processing in the School of Information Studies at Syracuse University. Research interests include question answering, information retrieval and text categorization. For further information and contact, see www.cnlp.org

Anne Diekema is a research assistant professor at the Center for Natural Language Processing in the School of Information Studies at Syracuse University. Her research interests include cross-language information retrieval, natural language processing, and information retrieval. For further information and contact, see www.cnlp.org

Lan He is a research analyst at the Center for Natural Language Processing in the School of Information Studies at Syracuse University. Research interests include coreference resolution and question-answering. For further information and contact, see www.cnlp.org

Sarah Harwell is a research analyst at the Center for Natural Language Processing in the School of Information Studies at Syracuse University. Research interests include question-answering, text classification and natural language processing. For further information and contact, see www.cnlp.org