Selecting Features for Paraphrasing Question Sentences

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

Abstract

In this paper, we investigate several schemes for selecting features which are useful for automatically classifying questions by their question type. We represent questions as a set of features, and compare the performance of the C5.0 machine learning algorithm using the different representations. Experimental results show a high accuracy rate in categorizing question types using a scheme based on NLP techniques as compared to a scheme based on IR techniques. The ultimate goal of this research is to use question type classification in order to help identify whether or not two questions are paraphrases of each other. We hypothesize that the identification of features which help identify question type will be useful in the generation of question paraphrases as well.

1 Introduction

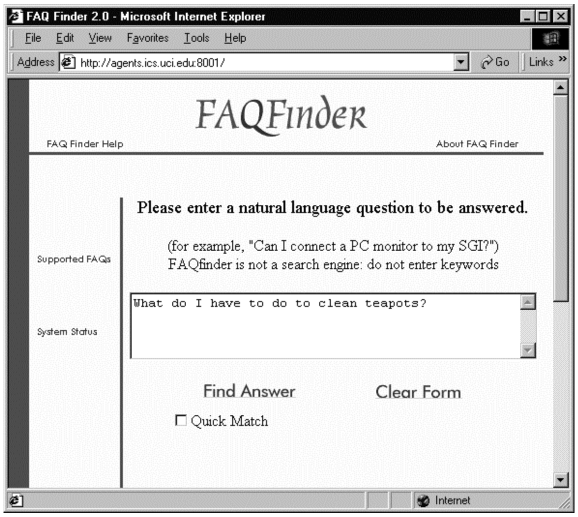
In recent years, techniques for paraphrasing have received much attention in Natural Language Processing (NLP), particularly in the area of text summarization and NL generation (e.g. (McKeown et al., 1999)). In simple sentence paraphrasing (without reducing the sentence length), restating a declarative sentence into another sentence can be done by applying some general transformation patterns at the surface level. Those transformation patterns include lexical substitution by synonyms at word level, passivization, verb alternations (Levin, 1993), and denominalization at sentence level. On the other hand, paraphrasing a question is more difficult than a declarative sentence, because interrogative words (e.g. "how" in the question "How do I clean teapots?") carry a meaning of their own, which is subject to paraphrasing in addition to the rest of the sentence ("(do) I clean teapots"). Moreover, paraphrasing the question part sometimes results in significant changes in the structure and words used in the original question. For example, "How can I clean teapots?" can be paraphrased as (among others):

- "In what way can we clean teapots?"
- "What do I have to do to clean teapots?"
- "What is the best way to clean teapots?"
- "What method is used for cleaning teapots?"
- "How do I go about cleaning teapots?"

Thus, with an additional element, questions require more flexible and complex paraphrasing patterns.

There are several interesting characteristics in the paraphrasing patterns of questions. First, they involve *non-content words*, consisting of many closed class words and some open class words. Second, those patterns seem to hold across paraphrases of the same *question type*. Third, there are some known, almost idiosyncratic patterns (e.g. "in what way", "what should I do to"), but there are also infinitely many others without fixed expressions.

This paper investigates several schemes for selecting features from questions in order to classify them by question type. The ultimate goal of this research is to improve our ability to identify whether or not two questions are paraphrases of each other. We examined three feature selection schemes: one based on *Gain Ratio* (Quinlan, 1994), an information-theoretic metric often used in Text Categorization; another based on words that appeared in particular kinds of phrases in a sentence; and finally one based on manual selection. In our experiment, we chose 35 questions of various question types from Usenet Frequently Asked Questions (FAQs), and collected paraphrases of those questions from a wide audience. Then we represented those paraphrases using the three sets of features, and compared their classification errors made by C5.0 (Quinlan, 1994), a decision tree classification system. The results we obtained showed a high accuracy rate in categorizing question types using a scheme based


FAQ Finder 2.0 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://agents.ics.uci.edu:8001/
Go
Links
FAQFinder
FAQ Finder Help
About FAQ Finder
Supported FAQs
System Status
Find Answer
Clear Form
Internet

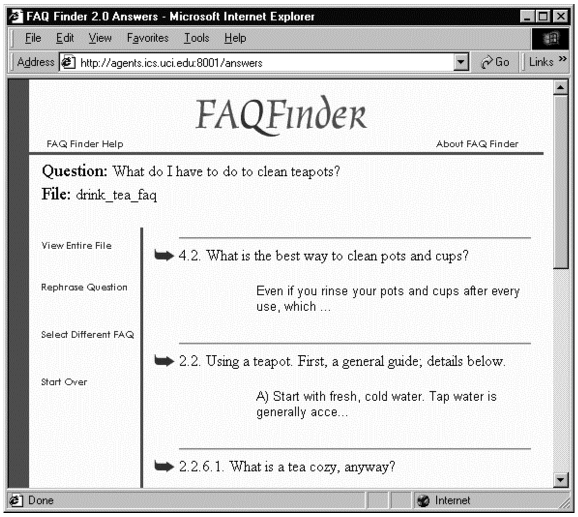
Please enter a natural language question to be answered.

(for example, "Can I connect a PC monitor to my SGI?")
FAQfinder is not a search engine: do not enter keywords

What do I have to do to clean teapots?

☐ Quick Match

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


FAQ Finder 2.0 Answers - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://agents.ics.uci.edu:8001/answers
Go
Links
FAQFinder
FAQ Finder Help
About FAQ Finder
View Entire File
Rephrase Question
Select Different FAQ
Start Over
Done
Internet

Question: What do I have to do to clean teapots?
File: drink_tea_faq

4.2. What is the best way to clean pots and cups?

Even if you rinse your pots and cups after every use, which ...

2.2. Using a teapot. First, a general guide; details below.

A) Start with fresh, cold water. Tap water is generally acce...

2.2.6.1. What is a tea cozy, anyway?

Figure 2: The 5 best-matching FAQ questions

on NLP techniques as compared to a scheme based on IR techniques.

Although our work here is essentially to derive features for recognizing (rather than generating) paraphrases of question sentences, selecting the appropriate features for recognition is itself a difficult task. Investigation of such features or words can discover where the meaning of the question part is in a given sentence, and will help us develop *transformation patterns* for automatic paraphrasing of question sentences.

Motivation behind the work we present here is to improve the retrieval accuracy of our system called FAQFinder (Burke et al., 1997; Lytinen, Tomuro, and Repede, 2000). FAQFinder is a web-based, natural language question-answering system which uses Usenet FAQ files to answer users' questions. Figures 1 and 2 show FAQFinder's I/O behavior. First, the user enters a question in natural language. The system then searches the FAQ files for questions that are similar to the user's. Based on the results of the search, FAQFinder displays 5 FAQ questions which are ranked the highest by the system's similarity measure. Thus, FAQFinder's task is to identify FAQ questions which are the best paraphrases of the user's question.

To measure the similarity of the two questions, FAQFinder currently uses a combination of Information Retrieval (IR) techniques (*tfidf* and cosine (Salton and McGill, 1983)) and linguistic/semantic knowledge (WordNet (Miller, 1990)). We are planning to add question type in the similarity measure and see how much it helps increase the recall and precision of the retrieval performance. Our work on question paraphrases here is the first step in this direction.

2 Question Types

In this work, we defined 12 question types below.

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

Although those types do not cover all possible questions completely, they do seem to cover the majority of questions entered in FAQFinder by the users. Descriptive definitions and examples of each type are found in the Appendix at the end of this paper.

Our question types are intended to cover a wide variety of questions. For that purpose, our 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 trivial pursuit type questions which ask for simple facts, 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 more general class of questions.

Our question types are determined based on the paraphrasing patterns. For instance, types `PRC` and `MNR` both include 'how' questions, such as "How should I store beer?" (PRC) and "How did the solar system form?" (MNR). Even the meanings of "how" in these sentences are the same: "In what manner or way" (Webster's Collegiate Dictionary, sense 1 of "how"). However, some of the paraphrasing patterns for PRC questions do not apply to MNR questions. For example,

* "What did the solar system have to do to form?"
* "What was the best way for the solar system to form?"

Also, we defined a type `ATR` (for ATRANS in Conceptual Dependency (Schank, 1973)) as a special case of `PRC`. An example question of this type would be "How can I get tickets for the Indy 500?". Not only do ATR questions undergo the paraphrasing patterns of PRC questions, they also allow rephrasings which ask for the (source) location or entity of the thing(s) being sought, for instance, "Where can I get tickets for the Indy 500?" and "Who sells tickets for the Indy 500?". Those ATR paraphrases in fact occurred very frequently in the FAQ files as well as in the FAQFinder user logs. Thus, we determined that ATR questions constitute an important question type for FAQFinder.

Notice that our categorization of questions is not lexically based in the sense that the type of a question cannot be predicted reliably by simply looking at the first word. It seems that even the notion of *question phrase* as a linguistic unit is sometimes difficult to specify, particularly for the types `MNR` and `INT`.

3 Features Selection Schemes

In our experiment, we selected a total of 35 questions from 5 FAQ categories/domains: `astronomy`, `copyright`, `gasoline`, `mutual-fund` and `tea`. Table 1 shows some of those sentences along with their question types.

To obtain paraphrases, we created a web site where users could enter paraphrases for any of the 35 questions. The site was made public for two weeks, and a total of 1000 paraphrases were entered. Then we inspected each entry and discarded ill-formed ones (such as keywords or boolean queries) and incorrect paraphrases. This process left us with 714 correct paraphrases. These examples constitute the base dataset for our experiments. The breakdown of the number of examples in each FAQ category is shown in Table 2.

Then, the example sentences were preprocessed by assigning each word a part-of-speech category using the Brill tagger (Brill, 1995), and stemming it to a base form.

Table 2: No of sentences in each FAQ Category

FAQ Category	No. of sentences
`astro`	215
`copyright`	117
`gasoline`	140
`mutual-fund`	84
`tea`	158
Total	714

In our current work, features were taken from the (stemmed) words in the example questions, which consisted of 543 unique words. We examined three feature selection schemes: (1) by using *Gain Ratio* (Quinlan, 1994); (2) by choosing words that appeared in some particular kinds of phrases; and (3) by manual selection. The first two schemes are automatic methods. Gain Ratio is an information-theoretic metric which has been frequently used in Text Categorization tasks, thus the first scheme essentially represents an (IR) approach. The second scheme analyzes the structure of each question and focuses on words in specific phrases, thus it represents an NLP approach. By comparing the performance of the three schemes, we will be able to see if NLP techniques have advantages over bag-of-words IR techniques, as well as any potential issues and difficulties in applying automatic techniques to question type identification.

3.1 Scheme (1): Gain Ratio

Gain Ratio (GR) is a metric often used in classification systems (notably in the C4.5 decision-tree classifier (Quinlan, 1994)) for measuring how well a feature predicts the categories of the examples. GR is a normalized version of another metric called *Information Gain* (IG), which measures the informativeness of a feature by the number of bits required to encode the examples if they are partitioned into two sets, based on the presence or absence of the feature.[1]

Let C denote the set of categories $c_1, .., c_m$ for which the examples are classified (i.e., target categories). Given a collection of examples S, the Gain Ratio of a feature A, $GR(S, A)$, is defined as:

$$GR(S, A) = \frac{IG(S, A)}{SI(S, A)}$$

where $IG(S, A)$ is the Information Gain defined

[1] The description of Information Gain here is for binary partitioning. Information Gain can also be generalized to m-way partitioning, for all $m >= 2$.

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?"

to be:

$$\begin{aligned} IG(S,A) = & -\sum_{i=1}^{m} Pr(c_i)\ log_2 Pr(c_i) \\ & + Pr(A) \sum_{i=1}^{m} Pr(c_i|A)\ log_2 Pr(c_i|A) \\ & + Pr(\overline{A}) \sum_{i=1}^{m} Pr(c_i|\overline{A})\ log_2 Pr(c_i|\overline{A}) \end{aligned}$$

and $SI(S,A)$ is the *Splitting Information* defined to be:

$$SI(S,A) = -Pr(A)\ log_2 Pr(A) - Pr(\overline{A})\ log_2 Pr(\overline{A})$$

Then, features which yield high GR values are good predictors. In previous work in text categorization, GR (or IG) has been shown to be one of the most effective methods for reducing dimensions (i.e., words to represent each text) (Yang and Pedersen, 1997; Spitters, 2000).

However, in applying GR to our problem of question type classification, there was an important issue to consider: how to distinguish content words from non-content words. This issue arose from the uneven distribution of the question types among the five FAQ domains chosen. Since not all question types were represented in every domain, if we chose the question types as the target categories, features which yield high GR values might include some domain-specific words. For example, the word "sun" yielded a high score for predicting the question type (INT), because it only appeared in the questions of that type. Such a content word would not make a good predictor when the classifier was applied to other domains. In effect, good predictors for our purpose are words which predict question types very well, but do not predict domains (i.e., non-content words). Therefore, we defined the GR score of a word to be the combination of two values: the GR value if the target categories were question types, minus the GR value if the target categories were domains.

The modified GR measure was applied to all 543 words in the example questions, and the top 270 words were selected as the feature set.

3.2 Scheme (2): Phrases

For the second scheme, we first applied a pattern-based phrase extraction algorithm to each question sentence, and extracted three phrases: WH phrase (WHP), subject noun phrase (NP) and main verb (V). A WHP consisted of all words from the beginning of the sentence up to and including the auxiliary, and NP and V were taken in the usual way. For example, in the question "How can I clean teapots?", extracted phrases were "How can" (WHP), "I" (NP), and "clean" (V). In the current work, object NP's were not considered in the extraction patterns, since most words in object nouns seemed to be content words. A total of 279 unique words were selected by this scheme.

3.3 Scheme (3): Manual Selection

For the third scheme, we manually selected a set of 90 words which seemed to predict question types. All words in this set were non-content words, and they were a mixture of closed-class words including interrogatives, modals and pronouns; and domain-indepenent words including common nouns (e.g. "reason", "effect", "way"), verbs (e.g. "do", "have", "get", "find"), adjectives (e.g. "long", "far"), and prepositions (e.g. "in", "for", "at").

4 Results

4.1 Training Set

To compare the different feature selection schemes, we created a dataset for each scheme by representing each example sentence in the 714 examples by a vector of length n, where n is the size of the respective feature set used. Values in a vector were binary (0 or 1), indicating the presence or absence of the feature/word.

To test the classification accuracy, we used a decision-tree supervised learning algorithm called C5.0 (the commercial version of C4.5, available at `http://www.rulequest.com`) on each dataset.[2] Ta-

[2] In our preliminary experiment, we also used a k-nearest neighbor (KNN) algorithm (Cost and Salzberg, 1993). The results we obtained from the two algorithms were very similar, thus we only used

Table 3: Classification error rates on the first question dataset

Scheme	#words	Error
`Gain Ratio`	270	12.9
`Phrases`	279	8.0
`Manual`	90	17.5
All	543	7.4

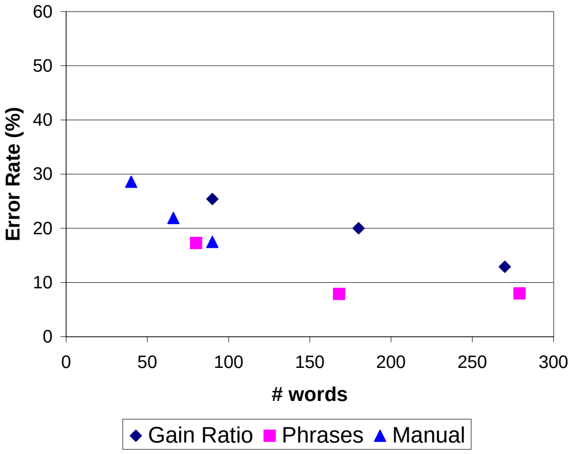


Figure 3: Error rates by various feature set sizes on the first question dataset

ble 3 shows the results. Each dataset was run using 5-fold cross-validation. The figures given are average error rates of 5 runs. As you can see, the overall best performance was achieved using the `Phrases` scheme, which yielded a 8.0% error rate. This figure is extremely low (indicating a very high classification accuracy), considering the baseline error rate for 12-way classification (by pure chance) would be 91.7% (≈ 11/12). As a note, we also ran a separate test using all 543 features, and the error rate was 7.4%. This means that words selected by the `Phrases` scheme achieved a comparable accuracy by using only half of the words. This result indicated that the NLP techniques had strong advantages over the IR techniques for question type identification.

In order to further examine the three schemes, for each scheme, we gradually reduced the size of the feature set, and observed how the error rates degraded. Features in the `Gain Ratio` scheme were divided into 3 subsets by taking the top 270 (the original set), 180 and 90 features according to the GR scores. Features in the `Phrases` scheme were also divided into 3 subsets by decreasing the *scope* of the phrases, from WHP+NP+V (the original set, 279 words) to WHP+NP (168 words) to WHP only (80 words). For features in the `Manual` scheme, we restricted the set in a similar way: from the original 90 words, we created the first subset (66 words) by removing verbs, and then the second subset (40 words) by removing nouns and adjectives from the first subset.

Figure 3 shows the result. As you can see, the error rates of the `Phrases` were the same for WHP+NP+V and WHP+NP. This means that the feature set could be further reduced to 168 words, and would still achieve the same, very low error rate.[3] This would make a 70% reduction from the original 543 unique words (168/543 ≈ .3). On the other hand, the error rates of the `Gain Ratio` were consistently higher than those of `Phrases`, giving a further support for the effectiveness of the NLP techniques over the IR techniques. As for the `Manual` scheme, the error rate by the full 90 word set was comparable to the error by the 80 word set (WHP) of the `Phrases` scheme, indicating that manual feature selection was no worse than the NLP techniques.

4.2 Test Sets

To see how well the selected feature sets would apply to other questions and domains, we also tested 2 additional sets of questions:

1. **tq1** – 160 additional questions from the same FAQ files as the original 35 questions.

2. **tq2** – 620 questions from other domains typed by FAQFinder users, taken from the FAQFinder server logs.

For each new test set, we constructed a C5.0 decision tree using the original dataset (of 714 questions) for each of the 3 selection schemes with varying feature set sizes, and measured their classification error rates on the new test sets.

Figure 4 and 5 show the results for the **tq1** and **tq2** respectively. As you see, error rates on both datasets were much higher than those on the first dataset for the `Gain Ratio` and `Phrases` schemes: around 50% on tq1 and 40% on tq2. Also for those schemes, error rates did not decrease even after more features were considered. This indicates that the automatic selection methods, based on IR or NLP techniques, were not successful in identifying important non-content words in the training set. Indeed, by inspecting the decision rules induced by C5.0, we discovered that words used in

C5.0 in this work.

[3] Note that, although this result seems to imply the verbs in the `Phrases` feature set did not contribute to the overall performance, we ran a separate test with words in WHP+V, and confirmed that the verbs did help decrease the error rate as well.

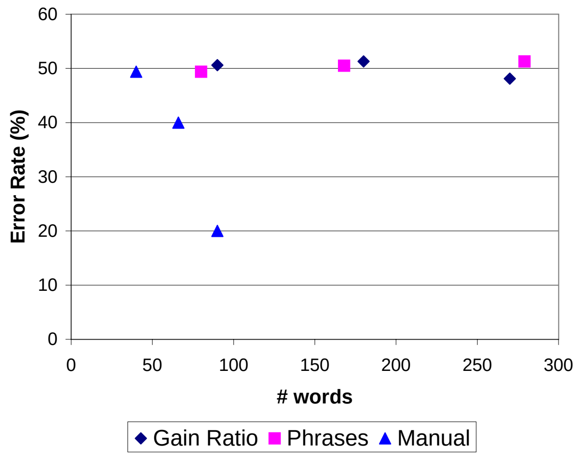


Figure 4: Error rates on the **tq1** testset

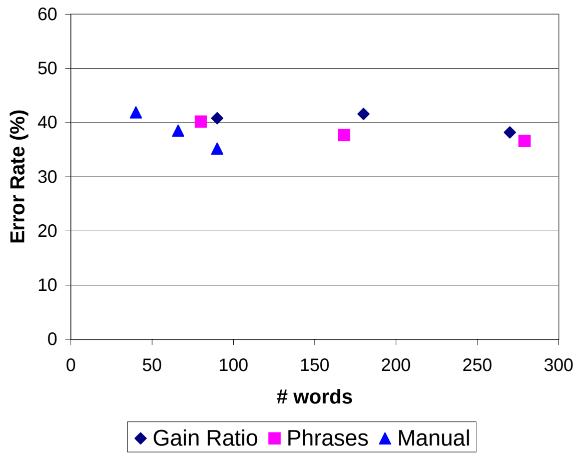


Figure 5: Error rates on the **tq2** testset

the rules contained many domain-specific content words such as "patent", "planet" and "caffeine".

On the other hand, the `Manual` scheme showed significantly lower error rates than the automatic schemes on the tq1 testset. In particular, when the full 90 words were used, the error rate was 20%, which is comparable to the error on the training set (17.5%). However on tq2, errors were only slightly less (35%) than the other schemes. This indicates that the manually selected words transferred to other questions in the same domains very well, but not to questions in other domains. From those results on external testsets, we see that it is quite difficult to derive a broad-coverage, *scalable* feature set for question type classification, whether it is done automatically or manually.

Lastly, one interesting observation was that the error rates of the `Manual` scheme decreased monotonically as the number of features increased on both training set and test sets. This means that all words in the set were critical in determining the question types. Thus, we can see that in general the semantics of a question is made of non-content words of various part-of-speech categories, including interrogatives, nouns, verbs and adjectives, therefore we must consider all such words in order to correctly identify question types.

5 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 breaking the sentence into phrases, and then looking at either the interrogative word and the semantic class of the head noun (Abney, Collins, and Singhal, 2000; Cardie et al., 2000; Harabagiu et al., 2000), or applying question patterns or *templates* (Hovy et al., 2001). In our work, we consider verbs and other part-of-speech categories as well as head nouns (taken from the original 543 unique words) in all schemes. As we discussed in the previous section, those additional words could make significant contributions in identifying question types for general question sentences.

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 (Miller, 1990).

As for the schemes for selecting features from questions, (Agichtei, Lawrence, and Gravano, 2001) describes a method which learns phrase features for classifying questions into question types.

Their method looks for common sequences of words (i.e., n-grams) anchored at the beginning of a sentence, and extracts sequences which occur more than some number of times in the training set. However, the n-gram method can extract idiosyncratic patterns, but it does not apply directly to questions without fixed expressions.

6 Conclusions and Future Work

In this paper, we have shown that NLP techniques are more effective than IR techniques in question type identification, but it is still very challenging to derive schemes for selecting features which generalize to broad domains from sample questions. Automatic categorization of question type is potentially quite useful in paraphrasing questions, or in identifying whether or not two questions are paraphrases of each other. We are currently adding the matching of question type to the other metrics which are used in FAQFinder to compute similarity between user and FAQ questions. Preliminary results show that the use of this additional information indeed improves the system's performance.

We are planning to apply our selection schemes to TREC-8 and 9 data, and compare results to other systems, in particular to those which used question templates (Hovy et al., 2001; Harabagiu et al., 2000). We also plan to investigate ways to learn question features. We would like to extend the n-gram method used in (Agichtei, Lawrence, and Gravano, 2001) by including collocational information (Wiebe, McKeever, and Bruce, 1998).

Finally, we would like to investigate incorporating the use of semantic information into question classification. In the current work, we used (stemmed) words as features. We can certainly experiment with semantic classes instead, by using a general lexical resource such as WordNet. The use of 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 scalable to a wide range of domains. We believe the semantics of the words will greatly assist in question classification for general question-answering systems.

References

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

Agichtei, E., S. Lawrence, and L. Gravano. 2001. Learning search engine specific query transformations for question answering. In *Proceedings of the 10th International World Wide Web Conference (WWW10)*, Hong Kong.

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

Cardie, C., V. Ng, D. Pierce, and C. Buckley. 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)*, Seattle, WA.

Cost, S. and S. Salzberg. 1993. A weighted nearest neighbor algorithm for learning with symbolic features. *Machine Learning*, 10(1).

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

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

Levin, B. 1993. *English Verb Classes and Alternations: A Preliminary Investigation.* The University of Chicago Press.

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

McKeown, K., J. Klavans, V. Hatzivassiloglou, R. Barzilay, and E. Eskin. 1999. Towards multidocument summarization by reformulation: Progress and prospects. In *Proceedings of the 16th National Conference on Artificial Intelligence (AAAI-99)*, Orlando, Florida.

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

Quinlan, R. 1994. *C4.5: Programs for Machine Learning.* Morgan Kaufman.

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

Schank, R. 1973. Identification of conceptualizations underlying natural language. In R. Schank and K. Colby, editors, *Computer Models of Thought and Language.* Freeman.

Spitters, M. 2000. Comparing feature sets for learning text categorization. In *Proceedings of the RIAO-000*, Paris, France.

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

Wiebe, J., K. McKeever, and R. Bruce. 1998. Mapping collocational properties into machine learning features. In *Proceedings of the 6th Workshop on Very Large Corpora (WVLC-98).*

Yang, Y. and J. Pedersen. 1997. A comparative study on feature selection in text categorization. In *Proceedings of the 14th International Conference on Machine Learning (ICML-97).*

Appendix: The 12 Question Types

1. `DEF (definition)` – question which asks for a definition of something. This type includes typical 'what-is/are' questions as well as questions which ask for descriptive definitions by 'what does X do' and 'how does X work'.
   - "What is CORBA?"
   - "How does RMI work?"

2. `REF (reference)` – question which asks for a simple reference, typically a fact. Fill-in-the-blank questions using 'what' or 'which' (excluding 'who', 'where' and 'when').
   - "What are the numbers for the U.S. Copyright Office?"
   - "Which are good races to speculate at?"

3. `TME (time)` – question which asks for a simple reference to a time in general. This type excludes questions of type 11 `INT`.
   - "When is the new moon?"
   - "What years are leap years?"

4. `LOC (location)` – question which asks for a simple reference to a location in general. This type excludes questions of type 10 `ATR`.
   - "Where is Sam Adams beer made?"
   - "What places should we visit in Italy?"

5. `ENT (entity)` – question which asks for a simple reference to an entity in general. This type excludes questions of type 10 `ATR`.
   - "Who created Mr. Bill?"
   - "Which companies have ftp sites?"

6. `RSN (reason)` – question which asks for a reason, causation or goal.
   - "Why did John go to New York?"
   - "What causes migraine headache?"

7. `PRC (procedure)` – question which asks for a procedure involved in an action. 'How-to' questions. Answers to this type of questions are prescriptive and/or instructional (as compared to question type 8 `MNR`).
   - "How should I store beer?"
   - "What should I do to prepare for a law school?"

8. `MNR (manner)` – question which ask for a manner of an action. Answers to this type of questions are descriptive.
   - "How do I deal with cursed items?"
   - "How did the solar system form?"
   - "What is the effect of altitude?"
   - "What happens if I replace the crystal?"

9. `DEG (degree)` – question which ask for a degree or extent. Most 'How+adj/adv' questions, including 'how-many' and 'how-much'.
   - "How accurate is my meter?"
   - "How much protein is in an egg?"
   - "What percentage of children are vaccinated?"

10. `ATR (atrans)` – question which ask for a procedure ('how-to') for obtaining something (physical object or information). Special case of `PRC` (type 7) and `LOC` (type 4) or `ENT` (type 5). Questions of this type can be rephrased by asking the location ('where') or entity ('who') of the source or destination.
    - "How can I get tickets for the Indy 500?"
    - "Where can I find out about sailing?"
    - "Who sells brand X equipment?"
    - "Which vendors are licensing OpenGL?"

11. `INT (interval)` – question which asks for a degree (DEG) that embodies the notion of interval of time. Special case of `DEG` (type 9) and `TME` (type 3). Questions of this type can be rephrased by asking for the time ('when') of the end points.
    - "How long do negative items stay on my credit report?"
    - "When will we run out of crude oil?"

12. `YNQ (yes-no)` – yes/no question.
    - "Did John go to New York?"