HITIQA: A Question Answering Analytical Tool

Tomek Strzalkowski, Sharon Small, Hilda Hardy, Boris Yamrom, Ting Liu, Paul Kantor[†], *K.B. Ng*[†], *Nina Wacholder*[†]

The State University of New York at Albany
1400 Washington Avenue
Albany, NY 12222
tomek@csc.albany.edu

Keywords: Question Answering, Information Extraction, Visualization, Open Source Intelligence (OSINT)

Abstract

HITIQA (High Quality Interactive Question Answering) is currently being developed to assist analysts in finding answers to complex intelligence problems, efficiently and thoroughly. The system uses event-based, 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 first version of the system has undergone a series of preliminary evaluations with the analysts from the US Naval Reserve, producing valuable usage and performance data. These evaluations suggest that HITIQA creates a measurable cognitive augmentation effect for the analyst. The second more advanced version of the system is currently being implemented.

1. How HITIQA works

HITIQA is an advanced question answering system that helps the analyst to produce higher quality reports for complex intelligence problems in less time and with lower cognitive load. The primary function of HITIQA is to supply composite answers to complex, exploratory questions such as "*What is the state of development of long range missiles in North Korea? Can they reach the U.S.?*" Submitting such questions to a conventional internet search produces tens of thousands of hits that include many related and unrelated documents of varying length and veracity. More questions may need to be asked along the way to fill information gaps or to explore related topics such as production capabilities or missile technology proliferation, etc. Each time a question is posed, substantial effort must be applied to retain relevant facts, note contradictions, ignore repetitions and discard unrelated and unreliable sources. The problem is not usually lack of information; more often it is too much information—fragmented, indirectly related, sometimes misleading—and the lack of skilled assistance to help the analyst wade through it.

HITIQA does not return long lists of documents, as keyword search does; instead, it retains only the most relevant passages and assembles them into a coherent composite answer. HITIQA selects its answer more carefully, too: keyword match may be a reasonable indication of potential relevance, but until we know *why* these words are found together, answer precision is likely to be low. In order to improve accuracy, HITIQA performs named entity extraction from candidate text passages and then uses several prototypical event templates (called *frames*) to arrive at the most likely interpretation. The text passage is thus rendered into an event frame which assigns event roles to the entities found in text. The list of available roles varies from one event type to another, but in general they include positions such as AGENT, LOCATION, DESTINATION, TIME, etc. Often, multiple interpretations are possible, and it is assumed that one interpretation is better than another if it explains the roles of more entities. In other words, a lower perplexity explanation is preferred, and better still if it matches the user's question. Multiple frames can be assigned to each text passage.

The framing process outlined above identifies a set of highly relevant passages which are then organized into a coherent answer by weeding out duplicate information and imposing a rudimentary rhetorical structure over it. Additionally, frames are converted into natural language headlines serving as high-level summaries and eye grabbers. However, this is only a part of what HITIQA does; the second important function of the system is to track other related, possibly fragmentary information located at the fringes of the current answer space and offer it for the analyst's consideration. These offers must be made in a way that is at the same time exhaustive (i.e., they bring the analyst's attention to key issues that may be missing from the current answer)

[†] Rutgers University, School of Library and Information Science, New Brunswick, NJ 08904

and non-invasive (i.e., the offers can be accepted or rejected with minimal cognitive overhead). In HITIQA, this is accomplished through an interactive dialogue with the analyst.

The need for dialogue arises from mismatches between the question as posed to the system and the content of the potential answer space. If the question scope and interpretation could be altered somewhat leading to a significant change of the answer space (expansion, reduction, or shift), it may be reasonable to ask the analyst if such a change is permissible. One way to bring up such issues is in a form of definite suggestions:

ANALYST: *Does N. Korea make missiles that can reach U.S.?*
HITIQA: *Would you be interested in information on Taepo-dong-2 missile?*

Another effective method of communicating mismatches is through a visual interface that shows in an abstract and simplified way the entire answer space and its immediate environs. A color-coded display, showing information considered relevant (blue) located next to what is currently assumed not relevant (red) may prompt the analyst to examine these regions (Figure 1).

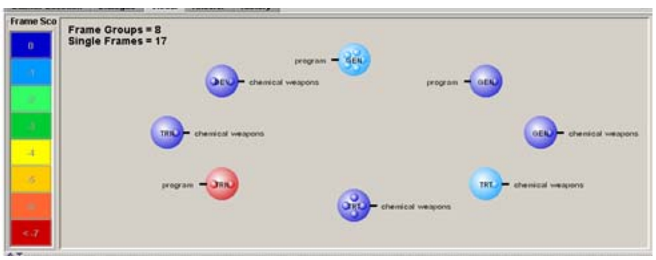

Figure 1: HITIQA visual display of answer space

In the following sections we outline some of the key technical aspects of the current system and discuss some future research directions. More details of HITIQA architecture can be found in (Small et al., 2004a, b).

2. Finding information: cast a wide net

HITIQA works with unstructured text data; therefore a document retrieval step is required to pre-fetch information that may relate to the user question. We use relatively simple information retrieval methods (term matching, etc.) to obtain perhaps 200 potentially relevant documents from a database. This gives us an initial information space to work with in order to determine the scope and complexity of the answer. In a traditional document retrieval system, the task of extracting the answer from the retrieved documents falls squarely upon the user, a significant cognitive burden, which the QA technology aims to reduce. The current version of HITIQA uses the INQUERY IR system (Callan et al., 1992), although we have also used SMART (Buckley, 1985) and can substitute other IR systems (such as Google).

An interesting question is how accurate the initial document selection needs to be. Although high recall is obviously desirable, a relatively low precision of the initial document set will have only limited impact on the final outcome of the QA process. This is because the low initial precision can be offset by more careful analysis of the now much smaller amount of text. In HITIQA, we first break up the retrieved documents into short passages and then cluster the passages using a combination of hierarchical clustering and *n*-bin classification (Hardy et al., 2002). Each cluster represents a topic theme within the retrieved set. Since clusters are built out of short text passages, HITIQA usually associates one event frame with each passage, thus providing a (simplified) interpretation of the text[1]. This process leads to a secondary grouping of passages around same-event frames, in effect combining complementary information supplied by different passages. At the cluster level, the frames serve as representatives of entire groups of passages.

3. Understanding information: event frames

The framing process imposes a partial structure on the text passages. This allows HITIQA to systematically compare different passages against one another and against the user question. Framing does not attempt to capture the entire meaning of the passage; it needs to be just complete enough to communicate with the user about key content differences between his or her 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. If these topics or aspects align closely with the user's question, (i.e., matching many of the salient attributes) HITIQA will attempt to make the user aware of them and let him/her decide if they should be included in the answer.

HITIQA starts text framing by building a *general frame* over the passages from each cluster as well as over each of the top *N* (currently *N=10*) scored passages 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, ORGANIZATION, DATE, etc. Attributes are extracted from text passages by BBN's Identifinder (Miller et al., 1999), which tags 24 types of named entities. The event/relation itself is a value of the TOPIC attribute, which is captured from the central verb or noun phrase of the passage and is not otherwise restricted in any way, e.g., *accident*, *pollution*, *trade*, etc. In the general frame, attributes have no assigned roles; instead, they are loosely grouped around the TOPIC. An example of general frame based on a text passage is shown in Figure 2. It should be noted that this frame does not represent any specific single event; instead, it may be seen as a bundle of possibly multiple events anchored at the common TOPIC attribute.

[1] Multiple frames may be assigned to each text passage, but they still would be ranked by the degree of fit. The current version of HITIQA uses only the top ranked frame.

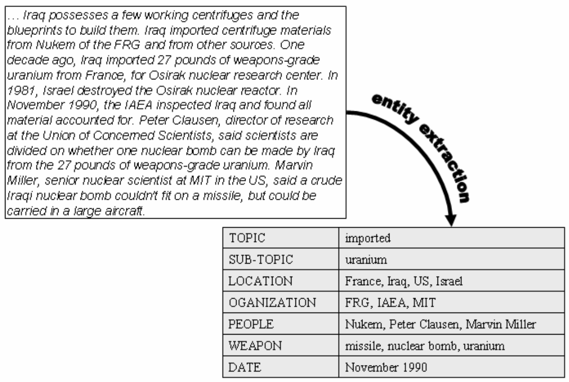

Figure 2: A general frame derived from text passage

We have developed a small collection of specialized *typed frames.* Typed frames represent a class of prototypical events which are automatically instantiated to specific events that cover a number of analytical domains, including weapons, technology, terrorism, finance and economy, ethnic strife, and possibly more. Typed frames are obtained from generic frames by matching the value of TOPIC attribute with an appropriate event type and assigning roles to some attributes. Examples of typed frames include: the *TRANSFER* frame with roles including SOURCE, DESTINATION and OBJECT; the *DEVELOP* frame with AGENT and OBJECT roles; the *ATTACK* frame, with roles including AGENT, TARGET, INSTRUMENT, etc. We have currently identified 10 frames that appear to sufficiently cover a number of domains of interest to intelligence applications[2]. These frames are listed in Figure 3.

We may note that the names given to the frames in Figure 3 are somewhat arbitrary in that they reflect the domains of the data we use in our experiments, and can be broadened or narrowed without losing any functionality of the framing process. Domain adaptation is desirable for obtaining more focused dialogue (see below), but it is not necessary for HITIQA to work.

In order to develop a collection of typed frames like those listed in Figure 3, we run "concordances" over pairs of named entities (persons, places and organizations) found in a representative text corpus. These concordances show us the contexts around NE pairs and sometimes indicate events that connect them. Here is an example: "*Yesterday, 7 December 1941 the United States of America was suddenly and deliberately attacked by naval and air forces of the Empire of Japan*." We examine the most common verbs (and occasionally nouns) occurring in the contexts; these are potential event triggers. We can classify the trigger words into event types; for instance, *arrested, detained, filed lawsuit* indicate LEGAL events; and *attack, bombed, airstrike* point to ATTACK events.

Event frame	Example events	Key roles (selected)
AGREE	treaty, agreement, sign	PARTIES, TYPE, INSTR, LOC
ASSIST(*)	help, support, assist, aid	TARGET, AGENT, TYPE, INSTR
ATTACK	attack, invade, destroy	TARGET, AGENT, TYPE, INSTR
CAPABLE(*)	possess, control, capable	AGENT, INSTR, QUANT
DEVELOP	construct, develop	OBJECT, AGENT, QUANT
FINANCIAL	fund, finance, pay	TARGET, SOURCE, QUANT
LEGAL(*)	inspect, embargo, detain	TARGET, AGENT, CHARGE
POLITICAL(*)	elect, appoint, resign	TARGET, AGENT, POSITION
THREAT(*)	threaten, fear, menace	TARGET, AGENT, TYPE, INSTR
TRANSFER	acquire, export, smuggle	OBJECT, AGENT, SOURCE, DEST

Figure 3: Prototypical event frames in HITIQA (*under development)

Once we have determined the initial set of prototypical event frames, we annotate a small number of documents from the corpus, using a simple software annotation tool to indicate events and the values for their attributes or roles. During the manual annotation step, training examples are gathered and the roles for each event are refined. From these training examples we write patterns or rules so that we can automatically extract event frames. The event patterns are based on trigger words, key prepositions and sometimes other terms, as well as information gleaned from Identifinder tags, parser output, and position in the sentence.

The resulting rule-based event-extraction shows promise in terms of precision (see *Accuracy*, below), but it cannot extract events containing new trigger words or new patterns. We are implementing a bootstrapping process (based on e.g., Strzalkowski & Wang, 1996; Yangarber, 2003) to take the initial rules as seeds and use unsupervised machine-learning to obtain additional rules, improving recall for each frame type.

[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. While classifications such as (Levin, 1993) or FrameNet (Fillmore, 2001) are certainly relevant here, we are aiming at a discourse-based *semantic* system that does not limit frame acquisition to sentence boundaries or *syntactic* patterns.

4. Understanding the question: frame matching

The framing process is also applied to the user's question, resulting in one or more *Goal frames*, which are subsequently compared to the frames obtained from retrieved text passages. A Goal frame can be a general frame or any of the typed frames. Figure 4 shows an example of matching text and question frames. 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, ranging from direct incompatibilities (e.g., different locations) to role mismatch (e.g., *from* Korea vs. *to* Korea) to missing or underspecified attribute values.

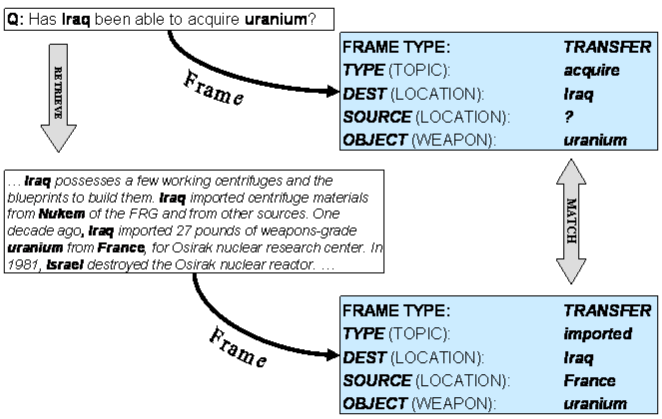

Figure 4: Question and a matching text frames.

As indicated above, text framing serves several important functions in HITIQA. These include:

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 supporting meaningful dialogue with the user.

In order to provide adequate support for all these functions, we continue to refine various aspects of the frame representation and the framing process itself. A few aspects of HITIQA framing process that are still under development are briefly discussed below.

Roles:
Figure 3 shows only the most important roles for each type of frame; however, additional, optional roles are defined for some frames to capture lesser aspects of an event whenever possible, e.g., QUANTITY, AMOUNT, CONVEYANCE, etc., as well as more universal roles such as TIME and LOCATION. Nonetheless, all potential roles cannot be anticipated for every event; therefore many frame attributes will remain unassigned. Unassigned attributes are not discarded and they remain accessible to the frame matching program from the underlying general frames.

Modalities:
Frames represent unqualified "core" events; however, in actual text these events may only be referred to as future, possible, hypothetical, or may be even denied as ever occurring. Here is a complex but not atypical example: "*The Security Council has discussed a plan to pressure Syria into ending the flow of weapons to Iraq*." The core event is a TRANSFER of *weapons* from *Syria* to *Iraq*, and it can be understood as actual or past, depending upon other context and time frame. In other words, certain modal operators may need to be applied to the frames in order to properly capture the information content from the passage. In the current version of HITIQA modalities are not explicitly represented in frames and all event references are treated equally. It is left up to the analyst to determine if any of these references are valuable, however, we are considering options during answer generation where some of the modal operators (e.g., negation) would be converted into appropriate rhetorical links ("…, *however* …").

Frame fusion:
The framing process is applied to short text passages in order to maximize salience of a single event while minimizing interference from references to different events that may be described in the balance of the document. This helps to increase precision of event extraction, but may produce incomplete frames when attributes are scattered over larger distances. HITIQA compensates by attempting to fuse frames representing the same event but built over distinct passages. Two frames may represent the same event if they are of the same type and their attributes are compatible. In order to accommodate irresolvable variants, and also to overcome some common extraction mistakes, HITIQA allows merged frames to have lists of alternative values for some attributes. These multi-valued attributes may be viewed as disjunctions of options; the resulting apparent ambiguity is resolved when the frame is compared against a question frame. At present, explicit frame fusion is limited to passages from the same document, although we allow event-level grouping on the visual display.

Accuracy:[3]
We have conducted preliminary assessment of the accuracy of the framing process, using the *ATTACK* frame as a test. After running the frame spotter over 100 unseen documents in our corpus, we evaluated the resulting 122 frames as follows: 66 were scored as "good" (the event and all attributes correctly assigned); 42 were scored as "medium" (AGENT or TARGET wrong or imprecise, but still a valid *ATTACK* event); and 14 were scored as "bad" (not an *ATTACK* at all, or most arguments wrong, etc.). Considering only the "good" frames,

[3] It should be noted that HITIQA processes are structured into a cycle of stepwise refinements thus avoiding the error propagation issues plaguing cascaded systems.

gives an (informal) precision rate of 54%. The combined rate of 88.5% is for both "good" and "medium" frames, which are usually sufficient to support effective clarification dialogue. On the other hand, the recall was only 49%, as tested on the same documents. We expect to improve the recall through unsupervised bootstrapping.

5. Interacting with the analyst

Framed information allows HITIQA to automatically judge text passages as fully or partially relevant and to conduct a meaningful dialogue with the user about their content. The purpose of the dialogue is to help the user navigate the answer space and to negotiate more precisely what information he or she is seeking. The dialogue includes clarifications (*did you mean…*), offers (*would you like…*), counter-offers (*… however …)*, and may also include suggestions, alerts, reminders, and other dialogue acts that are still under development. The dialogue is primarily content-oriented; that is, the system will ask questions related to the analyst's question topic, not about data elements. The reason for this is as follows: a well-posed, relevant question is not only helpful, but it requires relatively little additional cognitive effort on the part of the analyst to respond to (cf. the exchange about Taepodong-2 missile). The relevance information that is not captured in dialogue can be glimpsed from other actions performed by the analyst, including browsing of the visual panel and copying material for the final report.

HITIQA supports dialogue to help the analyst negotiate the scope of the information task that needs to be performed. Data frames which have no conflicts with the analyst's question form the initial answer space that is immediately available to the user for inspection. Depending upon the presence of other frames outside this set, the system may initiate a dialogue to clarify the scope of the question and/or to offer additional items to the analyst. HITIQA begins by asking the user about frames that have but one conflict with the original question; perhaps the location or time or event can't be matched.

ANALYST**:** *What is the history of the nuclear arms program between Russia and Iraq?*

HITIQA**:** *Do you want to see material on cooperation on development of weapons between Iraq and Russia?*

As these conflicts are cleared away, other frames with multiple conflicts may come into view; for example, if cooperation is relevant, then perhaps production facilities are also of interest, etc. The dialogue proceeds until all "near-miss" frames are processed.

HITIQA verbal dialogue is augmented by an interactive visual display which allows the analyst to manipulate the answer space based on visual rather than verbal cues. Both methods of interaction are tightly integrated and completely interchangeable. In the most recent version of the system verbal and visual dialogues proceed simultaneously from the same computer screen, where they are presented in adjacent panels (Figures 5 and 6). This allows the analyst to easily switch between dialogue modes as well as to see immediately the effects of each dialogue move on the answer space.

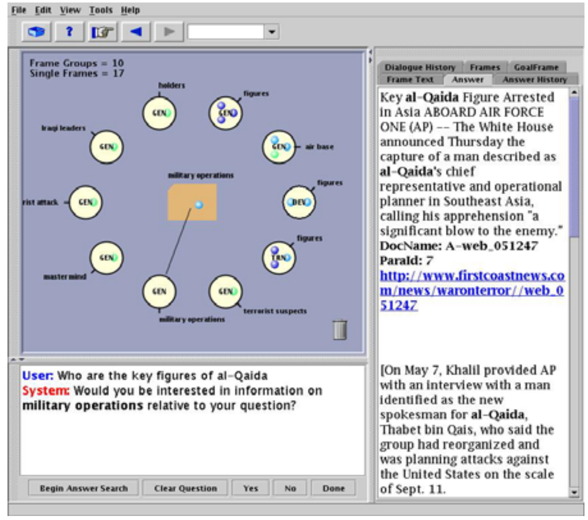

Figure 5: Visual Interface, cluster view

As the dialogue progresses, HITIQA answer space takes a more definite shape, while the information is organized to facilitate final report preparation. For additional clarity, significant structural changes to the visual display, such as formation of new folders or moving icons between folders, are animated.

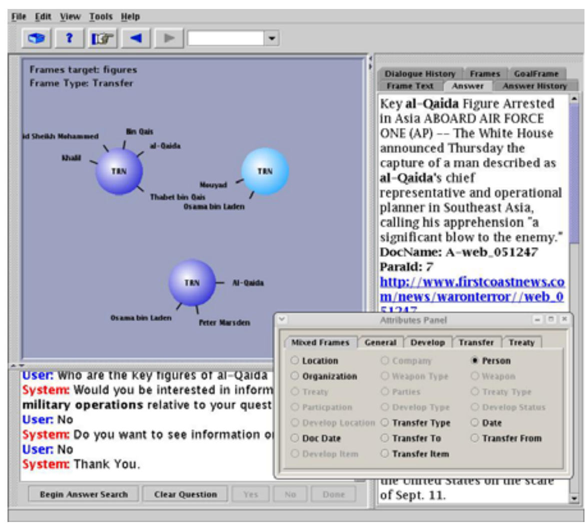

Figure 6: Visual Interface: frame view

6. Generating a coherent answer

In answer generation, text passages extracted from multiple sources need to be organized 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: without a deeper understanding of the material and a detailed domain model, our options are somewhat limited. We want to construct an acceptable rhetorical structure over a collection of largely disconnected information items (passages and their frames) using relatively weak semantic relations that may be computed among them: partial temporal order, approximate spatial collocation, possibly shared entities, as well as potential contradictions and other modal modifiers.

The relationship between two text passages may be calculated by comparing their frames and linking frame attributes. Some combinations of links may indicate elaboration, others would indicate contradiction, yet others might suggest a temporal sequence. This work is in early stages. Currently we organize the output around the key relevant event frames that emerged from the dialogue with the analyst. We also generate headline-like one-line summaries.

7. Evaluating HITIQA in realistic drills

We have evaluated HITIQA in a series of workshops with professional analysts in order to obtain an in-depth assessment of the system usability and performance. For the participating analysts, the primary activity at these workshops involved preparation of reports in response to "scenarios"—complex questions that often encompass multiple sub-questions, aspects and hypotheses. For example, in one scenario, analysts were asked to locate information about the al Qaida terrorist group: its membership, sources of funding and activities. We prepared a database of over 1GByte of text documents; it includes articles from the Center for Non-proliferation Studies (CNS) data collected for the AQUAINT program and similar data retrieved from the web. Over several on-site and on-line drills, a group of fifteen analysts generated multiple reports for 12 realistic intelligence problems, spending anywhere between 1 and 3 hours per report. Each session involved multiple questions posed to the system, as well as clarification dialogue, visual browsing and report construction. The evaluation instruments included questionnaires assessing analysts' opinions about various aspects of the system, as well as a cross-evaluation process where analysts scored each other's reports for completeness and organization.

While still preliminary, the evaluation suggests two important advantages of HITIQA over other approaches as well as over a document retrieval baseline using Google:

1) the HITIQA interactive approach is significantly more efficient because it requires the analyst to ask fewer questions (nearly 60% fewer than using Google) and consequently spend less time to obtain a report of equal or better content; and

2) HITIQA is more effective because it produces more usable information per user question, evidenced by analysts saving more material for their reports and doing so more often. It makes the collection process twice as effective as searching with Google.

These findings suggest that HITIQA has a potential to provide cognitive augmentation for the analyst thus allowing him or her to produce better reports using fewer resources than is the current practice. Further, larger scale evaluations are required to verify these findings.[4]

Acknowledgements

This paper is based on work supported in part by the Advanced Research and Development Activity (ARDA)'s Advanced Question Answering for Intelligence (AQUAINT) Program. Special thanks to Mike Blair, John Rogers, J. Steindl and USNR analysts for participation in the workshops. Additional thanks to Dr. Emile Morse for running the ARDA Metrics Challenge. Thanks also go to Google for extending their license, to Ralph Weischedel for the use of IdentiFinder, to Chuck Messenger and Peter LaMonica for assistance in development of the analytical scenarios, and to Bruce Croft for the use of INQUERY system.

References

Buckley, C. 1985. *Implementation of the Smart information retrieval system.* TR85-686, Computer Sci., Cornell Univ.

Callan, J. P., W. B. Croft, and S. M. Harding. 1992. *The INQUERY Retrieval System*, in Proceedings of the 3rd International Conference on Database and Expert Systems.

Fillmore, C., C. F. Baker. 2001. *Frame semantics for text understanding*. WordNet Workshop at NAACL.

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

Levin, B. 1993. ***English Verb Class and Alternations: A Preliminary Investigation***. Chicago: Univ. of Chicago Press.

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

Morse, E. 2004. *An Investigation of Evaluation Metrics for Analytic Question Answering*. Proceedings of AQUAINT Phase 2 6-month PI Meeting, Tampa.

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

Small, S., T. Strzalkowski, T. Liu, S. Ryan, R. Salkin, N. Shimizu, P. Kantor, D. Kelly, N. Wacholder. 2004b. *HITIQA: Towards Analytical Question Answering*. Proceedings of COLING 2004 Conference, Geneva, Switzerland.

Strzalkowski, T. and J. Wang. 1996. *A Self-Learning Universal Concept Spotter*. Proceedings of COLING 1996.

Wacholder, et al. (forthcoming) *A Model for Realistic Evaluation of an End-to-end Question Answering System.* Unpublished manuscript.

Yangarber, R. 2003. *Counter-Training in Discovery of Semantic Patterns*. Proceedings of ACL-2003, Sapporo, Japan.

[4] For details of the evaluation methodology and results, see (Morse, 2004; Wacholder et al., forthcoming). The AQUAINT Program-wide evaluations are currently being developed.