HITIQA: A Data Driven Approach to Interactive Analytical Question Answering

Sharon Small and Tomek Strzalkowski
The State University of New York at Albany
1400 Washington Avenue
Albany, NY 12222
{small,tomek}@cs.albany.edu

Abstract

In this paper we describe the analytic question answering system HITIQA (High-Quality Interactive Question Answering) which has been developed over the last 2 years as an advanced research tool for information analysts. HITIQA is an interactive open-domain question answering technology designed to allow analysts to pose complex exploratory questions in natural language and obtain relevant information units to prepare their briefing reports. The system uses novel data-driven semantics to conduct a clarification dialogue with the user that explores the scope and the context of the desired answer space. The system has undergone extensive hands-on evaluations by a group of intelligence analysts representing various foreign intelligence services. This evaluation validated the overall approach in HITIQA but also exposed limitations of the current prototype.

1 Introduction

Our objective in HITIQA is to allow the user to submit exploratory, analytical 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. Questions posed by professional analysts are aimed to probe the available data along certain dimensions. The results of these probes determine follow up questions, if necessary. Furthermore, at any stage clarifications may be needed to adjust the scope and intent of each question. Figure 1a shows a fragment of an analytical session with HITIQA; please note that these questions are *not* aimed at factoids, despite appearances. HITIQA project is part of the ARDA AQUAINT program that aims to make significant advances in state of the art of automated question answering.

User: *What is the history of the nuclear arms program between Russia and Iraq?*
HITIQA: [responses and clarifications]
User: *Who financed the nuclear arms program in Iraq?*
HITIQA:...
User: *Has Iraq been able to import uranium?*
HITIQA:...
User: *What type of debt does exist between Iraq and Russia?*

FIGURE 1a: A fragment of analytic session

2 Factoid vs. Analytical QA

The process of automated question answering is now fairly well understood for most types of factoid questions. Factoid questions display a fairly distinctive "answer type", which is the type of the information item needed for the answer, e.g., "person" or "country", etc. Most existing factoid QA systems deduct this expected answer type from the form of the question using a finite list of possible answer types. For example, "How long was the Titanic?" expects some length measure as an answer, probably in yards and feet, or meters. 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., 2002; Hovy et al., 2000; Prager at al., 2001).

This answer-typing process is not easily applied to analytical questions because the type of an answer for analytical questions cannot always be anticipated due to their inherently exploratory character. In contrast to a factoid question, an analytical question has an unlimited variety of syntactic forms with only a loose connection between their syntax 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. Therefore, the formation of an answer in analytical QA should instead be guided by the user's interest as expressed in the question, as well as through an interactive dialogue with the system.

In this paper we argue that the semantics of an analytical question is more likely to be deduced from the

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

information that is considered relevant to the question than through a detailed analysis of its particular form. 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 200 "relevant" documents from a database. This gives us an initial answer space to work from in order to determine the scope and complexity of the answer, but we are nowhere near the answer yet. 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 before we can attempt to formulate an answer.

3 Text Framing

In HITIQA we use a text framing technique to delineate the gap between the meaning of the user's question and the system's "understanding" of this question. The framing process does not attempt to capture the entire meaning of the passages; instead it imposes a partial structure on the text passages that would allow the system to systematically compare different passages against each other and against the question. Framing is just sufficient enough to communicate with the user about the differences in their question and the returned text. In particular, the framing process may uncover topics or aspects within the answer space which the user has not explicitly asked for, and thus may be unaware of their existence. If these topics or aspects align closely with the user's question, we may want to make the user aware of them and let him/her decide if they should be included in the answer.

Frames are built from the retrieved data, after clustering it into several topical groups. Retrieved documents are first broken down into passages, mostly exploiting the naturally occurring paragraph structure of the original sources, filtering out duplicates. The remaining passages are clustered using a combination of hierarchical clustering and n-bin classification (Hardy et al., 2002). Typically three to six clusters are generated. Each cluster represents a topic theme within the retrieved set: usually an alternative or complimentary interpretation of the user's question. Since clusters are built out of small text passages, we associate a frame with each passage that serves as a seed of a cluster. We subsequently merge passages, and their associated frames whenever anaphoric and other cohesive links are detected.

HITIQA starts by building a general frame on the seed passages of the clusters and any of the top *N* (currently *N=10*) scored passages that are not already in a cluster. The general frame represents an event or a relation involving any number of entities, which make up the frame's attributes, such as LOCATION, PERSON, COUNTRY, ORGANIZATION, etc. Attributes are extracted from text passages by BBN's Identifinder, which tags 24 types of named entities. The event/relation itself could be pretty much anything, e.g., *accident, pollution, trade*, etc. and it is captured into the TOPIC attribute from the central verb or noun phrase of the passage. In general frames, attributes have no assigned roles; they are loosely grouped around the TOPIC.

We have also defined three slightly more specialized *typed frames* by assigning *roles* to selected attributes in the general frame. These three "specialized" frames are: (1) a Transfer frame with three roles including FROM, TO and OBJECT; (2) a two-role Relation frame with AGENT and OBJECT roles; and (3) a one-role Property frame. These typed frames represent certain generic events/relationships, which then map into more specific event types in each domain. Other frame types may be defined if needed, but we do not anticipate there will be more than a handful all together.[2] Where the general frame is little more than just a "bag of attributes", the typed frames capture some internal structure of an event, but only to the extent required to enable an efficient dialogue with the user. Typed frames are "triggered" by appearance of specific words in text, for example the word *export* may trigger a Transfer frame. A single text passage may invoke one or more typed frames, or none at all. When no typed frame is invoked, the general frame is used as default. If a typed frame is invoked, HITIQA will attempt to identify the roles, e.g. FROM, TO, OBJECT, etc. This is done by mapping general frame attributes selected from text onto the typed attributes in the frames. In any given domain, e.g., weapon non-proliferation, both the trigger words and the role identification rules can be specialized from a training corpus of typical documents and questions. For example, the role-ID rules rely both on syntactic cues and the expected entity types, which are domain adaptable.

Domain adaptation is desirable for obtaining more focused dialogue, but it is not necessary for HITIQA to work. We used both setups under different conditions: the generic frames were used with TREC document collection to measure impact of IR precision on QA accuracy (Small et al., 2004). The domain-adapted frames were used for sessions with intelligence analysts working with the WMD Domain (see below). Currently, the adaptation process includes manual tuning followed by corpus bootstrapping using an unsupervised learning method (Strzalkowski & Wang, 1996). We generally rely on BBN's Identifinder for extraction of basic entities, and use bootstrapping to define additional entity types as well as to assign roles to attributes.

The version of HITIQA reported here and used by analysts during the evaluation has been adapted to the

[2] Scalability is certainly an outstanding issue here, and we are working on effective frame acquisition methods, which is outside of the scope of this paper.

Weapons of Mass Destruction Non-Proliferation domain (WMD domain, henceforth). Figure 1b contains an example passage from this data set. In the WMD domain, the typed frames were mapped onto *WMDTransfer* 3-role frame, and two 2-role frames *WMDTreaty* and *WMDDevelop*. Adapting the frames to WMD domain required only minimal modification, such as adding WEAPON entity to augment Identifinder entity set, specializing OBJECT attribute in *WMDTransfer* to WEAPON, generating a list of international weapon control treaties, etc.

HITIQA frames define top-down constraints on how to interpret a given text passage, which is quite different from MUC[3] template filling task (Humphreys et al., 1998). What we're trying to do here is to "fit" a frame over a text passage. This means also that multiple frames can be associated with a text passage, or to be exact, with a cluster of passages. Since most of the passages that undergo the framing process are part of some cluster of very similar passages, the added redundancy helps to reinforce the most salient features for extraction. This makes the framing process potentially less error-prone than MUC-style template filling[4].

> The Bush Administration claimed that Iraq was within one year of producing a nuclear bomb. On 30 November 1990... Leonard Spector said that Iraq possesses 200 tons of natural uranium imported and smuggled from several countries. Iraq possesses a few working centrifuges and the blueprints to build them. Iraq imported centrifuge materials from Nukem of the FRG and from other sources. One decade ago, Iraq imported 27 pounds of weapons-grade uranium from France, ...

FIGURE 1b: A text passage from the WMD domain data

A very similar framing process is applied to the user's question, resulting in one or more *Goal frames,* which are subsequently compared to the data frames obtained from retrieved text passages. A Goal frame can be a general frame or any of the typed frames. The Goal frame generated from the question, *"Has Iraq been able to import uranium*?" is shown in Figure 2. This frame is of *WMDTransfer* type, with 3 role attributes TRF_TO, TRF_FROM and TRF_OBJECT, plus the relation type (TRF_TYPE). Each role attribute is defined over an underlying general frame attribute (given in parentheses), which is used to compare frames of different types.

HITIQA automatically judges 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 with 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.

FRAME TYPE: *WMDTransfer*
TRF_TYPE (TOPIC): *import*
TRF_TO (LOCATION): *Iraq*
TRF_FROM (LOCATION, ORGANIZATION): *?*
TRF_OBJECT (WEAPON): *uranium*

FIGURE 2: A domain Goal frame from the Iraq question

FRAME TYPE: *WMDTransfer*
TRF_TYPE (TOPIC): *imported*
TRF_TO (LOCATION): *Iraq*
TRF_FROM (LOCATION): *France* [missed: *Nukem of FRG*]
TRF_OBJECT (WEAPON): *uranium*
CONFLICT SCORE*: 0*

FIGURE 3: A frame obtained from the text passage in Figure 1b, in response to the Iraq question

All other data frames are scored with an increasing value based on the number of conflicts, score 1 for frames with one conflict with the Goal frame, score 2 for two conflicts etc. Frames that conflict with all information found in the query are given the score 99 indicating the lowest rank. Currently, frames with a conflict score 99 are excluded from further processing as outliers. The frame in Figure 3 is scored as relevant to the user's query and included in the answer space.

4 Clarification Dialogue

Data frames with a conflict score of zero form the initial kernel answer space and HITIQA proceeds by generating an answer from this space. Depending upon the presence of other frames outside of this set, the system may initiate a dialogue with the user. HITIQA begins asking the user questions on these near-miss frame groups, groups with one or more conflicts, with the largest group first. In order to keep the dialogue from getting too winded, we set thresholds on number of conflicts and group size that are considered by the dialogue manager.

A *1-conflict* frame has only a single attribute mismatch with the Goal frame. This could be a mismatch on any of the general frame attributes, for example, LOCATION, ORGANIZATION, TIME, etc., or in one of the role-assigned attributes, TO, FROM, OBJECT, etc. A special case arises when the conflict occurs on the TOPIC attribute, which indicates the event type. Since all other attributes match, we may be looking at potentially different events of the same kind involving the same entities, possibly 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. Another special case arises when the Goal frame is of a different type than a data frame. The purpose of the clarification dialogue in this case is to see if the user wishes to expand the answer space to include events of a different type. This situation is illustrated in the ex-

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

[4] We do not have enough data to make a definite comparison at this time.

change shown in Figure 4. Note that the user can examine a partial answer prior to answering clarification questions.

User: *"Has Iraq been able to import uranium?"*
[a partial answer displayed in an answer window]
HITIQA: *"Are you also interested in background information on the uranium development program in Iraq?"*

FIGURE 4: Clarification question generated for the Iraq/uranium question

The clarification question in Figure 4 is generated by comparing the Goal frame in Figure 2 to a partly matching frame (Figure 5) generated from some other text passage. We note first that the Goal frame for this example is of *WMDTransfer* type, while the data frame in Figure 5 is of the type *WMDDevelop*. Nonetheless, both frames match on their general-frame attributes WEAPON and LOCATION. Therefore, HITIQA asks the user if it should expand the answer space to include development of uranium in Iraq as well.

During the dialogue, as new information is obtained from the user, the Goal frame is updated and the scores of all the data frames are reevaluated. If the user responds the equivalent of "yes" to the system clarification question in the dialogue in Figure 4, a corresponding *WMDDevelop* frame will be added to the set of active Goal frames and all *WMDDevelop* frames obtained from text passages will be re-scored for possible inclusion in the answer.

FRAME TYPE: *WMDDevelop*
DEV_TYPE (TOPIC): *development*, *produced*
DEV_OBJECT (WEAPON): *nuclear weapons*, *uranium*
DEV_AGENT (LOCATION): *Iraq*, *Tuwaitha*
CONFLICT SCORE: *2*
Conflicts with FRAME_TYPE *and* TOPIC

FIGURE 5: A 2-conflict frame against the Iraq/uranium question that generated the dialogue in Figure 4.

The user may end the dialogue at any point using the generated answer given the current state of the frames. Currently, the answer is simply composed of text passages from the zero conflict frames. In addition, HITIQA will generate a "headline" for the text passages in the answer space. This is done using a combination of text templates and simple grammar rules applied to the attributes of the passage frame.

5 HITIQA Qualitative Evaluations

In order to assess our progress thus far, and to also develop metrics to guide future evaluation, we invited a group of analysts employed by the US government to participate in two three-day workshops, held in September and October 2003.

The two basic objectives of the workshops were:

1. To perform a realistic assessment of the usefulness and usability of HITIQA as an end-to-end system, from the information seeker's initial questions to completion of a draft report.

2. To develop metrics to compare the answers obtained by different analysts and evaluate the quality of the support that HITIQA provides.

The analysts' primary task was preparation of reports in response to *scenarios* - complex questions that usually encompassed multiple sub-questions. The scenarios were developed in conjunction with several U.S. government offices. These scenarios, detailing information required for the final report, were not normally used directly as questions to HITIQA, instead, they were treated as a basis to issues possibly leading to a series of questions, as shown in Figure 1a.

The results of these evaluations strongly validated our approach to analytical QA. At the same time, we learned a great deal about how analysts work, and about how to improve the interface.

Analysts completed several questionnaires designed to assess their overall experience with the system. Many of the questions required the analysts to compare HITIQA to other tools they were currently using in their work. HITIQA scores were quite high, with mean score 3.73 out of 5. We scored particularly high in comparison to current analytic tools. We have also asked the analysts to cross-evaluate their product reports obtained from interacting with HITIQA. Again, the results were quite good with a mean answer quality score of 3.92 out of 5. While this evaluation was only preliminary, it nonetheless gave us confidence that our design is "correct" in a broad sense.[5]

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

Hardy, H., et al. 2002. *Cross-Document Summarization by Concept Classification*. Proceedings of SIGIR, Tampere, Finland.

Harabagiu, S., et. al. 2002. *Answering Complex, List and Context questions with LCC's Question Answering Server.* In Proceedings of Text Retrieval Conference (TREC-10).

Hovy, E., et al. 2000. *Question Answering in Webclopedia. Notebook.* Proceedings of Text Retrieval Conference (TREC-9).

Humphreys, R. et al. 1998. Description of the LaSIE-II System as Used for MUC-7. Proc. of 7th Message Under. Conf. (MUC-7.).

Prager, J. et al, 2003. *In Question-Answering Two Heads are Better Than One.* Proceedings of HLT-NAACL 2003, pp 24-31.

Strzalkowski, T and J. Wang. 1996. A self-learning Universal Concept Spotter. Proceedings of COLING-86, pp. 931-936.

Small S., Strzalkowski T., et al. 2004. A Data Driven Approach to Interactive Question Answering. In M. Maybury (ed). *Future Directions in Automated Question Answering*. MIT Press (to appear)

[5] Space limitations do not allow for more complete discussion of the analysts workshops and the results of the evaluations.