

# FAQFinder Question Answering Improvements Using Question/Answer Matching

Stanley J. Mlynarczyk

Steven L. Lytinen

School of Computer Science, Telecommunications, and Information Systems  
DePaul University  
243 S. Wabash  
Chicago, Illinois 60604  
United States of America

## Abstract

FAQFinder is a tool that was developed to provide answers to user questions through the retrieval of previously asked questions residing in Internet Frequently Asked Questions (FAQ) files, primarily USENET FAQ files. Instead of generating responses to a user question from scratch, the FAQFinder approach is to analyze a user's natural language query and to use this analysis to find a similar question that has been asked and answered previously. Our approach is to use a variety of techniques from information retrieval and natural language processing to match user questions with similar questions in FAQ files.

In our previous work, we have made no use of the answers in FAQ files other than to display them as potential answers to the questions that users pose to FAQFinder. In this paper, we report the results of incorporating a new matching process into FAQFinder, this time matching user questions directly with answers in the FAQ files. The matching process uses a set of question transformation patterns, which specify what the answers to various kinds of questions should look like. This new matching process uses a context-free parser to extract basic syntactic constituents from a user's question (e.g., the question's subject, main verb, and direct object). The constituents are then plugged into one or more transformation rules to predict the general forms that an answer to the question is likely to take.

## Introduction

We have developed a question-answering system called FAQFinder, which can answer many questions posed by users on a wide variety of topics. FAQFinder uses Frequently Asked Question (FAQ) files to enable it to answer users' questions. The program uses a variety of techniques from information retrieval and natural language processing to determine (a) which FAQ file(s) are most relevant to a user's question; and (b) which particular question(s) in a FAQ file are most similar to the user's question. If a FAQ question is found which is similar enough to the user's question, then FAQFinder presents the FAQ question's answer to the user as its answer to the original question.

FAQFinder has been under development for a number of years. Our previous work has focused on developing and refining techniques for matching user and FAQ questions;

that is, determining which FAQ question(s) are most similar to a user's question. In our previous work, we have made no use of the answers in FAQ files, other than to present a FAQ answer to the user as a likely answer to the user's question.

In this paper, we describe a new approach to finding the answer to a user question in a FAQ file. In the new approach, we incorporate a matching process which attempts to match a user question directly with an answer in a FAQ file. The approach uses a set of question transformation patterns, which specify what the answers to various kinds of questions might look like. This new matching process uses a context-free parser to extract basic syntactic constituents from a user's question (e.g., the question's subject, main verb, and direct object). The constituents are then plugged into one or more transformation rules to predict the form(s) that an answer to the question is likely to take.

In the remainder of this paper, we first review our previous work in FAQFinder, which has focused on matching user questions with FAQ questions (question/question matching). Then we describe our approach to directly matching user questions with FAQ answers (question/answer matching). Finally, we report experimental results which show an improvement in FAQFinder's performance based on the incorporation of question/answer matching using our SVO based approach.

## Question/Question Matching

Matching of a user question to a FAQ file question has been the primary approach to FAQFinder research since its inception. Two approaches have been used; one based on word document frequency and another on semantic analysis of words using WordNet distance.

## Vector Analysis Using Word Frequency

One of the methods used by FAQFinder computes weights for each term in a user question and is based on word document frequency (Korfhage, 1997). Each sentence's words (terms) are first stemmed and then tagged according to Part Of Speech (POS) using the Brill Tagger (Brill, 1995). Each FAQ question is then considered to be a "document" and a user's question will

be matched against the question in a FAQ file document (the question component of a FAQ file).

Question terms are converted to a term vector and a POS tagged "term set". The vector components' individual weights are computed using the standard *td.idf* formula (Salton and McGill, 1983):

$$W_i = (1 + \log(tf_i)) \log N/df_i$$

Where,  
*i* = the term index  
*tf<sub>i</sub>* = frequency this term occurs in the question  
*df<sub>i</sub>* = frequency of questions in which term *i* appears.

The above is computed for each term in both the user question's vector and for each FAQ question's vector in a given FAQ file. Comparisons between the user question vector and each FAQ file's question vector is accomplished through the use of the cosine measure (Salton and McGill, 1983):

$$\cos(v_{user}, v_{faq}) = \sum W_{useri} W_{faqi} / (\sqrt{\sum W_{useri}^2} \times \sqrt{\sum W_{faqi}^2})$$

Where,

*W<sub>i</sub>* = weight of term *i* (where there is a matching term)  
*v<sub>user</sub>* = weight vector for the user question  
*v<sub>faq</sub>* = weight vector for the FAQ question

### Semantic Similarity Using Word Distance

In addition to the word frequency approach described above, FAQFinder also uses simple lexical semantic information in its question/question matching. The source of this lexical semantic information is WordNet (Miller, 1990; Fellbaum, 1998).

FAQFinder uses WordNet for two purposes. First, it attempts to disambiguate each noun and verb in a question, by trying to determine which of the WordNet senses of a word is likely to be the intended sense. Word senses in WordNet are called "synonym sets" (or synsets). Synsets are connected to each other through various links which denote different types of semantic relationships. We use the *hypernym/hyponym* links, which encode "x is a kind of y" relationships among synsets. In particular, FAQFinder disambiguates words by selecting a synset for each word such that there is minimum distance between the synsets. The distance between two synsets is measured by counting the minimum number of hypernym/hyponym links which must be traversed to connect them. The following formula is used to find the minimal set of sense combinations:

$$\Delta(S) = \sum_{s_i \in S} \min_{s_j \in S, i \neq j} D(s_i, s_j)$$

Where S = Set of combinations of all synsets

*s<sub>i</sub>, s<sub>j</sub>* = synsets  
*D* = distance

Second, WordNet is also used to match user and FAQ questions. We compute a "semantic distance" between two questions by finding the sum of the distances between pairs of words in the two questions:

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

Where

$$I(u, f) = \sum_{u \in T} \frac{1}{1 + \min_{f \in T} d(u, f)}$$

And

$$I(f, u) = \sum_{f \in T} \frac{1}{1 + \min_{u \in T} d(f, u)}$$

|*T<sub>u</sub>*| and |*T<sub>f</sub>*| denote the size of *T<sub>u</sub>* and *T<sub>f</sub>* where:

*T<sub>u</sub>* = {*u<sub>1</sub>, ..., u<sub>n</sub>*}; a tagged term set representing the user question.

*T<sub>f</sub>* = {*f<sub>1</sub>, ..., f<sub>m</sub>*}; a tagged term set representing a FAQ question.

These two approaches (vector analysis and semantic analysis) give FAQFinder two measures of similarity between user and FAQ questions. The two metrics are combined to produce a composite measure of question similarity. FAQFinder then uses this composite similarity measure to select a FAQ question which best matches the user question, and presents that question's answer as its answer to the user's question.

### Question/Answer Matching

While vector and semantic approaches to question matching have proved to be effective, the formatting of some FAQ files results in sub optimal performance of question/question matching. Some FAQ files contain questions that require domain knowledge or contextual knowledge to properly extract semantic meaning. In some cases a FAQ file is structured to answer many anticipated questions without the presence of corresponding user questions. To address these issues research was undertaken to determine the best way to address such limitations.

A first attempt was to use word frequency analysis of the answer components of FAQ files. This proved to be of limited value because FAQ files contain groupings of questions/answers that are categorically similar. It became apparent that a much more fine-grained level of analysis was needed.

Our next, more successful, approach was to develop a simple taxonomy of question types along with a set of transformation rules which predict the possible form (s) of answers to each type of question. The general approach is similar to that used in Webclopedia

(Ravichandran and Hovy, 2002), although our taxonomy of question types is much simpler. We use a simple context-free parse of a question to identify key constituents, such as the question's subject, main verb, and direct object. This information is then used to "instantiate" a transformation rule (or possibly several rules) which then predict the likely form(s) of the question's answer.

Question transformation rules take the basic form of:

question type:question form:answer form

where the general form for the grammar for both question and answer is a combination of words and variables. Some of variables currently defined include:

\$SUBJ, \$VERB, \$OBJ – to accommodate identification of the key syntactic constituents in a question.

\$NN, \$ADJ, \$NNP, \$PRON, \$ADV, \$VB – the current parts-of-speech supported.

Linkage between a question form and potential answer form is supported by the ability to associate a variable in the answer form with a corresponding variable in the question form using the "..." notation; e.g.

HOW:How is/\$VERB [the] \$NNP/\$OBJ different from [the] \$NNP/\$OBJ: [The] \$NNP/\$SUBJ..\$OBJ differs/\$VERB from [the] \$NNP/\$OBJ..\$SUBJ

In the above example, there is a linkage between the potential answer's object (\$OBJ) with a question's subject (\$SUBJ). Similar support is provided for other variable types. Bracketed words are optional but do add to the match weighting when they are present. A word or variable can be further qualified using the '/'. For example, \$NNP/\$OBJ indicates that we are expecting a proper noun that will play the role of the sentence object.

There is a potential that multiple question forms may produce matches and thus each potential answer form must be applied to a target FAQ file answer component.

The question/answer match is weighted such that it is given preference when question/question matching is inconclusive and when question/answer has a strong candidate answer.

Proper Part Of Speech (POS) tagging is essential for proper matching of user questions with question/answer forms within the pattern files. WordNet, a dictionary based on the Webster Dictionary, and a small supplementary dictionary developed explicitly for FAQFinder are consulted for POS identification after a question is first tagged using the Brill Tagger. Tagging occurs not only on the user question but also on all text within an answer

component of a FAQ file. This is computationally intensive, however, the majority of the FAQ file answer component tagging will eventually be converted to a pre-parsed format.

To further enhance FAQFinder's ability to correctly parse user questions and FAQ questions and answers, translation logic has been implemented to parse phrase forms. Approximately 50,000 phrases (extracted from WordNet) have partially been classified. A similar approach is taken to identify named entities, acronyms, numbers and dates, although the effort to implement this is ongoing.

Two purposes are served by phrase support. First, phrases make parsing/tagging potentially less accurate and translation into a simpler form enhances the chances for proper tagging/parsing. Secondly, phrase translation decreases the complexity and number of match rules in our grammar class files.

## Evaluation of FAQ Answer Component

To evaluate the benefit of including question/answer matching in FAQFinder, we developed a set of 124 question transformation patterns for the HOW question type. This type was selected for initial testing because it poses a greater challenge over other question types such as: when, yes/no, where, who. This is because HOW questions are more open-ended, and therefore their answers can be more involved.

### Recall vs. Rejection

In our test, we selected a set of HOW 150 questions gathered from a web based question collection screen that allowed users to input questions related to a set of specific topics. Of the 150 selected questions, 100 were judged to have a potential answer and 50 to have no correct answer although these 50 questions were related to the FAQ file topics. There were a total of 733 FAQ questions in 38 FAQ files. The contribution of the FAQ answer component match contributed 9 matches.

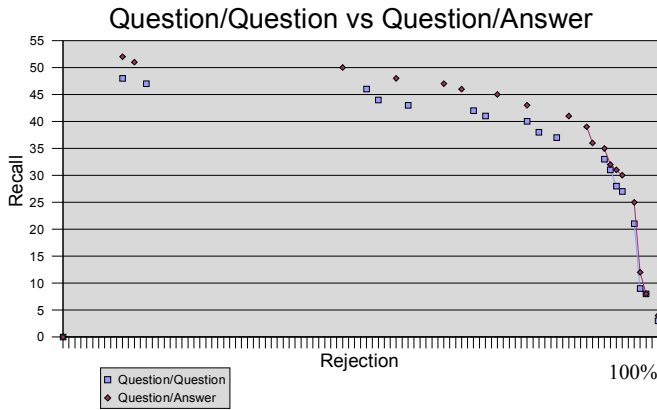
We evaluated FAQFinder's performance using a metric based on recall and rejection. We feel this is better suited to FAQFinder's task than precision.

Recall is computed as the percentage of test questions with at least one correct match. Rejection is computed as the percentage of test questions with no correct match for which FAQFinder finds no matches. Although not the standard way to compute recall (FAQFinder can achieve 100% recall by only displaying one appropriate match even though there may be many matches), we feel its better suited to the task, because if FAQFinder returns more than one match, the user does not care if "all" potential matches are found.

The graph shows FAQFinder's performance on the test question set both with the traditional approach

where we focus on matching a user question with a FAQ question, and also where consideration is given to a match between a user question and the answer component of a FAQ file.

Our results using this approach improved FAQFinder accuracy as shown below:



As can be seen from the above graph, the answer component match did increase recall. This increase is maintained as rejection increases.

In analyzing matches, we found that in some cases FAQfinder was able to find good matches that we were a surprise to us. These were valid matches that our manual efforts at identifying potential FAQ answers did not reveal. There were 7 such successful matches beyond those we expected.

### Conclusion and Future Work

Our conclusion is that matching of user questions with potential FAQ file answer candidates benefits from a variety of traditional text retrieval and semantic mechanisms. Our question/answer matching approach holds promise for additional work in this area.

We anticipate the need to further explore combining both question/question and question/answer approaches to increase FAQ file question matching. This will require additional discrimination functionality within the answer component approaches. We will explore further refinements of our current approach to include the use of additional semantics. This will include use of WordNet for word distance analysis between question text and answer text (currently we've only implemented distance analysis between user and FAQ questions) and additional enhancements to named entity support.

Lastly, our question matching functionality can potentially improve question/question matching because of the additional information provided by the context-free parse of questions and subsequent identification of key syntactic constituents in the questions. This is because one of the semantic methods used for question/question match depends on the accuracy of

identification of syntactic constituents.

In conclusion, our results indicate that deeper semantic analysis in FAQ answer component can improve FAQFinder accuracy without decreasing overall accuracy.

### References

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

Fellbaum, C. (ed). WordNet: An Electronic Lexical Database. MIT Press, Cambridge MA, 1998.

Korfhage, R. (1997). Information Storage and Retrieval. John Wiley & Sons, Inc. (1997)

Lytinen, S.; Tomuro, N.; Repede, T. (2000). The Use of WordNet Sense Tagging in FAQFinder. In *AAAI 2000 Workshop on AI and Web Search (2000)*

Lytinen, S.; Tomuro, N. (2002). The Use of Question Types to Match Questions in FAQFinder. In *AAAI 2002 Spring Symposium on Mining Answers From Text (2002)*

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

Ravichandran, D. and Hovy E. (2002). Learning Surface Text Patterns for a Question Answering System. In *Proceedings of the 40th Conference of the Association of Computational Linguistics (ACL-2002)*, Philadelphia PA, July 2002.

Tomuro, N. (1989). Semi-automatic Induction of Systematic Polysemy from WordNet. In *17th International Conference on Computational Linguistics (COLING '98)*

Tomuro, N. (2002). Question Terminology and Representation for Question Type Classification. In *19th International Conference on Computational Linguistics (COLING '02)*

Voorhees, E., and Harman, D (editors, 2001). *The Tenth Text REtrieval Conference (TREC 2001)*.ernment Printing Office, stock number SN003-003-03750-8