Experiments with Interactive Question-Answering

Sanda Harabagiu, Andrew Hickl, John Lehmann, and Dan Moldovan
Language Computer Corporation
Richardson, Texas USA
sanda@languagecomputer.com

Abstract

This paper describes a novel framework for interactive question-answering (Q/A) based on *predictive questioning*. Generated off-line from topic representations of complex scenarios, predictive questions represent requests for information that capture the most salient (and diverse) aspects of a topic. We present experimental results from large user studies (featuring a fully-implemented interactive Q/A system named FERRET) that demonstrates that surprising performance is achieved by integrating predictive questions into the context of a Q/A dialogue.

1 Introduction

In this paper, we propose a new architecture for *interactive* question-answering based on *predictive questioning*. We present experimental results from a currently-implemented interactive Q/A system, named FERRET, that demonstrates that surprising performance is achieved by integrating sources of topic information into the context of a Q/A dialogue.

In interactive Q/A, professional users engage in extended dialogues with automatic Q/A systems in order to obtain information relevant to a complex scenario. Unlike Q/A in isolation, where the performance of a system is evaluated in terms of how well answers returned by a system meet the specific information requirements of a single question, the performance of interactive Q/A systems have traditionally been evaluated by analyzing aspects of the dialogue as a whole. Q/A dialogues have been evaluated in terms of (1) efficiency, defined as the number of questions that the user must pose to find particular information, (2) effectiveness, defined by the relevance of the answers returned, (3) user satisfaction.

In order to maximize performance in these three areas, interactive Q/A systems need a predictive dialogue architecture that enables them to propose related questions about the relevant information that could be returned to a user, given a domain of interest. We argue that interactive Q/A systems depend on three factors: (1) the effective representation of the topic of a dialogue, (2) the dynamic recognition of the structure of the dialogue, and (3) the ability to return relevant answers to a particular question.

In this paper, we describe results from experiments we conducted with our own interactive Q/A system, FERRET, under the auspices of the ARDA AQUAINT[1] program, involving 8 different dialogue scenarios and more than 30 users. The results presented here illustrate the role of predictive questioning in enhancing the performance of Q/A interactions.

In the remainder of this paper, we describe a new architecture for interactive Q/A. Section 2 presents the functionality of several of FERRET's modules and describes the NLP techniques it relies upon. In Section 3, we present one of the dialogue scenarios and the topic representations we have employed. Section 4 highlights the management of the interaction between the user and FERRET, while Section 5 presents the results of evaluating our proposed

[1] AQUAINT is an acronym for *Advanced QUestion Answering for INTelligence.*

Proceedings of the 43rd Annual Meeting of the ACL, pages 205–214,
Ann Arbor, June 2005.

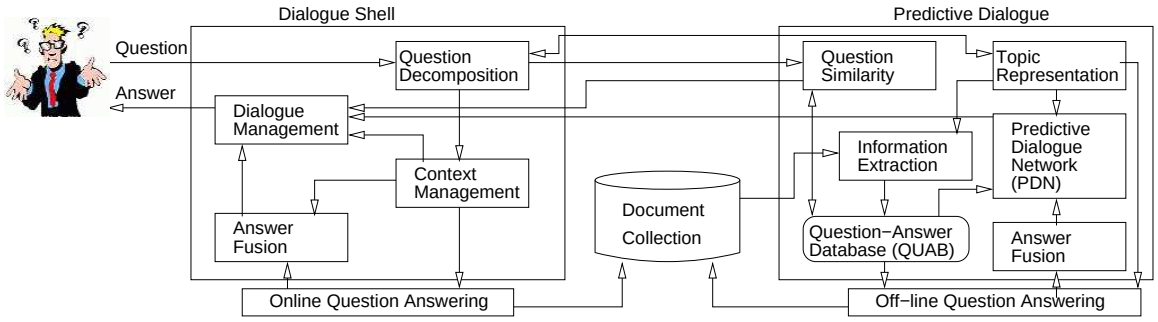

Figure 1: FERRET - A Predictive Interactive Question-Answering Architecture.

model, and Section 6 summarizes the conclusions.

2 Interactive Question-Answering

We have found that the quality of interactions produced by an interactive Q/A system can be greatly enhanced by predicting the range of questions that a user might ask in the context of a given topic. If a large database of topic-relevant questions were available for a wide variety of topics, the accuracy of a state-of-the-art Q/A system such as (Harabagiu et al., 2003) could be enhanced.

In FERRET, our interactive Q/A system, we store such "predicted" pairs of questions and answers in a database known as the *Question Answer Database* (or QUAB). FERRET uses this large set of topic-relevant question-and-answer pairs to improve the interaction with the user by suggesting new questions. For example, when a user asks a question like (Q1) (as illustrated in Table 1), FERRET returns an answer to the question (A1) and proposes (Q2), (Q3), and (Q4) as suggestions of possible continuations of the dialogue. Users then choose how to continue the interaction by either (1) ignoring the suggestions made by the system and proposing a different question, or by (2) selecting one of the proposed questions and examining its answer.

Figure 1 illustrates the architecture of FERRET. The interactions are managed by a *dialogue shell*, which processes questions by transforming them into their corresponding predicate-argument structures[2].

The data collection used in our experiments was made available by the Center for Non-Proliferation Studies (CNS)[3].

(Q1) What weapons are **included in** Egypt's stockpiles?
(A1) The Israelis point to comments made by former President Anwar Sadat, who in 1970 stated that Egypt has biological weapons stored in refrigerators ready to use against Israel if need be. The program might include **"plague, botulism toxin, encephalitis virus, anthrax, Rift Valley fever and mycotoxicosis."**
(Q2) Where did Egypt **inherit** its first stockpiles of chemical weapons?
(Q3) Is there evidence that Egypt has **dismantled** its stockpiles of weapons?
(Q4) Where are Egypt's weapons stockpiles **located**?
(Q5) Who **oversees** Egypt's weapons stockpiles?

Table 1: User question and proposed questions from QUABs

Modules from the FERRET's dialogue shell interact with modules from the predictive dialogue block. Central to the predictive dialogue is the topic representation for each scenario, which enables the population of a Predictive Dialogue Network (PDN). The PDN consists of a large set of questions that were asked or predicted for each topic. It is a network because questions are related by "similarity" links, which are computed by the Question Similarity module. The topic representation enables an Information Extraction module based on (Surdeanu and Harabagiu, 2002) to find topic-relevant information in the document collection and to use it as answers for the QUABs. The questions associated with each predicted answer are generated from patterns that are related to the extraction patterns used for identifying topic relevant information. The quality of the dialog between the user and FERRET depends on the quality of the topic representations and the coverage of the QUABs.

[2]We have employed the same representation of predicate-argument structures as those encoded in PropBank. We use a semantic parser (described in (Surdeanu et al., 2003)) that recognizes predicate-argument structures.

[3]The Center for Non-Proliferation Studies at the Monterrey Institute of International Studies distributes collections of print and online documents on weapons of mass destruction. More information at: http://cns.miis.edu.

GENERAL BACKGROUND

Serving as a background to the scenarios, the following list contains subject areas that may be relevant to the scenarios under examination, and it is provided to assist the analyst in generating questions.

1) Country Profile
2) Government: Type of, Leadership, Relations
3) Military Operations: Army, Navy, Air Force, Leaders, Capabilities, Intentions
4) Allies/Partners: Coalition Forces
5) Weapons: Chemical, Biological, Materials, Stockpiles, Facilities, Access, Research Efforts, Scientists
6) Citizens: Population, Growth Rate, Education
7) Industrial: Major Industrires, Exports, Power Sources
8) Economics: Growth Domestic Product, Growth Rate, Imports
9) Threat Perception: Border and Surrounding States, International, Terrorist Groups
10) Behaviour: Threats, Invasions, Sponsorship and Harboring of Bad Actors
11) Transportation Infrastructure: Kilometers of Road, Rail, Air Runways, Harbors and Ports, Rivers
12) Beliefs: Ideology, Goals, Intentions
13) Leadership:
14) Behaviour: Threats to use WMDs, Actual Usage, Sophistication of Attack, Anectodal or Simultaneous
15) Weapons: Chemical, Bilogical, Materials, Stockpiles, Facilities, Access

SCENARIO: Assessment of Egypt's Biological Weapons

As terrorist Activity in Egypt increases, the Commander of the United States Army believes a better understanding of Egypt's Military capabilities is needed. Egypt's biological weapons database needs to be updated to correspond with the Commander's request. Focus your investigation on Egypt's access to old technology, assistance received from the Soviet Union for development of their pharmaceutical infrastructure, production of toxins and BW agents, stockpiles, exportation of these materials and development technology to Middle Eastern countries, and the effect that this information will have on the United States and Coalition Forces in the Middle East. Please incorporate any other related information to your report.

Figure 2: Example of a Dialogue Scenario.

3 Modeling the Dialogue Topic

Our experiments in interactive Q/A were based on several scenarios that were presented to us as part of the ARDA Metrics Challenge Dialogue Workshop. Figure 2 illustrates one of these scenarios. It is to be noted that the *general background* consists of a list of subject areas, whereas the *scenario* is a narration in which several sub-topics are identified (e.g. *production of toxins* or *exportation of materials*). The creation of scenarios for interactive Q/A requires several different types of domain-specific knowledge and a level of operational expertise not available to most system developers. In addition to identifying a particular *domain of interest*, scenarios must specify the set of relevant *actors*, *outcomes*, and *related topics* that are expected to operate within the domain of interest, the salient *associations* that may exist between entities and events in the scenario, and the specific *timeframe* and *location* that bound the scenario in space and time. In addition, real-world scenarios also need to identify certain operational parameters as well, such as the identity of the scenario's *sponsor* (i.e. the organization sponsoring the research) and *audience* (i.e. the organization receiving the information), as well as a series of *evidence conditions* which specify how much verification information must be subject to before it can be accepted as fact. We assume the set of sub-topics mentioned in the general background and the scenario can be used together to define a topic structure that will govern future interactions with the Q/A system. In order to model this structure, the topic representation that we create considers separate *topic signatures* for each sub-topic.

The notion of topic signatures was first introduced in (Lin and Hovy, 2000). For each subtopic in a scenario, given (a) documents relevant to the sub-topic and (b) documents not relevant to the subtopic, a statistical method based on the likelihood ratio is used to discover a weighted list of the most topic-specific concepts, known as the topic signature. Later work by (Harabagiu, 2004) demonstrated that topic signatures can be further enhanced by discovering the most relevant relations that exist between pairs of concepts. However, both of these types of topic representations are limited by the fact that they require the identification of topic-relevant documents prior to the discovery of the topic signatures. In our experiments, we were only presented with a set of documents relevant to a particular scenario; no further relevance information was provided for individual subject areas or sub-topics.

In order to solve the problem of finding relevant documents for each subtopic, we considered four different approaches:

- **Approach 1**: All documents in the CNS collection were initially clustered using K-Nearest Neighbor (KNN) clustering (Dudani, 1976). Each cluster that contained at least one keyword that described the sub-topic was deemed relevant to the topic.

- **Approach 2**: Since individual documents may contain discourse segments pertaining to different sub-topics, we first used TextTiling (Hearst, 1994) to automatically segment all of the documents in the CNS collection into individual text tiles. These individual discourse segments

then served as input to the KNN clustering algorithm described in Approach 1.

- **Approach 3**: In this approach, relevant documents were discovered simultaneously with the discovery of topic signatures. First, we associated a binary *seed relation* r_i for each each sub-topic S_i. (Seed relations were created both by hand and using the method presented in (Harabagiu, 2004).) Since seed relations are by definition relevant to a particular subtopic, they can be used to determine a binary partition of the document collection C into (1) a relevant set of documents R_i (that is, the documents relevant to relation r_i) and (2) a set of non-relevant documents C-R_i. Inspired by the method presented in (Yangarber et al., 2000), a topic signature (as calculated by (Harabagiu, 2004)) is then produced for the set of documents in R_i. For each subtopic S_i defined as part of the dialogue scenario, documents relevant to a corresponding seed relation r_i are added to R iff the relation r_i meets the *density criterion* (as defined in (Yangarber et al., 2000)). If D represents the set of documents where r_i is recognized, then the density criterion can be defined as: $\frac{|D \cap R|}{D \cap C} \gg \frac{|R|}{|C|}$. Once D is added to R_i, then a new topic signature is calculated for R. Relations extracted from the new topic signature can then be used to determine a new document partition by re-iterating the discovery of the topic signature and of the documents relevant to each subtopic.

- **Approach 4**: Approach 4 implements the technique described in Approach 3, but operates at the level of discourse segments (or texttiles) rather than at the level of full documents. As with Approach 2, segments were produced using the TextTiling algorithm.

In modeling the dialogue scenarios, we considered three types of topic-relevant relations: (1) *structural relations*, which represent hypernymy or meronymy relations between topic-relevant concepts, (2) *definition relations*, which uncover the characteristic properties of a concept, and (3) *extraction relations*, which model the most relevant events or states associated with a sub-topic. Although structural relations and definition relations are discovered reliably using patterns available from our Q/A system (Harabagiu et al., 2003), we found only extraction relations to be useful in determining the set of documents relevant to a subtopic. Structural relations were available from concept ontologies implemented in the Q/A system. The definition relations were identified by patterns used for processing definition questions.

Extraction relations are discovered by processing documents in order to identify three types of relations, including: (1) syntactic attachment relations (including subject-verb, object-verb, and verb-PP relations), (2) predicate-argument relations, and (3) salience-based relations that can be used to encode long-distance dependencies between topic-relevant concepts. (Salience-based relations are discovered using a technique first reported in (Harabagiu, 2004) which approximates a Centering Theory-style approach (Kameyama, 1997) to the resolution of coreference.)

Subtopic: Egypt's production of toxins and BW agents

Topic Signature:
produce − phosphorous trichloride (TOXIN)
house − ORGANIZATION
cultivate − non−pathogenic Bacilus Subtilis (TOXIN)
produce − mycotoxins (TOXIN)
acquire − FACILITY

Subtopic: Egypt's allies and partners

Topic Signature:	
provide − COUNTRY	cooperate − COUNTRY
cultivate − COUNTRY	train − PERSON
supply − precursors	supply − know−how

Figure 3: Example of two topic signatures acquired for the scenario illustrated in Figure 2.

We made the extraction relations associated with each topic signature more general (a) by replacing words with their (morphological) root form (e.g. *wounded* with *wound*, *weapons* with *weapon*), (b) by replacing lexemes with their subsuming category from an ontology of 100,000 words (e.g. *truck* is replaced by VEHICLE, ARTIFACT, or OBJECT), and (c) by replacing each name with its name class (*Egypt* with COUNTRY). Figure 3 illustrates the topic signatures resulting for the scenario illustrated in Figure 2.

Once extraction relations were obtained for a particular set of documents, the resulting set of relations were ranked according to a method proposed in (Yangarber, 2003). Under this approach,

the score associated with each relation is given by: $score(r) = \frac{Sup(r)}{|D|} * \log_2 Sup(r)$, where $|D|$ represents the cardinality of the documents where the relation is identified, and $Sup(r)$ represents support associated with the relation r. $Sup(r)$ is defined as the sum of the relevance of each document in D: $Sup(r) = \sum_{d \in D} Rel(d)$. The relevance of a document that contains a topic-significant relation can be defined as: $Rel(d) = 1 - \prod_{r \in TS}(1 - Prec(r))$, where TS represents the topic signature of the subtopic[4]. The accuracy of the relation, then, is given by: $Prec(r) = \frac{1}{|D|}(\sum_{d \in D} Rel^{s_i}(d) - \sum_{j \neq i} Rel^{s_j}(d))$. Here, $Rel^{S_i}(d)$ measures the relevance of a subtopic S_i to a particular document d, while $Rel^{S_j}(d)$ measures the relevance of d to another subtopic, S_j.

We use a different learner for each subtopic in order to train simultaneously on each iteration. (The calculation of topic signatures continues to iterate until there are no more relations that can be added to the overall topic signature.) When the precision of a relation to a subtopic S_i is computed, it takes into account the *negative* evidence of its relevance to any other subtopic $S_i \neq S_j$. If $Prec(r) \leq 0$, the relation is not included in the topic signature, where relations are ranked by the score $Score(r) = Prec(r) * log(Sup(r))$.

Representing topics in terms of relevant concepts and relations is important for the processing of questions asked within the context of a given topic. For interactive Q/A, however, the ideal topic-structured representation would be in the form of question-answer pairs (QUABs) that model the individual segments of the scenario. We have currently created two sets of QUABs: a handcrafted set and an automatically-generated set. For the manually-created set of QUABs, 4 linguists manually generated 3210 question-answer pairs for each of the 8 dialogue scenarios considered in our experiments.

In a separate effort, we devised a process for automatically populating the QUAB for each scenario. In order to generate question-answer pairs for each subtopic, we first identified relevant text passages in the document collection to serve as "answers" and then generated individual questions that could be answered by each answer passage.

◇ **Answer Identification**: We defined an *answer passage* as a contiguous sequence of sentences with a positive *answer rank* and a *passage price* of ≤ 4. To select answer passages for each subtopic S_i, we calculate an *answer rank*, $rank(a) = \sum_{r_i} score(r_i)$, that sums across the scores of each relation from the topic signature that is identified in the same text window. Initially, the text window is set to one sentence. (If the sentence is part of a quote, however, the text window is immediately expanded to encompass the entire sentence that contains the quote.) Each passage with $rank(a) > 0$ is then considered to be a *candidate answer passage*. The text window of each candidate answer passage is then expanded to include the following sentence. If the answer rank does not increase with the addition of the succeeding sentence, then the *price* (p) of the candidate answer passage is incremented by 1, otherwise it is decremented by 1. The text window of each candidate answer passage continues to expand until $p = 4$. Before the ranked list of candidate answers can be considered by the Question Generation module, answer passages with a positive price p are stripped of the last p sentences.

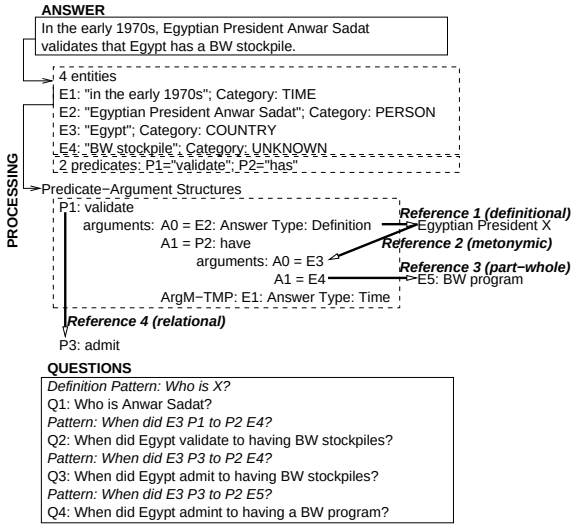

Figure 4: Associating Questions with Answers.

◇ **Question Generation**: In order to automatically generate questions from answer passages, we considered the following two problems:

- **Problem 1**: Every word in an answer passage can refer to an entity, a relation, or an event. In order for question generation be successful, we must determine whether a particular reference

[4]Initially, TS contains only the seed relation. Additional relations can be added with each iteration.

is "interesting" enough to the scenario such that it deserves to be mentioned in a topic-relevant question. For example, Figure 4 illustrates an answer that includes two predicates and four entities. In this case, four types of reference are used to associate these linguistic objects with other related objects: (a) *definitional reference*, used to link entity (E1) "*Anwar Sadat*" to a corresponding attribute "*Egyptian President*", (b) *metonymic reference*, since (E1) can be coerced into (E2), (c) *part-whole reference*, since "*BW stockpiles*"(E4) necessarily imply the existence of a "*BW program*"(E5), and (d) *relational reference*, since *validating* is subsumed as part of the meaning of *declaring* (as determined by WordNet glosses), while *admitting* can be defined in terms of *declaring*, as in *declaring [to be true]*.

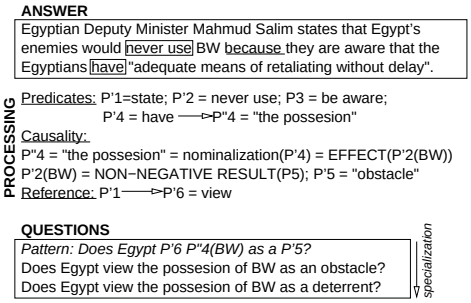

Figure 5: Questions for Implied Causal Relations.

- **Problem 2**: We have found that the identification of the association between a candidate answer and a question depends on (a) the recognition of predicates and entities based on both the output of a named entity recognizer and a semantic parser (Surdeanu et al., 2003) and their structuring into predicate-argument frames, (b) the resolution of reference (addressed in Problem 1), (c) the recognition of implicit relations between predications stated in the answer. Some of these implicit relations are referential, as is the relation between predicates P_1 and P_3 illustrated in Figure 4. A special case of implicit relations are the causal relations. Figure 5 illustrates an answer where a causal relation exists and is marked by the cue phrase *because*. Predicates – like those in Figure 5 – can be phrasal (like P_3') or negative (like P_2'). Causality is established between predicates P_2' and P_4' as they are the ones that ultimately determine the selection of the answer. The predicate p_4' can be substituted by its nominalization since Arg_1 of P_2 is *BW*, the same argument is transferred to P_4''. The causality implied by the answer from Figure 5 has two components: (1) the effect (i.e. the predicate P_4'') and (2) the result, which eliminates the semantic effect of the negative polarity item *never* by implying the predicate p_5, *obstacle*. The questions that are generated are based on question patterns associated with causal relations and therefore allow different degrees for the specificity of the resultative, i.e *obstacle* or *deterrent*.

We generated several questions for each answer passage. Questions were generated based on patterns that were acquired to model interrogations using relations between predicates and their arguments. Such interrogations are based on (1) associations between the answer type (e.g. DATE) and the question stem (e.g. "*when*" and (2) the relation between predicates, question stem and the words that determine the answer type (Narayanan and Harabagiu, 2004). In order to obtain these predicate-argument patterns, we used 30% (approximately 1500 questions) of the handcrafted question-answer pairs, selected at random from each of the 8 dialogue scenarios. As Figures 4 and 5 illustrate, we used patterns based on (a) embedded predicates and (b) causal or counterfactual predicates.

4 Managing Interactive Q/A Dialogues

As illustrated in Figure 1, the main idea of managing dialogues in which interactions with the Q/A system occur is based on the notion of predictions, i.e. by proposing to the user a small set of questions that tackle the same subject as her question (as illustrated in Table 1). The advantage is that the user can follow-up with one of the pre-processed questions, that has a correct answer and resides in one of the QUABs. This enhances the effectiveness of the dialogue. It also may impact on the efficiency, i.e. the number of questions being asked if the QUABs have good coverage of the subject areas of the scenario. Moreover, complex questions, that generally are not processed with high accuracy by current state-of-the-art Q/A systems, are associated with predictive questions that represent decompositions based on

similarities between predicates and arguments of the original question and the predicted questions.

The selection of the questions from the QUABs that are proposed for each user question is based on a similarity-metric that ranks the QUAB questions. To compute the similarity metric, we have experimented with seven different metrics. The first four metrics were introduced in (Lytinen and Tomuro, 2002).

- **Similarity Metric 1** is based on two processing steps:
  (a) the content words of the questions are weighted using the $tfidf$ measure used in Information Retrieval $w_i = w(t_i) = (1 + \log(tf_i))\frac{\log N}{df_i}$, where N is the number of questions in the QUAB, df_i is the number of questions containing t_i and tf_i is the number of times t_i appears in the question. This allows the user question and any QUAB question to be transformed into two vectors, $v_u = \langle w_{u_1}, w_{u_2}, ..., w_{u_n} \rangle$ and $v_q = \langle w_{q_1}, w_{q_2}, ..., w_{q_m} \rangle$;
  (b) the term vector similarity is used to compute the similarity between the user question and any question from the QUAB: $\cos(v_u, v_q) = (\sum_i w_{u_i} w_{q_i}) / ((\sum_i w_{u_i}^2)^{\frac{1}{2}} \times (\sum_i w_{q_i}^2)^{\frac{1}{2}})$

- **Similarity Metric 2** is based on the percent of user question terms that appear in the QUAB question. It is obtained by finding the intersection of the terms in the term vectors of the two questions.

- **Similarity Metric 3** is based on semantic information available from WordNet. It involves:
  (a) finding the minimum path between WordNet concepts. Given two terms t_1 and t_2, each with n and m WordNet senses $S_1 = \{s_1, ..., s_n\}$ and $S_2 = \{r_1, ..., r_m\}$. The semantic distance between the terms $\delta(t_1, t_2)$ is defined by the minimum of all the possible pairwise semantic distances between S_1 and S_2: $\delta(t_1, t_2) = \min_{s_i \in S_1, r_j \in S_2} D(s_i, r_j)$, where $D(s_i, r_j)$ is the path length between s_i and r_j.
  (b) the semantic similarity between the user question $T_u = \langle u_1, u_2, ..., u_n \rangle$ and the QUAB question $T_q = \langle b_1, b_2, ..., b_m \rangle$ to be defined as $sem(T_u, T_q) = \frac{I(T_u, T_q) + I(T_q, T_u)}{|T_u| + |T_q|}$, where $I(T_x, T_y) = \sum_{x \in T_x} \frac{1}{1 + \min_{y \in T_y} \delta(x,y)}$

- **Similarity Metric 4** is based on the question type similarity. Instead of using the question class, determined by its stem, whenever we could recognize the answer type expected by the question, we used it for matching. As backoff only, we used a question type similarity based on a matrix akin to the one reported in (Lytinen and Tomuro, 2002)

- **Similarity Metric 5** is based on question concepts rather than question terms. In order to translate question terms into concepts, we replaced (a) *question stems* (i.e. a *WH*-word + NP construction) with expected answer types (taken from the answer type hierarchy employed by FERRET's Q/A system) and (b) *named entities* with corresponding their corresponding classes. Remaining nouns and verbs were also replaced with their WordNet semantic classes, as well. Each concept was then associated with a weight: concepts derived from named entities classes were weighted heavier than concepts from answer types, which were in turn weighted heavier than concepts taken from WordNet clases. Similarity was then computed across "matching" concepts. [5] The resultant similarity score was based on three variables:
  x = sum of the weights of all concepts matched between a *user query* (Q_u) and a *QUAB query* (Q_b);
  y = sum of the weights of all unmatched concepts in Q_u;
  z = sum of the weights of all unmatched concepts in Q_b;
  The similarity between Q_u and Q_b was calculated as $x - (p_u \times y) - (p_b \times z)$, where p_u and p_b were used as coefficients to penalize the contribution of unmatched concepts in Q_u and Q_b respectively. [6]

- **Similarity Metric 6** is based on the fact that the

[5] In the case of ambiguous nouns and verbs associated with multiple WordNet classes, all possible classes for a term were considered in matching.

[6] We set p_u = 0.4 and p_b = 0.1 in our experiments.

Q1: Does Iran have an indigenous CW program?	**QUABs:** (1a) How did Iran start its CW program? (1b) Has the plant at Qazvin been linked to CW production? (1c) What CW does Iran produce?
Answer (A1):	
Although Iran is making a concerted effort to attain an independent production capability for all aspects of chemical weapons program, it remains dependent on foreign sources for chemical warfare−related technologies.	
Q2: Where are Iran's CW facilities located?	**QUABs:** (2a) What factories in Iran could produce CW? (2b) Where are Iran's stockpiles of CW? (2c) Where has Iran bought equipment to produce CW?
Answer(A2):	
According to several sources, Iran's primary suspected chemical weapons production facility is located in the city of Damghan.	
Q3: What is Iran's goal for its CW program?	**QUABs:** (3a) What motivated Iran to expand its chemical weapons program? (3b) How do CW figure into Iran's long−term strategic plan? (3c) What are Iran's future CW plans?
Answer(A3):	
In their pursuit of regional hegemony, Iran and Iraq probably regard CW weapons and missiles as necessary to support their political and military objectives. Possession of chemical weapons would likely lead to increased intimidation of their Gulf, neighbors, as well as increased willingness to confront the United States.	

Figure 6: A sample interactive Q/A dialogue.

QUAB questions are clustered based on their mapping to a vector of important concepts in the QUAB.The clustering was done using the K-Nearest Neighbor (KNN) method (Dudani, 1976). Instead of measuring the similarity between the user question and each question in the QUAB, similarities are computed only between the user question and the centroid of each cluster.

- **Similarity Metric 7** was derived from the results of Similarity Metrics 5 and 6 above. In this case, if the QUAB question (Q_b) that was deemed to be most similar to a user question (Q_u) under Similarity Metric 5 is contained in the cluster of QUAB questions deemed to be most similar to Q_u under Similarity Metric 6, then Q_b receives a *cluster adjustment score* in order to boost its ranking within its QUAB cluster. We calculate the cluster adjustment score as $score_{adj}(Q_b) = (sim_5 * (1 - C_f)) + (sim_6 * C_f)$, where C_f represents the difference in rank between the centroid of the cluster and the previous rank of the QUAB question Q_b.

In the currently-implemented version of FERRET, we used Similarity Metric 5 to automatically identify the set of 10 QUAB questions that were most similar to a user's question. These question-and-answer pairs were then returned to the user – along with answers from FERRET's automatic Q/A system – as potential continuations of the Q/A dialogue. We used the remaining 6 similarity metrics described in this section to manually assess the impact of similarity on a Q/A dialogue.

5 Experiments with Interactive Q/A Dialogues

To date, we have used FERRET to produce over 90 Q/A dialogues with human users. Figure 6 illustrates three turns from a real dialogue from a human user investigating Iran's chemical weapons prorgram. As it can be seen coherence can be established between the user's questions and the system's answers (e.g. Q3 is related to both A1 and A3) as well as between the QUABs and the user's follow-up questions (e.g. QUAB (1b) is more related to Q2 than either Q1 or A1). Coherence alone is not sufficient to analyze the quality of interactions, however.

In order to better understand interactive Q/A dialogues, we have conducted three sets of experiments with human users of FERRET. In these experiments, users were allotted two hours to interact with Ferret to gather information requested by a dialogue scenario similar to the one presented in Figure 2. In Experiment 1 (E1), 8 U.S. Navy Reserve (USNR) intelligence analysts used FERRET to research 8 different scenarios related to chemical and biological weapons. Experiment 2 and Experiment 3 considered several of the same scenarios addressed in E1: E2 included 24 mixed teams of analysts and novice users working with 2 scenarios, while E3 featured 4 USNR analysts working with 6 of the original 8 scenarios. (Details for each experiment are provided in Table 2.) Users were also given a task to focus their

research; in E1 and E3, users prepared a short report detailing their findings; in E2, users were given a list of "challenge" questions to answer.

Exp	Users	QUABs?	Scenarios	Topics
E1	8	Yes	8	Egypt BW, Russia CW, South Africa CW, India CW, North Korea CBW, Pakistan CW, Libya CW, Iran CW
E2	24	Yes	2	Egypt BW, Russia CW
E3	4	No	6	Egypt BW, Russia CW, North Korea CBW, Pakistan CW India CW, Libya CW, Iran CW

Table 2: Experiment details

In E1 and E2, users had access to a total of 3210 QUAB questions that had been hand-created by developers for each the 8 dialogue scenarios. (Table 3 provides totals for each scenario.) In E3, users performed research with a version of FERRET that included no QUABs at all.

Scenario	Handcrafted QUABs
INDIA	460
LIBYA	414
IRAN	522
NORTH KOREA	316
PAKISTAN	322
SOUTH AFRICA	454
RUSSIA	366
EGYPT	356
Testing Total	3210

Table 3: QUAB distribution over scenarios

We have evaluated FERRET by measuring efficiency, effectiveness, and user satisfaction:

Efficiency FERRET's QUAB collection enabled users in our experiments to find more relevant information by asking fewer questions. When manually-created QUABs were available (E1 and E2), users submitted an average of 12.25 questions each session. When no QUABs were available (E3), users entered a total of 44.5 questions per session. Table 4 lists the number of QUAB question-answer pairs selected by users and the number of user questions entered by users during the 8 scenarios considered in E1. In E2, freed from the task of writing a research report, users asked significantly ($p < 0.05$) fewer questions and selected fewer QUABs than they did in E1. (See Table 5).

Effectiveness QUAB question-answer pairs also improved the overall accuracy of the answers returned by FERRET. To measure the effectiveness of a Q/A dialogue, human annotators were used to perform a post-hoc analysis of how relevant the QUAB pairs returned by FERRET were to each question

Country	n	QUAB (avg.)	User Q (avg.)	Total (avg.)
India	2	21.5	13.0	34.5
Libya	2	12.0	9.0	21.0
Iran	2	18.5	11.0	29.5
N.Korea	2	16.5	7.5	34.0
Pakistan	2	29.5	15.5	45.0
S.Africa	2	14.5	6.0	20.5
Russia	2	13.5	15.5	29.0
Egypt	2	15.0	20.5	35.5
TOTAL(E1)	16	17.63	12.25	29.88

Table 4: Efficiency of Dialogues in Experiment 1

Country	n	QUAB (avg.)	User Q (avg.)	Total (avg.)
Russia	24	8.2	5.5	13.7
Egypt	24	10.8	7.6	18.4
TOTAL(E2)	48	9.50	6.55	16.05

Table 5: Efficiency of Dialogues in Experiment 2

entered by a user: each QUAB pair returned was graded as "relevant" or "irrelevant" to a user question in a forced-choice task. Aggregate relevance scores were used to calculate (1) the percentage of relevant QUAB pairs returned and (2) the mean reciprocal rank (MRR) for each user question. MRR is defined as $\frac{1}{n}\sum_{i=1}\frac{1}{r_i}$, whree r_i is the lowest rank of any relevant answer for the i^{th} user query[7]. Table 6 describes the performance of FERRET when each of the 7 similarity measures presented in Section 4 are used to return QUAB pairs in response to a query. When only answers from FERRET's automatic Q/A system were available to users, only 15.7% of system responses were deemed to be relevant to a user's query. In contrast, when manually-generated QUAB pairs were introduced, as high as 84% of the system's responses were deemed to be relevant. The results listed in Table 6 show that the best metric is Similarity Metric 5. Thse results suggest that the selection of relevant questions depends on sophisticated similarity measures that rely on conceptual hierarchies and semantic recognizers.

We evaluated the quality of each of the four sets of automatically-generated QUABs in a similar fashion. For each question submitted by a user in E1, E2, and E3, we collected the top 5 QUAB question-answer pairs (as determined by Similarity Metric 5) that FERRET returned. As with the manually-generated QUABs, the automatically-

[7]We chose MRR as our scoring metric because it reflects the fact that a user is most likely to examine the first few answers from any system, but that all correct answers returned by the system have some value because users will sometimes examine a very large list of query results.

	% of Top 5 Responses Relevant to User Q	% of Top 1 Responses Relevant to User Q	MRR
Without QUAB	15.73%	26.85%	0.325
Similarity 1	82.61%	60.63%	0.703
Similarity 2	79.95%	58.45%	0.681
Similarity 3	79.47%	56.04%	0.664
Similarity 4	78.26%	46.14%	0.592
Similarity 5	**84.06%**	**68.36%**	**0.753**
Similarity 6	81.64%	56.04%	0.671
Similarity 7	84.54%	64.01%	0.730

Table 6: Effectiveness of dialogs

generated pairs were submitted to human assessors who annotated each as "relevant" or irrelevant to the user's query. Aggregate scores are presented in Table 7.

	Egypt		Russia	
Approach	% of Top 5 Responses Rel. to User Q	MRR	% of Top 5 Responses Rel. to User Q	MRR
Approach 1	40.01%	0.295	60.25%	0.310
Approach 2	36.00%	0.243	72.00%	0.475
Approach 3	44.62%	0.271	60.00%	0.297
Approach 4	68.05%	0.510	68.00%	0.406

Table 7: Quality of QUABs acquired automatically

User Satisfaction Users were consistently satisfied with their interactions with FERRET. In all three experiments, respondents claimed that they found that FERRET (1) gave meaningful answers, (2) provided useful suggestions, (3) helped answer specific questions, and (4) promoted their general understanding of the issues considered in the scenario. Complete results of this study are presented in Table 8[8].

Factor	E1	E2	E3
Promoted understanding	3.40	3.20	3.75
Helped with specifi c questions	3.70	3.60	3.25
Make good use of questions	3.40	3.55	3.0
Gave new scenario insights	3.00	3.10	2.2
Gave good collection coverage	3.75	3.70	3.75
Stimulated user thinking	3.50	3.20	2.75
Easy to use	3.50	3.55	4.10
Expanded understanding	3.40	3.20	3.00
Gave meaningful answers	4.10	3.60	2.75
Was helpful	4.00	3.75	3.25
Helped with new search methods	2.75	3.05	2.25
Provided novel suggestions	3.25	3.40	2.65
Is ready for work environment	2.85	2.80	3.25
Would speed up work	3.25	3.25	3.00
Overall like of system	3.75	3.60	3.75

Table 8: User Satisfaction Survey Results

6 Conclusions

We believe that the quality of Q/A interactions depends on the modeling of scenario topics. An ideal model is provided by question-answer databases (QUABs) that are created off-line and then used to make suggestions to a user of potential relevant continuations of a discourse. In this paper, we have presented FERRET, an interactive Q/A system which makes use of a novel Q/A architecture that integrates QUAB question-answer pairs into the processing of questions. Experiments with FERRET have shown that, in addition to being rapidly adopted by users as valid suggestions, the incorporation of QUABs into Q/A can greatly improve the overall accuracy of an interactive Q/A dialogue.

[8]Evaluation scale: 1-does not describe the system, 5-completely describes the system

References

S. Dudani. 1976. The distance-weighted k-nearest-neighbour rule. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-6(4):325–327.

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

Sanda Harabagiu. 2004. Incremental Topic Representations. In *Proceedings of the 20th COLING Conference*, Geneva, Switzerland.

Marti Hearst. 1994. Multi-Paragraph Segmentation of Expository Text. In *Proceedings of the 32nd Meeting of the Association for Computational Linguistics*, pages 9–16.

Megumi Kameyama. 1997. Recognizing Referential Links: An Information Extraction Perspective. In *Workshop of Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts, (ACL-97/EACL-97)*, pages 46–53.

Chin-Yew Lin and Eduard Hovy. 2000. The Automated Acquisition of Topic Signatures for Text Summarization. In *Proceedings of the 18th COLING Conference*, pages 495–501.

S. Lytinen and N. Tomuro. 2002. The Use of Question Types to Match Questions in FAQFinder. In *Papers from the 2002 AAAI Spring Symposium on Mining Answers from Texts and Knowledge Bases*, pages 46–53.

Srini Narayanan and Sanda Harabagiu. 2004. Question Answering Based on Semantic Structures. In *Proceedings of the 20th COLING Conference*, Geneva, Switzerland.

Mihai Surdeanu and Sanda M. Harabagiu. 2002. Infratructure for open-domanin information extraction. In *Conference for Human Language Technology (HLT-2002)*.

Mihai Surdeanu, Sanda M. Harabagiu, John Williams, and Paul Aarseth. 2003. Using predicate-argument structures for information extraction. In *ACL*, pages 8–15.

Roman Yangarber, Ralph Grishman, Pasi Tapanainen, and Silja Huttunen. 2000. Automatic Acquisition of Domain Knowledge for Information Extraction. In *Proceedings of the 18th COLING Conference*, pages 940–946.

Roman Yangarber. 2003. Counter-Training in Discovery of Semantic Patterns. In *Proceedings of the 41th Meeting of the Association for Computational Linguistics*, pages 343–350.