Enhanced Interactive Question-Answering with Conditional Random Fields

Andrew Hickl and Sanda Harabagiu
Language Computer Corporation
Richardson, Texas 75080
andy@languagecomputer.com

Abstract

This paper describes a new methodology for enhancing the quality and relevance of suggestions provided to users of interactive Q/A systems. We show that by using Conditional Random Fields to combine relevance feedback gathered from users along with information derived from discourse structure and coherence, we can accurately identify irrelevant suggestions with nearly 90% F-measure.

1 Introduction

Today's interactive question-answering (Q/A) systems enable users to pose questions in the context of extended dialogues in order to obtain information relevant to complex research scenarios. When working with an interactive Q/A system, users formulate sequences of questions which they believe will return answers that will let them reach certain information goals.

Users need more than answers, however: while they might be cognizant of many of the different types of information that they need, few – if any – users are capable of identifying all of the questions that must be asked and answered for a particular scenario. In order to take full advantage of the Q/A capabilities of current systems, users need access to sources of domain-specific knowledge that will expose them to new concepts and ideas and will allow them to ask better questions.

In previous work (Hickl et al., 2004; Harabagiu et al., 2005a), we have argued that interactive question-answering systems should be based on a *predictive dialogue architecture* which can be used to provide users with both precise answers to their questions as well as suggestions of relevant research topics that could be explored throughout the course of an interactive Q/A dialogue.

Typically, the quality of interactive Q/A dialogues has been measured in three ways: (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 answer returned, and (3) user satisfaction (Scholtz and Morse, 2003).

In our experiments with an interactive Q/A system, (known as FERRET), we found that performance in each of these areas improves as users are provided with suggestions that are relevant to their domain of interest. In FERRET, suggestions are made to users in the form of predictive question-answer pairs (known as QUABs) which are either generated automatically from the set of documents returned for a query (using techniques first described in (Harabagiu et al., 2005a)), or are selected from a large database of questions-answer pairs created off-line (prior to a dialogue) by human annotators.

Figure 1 presents an example of ten QUABs that were returned by FERRET in response to the question "*How are EU countries responding to the worldwide increase of job outsourcing to India?*".

While FERRET's QUABs are intended to provide users with relevant information about a domain of interest, we can see from Figure 1 that users do not always agree on which QUAB suggestions are relevant. For example, while someone unfamiliar to the notion of "job outsourcing" could benefit from

Proceedings of the Interactive Question Answering Workshop at HLT-NAACL 2006, pages 25–32,
New York City, NY, USA. June 2006.

Relevant?		QUAB Question
$User_1$	$User_2$	
NO	YES	$QUAB_1$: What EU countries are outsourcing jobs to India?
YES	YES	$QUAB_2$: What EU countries have made public statements against outsourcing jobs to India?
NO	YES	$QUAB_3$: What is job outsourcing?
YES	YES	$QUAB_4$: Why are EU companies outsourcing jobs to India?
NO	NO	$QUAB_5$: What measures has the U.S. Congress taken to stem the tide of job outsourcing to India?
YES	NO	$QUAB_6$: How could the anti-globalization movements in EU countries impact the likelihood that the EU Parliament will take steps to prevent job outsourcing to India?
YES	YES	$QUAB_7$: Which sectors of the EU economy could be most affected by job outsourcing?
YES	YES	$QUAB_8$: How has public opinion changed in the EU on job outsourcing issues over the past 10 years?
YES	YES	$QUAB_9$: What statements has French President Jacques Chirac made about job outsourcing?
YES	YES	$QUAB_{10}$: How has the EU been affected by anti-job outsourcing sentiments in the U.S.?

Figure 1: Examples of QUABs.

a QUAB like $QUAB_3$: "*What is job outsourcing?*", we expect that a more experienced researcher would find this definition to be uninformative and potentially irrelevant to his or her particular information needs. In contrast, a complex QUAB like $QUAB_6$: "*How could the anti-globalization movements in EU countries impact the likelihood that the EU Parliament will take steps to prevent job outsourcing to India?*" could provide a domain expert with relevant information, but would not provide enough background information to satisfy a novice user who might not be able to interpret this information in the appropriate context.

In this paper, we present results of a new set of experiments that seek to combine feedback gathered from users with a relevance classifier based on conditional random fields (CRF) in order to provide suggestions to users that are not only related to the topic of their interactive Q/A dialogue, but provide them with the new types of information they need to know.

Section 2 presents the functionality of several of FERRET's modules and describes the NLP techniques for processing questions as well as the framework for acquiring domain knowledge. In Section 3 we present two case studies that highlight the impact of user background. Section 4 describes a new class of user interaction models for interactive Q/A and presents details of our CRF-based classifier. Section 5 presents results from experiments which demonstrate that user modeling can enhance the quality of suggestions provided to both expert and novice users. Section 6 summarizes the conclusions.

2 The FERRET Interactive Question-Answering System

We believe that the quality of interactions produced by an interactive Q/A system can be enhanced by predicting the range of questions that a user might ask while researching a particular topic. By providing suggestions from a large database of question-answer pairs related to a user's particular area of interest, interactive Q/A systems can help users gather the information they need most – without the need for complex, mixed-initiative clarification dialogues.

FERRET uses a large collection of QUAB question-answer pairs in order to provide users with suggestions of new research topics that could be explored over the course of a dialogue. For example, when a user asks a question like *What is the result of the European debate on outsourcing to India?* (as illustrated in (Q1) in Table 1), FERRET returns a set of answers (including (A1) and proposes the questions in (Q2), (Q3), and (Q4) as suggestions of possible continuations of the dialogue. Users then have the freedom to choose how the dialogue should be continued, either by (1) ignoring the suggestions made by the system, (2) selecting one of the proposed QUAB questions and examining its associated answer, or (3) resubmitting the text of the QUAB question to FERRET's automatic Q/A system in order to retrieve a brand-new set of answers.

(Q1) What is the result of the European debate on outsourcing to India?
(A1) Supporters of economic openness understand how outsourcing can strengthen the competitiveness of European companies, as well as benefi t jobs and growth in India.
(Q2) Has the number of customer service jobs outsourced to India increased since 1990?
(Q3) How many telecom jobs were outsourced to India from EU-based companies in the last 10 years?
(Q4) Which European Union countries have experienced the most job losses due to outsourcing over the past 10 years?

Table 1: Sample Q/A Dialogue.

FERRET was designed to evaluate how databases of topic-relevant suggestions could be used to enhance the overall quality of Q/A dialogues. Figure 2 illustrates the architecture of the FERRET system. Questions submitted to FERRET are initially processed by a *dialogue shell* which (1) decomposes complex questions into sets of simpler questions (using techniques first described in (Harabagiu et al., 2005a)), (2) establishes discourse-level relations between the current question and the set of questions

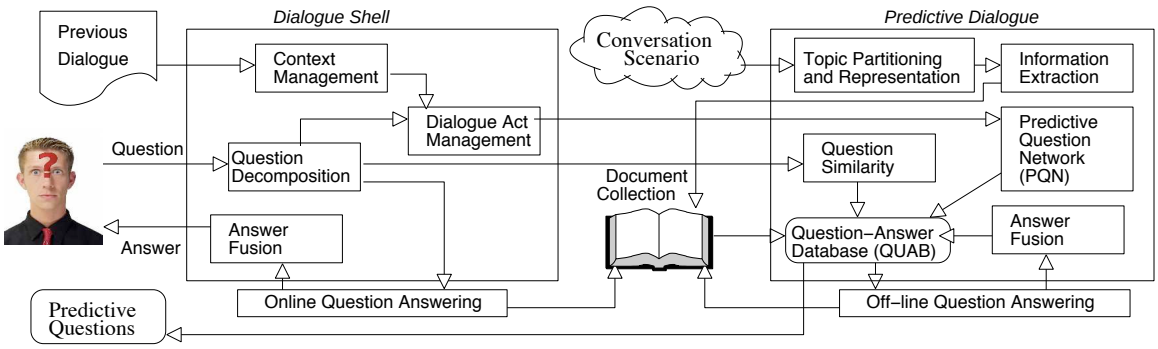

Figure 2: FERRET - A Predictive Interactive Question-Answering Architecture.

already entered into the discourse, and (3) identifies a set of basic dialogue acts that are used to manage the overall course of the interaction with a user.

Output from FERRET's dialogue shell is sent to an *automatic question-answering system* which is used to find answers to the user's question(s). FERRET uses a version of LCC's PALANTIR question-answering system (Harabagiu et al., 2005b) in order to provide answers to questions in documents. Before being returned to users, answer passages are submitted to an *answer fusion* module, which filters redundant answers and combines answers with compatible information content into single coherent answers.

Questions and relational information extracted by the *dialogue shell* are also sent to a *predictive dialogue* module, which identifies the QUABs that best meet the user's expected information requirements. At the core of the FERRET's *predictive dialogue* module is the *Predictive Dialogue Network* (PQN), a large database of QUABs that were either generated off-line by human annotators or created on-line by FERRET (either during the current dialogue or during some previous dialogue)[1]. In order to generate QUABs automatically, documents identified from FERRET's automatic Q/A system are first submitted to a *Topic Representation* module, which computes both topic signatures (Lin and Hovy, 2000) and enhanced topic signatures (Harabagiu, 2004) in order to identify a set of topic-relevant passages. Passages are then submitted to an *Information Extraction* module, which annotates texts with a wide range of lexical, semantic, and syntactic information, including (1) morphological information, (2) named entity information from LCC's CICEROLITE named entity recognition system, (3) semantic dependencies extracted from LCC's PropBank-style semantic parser, and (4) syntactic parse information. Passages are then transformed into natural language questions using a set of question formation heuristics; the resultant QUABs are then stored in the PQN. Since we believe that the same set of relations that hold between questions in a dialogue should also hold between pairs of individual questions taken in isolation, discourse relations are discovered between each newly-generated QUAB and the set of QUABs stored in the PQN. FERRET's *Question Similarity* module then uses the similarity function described in (Harabagiu et al., 2005a) – along with relational information stored in the PQN – in order to identify the QUABs that represent the most informative possible continuations of the dialogue. QUABs are then ranked in terms of their relevance to the user's submitted question and returned to the user.

3 Two Types of Users of Interactive Q/A Systems

In order to return answers that are responsive to users' information needs, interactive Q/A systems need to be sensitive to the different questioning strategies that users employ over the course of a dialogue. Since users gathering information on the same topic can have significantly different information needs, interactive Q/A systems need to be able to accommodate a wide range of question types in order to help users find the specific information that

[1]Techniques used by human annotators for creating QUABs were first described in (Hickl et al., 2004); full details of FERRET's automatic QUAB generation components are provided in (Harabagiu et al., 2005a).

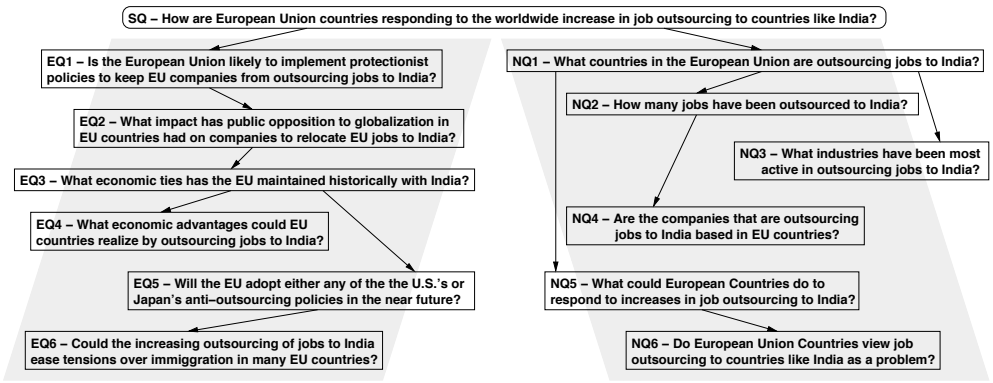

Figure 3: Expert User Interactions Versus Novice User Interactions with a Q/A System.

they are looking for.

In past experiments with users of interactive Q/A systems (Hickl et al., 2004), we have found that a user's access to sources of domain-specific knowledge significantly affects the types of questions that a user is likely to submit to a Q/A system. Users participate in information-seeking dialogues with Q/A systems in order to learn "new" things – that is, to acquire information that they do not currently possess. Users initiate a set of speech acts which allow them to maximize the amount of new information they obtain from the system while simultaneously minimizing the amount of redundant (or previously-acquired) information they encounter. Our experiments have shown that Q/A systems need to be sensitive to two kinds of users: (1) expert users, who interact with a system based on a working knowledge of the conceptual structure of a domain, and (2) novice users, who are presumed to have limited to no foreknowledge of the concepts associated with the domain. We have found that novice users that possess little or no familiarity with a domain employ markedly different questioning strategies than expert users who possess extensive knowledge of a domain: while novices focus their attention in queries that will allow them to discover basic domain concepts, experts spend their time asking questions that enable them to evaluate their hypotheses in the context of a the currently available information. The experts tend to ask questions that refer to the more abstract domain concepts or the complex relations between concepts. In a similar fashion, we have discovered that users who have access to structured sources of domain-specific knowledge (e.g. knowledge bases, conceptual networks or ontologies, or mixed-initiative dialogues) can end up employing more "expert-like" questioning strategies, despite the amount of domain-specific knowledge they possess.

In real-world settings, the knowledge that expert users possess enables them to formulate a set of hypotheses – or belief states – that correspond to each of their perceived information needs at a given moment in the dialogue context. As can be seen in the dialogues presented in Figure 3, expert users generally formulate questions which seek to validate these belief states in the context of a document collection. Given the global information need in S_1, it seems reasonable to presume that questions like EQ_1 and EQ_2 are motivated by a user's expectation that *protectionist policies or* public opposition to globalization could impact a European Union country's willingness to take steps to stem job outsourcing to India. Likewise, questions like EQ_5 are designed to provide the user with information that can decide between two competing belief states: in this case, the user wants to know whether the European Union is more likely to model the United States or Japan in its policies towards job outsourcing. In contrast, without a pre-existing body of domain-specific knowledge to derive reasonable hypotheses from, novice users ask questions that enable them to discover the concepts (and the relations between concepts) needed to formulate new, more specific hypotheses and questions. Returning again to Figure 3, we can see that questions like NQ_1 and NQ_3 are designed to discover new knowledge that the user does not currently possess, while questions like NQ_6 try to

establish whether or not the user's hypothesis (i.e. namely, that EU countries view job outsourcing to India as an problem) is valid and deserves further consideration.

4 User Interaction Models for Relevance Estimation

Unlike systems that utilize mixed initiative dialogues in order to determine a user's information needs (Small and Strzalkowski, 2004), systems (like FERRET) which rely on interactions based on predictive questioning have traditionally not incorporated techniques that allow them to gather relevance feedback from users. In this section, we describe how we have used a new set of user interaction models (UIM) in conjunction with a relevance classifier based on conditional random fields (CRF) (McCallum, 2003; Sha and Pereira, 2003) in order to improve the relevance of the QUAB suggestions that FERRET returns in response to a user's query.

We believe that systems based on predictive questioning can derive feedback from users in three ways. First, systems can learn which suggestions or answers are relevant to a user's domain of interest by tracking which elements users select throughout the course of a dialogue. With FERRET, each answer or suggestion presented to a user is associated with a hyperlink that links to the original text that the answer or QUAB was derived from. While users do not always follow links associated with passages they deem to be relevant to their query, we expect that the set of selected elements are generally more likely to be relevant to the user's interests than unselected elements. Second, since interactive Q/A systems are often used to gather information for inclusion in written reports, systems can identify relevant content by tracking the text passages that users copy to other applications, such as text editors or word processors. Finally, predictive Q/A systems can gather explicit feedback from users through the graphical user interface itself. In a recent version of FERRET, we experimented with adding a "relevance checkbox" to each answer or QUAB element presented to a user; users were then asked to provide feedback to the system by selecting the checkboxes associated with answers that they deemed to be particularly relevant to the topic they were researching.

4.1 User Interaction Models

We have experimented with three models that we have used to gather feedback from users of FERRET. The models are illustrated in Figure 4.

$\underline{\text{UIM}_1}$: Under this model, the set of QUABs that users copied from were selected as relevant; all QUABs not copied from were annotated as irrelevant.
$\underline{\text{UIM}_2}$: Under this model, QUABs that users viewed were considered to be relevant; QUABs that remained unviewed were annotated as irrelevant.
$\underline{\text{UIM}_3}$: Under this model, QUABs that were either viewed or copied from were marked as relevant; all other QUABs were annotated as irrelevant.

Figure 4: User Interaction Models.

With FERRET, users are presented with as many as ten QUABs for every question they submit to the system. QUABs – whether they be generated automatically by FERRET's QUAB generation module, or selected from FERRET's knowledge base of over 10,000 manually-generated question/answer pairs – are presented in terms of their conceptual similarity to the original question. Conceptual similarity (as first described in (Harabagiu et al., 2005a)) is calculated using the version of the cosine similarity formula presented in Figure 5.

Conceptual Similarity weights content terms in Q_1 and Q_2 using $tfidf$ ($w_i = w(t_i) = (1 + \log(tf_i))\frac{\log N}{df_i}$), where N is the number of questions in the QUAB collection, while df_i is equal to the number of questions containing t_i and tf_i is the number of times t_i appears in Q_1 and Q_2. The questions Q_1 and Q_2 can be transformed into two vectors, $v_q = \langle w_{q_1}, w_{q_2}, ..., w_{q_m} \rangle$ and $v_u = \langle w_{u_1}, w_{u_2}, ..., w_{u_n} \rangle$; The similarity between Q_1 and Q_2 is measured as the cosine measure between their corresponding vectors:

$\cos(v_q, v_u) = (\sum_i w_{q_i} w_{u_i}) / ((\sum_i w_{q_i}^2)^{\frac{1}{2}} \times (\sum_i w_{u_i}^2)^{\frac{1}{2}})$

Figure 5: Conceptual Similarity.

In the three models from Figure 4, we allowed users to perform research as they normally would. Instead of requiring users to provide explicit forms of feedback, features were derived from the set of hyper-links that users selected and the text passages that users copied to the system clipboard.

Following (Kristjansson et al., 2004) we analyzed the performance of each of these three models using a new metric derived from the number of relevant QUABs that were predicted to be returned for each model. We calculated this metric – which we refer to as the Expected Number of Irrelevant QUABs – using the formula:

$$p_0(n) = \sum_{k=1}^{10} k p_0(k) \quad (1)$$

$$p_1(n) = (1 - p_0(0)) + \sum_{k=1}^{10} k p_1(k) \quad (2)$$

where $p_m(n)$ is equal to the probability of finding n irrelevant QUABs in a set of 10 suggestions returned to the user given m rounds of interaction. $p_0(n)$ (equation 1) is equal to the probability that all QUABs are relevant initially, while $p_1(n)$ (equation 2) is equal to the probability of finding an irrelevant QUAB after the set of QUABs has been interacted with by a user. For the purposes of this paper, we assumed that all of the QUABs initially returned by FERRET were relevant, and that $p_0(0) = 1.0$. This enabled us to calculate p_1(n) for each of the three models provided in Figure 4.

4.2 Relevance Estimation using Conditional Random Fields

Following work done by (Kristjansson et al., 2004), we used the feedback gathered in Section 4.1 to estimate the probability that a QUAB selected from FERRET'S PQN is, in fact, relevant to a user's original query. We assume that humans gauge the relevance of QUAB suggestions returned by the system by evaluating the informativeness of the QUAB with regards to the set of queries and suggestions that have occurred previously in the discourse. A QUAB, then, is deemed relevant when it conveys content that is sufficiently informative to the user, given what the user knows (i.e. the user's level of expertise) and what the user expects to receive as answers from the system.

Our approach treats a QUAB suggestion as a single node in a sequence of questions $\langle Q_{n-1}, Q_n, QUAB \rangle$ and classifies the QUAB as relevant or irrelevant based on features from the entire sequence.

We have performed relevance estimation using Conditional Random Fields (CRF). Given a random variable x (corresponding to data points $\{x_1, \ldots, x_n\}$) and another random variable y (corresponding to a set of labels $\{y_1, \ldots, y_n\}$), CRFs can be used to calculate the conditional probability $p(y|x)$. Given a sequence $\{x_1, \ldots, x_n\}$ and set of labels $\{y_1, \ldots, y_n\}$, $p(y|x)$ can be defined as:

$$p(y|x) = \frac{1}{z_0} \exp\left(\sum_{n=1}^{N} \sum_{k} \lambda_k f_k(y_{i-1}, y_i, x, n)\right) \quad (3)$$

where z_0 is a normalization factor and λ_k is a weight learned for each feature vector $f_k(y_{i-1}, y_i, x, n)$.

We trained our CRF model in the following way. If we assume that Λ is a set of feature weights $(\lambda_0, \ldots, \lambda_k)$, then we expect that we can use maximum likelihood to estimate values for Λ given a set of training data pairs (x, y).

Training is accomplished by maximizing the log-likelihood of each labeled data point as in the following equation:

$$w_\Lambda = \sum_{i=1}^{N} \log(p_\Lambda(x_i|y_i)) \quad (4)$$

Again, following (Kristjansson et al., 2004), we used the CRF Viterbi algorithm to find the most likely sequence of data points assigned to each label category using the formula:

$$y^* = \arg\max_y p_\Lambda(y|x) \quad (5)$$

Motivated by the types of discourse relations that appear to exist between states in an interactive Q/A dialogue, we introduced a large number of features to estimate relevance for each QUAB suggestion. The features we used are presented in Figure 6

(a) Rank of QUAB: the rank (1, ..., 10) of the QUAB in question.
(b) Similarity: similarity of QUAB, Q_n and QUAB, Q_{n-1}.
(c) Relation likelihood: equal to the likelihood of each predicate-argument structure included in QUAB given all QUABs contained in FERRET's QUAB; calculated for Arg-0, Arg-1, and ArgM-TMP for each predicate found in QUAB suggestions. (Predicate-argument relations were identifi ed using a semantic parser trained on PropBank (Palmer et al., 2005) annotations.)
(d) Conditional Expected Answer Type likelihood: equal to the joint probability $p(EAT_{QUAB}\|EAT_{question})$ calculated from a corpus of dialogues collected from human users of FERRET.
(e) Terms in common: real-valued feature equal to the number of terms in common between the QUAB and both Q_n and Q_{n-1}.
(f) Named Entities in common: same as terms in common, but calculated for named entities detected by LCC's CIEROLITE named entity recognition system.

Figure 6: Relevance Features.

In the next section, we describe how we utilized the user interaction model described in Subsection 4.1 in conjunction with this subsection in order to improve the relevance of QUAB suggestions returned to users.

5 Experimental Results

In this section, we describe results from two experiments that were conducted using data collected from human interactions with FERRET.

In order to evaluate the effectiveness of our relevance classifier, we gathered a total of 1000 questions from human dialogues with FERRET. 500 of

these came from interactions (41 dialogues) where the user was a self-described "expert" on the topic; another selection of 500 questions came from a total of 23 dialogues resulting from interactions with users who described themselves as "novice" or were otherwise unfamiliar with a topic. In order to validate the user's self-assessment, we selected 5 QUABs at random from the set of manually created QUABs assembled for each topic. Users were asked to provide written answers to those questions. Users that were judged to have correctly answered three out of five questions were considered "experts" for the purpose of our experiments. Table 2 presents the breakdown of questions across these two conditions.

User Type	Unique Topics	# Dialogues	Avg # of Qs/dialogue	Total Qs
Expert	12	41	12.20	500
Novice	8	23	21.74	500
Total	12	64	15.63	1000

Table 2: Question Breakdown.

Each of these experiments were run using a version of FERRET that returned the top 10 most similar QUABs from a database that combined manually-created QUABs with the automatically-generated QUABs created for the user's question. While a total of 10,000 QUABs were returned to users during these experiments, only 3,998 of these QUABs were unique (39.98%).

We conducted two kinds of experiments with users. In the first set of experiments, users were asked to mark all of the relevant QUABs that FERRET returned in response to questions submitted by users. After performing research on a particular scenario, expert and novice users were then supplied with as many as 65 questions (and associated QUABs) taken from previously-completed dialogues on the same scenario; users were then asked to select checkboxes associated with QUABs that were relevant. In addition, we also had 2 linguists (who were familiar with all of the research scenarios but did not research any of them) perform the same task for all of the collected questions and QUABs. Results from these three sets of annotations are found in Table 3.

User Type	Users	# Qs	# QUABs	# rel. QUABs	% relevant	ENIQ(P_1)
Expert	6	250	2500	699	27.96%	5.88
Novice	4	250	2500	953	38.12%	3.73
Linguists	2	500	5000	2240	44.80%	3.53

Table 3: User Comparison.

As expected, experts believed QUABs to be significantly ($p < 0.05$) less relevant than novices, who found approximately 38.12% of QUABs to be relevant to the original question submitted by a user. In contrast, the two linguists found 44.8% of the QUABs to be relevant. This number may be artificially high: since the linguists did not engage in actual Q/A dialogues for each of the scenarios they were annotating, they may not have been appropriately prepared to make a relevance assessment.

In the second set of experiments, we used the UIM in Figure 4 to train CRF-based relevance classifiers. We obtained training data for UIM_1 ("copy-and-paste"-based), UIM_2 ("click"-based), and UIM_3 ("hybrid") from 16 different dialogue histories collected from 8 different novice users. During these dialogues, users were asked to perform research as they normally would; no special instructions were given to users to provide additional relevance feedback to the system. After the dialogues were completed, QUABs that were copied from or clicked were annotated as "relevant" examples (according to each UIM); the remaining QUABs were annotated as "irrelevant". Once features (as described in Table 3) were extracted and the classifiers were trained, they were evaluated on a set of 1000 QUABs (500 "relevant", 500 "irrelevant") selected at random from the annotations performed in the first experiment. Table 4 presents results from these two classifiers.

UIM_1	P	R	F ($\beta = 1$)
Irrelevant	0.9523	0.9448	0.9485
Relevant	0.3137	0.3478	0.3299
UIM_2	P	R	F ($\beta = 1$)
Irrelevant	0.8520	0.8442	0.8788
Relevant	0.3214	0.4285	0.3673
UIM_3	P	R	F ($\beta = 1$)
Irrelevant	0.9384	0.9114	0.9247
Relevant	0.3751	0.3961	0.3853

Table 4: Experimental Results from 3 User Models.

Our results suggest that feedback gathered from a user's "normal" interactions with FERRET could be used to provide valuable input to a relevance classifier for QUABs When "copy-and-paste" events were used to train the classifier, the system detected instances of irrelevant QUABs with over 80% F. When the much more frequent "clicking" events were used to train the classifier, irrelevant QUABs were detected at over 90%F for both UIM_2 and UIM_3. In each of these three cases, however, detection of rel-

evant QUABs lagged behind significantly: relevant QUABs were detected at 42% F in UIM_1 at nearly 33% F under UIM_2 and at 39% under UIM_3.

We feel that these results suggest that the detection of relevant QUABs (or the filtering of irrelevant QUABs) may be feasible, even without requiring users to provide additional forms of explicit feedback to the system. While we acknowledge that training models on these types of events may not always provide reliable sources of training data – especially as users copy or click on QUAB passages that may not be relevant to their interests in the research scenario, we believe the initial performance of these suggests that accurate forms of relevance feedback can be gathered without the use of mixed-initiative clarification dialogues.

6 Conclusions

In this paper, we have presented a methodology that combines feedback that was gathered from users in conjunction with a CRF-based classifier in order to enhance the quality of suggestions returned to users of interactive Q/A systems. We have shown that the irrelevant QUAB suggestions can be identified at over 90% when systems combine information from a user's interaction with semantic and pragmatic features derived from the structure and coherence of an interactive Q/A dialogue.

7 Acknowledgments

This material is based upon work funded in whole or in part by the U.S. Government and any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U.S. Government.

References

Sanda Harabagiu, Andrew Hickl, John Lehmann, and Dan Moldovan. 2005a. Experiments with Interactive Question-Answering. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*.

S. Harabagiu, D. Moldovan, C. Clark, M. Bowden, A. Hickl, and P. Wang. 2005b. Employing Two Question Answering Systems in TREC 2005. In *Proceedings of the Fourteenth Text REtrieval Conference*.

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

Andrew Hickl, John Lehmann, John Williams, and Sanda Harabagiu. 2004. Experiments with Interactive Question-Answering in Complex Scenarios. In *Proceedings of the Workshop on the Pragmatics of Question Answering at HLT-NAACL 2004*.

T. Kristjansson, A. Culotta, P. Viola, and A. McCallum. 2004. Interactive information extraction with constrained conditional random fi elds. In *Proceedings of AAAI-2004*.

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

A. McCallum. 2003. Effi ciently inducing features of conditional random fi elds. In *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence (UAI03)*.

M. Palmer, D. Gildea, and P. Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. In *Computational Linguistics*, 31(1):71–106.

Jean Scholtz and Emile Morse. 2003. Using consumer demands to bridge the gap between software engineering and usability engineering. In *Software Process: Improvement and Practice*, 8(2):89–98.

F. Sha and F. Pereira. 2003. *Shallow parsing with conditional random fields*. In *Proceedings of HLT-NAACL-2003*.

Sharon Small and Tomek Strzalkowski. 2004. HITIQA: Towards analytical question answering. In *Proceedings of Coling 2004*.