HITIQA: An Interactive Question Answering System
A Preliminary Report

Sharon G. Small, Nobuyuki Shimizu, Tomek Strzalkowski, Liu Ting
ILS Institute
The State University of New York at Albany
1400 Washington Avenue
Albany, NY 12222
{small,ns3203,tomek,tl7612}@albany.edu
(518)442-3137
(518)442-2606(fax)

Summary

HITIQA is an interactive question answering technology designed to allow intelligence analysts and other users of information systems to pose questions in natural language and obtain relevant answers, or the assistance they require in order to perform their tasks. Our objective in HITIQA is to allow the user to submit exploratory, analytical, non-factual questions, such as *"What has been Russia's reaction to U.S. bombing of Kosovo?"* There are very significant differences between factual, or fact-finding, and analytical question answering. A factual question seeks pieces of information that would make a corresponding statement true (i.e., they become facts): "How many states are in the U.S.?" / "There are X states in the U.S." The distinguishing property of analytical questions is that one cannot generally anticipate what might constitute the answer. While certain types of things may be expected (e.g., diplomatic statements), the answer is heavily conditioned by what information is in fact available on the topic. From a practical viewpoint, analytical questions are often underspecified, thus casting a broad net on a space of possible answers. Therefore, clarification dialogue is often needed to negotiate with the user the exact scope and intent of the question.

Keywords

Analytical Question-Answering System
Dialogue

Abstract

HITIQA is an interactive question answering technology designed to allow intelligence analysts and other users of information systems to pose questions in natural language and obtain relevant answers, or the assistance they require in order to perform their tasks. Our objective in HITIQA is to allow the user to submit exploratory, analytical, non-factual questions, such as *"What has been Russia's reaction to U.S. bombing of Kosovo?"* The distinguishing property of such questions is that one cannot generally anticipate what might constitute the answer. While certain types of things may be expected (e.g., diplomatic statements), the answer is heavily conditioned by what information is in fact available on the topic. From a practical viewpoint, analytical questions are often underspecified, thus casting a broad net on a space of possible answers. Therefore, clarification dialogue is often needed to negotiate with the user the exact scope and intent of the question.

1 Introduction

HITIQA project is part of the ARDA AQUAINT program that aims to make significant advances in the state of the art of automated question answering. In this paper we focus on two aspects of our work:

1. Question Semantics: how the system "understands" user requests.
2. Human-Computer Dialogue: how the user and the system negotiate this understanding.

We will also discuss very preliminary evaluation results from a series of pilot tests of the system conducted by intelligence analysts via a remote internet link.

2 Factual vs. Analytical

The objective in HITIQA is to allow the user to submit and obtain answers to exploratory, analytical, non-factual questions. There are very significant differences between factual, or fact-finding, and analytical question answering. A factual question seeks pieces of information that would make a corresponding statement true (i.e., they become facts): "How many states are in the U.S.?" / "There are X states in the U.S." In this sense, a factual question usually has just one correct answer that can generally, be judged for its truthfulness. By contrast, an analytical question is when the "truth" of the answer is more a matter of opinion and may depend upon the context in which the question is asked. Answers to analytical questions are rarely unilateral, indeed, a mere "correct" answer may have limited value, and in some cases may not even be determinate ("Which college is the best?", "How do I stop my baby's crying?"). Instead, answers to analytical questions are often judged as helpful, or useful, or satisfactory, etc. "Technically correct" answers (e.g., "feed the baby milk") may be considered as irrelevant or at best unresponsive.

The distinction between factual and analytical questions depends primarily on the intention of the person who is asking, however, the form of a question is often indicative of which of the two classes it is more likely to belong to. Factual questions can be classified into a number of syntactic formats ("question typology") that aids in automatic processing.

Factual questions display a fairly distinctive "answer type", which is the type of the information piece needed to fulfill the statement. Recent automated systems for answering factual questions deduct this expected answer type from the form of the question and a finite list of possible answer types. For example, "Who was the first man in space" expects a "person" as the answer, while "How long was the Titanic?" expects some length measure as an answer, probably in yards and feet, or meters. (Prager, 2001). This is generally a very good strategy, that has been exploited successfully in a number of automated QA systems that appeared in recent years, especially in the context of TREC QA[1] evaluations (Harabagiu et al., 2000; Hovy et al., 2000; Prager at al., 2001).

This process is not easily applied to analytical questions. This is because the type of an answer for analytical questions cannot always be anticipated due to their inherently exploratory character. In contrast to a factual question, an analytical question has an unlimited variety of syntactic forms with only a loose connection between their syntax

[1] TREC QA is the annual Question Answering evaluation sponsored by the U.S. National Institute of Standards and Technology www.trec.nist.gov.

and the expected answer. Given the unlimited potential of the formation of analytical questions, it would be counter-productive to restrict them to a limited number of question/answer types. Even finding a non-strictly factual answer to an otherwise simple question about Titanic length (e.g., “two football fields”) would push the limits of the answer-typing approach. Therefore, the formation of an answer should instead be guided by the topics the user is interested in, as recognized in the query and/or through the interactive dialogue, rather than by a single type as inferred from the query in a factual system.

This paper argues that the semantics of an analytical question is more likely to be deducted from the information that is considered relevant to the question than through a detailed analysis of their particular form. While this may sound circular, it needs not be. Determining “relevant” information is not the same as finding an answer; indeed we can use relatively simple information retrieval methods (keyword matching, etc.) to obtain perhaps 50 or 100 “relevant” documents from a database. This gives us an initial answer space to work on in order to determine the scope and complexity of the answer. In our project, we use structured templates, which we call *frames* to map out the content of pre-retrieved documents, and subsequently to delineate the possible meaning of the question (Section 6).

3 Document Retrieval

In the experiments with the HITIQA prototype, see Figure 1, we are retrieving the top fifty documents from three gigabytes of newswire (AQUAINT corpus plus web-harvested documents).

4 Data Driven Semantics of Questions

The set of documents and text passages returned from the initial search is not just a random subset of the database. Depending upon the quality (recall and precision) of the text retrieval system available, this set can be considered as a first stab at understanding the user’s question by the machine. Again, given the available resources, this is the best the system can do under the circumstances. Therefore, we may as well consider this collection of retrieved texts (*the Retrieved Set*) as the meaning of the question as understood by the system. This is a fair assessment: the better our search capabilities, the closer this set would be to what the user may accept as an answer to the question.

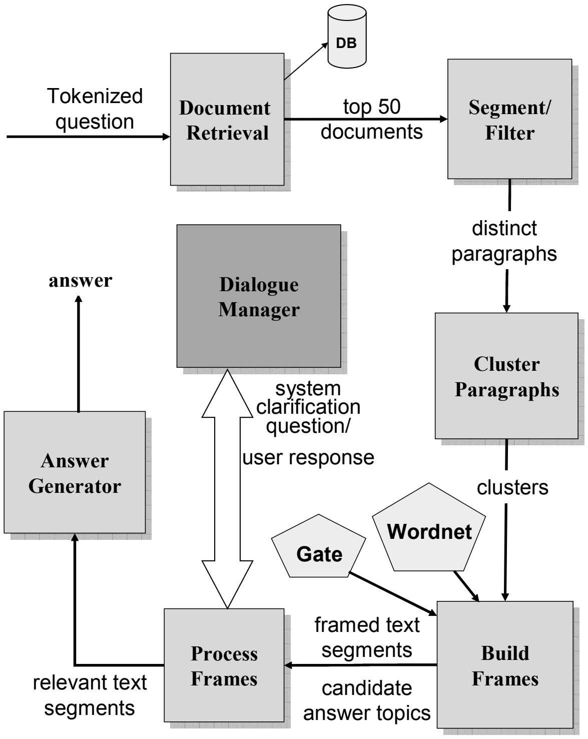

Figure 1: HITIQA preliminary architecture

We can do better, however. We can perform automatic analysis of the retrieved set, attempting to uncover if it is a fairly homogenous bunch (i.e., all texts have very similar content), or whether there are a number of diverse topics represented there, somehow tied together by a common thread. In the former case, we may be reasonably confident that we have the answer, modulo the retrievable information. In the latter case, we know that the question is more complex than the user may have intended, and a negotiation process is needed.

We can do better still. We can measure how well each of the topical groups within the retrieved set is “matching up” against the question. This is accomplished through a framing process described later in this paper. The outcome of the framing process is twofold: firstly, the alternative interpretations of the question are ranked within 3 broad categories: *on-target, near-misses and outliers*. Secondly, salient concepts and attributes for each topical group are extracted into topic frames. This enables the system to conduct a meaningful dialogue with the user, a dialogue which is wholly content oriented, and thus entirely data driven.

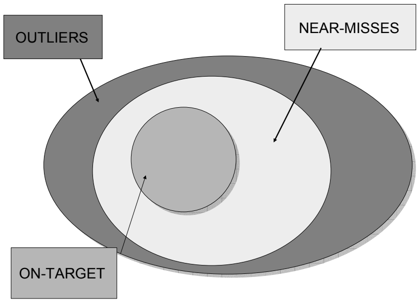

Figure 2: Answer Space Topology. The goal of interactive QA it to optimize the ON-TARGET middle zone.

5 Clustering

Section will be included in full paper.

6 Framing

In HITIQA we use a *text framing* technique to delineate the gap between the meaning of the user's question and the system "understanding" of this question. The framing is an attempt to impose a partial structure on the text that would allow the system to systematically compare different text pieces against each other and against the question, and also to communicate with the user about this. In particular, the framing process may uncover topics and themes within the retrieved set which the user has not explicitly asked for, and thus may be unaware of their existence. Nonetheless these may carry important information – the NEAR-MISSES in Figure 2.

In the current version of the system, frames are fairly generic templates, consisting of a small number of attributes, such as LOCATION, PERSON, COUNTRY, ORGANIZATION, etc. Future versions of HITIQA will add domain specialized frames, for example, we are currently constructing frames for the Weapons Non-proliferation Domain. Most of the frame attributes are defined in advance, however, dynamic frame expansion is also possible. Each of the attributes in a frame is equipped with an extractor function which specializes in locating and extracting instances of this attribute in the running text. Therefore, the framing process resembles strongly the template filling task in information extraction (cf. MUC[3] evaluations), with one significant exception: while the MUC task was to fill in a template using potentially any amount of source text (Humphreys et al., 1998), the framing is essentially an inverse process. In framing, potentially multiple frames can be associated with a small chunk of text (a passage or a short paragraph). Furthermore, this chunk of text is part of a cluster of very similar text chunks that further reinforce some of the most salient features of these texts. This makes the frame filling a significantly less error-prone task – our experience has been far more positive than the MUC evaluation results may indicate. This is because, rather than trying to find the most appropriate values for attributes from among many potential candidates, we in essence fit the frames over small passages[4].

TOPIC:*[pollution, industry, sources]*
LOCATION: *[Black Sea]*
INDUSTRY:*[fishing]*

Figure 3: HITIQA generated Goal Frame

TOPIC: *pollution*
SUB-TOPIC: *[sources]*
LOCATION: *[Black Sea]*
INDUSTRY :*[fisheries, tourism]*
TEXT: *[In a period of only three decades (1960's-1980's), the* ***Black Sea*** *has suffered the catastrophic degradation of a major part of its natural resources. Particularly acute problems have arisen as a result of* ***pollution*** *(notably from nutrients, fecal material, solid waste and oil), a catastrophic decline in commercial fish stocks, a severe decrease in* ***tourism*** *and an uncoordinated approach towards coastal zone management. Increased loads of nutrients from rivers and coastal* ***sources*** *caused an overproduction of phytoplankton leading to extensive eutrophication and often extremely low dissolved oxygen concentrations. The entire ecosystem began to collapse. This problem, coupled with* ***pollution*** *and irrational exploitation of fish stocks, started a sharp decline in* ***fisheries*** *resources.]*
RELEVANCE: *Matches on all elements found in goalframe*

Figure 4: A HITIQA generated date frame. Words in bold were used to fill the Frame.

[3] MUC, the Message Understanding Conference, funded by ARPA, involved the evaluation of information extraction systems applied to a common task.

[4] We should note that selecting the right frame type for a passage is an important pre-condition to "understanding".

A very similar process is applied to the user's question, resulting in a *Goal Frame* which can be subsequently compared to the data frames obtained from retrieved data. For example, the Goal Frame generated from the question, "*How has pollution in the Black Sea affected the fishing industry, and what are the sources of this pollution?*" is shown in Figure 3.

7 Judging Frame Relevance

We judge a particular data frame as relevant, and subsequently the corresponding segment of text as relevant, by comparison to the Goal Frame. The data frames are scored based on the number of conflicts found between them and the Goal Frame. The conflicts are mismatches on values of corresponding attributes. If a data frame is found to have no conflicts, it is given the highest relevance rank, and a conflict score of zero. All other data frames are scored with an incrementing conflict value, one for frames with one conflict with the Goal Frame, two for two conflicts etc

8 Enabling Dialogue with the User

Framed information allows HITIQA to automatically judge some text as relevant and to conduct a meaningful dialogue with the user as needed on other text. The purpose of the dialogue is to help the user to navigate the answer space and to solicit from the user more details as to what information he or she is seeking. The main principle here is that the dialogue is at the information semantic level, not at the information organization level. Thus, it is okay to ask the user whether information about the AIDS conference in Cape Town should be included in the answer to a question about combating AIDS in Africa. However, the user should never be asked if a particular keyword is useful or not, or if a document is relevant or not. We have developed a 3-pronged strategy:

1. Narrowing dialogue: ask questions that would allow the system to reduce the size of the answer set.
2. Expanding dialogue: ask questions that would allow the system to decide if the answer set needs to be expanded by information just outside of it (near-misses).
3. Fact seeking dialogue: allow the user to ask questions seeking additional facts and specific examples, or similar situations.

Of the above, we have thus far implemented the first two options as part of the preliminary clarification dialogue.

9 Clarification Dialogue

Data frames with a conflict score of zero form the initial kernel answer space. Depending upon the size of this set and the presence of other frames outside of it, the system either proceeds to generate the answer or initiates a dialogue with the user. For example, if the answer space appears too large or varied, e.g. consists of many different topics, the system may ask the user how to narrow it. Alternatively, the presence of large groups of texts frames with near-miss frames assigned to them (i.e., frames with 1 or 2 conflicts with the Goal Frame) may indicate that the answer space is actually larger, and the user will be consulted about a possible broadening of the question. Currently, we only initiate a clarification dialogue for 1-conflict frames.

A 1-conflict frame has only a single attribute mismatch with the Goal Frame. This could be a mismatch on any attribute, for example, LOCATION, or ORGANIZATION, or TIME, etc. A special case arises when the conflict occurs on the TOPIC attribute. Since all other attributes match, we may be looking at potentially different events or situations involving the same entities, or occurring at the same location or time. The purpose of the clarification dialogue in this case is to probe which of these topics may be of interest to the user. This is illustrated in the exchange below recorded during an evaluation session with an intelligence analyst:

User: *"Who is Elizardo Sanchez?"*
HITIQA*: "Are you interested in seeing information about civil rights as it is related to Elizardo Sanchez?*

ONE-CONFLICT FRAME

TOPIC**: *civil_rights***
SUB-TOPIC*: []*
LOCATION: *[Cuba, Cuba, Cuba]*
COMPANIES: *[]*
PEOPLE: *[Sanchez, Sanchez]*
DOCUMENT DATE: *[2000, 1, 11]*
SOURCE*: HAVANA (AP) –*

TEXT: *[``I consider that the situation for civil and political rights in* ***Cuba*** *has worsened over the* ***past year****... owing to that* ***Cuba*** *continues to be the only closed society in this hemisphere,"* ***Sanchez*** *said. ``There have been no significant release of prisoners, the number of people sanctioned or processed for political motives increased.* ***Sanchez****, who himself spent many years in* ***Cuban*** *prisons, ...]*

Figure 5: One of the Frames that were used in generating Sanchez dialogue. Words in bold were used to fill the Frame.

In order to understand what happened here, we need to note first that the Goal Frame for the user question does not have any specific value assigned to its TOPIC attribute. This of course is as we would expect it: the question does not give us a hint as to what information we need to look for or may be hoping to find about Sanchez. This also means that all the text frames obtained from the retrieved set for this question will have at least one conflict, near-misses. One such text frame is shown in Figure 5: its topic is “civil rights” and it about Sanchez. HITIQA thus asks if “civil rights” is a topic of interest to the user. If the user responds positively, this topic will be added to the answer space.

The clarification dialogue will continue on the topic level until all the significant sets of NEAR-MISS frames are either included in the answer space (through user broadening the scope of the question that removes the initial conflicts) or dismissed as not relevant. When the number of frames is within the acceptable range, HITIQA will generate the answer using the text from the frames in the current answer space. The user may end the dialogue at any point and have an answer generated given the current state of the frames.

10 Answer Generation

Currently, the answer is simply composed of text passages from the zero conflict frames. The text of these frames are ordered by date and outputted to the user.

11 Evaluations

Section will be included in full paper

12 Future Work

Section will be included in full paper

Acknowledgements

This paper is based on work supported by the Advanced Research and Development Activity (ARDA)’s Advanced Question Answering for Intelligence (AQUAINT) Program under contract number 2002-H790400-000.

References

Harabagiu, S., M. Pasca and S. Maiorano. 2000. Experiments with Open-Domain Textual Question Answering. In *Proc. of COLING-2000*. 292-298.

Humphreys, R. Gaizauskas, S. Azzam, C. Huyck, B. Mitchell, H. Cunningham, Y. Wilks. 1998. Description of the LaSIE-II System as Used for MUC-7. *In Proceedings of the Seventh Message Understanding Conference (MUC-7.)*

Hovy, E., L. Gerber, U. Hermjakob, M. Junk, C-Y. Lin. 2000. Question Answering in Webclopedia. *Notebook Proceedings of Text Retrieval Conference (TREC-9).*

Miller, G.A. 1995. WordNet: A Lexical Database. *Comm. of the ACM*, 38(11):39-41.

S. Seneff and J. Polifroni, ``Dialogue Management in the MERCURY Flight Reservation System," *Proc. ANLP-NAACL 2000, Satellite Workshop*, 1-6, Seattle, WA, 2000.

Marilyn A. Walker. An Application of Reinforcement Learning to Dialogue Strategy Selection in a Spoken Dialogue System for Email . *Journal of Artificial Intelligence Research*.12:387-416.

W. Ward and B. Pellom. 1999. The CU Communicator System. *IEEE ASRU*. 341-344.