The Use of Question Types to Match Questions in FAQFinder

Steven L. Lytinen and Noriko Tomuro
DePaul University
School of Computer Science, Telecommunications and Information Systems
243 S. Wabash Ave.
Chicago, IL 60604
{lytinen,tomuro}@cs.depaul.edu

Abstract

One useful way to find the answer to a question is to search a library of previously-answered questions. This is the idea behind FAQFinder, a Web-based natural language question-answering system which uses Frequently Asked Questions (FAQ) files to answer users' questions. FAQFinder tries to answer a user's question by retrieving a similar FAQ question, if one exists, and its answer. FAQFinder uses several metrics to judge the similarity of user and FAQ questions. In this paper, we discuss a metric based on question type, which we recently added to the system. We discuss the taxonomy of question types used, and present experimental results which indicate that the incorporation of question type information has substantially improved FAQFinder's performance.

Introduction

One useful way to find the answer to a question is to search a library of previously-answered questions. If a similar question can be found in the library, then its answer will likely serve as a good answer to the new question as well. Usenet Frequently Asked Questions (FAQ) files are built with this in mind, and as such they serve as a wide-ranging library of previously-answered questions.

This paper discusses FAQFinder (Burke *et al.* 1997; Lytinen, Tomuro, & Repede 2000), a Web-based natural language question-answering system which uses FAQ files to answer users' questions. Since FAQ files are written in question-and-answer format, FAQFinder tries to answer a user's question by retrieving the answer of a similar FAQ question, if one exists. FAQFinder uses a library of over 600 FAQ files, allowing it to answer questions about a broad range of subjects. FAQFinder can be found at `http://faqfinder.ics.uci.edu`.

Figures 1-3 show an example session with FAQFinder. After the user has typed a question (figure 1), FAQFinder matches it with a FAQ question in 2 stages. In the first stage, the system displays those FAQ files which are judged most likely to be relevant to the user's question (figure 2). The SMART information retrieval system (Salton 1971) is used to select these files. In the second stage, after the user chooses one of the FAQ files from this list, the individual questions from that file which are judged to be most similar to the user question are displayed, with their answers (figure 3). The second stage of processing uses a combination of 4 metrics to judge the similarity of user and FAQ questions. These metrics incorporate methods taken from both information retrieval and natural language processing.

In this paper, we focus on one of the 4 metrics used in question matching, which is based on *question type*. We recently incorporated the use of this metric into FAQFinder, and found that it substantially improves the system's ability to accurately judge question similarity. FAQFinder automatically classifies user and FAQ questions according to type using a distance-weighted k-nearest-neighbor (KNN) algorithm (Dudani 1976). The set of question types used by the system is shown in figure 4. The question type of the user question and a FAQ question determine the value of the metric through a *similarity matrix* which is defined for the question types.

In the rest of the paper, we first discuss the set of question types used in FAQFinder, and the use of KNN to classify questions by type. Then we discuss the incorporation of the question type metric into FAQFinder's question matching algorithm. Finally, we present results of empirical testing, which demonstrate the improved performance of the system with the addition of the question type metric.

Question types

Generally, questions which are paraphrases of each other can be answered in the same way. If one examines sets of paraphrases, one discovers that there are many ways to ask the same question. However, it is often the case that certain keywords (e.g. interrogatives such as "What", "When" and "How"), closed-class words, idioms, or syntactic constructions can be found which they share. Moreover, these cues seem just as important as the actual content words contained in the questions. For example, consider the following paraphrases:

How did the solar system form?
In what way was our solar system created?
How was the solar system created?
How did the solar system come into existence?
What happened in the Big Bang?

The interrogative "how" is used in many, but not all, of these paraphrases. The phrase "in what way" appears to be

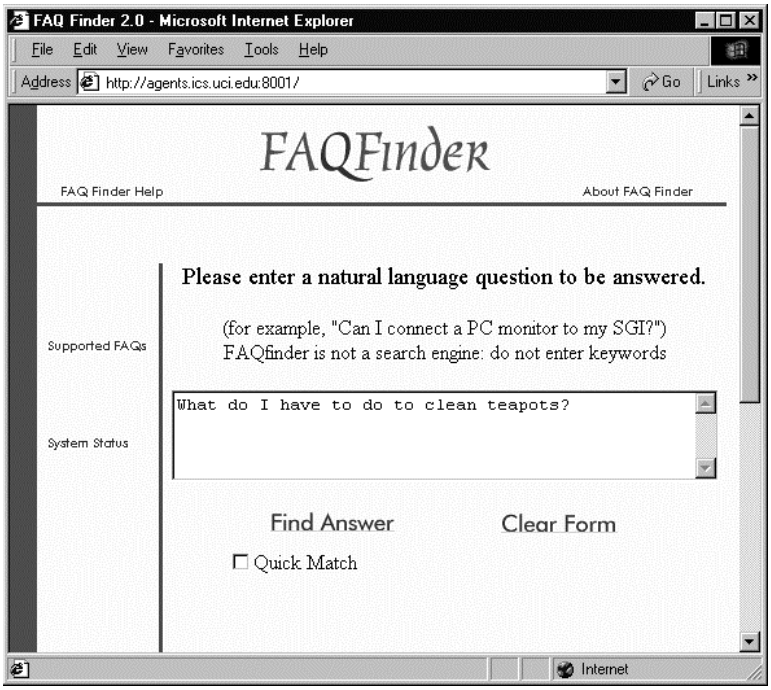

Figure 1: User question entered as a natural language query to FAQFinder

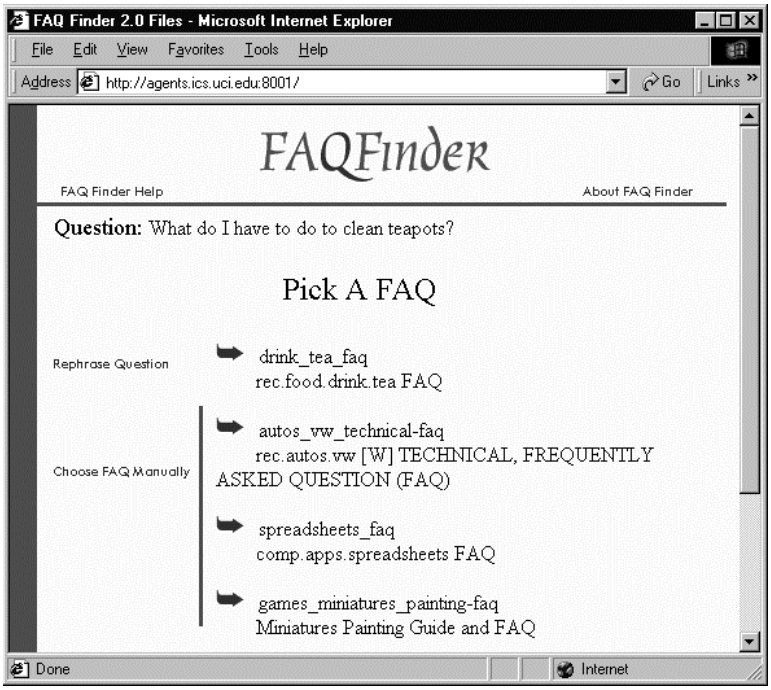

Figure 2: The 5 highest-ranked FAQ files

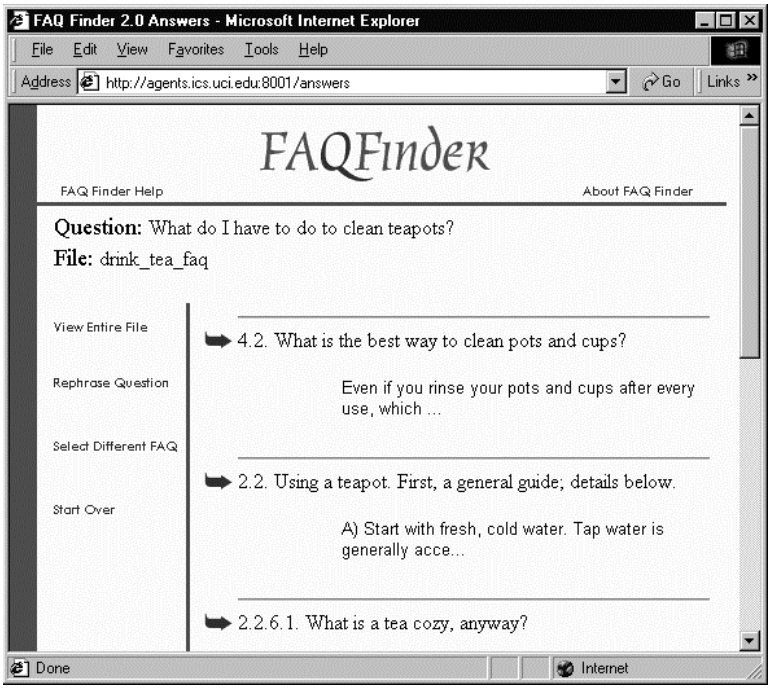

Figure 3: The 5 best-matching FAQ questions

synonymous with "how", and "what happened" likewise appears to be used in the same way.

On the other hand, varying some of the cue words can drastically change the meaning of the question, and therefore the way that it should be answered. Consider these variants of the above questions:

When did the solar system form?
In what galaxy was our solar system created?
What solar system are we in?

Intuitively, then, questions can be categorized into different types which roughly correspond to the keywords, idioms, or syntactic constructions discussed above. This intuition has been developed in previous work, in which several different taxonomies have been proposed for question types (e.g. (Lehnert 1978; Harabagiu *et al.* 2000)).

We used Lehnert's *conceptual question categories* as a starting point for developing a taxonomy of question types that would be of use to FAQFinder. In order to further refine this taxonomy, we picked 35 questions from 5 FAQ files[1] which we judged to represent a variety of question types, and we posted them on a Web site. Table 1 shows examples of the 35 sentences along with their question types. Visitors to the site were shown a random sampling of the 35 questions and were asked to rephrase them. After leaving the site on the Web for a period of 2 weeks, we had gathered a set of 679 example questions, or approximately 20 paraphrases for each of the 35 original questions. After examination of this sample data, we arrived at a set of 12 question types as displayed in figure 4.

1. `DEF (definition)`	7. `PRC (procedure)`
2. `REF (reference)`	8. `MNR (manner)`
3. `TME (time)`	9. `DEG (degree)`
4. `LOC (location)`	10. `ATR (atrans)`
5. `ENT (entity)`	11. `INT (interval)`
6. `RSN (reason)`	12. `YNQ (yes-no)`

Figure 4: Question types used in FAQFinder

Notice our question types are more general than those used in some of the systems which competed in the Text Retrieval Evaluation Conference (TREC) Question-Answering track (Voorhees 1999). Most sentences given in the TREC Q&A track are questions which ask for simple facts, can be answered by a single noun phrase, and would fall under our `REF`, `TME`, `LOC` and `ENT` categories. On the other hand, FAQFinder is a general Q&A system. Therefore we need a comprehensive set of question types which cover a general class of questions. Also notice that our categorization of questions is not lexically based in the sense that the type of a question can not be predicted reliably by simply looking at the first word.

Training for question type classification

The next task was to develop an algorithm for automatically classifying a question according to its type. We chose to use a distance-weighted k-nearest-neighbor (KNN) algorithm (Dudani 1976) for this task. The 679 examples col-

[1]The FAQ files were `astronomy`, `copyright`, `gasoline`, `mutual-funds` and `tea`.

Table 1: Examples of the original FAQ questions

Question Type	Question
DEF	"What does "reactivity" of emissions mean?"
REF	"What do mutual funds invest in?"
TME	"What dates are important when investing in mutual funds?"
ENT	"Who invented Octane Ratings?"
RSN	"Why does the Moon always show the same face to the Earth?"
PRC	"How can I get rid of a caffeine habit?"
MNR	"How did the solar system form?"
ATR	"Where can I get British tea in the United States?"
INT	"When will the sun die?"
YNQ	"Is the Moon moving away from the Earth?"

Table 2: No. of sentences in each FAQ Category

FAQ Category	No. of sentences
astro	210
copyright	111
gasoline	131
mutual-fund	77
tea	150
Total	679

lected from the Web, which we classified manually, served as the training set for the algorithm. The breakdown of the number of examples in each FAQ category is shown in Table 2.

In order to use KNN, we needed to select a set of words (i.e., features) to serve as the feature set for the classification task. We manually selected a set of 90 words from the example questions (containing 543 unique words), which we intuitively felt were most predictive of question type.[2] All words in this set were domain-independent, so that the same feature set could be used for any FAQ file. Most were closed-class words, including a mixture of interrogative words, modals, pronouns, prepositions. We also included domain-independent common nouns (e.g. "reason", "effect", "way"), verbs (e.g. "do", "have", "get", "find"), and adjectives (e.g. "long", "far"). Word order was not a factor in the system's classification of questions.[3]

Our implementation of KNN calculates *distance* between examples as a weighted sum of the difference between each feature value. During training, these weights are adjusted to optimize performance based on 5-fold cross-validation. The training algorithm examines the training set in multiple iterations, and after each iteration it incrementally adjusts the weight of each feature so as to minimize classification error. The training algorithm is similar to that used in (Lowe 1995).

After training, the error rate of question classification on the training set was approximately 23%. Considering the difficulty of 12-way classification, this is a positive result. We also tested the domain independence of the features and feature weighting by classifying examples from a randomly selected of question from other FAQ files. On this more general test, the error rate of question classification was about 30%.

[2]Words were stemmed using the WordNet morphing function, as we describe in the next section.

[3]Discussions on various feature selection schemes, including some automatic methods, on our FAQ data are found in (Tomuro & Lytinen 2001).

Computing question similarity

FAQFinder rates the similarity of a FAQ question to a user question by using a combination of four metrics: term vector similarity, coverage, semantic similarity, and question type similarity. Each metric is normalized to produce a value of between 0 and 1, where 1 indicates the strongest similarity. In the current version of FAQFinder, overall similarity is computed by averaging the 4 metrics together. In previous work, we used a weighted sum of the 4 metrics, where the weights sum up to 1 and their distribution was derived from training. However, in our current experiment, preliminary testing showed that training did not significantly improve the performance of the system, so we used a uniform distribution, by assigning each metric a weight of .25.

Preprocessing

In order to compute the 4 metrics, FAQFinder preprocesses each FAQ question by assigning each word a part-of-speech category using the Brill tagger (Brill 1995), and stemming it to a base form by using the WordNet morphing function. Then for each question, FAQFinder stores the results as a term vector and a question type.

A term vector is a vector of weights for the terms (i.e., stemmed words) in a question. A weight for a term is obtained by *tfidf* (Salton & McGill 1983), a measure often used in Information Retrieval (IR), which specifies the weight w_i for each term t_i to be:

$$w_i = (1 + log(tf_i)) \frac{logN}{df_i}$$

Here, a "document" is a single question; thus, N is the number of questions in the example set (679), df_i is the number of questions in which t_i appears, and tf_i is the number of times t_i appears in the question (usually 1). Note that tfidf is applied to terms in a question after some closed-class terms are discarded using a stop list, as is standard in IR.

In our current work, question type of a FAQ question is assigned manually, using the data we prepared for the training of the KNN algorithm (described in the last section).

Question similarity

On-line processing proceeds as follows: for the user question entered, words in the question are first converted to base forms in the same manner as the FAQ questions, a term vector and a question type. To compute tfidf, the user question is considered to be one of the "documents"; thus N in the above equation is increased by 1, and all FAQ term vectors are adjusted to reflect the addition of the user question. The question type is obtained by running the KNN classifier (which uses the same feature set as the training) on the question.

Next, the user question is compared with each FAQ question, and the four similarity metrics are computed. The first metric, term vector similarity, is computed as follows. Let $v_u = \langle w_{u1}, w_{u2}, ..., w_{un} \rangle$ be the term vector representing the user question (after stop-list is applied), and let $v_f = \langle w_{f1}, w_{f2}, ..., w_{fn} \rangle$ be the term vector representing a FAQ question. Term vector similarity is computed using the cosine measure:

$$cos(v_u, v_f) = \frac{\sum w_{ui} w_{fi}}{\sqrt{\sum w_{ui}^2} \sqrt{\sum w_{fi}^2}}$$

The second metric, coverage, is the percentage of user question terms that appear in the FAQ question. It is obtained by finding the intersection of the (stemmed and stop list-filtered) terms in the term vectors of the two questions.

The third metric, semantic similarity, is calculated using WordNet (Miller 1990), and involves finding the minimum path length between WordNet concepts (called *synonym sets* or *synsets*) referred to by terms in the user and FAQ questions. The minimum distance between synsets is calculated for pairs of terms, one term from the user question and the other from the FAQ question. In general, $\delta(t_1, t_2)$, the semantic distance between two terms t_1 and t_2, each of which has n and m WordNet senses[4] $S1 = \{s_1, .., s_n\}$ and $S2 = \{r_1, .., r_m\}$, is the minimum of all possible pair-wise semantic distances between S_1 and S_2, that is,

$$\delta(t_1, t_2) = \min_{si \in S1, rj \in S2} D(s_i, r_j)$$

where $D(s_i, r_j)$ is a path length between WordNet synsets s_i and r_j. For example, $\delta(bug, termite)$ is 2, because there is a hypernym (is-a) link between "bug" (noun sense 1) and "insect" (noun sense 1), and a hyponym (inverse is-a) link between "insect" (noun sense 1) and "termite" (noun sense 1). If there is no path between any of the synsets of t_1 and t_2, then $\delta(t_1, t_2) = \infty$.

Then, the semantic similarity between the user question $T_u = \{u_1, .., u_n\}$ and a FAQ question $T_f = \{f_1, .., f_m\}$ is defined as follows:

$$sem(T_u, T_f) = \frac{I(T_u, T_f) + I(T_f, T_u)}{|T_u| + |T_f|}$$

where

$$I(T_x, T_y) = \sum_{x \in T_x} \frac{1}{1 + \min_{y \in T_y} \delta(x, y)}$$

and $|T_x|, |T_y|$ denote the size of T_x and T_y. Thus, $sem(T_x, T_y)$ is essentially a metric which is the normalized sum of the inverse of pair-wise semantic distances between all words in T_x and T_y measured from both directions.

Finally, the fourth metric, question type similarity, is computed by comparing the question type of user and FAQ questions. The similarity value between two types is defined in a *similarity matrix* shown in figure 5.[5] The similarity matrix essentially reflects the degree of closeness between question types, and is incorporated in the system in order not to overly penalize classification errors between very close question types. For instance, types PRC (procedure) and MNR (manner) include mainly 'how' questions, and questions of those types are sometimes quite difficult to distinguish. An example question would be "How can I deal with cursed items?". This question should be a MNR question, since the answer would be descriptive rather than procedural. But this question also allows all paraphrasing patterns of the PRC questions, such as "What is the best way to deal with cursed items?", although some of the PRC paraphrasing patterns do not usually apply to MNR questions (e.g. "What was the best way for the solar system to form?").

Another purpose of the similarity matrix is to match two questions of different types for which the answer to one question type is often included in the answer to the other type. For instance, the answer to a user question "Which credit reporting agencies can I write to to get free credit reports?" is included in the answer to the FAQ question "Can I get a free copy of my own credit report?".[6] In general, practically any question can be reformulated as a yes/no question to ask the question in a more general way; therefore YNQ is given a non-zero value (.2) to all other question categories. Similarly, many questions types can be reformulated as a REF (reference) question, which asks for the referent of the interrogative "what" or "which". For example:

How did the solar system begin? (MNR)
What event started the solar system? (REF)

Other values in the matrix reflect our judgments of how similar each question type is to other types.

Evaluation of FAQFinder

To test the effect of question type similarity on the overall performance of the system, we ran the system on the set of 679 paraphrases gathered from the Web, first using all 4 similarity metrics, and again omitting the question type similarity metric. Since the 679 paraphrases were all generated from the original 35 FAQ questions from 5 of the FAQ files, we determined that the correct match for each paraphrase

[4]FAQFinder also uses a WordNet sense tagging algorithm to restrict the possible senses or a word which are used in this calculation; see (Lytinen, Tomuro, & Repede 2000) for details.

[5]Since the similarity matrix is symmetric, only the lower left half of the matrix is shown.

[6]While the latter question is a YNQ (yes/no question), these questions are often indirect requests for other types of information, and a good answer is rarely a simple yes/no response (Searle 1975).

	YNQ	DEG	TME	LOC	ENT	PRC	MNR	RSN	REF	DEF	INT	ATR
YNQ	1											
DEG	.2	1										
TME	.2	0	1									
LOC	.2	0	0	1								
ENT	.2	0	0	0	1							
PRC	.2	0	0	0	0	1						
MNR	.2	0	0	0	0	.5	1					
RSN	.2	0	0	0	0	0	.5	1				
REF	.2	.1	.1	.1	.1	.1	.1	.1	1			
DEF	.2	0	0	0	0	0	0	0	.5	1		
INT	.2	.6	.6	.6	0	0	0	0	.1	0	1	
ATR	.2	0	0	.6	.6	.6	0	0	.1	0	0	1

Figure 5: Similarity matrix for the 12 question types

was the original FAQ question that it was generated from. Recall improved slightly with the use of question type, from 91% without the use of the question type metric to 94% with inclusion of this metric.

As a note, in our previous work (Lytinen, Tomuro, & Repede 2000), we reported that FAQFinder achieved recall performance of approximately 60-65% using the original 3 metrics. The difference in performance between the previous and current work is due to the nature of the test questions used in this experiment; many of them (paraphrases entered by the visitors of our Web site) were reformulations of the original FAQ questions using passivization, denominalization (e.g. "Who is the owner of X?" to "Who owns X?") and change of idiomatic expressions (e.g. "Why did X happen?" to "How come X happened?"), but still used the same words in the original questions. Therefore, it was easier for FAQFinder to match them than previous test data, which was taken from FAQFinder user logs. In the user logs, many questions were either very short or did not have a large number of overlapping words with the FAQ questions.

Despite the differences in test data, the current experiment is still informative, because we intentionally selected each of the 35 FAQ questions such that there were other questions (of different type) in the same FAQ file that used the same words. Thus, previous versions of FAQFinder would have given high scores to those near-misses. Therefore, the results of the current work are useful in comparing performance of FAQFinder with and without the use of the question type metric.

To further explore the effect of question type on system performance, we examined the trade-off between recall and *rejection*, with and without the use of question type. Rejection is a metric similar to precision, but one which we feel better represents performance in FAQFinder's task than precision. It is defined as the percentage of *unanswerable* user questions (i.e., questions for which there is no equivalent FAQ question) for which FAQFinder displays no matches to the user.[7] While FAQFinder generally displays up to 5 questions from a FAQ file, a question is only displayed if its similarity metric exceeds a threshold. If no FAQ question's similarity metric exceeds the threshold, then no questions are displayed. Thus, we can measure the trade-off between recall and rejection by adjusting the threshold for displaying a question.

Since our test set of paraphrases did not include any unanswerable questions, we judged system performance on unanswerable questions by running the system on the same set of 679 test questions with the best-matching FAQ question removed. Figure 6 shows FAQFinder's performance on the test set with and without the use of the question type similarity metric. Notice that FAQFinder with question type retains recall as the threshold (and therefore rejection) is increased. When rejection rates are between 30-80%, recall is 10-15% higher when question type is used than when it is not. Indeed, the paired t-test yielded the p-value $< .001$, indicating the increase in the recall values was statistically significant.[8]

To examine further the effects of each of the 4 metrics on FAQFinder's performance, we conducted an ablation study, in which we ran the system using only one metric at a time. Figure 7 shows the results. While the term vector, coverage, and semantic similarity metrics all exhibit a trade-off between recall and rejection, it is interesting to note that the question type metric does not; the recall performance of this metric is approximately the same at 10% rejection as it is at 100%. This suggests that question type is a better metric for eliminating candidate matches between user and FAQ questions than for identifying good matches. Intuitively, this makes sense; two questions of the same type may be highly dissimilar if their content words are not related, but two questions with similar content words may still not be good matches if they are of different question types. Thus, the question type metric nicely complements the other 3 metrics used by the system, by measuring similarity between questions along another dimension not well-captured by term vector, coverage, or semantic similarity.

[7] Rejection is a better measure of FAQFinder's response to unanswerable questions than precision: although precision is affected by system performance on unanswerable questions, performance on these questions may be overshadowed by the fact that precision also penalizes the system for displaying 5 matches for answerable questions instead of just the single best match.

[8] Here, we used the upper-tailed test, with the null hypothesis that the mean of the differences by the use of question type was 0, versus the alternative hypothesis that the mean was greater than 0.

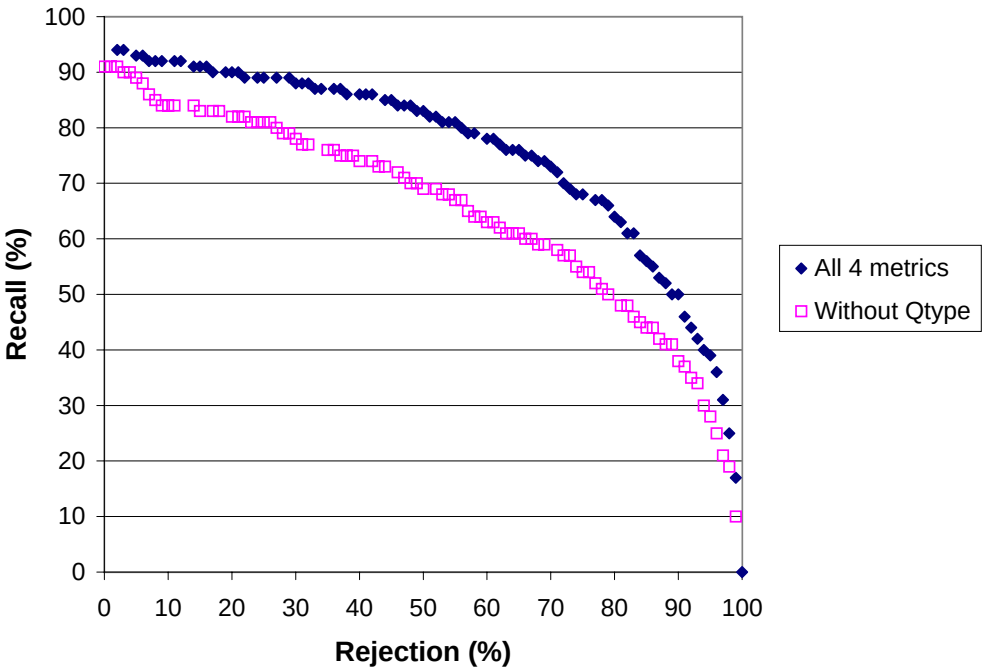

Figure 6: Recall vs. Rejection for FAQFinder with and without the use of question types

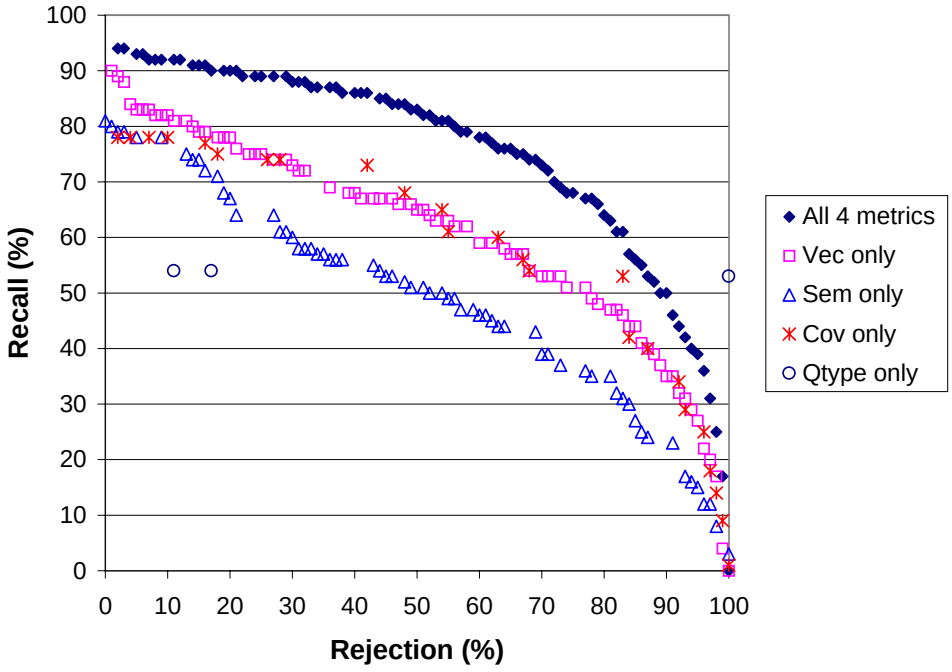

Figure 7: Ablation study

Related Work

In recent years, question types have been used in several Question-Answering systems. Among them, systems which competed in the TREC-8 and 9 Q&A track used question types to identify the kind of entity being asked. Due to the nature of the task (which is to extract a short, specific answer to the question), their categories were strongly tied to the *answer types*, such as `PERSON`, `MONEY` and `PERCENTAGE`. The type of a question is typically identified by first applying a pattern-based phrase extraction algorithm or parsing to identify phrases, and then looking at either the interrogative word or the semantic class of the head noun (Abney, Collins, & Singhal 2000; Cardie *et al.* 2000; Harabagiu *et al.* 2000). Some systems apply (hand-built) question patterns or *templates* (Hovy *et al.* 2001; Hermjakob 2001). In our work, we deal with general questions from broad domains, thus the identification of question types is much more difficult.

As for paraphrasing questions, AskJeeves (`http://www.askjeeves.com`) utilizes question templates to transform user questions into more specific ones (for more accurate answer retrieval). For instance, a question "How can I find out about sailing?" is matched with a template "Where can I find a directory of information related to X?", and X is instantiated with a list of choices (in this case, "boat" as the first choice). However, their templates are predefined and the coverage is limited, thus the system quite often retrieves incorrect templates. For example, a user question "How can I get tickets for the Indy 500?" is matched with a template "Who won the Indy 500 in X (1991)?".

Among the TREC Q&A systems, (Harabagiu *et al.* 2000) applies reformulation rules to a question, and expands the open-class words in the question by their synonyms and hypernyms using WordNet. Their result indicates improved answer retrieval performance by categorizing questions.

Conclusions and Future Work

We have shown that automatic classification of question type can be used to improve performance of the FAQFinder system. For future work, we plan to investigate automatic feature selection schemes for identifying question types. Currently FAQFinder uses a manually selected set of features. Although the results we obtained in this work are very encouraging, we must also consider how to derive a feature set which is scalable to a wide range of domains. Our preliminary results show that this task is quite challenging (Tomuro & Lytinen 2001). To improve upon those results, we are planning to incorporate semantic information, by using a general lexical resource such as WordNet. The use of (abstract) semantic classes has two major advantages: first, it reduces the number of features in the feature set; and second, it can make the feature set flexible to unseen examples from different domains. Also, we can obtain the semantic classes of words with no extra cost, since they are already computed in the calculation of the semantic similarity. We believe the semantics of the words will greatly assist in question classification for general question-answering systems.

References

Abney, S.; Collins, M.; and Singhal, A. 2000. Answer extraction. In *Proceedings of the 6th Applied Natural Language Processing (ANLP-2000)*.

Brill, E. 1995. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics* 21(4).

Burke, R.; Hammond, K.; Kulyukin, V.; Lytinen, S.; Tomuro, N.; and Schoenberg, S. 1997. Question answering from frequently asked question files: Experiences with the faqfinder system. *AI Magazine* 18(2).

Cardie, C.; Ng, V.; Pierce, D.; and Buckley, C. 2000. Examining the role of statistical and linguistic knowledge sources in a general-knowledge question-answering system. In *Proceedings of the 6th Applied Natural Language Processing (ANLP-2000)*.

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

Harabagiu, S.; Moldovan, D.; Pasca, M.; Mihalcea, R.; Surdeanu, M.; Bunescu, R.; Girju, R.; Rus, V.; and Morarescu, P. 2000. Falcon: Boosting knowledge for answer engines. In *Proceedings of TREC-9*.

Hermjakob, U. 2001. Parsing and question classification for question answering. In *Proceedings of the Workshop on Open Domain Question Answering at the ACL-01*.

Hovy, E.; Gerber, L.; Hermjakob, U.; Lin, C.; and Ravichandran, D. 2001. Toward semantics-based answer pinpointing. In *Proceedings of the DARPA Human Language Technologies (HLT)*.

Lehnert, W. 1978. *The Process of Question Answering*. Lawrence Erlbaum Associates.

Lowe, D. 1995. Similarity metric learning for a variable-kernel classifier. *Neural Computation* SMC-7(1):72–85.

Lytinen, S.; Tomuro, N.; and Repede, T. 2000. The use of wordnet sense tagging in faqfinder. In *Proceedings of the workshop on Artificial Intelligence for Web Search at AAAI-2000*.

Miller, G. 1990. Wordnet: An online lexical database. *International Journal of Lexicography* 3(4).

Salton, G., and McGill, M. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill.

Salton, G. 1971. *The SMART Retrieval System: Experiments in Automatic Document Processing*. Prentice-Hall.

Searle, J. 1975. Indirect speech acts. In Cole, P., and Morgan, J., eds., *Syntax and Semantics 3, Speech Acts*. Academic Press.

Tomuro, N., and Lytinen, S. 2001. Selecting features for paraphrasing question sentences. In *Proceedings of the workshop on Automatic Paraphrasing at NLP Pacific Rim 2001 (NLPRS-2001)*.

Voorhees, E. 1999. The trec-8 question answering track report. In *Proceedings of TREC-8*.