Experiments with Interactive Question Answering in Complex Scenarios

Andrew Hickl, John Lehmann, John Williams and Sanda Harabagiu
Language Computer Corporation
1701 N. Collins Suite 2000
Richardson TX 75080
andy@languagecomputer.com

Abstract

This paper addresses the pragmatic challenges that state-of-the-art question/answering systems face in trying to decompose complex information-seeking scenarios. We propose that question decomposition can be approached in one of two ways: either by approximating the domain-specific knowledge for a particular set of domains, or by identifying the decomposition strategies employed by human users. We also present preliminary results from experiments that confirm the viability of each of these approaches within an interactive Q/A context.

1 Introduction

Over the past five years, much research has focused on the different challenges question-answering systems face when answering questions in isolation as opposed to questions presented as part of a contextualized interaction with a user.

The domain of interactive question answering is typically concerned with two tasks: the decomposition of complex questions into questions that can be processed by current Question/Answering (Q/A) systems and the dynamic representation of dialogue between a user and a Q/A system. In this paper, we argue that the effective decomposition of questions is much more valuable to the performance and the development of interactive Q/A systems than any potential advances in dialogue processing alone. We believe that this is due to the fact that dialog systems do not operate under the requirement of finding information from vast text collections through a sequence of questions and answers, which is the operating principle of interactive Q/A systems.

We also believe that the quality of interactive Q/A systems largely depends on the precision with which they can find answers to questions. Unless a Q/A system can reliably process questions in isolation with a high degree of accuracy, it is very unlikely that that system will be able to answer questions in an interactive context with any degree of precision. Although dialogue processing may improve the quality of answers generated by a question-answering system (by resolving ambiguities in a set of possible answers, for example), recent advances in dialogue processing have not contributed to overall improvements in answer retrieval or extraction in any meaningful way.

However, we are not suggesting that dialogue processing should play no role in the development of Q/A technology. A system that decomposes questions well must continue to be responsive to users in order to gather information about the user's level of expertise and to anticipate the user's information needs throughout the course of the interaction.

In this paper, we present the framework for a new way of decomposing complex questions that benefits from the latest technology in both the answering of simple questions and the processing of interactions with users of several levels of expertise.

We present an analysis of two years' worth of experiments that we have conducted in interactive Q/A that takes advantage of a state-of-the-art question answering system. Our experiments show that tackling complex questions requires several ways of modeling the domain of a complex topic as well as meaningful ways of finding topic-relevant information in text collections.

We have also found that the interactions required by the resolution of complex questions or scenarios engender many more forms of questions than the types of questions (i.e. factoid, definition, list) evaluated in the past TRECs. We believe this is due to the fact that complex scenarios may consist of many different possible relations between complex questions that need to be topologically modeled as complex answer structures. Furthermore, our experiments have identified three means of decomposing complex scenarios into simple questions that can be processed reasonably well by current question-answering systems.

In addition, we have found that ontological requirements and real-time constraints are major factors in developing interactive question-answering systems that realistically satisfy the needs of expert users. The NLP techniques required to process questions interactively consist of (1) several syntactic and semantic processes that lead to the identification of the structure of the expected answer; (2) context modeling that includes coreference resolution; and (3) the coherent decomposition

of a complex scenario into a series of simple questions that are likely to be asked by users.

The remainder of the paper is structured as follows. Section 2 presents the state-of-the-art in question answering technology. In Section 3, we detail our method of processing domain-dependent complex questions and the two classes of users on which we tested our methods. In Section 4, we describe our approach to question decompositions whereas in section 5, we present an analysis of the automatic decompositions generated by our system and results of our scenario-processing experiments involving expert information analysts. Section 6 presents our conclusions and directions for future work.

2 Domain-Dependent Complex Questions

The decomposition of complex information-seeking scenarios represents a trio of pragmatic challenges for Q/A systems.

First, effective question decomposition depends on the acknowledgment of the intentions that underlie a user's interaction with a Q/A system. Individuals participate in information-seeking dialogues (whether with other humans or with interactive Q/A systems) in order to learn new things – that is, to gather information that they do not currently possess. A user's behavior in a dialogue focuses on that set of speech acts which allow them to maximize the new information they obtain from the conversation while (at the same time) minimizing the amount of redundant or previously-established information that they encounter. We expect these same principles to govern the decomposition of complex scenarios as well: the decompositions generated by a user will focus on returning the domain-specific information that the user currently does not possess. Expert users (who are assumed to be familiar with a domain) will use interactive Q/A systems to (1) evaluate their existing knowledge with regards to changes in the context or (2) seek new information about known entities or events within the domain. In contrast, non-expert users (who remain unfamiliar with much of the ontological structure of a complex domain) will have very broad and potentially poorly-defined informational goals; in these cases, interactive Q/A systems will need to return information which will facilitate the novice users' exploration of the domain.

Second, we suggest that question decomposition will depend on the development of semantic ontologies that are articulated enough to address the domain-specific questions characteristic of most complex information-seeking scenarios. Current Q/A technologies are unable to process (or decompose) complex questions without access to a large amount of domain-specific knowledge. Modeling domain-specific knowledge for complex domains, however, is an arduous task: complex domains necessarily consist of sets of structured concepts linked by classes of semantic relations. Although this kind of domain modeling is traditionally considered to be tangential to research in NLP, we believe that interactive Q/A systems must have access to not only the ontological structure of answers and complex semantic information, but also modes of probabilistic reasoning that can be used to induce categories of meanings between domain concepts.

Finally, since complex questions represent such diverse informational goals, it should not be assumed that even the decompositions produced by expert users will be sufficiently simple enough to be processed by current Q/A systems. We propose that careful study needs to be conducted to identify the new types of context-dependent questions that are generated as part of interactive Q/A.

The rest of this section is ordered as follows. Section 2.1 describes three new types of questions found in the sets of decompositions generated by human users. Section 2.2 details a simple solution that expands the coverage of interactive Q/A systems for specific topic domains. Finally, Section 2.3 distinguishes between two idealized types of users of interactive Q/A systems: experts and novices.

2.1 Scenarios and Questions

Since complex questions represent such diverse informational goals, it should not be assumed that even the decompositions produced by expert users will be sufficiently simple enough to be processed by current Q/A systems.

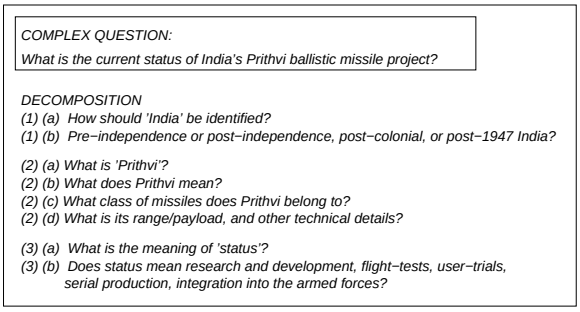


Figure 1: Scenario decomposition for the topic focused on Prithvi missiles.

The decompositions presented in Figure 1 introduce a number of novel challenges for Q/A systems. Three are discussed below:

Clarification Questions. Questions like *What is the meaning of "status"?* represent a new challenge for current Q/A systems. Unlike TREC-style definition questions, this class of questions (which we refer to as clarification questions) seek to identify the most domain-specific characterization available for the concept, entity, or term in focus. Although informationally "simple",

answers to these questions depend on implicit domain-specific knowledge that can only be supplied by an interactive Q/A system. In order to answer a question like *What is the meaning of "status"?*, a system must be able (1) to identify the differences between the domain-specific and the domain-general characterization of the focal item, (2) to recognize which domain-specific sense the user is seeking, and finally (3) to return information that will help the user understand all of the domain-dependent semantic entailments of the term.

Alternative Set Questions. Questions produced as part of a scenario decomposition often ask a system to distinguish between several different possible alternatives for the characterization of an entity. Faced with a question like *How should "India" be identified? Pre-independence or post-independence? Post-colonial or post-1947 India?*, the Q/A system must not only identify the named entity *India* but must also return enough contextual information to be able to determine which of the named set of entities should be considered most relevant to the current contextual scenario.

Although the set of alternatives can be overtly stipulated by the user, an interactive Q/A system should ideally possess the domain-specific knowledge and the inferential capacity to be able to generate these kinds of alternative sets automatically. Although a set of alternatives may be extractable from a highly-specified semantic ontology for a question like *What is the meaning of "status"?*, it is unlikely that such an ontology can be used to derive the different instantiations of *India* listed in *How should "India" be identified?*. In this latter case, the system would have to (1) decide whether some sort of differentiation was necessary between the available instantiations, (2) identify which of the set of instantiations were the most relevant alternatives, and finally, (3) determine which instantiation should be used to retrieve the answer.

Contextual-Dependent Ellipsis. Questions that involve syntactic ellipsis must be answered in context. With a question like *What does "Prithvi" mean?*, the system must recognize that semantic meaning is evaluated with regards to a language (here, any of those spoken on the Indian sub-continent). The system must also be able to (1) identify examples of implicit syntactic ellipsis, (2) determine the semantic type of the syntactically-elided information, and finally, (3) supply the contextually-relevant members of that semantic class needed to return the answer.

Based on the initial observations above, we conclude that a careful analysis of the questions generated by scenario decompositions needs to be conducted to identify new types of questions that cannot be processed by current Q/A systems. By expanding the coverage of Q/A systems for these kinds of "informationally simple" questions, we expect future Q/A systems to be better positioned to process questions with more complex informational goals. A careful examination of the question decompositions generated by expert users can help us better understand what kinds of domain-specific knowledge should be made available to an interactive Q/A system.

2.2 A Practical Solution

The goal of question decomposition is to translate complex questions into simpler questions that have identifiable answer types. Effective question decomposition does not guarantee answers, however: current Q/A systems are only able to provide answers for approximately 55% of simple (i.e. factoid, definition, and list) questions.

For most state-of-the-art Q/A systems, correct answers are returned iff the system identifies the correct answer type from the syntax and semantics of the question itself. Although current answer-type hierarchies can be fairly broad in their coverage of concepts, they do not provide an exhaustive treatment of all of the types of information that users can request for any particular domain. In LCC's current Q/A system (Harabagiu, Moldovan, et al., 2003), no answer type could be detected for questions like *What business was the source of John D. Rockefeller's fortune?* (TREC-1909) or *What 1857 U.S. Supreme Court decision denied that blacks were citizens?* (TREC-2259). The failure of our system to return answer types for these questions was attributed to identifiable gaps in our semantic ontology of answer types. By revising our answer type hierarchy to include classes of *businesses* or *Supreme Court decisions*, we could presumably enable our system to identify a viable answer type for each of these questions, and thereby improve our chances of returning a correct answer to the user.

However, the challenge of expanding an answer type hierarchy becomes considerably more difficult when we start considering the very specific semantic ontologies that would need to be added to a hierarchy to account for specific domains such as *the development of Prithvi missiles in India*, *opium production in Afghanistan*, or *AIDS in Africa*. Without expert input into ontology creation for each of these domains, NLP researchers can have only a limited idea of the conceptual knowledge that necessarily needs to be added to the answer type hierarchy in order to improve Q/A for each of these domains.

Given these considerations, we were able to improve our coverage of domain-specific factoid questions by incorporating a database of 342 question-answer pairs (related to a series of specific domains) into our Q/A system. We used this database, known as the QUestion-Answer Base or QUAB, to measure the conceptual similarity of new questions to question-answer pairs already listed in QUAB. In the absence of a highly-articulated answer-type hierarchy, we assumed that questions that exhibited a high degree of similarity necessarily encoded

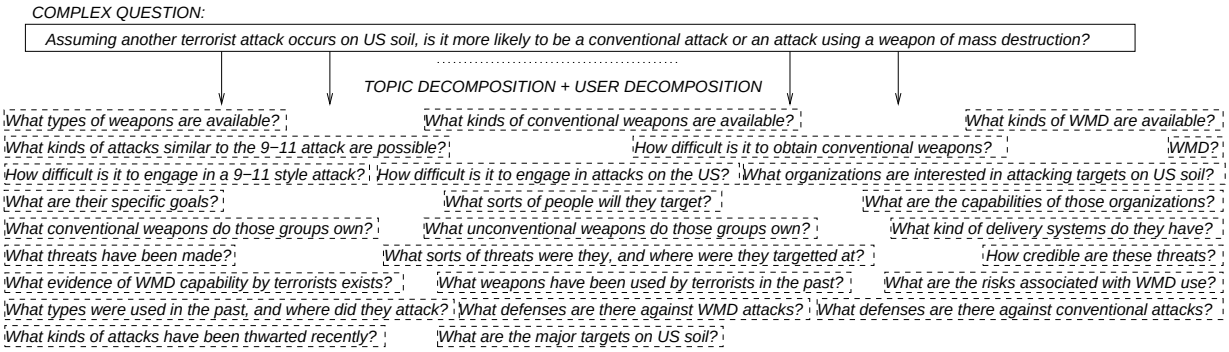


Figure 2: Example of questions entered in QUAB.

similar informational goals and could be answered with similar types of information. When a new question was judged to be conceptually similar to a question in QUAB, the QUAB question's answer was returned to users as a potential answer. Questions that were not similar to any existing question in QUAB were submitted automatically to our Q/A system without providing any additional feedback to users. This type of methodology allowed us to develop a series of Just-in-Time-Information-Search-Agents (JITISA) which exploited different measures of conceptual and lexical similarity in order to identify answers to domain-specific questions.

2.3 Users of Complex Q/A

The performance of interactive Q/A systems can be improved by identifying what strategies different users employ to reach their informational goals. We define an informational goal as the propositional knowledge that a user is trying to obtain by participating in a dialogue with a Q/A system. We suggest that the representation of an informational goal depends crucially on the specific knowledge that individual users bring to the interaction with the Q/A system. Users that possess little or no knowledge of a particular domain will necessarily seek a different level of information than users who are intimately familiar with the domain. Based on these assumptions, we propose that interactive Q/A systems be sensitive to two kinds of users: (1) expert users, who may be expected to interact with the system based on a working knowledge of the semantic ontology underlying a domain, and (2) novice users, who are expected to have no foreknowledge of the ontological categories specific to the domain. By examining the differences in information-seeking techniques employed by expert users (such as intelligence analysts) and novice users (such as NLP researchers), we can better identify user intentions and work towards anticipating the information needs of any user.

3 Question Decompositions

We suggest that question decomposition can be approached in two ways. The first approach generates a set of questions related to the complex question by maximizing the extraction of information related to the domain. In this way, the Q/A user is provided with full coverage of the information associated with the concepts expressed in the complex question. This methodology seeks to approximate domain-specific knowledge. The idea is that by caching information associated with the domain, the domain coverage is maximized and the likelihood that the retrieved answers meet the users' information needs is enhanced. The questions that extract relevant domain information are clustered in related sub-topics and generate a bottom-up decomposition of the complex question.

The second approach generates a top-down decomposition by monitoring user strategies towards decomposition. The purpose is to derive general relations between topic-specific questions and the subquestions that they entail. Such relations are discovered by combining domain specific knowledge with general coherence relations. The domain knowledge selects decompositions viable in the context of a domain scenario, whereas the coherence relations connect questions of different levels of complexity.

In recognition of these diverse goals, we hypothesize that research in question decomposition should follow two parallel tracks: topic-centric and user-centric. These two proposed strategies have different strengths. The user-centric strategy mimics the user's intentions when resolving an information-seeking task but may miss relevant information since not all the right questions may be covered. In contrast, the topic-centric strategy generates good recall, but it relies on similarity functions that are hard to encode. Section 3.1 presents topic-centric work.

The rest of the section is organized in the following way. Section 3.2 presents user-centric work. Section 3.3 speculates about the contribution of each form of decomposition to interactions with the Q/A system. We argue that there are optimal ways for combining advances in

each approach to provide a unified treatment of question decomposition.

3.1 Topic-Centric Method

Answers to a complex question are retrieved from a set of topic-relevant documents. In our experiments, we have used two sets of such documents. In the first pilot evaluations, we have created our own corpus of documents relevant to the topics proposed. The corpus combined documents from Lexus-Nexus with documents we have gathered from the Internet. The relevance of the documents was provided by the presence of certain concepts we deemed characteristic for each domain. In the second pilot, we used the documents provided by the Center for Non-Proliferation Studies (CNS) that were considered relevant based on the concepts that could be derived from the complex questions and their decompositions.

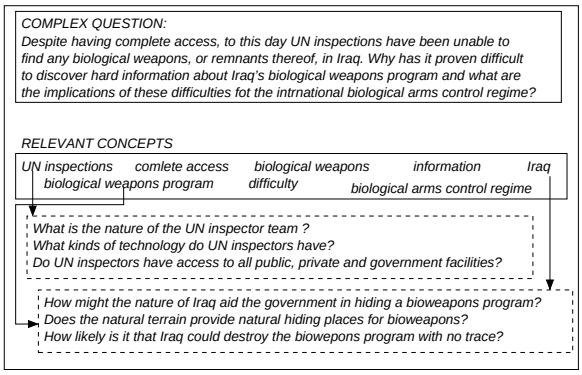


Figure 3: Topic-centric decomposition.

As illustrated in Figure 3, topic-relevant concepts guide the generation of questions than are easier to process. Because the simpler questions contain concepts used also in the complex question, they are related to it. Figure 3 shows several questions created by question patterns for which (1) there was at least one text snippet in the collection that matched a trigger word; and (2) contained at least one of the relevant concepts. When the relevant concept was *"UN inspectors"* the trigger words were: *"team"*, *"technology"* and *"facilities"*, which are typical of inspections.

Questions generated by topic-centric decomposition could be related to multiple relevant concepts. Figure 3 illustrates also a set of questions that related both to the *"Iraq"* concept and to the *"bioweapons program"* concept. The latter concept is a synonym of the concept *"biological weapons program"*.

3.2 User-Centric Method

Different users might decompose a complex question differently. By producing an analysis of the complex question and of the questions produced by each user we can explore several different paths of searching for the relevant information of a complex question. Additionally, the different paths indicate the kind of topic knowledge each user has available. It also indicates the level of expertise of each user.

In analyzing the complex question, we focus on (1) the focus of the question; (2) the context of the question and (3) the implied results. Since complex questions may consist of multiple sentences and interrogatives, we produce such tree-dimensional structures for each sentence/interrogative of the complex question, as illustrated in Figure 4. Figure 4 also lists a set of questions that may be derived from the structure associated with each question constituent. It may be noticed that these questions have multiple natures. Some can be cast as definition questions, e.g. *What is a biological weapon ?*. Other questions are based on knowledge of each sub-topic. For example, *Did Iraq violate any international law?*, implies that international laws govern the international biological arms control regime.

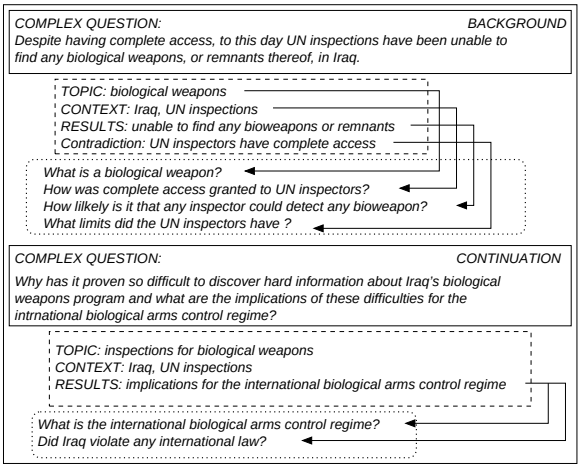


Figure 4: User-centric decomposition.

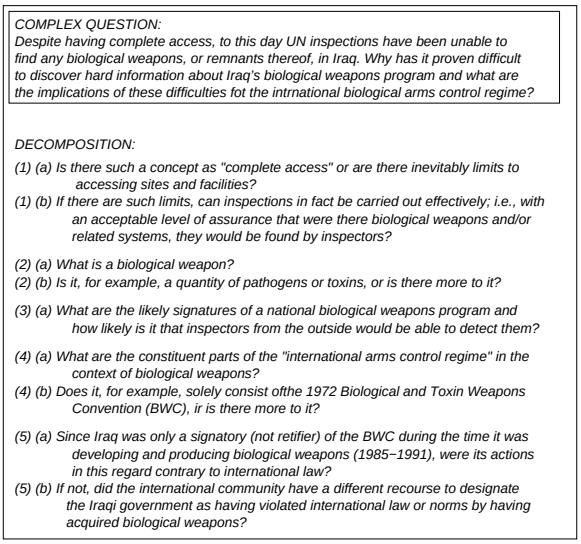
User-centric decompositions are based on the idea that each user generates a sequence of questions that represents a path from the complex question to a series of questions that are connected through coherence relations of the type ELABORATION or CAUSE-EFFECT. Since definition and list questions are also used, the set of coherence relations needs to be adapted for the task of interactive Q/A.

3.3 Experiments with Expert Users

This section presents a brief case study comparing decompositions produced by three users of different skill levels.

We collected three decompositions of the following complex scenario: *Despite having complete access, to this day UN inspections have been unable to find any biological weapons, or remnants thereof, in Iraq. Why*

has it proven difficult to discover hard information about Iraq's biological weapons program and what are the implications of these difficulties for the international biological arms control regime?. This scenario asks users to elaborate about a state of affairs: namely, the failure of UN weapons inspectors to find evidence of a biological weapons program in Iraq. In addition, users are asked to return information about (1) the potential causes of this state as well as (2) the expected effects of the continued duration of the state on the "biological arms control regime".



COMPLEX QUESTION:
Despite having complete access, to this day UN inspections have been unable to find any biological weapons, or remnants thereof, in Iraq. Why has it proven difficult to discover hard information about Iraq's biological weapons program and what are the implications of these difficulties fot the intrnational biological arms control regime?

DECOMPOSITION:

(1) (a) Is there such a concept as "complete access" or are there inevitably limits to accessing sites and facilities?
(1) (b) If there are such limits, can inspections in fact be carried out effectively; i.e., with an acceptable level of assurance that were there biological weapons and/or related systems, they would be found by inspectors?

(2) (a) What is a biological weapon?
(2) (b) Is it, for example, a quantity of pathogens or toxins, or is there more to it?

(3) (a) What are the likely signatures of a national biological weapons program and how likely is it that inspectors from the outside would be able to detect them?

(4) (a) What are the constituent parts of the "international arms control regime" in the context of biological weapons?
(4) (b) Does it, for example, solely consist ofthe 1972 Biological and Toxin Weapons Convention (BWC), ir is there more to it?

(5) (a) Since Iraq was only a signatory (not retifier) of the BWC during the time it was developing and producing biological weapons (1985−1991), were its actions in this regard contrary to international law?
(5) (b) If not, did the international community have a different recourse to designate the Iraqi government as having violated international law or norms by having acquired biological weapons?

Figure 5: Scenario decomposition provided by NIST.

We predict that the domain-specific knowledge that users possess will directly influence how they perform question decomposition. If a user has overt knowledge that causality can exist between the state described in the scenario and another set of states or events, then we should expect decomposition to proceed in an evidentiary mode. Since the user has evidence that two states may be causally linked in another domain, subquestions are asked in order to gather information that describes how this causal relationship is instantiated in the current domain of interest. However, if a user only has a belief or an expectation (and no overt knowledge) that a causal link can be established between two states, we expect decomposition to be more general and epistemic in nature. Since the user has only a belief that causality exists between two states, they must first confirm that this expectation is viable before they can turn to gathering information which supports their claim. Since they have (by definition) a better conception of the semantic ontology for the domain, expert users will ask a higher percentage of factoid and evidence-seeking questions than novice users. Likewise, we expect that that the decompositions of novice users will be characterized by more general questions that seek to evaluate which ontological relationships are available in a particular domain.

Although these predictions may prove difficult to evaluate in many real texts, they do appear to borne out in the following decompositions.

NIST Decomposition. Figure 5 presents the scenario decomposition generated by NIST as part of the ARDA AQUAINT project. This decomposition focuses on four major topics: (1) the nature of *"complete access"* in terms of the UN inspections in Iraq, (2) the definition of the term *"biological weapon"*, (3) the potential sources of evidence which would point to the existence of a biological weapons program, and finally, (4) the clarification of international laws concerning biological weapons. Although these four topics are clearly central to the domain, it is notable that this decomposition does not include any questions that address the issue of finding biological weapons in Iraq.

What was the scope of Iraq's biological weapons program? In the past? Immediately prior to US invasion?

What quantities of biological weapons has Iraq used in past wars? In other periods? Within Iraq? Against Iran?

Does Iraq have the infrastructure necessary for destroying biological weapons safely? For creating biological weapons?

Does Iraq have the capacity to store and/or transport biological weapons? By land? By air? By sea? How has that capacity chanced since 1991?

Are there personnel within the Iraqi government responsible for destroying biological weapons? Are these people civilians or military personnel?

Are there Iraqi personnel (scientist, clerks, military) that we can identify who have been traditionally associated with the Iraqi bioweapons program? What are their names? In what capacity did they participate in the bioweapons program?

Is there evidence from Iraqi military medical records for possible signs of biological warfare sickness or contamination?

Is there evidence from Iraqi civilian hospital records of doctors who have treated possible biological weapon sicknesses? Are there individuals who have witnessed cases of biological weapon sicknesses?

Has Iraqi military trained personnel in the use of biological weapons? At any time in the past 12 years?

Does Iraq have any military units tasked with using biological weapons? Are those units still active? When were they disbanded?

Which countries have been formally allied with Iraq? Since 1991?

Is there evidence that countries may have stored bioweapons for Iraq? Is there evidence that other countries have engaged in similar kinds of deals with Iraq in the past?

Figure 6: Scenario decomposition created by an Intelligence Analyst.

Analyst Decomposition. Figure 6 presents a scenario decomposition generated by an intelligence analyst as part of a pilot study conducted by LCC. (For more details on this pilot study, see Section 4.2) In contrast with the NIST decomposition, this decomposition focuses on establishing factual evidence for several different hypotheses concerning the failure of inspectors to find bioweapons in Iraq.

LCC Decomposition. Finally, Figure 7 presents a scenario decomposition created by a novice LCC researcher who had no specific training in analysis techniques and no specific background in the domain. Although the LCC researcher produced considerably more questions than the two experts, these questions focus mostly on discovering the classes of hypotheses and conceptual relations that are found in the domain. Questions like *What do*

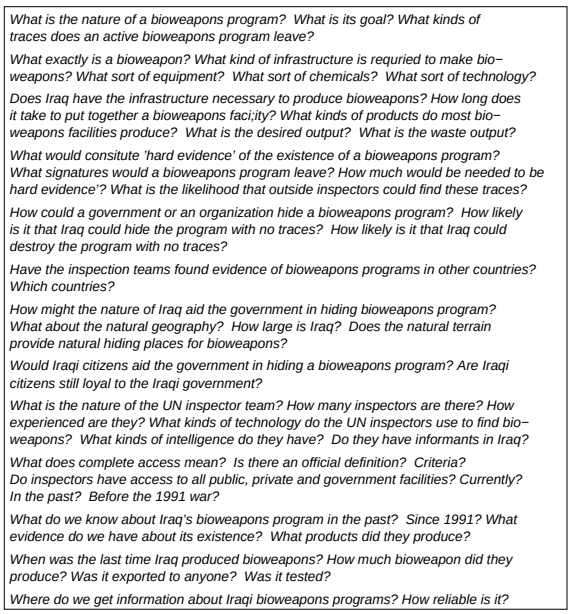


Figure 7: Scenario decomposition created by LCC researchers.

we know about Iraq's bioweapons program in the past? suggest the researcher's informational goals were defined at a much more general level than either of the two experts. In addition, the LCC researcher's questions provided a broader coverage of the topics within the domain; although this demonstrates that the researcher did have some familiarity with issues central to the domain, it also signifies that he most likely did not have access to knowledge that would have allowed him to evaluate which concepts were most central to the informational focus defined in the scenario.

Comparison. Although all three of the decompositions above cover many of the same topics (e.g. the nature of bioweapons and bioweapons programs, the Iraqi infrastructure for supporting bioweapons programs, etc.), they differ in the the level of specificity of their questions. While both expert and novice decompositions do include questions that establish domain-specific definitions for particular keywords or phrases (e.g. *What constitutes "complete access" for inspectors?*), questions in the expert decompositions appear to focus more on gathering evidence for particular hypotheses, while questions in novice decompositions focus more on establishing which kinds of hypotheses are viable in the given domain. This observation is supported by comparing analogous examples from the decompositions above. While the expert analyst was able to ask a rather specific question about Iraq's past use of biological weapons that demonstrated an in-depth knowledge of the geopolitical entities in the domain (i.e. *What quantities of biological weapons has Iraq used in past wars? In other periods? Within Iraq? Against Iran?*), the novice LCC user was only able to question whether or not the event had occurred at all (i.e. *Has Iraq ever used biological weapons?*).

The results from this case study appears to confirm that interactive Q/A systems need to be sensitive to the intentions users bring to their interaction with the system. The Gricean maxim of quantity is supported in these cases: users participate in dialogues in order to obtain information that they do not already possess. Given this assumption, we predict that users' decompositions of complex scenarios focus on questions that allow them to maximize the new information obtained from the system while minimizing the amount of old (or previously-known) information that the system returns.

4 Lessons Learned

This section presents preliminary results from two experiments examining scenario decomposition in an interactive Q/A context. In Section 4.1 we discuss results which confirm that a database of question/answer pairs can be used to approximate the types of specific semantic knowledge necessary to process (and answer) domain-dependent complex questions. In Section 4.2, we outline five strategies for question decomposition employed by experts that could be use to improve the automatic processing of complex information-seeking scenarios.

4.1 Results of the Interactions based on QUAB

Recent research has shown that the precision of Q/A systems is dependent on the semantic coverage of their answer type hierarchies. For most current interactive Q/A systems, correct answers can only be returned iff a system is able to identify the answer type that most closely approximates a question's informational goal. In most cases, if the Q/A system cannot identify an appropriate answer type – or if the answer type does not exist in the semantic ontology – no answer can be returned.

However, as we pointed out in Section 3.2, ontology creation may not be possible (or effective) for every semantic domain that users ask about. In order to answer domain-dependent questions, interactive Q/A systems need to incorporate ways of approximating the domain-specific information that their answer type hierarchies may lack. In this section, we present results that show that a database of question/answer pairs (known in our system as QUAB) can be used to improve interactive Q/A for domains that may not have completely specified answer-type hierarchies.

The utility of QUAB was evaluated in a series of two "Wizard-of-Oz"-style dialogue pilot experiments conducted as part of the ARDA AQUAINT project. In each pilot, professional intelligence analysts interacted with LCC developers (and the LCC interactive Q/A system) through an Internet chat-style interface. LCC used

Domain	Answers	User Qs	QUAB Qs	%Q from QUAB	%A from QUAB
India - Prithvi	8	7	4	36.4%	50.0%
Iraq - Bioweapons	18	9	9	50.0%	50.0%
Russia - Nuclear Thefts	18	14	6	30.0%	33.3%
China - Arms Control	20	7	2	33.3%	10.0%
Iraq - Nuclear Program	25	5	2	28.6%	8.0%
Russia - North Korea	27	8	4	33.3%	14.8%
Total	116	50	27	35.1%	23.3%

Table 2: Results from the second pilot

Domain	Full	Partial	Not at all	Total
Opium in Afghanistan	3	2	1	6
AIDS in Africa	3	2	3	8
Black Sea Pollution	5		2	7
FARC/Colombia	2	3		5
Indonesian Economy	8	3	2	13
Cell Phones in Ivory Coast	3	6	2	11
Japanese Joint Ventures	2		4	6
Microsoft and Viruses	3		1	4
Elizardo Sanchez	5		1	6
Robotic Surgery	5		4	9
Total	39	16	20	75

Table 1: Pilot 1 System Effectiveness

the preparation time prior to the first experiment to seed QUAB with 140 domain-specific question/answer pairs. 182 additional question/answer pairs (based on 6 of the 12 Spring 2003 AQUAINT domains) were added to QUAB prior to the second pilot experiment as well. QUAB was primarily used to return answers for domain-specific questions that our interactive Q/A system could not process. Each user question was evaluated in terms of keyword and conceptual similarity with all of the question-answer pairs contained in QUAB; if no answer could be provided to the user's question, the most similar QUAB answers were returned. In results compiled from both pilots, QUAB provided exactly the correct answer 52% (39/75) of the time, and either exactly or partially the correct answer 73% (55/75) of the time. Table 1 presents these results organized by question domain.

QUAB was also used to provide an interactive component to our Q/A system as well. Each question submitted by a user was compared to the database of question/answer pairs already contained in QUAB. If the user's question was deemed to be conceptually similar to an entry in QUAB, the user was informed that the system could return information "related" to the user's question. If a user requested this related information, the QUAB entry was presented to the user in the form of a question/answer pair. For example, when a user asked the question *What facilities has Iraq used to produce biological weapons?*, QUAB offered the answer to *How does the US know about the existence of biological weapons plants in Iraq?* as related information that could potentially facilitate the user's research.

In the second dialogue pilot, question/answer pairs from QUAB were presented to the users a total of 27 times in 6 different dialogues. On average, contributions from QUAB made up approximately 35% of the questions and about 23% of the answers considered by users throughout the course of the dialogue. Table 2 presents results from the 6 domains considered in the second dialogue pilot. (It is important to note that in this pilot, users could ask the Q/A system to return more answers for any question; this explains why there are often more answers than questions in each of the dialogues.)

The success of a relatively small QUAB suggest that this type of database construction may be an efficient way to augment interactive Q/A and answer-type detection for very domain-specific questions.

4.2 Results Produced by Experts

In this section, we present work from a pilot study that examined how intelligence analysts performed question decompositions for domains within their areas of expertise. After presenting a case study comparing their individual decompositions of a domain, we identify five different decomposition strategies employed by the analysts.

In order to obtain more high-quality data for analysis, we invited three intelligence analysts from the Naval Reserve to LCC for three days of study. We were interested in (1) determining whether users of a specific level of expertise performed decompositions of complex questions in a similar fashion and (2) identifying possible patterns in their research styles that could be used in the development of automatic question decomposition strategies.

Analysts participated in three tasks. For the first task, analysts were asked to create short outlines (dubbed "skeleton reports") of answers to complex questions using only publicly-available web-based resources. For the second and third tasks, analysts were asked to provide decompositions of complex questions. In the second task, analysts were asked to list the questions that they anticipated they would answer prior to starting their research; in the third task, analysts decomposed questions without

any other special instructions. For purposes of comparison, a LCC developer participated also participated in the decomposition tasks.

Despite their similar levels of training and expertise, the differences in the analysts' individual styles were striking. When asked to decompose the "Iraqi bioweapons" scenario presented in Figures 5 , 6, and 7, analysts produced questions that demonstrated broad differences in their interpretation of the scenario itself.

Analyst 1. Analyst 1's decomposition focused on four specific aspects of Iraq's bioweapons program: (1) the history of the bioweapons program, (2) the alleged products of the program, (3) the personnel involved with creation of the program, and (4) the potential locations for program. Although this analyst's 10 questions were well-balanced, his decomposition centered on the nature of the program itself, and provided no potential for an explanation of why the weapons were difficult to find.

Analyst 2. Analyst 2's decomposition questioned many of the implicit assumptions set forth in the topic question itself. Instead of providing subquestions that could have led to potential answers for this topic question, his decomposition suggested that he had rejected the propositions that the topic question was based on. In his 18 subquestions, he questioned the presuppositions of the scenario itself, generating subquestions such as *Is it the case that the UN inspectors are really being denied "complete access"?* and *Can we be sure that Iraq had bioweapons at any point in the past?*.

Analyst 3. Analyst 3's decomposition focused on the reasons he believed were responsible for the difficulty in finding Iraq's bioweapons. In his 17 questions, he discussed three real hypotheses: (1) the weapons do not exist, (2) the weapons are well hidden in Iraq, (3) the weapons have been moved outside of Iraq. After identifying these three hypotheses, Analyst 3 asked a variety of subquestions that gathered evidence for (or against) each of these three possibilities.

Although the analysts produced roughly similar numbers of subquestions, there was little overlap in the content that they covered. This suggests that even expert analysts differ markedly in their expectations of what constitutes the informational goal of a complex information-seeking scenario.

In examining the decompositions produced by the analysts, we discovered that the analysts employed a number of distinct decomposition strategies. Four of them are discussed below:

Syntactic Decomposition. Analysts split questions that featured syntactic coordination into subquestions that contained each of the individual conjuncts. For example, a question like *How do we know that the UN has not found any biological or chemical weapons?* was decomposed as *How do we know that the UN has not found any biological weapons?* and *How do we know that the UN has not found any chemical weapons?* In the data we collected, we only found examples of adjective phrase and noun phrase conjunction; we expect analysts to decompose examples of sentence or verb phrase coordination in a similar fashion.

Entity Motivations. Analysts asked questions about an entity's political or economic motives if the topic question involved a predicate that implied that the entity had volitional control over its actions. For example, a topic question like *Why does China dispute Taiwan's independence?* was decomposed into questions like *What are China's economic motives for disputing Taiwan's independence?* or *What are China's political motives for disputing Taiwan's independence?*.

State Discovery. When faced with a question about the existence of a property or past state, analysts generated decompositions that contrasted the previous status and its current status. For example, questions like *What type of nuclear assistance did China give to the Middle East between 1980 and 1990?* were routinely decomposed into questions of the form *How does the nuclear assistance given by China to the Middle East from 1980 to 1990 compare to nuclear assistance it provides to the Middle East today?*.

These subquestions took three forms. Analysts wanted to know: (1) how the situation in the past differs from the present situation, (2) what caused the change from the past to the present, and (3) what impact the past events have on the present. We hypothesize that the above subquestions are part of a larger class of subquestions known as state discovery questions. Unlike events, which represent a particular moment in time (or set of moments), states are inherently durative and therefore are subject to a wider variety of changes in scope, level, or status over time. We believe that questions that make reference to a property or a state of being necessarily make an implicit comparison between periods in time: i.e. an identified point (such as the years between 1980 - 1990) and some other reference point (either the current moment or some other salient period).

Meronymy. We found that analysts were sensitive to the internal structure of many of the named entities referenced in topic questions. In general, analysts generated questions about the subparts of an entity if and only if information about those subparts proved informative in answering the topic question as a whole. Given an example like *Where are Prithvi missiles manufactured?*, analysts generated decompositions like *Where are the guidance systems for Prithvi missiles manufactured?* or *Where are the warheads for Prithvi missiles manufactured?*. Further research is needed to determine when an entity's component parts should be considered as part of the informational goal of a question.

When faced with topic questions that present a controversial or empirically-unverified proposition, we found that analysts generated decompositions that questioned the relative truth of the proposition before generating other decompositions. Faced with a complex question like *How much nuclear material was stolen from the Soviet military after the fall of the Soviet Union?* (which necessarily entails that nuclear material was, in fact, stolen from the Soviet military), analysts generated decompositions like *How much nuclear material is known to have been stolen from the Soviet military?* or *How much nuclear material is suspected to have been stolen from the Soviet military?* Analysts did not generate these types of questions, however, when the question under discussion reflected a publicly accepted proposition or an empirically-verifiable state.

5 Conclusions

In this paper we presented a new framework for the decomposition of complex information-seeking scenarios. We proposed that question decomposition for interactive Q/A could be approached in one of two ways: either (1) by approximating the domain-specific knowledge for a particular set of domains or (2) by identifying the decomposition strategies employed by human users. In addition, we presented preliminary experimental results that confirmed the viability of both of these approaches. We discussed two years' worth of experiments that investigated how users of varying levels of expertise decomposed complex scenarios, as well as work that described how a database of question/answer pairs could be used to improve the coverage of Q/A systems for domain-specific questions.

References

C. Fellbaum. 1998. WordNet: An Electronic Lexical Database and Some of its Applications. *MIT Press*.

S. Harabagiu, D. Moldovan, C. Clark, M. Bowden, J. Williams, J. Besley. 2003. Answer Mining by Combining Extraction Techniques with Abductive Reasoning. In *Proceedings of the Twelfth Text Retrieval Conference (TREC 2003)*

D. Moldovan, C. Clark, S. Harabagiu, S. Maiorano. 2003. COGEX: A Logic Prover for Question Answering. In *Proceedings of the Human Language Technology and North American Chapter of the Association for Computational Linguistics Conference (HLT-NAACL 2003)*

E. Voorhees. 2003. Overview of the TREC 2003 Question Answering Track. In *Proceedings of the Text REtrieval Conference (TREC-2003)*, pages 14-27, 2003.