Paraphrasing with Search Engine Query Logs

Shiqi Zhao[†‡], **Haifeng Wang**[†], **and Ting Liu**[‡]
[†]Baidu Inc.
[‡]HIT Center for Information Retrieval, Harbin Institute of Technology
{zhaoshiqi, wanghaifeng}@baidu.com, tliu@ir.hit.edu.cn

Abstract

This paper proposes a method that extracts paraphrases from search engine query logs. The method first extracts paraphrase query-title pairs based on an assumption that a search query and its corresponding clicked document titles may mean the same thing. It then extracts paraphrase query-query and title-title pairs from the query-title paraphrases with a pivot approach. Paraphrases extracted in each step are validated with a binary classifier. We evaluate the method using a query log from Baidu[1], a Chinese search engine. Experimental results show that the proposed method is effective, which extracts more than 3.5 million pairs of paraphrases with a precision of over 70%. The results also show that the extracted paraphrases can be used to generate high-quality paraphrase patterns.

1 Introduction

The use of paraphrases is ubiquitous in human languages, which also presents a challenge for natural language processing (NLP). Previous studies have shown that paraphrasing can play important roles in plenty of areas, such as machine translation (MT) (Callison-Burch et al., 2006; Kauchak and Barzilay, 2006), question answering (QA) (Duboue and Chu-Carroll, 2006; Riezler et al., 2007), natural language generation (NLG) (Iordanskaja et al., 1991), and so on. As a result, the research on paraphrasing and its applications have attracted significant interest.

[1]www.baidu.com

This paper proposes a method that uses search engine query logs for extracting paraphrases, which is illustrated in Figure 1. Specifically, three kinds of paraphrases can be extracted with our method, which include (1) query-title (Q-T): a query and a document title that users clicked on; (2) query-query (Q-Q): two queries, for which users clicked on the same document title; (3) title-title (T-T): two titles that users clicked on for the same query. We train a classifier for each kind to filter incorrect pairs and refine the paraphrases.

Extracting paraphrases using query logs has many advantages. First, query logs keep growing, which have no scale limitation. Second, query logs reflect web users' real needs, hence the extracted paraphrases may be more useful than that from other kinds of corpora. Third, paraphrases extracted from query logs can be directed applied in search engines for query suggestion and document reranking. In addition, we find that both queries and titles contain a good many question sentences, which can be useful in developing QA systems.

We conduct experiments using a query log of a commercial Chinese search engine Baidu, from which we extracted about 2.7 million pairs of paraphrase Q-T, 0.4 million pairs of paraphrase Q-Q, and 0.4 million pairs of paraphrase T-T. The precision of the paraphrases is above 70%. In addition, we generate paraphrase patterns using the extracted paraphrases. The results show that 73,484 pairs of paraphrase patterns have been generated, with a precision of over 78%.

In the rest of the paper, we first review related work in Section 2. Section 3 describes our method in detail. Section 4 presents the evaluation and re-

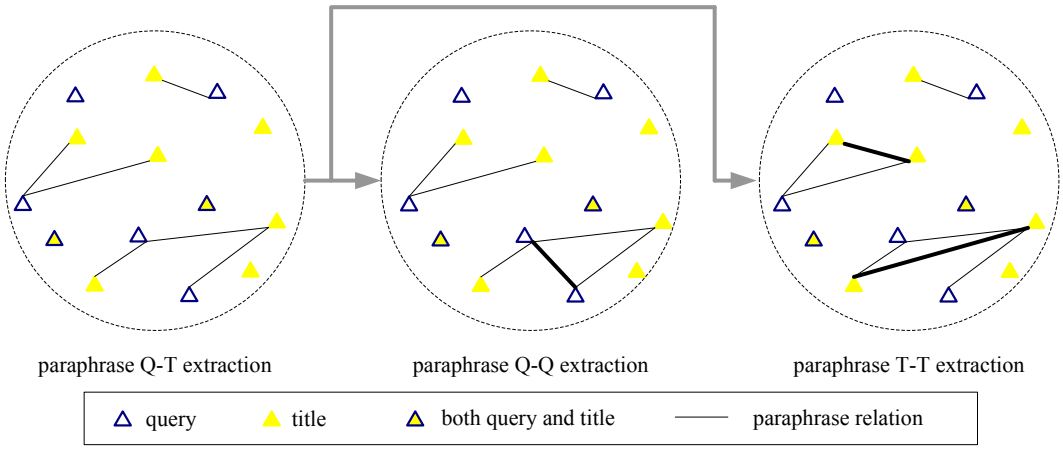

Figure 1: Illustration of the proposed method.

sults. Section 5 concludes the paper and discusses future directions.

2 Related Work

In this section, we briefly review previous studies on paraphrase extraction and query log mining in information retrieval (IR).

2.1 Paraphrase Extraction

A variety of data resources have been exploited for paraphrase extraction. For example, some researchers extract paraphrases from multiple translations of the same foreign novel (Barzilay and McKeown, 2001; Ibrahim et al., 2003), while some others make use of comparable news articles that report on the same event within a small time interval (Shinyama et al., 2002; Barzilay and Lee, 2003; Dolan et al., 2004). Besides the monolingual corpora, bilingual parallel corpora have also been used for extracting paraphrases (Bannard and Callison-Burch, 2005; Callison-Burch, 2008; Zhao et al., 2008). Their basic assumption is that phrases that align with the same foreign phrase may have the same meaning.

The above methods have achieved promising results. However, their performances are usually constrained due to the scale and domain limitation. As an alternative, researchers have tried to acquire paraphrases from large-scale web corpora (Lin and Pantel, 2001; Paşca and Dienes, 2005; Bhagat and Ravichandran, 2008) or directly based on web mining (Ravichandran and Hovy, 2002). These methods are guided by an extended version of distributional hypothesis, namely, if two phrases often occur in similar contexts, their meanings tend to be similar. The disadvantage of these methods is that the underlying assumption does not always hold. Phrases with opposite meanings can also occur in similar contexts, such as "X solves Y" and "X worsens Y" (Lin and Pantel, 2001). In addition, the extracted paraphrases are generally short fragments with two slots (variables) at both ends.

2.2 Query Log Mining in IR

Query logs are widely used in the IR community, especially for mining similar queries. For example, Wen et al. (2002) clustered queries based on user click information. Their basic idea is that if some queries result in similar user clicks, the meanings of these queries should be similar. Such methods have also been investigated in (Gao et al., 2007) for cross-lingual query suggestion and (Zhao et al., 2007) for synonymous questions identification. This paper is partly inspired by their studies. However, we do not simply use click information as clues for mining similar queries. Instead, we mine paraphrases across queries and clicked document titles.

In addition, query logs can be used for query expansion. For instance, Cui et al. (2002) extract probabilistic correlations between query terms and document terms by analyzing query logs, which are then used to select high-quality

H1:	If a query q hits a title t, then q and t are likely to be paraphrases.
H2:	If queries q_1 and q_2 hit the same title t, q_1 and q_2 are likely to be paraphrases.
H3:	If a query q hits titles t_1 and t_2, then t_1 and t_2 are likely to be paraphrases.

Table 1: Hypotheses for extracting paraphrases.

expansion terms for new queries. Note that the expansion terms are merely related terms of the queries, not necessarily paraphrases.

There are other studies that use query logs for constructing ontologies (Sekine and Suzuki, 2007), learning named entities (Paşca, 2007), building user profiles (Richardson, 2008), correcting spelling errors (Ahmad and Kondrak, 2005), and so forth.

3 The Proposed Method

3.1 Basic Idea

Nowadays, more and more users tend to search long queries with search engines. Many users even directly search questions to get exact answers. By analyzing our query log that records rich information including user queries, clicked urls, titles, etc., we find that most titles of clicked documents are highly related with search queries. Especially, paraphrases can be easily found from long queries and the corresponding clicked titles. This motivates us to extract paraphrases from query-title pairs. Here we introduce a concept *hit* that will be frequently used: given a query q, a web document d, and d's title t, if there exist some users that click on d when searching q, then we say q *hits* t.

The hypothesis for extracting paraphrase Q-T is shown in Table 1 (H1). In addition, we find that when several queries hit the same title, the queries are likely to be paraphrases of each other. The other way round, when a query hits several titles, paraphrases can also be found among the titles. We therefore further extract paraphrase Q-Q and T-T from the paraphrase Q-T. The underlying hypotheses can be found in Table 1 (H2 and H3). Note that, based on H2 and H3, paraphrase Q-Q and T-T can be directly extracted from raw Q-T pairs. However, in consideration of precision, we extract them from paraphrase Q-T. We call our paraphrase Q-Q and T-T extraction approach as a pivot approach, since we use titles as *pivots* (queries as *targets*) when extracting paraphrase Q-Q and use queries as *pivots* (titles as *targets*) when extracting paraphrase T-T.

INPUT: $\mathcal{Q}$: query space, $\mathcal{T}$: title space
OUTPUT: P_{qt}: the set of paraphrase Q-T,
P_{qq}: the set of paraphrase Q-Q,
P_{tt}: the set of paraphrase T-T,
$ParaSet$: the set of paraphrases

1. **FOR** any $q \in \mathcal{Q}$ and $t \in \mathcal{T}$
2. **IF** q hits t
3. **IF** $IsParaphrase(q, t)$
4. Add $\langle q, t \rangle$ to P_{qt}
5. **END IF**
6. **END IF**
7. **END FOR**

8. **FOR** any $q_1, q_2 \in \mathcal{Q}$ and $t \in \mathcal{T}$
9. **IF** $\langle q_1, t \rangle \in P_{qt}$ and $\langle q_2, t \rangle \in P_{qt}$
10. **IF** $IsParaphrase(q_1, q_2)$
11. Add $\langle q_1, q_2 \rangle$ to P_{qq}
12. **END IF**
13. **END IF**
14. **END FOR**

15. **FOR** any $t_1, t_2 \in \mathcal{T}$ and $q \in \mathcal{Q}$
16. **IF** $\langle q, t_1 \rangle \in P_{qt}$ and $\langle q, t_2 \rangle \in P_{qt}$
17. **IF** $IsParaphrase(t_1, t_2)$
18. Add $\langle t_1, t_2 \rangle$ to P_{tt}
19. **END IF**
20. **END IF**
21. **END FOR**

22. **RETURN** $ParaSet = P_{qt} \cup P_{qq} \cup P_{tt}$

Table 2: Algorithm for extracting paraphrases.

3.2 Algorithm

Our paraphrase extraction algorithm is shown in Table 2. In particular, lines 1∼7 extract para-

phrase Q-T from the query log. Lines 8∼14 and 15∼21 extract paraphrase Q-Q and T-T, respectively. Line 22 combines the paraphrase Q-T, Q-Q, and T-T together. To filter noise, the extracted Q-T, Q-Q, and T-T pairs are all validated using a function $IsParaphrase(s_1, s_2)$. In this work, we recast paraphrase validation as a binary classification problem. Any pair of $\langle s_1, s_2 \rangle$ is classified as 1 (paraphrase) or 0 (non-paraphrase) with a support vector machine (SVM) classifier. The features used for classification will be detailed in Section 3.3.

In practice, we exploit a query log that contains 287 million Q-T pairs, which are then filtered using the following constraints: (1) exclude Q-T pairs that are too short, i.e., either query q or tittle t contains less than three terms; (2) exclude Q-T pairs where q subsumes t or vice versa, e.g., "牛肉 (beef)" and "牛肉 的 做法 (cooking method of beef)"; (3) exclude Q-T pairs in which the similarity between q and t is below a predefined threshold T[2]; (4) exclude Q-T pairs whose t contains frequent internet terms, such as "主页 (home page)", "网站 (web site)", "在线 (online)", since such titles are mostly organization home pages, online videos, downloadable resources, etc., which are useless for our purpose of paraphrase extraction.

3.3 Features for Paraphrase Validation

Given a pair of candidate paraphrases $\langle s_1, s_2 \rangle$, in which s_1 and s_2 can be either a query or a title, we exploit the following features in the classification-based paraphrase validation.

• **Frequency Feature** F_F. F_F is defined based on each $\langle s_1, s_2 \rangle$'s frequency. We expect that more frequent $\langle s_1, s_2 \rangle$ should be more reliable.

$$F_F(s_1, s_2) = \begin{cases} \frac{c(s_1,s_2)}{C} & if \;\; c(s_1, s_2) < C \\ 1 & if \;\; c(s_1, s_2) \geq C \end{cases} \quad (1)$$

where $c(s_1, s_2)$ denotes the number of times that the $\langle s_1, s_2 \rangle$ pair occurs in the corpus. C is a normalizing factor ($C = 10$ in our experiments).

[2]The similarity is computed based on word overlap rate, which will be described in detail in section 3.3. We set $T = 0.6$ in the experiments.

• **Length Rate Feature** F_{LR}:

$$F_{LR}(s_1, s_2) = \frac{min\{c_w(s_1), c_w(s_2)\}}{max\{c_w(s_1), c_w(s_2)\}} \quad (2)$$

where $c_w(s)$ denotes the number of words in s.

• **Word Overlap Rate Feature** F_{WOR}:

$$F_{WOR}(s_1, s_2) = \frac{c_w(s_1 \cap s_2)}{max\{c_w(s_1), c_w(s_2)\}} \quad (3)$$

where "$s_1 \cap s_2$" is the intersection of s_1 and s_2.

• **Character Overlap Rate Feature** F_{COR}. Chinese words are composed of characters. It is quite often that words with similar characters share similar meanings, such as "爽快 (comfortable)" and "痛快 (comfortable)", "出售 (sell)" and "销售 (sell)". Here we use F_{COR} to measure the similarity between s_1 and s_2 at the character level. Detailedly, we segment s_1 and s_2 into sets of characters and compute the overlap rate based on Equation (3)[3].

• **Cosine Similarity Feature** F_{CS}. In F_{CS}, both s_1 and s_2 are represented as vectors and their cosine similarity is computed as:

$$F_{CS}(s_1, s_2) = \frac{vec_w(s_1) \cdot vec_w(s_2)}{\|vec_w(s_1)\| \times \|vec_w(s_2)\|} \quad (4)$$

where $vec_w(s)$ is the vector of words in s, "$\cdot$" denotes the dot product of two vectors, $\|vec_w(s)\|$ is the norm of a vector. Here, the weight of each word w in a vector is computed using a heuristic similar to tf-idf:

$$W(w) = tf(w) \times log(\frac{N}{c(w)} + 0.1) \quad (5)$$

where $tf(w)$ is the frequency of w in the given s, $c(w)$ is the number of times that w occurs in the corpus, $N = \max_w c(w)$.

• **Edit Distance Feature** F_{ED}. Let $ED(s_1, s_2)$ be the edit distance at the word level between s_1 and s_2, we compute F_{ED} as follows:

$$F_{ED}(s_1, s_2) = 1 - \frac{ED(s_1, s_2)}{max\{c_w(s_1), c_w(s_2)\}} \quad (6)$$

[3]In F_{COR}, $c_w(s)$ of Equation (3) denotes the number of characters in s.

• **Named Entity (NE) Similarity Feature** F_{NE}. NE information is critical in paraphrase identification (Shinyama et al., 2002). We therefore compute the NE similarity between s_1 and s_2 and take it as a feature. We employ a Chinese NE recognition tool that can recognize *person names*, *locations*, *organizations*, and *numerals*. The NE similarity is computed as:

$$F_{NE}(s_1, s_2) = \frac{c_{ne}(s_1 \cap s_2) + 1}{max\{c_{ne}(s_1), c_{ne}(s_2)\} + 1} \quad (7)$$

where $c_{ne}(s)$ denotes the number of NEs in s. Equation (7) guarantees $F_{NE} = 1$ if there are no NEs in either s_1 or s_2.

• **Pivot Fertility Feature** F_{PF}**:** F_{PF} is a feature specially designed for paraphrase Q-Q and T-T extraction, which are based on the pivot approach[4]. Specifically, we define *fertility* of a pivot as the number of targets it corresponds to. Our observation indicates that the larger the fertility of a pivot is, the more noisy the targets are. Hence we define F_{PF} as:

$$F_{PF}(s_1, s_2) = \max_p \frac{1}{f(p)} \quad (8)$$

where $s_1 = q_1$, $s_2 = q_2$, $p = t$ when classifying Q-Q, while $s_1 = t_1$, $s_2 = t_2$, $p = q$ when classifying T-T. $f(p)$ denotes the fertility of the pivot p. The value is maximized over p if s_1 and s_2 can be extracted with multiple pivots.

3.4 Generating Paraphrase Patterns

A key feature of our method is that the extracted paraphrases are particularly suitable for generating paraphrase patterns, especially for the hot domains that are frequently searched. For example, there are quite a few paraphrases concerning the therapy of various diseases, from which we can easily induce patterns expressing the meaning of "How to treat [X] disease", such as "[X] 病 如何 治疗", "怎么 治疗 [X] 病", and "[X] 病 的 治疗 方法". Therefore, in this work, we try to generate paraphrase patterns using the extracted paraphrases.

In our preliminary experiments, we only induce paraphrase patterns from paraphrases that contain no more than 6 words. In addition, only one slot is allowed in each pair of paraphrase patterns. Let s_1 and s_2 be a pair of paraphrases extracted above. If there exist words $w \in s_1$ and $v \in s_2$ that satisfy (1) $w = v$, (2) w and v are not stop words, then we can induce a pair of paraphrase patterns by replacing w in s_1 and v in s_2 with a slot "[X]". It is obvious that several pairs of paraphrase patterns may be induced from one pair of paraphrases.

[4] F_{PF} is not used in paraphrase Q-T validation.

	SAME	**RELA**	**DIFF**
percent (%)	55.92	44.08	-

Table 3: Human labeling of candidate Q-T.

4 Experiments

We experiment with a query log that contains a total of 284,316,659 queries. Statistics reveal that 170,315,807 queries (59.90%) lead to at least one user click, each having 1.69 clicks on average. We extract 287,129,850 raw Q-T pairs using the query log, from which 4,448,347 pairs of candidate Q-T are left after filtering as described in Section 3.2. Almost all queries and titles are written in Chinese, though some of them contain English or Japanese words. The preprocessing of candidate Q-T includes Chinese word segmentation (WSeg) and NE recognition (NER). Our WSeg tool is implemented based on forward maximum matching, while the NER tool is based on a NE dictionary mined from the web.

4.1 Evaluation of Candidate Q-T

We first evaluate candidate Q-T without validation. To this end, we randomly sampled 5000 pairs of candidate Q-T and labeled them manually. Each pair is labeled into one of the 3 classes: SAME - q and t have the same meaning; RELA - q and t have related meanings; DIFF - q and t have clearly different meanings. The labeling results are listed in Table 3. We can see that no candidate Q-T is in the DIFF class. This is not surprising, since users are unlikely to click on web pages unrelated to their queries.

To gain a better insight into the data, we analyzed the subtle types of candidate Q-T in both SAME and RELA classes. In detail, we sampled

1000 pairs of candidate Q-T from the 5000 pairs labeled above, in which 563 are in the SAME class, while the other 437 are in the RELA class. Our analysis suggests that candidate Q-T in the SAME class can be divided into 4 subtle types:

- Trivial change (12.61%): changes of punctuation or stop words, such as "考研 失败 怎么 办" and "考研 失败 怎么 办 ？ ".
- Word or phrase replacement (68.38%): replacements of synonymous words or phrases, such as "咖啡 斑 的 治疗 多少 钱 (how mach is ...)" and "咖啡 斑 的 治疗 费用 是 多少 (what is the price of ...)".
- Structure change (7.10%): changes of both words and word orders, such as "减肥 中 水果 可以 吃 什么 (what fruit can I eat on a diet)" and "吃 什么 水果 可以 瘦身 (what fruit can help loss weight)".
- Others (11.90%): candidate Q-T that cannot be classified into the 3 types above.

The above analysis reveals that more than two thirds of candidate Q-T in the SAME class are in the "word or phrase replacement" type, while the ones with structure changes are slightly more than 7%. We believe this is mainly because queries and titles are relatively short and their structures are simple. Thus structure rewriting can hardly be conducted. This distribution is in line with that reported in (Zhao et al., 2008).

As for the RELA class, we find that 42.33% of such candidate Q-T share a problem of named entity mismatch, such as "美国 (US) 大型 水利 工程" and "中国 (China) 急需 大型 水利 工程". This indicates that the NE similarity feature is necessary in paraphrase validation.

4.2 Evaluation of Paraphrase Q-T

The candidate Q-T extracted above are classified with a SVM classifier[5] under its default setting. To evaluate the classifier, we run 5-fold cross validation with the 5000 human annotated data, in which we use 4000 for training and the rest 1000 for testing in each run. The evaluation criteria are precision (P), recall (R), and f-measure (F), which are defined as follows:

$$P = \frac{\|S_a \cap S_m\|}{\|S_a\|} \tag{9}$$

$$R = \frac{\|S_a \cap S_m\|}{\|S_m\|} \tag{10}$$

$$F = \frac{2 \times P \times R}{P + R} \tag{11}$$

where S_a is the set of paraphrases automatically recognized with the classifier, S_m is the set of paraphrases manually annotated. Precision, recall, and f-measure are averaged over 5 runs in the 5-fold cross validation.

Figure 2 (a) shows the classification results (dark bars). For comparison, we also show the precision, recall[6], and f-measure of the candidate Q-T (light bars). As can be seen, the precision is improved from 0.5592 to 0.7444 after classification. F-measure is also evidently enhanced. This result indicates that the classification-based paraphrase validation is effective. We then use all of the 5000 annotated data to train a classifier and classify all the candidate Q-T. Results show that 2,762,291 out of 4,448,347 pairs of candidate Q-T are classified as paraphrases.

4.3 Evaluation of Paraphrase Q-Q and T-T

From the paraphrase Q-T, we further extracted 934,758 pairs of candidate Q-Q and 438,954 pairs of candidate T-T (without validation). We randomly sampled 5000 from each for human annotation. The results show that the precisions of candidate Q-Q and T-T are 0.4672 and 0.6860, respectively. As can be seen, the precision of candidate Q-Q is much lower than that of candidate T-T. Our analysis reveals that it is mainly because candidate Q-Q are more noisy, since user queries contain quite a lot of spelling mistakes and informal expressions.

The candidate Q-Q and T-T are also refined based on classification. We first evaluate the classification performance using the 5000 human labeled data. The experimental setups for Q-Q and

[5]We use libsvm-2.82 toolkit, which can be downloaded from http://www.csie.ntu.edu.tw/ cjlin/libsvm/

[6]We assume all possible paraphrases are included in the candidates, thus its recall is 100%.

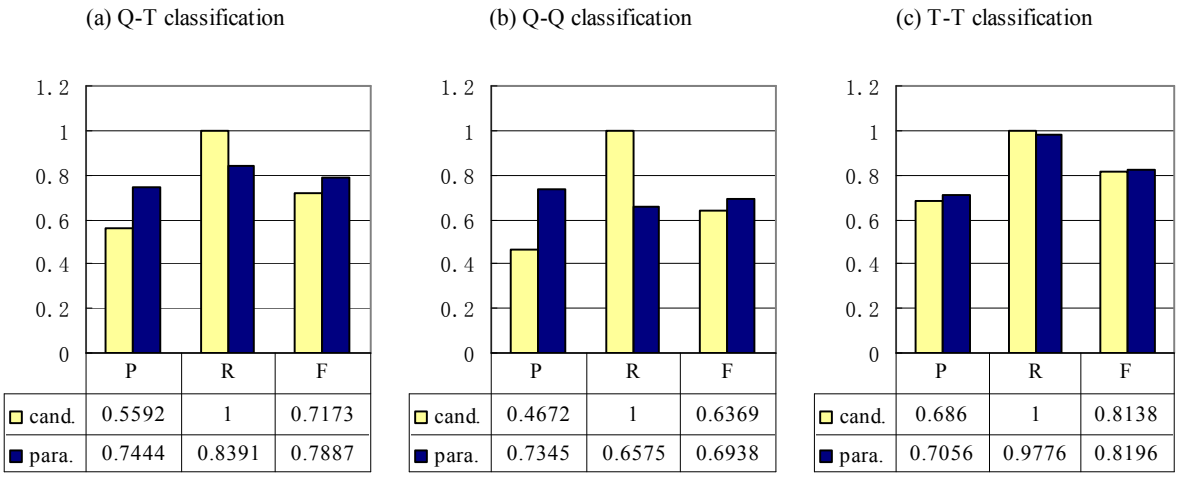

Figure 2: Classification precision (P), recall (R), and f-measure (F).

T-T classification are the same as that of Q-T classification, in which we run 5-fold cross validation with a SVM classifier using its default parameters. Figure 2 (b) and (c) give the classification results (dark bars) as well as the precision, recall, and f-measure of the candidates (light bars).

We can see that the precision of Q-Q is significantly enhanced from 0.4672 to 0.7345 after classification, which suggests that a substantial part of errors and noise are removed. The increase of f-measure demonstrates the effectiveness of classification despite the decrease of recall. Meanwhile, the quality of candidate T-T is not clearly improved after classification. The reason should be that the precision of candidate T-T is already pretty high. We then use all 5000 human labeled data to train a classifier for Q-Q and T-T respectively and classify all candidate Q-Q and T-T. Results show that 390,920 pairs of paraphrase Q-Q and 415,539 pairs of paraphrase T-T are extracted after classification.

4.4 Evaluation of Paraphrase Patterns

Using the method introduced in Section 3.4, we have generated 73,484 pairs of paraphrase patterns that appear at least two times in the corpus. We randomly selected 500 pairs and labeled them manually. The results show that the precision is 78.4%. Two examples are shown in Table 4, in which p_1 and p_2 are paraphrase patterns. Some slot fillers are also listed below. We realize that the method currently used for inducing paraphrase patterns is simple. Hence we will improve the method in our following experiments. Specifically, multiple slots will be allowed in a pair of patterns. In addition, we will try to apply the alignment techniques in the generation of paraphrase patterns, as Zhao et al. (2008) did.

p_1	[X] 文件 怎么 打开
p_2	如何 打开 [X] 文件
	(how to open [X] file)
slot	7z; ashx; aspx; bib; cda; cdfs; cmp; cpi; csf; csv; cur; dat; dek...
p_1	关于 [X] 的 诗词
p_2	有关 [X] 的 诗歌
	(poems about [X])
slot	草原 (prairies); 长江 (Yangtze River); 泰山 (Mount Tai); 乡愁 (nostalgia)...

Table 4: Examples of paraphrase patterns.

4.5 Analysis

Feature Contribution. To investigate the contributions of different features used in classification, we tried different feature combinations for each of our three classifiers. The results are shown in Table 5, in which "+" means the feature has contribution to the corresponding classifier. As can be seen, the character overlap rate feature (F_{COR}), cosine similarity feature (F_{CS}), and NE similarity

Feature	Q-T	Q-Q	T-T
F_F	+		
F_{LR}		+	
F_{WOR}			
F_{COR}	+	+	+
F_{CS}	+	+	+
F_{ED}		+	
F_{NE}	+	+	+
F_{PF}		+	

Table 5: Feature contribution.

feature (F_{NE}) are the most useful, which play important roles in all the three classifiers. The other features are useful in some of the classifiers except the word overlap rate feature (F_{WOR}). The classification results reported in prior sections are all achieved with the optimal feature combination.

Analysis of the Paraphrases. We combine the extracted paraphrase Q-T, Q-Q and T-T and get a total of 3,560,257 pairs of unique paraphrases. Statistics show that only 8380 pairs (0.24%) are from more than one source, which indicates that the intersection among the three sets is very small. Further statistics show that the average length of the queries and titles in the paraphrases is 6.69 (words).

To have a detailed analysis of the extracted paraphrases, we randomly selected 1000 pairs and manually labeled the precision, types, and domains. It is found that more than 43% of the paraphrases are paraphrase questions, in which *how* (36%), *what* (19%), and *yes/no* (14%) questions are the most common. In addition, we find that the precision of paraphrase questions (84.26%) is evidently higher than non-question paraphrases (65.14%). Those paraphrase questions are useful in question analysis and expansion in QA, which can hardly be extracted from other kinds of corpora.

As expected, the paraphrases we extract cover a variety of domains. However, around 50% of them are in the 7 most popular domains[7], including: (1) health and medicine, (2) documentary download, (3) entertainment, (4) software, (5) education and study, (6) computer game, (7) economy and finance. This analysis reflects what web users are most concerned about. These domains, especially (4) and (6), are not well covered by the parallel and comparable corpora previously used for paraphrase extraction.

[7]Note that pornographic queries have been filtered from the query log beforehand.

5 Conclusions and Future Directions

In this paper, we put forward a novel method that extracts paraphrases from search engine query logs. Our contribution is that we, for the first time, propose to extract paraphrases from user queries and the corresponding clicked document titles. Specifically, three kinds of paraphrases are extracted, which can be (1) a query and a hit title, (2) two queries that hit the same title, and (3) two titles hit by the same query. The extracted paraphrases are refined based on classification. Using the proposed method, we extracted over 3.5 million pairs of paraphrases from a query log of Baidu. Human evaluation results show that the precision of the paraphrases is above 70%. The results also show that we can generate high-quality paraphrase patterns from the extracted paraphrases.

Our future research will be conducted along the following directions. Firstly, we will use a much larger query log for paraphrase extraction, so as to enhance the coverage of paraphrases. Secondly, we plan to have a deeper study of the transitivity of paraphrasing. Simply speaking, we want to find out whether we can extract $\langle s_1, s_3 \rangle$ as paraphrases given that $\langle s_1, s_2 \rangle$ and $\langle s_2, s_3 \rangle$ are paraphrases.

6 Acknowledgments

We would like to thank Wanxiang Che, Hua Wu, and the anonymous reviewers for their useful comments on this paper.

References

Farooq Ahmad and Grzegorz Kondrak. 2005. Learning a Spelling Error Model from Search Query Logs. In *Proceedings of HLT/EMNLP*, pages 955-962.

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with Bilingual Parallel Corpora. In *Proceedings of ACL*, pages 597-604.

Regina Barzilay and Lillian Lee. 2003. Learning to Paraphrase: An Unsupervised Approach Using Multiple-Sequence Alignment. In *Proceedings of HLT-NAACL*, pages 16-23.

Regina Barzilay and Kathleen R. McKeown. 2001. Extracting Paraphrases from a Parallel Corpus. In *Proceedings of ACL/EACL*, pages 50-57.

Rahul Bhagat and Deepak Ravichandran. 2008. Large Scale Acquisition of Paraphrases for Learning Surface Patterns. In *Proceedings of ACL-08: HLT*, pages 674-682.

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006. Improved Statistical Machine Translation Using Paraphrases. In *Proceedings of HLT-NAACL*, pages 17-24.

Chris Callison-Burch. 2008. Syntactic Constraints on Paraphrases Extracted from Parallel Corpora. In *Proceedings of EMNLP*, pages 196-205.

Hang Cui, Ji-Rong Wen, Jian-Yun Nie, Wei-Ying Ma. 2002. Probabilistic Query Expansion Using Query Logs In *Proceedings of WWW*, pages 325-332.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources. In *Proceedings of COLING*, pages 350-356.

Pablo Ariel Duboue and Jennifer Chu-Carroll. 2006. Answering the Question You Wish They Had Asked: The Impact of Paraphrasing for Question Answering. In *Proceedings of HLT-NAACL*, pages 33-36.

Wei Gao, Cheng Niu, Jian-Yun Nie, Ming Zhou, Jian Hu, Kam-Fai Wong, and Hsiao-Wuen Hon. 2007. Cross-Lingual Query Suggestion Using Query Logs of Different Languages. In *Proceedings of SIGIR*, pages 463-470.

Ali Ibrahim, Boris Katz, Jimmy Lin. 2003. Extracting Structural Paraphrases from Aligned Monolingual Corpora. In *Proceedings of IWP*, pages 57-64.

Lidija Iordanskaja, Richard Kittredge, and Alain Polguère. 1991. Lexical Selection and Paraphrase in a Meaning-Text Generation Model. In Cécile L. Paris, William R. Swartout, and William C. Mann (Eds.): Natural Language Generation in Artificial Intelligence and Computational Linguistics, pages 293-312.

David Kauchak and Regina Barzilay. 2006. Paraphrasing for Automatic Evaluation. In *Proceedings of HLT-NAACL*, pages 455-462.

De-Kang Lin and Patrick Pantel. 2001. Discovery of Inference Rules for Question Answering. In *Natural Language Engineering* 7(4): 343-360.

Marius Paşca and Péter Dienes. 2005. Aligning Needles in a Haystack: Paraphrase Acquisition Across the Web. In *Proceedings of IJCNLP*, pages 119-130.

Marius Paşca. 2007. Weakly-supervised Discovery of Named Entities using Web Search Queries. In *Proceedings of CIKM*, pages 683-690.

Deepak Ravichandran and Eduard Hovy. 2002. Learning Surface Text Patterns for a Question Answering System. In *Proceedings of ACL*, pages 41-47.

Matthew Richardson. 2008. Learning about the World through Long-Term Query Logs. In *ACM Transactions on the Web* 2(4): 1-27.

Stefan Riezler, Alexander Vasserman, Ioannis Tsochantaridis, Vibhu Mittal and Yi Liu. 2007. Statistical Machine Translation for Query Expansion in Answer Retrieval. In *Proceedings of ACL*, pages 464-471.

Satoshi Sekine and Hisami Suzuki. 2007. Acquiring Ontological Knowledge from Query Logs. In *Proceedings of WWW*, pages 1223-1224.

Yusuke Shinyama, Satoshi Sekine, and Kiyoshi Sudo. 2002. Automatic Paraphrase Acquisition from News Articles. In *Proceedings of HLT*, pages 40-46.

Ji-Rong Wen, Jian-Yun Nie, and Hong-Jiang Zhang. 2002. Query Clustering Using User Logs. In *ACM Transactions on Information Systems* 20(1): 59-81, 2002.

Shiqi Zhao, Haifeng Wang, Ting Liu, and Sheng Li. 2008. Pivot Approach for Extracting Paraphrase Patterns from Bilingual Corpora. In *Proceedings of ACL-08:HLT*, pages 780-788.

Shiqi Zhao, Ming Zhou, and Ting Liu. 2007. Learning Question Paraphrases for QA from Encarta Logs. In *Proceedings of IJCAI*, pages 1795-1800.