Natural Language Engineering **15** (1): 31–54. © 2008 Cambridge University Press
doi:10.1017/S1351324908004890 Printed in the United Kingdom

HITIQA: High-quality intelligence through interactive question answering

S. SMALL[1] and T. STRZALKOWSKI[2]

[1]*Language Analytic Corporation, Schenectady, NY, USA*
[2]*Institute of Informatics, Logics and Security Studies, University at Albany, State University of New York, Albany, NY, USA*
e-mail: small@laancor.com; tomek@csc.albany.edu

(*Received 1 March 2007*)

Abstract

We describe an interactive question answering system, HITIQA, which helps users find answers to complex analytical problems. Such problems often necessitate the user to submit not one but an entire series of questions, both simple and complex, and then to negotiate the final content and form of the answer. HITIQA advances research in human–computer dialogue by enabling topical, mixed initiative interaction over unstructured data. HITIQA uses the process of text framing to bring a level of semantic representation to open-domain data in order to facilitate meaningful dialogue with the user. In this paper we give an overview of HITIQA's design and explain the workings of its main components with particular attention given to its dialogue capabilities. We also present results of end-to-end system evaluations that demonstrate the effectiveness of the system as a whole, as well as contributions of the individual components and specifically the benefits of our dialogue-based approach. While our research continues, a number of HITIQA prototypes have recently been deployed at various government agencies where they are being tested under real operational conditions.

1 Introduction

High-quality intelligence through interactive question answering (HITIQA) is an automated question answering (QA) system developed to assist information professionals in finding answers to complex, analytical problems, both efficiently and thoroughly. In a typical session with HITIQA, a user may pose a series of interlocking questions over an extended period of time, usually an hour or more. The system uses data-driven semantic text processing (which we call framing), augmented by natural language dialogue as well as an advanced information visualization interface, to assemble accurate answers to each question out of text snippets extracted from massive data sources. The system also helps the users to organize their findings into a comprehensive report. HITIQA has undergone a series of evaluations with active duty analysts from the US Naval Reserve, producing valuable usage and performance data. A recent analysis of the accumulated evaluation data demonstrated that HITIQA creates a cognitive augmentation effect for the analyst

https://doi.org/10.1017/S1351324908004890 Published online by Cambridge University Press

by helping them to complete their tasks in less time and with less effort than would be required using conventional internet search. This effect is measured by the number of questions asked, the number of information items saved per question and per time unit and by the overall time to complete the task. In addition, we have performed component level evaluations to measure the effectiveness and accuracy of each of the system components.

1.1 Analytical question answering

Information professionals need advanced tools to search for nuggets of relevant information that may be buried in massive, constantly changing data sources in order to assemble a report, or to verify a hypothesis or to fulfill a specific information need. In addition to collecting passages or excerpts, they are required to analyze this data, form conclusions and predictions and be able to support or defend their findings to customers or managers. While this task is similar in a variety of professional areas, including business analyst, research librarian, lawyer, historian, detective, news reporter, etc., our focus will be specifically on tasks facing intelligence analysts. The task is assigned to an analyst in a context that we call a scenario, which may look something like the following:

> *Scenario 1: The al-Qaida Terrorist Group*
> Your division chief has ordered a detailed report on the al-Qaida Terrorist Group due in three weeks. This report should present information regarding the most essential concerns, including who are the key figures involved with al-Qaida along with other organizations, countries, and members that are affiliated, any trades that al-Qaida has made with organizations or countries, what facilities they possess, where they receive their financial support, what capabilities they have (CBW program, other weapons, etc.) and how have they acquired them, what is their possible future activity, how their training program operates, who their new members are.

There are several ways to approach complex information problems and the choice often depends upon the analyst's expertise and familiarity with the subject. Experienced analysts tend to decompose the problem into subproblems and aspects and then deal with each one in turn. For example, the scenario above suggests a number of analytical dimensions for the 'al-Qaida' topic, including key figures, facilities or financial resources. Less experienced analysts facing an unfamiliar problem may attempt to learn about the topic first by asking exploratory questions about the concepts mentioned in the scenario. A successful system should support both modes of operation. Whilst analysts usually possess significant domain expertise, they cannot always anticipate all relevant aspects of the current problem, which may differ in unexpected ways from prior cases. When assigned an information task, the analyst typically forms a hypothesis (or a series of hypotheses) about the structure of the problem. Such hypotheses can be 'verified' against available data by asking a number of questions to confirm the assumptions made. While this approach seems

https://doi.org/10.1017/S1351324908004890 Published online by Cambridge University Press

reasonable and efficient, it also presents significant dangers of straightjacket analysis and analytical bias. This is a serious problem because analysts frequently overlook, for lack of time or motivation to consider alternative, perhaps 'inconvenient' answers, while focusing on those that directly support their current hypotheses. A successful analytical QA (AQA) system needs to address the problem of analytical bias. It must not only provide direct answers to the analyst's questions, but it should also uncover highly related contextual information that may reveal unforeseen aspects of the data. In other words, an AQA system should return information that the analyst is expecting, as well as additional information which, in context, may be judged as relevant although possibly unexpected, and even further information that was never anticipated. This latter information may be difficult to judge as relevant without some further inquiry; nonetheless, it may expose additional facts that can strengthen or weaken the current hypothesis. Our objective in this work was to design a system that could return all these kinds of information in response to a user's analytical questions.

1.2 Analytical questions

In recent years QA has been often categorized as either factoid (or factual) QA or complex, AQA. There are significant differences between these two modes of QA, at least from the viewpoint of automated system design. Factoid questions typically seek salient, isolated facts expressed by a few words: a person name, a birth date, a distance between two points that need to be pinpointed among many more nonanswer words. The process of automated QA for most types of factoid questions is now fairly well understood. It has been noted early that most factoid questions display a distinctive 'expected answer type' (or EAT), which is the type of the information piece required to answer it. EAT can be deduced from the form of the question using a finite list of patterns that take into account the question word (where, when, how, who, etc.) as well as the adjective or noun they modify (e.g., how many, which designer, etc.). For example, 'Who was the first man in space' expects a 'person' as the answer type (who question), while 'How long was the Titanic?' expects some length measure as an answer, probably in yards and feet or meters. Finding an answer is thus reduced to locating a word or phrase with an appropriate EAT near other words that are mentioned in the question (e.g., Titanic). In other words, designing a good factoid QA system requires an exhaustive and carefully designed typology of questions and a corresponding list of answer types. 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 evaluations (Hovy *et al.*, 2000; Harabagiu *et al.*, 2002; Prager *et al.*, 2003).

These evaluations also made it quite clear that the more fine-grained the question answer typology the better the results and that embedding a significant amount of domain knowledge into it can make a system perform extremely well. As may be expected, capturing detailed domain knowledge into question-answer patterns is an expensive and time consuming process. Analytical questions require a very different

https://doi.org/10.1017/S1351324908004890 Published online by Cambridge University Press

kind of processing than isolated, fact recall questions. An answer to an analytical question is not a single fact to be plucked out of pages of text, but potentially many such facts that need to be assembled together to form an answer. Moreover, the 'facts' are not just entities or their properties, but more often relationships or events that tie them together. Many such events may appear disconnected and insignificant when taken out of context, therefore an AQA system needs to supply sufficient contextual information with each answer piece. This is necessary to enable the users to reason about the information just presented, perform additional data collection if necessary and ultimately being able to defend their arguments. This need for contextual information places significant emphasis on the users' intentions when a question is asked. Many analytical questions may appear to call for specific facts or simple confirmation, for example, 'Is North Korea building nuclear weapons that can reach the United States?' A simple 'Yes' response to this question, in the context of the analytical task to assess the US defense needs, would probably be inadequate. Instead, some evidence supporting the answer would be expected, for example, technical specifications of long range missiles in the adversary's possession, etc. What's more, an apparently simple 'factoid' question such as: 'How far is Basra from Baghdad?' when asked in an analytical context of, for example, supplies delivery planning, would likely take on an entirely different meaning that may completely change its expected answer type: the analyst may be actually interested in how long it would take to deliver the supplies or how many armored vehicles to send along for protection. To decide this, the analyst may require such contextual information as presence of minefields along the way, possible insurgents lurking around, the current condition of the road, etc. It should be understood then that whether a question is analytical or not, may not always be obvious or discernable automatically but it is truly in the intention of the questioner given the context of the task. Historically, intent and context have been studied a great deal in speech acts theory. Context is vital to understanding the meaning or intention of an utterance or question (Searle, 1969). For our purpose, an analytical question is one that requires an answer to be assembled (not simply located in text), and this is influenced by the overall analytical task as we argued above. This further means that a successful AQA system must learn as much as possible about the overall task the user is pursuing, which we claim may be partially inferred from the current state of analysis: data collected, questions asked, saved material, etc., as well as through an explicit interactive dialogue with the analyst.

1.3 The implicit analytic scenario

An additional challenge for AQA is therefore the implicit scenario problem. It means that the complete specification of the analytical task (the scenario) is generally not available to the system. While the analysts normally have a task description, they may be pursuing any number of intermediate hypotheses that are often only partially formed in their mind. To make things harder, the data set is not static and it changes at various speeds as new information is added continuously and older outdated information is removed or superseded. As information changes, so

https://doi.org/10.1017/S1351324908004890 Published online by Cambridge University Press

does the structure and scope of the task that the analyst is working on. This fluid nature of the analytical domain virtually precludes any advance domain-modeling approach – what may be considered costly and time consuming for 'factoid' QA is simply not workable for AQA; even if models could be built, with great effort and cost, for every domain of analytical interest, they would soon become obsolete thus restricting rather than extending system analytic capabilities with outdated knowledge structures.

2 The HITIQA system

HITIQA is an advanced QA system that helps analysts to produce higher quality reports for complex intelligence problems in less time and with lower cognitive load. HITIQA uses data-driven semantic processing and natural language dialogue, coupled with an advanced information visualization interface, to deliver accurate answers to the analyst's questions along with related contextual information. The primary function of HITIQA is to supply composite answers to complex, exploratory questions that analysts pose while working on their tasks. HITIQA's basic architecture consists of several components that include document retrieval, passage segmentation, named entity extraction, event framing and a multimodal dialogue manager. Below we describe each of them in detail.

2.1 Document retrieval

When the analyst poses a complex question against a massive source of unstructured text data, an AQA system needs to first narrow this inquiry to a smaller but possibly still quite large set of items where the answer is likely to be found. The objective here is to cast a net wide enough to assure the maximum recall, while leaving the task of sorting things out to later, more accurate (and more costly) processes. For most QA systems including HITIQA, keyword-based document retrieval is a method of choice because it guarantees fast though definitely noisy results. The current version of HITIQA uses the INQUERY system (Callan, Croft and Harding, 1992) to retrieve the top fifty documents from the data source. The retrieved documents are subsequently broken into smaller passages, in most cases using naturally occurring paragraph breaks. The system then recognizes and removes duplicate paragraphs using an n-gram matching algorithm (Clough *et al*., 2002). Out of the remaining passages, the first 200 are retained and passed to the passage clustering component.

2.2 Passage clustering

HITIQA clusters the retained text passages into semantically correlated groups using a version of hierarchical clustering (Hardy *et al*., 2002) with a strong similarity function based on n-gram overlap, with n-grams up to the length of six. The method first produces a small number (usually six or fewer) tight seed clusters of three or more elements, upon which the remaining passages are distributed among the existing clusters as satellites until the lower similarity threshold is reached. At this

https://doi.org/10.1017/S1351324908004890 Published online by Cambridge University Press

point, each cluster represents a salient theme within the set of passages; however some of these themes may be only tangentially related to the user question due to the low precision of the keyword retrieval. The clustering step accounts for some of the passages returned from the data source, perhaps 20%, and therefore many more remain unassigned. From these, HITIQA selects a small number of passages (currently ten) that display the highest similarity to the original question (the top ten clusters) but did not make it into any of the seed clusters. These singleton passages along with the cluster seeds and their satellites are sent to the next component, the text framing step.

2.3 Text framing

We developed a dynamic method for imposing a partial structure on the underlying text data, which we call text framing. These structures (or frames) provide crucial metadata that allows HITIQA to systematically compare different text passages, both against each other and against the user question. Framing does not attempt to capture the entire meaning of the passage; instead it makes salient that portion of the meaning which appears to match the analyst question most closely. The frame also provides a handle to the underlying text that enables HITIQA to communicate its content to the user and to highlight any differences between the question and a potential answer. This, in turn, allows HITIQA to inquire of the user if such a potential answer is germane. Frames derived from user questions are called the goal frames. Text framing consists of two subprocesses: building a general frame over a selected text passage and if possible deriving one or more event frames from it.

A general frame represents one or more relationships involving any number of entities referenced in a text passage, such as LOCATION, PERSON, ORGANIZATION, DATE and so on, which make up the frame's attributes. Attributes are extracted from the passage text using any good quality named entity (NE) extraction system such as BBN's IdentiFinder (Miller *et al*., 1999), which tags twenty-four types of named entity classes or University of Sheffield's GATE toolkit. After named entities are tagged, any remaining well-formed expressions, including continuous noun groups and verb groups are considered as relation indicators between the entities in the passage (Small *et al*., 2003). The tacit assumption here is that a relationship must exist and that is why the entities are mentioned together. Basically, any continuous noun or verb group sequence, except for the words referring to the named entities themselves and certain excluded stop-words, can become values of the RELATION attribute. This data-driven approach allows the system to work effectively in any domain and with any text genre, without necessitating costly knowledge engineering. When a question is submitted and processed into a goal frame, HITIQA compares the relation indicators in all the available data frames against the relation indicators in the goal frame. This process supports selection of the answer elements that contain relations of interest to the analyst. In order to achieve a match, HITIQA uses a variety of an approximate substring matching techniques while consulting the Wordnet lexical database to substitute synonyms or hypernyms, if necessary. If no match is found we assume there is no common

https://doi.org/10.1017/S1351324908004890 Published online by Cambridge University Press

Table 1. *Prototypical event frames in HITIQA*

Event	Example triggers	Key RoJes
Agree	treaty, agreement, sign, ratify	PARTIES, INS TR
Assist	helped, supporting, assisted, aid	AGENT, TARGET, INS TR
Attack	attacked, invaded, bombed, destroyed	AGENT, TARGET, INS TR
Develop	construct, develop, manufacture	AGENT, OBJECT
Financial	funded, financed, laundered money	SOURCE, TARGET, QUANTITY
Law-Cnfninai	arrested, detained, caught, charges	AGENT, TARGET, CHARGE
Law-NatWnt'l	inspectors, embargo, legislation	AGENT, TARGET, WHAT, CHARGE
Political	election, fired, hired, appointed, voted	AGENT, TARGET, POSITION
Threat	threaten, fear, warn	AGENT, TARGET, INS TR
Transfer	acquire, seize, steal, export, smuggle	SOURCE, OBJECT, DEST

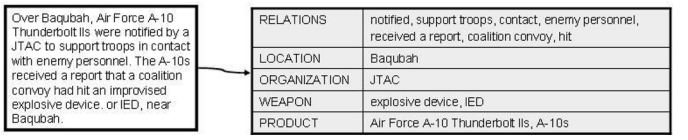


Fig. 1. A general frame derived from a text passage may represent one or more relationships between its attributes.

relation between the goal frame and the data frame, in which case the data frame inherits the value of the cluster relation from the passage cluster where the current passage belongs. The cluster relation is the relation indicator phrase that occurs most frequently in the cluster. Obviously, if this last option is used, the data frame just completed will have at least one conflict with the goal frame. An example of a general frame derived from a text passage is shown in Figure 1.

While general frames enable us to capture a wide range of potential events, we can achieve further improvement in representation accuracy and efficiency by recognizing several specific entity relationships in greater detail. We define a handful (currently ten) prototypical events (Small *et al*., 2004; Hardy, Kanchakouskaya and Strzalkowski, 2006) that have been empirically determined to underlie many reported events across a number of analytic domains (Table 1).

When an event of any of these types is mentioned in a question or in a text passage retrieved from the database, a more specialized frame is built that (a) explicitly recognizes the type of event occurring, (b) selects attributes from among those available in the general frame (or possibly extracting additional items from text) that will be part of the specialized frame and (c) assign roles to these attributes with respect to the frame relation, for example, Agent, Target, Source, Destination, etc. Therefore, unlike in a general frame where the relationships among the entities and their roles remain underspecified, in the event frame we impose a predefined semantic structure upon the attributes: a ready-made template is selected from a repository and then instantiated against the elements of the general frame. We

https://doi.org/10.1017/S1351324908004890 Published online by Cambridge University Press

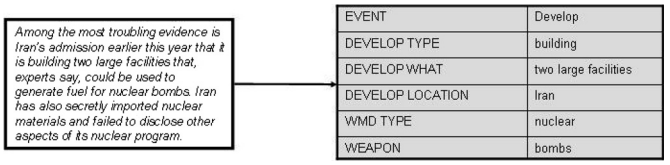


Fig. 2. Develop frame built from a retrieved passage for the question: 'Is there evidence that Iran is developing nuclear materials?'

should note here that an event frame may involve only a subset of named entity attributes found in the general frame; it is therefore possible to obtain more than one event frame from a single general frame.

Where the general frames are assembled directly from textual elements, the event frames require predetermined extraction rules that can either be manually constructed or acquired from annotated text corpora. In a manual approach, a set of trigger words is first identified for each type of event. For example, for the TRANSFER event, verbs such as move, transfer, export, import, sell, transport, acquire, deliver, etc. may form an initial set of triggers. These triggers are used to pull out a large number of example sentences and passages from a training text corpus. Among these sentences, many will indeed describe transfer type events, and these sentences are then manually annotated to indicate the event phrase, the attributes and their roles. From these representative examples (as well as from some negative examples captured) we construct extraction patterns for event frames of each type, based on trigger words, key prepositions, named entity tags and syntactic structures. These manual rules can be quite accurate with approximately 90% precision although much lower recall at about 50%; the accuracy of role assignments can vary from low 60% to near 100%. In order to improve frame recall we have also developed an automated acquisition method which extracts initial patterns from the annotated training corpus and then uses unsupervised bootstrapping (based on (Strzalkowski and Wang, 1996) and see also (Yangarber *et al.*, 2000)) to refine the rules with the objective of improving their recall. The bootstrapping approach was first attempted on the set of manual rules to increase their recall rate, but this turned out to be difficult to accomplish due to inherent incompatibilities of the two methods. However, when applied to the syntactic patterns derived directly from the training corpus, the bootstrapping produces an extended set of patterns that show good performance with recall increasing to 70%. As indicated before, a single text passage with its general frame may yield several event frames involving different event types and different attributes. In the absence of a user question any of these alternative frames are potential candidates for representing this passage, and thus they remain 'folded' within the general frame. When the passage is retrieved in response to a user's question, HITIQA will select and rank those potential event frames that best match the question frame. For example, if the analyst is interested in some type of transfer, and a transfer event frame can be assigned to the passage, then this frame will be selected over other event frames that may exist in the same passage. Figure 2 shows an example of an event frame instantiated in response to a specific user question.

https://doi.org/10.1017/S1351324908004890 Published online by Cambridge University Press



Fig. 3. A general goal frame built for the question: 'How does the training program of Al-Qaida operate?

2.4 The goal frame

As already signaled, a very similar framing process is applied to the questions posed to the system. This allows for systematic comparison between the analyst's question and the retrieved text passages and for selection of instantiated frames for answer passages. For ease of reference we shall call the frame built for a question the goal frame, while the frames built over text passages in the database will be called data frames. As before, we extract named entities from the question and identify relation indicator strings. A goal frame can be a general frame or it can produce one or more event frames. For example, given the question: 'How does the training program of Al-Qaida operate?', HITIQA will identify one named entity Al-Qaida (an ORGANIZATION) and two possible relation indicators: training program and operate, and create a general goal frame as shown in Figure 3.

2.5 Comparing of frames

A goal frame is compared to data frames in the retrieved passages to select the best matching representations for each passage, and then to rank the passages by the matching score. We should note that frame-to-frame matching is done at the level of frame attributes, including the frame (or event) type, not by comparing individual words. In order to match specific attributes a selection of lexical and syntactic string matching techniques are utilized and supplemented by available lexical resources, such as Wordnet. When comparing the values of corresponding attributes we are utilizing a variety of string matching techniques, expansions to synonyms using Wordnet, as well as gazetteers for typical location attributes. This takes care of most standard surface variations, including synonyms, some insertions and deletions, word-order changes and in the case of geo-political entities the inclusion information. In this way, the comparison between frames can be reduced to the number of mismatches at the attribute level. Specifically, data frames are scored against the goal frame by counting the number of attribute mismatches, or conflicts, between them. When there are no conflicts, we shall say that the data frame has a 0-conflicts score with the question; when there is a mismatch on a single attribute, we shall say that the data frame has a 1-conflict score with the question, etc. When all attributes mismatch, we shall say that the data frame has a 99-conflicts score with the question. All framed passages irrespective of their conflict score form HITIQA's answer space with respect to the current question. Figure 4 illustrates the question to passage matching.

The passages with 0-conflict frames form a direct answer to the user's question. However, the other frames in the answer space provide a mechanism through which HITIQA can uncover additional, perhaps tangentially related topics that the user may not have explicitly asked for. If these topics or aspects align closely with the

https://doi.org/10.1017/S1351324908004890 Published online by Cambridge University Press

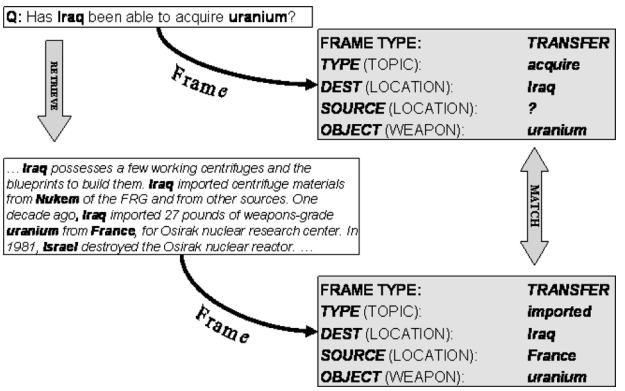


Fig. 4. Question and a matching text frames.

user's question, (i.e., matching many of the goal frame attributes) HITIQA will attempt to make the user aware of them and let them decide if they should be included in the answer. Clearly, data frames with low conflict scores, 1 or 2, are prime candidates for an extended answer, but it may not be always obvious if they should be shown to the user or not. Perhaps these are the answers to some future question the user may ask within the current task; or they may suggest a whole new line of questions for the user. On the other hand these may be simply some peripheral items that the user is not interested in at all. In order to resolve such issues a dialogue with the user is necessary. We will return to this problem in a subsequent section.

In summary, text framing serves several important functions in HITIQA, including:

(1) a simplified semantic representation of otherwise unstructured language data;
(2) a more accurate assessment of relevancy and degree of relevancy of various text passages;
(3) a conceptual structure to support meaningful dialogue with the user.

2.6 Generating an answer

In answer generation, we wish to organize the text passages extracted from multiple sources into a reasonably coherent answer before they are presented to the analyst. For example, reporting events arranged in time and/or space would make a better answer than simply a list ordered by putative relevance. Similarly, contradictory or hypothetical information may be better presented in a single block to achieve an appropriate contrast effect. This is not an easy task, and quite possibly somewhat futile since analysts would normally assemble reports in order to support a particular hypothesis. Moreover, the information collected during a QA session with the system may provide only part of the argument structure that the analyst is looking for, with other elements supplied from prior knowledge and/or from different sources. In the current version of HITIQA we do not attempt to generate a single narrative answer; instead, a nested structure of folders is constructed, each folder representing

https://doi.org/10.1017/S1351324908004890 Published online by Cambridge University Press

a particular subtopic encountered during a QA session. A subtopic may be formed when a number of semantically correlated data items are retrieved (e.g., reports of weapons shipments to Iraq) or when the analyst explicitly identifies it, for example, by accepting the system's clarification offer ('Would you be also interested in?'). Data items (text passages and their frames) are then assigned to appropriately labeled folders, which are created on the fly. These folders may subsequently be updated as new information is discovered. We should note that the folders mechanism allows the user a partial (though indirect) control over how the answer space is organized: in most cases observed during analytic exercises the final structure of folders differs from the system-generated topic clusters (or from the organization of the visual display discussed later). Analysts often assemble their reports by selecting items from the relevant folders.

3 Interacting with the analyst

3.1 Coping with the implicit analytical scenario

Earlier we outlined some of the problems that arise when answering complex analytical questions without a priori knowledge of either the domain or analytical task. In HITIQA we attempt to address these issues by examining which of the topics captured in the system's answer space but outside of the 0-conflict set, that is, the passages represented by only partially matching frames, may be relevant to the user's task. If a topic represents a part of the problem the user is trying to solve – or we may say, a part of an appropriate domain model – the user should be able to recognize whether the corresponding passage belongs to the current answer or not. In other words, if the system asks the user whether a certain apparently related piece of information is relevant, this question should not come as a surprise or a distraction. In order to maximize the chances of asking the right clarification question HITIQA calculates how closely each passage may be related to the analyst's information need. Clearly, passages with 1- or 2-conflict frames are the primary candidates for clarification questions. In a later evaluation section we shall demonstrate that this is indeed an optimal choice. To further illustrate our approach it may help to schematically visualize the answer space for an analytical question, as consisting of three zones (Figure 5). The center zone contains information that directly answers the analyst's question; in other words, the passages represented with 0-conflict frames. The middle zone represents information that appears highly related but not perfectly matching the question: these are the passages represented by frames with some conflicts with the question. The outmost zone consists of all remaining passages in the answer space which still may share some common terms with the question, but have too many conflicts with the goal frame and usually no common attributes on the frame level.

The three zones represent the system's interpretation of the answer space; what the analyst is interested in, however, is depicted by the dashed area that cuts through the zone boundaries – ideally this area and the central zone should converge. In order to accomplish this, the system initiates a series of clarification questions to negotiate the expected scope of the answer with the analyst.

https://doi.org/10.1017/S1351324908004890 Published online by Cambridge University Press

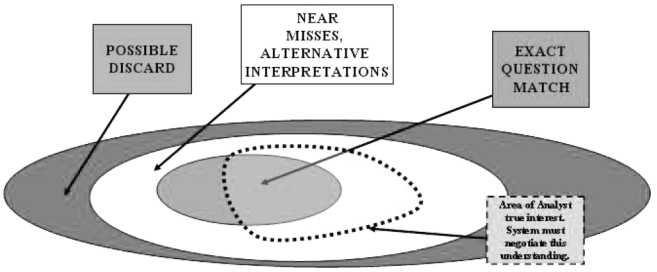


Fig. 5. Analytical question answer space; system and analyst interpretation.

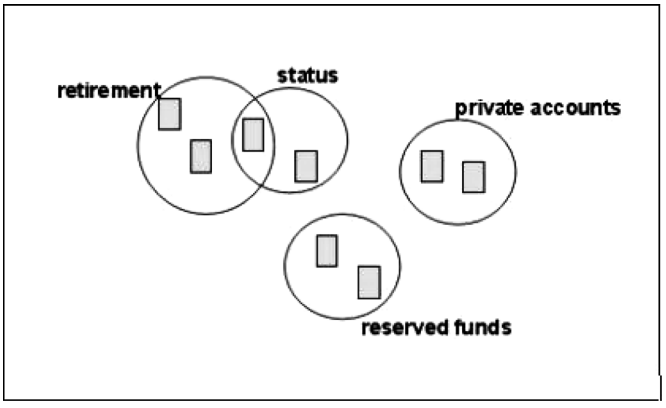


Fig. 6. Cluster topics discovered in the answer space for the question: 'What is the status of the social security system?'

3.2 Negotiating the answer scope with the user

HITIQA uses interactive dialogue with the analyst to negotiate the scope of the answer and to address the problems of the implicit scenario. Our experiment with several groups of analysts showed that users accept system suggestions regarding both the current question scope as well as these that relate to other aspects of their task – thus speeding up their work. In HITIQA the dialogue is enabled through the frame-based representation of content and more specifically through the mismatches between the goal frame and the data frames in the answer space. In addition, we exploit the passage clusters obtained within the answer space in order to reach beyond the data items that closely match the question itself. Each cluster represents a topic within the answer space: possibly reflecting several alternative or complementary interpretations of the analyst's question, as shown in Figure 6. It is important to note that the topic/cluster boundaries may cross the boundaries of the conflict zones in Figure 5 in arbitrary ways. The resulting dual structure opens up additional possibilities for the clarification dialogue: beyond discovering what potentially relevant items may lie nearby, the clarification dialogue can also be used to tacitly solicit the user's help in properly circumscribing his or her actual (but underspecified) information needs beyond the (apparent) scope of the current question.

https://doi.org/10.1017/S1351324908004890 Published online by Cambridge University Press

Analyst: *"What is the status of the social security system?"*
[exact match answer passages are displayed]
HITIQA: *"Would you be interested in information on* ***retirement*** *relative to your question?"*
Analyst: *"Yes"*
HITIQA: *"Do you want to see information on* ***private accounts****?"*
Analyst: *"Yes"*
HITIQA: *"Thank you, Please view your answer"*
[final selection of passages are displayed]

Fig. 7. Question interaction with the HITIQA system as demonstrated live to the ARDA AQUAINT community, Washington, DC.

RELATIONS	status
ORGANIZATION	Social Security System

Fig. 8. Goal frame for the question: 'What is the status of the social security system?'

Last week, the General Accounting Office, the auditing arm of Congress, issued a chilling report. Without badly needed reform, Social Security is rapidly headed for financial ruin. By 2017, when baby boomers are well into ***retirement,*** *the program will start running annual deficits that will eventually bankrupt the system, the GAO warned.*

RELATION	retirement
ORGANIZATION	General Accounting Office, Social Security, Congress, GAO
DATE	2017

Social Security is sound for today's seniors and for those nearing ***retirement****, but it needs to be fixed for younger workers – our children and grandchildren. The government has made promises it cannot afford to pay for with the current pay-as-you-go system.*

RELATION	retirement
ORGANIZATION	Social Security

Fig. 9. Two cluster seed passages and their corresponding frames relative to the retirement clarification question.

We now review in detail two examples that illustrate how HITIQA interactive process works. The first example is shown in Figure 7. In order to understand how this interaction was generated we need to examine the goal frame created for the question (shown in Figure 8) and the frames built for each of the retrieved passages (shown in Figures 9 and 10).

The passages in Figures 9 and 10 are represented by 1-conflict general data frames in the analyst's answer space. The conflicts on the relation attribute were, respectively: status versus retirement and status versus private accounts. These relations were obtained from the passage's cluster relations: one cluster containing retirement (Figure 9) and the other cluster containing private accounts (Figure 10). From the system's viewpoint each of these clusters represents an alternative interpretation of the user's question about the status of social security. This explains the clarification dialogue with the user shown in Figure 7. We may note that both clarification questions are very much on target and that the user accepts both suggestions.

The next example shows how HITIQA can utilize event frames (as opposed to just general frames) to generate very specific suggestions for the user (Figure 11).

https://doi.org/10.1017/S1351324908004890 Published online by Cambridge University Press

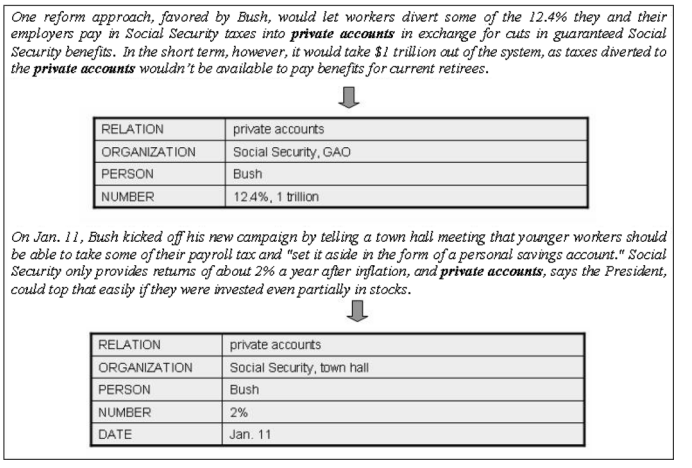


Fig. 10. Two cluster passages and their corresponding frames relative to the private accounts clarification question.

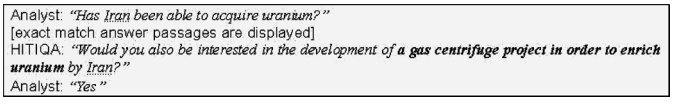


Fig. 11. Question interaction with the HITIQA system, expanding to a different event type.

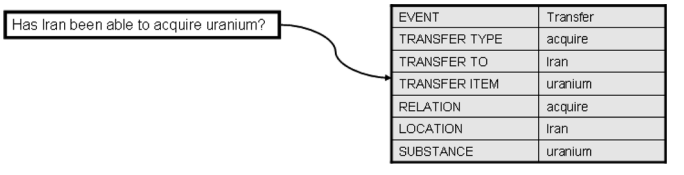


Fig. 12. Goal frame for the question: 'Has Iran been able to acquire uranium?'

The goal frame for the user's initial question is a TRANSFER event frame shown in Figure 12. We should note that all general frame attributes are still accessible for comparison.

One of the passage returned for this question is represented by a DEVELOP data frame, so obviously there is a conflict here, even though many of the attributes are shared (Figure 13).

By matching off similar role attributes, that is, transfer_item to develop_item and transfer_to to develop_agent, HITIQA is able to reduce the number of conflicts to just one. This is possible because HITIQA can match frame attributes on the general frame level when it attempts to compare event frames with incompatible event types.

3.3 Interactive visualization

The system's verbal dialogue is augmented by an interactive visual display, which allows the analyst to view the answer space based on a visual rendering of the

https://doi.org/10.1017/S1351324908004890 Published online by Cambridge University Press

The IAEA is likely to renew its investigations into whether or not Iran is developing a gas centrifuge project in order to enrich uranium. An IAEA team of investigators is scheduled to travel to Iran to investigate suspicions that the Iranians are setting up a secret plant to enrich uranium through centrifugation which could be used for the development of nuclear weapons.

EVENT	Develop
DEVELOP TYPE	developing
DEVELOP ITEM	a gas centrifuge
DEVELOP AGENT	Iran
RELATION	developing
LOCATION	Iran
SUBSTANCE	uranium
ORGANIZATION	IAEA
Conflict with [event_type]; Matches with [relation, location, substance, develop_item to transfer_item, develop_agent to transfer_to	

Fig. 13. Data frame built on a passage retrieved for the question, 'Has Iran been able to acquire uranium?'

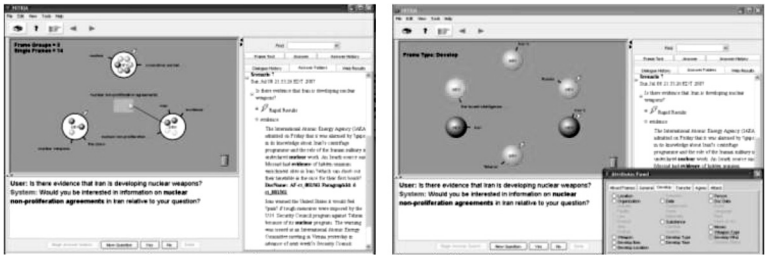


Fig. 14. Visual Interface, group view (left) and frame view (right).

frames (Small *et al*., 2004). In the most recent version of the system verbal dialogue and visual representation are seen simultaneously on the same screen, where they are presented in adjacent panels. This allows the analyst to easily switch between interaction modes, as well as to simultaneously see the effects of each dialogue move on the answer space. The visualization portion of this panel uses color-coded circles to represent each of the frames in the answer space. Each frame corresponds to one solid colored circle in this display where the color indicates frame relevance to the question. The dark blue represents a 0-conflict frame, a light blue circle represents a 1-conflict frame, a green circle represents a 2-conflict frame and so forth (the actual color scale can be adjusted to the users' preferences). At the other end of the spectrum, a red colored icon represents a 99-conflict frame. The view shown on the left in Figure 14 is a visualization of the entire answer space at its highest level, where all frames in the answer space are represented; we call this our group view. There are two view levels currently available: the group view and the individual frame view. In the group view, the frames are grouped by event type (with a three letter abbreviation used for the event type) and the value of the selected key attributes, which are displayed on the spokes: the develop type attribute for DEVELOP frames, the transfer type for TRANSFER frames, the agreement type for AGREE frames and so forth, and finally the relation attribute for the general frame. This organization of the display helps the user to quickly see the amount and type of information in the answer space. Additionally, the folder icon displayed in the center shows the frame(s) that are the subject of the current dialogue clarification question. By clicking on this center group the analyst can view the text underlying the frame prompting the current question.

https://doi.org/10.1017/S1351324908004890 Published online by Cambridge University Press

While the system may seek explicit clarification about some of the 1- and 2-conflict frames (light blue and green), the users are free to explore the visual answer space on their own. We facilitate this by making the visualization interactive: users can 'drill-down' into the individual frame view (Figure 14, right) by clicking on a frame group or the center dialogue group. From this view users can focus on individual frames, read the underlying text passages and change their relevance scoring with a mouse click. The analyst can use this panel to quickly see the values of the attributes that exist in the underlying passage for each frame.

4 Evaluations

The HITIQA system has undergone a variety of evaluations. We will give the results of two separate evaluations: (1) the ARDA Challenge workshop results will be used to show overall performance when compared to a keyword search baseline system and (2) the component level evaluation where we focus on impact of the clarification dialogue on the content of the final answer.

4.1 ARDA Metrics Challenge workshop

Our system participated in the ARDA Metrics Challenge workshop, run by the National Institute of Standards and Technology (NIST) over several weeks in June 2004. We report here the results of our most recent data analysis from this workshop. The workshop focused on two aspects of evaluation: the analysts' process and the analysts' products. The objective measures of efficiency and effectiveness are measures of the process, whereas the more subjective rating of the analysts' reports assess if better quality information is being located and used in the final product. The data set for this evaluation was created by a team of NIST researchers and was approximately 4 GB in size. The majority of files had been mined from the web and centered on the subject of nonproliferation. This was augmented with a smaller set of completed reports received from the Center for Nonproliferation Studies (CNS). Four systems participated in the study, which included the NIST chosen baseline system, GNIST, which was Google search limited to the test collection. Eight analysts were recruited from the US Naval Reserve to be the subjects in the study. Air Force Rome Labs (AFRL) created eight scenarios, each of which was reviewed by a panel of experts. Analyst had two days to work with each system, following a day of training and warm up exercises. Analysts used each system in a series of 2.5-hour sessions during which they were tasked to collect data and prepare an analytical report for a given problem (a scenario). Following each session the participants were asked to complete a number of evaluation instruments. These instruments included questionnaires, report cross-evaluation forms and an effort assessment tool. The cross-evaluation method was used to obtain judgments of report quality where analysts would score each other reports in a number of quality categories including content, coverage and analytical value. Below we discuss selected results and statistics collected from this evaluation that are particularly relevant to the work reported here.

https://doi.org/10.1017/S1351324908004890 Published online by Cambridge University Press

3. The [X] system helped me to better understand the scenario by the information that it provided.
7. The [X] system helped me to think about the scenario in new ways.
11. The [X] system stimulated my thinking about the scenario
15. The [X] system expanded my understanding of the scenario.

Fig. 15. Four session evaluation questions related to NIST Hypothesis 10: QA systems should help the analyst recognize gaps in their thinking.

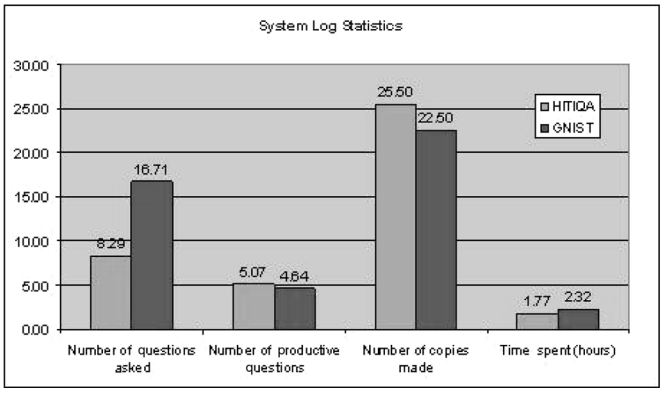


Fig. 16. System log statistics; average per scenario.

4.1.1 Selected evaluation results

HITIQA scored significantly higher than the baseline on a number of session evaluation questions (Figure 15). These questions assessed the system's ability to uncover contextual information related to the scenario, thereby helping analysts to avoid bias and to recognize potential gaps in their thinking.

HITIQA also scored significantly higher than the baseline on several system evaluation questions. For the question: 'The system provided me with suggestions about which I had not thought', HITIQA's score was 3.57 (out of 5), significantly above the baseline GNIST's 1.86. In addition for the question: 'The system often presented me with novel information', HITIQA scored significantly higher at 3.00, than GNIST's 2.29. Providing new and related information to the analyst turned out to be one of the most visible advantages of the HITIQA system.

4.1.2 System statistics

A variety of system usage statistics were computed from the system logs that were automatically recorded during analytic sessions. The chart in Figure 16 compares the total number of questions submitted to each system to the number of productive questions, that is, the questions that resulted in an analyst finding and copying information to their report.

We note that the analysts working with GNIST asked on average about twice as many questions as the analysts using HITIQA. At the same time, most of the questions, submitted to GNIST resulted in no useful information that the analyst

https://doi.org/10.1017/S1351324908004890 Published online by Cambridge University Press

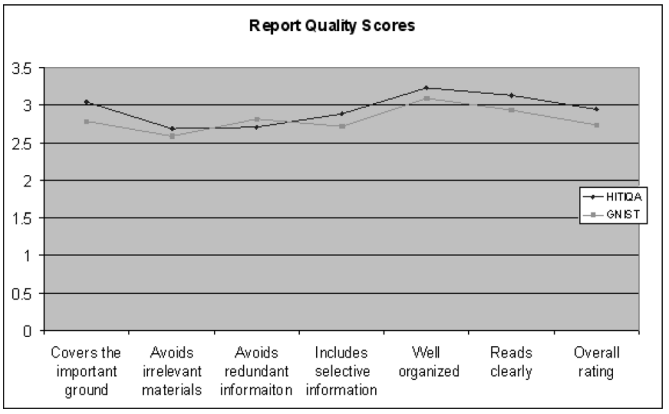


Fig. 17. Cross evaluation results: Scale 0–5; inverted so all higher scores are best.

would copy into their reports and only 27.8% of them were productive. By contrast, 61.2% of questions that analysts asked of HITIQA were productive. In addition, the analysts spent approximately 30% less time on their tasks while using HITIQA than when using GNIST; nonetheless, HITIQA users still produced reports with the higher overall quality (Figure 17).

4.1.3 Report quality

Analytic reports were evaluated using cross-evaluation forms completed immediately following each session. The analysts were asked to rate all reports (including their own) on a number of criteria, including coverage, precision, nonredundancy and organization. Figure 17 lists all seven criteria and shows that the reports produced by analysts using HITIQA were rated significantly higher than those produced by analysts using GNIST on five out of the seven quality criteria, including the overall score, while the difference on the remaining two criteria was not statistically significant. When combined with the performance statistics in Figure 16, these results suggests that HITIQA delivers a degree of cognitive augmentation to its users.

We must note here that while these results indicate that HITIQA performs better than the baseline GNIST, the report quality and other performance metrics are affected by a number of independent factors including difficulty of the task, data available to the system, capabilities of individual analysts, among others. Whenever possible, the results reported here have been normalized to separate impact of these effects and thus should be considered reliable. For further details the reader is referred to the original report from the ARDA workshop (Morse and Sholtz, 2004).

4.1.4 ARDA workshop conclusions

The combination of performance statistics and report quality scores demonstrates that HITIQA enables analysts to produce higher quality reports in significantly less time and with less effort than the current keyword-based search technology.

https://doi.org/10.1017/S1351324908004890 Published online by Cambridge University Press

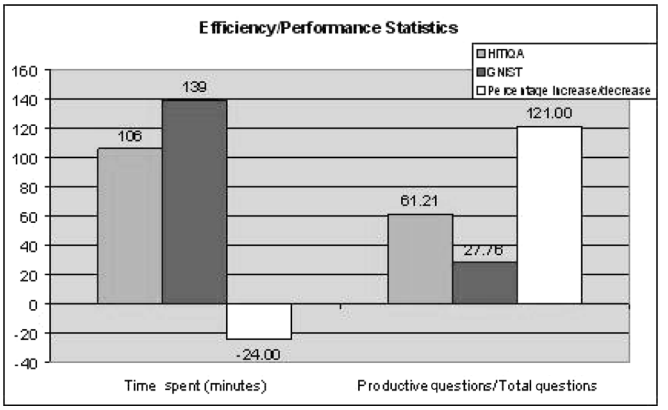


Fig. 18. Summary of performance statistics HITIQA versus GNIST.

HITIQA is addressing some of the major deficiencies of keyword search, where users are spending significant amount of time and effort for relatively low pay off: 72.2% of the questions they ask results in no usable information. This problem is particularly acute with analysts who work under time pressure to produce their reports. By contrast, analysts working with HITIQA spent on average 24% less time and their questions led to reportable information significantly more often: they needed to ask only about half as many questions to obtain a comparable quality and quantity of information. This means that HITIQA helps analysts to be more productive, as shown in Figure 18.

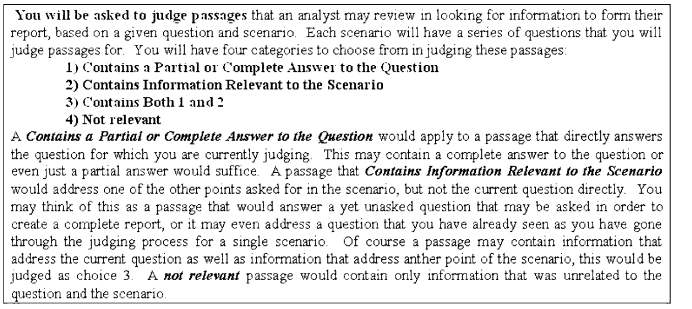
4.2 Component evaluation

We have also conducted component level evaluations to assess the performance of individual HITIQA modules. In these evaluations we were primarily interested in measuring the accuracy of the emerging answer space, in terms of recall and precision relative to each question as well as to the overall task scenario. In order to support this evaluation, we created a judged text corpus and used it as a test database for HITQA.

4.2.1 The corpus

We have created an annotated corpus of questions and text documents. The documents were manually annotated by a panel of human assessors to provide a gold standard against which the performance of each component could be measured. In order to create a representative sample we randomly collected hundred user questions from HITIQA logs that were acquired through sessions with USNR analysts. The test corpus was assembled by pulling together fifty (or fewer) documents for each question from a larger data set of approximately 750 MB consisting mostly of web-mined documents as well as a small subcollection of intelligence reports from the Center for Nonproliferation Studies (CNS). The documents were retrieved using the

https://doi.org/10.1017/S1351324908004890 Published online by Cambridge University Press


You will be asked to judge passages that an analyst may review in looking for information to form their report, based on a given question and scenario. Each scenario will have a series of questions that you will judge passages for. You will have four categories to choose from in judging these passages:

1) Contains a Partial or Complete Answer to the Question
2) Contains Information Relevant to the Scenario
3) Contains Both 1 and 2
4) Not relevant

A ***Contains a Partial or Complete Answer to the Question*** would apply to a passage that directly answers the question for which you are currently judging. This may contain a complete answer to the question or even just a partial answer would suffice. A passage that ***Contains Information Relevant to the Scenario*** would address one of the other points asked for in the scenario, but not the current question directly. You may think of this as a passage that would answer a yet unasked question that may be asked in order to create a complete report, or it may even address a question that you have already seen as you have gone through the judging process for a single scenario. Of course a passage may contain information that address the current question as well as information that address anther point of the scenario, this would be judged as choice 3. A ***not relevant*** passage would contain only information that was unrelated to the question and the scenario.

Fig. 19. Instruction excerpt provided to judges.

INQUERY document retrieval system with the objective of obtaining approximately 200 text passages per question, thus sometimes fewer than fifty documents were needed. The average number of passages retained per question was 214, thus the entire corpus consisted of about 21,400 passages of varying length. All passages in the corpus were subsequently judged for relevance to (1) the questions that were used to retrieve them and (2) the scenarios for which these questions were generated.

4.2.2 The Judges

Judges were recruited from among graduate students and faculty of information science. They were asked to annotate the entire collection and given a set of instructions as shown in Figure 19.

We asked three subject domain experts to establish an annotation baseline using only two questions. Not surprisingly, the interannotator agreement among the domain experts was very high: 80%–90%, but it was more varied among the other judges; we thus selected only a subset of judges whose annotations matched most closely those of the experts, with the interannotator agreement in the 62%–75% range. Two different agreement scores were created, a 'hard score' that counted the number of exact judgments among all four possible relevancy selections and a 'lenient score', that assumed a passage could be only 'relevant' or 'nonrelevant', in other words the selections 1, 2 and 3 in the instructions were treated as equivalent. The five best performing annotators (hard score: 62%–75% and lenient score: 75%–82%) were retained to annotate the full corpus.

4.2.3 Evaluation

In evaluating the performance of various components we are primarily interested in changes in precision of the set of passages passed down from one component to the next. Recall is less interesting because it can only be measured relative to the maximum number of passages returned by each component. For example, the passage clustering component returns on average 34 passages or fewer which is only a small fraction of about 200 passages it receives as input. Three distinct precision scores were calculated for each passage set of passages S:

https://doi.org/10.1017/S1351324908004890 Published online by Cambridge University Press

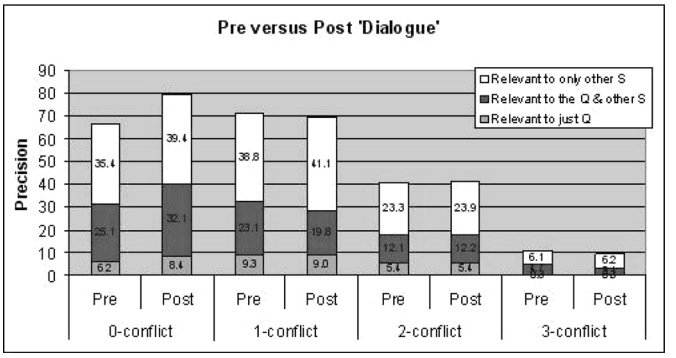


Fig. 20. Precision before and after 'Dialogue'.

Precision(S, n) Number passages judged into category n total number of passages in S

where n is either 1, 2 or 3 as per instructions in Figure 19.

4.3 Effectiveness of clarification dialogue

The purpose of interactive dialogue in HITIQA is to offer to the user the information which falls outside of the initial answer due to various mismatches with the question. As we discussed before, such mismatches may be caused by lexical or syntactic variation or more likely by semantic differences; nonetheless the information may still be relevant to the question or to the task scenario. We were interested in evaluating the impact of the clarification dialogue on the precision of the core answer set (the 0-conflict zone). To estimate the lower bound of the dialogue effect we ran a simulated 'dialogue' experiment where the system was configured to automatically reply 'Yes' to all of its own clarification questions. While this experiment is quite informative, it is important to keep in mind that in real operation the analyst guides the dialogue process, in that each response influences the next question that the system generates. Therefore, the evaluation results in this section can only be seen as a lower baseline. We compared the precision of the set of passages in the 0-conflict zone (c.f., Figure 5) before the dialogue to its precision after the dialogue. Figure 20 summarizes this.

The experiment showed a substantial increase in precision of the 0-conflict zone following the 'dialogue' across all three types of relevance. This confirmed that the system was indeed finding and offering more relevant passages (accepted offers) than nonrelevant passages (rejected offers). Specifically, we note a 26.2% increase in the precision with respect to the current question and a 10.1% increase in the precision with respect to other parts of the scenario, for an overall increase of 21.8% in answer precision following the dialogue. These results attest to the effectiveness of the dialogue moves in HITIQA: selecting a random subset of passages from the 1-conflict zone and moving them into the 0-conflict zone is expected to increase precision by no more than 3%.

https://doi.org/10.1017/S1351324908004890 Published online by Cambridge University Press

5 Related research

Research in IQA is still in the early stages and few systems exist today that can be directly compared to HITIQA. However, there has been some notable work in the field of IQA (Bertomeau *et al*., 2006; Kelly *et al*., 2006). The currently dominating model of interactivity for QA has been adapted from earlier work in interactive information retrieval (e.g., (Koenemman and Belkin, 1996)). In this model, the users provide relevant feedback and revise their queries while the system remains largely passive. While relevant feedback has been shown to improve retrieval accuracy in laboratory tests, the process is very inefficient from the user perspective; worse still, since the users are not very good at revising their queries the process tends to produce highly biased and low recall results. Nonetheless, recent advances in QA research sparked interest in systems that can sustain longer and more meaningful interaction with the user, for example, by answering a series of interlocking questions, or by allowing the user to submit follow-up questions. TREC QA evaluations, the traditionally strong engine of progress, has introduced tracks that allow limited forms of interaction, although these remain hobbled by the overall test conditions in TREC which require batch submission of results. Research into AQA has been facilitated since 2001 by government-funded programs such as AQUAINT, part of which has focused on IQA. Besides HITIQA which has been developed largely under AQUAINT, few other systems offer a comparable degree of interactivity for the user. The FERRET system (Harabagiu *et al*., 2005) offers a limited form of interaction through a prebuilt structure of question/answer pairs. User questions submitted to FERRET are compared against the list of known questions (Ask-Jeeves style) which are then offered to the user, along with their answers, for possible selection. HITIQA's approach to interactivity is strongly influenced by the research in natural language dialogue. Specifically, the use of event frames to drive clarification dialogue can be seen as an extension of the slot-filler approach in spoken language dialogue systems such as AMITIES (Hardy, Strzalkowski and Wu, 2003) and CU-Communicator (Ward and Pellom, 1999). Other research that influenced the HITIQA design includes intention modeling work for indirect speech acts, for example, TRAINS/TRIPS systems (Allen *et al*., 1995) and, perhaps somewhat more directly, the JUPITER system (Glass *et al*., 1995). Hybrid information seeking systems are also closely related to HITIQA, for example, the Yellow Pages Assistant (De Roeck *et al*., 2000), which allows users to ask natural language questions over a classified advertising directory and then uses dialogue to guide them to appropriate entries.

6 Conclusion

In this paper we described an IQA system, HITIQA, which helps users to find answers to complex analytical problems. Such problems often necessitate the user to submit not one but an entire series of questions, both simple and complex, and then negotiate the final content and form of the answer. HITIQA advances research in human–computer dialogue by enabling topical, mixed initiative interaction over unstructured data. HITIQA uses the process of text framing to bring a level of

https://doi.org/10.1017/S1351324908004890 Published online by Cambridge University Press

semantic representation to open-domain data and facilitates a meaningful dialogue with the user. The system exploits mismatches between semantic interpretations of the user questions and the available data items to generate efficient and effective dialogue moves. This highly targeted and topical dialogue helps the system to rapidly shape the answer space satisfying the user's information need. We demonstrated through user-centered evaluations that this approach is quite effective and that it produces measurable benefits in search efficiency and answer quality over currently dominant information retrieval approaches. We recognize, however, that the current version of HITIQA is just a first step in building a truly IQA system. Planned future work includes modeling discourse context to allow for more expressive forms of dialogue as well as investigation of multimodal interactivity including extended forms of data exploration. More research is also needed into user modeling that would enable the system over time to adapt to the user interaction style and expectations. Developing models of human-system interactions, as opposed to human–human interactions, will enable more effective dialogue design. This work is already starting: HITIQA system logs created over the last 4 years of analytical exercises provide an important though still relatively small data set to start with.

7 Acknowledgment

This paper is based on work supported in part by the Disruptive Technology Office under contracts 2002-H790400-000, NBCHC040028 and 2006-1057203-000 (AQUAINT Program) to SUNY, Albany.

References

Allen, J., Schubert, L., Ferguson, G., Heeman, P., Hwang, C., Kato, T., Light, M., Martin, N., Miller. B., Posesio, M., and Traum, D. 1995. The TRAINS project: a case study in building a conversational planning agent. *Journal of Experimental and Theoretical Artificial Intelligence* **7**: 7–48.

Bertomeau, N., Uszkoreit, H., Frank, A., Krieger, H., and Jorg, B. 2006. Contextual phenonmena and thematic relations in database qa dialogues: results from a wizard-of-oz experiment. In *Worshop on Interactive Question Answering, HLT-NAACL 2006*, New York.

Callan, J., Croft, W., and Harding, S. 1992. The inquery retrieval system. In *Proceedings of the 3rd International Conference on Database and Expert Systems*, Valencia, Spain.

Clough, P., Gaizauskas, R., Piao, S., and Wilks, Y. 2002. Meter: measuring text reuse. In *Proceedings of the 40th Anniversary Meeting for the Association for Computational Linguistics (ACL-02)*, University of Pennsylvania, Philadelphia, USA, pp. 152–59.

De Roeck, A., Kruschwitz, U., Scott, P., Steel, S., Turner, R., and Webb, N. 2000. The YPA – An assistant for classified directory enquiry. In *Intelligent Systems and Soft Computing: Prospects, Tools and Applications. Lecture Notes in Artificial Intelligence (LNAI)*, vol. 1804. Springer Verlag.

Glass, J., Flammia, G., Goodine, D., Phillips, M., Polifroni, J., Sakai, S., Seneff, S., and Zue, V. 1995. Multilingual spoken-language understanding in the MIT VOYAGER System. *Speech Communication* **17**: 1–18.

Harabagiu, S., Hickl, A., Lehmann, J., and Moldovan, D. 2005. Experiments with interactive question-answering. In *Proceedings of the 43rd AnnualMeeting of the ACL*, Ann Arbor, MI, pp. 205–14.

https://doi.org/10.1017/S1351324908004890 Published online by Cambridge University Press

Harabagiu, S., Moldovan, D., Pasca, M., Surdeanu, M., Mihalcea, R., Girju, R., Rus, V., Lacatusu, F., Morarescu, P., and Bunescu, R. 2002. Answering complex, list and context questions with lccs question answering server. In *Proceedings of Text Retrieval Conference (TREC-10)*, Gaithersburg, MD.

Hardy, H., Kanchakouskaya, V., and Strzalkowski, T. 2006. Automatic event classification using surface text features. In *Proceedings of the AAAI Workshop on Event Extraction and Synthesis*, Boston, MA.

Hardy, H., Shimizu, N., Strzalkowski, T., Wise, B., and Zhang, X. 2002. Cross-document summarization by concept classification. In *Proceedings of SIGIR-2002*, Tampere.

Hardy, H., Strzalkowski, T., and Wu, M. 2003. Dialogue management for an automated multilingual call center. In *Proceedings of the HLT-NAACL 2003 Workshop: Research Directions in Dialogue Processing*, Edmonton, Alberta, Canada, pp. 10–12.

Hovy, E., Gerber, L., Hermjakob, U., Junk, M., and Lin, C. 2000. Question answering in webclopedia. In *Proceedings of Text Retrieval Conference (TREC-9)*, Gaithersburg, MD.

Kelly, D., Kantor, P., Morse, E., Scholtz, J., and Sun, Y. 2006. User-centered evaluation of interactive question answering systems. In *Worshop on Interactive Question Answering, HLT-NAACL 2006*, New York.

Koenemman, J., and Belkin, N. 1996. A case for interaction: a study of interactive information retrieval behavior and effectiveness. In *CHI'96 Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, New York: ACM, pp. 205–12.

Miller, D., Schwartz, R., Weischedel, R., and Stone, R. 1999. Named entity extraction from broadcast news. In *Proceedings of DARPA Broadcast News Workshop*, Herndon, VA.

Morse, E., and Sholtz, J. 2004. An investigation of evaluation metrics for analytic question answering. In *Proceedings of AQUAINT Phase 2 6-month PI Meeting*, Tampa.

Prager, J., Chu-Carroll, J., Czuba, K., and Ittyercheriah, A. 2003. In question-answering two heads are better than one. In *Proceedings of HLT-NAACL 2003*, Edmonton, AB, Canada, pp. 24–31.

Searle, J. R. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge: Cambridge University Press.

Small, S., Shimizu, N., Strzalkowski, T., and Liu, T. 2003. Hitiqa: a data driven approach to interactive question answering, a preliminary report. In *Proceedings of AAAI New Directions in Question Answering Spring Symposium 2003*, San Jose, California, pp. 94–104.

Small, S., Strzalkowski, T., Liu, T., Shimizu, N., and Yamrom, B. 2004. A data driven approach to interactive QA. In M. Maybury (ed.), *New Directions in Question Answering*, pp. 129–40. MIT press.

Strzalkowski, T., and Wang, J. 1996. A self-learning universal concept spotter. In *Proceedings of the 16th conference on Computational Lingusitics*, Association for Computational Linguistics, Copenhagen, Denmark, vol. 2, pp. 931–36.

Ward, W., and Pellom, B. 1999. The CU Communicator system. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, Keystone, CO, pp. 341–44.

Yangarber, R., Grishman, R., Tapanainen, P., and Huttunen, S. 2000. Unsupervised discover of scenario-level patterns for information extraction. In *Proceedings of the 6th conference on Applied Natural Language Processing*, Seattle, WA.

https://doi.org/10.1017/S1351324908004890 Published online by Cambridge University Press