FAQ Finder

We are developing a class of systems, called FAQ Finder systems, that use a natural language question-based interface to access distributed text information sources, specifically text files organized as question/answer pairs such as FAQ files (Hammond, et al. 1995). In using these systems, a user enters a question in natural language and the system attempts to find an information source that answers the question, and then find the closest matching question/answer pair. As previous publications have described, these systems combine three technologies: statistically based IR engines, syntactic natural language analysis and semantic networks. In particular, they combine the SMART information retrieval system, a natural language parser, a semantic net derived from Princeton's WORDNET, and question-answering strategies that handle questions that cannot be answered by straightforward question matching.

The power of our approach rises out of two features: We are using knowledge sources that have already been designed to “answer” the commonly asked questions in a domain and as such are more highly organized than free text. We do not need our systems to actually comprehend the queries they receive. They only have to identify the files that are relevant to the query and then match against the segments of text that are used to organize the files themselves (e.g., questions, section headings, key words, etc.).

The most natural kind of interface to a database of answers is the question, stated in natural language. While the general problem of understanding questions stated in natural language remains open, we believe that the simpler task of matching questions to corresponding question/answer pairs is feasible and practical.

As it stands, the FAQ Finder system is an automated question-answering system that uses the files of “Frequently-Asked Questions” (FAQs) associated with many USENET newsgroups [1,2]. These files are compendiums of the accumulated wisdom of the newsgroup on topics that are of frequent interest. Users connect to FAQ Finder using the World-Wide Web. The system takes a user’s query on any topic, attempts to find the FAQ file most likely to yield an answer, searches within that file for similar questions, and returns the given answers.

The FAQ Finder system

The technology used to develop the basic FAQ Finder system is fairly simple (See figure 1). We have combined four central ideas in constructing the system:

- Statistical information retrieval, embodied in SMART [3], is used to select FAQ files, given a particular question.
- Syntactic parsing, embodied in a bottom up chart parser, is used to construct a simple parse tree and identify the primary verb and noun phrases in a question.
- A taxonomy of question types, describing the different types of commonly-asked questions, is used to focus the system’s search for similarity between questions. Each question type has a recognizer that operates on the parse tree.
- Semantic concept matching, through the use of the WORDNET network of lexical semantics and the MOBY machine-readable thesaurus, is used to select possible matches between the query and target questions in the FAQ files.
- Question matching strategies, for answering questions when there is no closely-matching question in the file.

Figure 2 shows the World Wide Web interface to FAQFinder. Here the user enters the question:

*Is downshifting a good way to slow down my car?*
FAQ Finder uses SMART to identify FAQs that are possible sources of the answer. In Figure 3, the system presents the files, ordered by degree of match. The best candidate, *Automobile Consumer Frequently Asked Questions* file, is obviously the best answer in this case, but if the user is unsure, the system provides feedback about what words in the question prompted the retrieval. See Figure 4.

Once an appropriate FAQ file has been identified, the system parses the query into a syntax tree. The goal here is not to parse into an unambiguous structure, but instead to parse into a rough representation that can be used to support matching. After this parse is completed, the question-type recognizers operate on the parse tree to identify the appropriate category for the question. For example, questions beginning with the phrase “What is the difference between...” tend to fall into the category Q-COMPARISON.

The matching algorithm is based on the classical marker-passing algorithms of Quillian [4]. In Quillian’s system, marker-passing in semantic space was used to identify candidate structures which were then compared to *form tests* to judge their linguistic accuracy. For example, the input phrase “lawyer’s client” would cause *marker* data structures to be passed through a network from the *lawyer* and *client* concepts. One concept discovered by this search would be the *employment* concept, with the form test: “*first’s second*”. The form test verifies that the input actually was of the proper form to identify the *employment* concept.

From the point of view of FAQ Finder, Quillian’s basic algorithm had the particularly useful feature that it was *fast*. The marker-passing phase relies solely on
the shallow lexical semantics of Wordnet and the natural language thesaurus; the annotations of each FAQ link are not taken into account. Because there is no checking to make sure that complex semantic or syntactic relations are satisfied, this marker-passing phase is very fast.

Marker passing in general produces an abundance of matches. FAQ Finder filters the answers by using several criteria. First, it eliminates questions that have a different question type from the question as asked. A question “How often should I change my oil?” (type Q-FREQUENCY) is not well answered by the answer to the question “How much should I expect to pay to get my oil changed?” (type Q-QUANTITY). Simple question-type pruning eliminates such confusion.

A second stage of filtering is performed using the results of parsing. The kind of mismatch we are hoping to prevent is that between questions that use similar words, but in different ways. For example, the question “How do I change the used oil in my motor?” and the question “How does motor oil change as it is used?” required quite different answers, even though they are both Q-HOW questions and match against the words “oil,” “motor,” “change,” and “user.” Because the parser identifies “I” as the subject in the first sentence and “oil” in the second, a possible match between them can be rejected. Note that all of the questions in each FAQ file are pre-parsed off-line so that this comparison does not involved additional parsing of the questions from the FAQ file.

When a match is found, the matching question/answer pair is presented to the user (Figure 5). At every stage of this process, the user is given some control over disambiguation and selection of the final answer.

Inevitably, there are situations in which the information needed to answer a user's question is in the file, but it is associated with a question that is not a good match to the user's stated question. This can happen for several reasons.

- FAQ writers often combine the answers to many questions into a single long answer intended as an introduction. For example, in the Rose Gardening FAQ, there is the question “How do I propagate roses?” with an answer that covers several common methods. If the user asks, “How do I perform bud grafting on my rose bushes?” it will not match any questions in the file.

- Question matching may also fail in situations where a FAQ file is not entirely question/answer-based. For example, many FAQs contain a special section in which terms are defined. If a user wants to know, “What is an aubergine?” there will be no corresponding question in file, although there is a list of food names in which the term is defined.

- FAQ writers take advantage of the way that readers will scan FAQ files when writing them. For example, in the investment FAQ, there is a section on “technical analysis” of stock market trends. One of the questions in that section reads “Does it have any chance of working?” Nothing in the question refers to the topic, making it difficult to match against.

- Finally, there is the problem of the phrasing of a question. Any given question may be asked in a multitude of ways.

We have attempted to address these situations by employing in FAQ Finder a library of question-answering strategies. If the system cannot match
the user's question, it looks for a strategy that can be employed to transform the user's question. The result is a different question or a set of questions that may have an answer in the FAQ file.

For example, if the user asks "What is the difference between ale and lager?" the question-matching algorithm will fail to answer because there is no comparison question for different types of beer in the beer FAQ file. However, there are descriptions of different types of beers, which could help the user is forming his or her own comparison. The comparison strategy therefore calls for the transformation of a comparison question into a set of "What is..." questions according to what is being compared: "What is the difference between ale and lager?" will become "What is ale?" and "What is lager?" As it turns out, there are answers to these questions in the FAQ file.

**Question-answering strategies**

As part of our initial testing of FAQFinder, we asked college students to contribute questions on a range of topics, and received around 200 questions. We manually compared these questions with the questions found in the files. We found that about 1/3 of the questions were close enough that we expect question-to-question matching to be successful. Where there was no straightforward match, we tried to determine whether useful information was nonetheless present in the file.

In many cases, simple transformations of questions were sufficient to render user's questions answerable. The "comparison to 'What is'" transformation in the ale/lager example is one such question-answering strategy. At this point, we are exploring the space of possible strategies, by developing and extending them as required. We are concentrating on strategies that can be employed without employing deep semantic knowledge. The following partial list gives some sense of the different types of strategies we are examining.

**Comparison breakdown:** This strategy converts comparison questions into individual identification questions, as in the "vs ale" question above.

**Definition search:** If a direct answer to an identification question, such as "What are CFCs?" cannot be found in the FAQ, the system looks for a glossary in the FAQ where terms are defined.

**"When" to "How":** Convert unanswered temporal questions into procedural questions. For example, "How often should I synchronize my carbs?" becomes "How do I synchronize my carbs?" Answers to procedural questions often contain a comment like "You should do this every...

**Problem to solution:** Questions about causes of problems can sometimes be successfully answered by referring to problem fixes. "What causes my engine to knock?" may be answered in an answer to the question "How do I fix engine knock?"

**Problem to prevention:** Similar to above, but looks for prevention questions. "What causes my engine to knock?" becomes "How do I keep my engine from knocking?"

**Instrument to plan:** As described above, converts questions about the instruments used to achieve some plan into "How to" questions. "Can I use bottles with twist-off caps to store my beer?" becomes "How do I store my beer?"

Even with such strategies, we find that a significant subset of questions are probably beyond the scope of a simple question-matching approach. Some questions require either in-depth domain understanding or an understanding of the questioner's planning process. For example, one questioner asked "Can I scan images from magazines, alter them, and then use them for my own purposes?" The answer is found in the "Copyright FAQ" under the question "What is a fair use?" In order to make this connection, however, the system would have to be able to know that the operations described in the question constitute a kind of copying and that the concept of fair use is relevant to copying. These topics are beyond the scope of what we expect from FAQ Finder.

**Future Work**

Our continuing work on FAQ Finder pushes in two directions. First, we are building FAQ Minder, a semi-automated tool for handling the diversity of FAQ file formats. This interactive FAQ file maintenance and indexing tool will allow a user to create FAQ files and to build the annotations required by FAQ Finder. Our long-term aim is the distribution of this tool to individual FAQ maintainers.

Second, we are expanding the amount of information being brought to bear on the matching process. We are building matching procedures that use the structure of the FAQ file to resolve ambiguous references ("What is it?") by referring to topic headings or previous questions. We are integrating the information in the MOBY Thesaurus to expand the semantic information found in WordNet. We are also exploring techniques for extracting domain-specific terminology from FAQ files themselves. General-purpose dictionaries are being used for general topics, but less useful in technical discussions that contain highly-specialized terms.
Bibliography


