Task Orientation in Question Answering

Vanessa Murdock
Center for Intelligent Information Retrieval
Computer Science Department
University of Massachusetts
Amherst, MA 01003
vanessa@cs.umass.edu

W. Bruce Croft
Center for Intelligent Information Retrieval
Computer Science Department
University of Massachusetts
Amherst, MA 01003
croft@cs.umass.edu

Categories and Subject Descriptors

H.3.4 [Information Storage and Retrieval]: Systems and Software - *Performance evaluation*

General Terms

Measurement and Performance

Keywords

Query Classification, Task-Orientation

1. INTRODUCTION

Information retrieval techniques treat questions democratically: all questions, regardless of their grammar or orientation, are processed by the same rules and manipulated in similar ways. In about 12% of the questions in our data users ask about a process. Treating a request for how to do something as a request for information about a topic disregards an important subtlety in the question. We use the knowledge that a question asks about a process, rather than a fact, to develop categories of information need implied by the questions to aid in retrieval. We focus on questions because by their nature questions carry more information.

A well-known technique for improving retrieval is query expansion. A variety of query expansion and disambiguation techniques have been studied [1] [2] [3] [5], and although none distinguish between fact-orientation and task-orientation, several do distinguish categories of information need not explicit in the queries.

Understanding that the user has already made their information need clear by using a question rather than a keyword query, we utilized information encoded in the fact- or task-orientation to infer the type of information need. We found a measurable difference between task- and fact-oriented questions, and preliminary results indicate that treating fact- and task-oriented questions differently improves retrieval.

2. FACT AND TASK QUESTIONS

Our corpus of 4100 questions came from the query logs of the search engine, Govbot [1]. About 4 in every 1000 queries to Govbot were well-formed questions, with an average length of 8 words, and a total vocabulary of 3700 unique words pertaining to government. Variation in the grammatical structure of the sentence was reduced because every question began with a question word (who, what, where, when, why, how).

To get an initial estimate of the number of fact-oriented and task-oriented questions in our corpus, we developed rules based on the question words. We divided the 4100 questions into a fact-oriented set and a task-oriented set. According to these rules, 700 of the questions were task-oriented, and 3400 were fact-oriented. We did not filter out ambiguous questions; they were included in either the fact-oriented set or the task-oriented set depending on their question words.

In order to analyze the question structure in more depth, we constructed parse trees of the questions using a statistical parser, SIFT, from BBN [4]. From our set of 700 task-oriented questions, 630 contained the same distinctive parse of the verb phrase. Of our 3400 fact-oriented questions, only 490 contained that specific internal structure.

This suggests fact-oriented questions and task-oriented questions are distinct. It also suggests that task-oriented questions relate the verb to the noun phrase more directly than fact-oriented questions. Traditional information retrieval techniques do not give the verb any special consideration. The verb is frequently stemmed into a noun, or removed as a stop-word, altering the relationship between the verb and the noun phrase.

3. CLASSIFYING QUESTIONS USING LANGUAGE MODELS

There are many approaches to classification. Because we are interested in applications of language modeling in information retrieval, we built a classifier using language models, with perplexity as our metric.

Examining each question by hand, we created a training corpus of 580 task-oriented questions, and a test corpus of 60 task-oriented questions. We created two training corpora of 720 and 2930 fact-oriented questions, and a test corpus of 80 fact-oriented questions. For the sake of the accuracy of the language model, we eliminated questions that were ambiguous because they could not be answered with a fact or a process, such as those asking for opinions.

Table 1 shows the percentage of questions correctly classified by a variety of techniques. Trigram models performed the best, in terms of classification accuracy. Increasing the

[1] Govbot was a search engine for .gov and .mil domains, in operation from 1995 to 2001 at the Center for Intelligent Information Retrieval, at the University of Massachusetts.

SIGIR'02, August 11-15, 2002, Tampere, Finland.
ACM 1-58113-561-0/02/0008.

Technique	Accuracy	
	task	fact
unigram model	75%	59%
bigram model	88%	70%
trigram model	88%	74%
4-gram model	88%	73%
Increasing training set size	93%	73%
using VB-NP for task questions	28%	89%
fitting the perplexity curve	95%	75%
parse trees	77%	87%
rule-based approach	81%	99%

Table 1: Percent of questions correctly classified using a variety of classification techniques.

training-set size of the fact-oriented model from 720 questions to 2930 questions produced an improvement in accuracy of 5% for task-oriented questions.

Questions may use the same vocabulary but have a different orientation, as in "How do I apply for a passport" and "How long does it take to apply for a passport." The similarity in the vocabulary makes it nearly impossible for a language-model based classifier to distinguish between them. To reduce the similarity between the two training corpora, we reduced the task-oriented training set to its verb-noun phrase component. The fact-oriented training set could not be reduced because it doesn't share this construction. We trained a language model on the verb-noun phrase pairs from the task-oriented training set, and used a trigram model trained on 720 fact-oriented questions. The fact oriented questions showed an improvement of 15%.

Each of these classifiers assumes, by choosing the strictly lower perplexity score, that the threshold dividing the two classes is the line at which the perplexity scores of the two models is the same, i.e. the line $y = x$. But the threshold between the two classes may be a curve, or it may be a line with a different slope. To improve upon the simple trigram model we combined the best classifier for the task-oriented questions, and the best classifier for fact-oriented questions. The curve that best incorporates both classifiers is:

$$0.1 * [(0.95C_1)^3 + (0.05C_2)^3]^{1/3}$$

where C_1 is the classifier most accurate for task questions and C_2 is the classifier most accurate for fact questions. Using this classifier, we were able to correctly classify 95% of the task questions, and 75% of the fact questions.

From the parse trees, we know that the task questions have a more regular syntactic structure, whereas the fact questions have much more structural variation. In addition there were more than five times as many fact questions as task questions, introducing more uncertainty in those language models. For these reasons, the classification of fact questions is much less reliable than task questions. Choosing training sets with less linguistic variety, increasing the size of the task-oriented training set, or incorporating syntactic information into the language model may improve the accuracy of the classification.

Looking only for the verb-noun phrase structure we correctly classified 87% of fact questions, and 77% of task questions. Using only the first three words of the questions we had 99% accuracy in fact-oriented questions and 81% accuracy in task-oriented questions. This isn't surprising considering that we devised the rules by looking at a subset of the training set, and the questions in the corpus were selected from the query logs with criteria very similar to the rules in our classifier. In other words, the high accuracy of the rule-based approach reflects over-fitting the data.

That said, there is a strong correlation between the first three words of the sentence and the orientation of the sentence. There is a strong relationship between the syntactic structure of the questions and their orientation. There is a high degree of similarity between questions of one orientation, as suggested by the language model classification. This gives us three pieces of strong evidence that there is a measurable and significant difference between the two types of questions, a difference we can exploit to improve retrieval.

4. CONCLUSION

We have shown that we can classify questions as fact- or task-oriented in three ways: based on the question words, using their grammatical structure, and by training language models. Knowing that the information need implied by the question is different depending on its orientation allows us to develop techniques specific to each genre of question.

A manual analysis of ten questions suggested certain genres of document (specifically forms and FAQs) were almost always relevant to task-oriented questions. We are currently examining whether query expansion by adding genre-specific terms will improve retrieval for task-oriented questions. Since the syntactic difference between the two question types is consistent, we are also looking at whether preserving the relationship between verb and noun phrase, or weighting the verb in the query, improves retrieval.

Acknowledgments

This work was supported in part by the Center for Intelligent Information Retrieval and in part by NSF grant #IIS-9907018, and in part by Advanced Research and Development Activity under contract number MDA904-01-C-0984. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsor.

5. REFERENCES

[1] A. Berger, R. Caruana, D. Cohn, D. Freitag, and V. Mittal. Bridging the lexical chasm: Statistical approaches to answer-finding. In *Proceedings of the 23rd Annual Conference on Research and Development in Information Retrieval (ACM SIGIR)*, pages 192–199, 2000.

[2] C. Buckley, G. Salton, J. Allan, and A. Singhal. Automatic query expansion using SMART : TREC 3. In D. K. Harmon, editor, *NIST Special Publication 500-225: The Third Text Retrieval conference (TREC 3)*, pages 69–80, 1995.

[3] E. J. Glover, S. Lawrence, M. D. Gordon, W. P. Birmingham, and C. L. Giles. Web search your way. *Communications of the ACM*, 44(12):97–102, December 2001.

[4] S. Miller, M. Crystal, H. Fox, L. Ramshaw, R. Schwartz, R. Stone, R. Weischedel, and the Annotation Group. Algorithms that learn to extract information–BBN: Description of the SIFT system as used for MUC-7. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, 1998.

[5] J. Xu and B. Croft. Query expansion using local and global document analysis. In *Proceedings of the 19th Annual Conference on Research and Development in Information Retrieval (ACM SIGIR)*, pages 4–11, 1996.