Question Terminology and Representation for Question Type Classification

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

Abstract

Question terminology is a set of terms which appear in keywords, idioms and fixed expressions commonly observed in questions. This paper investigates ways to automatically extract question terminology from a corpus of questions and represent them for the purpose of classifying by *question type*. Our key interest is to see whether or not semantic features can enhance the representation of strongly lexical nature of question sentences. We compare two feature sets: one with lexical features only, and another with a mixture of lexical and semantic features. For evaluation, we measure the classification accuracy made by two machine learning algorithms, C5.0 and PEBLS, by using a procedure called *domain cross-validation*, which effectively measures the *domain transferability* of features.

1 Introduction

In Information Retrieval (IR), text categorization and clustering, documents are usually indexed and represented by domain terminology: terms which are particular to the domain/topic of a document. However, when documents must be retrieved or categorized according to criteria which do not correspond to the domains, such as *genre* (text style) (Kessler et al., 1997; Finn et al., 2002) or *subjectivity* (e.g. opinion vs. factual description) (Wiebe, 2000), we must use different, domain-independent features to index and represent documents. In those tasks, selection of the features is in fact one of the most critical factors which affect the performance of a system.

Question type classification is one of such tasks, where categories are question types (e.g. 'how-to', 'why' and 'where'). In recent years, question type has been successfully used in many Question-Answering (Q&A) systems for determining the kind of entity or concept being asked and extracting an appropriate answer (Voorhees, 2000; Harabagiu et al., 2000; Hovy et al., 2001). Just like genre, question types cut across domains; for instance, we can ask 'how-to' questions in the cooking domain, the legal domain etc. However, features that constitute question types are different from those used for genre classification (typically part-of-speech or meta-lingusitic features) in that features are strongly *lexical* due to the large amount of idiosyncrasy (keywords, idioms or syntactic constructions) that is frequently observed in question sentences. For example, we can easily think of question patterns such as "What is the best way to .." and "What do I have to do to ..". In this regard, terms which identify question type are considered to form a terminology of their own, which we define as *question terminology*.

Terms in question terminology have some characteristics. First, they are mostly domain-independent, *non-content* words. Second, they include many closed-class words (such as interrogatives, modals and pronouns), and some open-class words (e.g. the noun "way" and the verb "do"). In a way, question terminology is a complement of domain terminology.

Automatic extraction of question terminology is a rather difficult task, since question terms are mixed in with content terms. Another complicating factor is paraphrasing – there are many ways to ask the same question. For example,

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

In this paper, we present the results of our investigation on how to automatically extract

question terminology from a corpus of questions and represent them for the purpose of classifying by question type. It is an extension of our previous work (Tomuro and Lytinen, 2001), where we compared automatic and manual techniques to select features from questions, but only (stemmed) words were considered for features. The focus of the current work is to investigate the *kind(s)* of features, rather than selection techniques, which are best suited for representing questions for classification. Specifically, from a large dataset of questions, we automatically extracted two sets of features: one set consisting of terms (i.e., lexical features) only, and another set consisting of a mixture of terms and semantic concepts (i.e., semantic features). Our particular interest is to see whether or not semantic concepts can enhance the representation of strongly lexical nature of question sentences. To this end, we apply two machine learning algorithms (C5.0 (Quinlan, 1994) and PEBLS (Cost and Salzberg, 1993)), and compare the classification accuracy produced for the two feature sets. The results show that there is no significant increase by either algorithm by the addition of semantic features.

The original motivation behind our work on question terminology was to improve the retrieval accuracy of our system called FAQFinder (Burke et al., 1997; Lytinen and Tomuro, 2002). FAQFinder is a web-based, natural language Q&A system which uses Usenet Frequently Asked Questions (FAQ) files to answer users' questions. Figures 1 and 2 show an example session with FAQFinder. 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 a maximum of 5 FAQ questions which are ranked the highest by the system's similarity measure. Currently FAQFinder incorporates question type as one of the four metrics in measuring the similarity between the user's question and FAQ questions.[1] In the present implementation, the system uses a small set of manually selected words to determine the type of a question. The goal of our work here is to derive optimal features which would produce improved classification accuracy.

[1] The other three metrics are vector similarity, semantic similarity and coverage (Lytinen and Tomuro, 2002).

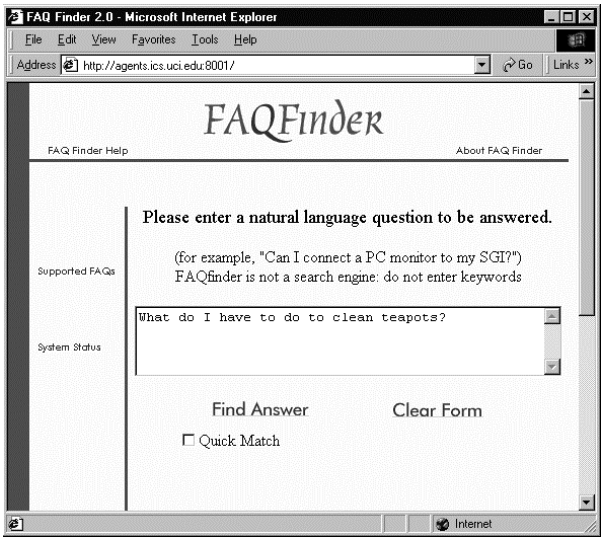

Figure 1: User question entered as a natural language query to FAQFinder

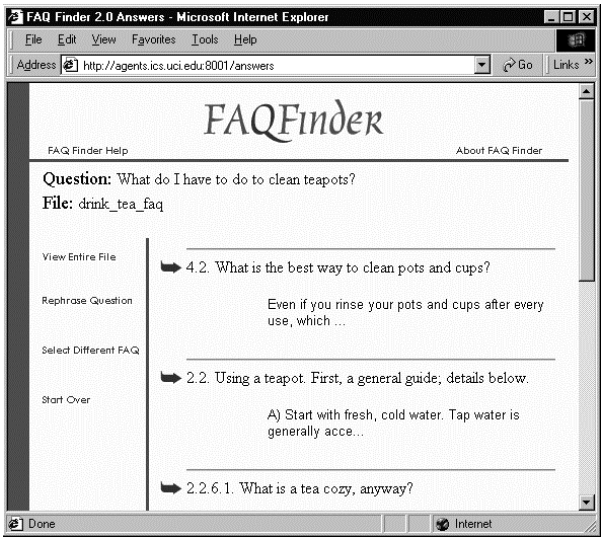

Figure 2: The 5 best-matching FAQ questions

2 Question Types

In our work, we defined 12 question types below.

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)

Descriptive definitions of these types are found in (Tomuro and Lytinen, 2001). Table 1 shows example FAQ questions which we had used to develop the question types. Note that

our question types are general question categories. They are aimed to cover a wide variety of questions entered by the FAQFinder users.

3 Selection of Feature Sets

In our current work, we utilized two feature sets: one set consisting of lexical features only (LEX), and another set consisting of a mixture of lexical features and semantic concepts (LEXSEM). Obviously, there are many known keywords, idioms and fixed expressions commonly observed in question sentences. However, categorization of some of our 12 question types seem to depend on open-class words, for instance, "What does mpg mean?" (DEF) and "What does Belgium import and export?" (REF). To distinguish those types, semantic features seem effective. Semantic features could also be useful as *back-off* features since they allow for generalization. For example, in WordNet (Miller, 1990), the noun "know-how" is encoded as a hypernym of "method", "methodology", "solution" and "technique". By selecting such abstract concepts as semantic features, we can cover a variety of paraphrases even for fixed expressions, and supplement the coverage of lexical features.

We selected the two feature sets in the following two steps. In the first step, using a dataset of 5105 example questions taken from 485 FAQ files/domains, we first manually tagged each question by question type, and then automatically derived the *initial lexical set* and *initial semantic set*. Then in the second step, we refined those initial sets by pruning irrelevant features and derived two subsets: LEX from the initial lexical set and LEXSEM from the union of lexical and semantic sets.

To evaluate various subsets tried during the selection steps, we applied two machine learning algorithms: C5.0 (the commercial version of C4.5 (Quinlan, 1994), available at `http://www.rulequest.com`), a decision tree classifier; and PEBLS (Cost and Salzberg, 1993), a k-nearest neighbor algorithm.[2] We also measured the classification accuracy by a procedure we call *domain cross-validation* (DCV). DCV is a variation of the standard cross-validation (CV) where the data is partitioned according to domains instead of random choice. To do a k-fold DCV on a set of examples from n domains, the set is first broken into k non-overlapping blocks, where each block contains examples exactly from $m = \frac{n}{k}$ domains. Then in each fold, a classifier is trained with $(k-1) * m$ domains and tested on examples from m unseen domains. Thus, by observing the classification accuracy of the target categories using DCV, we can measure the *domain transferability*: how well the features extracted from some domains transfer to other domains. Since question terminology is essentially domain-independent, DCV is a better evaluation measure than CV for our purpose.

3.1 Initial Lexical Set

The initial lexical set was obtained by ordering the words in the dataset by their *Gain Ratio* scores, then selecting the subset which produced the best classification accuracy by C5.0 and PEBLS. Gain Ratio (GR) is a metric often used in classification systems (notably in C4.5) 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.[3]

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 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:

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

[2] We used k = 3 and majority voting scheme for all experiments in our current work.

[3] 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: Example 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?"

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).

Here in applying GR, there was one issue we had to consider: how to distinguish content words from non-content words. This issue arose from the uneven distribution of the question types in the dataset. Since not all question types were represented in every domain, if we chose question type as the target category, features which yield high GR values might include some domain-specific words. In effect, good predictors for our purpose are words which predict question types very well, but do not predict domains. Therefore, we defined the GR score of a word to be the combination of two values: the GR value when the target category was question type, minus the GR value when the target category was domain.

We computed the (modified) GR score for 1485 words which appeared more than twice in the dataset, and applied C5.0 and PEBLS. Then we gradually reduced the set by taking the top n words according to the GR scores and observed changes in the classification accuracy. Figure 3 shows the result. The evaluation was done by using the 5-fold DCV, and the accuracy percentages indicated in the figure were an average of 3 runs. The best accuracy was achieved by the top 350 words by both algorithms; the remaining words seemed to have caused overfitting as the accuracy showed slight decline. Thus, we took the top 350 words as the initial lexical feature set.

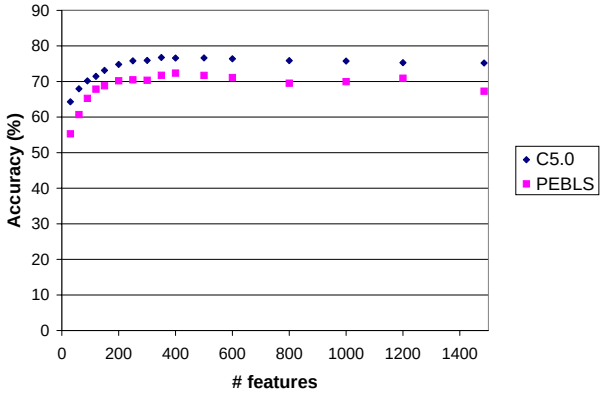

Figure 3: Classification Accuracy (%) on the training data measured by Domain Cross Validation (DCV)

3.2 Initial Semantic Set

The initial semantic set was obtained by automatically selecting some nodes in the WordNet (Miller, 1990) noun and verb trees. For each question type, we chose questions of certain structures and applied a shallow parser to extract nouns and/or verbs which appeared at a specific position. For example, for all question types (except for YNQ), we extracted the head noun from questions of the form "What is NP ..?". Those nouns are essentially the denominalization of the question type. The nouns extracted included "way", "method", "procedure", "process" for the type PRC, "reason", "advantage" for RSN, and "organization", "restaurant" for ENT. For the types DEF and MNR, we also extracted the main verb from questions of the form "How/What does NP V ..?". Such verbs included "work", "mean" for DEF, and "affect" and "form" for MNR.

Then for the nouns and verbs extracted for each question type, we applied the sense disambiguation algorithm used in (Resnik, 1997) and derived semantic classes (or nodes in the WordNet trees) which were their abstract generalization. For each word in a set, we traversed the WordNet tree upward through the hypernym links from the nodes which corresponded to the first two senses of the word, and assigned each ancestor a value which equaled to the inverse of the distance (i.e., the number of links traversed) from the original node. Then we accumulated the values for all ancestors, and selected ones (excluding the top nodes) whose value was above a threshold. For example, the set of nouns extracted for the type PRC were "know-how" (an ancestor of "way" and "method") and "activity" (an ancestor of "procedure" and "process").

By applying the procedure above for all question types, we obtained a total of 112 semantic classes. This constitutes the initial semantic set.

3.3 Refinement

The final feature sets, LEX and LEXSEM, were derived by further refining the initial sets. The main purpose of refinement was to reduce the union of initial lexical and semantic sets (a total of 350 + 112 = 462 features) and derive LEXSEM. It was done by taking the features which appeared in more than half of the decision trees induced by C5.0 during the iterations of DCV.[4] Then we applied the same procedure to the initial lexical set (350 features) and derived LEX. Now both sets were (sub) optimal subsets, with which we could make a fair comparison. There were 117 features/words and 164 features selected for LEX and LEXSEM respectively.

Our refinement method is similar to (Cardie, 1993) in that it selects features by removing ones that did not appear in a decision tree. The difference is that, in our method, each decision tree is induced from a strict subset of the domains of the dataset. Therefore, by taking the intersection of multiple such trees, we can effectively extract features that are domain-independent, thus transferable to other unseen domains. Our method is also computationally less expensive and feasible, given the number of features expected to be in the reduced set (over a hundred by our intuition), than other feature subset selection techniques, most of which require expensive search through model space (such as *wrapper* approach (John et al., 1994)).

Table 2: Classification accuracy (%) on the training set by using reduced feature sets

Feature set	# features	C5.0	PEBLS
Initial lex	350	76.7	71.8
LEX (reduced)	117	77.4	74.5
Initial lex + sem	462	76.7	71.8
LEXSEM (reduced)	164	77.7	74.7

Table 2 shows the classification accuracy measured by DCV for the training set. The increase of the accuracy after the refinement was minimal using C5.0 (from 76.7 to 77.4 for LEX, from 76.7 to 77.7 for LEXSEM), as expected. But the increase using PEBLS was rather significant (from 71.8 to 74.5 for LEX, from 71.8 to 74.7 for LEXSEM). This result agreed with the findings in (Cardie, 1993), and confirmed that LEX and LEXSEM were indeed (sub) optimal. However, the difference between LEX and LEXSEM was not statistically significant by either algorithm (from 77.4 to 77.7 by C5.0, from 74.5 to 74.7 by PEBLS; p-values were .23 and .41 respectively[5]). This means the semantic features did not help improve the classification accuracy.

As we inspected the results, we discovered that, out of the 164 features in LEXSEM, 32 were semantic features, and they did occur in 33% of the training examples (1671/5105 $\approx$ .33). However in most of those examples, key terms were already represented by lexical features, thus semantic features did not add any more information to help determine the question type. As an example, a sentence "What are the dates of the upcoming Jewish holidays?" was represented by lexical features "what", "be", "of" and "date", and a semantic feature "time-unit" (an ancestor of "date").

The 117 words in LEX are listed in the Appendix at the end of this paper.

[4] We have in fact experimented various threshold values. It turned out that .5 produced the best accuracy.

[5] P-values were obtained by applying the t-test on the accuracy produced by all iterations of DCV, with a null hypothesis that the mean accuracy of LEXSEM was higher than that of LEX.

Table 3: Classification accuracy (%) on the testsets

Feature set	# features	FAQFinder		AskJeeves	
		C5.0	PEBLS	C5.0	PEBLS
LEX	117	67.8	66.6	77.3	73.9
LEXSEM	164	67.5	67.1	73.7	71.1

3.4 External Testsets

To further investigate the effect of semantic features, we tested LEX and LEXSEM with two external testsets: one set consisting of 620 questions taken from FAQFinder user log, and another set consisting of 3485 questions taken from the AskJeeves (`http://www.askjeeves.com`) user log. Both datasets contained questions from a wide range of domains, therefore served as an excellent indicator of the domain transferability for our two feature sets.

Table 3 shows the results. For the FAQFinder data, LEX and LEXSEM produced comparable accuracy using both C5.0 and PEBLS. But for the AskJeeves data, LEXSEM did worse than LEX consistently by both classifiers. This means the additional semantic features were interacting with lexical features.

We speculate the reason to be the following. Compared to the FAQFinder data, the AskJeeves data was gathered from a much wider audience, and the questions spanned a broad range of domains. Many terms in the questions were from vocabulary considerably larger than that of our training set. Therefore, the data contained quite a few words whose hypernym links lead to a semantic feature in LEXSEM but did not fall into the question type keyed by the feature. For instance, a question in AskJeeves "What does Hanukah mean?" was mis-classified as type TME by using LEXSEM. This was because "Hanukah" in WordNet was encoded as a hyponym of "time_period". On the other hand, LEX did not include "Hanukah", thus correctly classified the question as type DEF.

4 Related Work

Recently, with a need to incorporate user preferences in information retrieval, several work has been done which classifies documents by genre. For instance, (Finn et al., 2002) used machine learning techniques to identify subjective (opinion) documents from newspaper articles. To determine what feature adapts well to unseen domains, they compared three kinds of features: words, part-of-speech statistics and manually selected meta-linguistic features. They concluded that the part-of-speech performed the best with regard to domain transfer. However, not only were their feature sets pre-determined, their features were distinct from words in the documents (or features were the entire words themselves), thus no feature subset selection was performed.

(Wiebe, 2000) also used machine learning techniques to identify subjective sentences. She focused on adjectives as an indicator of subjectivity, and used corpus statistics and lexical semantic information to derive adjectives that yielded high precision.

5 Conclusions and Future Work

In this paper, we showed that semantic features did not enhance lexical features in the representation of questions for the purpose of question type classification. While semantic features allow for generalization, they also seemed to do more harm than good in some cases by interacting with lexical features. This indicates that question terminology is strongly lexical indeed, and suggests that enumeration of words which appear in typical, idiomatic question phrases would be more effective than semantics.

For future work, we are planning to experiment with synonyms. The use of synonyms is another way of increasing the coverage of question terminology; while semantic features try to achieve it by generalization, synonyms do it by lexical expansion. Our plan is to use the synonyms obtained from very large corpora reported in (Lin, 1998). We are also planning to compare the (lexical and semantic) features we derived automatically in this work with manually selected features. In our previous work, manually selected (lexical) fea-

tures showed slightly better performance for the training data but no significant difference for the test data. We plan to manually pick out semantic as well as lexical features, and apply to the current data.

References

R. Burke, 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).

C. Cardie. 1993. Using decision trees to improve case-based learning. In *Proceedings of the 10th International Conference on Machine Learning (ICML-93)*.

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

A. Finn, N. Kushmerick, and B. Smyth. 2002. Genre classification and domain transfer for information filtering. In *Proceedings of the European Colloquium on Information Retrieval Research*, Glasgow.

S. Harabagiu, 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*.

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

G. John, R. Kohavi, and K. Pfleger. 1994. Irrelevant features and the subset selection problem. In *Proceedings of the 11th International Conference on Machine Learning (ICML-94)*.

K. Kessler, G. Nunberg, and H. Schutze. 1997. Automatic detection of text genre. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL-97)*.

D. Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics (ACL-98)*.

S. Lytinen and N. Tomuro. 2002. The use of question types to match questions in faqfinder. In *Papers from the 2002 AAAI Spring Symposium on Mining Answers from Texts and Knowledge Bases*.

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

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

P. Resnik. 1997. Selectional preference and sense disambiguation. In *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics*, Washington D.C.

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

E. Voorhees. 2000. The trec-9 question answering track report. In *Proceedings of TREC-9*.

J. Wiebe. 2000. Learning subjective adjectives from corpora. In *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-2000)*, Austin, Texas.

Y. Yang 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 LEX Set

"about" "address" "advantage" "affect" "and" "any" "archive" "available" "bag" "be" "begin" "benefit" "better" "buy" "can" "cause" "clean" "come" "company" "compare" "contact" "contagious" "copy" "cost" "create" "date" "day" "deal" "differ" "difference" "do" "effect" "emission" "evaporative" "expense" "fast" "find" "for" "get" "go" "good" "handle" "happen" "have" "history" "how" "if" "in" "internet" "keep" "know" "learn" "long" "make" "many" "mean" "milk" "much" "my" "name" "number" "obtain" "of" "often" "old" "on" "one" "or" "organization" "origin" "people" "percentage" "place" "planet" "price" "procedure" "pronounce" "purpose" "reason" "relate" "relationship" "shall" "shuttle" "site" "size" "sky" "so" "solar" "some" "start" "store" "sun" "symptom" "take" "tank" "tax" "that" "there" "time" "to" "us" "way" "web" "what" "when" "where" "which" "who" "why" "will" "with" "work" "world_wide_web" "wrong" "www" "year" "you"