7-2007

Cross-lingual query suggestion using query logs of different languages

Wei GAO
Singapore Management University, weigao@smu.edu.sg

Cheng NIU

Jian-Yun NIE

Ming ZHOU

Jian HU

See next page for additional authors

Author

Wei GAO, Cheng NIU, Jian-Yun NIE, Ming ZHOU, Jian HU, Kam-Fai WONG, and Hsiao-Wuen HON

Cross-Lingual Query Suggestion Using Query Logs of Different Languages

Wei Gao[1*], Cheng Niu[2], Jian-Yun Nie[3], Ming Zhou[2], Jian Hu[2], Kam-Fai Wong[1],
Hsiao-Wuen Hon[2]

[1]The Chinese University of Hong Kong, Hong Kong, China

{wgao, kfwong}@se.cuhk.edu.hk

[2]Microsoft Research Asia, Beijing, China

{chengniu, mingzhou, jianh, hon}@microsoft.com

[3]Université de Montréal, Montréal, QC, Canada

nie@iro.umontreal.ca

ABSTRACT

Query suggestion aims to suggest relevant queries for a given query, which help users better specify their information needs. Previously, the suggested terms are mostly in the same language of the input query. In this paper, we extend it to cross-lingual query suggestion (CLQS): for a query in one language, we suggest similar or relevant queries in other languages. This is very important to scenarios of cross-language information retrieval (CLIR) and cross-lingual keyword bidding for search engine advertisement. Instead of relying on existing query translation technologies for CLQS, we present an effective means to map the input query of one language to queries of the other language in the query log. Important monolingual and cross-lingual information such as word translation relations and word co-occurrence statistics, etc. are used to estimate the cross-lingual query similarity with a discriminative model. Benchmarks show that the resulting CLQS system significantly outperforms a baseline system based on dictionary-based query translation. Besides, the resulting CLQS is tested with French to English CLIR tasks on TREC collections. The results demonstrate higher effectiveness than the traditional query translation methods.

Categories and Subject Descriptors

H.3.3 [**Information storage and retrieval**]: Information Search and Retrieval – Query formulation

General Terms

Algorithms, Performance, Experimentation, Theory.

Keywords

cross-language information retrieval, query logs, query translation, query suggestion, query expansion

1. INTRODUCTION

Query suggestion is a functionality to help users of a search engine to better specify their information need, by narrowing down or expanding the scope of the search with synonymous queries and relevant queries, or by suggesting related queries that have been frequently used by other users. Search engines, such as *Google, Yahoo!, MSN, Ask Jeeves,* all have implemented query suggestion functionality as a valuable addition to their core search method. In addition, the same technology has been leveraged to recommend bidding terms to online advertiser in the pay-for-performance search market [12].

Query suggestion is closely related to query expansion which extends the original query with new search terms to narrow the scope of the search. But different from query expansion, query suggestion aims to suggest full queries that have been formulated by users so that the query integrity and coherence are preserved in the suggested queries.

Typical methods for query suggestion exploit query logs and document collections, by assuming that in the same period of time, many users share the same or similar interests, which can be expressed in different manners [12, 14, 26]. By suggesting the related and frequently used formulations, it is hoped that the new query can cover more relevant documents. However, all of the existing studies dealt with monolingual query suggestion and to our knowledge, there is no published study on cross-lingual query suggestion (CLQS). CLQS aims to suggest related queries but in a different language. It has wide applications on World Wide Web: for cross-language search or for suggesting relevant bidding terms in a different language.

CLQS can be approached as a query translation problem, i.e., to suggest the queries that are translations of the original query. Dictionaries, large size of parallel corpora and existing commercial machine translation systems can be used for translation. However, these kinds of approaches usually rely on static knowledge and data. It cannot effectively reflect the quickly shifting interests of Web users. Moreover, there are some problems with translated queries in target language. For instance,

---

the translated terms can be reasonable translations, but they are not popularly used in the target language. For example, the French query "aliment biologique" is translated into "biologic food" by Google translation tool[2], yet the correct formulation nowadays should be "organic food". Therefore, there exist many mismatch cases between the translated terms and the really used terms in target language. This mismatch makes the suggested terms in the target language ineffective.

A natural thinking of solving this mismatch is to map the queries in the source language and the queries in the target language, by using the query log of a search engine. We exploit the fact that the users of search engines in the same period of time have similar interests, and they submit queries on similar topics in different languages. As a result, a query written in a source language likely has an equivalent in a query log in the target language. In particular, if the user intends to perform CLIR, then original query is even more likely to have its correspondent included in the target language query log. Therefore, if a candidate for CLQS appears often in the query log, then it is more likely the appropriate one to be suggested.

In this paper, we propose a method of calculating the similarity between source language query and the target language query by exploiting, in addition to the translation information, a wide spectrum of bilingual and monolingual information, such as term co-occurrences, query logs with click-through data, etc. A discriminative model is used to learn the cross-lingual query similarity based on a set of manually translated queries. The model is trained by optimizing the cross-lingual similarity to best fit the monolingual similarity between one query and the other query's translation. Besides being benchmarked as an independent module, the resulting CLQS system is tested as a new means of query "translation" in CLIR task on TREC collections. The results show that this new "translation" method is more effective than the traditional query translation method.

The remainder of this paper is organized as follows: Section 2 introduces the related work; Section 3 describes in detail the discriminative model for estimating cross-lingual query similarity; Section 4 presents a new CLIR approach using cross-lingual query suggestion as a bridge across language boundaries. Section 5 discusses the experiments and benchmarks; finally, the paper is concluded in Section 6.

2. RELATED WORK

Most approaches to CLIR perform a query translation followed by a monolingual IR. Typically, queries are translated either using a bilingual dictionary [22], a machine translation software [9] or a parallel corpus [20].

Despite the various types of resources used, out-of-vocabulary (OOV) words and translation disambiguation are the two major bottlenecks for CLIR [20]. In [7, 27], OOV term translations are mined from the Web using a search engine. In [17], bilingual knowledge is acquired based on anchor text analysis. In addition, word co-occurrence statistics in the target language has been leveraged for translation disambiguation [3, 10, 11, 19].

---

[2] http://www.google.com/language_tools

Nevertheless, it is arguable that accurate query translation may not be necessary for CLIR. Indeed, in many cases, it is helpful to introduce words even if they are not direct translations of any query word, but are closely related to the meaning of the query. This observation has led to the development of cross-lingual query expansion (CLQE) techniques [2, 16, 18]. [2] reports the enhancement on CLIR by post-translation expansion. [16] develops a cross-lingual relevancy model by leveraging the cross-lingual co-occurrence statistics in parallel texts. [18] makes performance comparison on multiple CLQE techniques, including pre-translation expansion and post-translation expansion. However, there is lack of a unified framework to combine the wide spectrum of resources and recent advances of mining techniques for CLQE.

CLQS is different from CLQE in that it aims to suggest full queries that have been formulated by users in another language. As CLQS exploits up-to-date query logs, it is expected that for most user queries, we can find common formulations on these topics in the query log in the target language. Therefore, CLQS also plays a role of adapting the original query formulation to the common formulations of similar topics in the target language.

Query logs have been successfully used for monolingual IR [8, 12, 15, 26], especially in monolingual query suggestions [12] and relating the semantically relevant terms for query expansion [8, 15]. In [1], the target language query log has been exploited to help query translation in CLIR.

3. ESTIMATING CROSS-LINGUAL QUERY SIMILARITY

A search engine has a query log containing user queries in different languages within a certain period of time. In addition to query terms, click-through information is also recorded. Therefore, we know which documents have been selected by users for each query. Given a query in the source language, our CLQS task is to determine one or several similar queries in the target language from the query log.

The key problem with cross-lingual query suggestion is how to learn a similarity measure between two queries in different languages. Although various statistical similarity measures have been studied for monolingual terms [8, 26], most of them are based on term co-occurrence statistics, and can hardly be applied directly in cross-lingual settings.

In order to define a similarity measure across languages, one has to use at least one translation tool or resource. So the measure is based on both translation relation and monolingual similarity. In this paper, as our purpose is to provide up-to-date query similarity measure, it may not be sufficient to use only a static translation resource. Therefore, we also integrate a method to mine possible translations on the Web. This method is particularly useful for dealing with OOV terms.

Given a set of resources of different natures, the next question is how to integrate them in a principled manner. In this paper, we propose a discriminative model to learn the appropriate similarity measure. The principle is as follows: we assume that we have a reasonable monolingual query similarity measure. For any training query example for which a translation exists, its similarity measure (with any other query) is transposed to its translation.

Therefore, we have the desired cross-language similarity value for this example. Then we use a discriminative model to learn the cross-language similarity function which fits the best these examples.

In the following sections, let us first describe the detail of the discriminative model for cross-lingual query similarity estimation. Then we introduce all the features (monolingual and cross-lingual information) that we will use in the discriminative model.

3.1 Discriminative Model for Estimating Cross-Lingual Query Similarity

In this section, we propose a discriminative model to learn cross-lingual query similarities in a principled manner. The principle is as follows: for a reasonable monolingual query similarity between two queries, a cross-lingual correspondent can be deduced between one query and another query's translation. In other words, for a pair of queries in different languages, their cross-lingual similarity should fit the monolingual similarity between one query and the other query's translation. For example, the similarity between French query "pages jaunes" (i.e., "yellow page" in English) and English query "telephone directory" should be equal to the monolingual similarity between the translation of the French query "yellow page" and "telephone directory". There are many ways to obtain a monolingual similarity measure between terms, e.g., term co-occurrence based mutual information and χ^2. Any of them can be used as the target for the cross-lingual similarity function to fit. In this way, cross-lingual query similarity estimation is formulated as a regression task as follows:

Given a source language query q_f, a target language query q_e, and a monolingual query similarity sim_{ML}, the corresponding cross-lingual query similarity sim_{CL} is defined as follows:

$$sim_{CL}(q_f, q_e) = sim_{ML}(T_{q_f}, q_e) \qquad (1)$$

where T_{q_f} is the translation of q_f in the target language.

Based on Equation (1), it would be relatively easy to create a training corpus. All it requires is a list of query translations. Then an existing monolingual query suggestion system can be used to automatically produce similar query to each translation, and create the training corpus for cross-lingual similarity estimation. Another advantage is that it is fairly easy to make use of arbitrary information sources within a discriminative modeling framework to achieve optimal performance.

In this paper, support vector machine (SVM) regression algorithm [25] is used to learn the cross-lingual term similarity function. Given a vector of feature functions f between q_f and q_e, $sim_{CL}(t_f, t_e)$ is represented as an inner product between a weight vector and the feature vector in a kernel space as follows:

$$sim_{CL}(t_f, t_e) = w \bullet \phi(f(t_f, t_e)) \qquad (2)$$

where ϕ is the mapping from the input feature space onto the kernel space, and w is the weight vector in the kernel space which will be learned by the SVM regression training. Once the weight

vector is learned, the Equation (2) can be used to estimate the similarity between queries of different languages.

We want to point out that instead of regression, one can definitely simplify the task as a binary or ordinal classification, in which case CLQS can be categorized according to discontinuous class labels, e.g., *relevant* and *irrelevant*, or a series of levels of relevancies, e.g., *strongly relevant, weakly relevant,* and *irrelevant*. In either case, one can resort to discriminative classification approaches, such as an SVM or maximum entropy model, in a straightforward way. However, the regression formalism enables us to fully rank the suggested queries based on the similarity score given by Equation (1).

The Equations (1) and (2) construct a regression model for cross-lingual query similarity estimation. In the following sections, the monolingual query similarity measure (see Section 3.2) and the feature functions used for SVM regression (see Section 3.3) will be presented.

3.2 Monolingual Query Similarity Measure Based on Click-through Information

Any monolingual term similarity measure can be used as the regression target. In this paper, we select the monolingual query similarity measure presented in [26] which reports good performance by using search users' click-through information in query logs. The reason to choose this monolingual similarity is that it is defined in a similar context as ours − according to a user log that reflects users' intention and behavior. Therefore, we can expect that the cross-language term similarity learned from it can also reflect users' intention and expectation.

Following [26], our monolingual query similarity is defined by combining both query content-based similarity and click-through commonality in the query log.

First the content similarity between two queries p and q is defined as follows:

$$similarity_{content}(p, q) = \frac{KN(p, q)}{Max(kn(p), kn(q))} \qquad (3)$$

where $kn(x)$ is the number of keywords in a query x, $KN(p,q)$ is the number of common keywords in the two queries.

Secondly, the click-through based similarity is defined as follows,

$$similarity_{click-through}(p, q) = \frac{RD(p, q)}{Max(rd(p), rd(q))} \qquad (4)$$

where $rd(x)$ is the number of clicked URLs for a query x, and $RD(p,q)$ is the number of common URLs clicked for two queries.

Finally, the similarity between two queries is a linear combination of the content-based and click-through-based similarities, and is presented as follows:

$$similarity(p, q) = \alpha * similarity_{content}(p, q) + \beta * similarity_{click-through}(p, q) \qquad (5)$$

where α and β are the relative importance of the two similarity measures. In this paper, we set $\alpha = 0.4$, and $\beta = 0.6$ following the

practice in [26]. Queries with similarity measure higher than a threshold with another query will be regarded as relevant monolingual query suggestions (MLQS) for the latter. In this paper, the threshold is set as 0.9 empirically.

3.3 Features Used for Learning Cross-Lingual Query Similarity Measure

This section presents the extraction of candidate relevant queries from the log with the assistance of various monolingual and bilingual resources. Meanwhile, feature functions over source query and the cross-lingual relevant candidates are defined. Some of the resources being used here, such as bilingual lexicon and parallel corpora, were for query translation in previous work. But note that we employ them here as an assistant means for finding relevant candidates in the log rather than for acquiring accurate translations.

3.3.1 Bilingual Dictionary

In this subsection, a built-in-house bilingual dictionary containing 120,000 unique entries is used to retrieve candidate queries. Since multiple translations may be associated with each source word, co-occurrence based translation disambiguation is performed [3, 10]. The process is presented as follows:

Given an input query $q_f = \{w_{f1}, w_{f2}, \ldots w_{fn}\}$ in the source language, for each query term w_{fi}, a set of unique translations are provided by the bilingual dictionary D : $D(w_{fi}) = \{t_{i1}, t_{i2}, \ldots, t_{im}\}$. Then the cohesion between the translations of two query terms is measured using mutual information which is computed as follows:

$$MI(t_{ij}, t_{kl}) = P(t_{ij}, t_{kl}) \log \frac{P(t_{ij}, t_{kl})}{P(t_{ij})P(t_{kl})} \qquad (6)$$

where $P(t_{ij}, t_{kl}) = \frac{C(t_{ij}, t_{kl})}{N}$, $P(t) = \frac{C(t)}{N}$.

Here $C(x, y)$ is the number of queries in the log containing both x and y, $C(x)$ is the number of queries containing term x, and N is the total number of queries in the log.

Based on the term-term cohesion defined in Equation (6), all the possible query translations are ranked using the summation of the term-term cohesion $S_{dict}(T_{q_f}) = \sum_{i,k,i \neq k} MI(t_{ij}, t_{kl})$. The set of top-4 query translations is denoted as $S(T_{q_f})$. For each possible query translation $T \in S(T_{q_f})$, we retrieve all the queries containing the same keywords as T from the target language log. The retrieved queries are candidate target queries, and are assigned $S_{dict}(T)$ as the value of the feature *Dictionary-based Translation Score*.

3.3.2 Parallel Corpora

Parallel corpora are precious resources for bilingual knowledge acquisition. Different from the bilingual dictionary, the bilingual knowledge learned from parallel corpora assigns probability for each translation candidate which is useful in acquiring dominant query translations.

In this paper, the Europarl corpus (a set of parallel French and English texts from the proceedings of the European Parliament) is used. The corpus is first sentence aligned. Then word alignments are derived by training an IBM translation model 1 [4] using GIZA++ [21]. The learned bilingual knowledge is used to extract candidate queries from the query log. The process is presented as follows:

Given a pair of queries, q_f in the source language and q_e in the target language, the *Bi-Directional Translation Score* is defined as follows:

$$S_{IBM1}(q_f, q_e) = \sqrt{p_{IBM1}(q_f \mid q_e) p_{IBM1}(q_e \mid q_f)} \qquad (7)$$

where $p_{IBM1}(y \mid x)$ is the word sequence translation probability given by IBM model 1 which has the following form:

$$p_{IBM1}(y \mid x) = \frac{1}{(\mid x \mid + 1)^{\mid y \mid}} \prod_{j=1}^{\mid y \mid} \sum_{i=0}^{\mid x \mid} p(y_j \mid x_i) \qquad (8)$$

where $p(y_j \mid x_i)$ is the word to word translation probability derived from the word-aligned corpora.

The reason to use bidirectional translation probability is to deal with the fact that common words can be considered as possible translations of many words. By using bidirectional translation, we test whether the translation words can be translated back to the source words. This is helpful to focus on the translation probability onto the most specific translation candidates.

Now, given an input query q_f, the top 10 queries $\{q_e\}$ with the highest bidirectional translation scores with q_f are retrieved from the query log, and $S_{IBM1}(q_f, q_e)$ in Equation (7) is assigned as the value for the feature *Bi-Directional Translation Score*.

3.3.3 Online Mining for Related Queries

OOV word translation is a major knowledge bottleneck for query translation and CLIR. To overcome this knowledge bottleneck, web mining has been exploited in [7, 27] to acquire English-Chinese term translations based on the observation that Chinese terms may co-occur with their English translations in the same web page. In this section, this web mining approach is adapted to acquire not only translations but semantically related queries in the target language.

It is assumed that if a query in the target language co-occurs with the source query in many web pages, they are probably semantically related. Therefore, a simple method is to send the source query to a search engine (Google in our case) for Web pages in the target language in order to find related queries in the target language. For instance, by sending a French query "pages jaunes" to search for English pages, the English snippets containing the key words "yellow pages" or "telephone directory" will be returned. However, this simple approach may induce significant amount of noise due to the non-relevant returns from the search engine. In order to improve the relevancy of the bilingual snippets, we extend the simple approach by the following query modification: the original query is used to search with the dictionary-based query keyword translations, which are unified by the $\land$ (*and*) $\lor$ (*OR*) operators into a single Boolean query. For example, for a given query $q = abc$ where the set of

translation entries in the dictionary of for a is $\{a_1, a_2, a_3\}$, b is $\{b_1, b_2\}$ and c is $\{c_1\}$, we issue $q \wedge (a_1 \vee a_2 \vee a_3) \wedge (b_1 \vee b_2) \wedge c_1$ as one web query.

From the returned top 700 snippets, the most frequent 10 target queries are identified, and are associated with the feature *Frequency in the Snippets*.

Furthermore, we use *Co-Occurrence Double-Check (CODC) Measure* to weight the association between the source and target queries. CODC Measure is proposed in [6] as an association measure based on snippet analysis, named Web Search with Double Checking (WSDC) model. In WSDC model, two objects a and b are considered to have an association if b can be found by using a as query (forward process), and a can be found by using b as query (backward process) by web search. The forward process counts the frequency of b in the top N snippets of query a, denoted as $freq(b @ a)$. Similarly, the backward process count the frequency of a in the top N snippets of query b, denoted as $freq(a @ b)$. Then the CODC association score is defined as follows:

$$S_{CODC}(q_f, q_e) = \begin{cases} 0, \text{if } freq(q_e @ q_f) \times freq(q_f @ q_e) = 0 \\ e^{\left[\log \frac{freq(q_e @ q_f)}{freq(q_f)} \times \frac{freq(q_f @ q_e)}{freq(q_e)} \right]^{\alpha}}, \text{ otherwise} \end{cases} \quad (9)$$

CODC measures the association of two terms in the range between 0 and 1, where under the two extreme cases, q_e and q_f are of no association when $freq(q_e @ q_f) = 0$ or $freq(q_f @ q_e) = 0$, and are of the strongest association when $freq(q_e @ q_f) = freq(q_f)$ and $freq(q_f @ q_e) = freq(q_e)$. In our experiment, α is set at 0.15 following the practice in [6].

Any query q_e mined from the Web will be associated with a feature *CODC Measure* with $S_{CODC}(q_f, q_e)$ as its value.

3.3.4 Monolingual Query Suggestion

For all the candidate queries Q_0 being retrieved using dictionary (see Section 3.3.1), parallel data (see Section 3.3.2) and web mining (see Section 3.3.3), monolingual query suggestion system (described in Section 3.1) is called to produce more related queries in the target language. For each target query q_e, its monolingual source query $SQ_{ML}(q_e)$ is defined as the query in Q_0 with the highest monolingual similarity with q_e, i.e.,

$$SQ_{ML}(q_e) = \arg\max_{q'_e \in Q_0} sim_{ML}(q_e, q'_e) \quad (10)$$

Then the monolingual similarity between q_e and $SQ_{ML}(q_e)$ is used as the value of the q_e 's *Monolingual Query Suggestion Feature*. For any target query $q \in Q_0$, its *Monolingual Query Suggestion Feature* is set as 1.

For any query $q_e \notin Q_0$, its values of *Dictionary-based Translation Score, Bi-Directional Translation Score, Frequency in the Snippet,* and *CODC Measure* are set to be equal to the feature values of $SQ_{ML}(q_e)$.

3.4 Estimating Cross-lingual Query Similarity

In summary, four categories of features are used to learn the cross-lingual query similarity. SVM regression algorithm [25] is used to learn the weights in Equation (2). In this paper, LibSVM toolkit [5] is used for the regression training.

In the prediction stage, the candidate queries will be ranked using the cross-lingual query similarity score computed in terms of $sim_{CL}(t_f, t_e) = w \bullet \phi(f(t_f, t_e))$, and the queries with similarity score lower than a threshold will be regarded as non-relevant. The threshold is learned using a development data set by fitting MLQS's output.

4. CLIR BASED ON CROSS-LINGUAL QUERY SUGGESTION

In Section 3, we presented a discriminative model for cross lingual query suggestion. However, objectively benchmarking a query suggestion system is not a trivial task. In this paper, we propose to use CLQS as an alternative to query translation, and test its effectiveness in CLIR tasks. The resulting good performance of CLIR corresponds to the high quality of the suggested queries.

Given a source query q_f, a set of relevant queries $\{q_e\}$ in the target language are recommended using the cross-lingual query suggestion system. Then a monolingual IR system based on the BM25 model [23] is called using each $q \in \{q_e\}$ as queries to retrieve documents. Then the retrieved documents are re-ranked based on the sum of the BM25 scores associated with each monolingual retrieval.

5. PERFORMACNCE EVALUATION

In this section, we will benchmark the cross-lingual query suggestion system, comparing its performance with monolingual query suggestion, studying the contribution of various information sources, and testing its effectiveness when being used in CLIR tasks.

5.1 Data Resources

In our experiments, French and English are selected as the source and target language respectively. Such selection is due to the fact that large scale query logs are readily available for these two languages. A one-month English query log (containing 7 million unique English queries with occurrence frequency more than 5) of MSN search engine is used as the target language log. And a monolingual query suggestion system is built based on it. In addition, 5,000 French queries are selected randomly from a French query log (containing around 3 million queries), and are manually translated into English by professional French-English translators. Among the 5,000 French queries, 4,171 queries have their translations in the English query log, and are used for CLQS training and testing. Furthermore, among the 4,171 French queries, 70% are used for cross-lingual query similarity training, 10% are used as the development data to determine the relevancy threshold, and 20% are used for testing. To retrieve the cross-lingual related queries, a built-in-house French-English bilingual lexicon (containing 120,000 unique entries) and the Europarl corpus are used.

Besides benchmarking CLQS as an independent system, the CLQS is also tested as a query "translation" system for CLIR

tasks. Based on the observation that the CLIR performance heavily relies on the quality of the suggested queries, this benchmark measures the quality of CLQS in terms of its effectiveness in helping CLIR. To perform such benchmark, we use the documents of TREC6 CLIR data (AP88-90 newswire, 750MB) with officially provided 25 short French-English queries pairs (CL1-CL25). The selection of this data set is due to the fact that the average length of the queries are 3.3 words long, which matches the web query logs we use to train CLQS.

5.2 Performance of Cross-lingual Query Suggestion

Mean-square-error (MSE) is used to measure the regression error and it is defined as follows:

$$ MSE = \frac{1}{l} \sum_i \left(sim_{CL}(q_{fi}, q_{ei}) - sim_{ML}(T_{q_{fi}}, q_{ei}) \right)^2 $$

where l is the total number of cross-lingual query pairs in the testing data.

As described in Section 3.4, a relevancy threshold is learned using the development data, and only CLQS with similarity value above the threshold is regarded as truly relevant to the input query. In this way, CLQS can also be benchmarked as a classification task using precision (P) and recall (R) which are defined as follows:

$$ P = \frac{\left| S_{CLQS} \cap S_{MLQS} \right|}{\left| S_{CLQS} \right|}, \quad R = \frac{\left| S_{CLQS} \cap S_{MLQS} \right|}{\left| S_{MLQS} \right|} $$

where S_{CLQS} is the set of relevant queries suggested by CLQS, S_{MLQS} is the set of relevant queries suggested by MLQS (see Section 3.2).

The benchmarking results with various feature configurations are shown in Table 1.

Features	Regression	Classification	
	MSE	P	R
DD	0.274	0.723	0.098
DD+PC	0.224	0.713	0.125
DD+PC+ Web	0.115	0.808	0.192
DD+PC+ Web+ML QS	0.174	0.796	0.421

Table 1. CLQS performance with different feature settings

(*DD*: dictionary only; *DD+PC*: dictionary and parallel corpora; *DD+PC+Web*: dictionary, parallel corpora, and web mining; *DD+PC+Web+MLQS*: dictionary, parallel corpora, web mining and monolingual query suggestion)

Table 1 reports the performance comparison with various feature settings. The baseline system (DD) uses a conventional query translation approach, i.e., a bilingual dictionary with co-occurrence-based translation disambiguation. The baseline system only covers less than 10% of the suggestions made by MLQS. Using additional features obviously enables CLQS to generate more relevant queries. The most significant improvement on recall is achieved by exploiting MLQS. The final CLQS system is able

to generate 42% of the queries suggested by MLQS. Among all the feature combinations, there is no significant change in precision. This indicates that our methods can improve the recall by effectively leveraging various information sources without losing the accuracy of the suggestions.

Besides benchmarking CLQS by comparing its output with MLQS output, 200 French queries are randomly selected from the French query log. These queries are double-checked to make sure that they are not in the CLQS training corpus. Then CLQS system is used to suggest relevant English queries for them. On average, for each French query, 8.7 relevant English queries are suggested. Then the total 1,740 suggested English queries are manually checked by two professional English/French translators with cross-validation. Among the 1,747 suggested queries, 1,407 queries are recognized as relevant to the original ones, hence the accuracy is 80.9%. Figure 1 shows an example of CLQS of the French query "terrorisme international" ("international terrorism" in English).

international terrorism (0.991); what is terrorism (0.943);

counter terrorism (0.920); terrorist (0.911);

terrorist attacks (0.898); international terrorist (0.853);

world terrorism (0.845); global terrorism (0.833);

transnational terrorism (0.821); human rights (0.811);

terrorist groups (0. 777); patterns of global terrorism (0.762)

september 11 (0.734)

Figure 1. An example of CLQS of the French query "terrorisme international"

5.3 CLIR Performance

In this section, CLQS is tested with French to English CLIR tasks. We conduct CLIR experiments using the TREC 6 CLIR dataset described in Section 5.1. The CLIR is performed using a query translation system followed by a BM25-based [23] monolingual IR module. The following three different systems have been used to perform query translation: (1) CLQS: our CLQS system; (2) MT: Google French to English machine translation system; (3) DT: a dictionary based query translation system using co-occurrence statistics for translation disambiguation. The translation disambiguation algorithm is presented in Section 3.3.1. Besides, the monolingual IR performance is also reported as a reference. The average precision of the four IR systems are reported in Table 2, and the 11-point precision-recall curves are shown in Figure 2.

IR System	Average Precision	% of Monolingual IR
Monolingual	0.266	100%
MT	0.217	81.6%
DT	0.186	69.9%
CLQS	0.233	87.6%

Table 2. Average precision of CLIR on TREC 6 Dataset

(*Monolingual*: monolingual IR system; *MT*: CLIR based on machine translation; *DT*: CLIR based on dictionary translation; *CLQS*: CLQS-based CLIR)
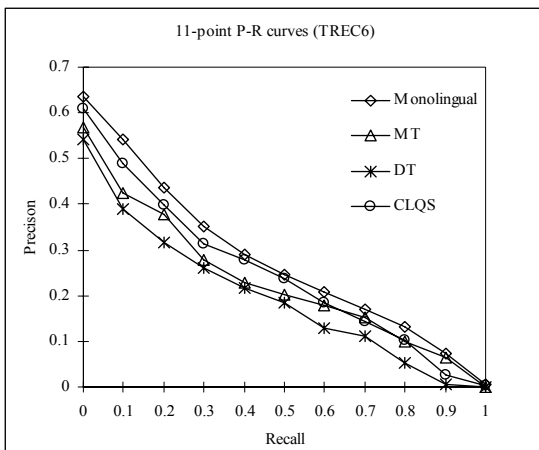
Figure 2. 11 points precision-recall on TREC6 CLIR data set

The benchmark shows that using CLQS as a query translation tool outperforms CLIR based on machine translation by 7.4%, outperforms CLIR based on dictionary translation by 25.2%, and achieves 87.6% of the monolingual IR performance.

The effectiveness of CLQS lies in its ability in suggesting closely related queries besides accurate translations. For example, for the query CL14 "terrorisme international" ("international terrorism"), although the machine translation tool translates the query correctly, CLQS system still achieves higher score by recommending many additional related terms such as "global terrorism", "world terrorism", etc. (as shown in Figure 1). Another example is the query "La pollution causée par l'automobile" ("air pollution due to automobile") of CL6. The MT tool provides the translation "the pollution caused by the car", while CLQS system enumerates all the possible synonyms of "car", and suggest the following queries "car pollution", "auto pollution", "automobile pollution". Besides, other related queries such as "global warming" are also suggested. For the query CL12 "La culture écologique" ("organic farming"), the MT tool fails to generate the correct translation. Although the correct translation is neither in our French-English dictionary, CLQS system generates "organic farm" as a relevant query due to successful web mining.

The above experiment demonstrates the effectiveness of using CLQS to suggest relevant queries for CLIR enhancement. A related research is to perform query expansion to enhance CLIR [2, 18]. So it is very interesting to compare the CLQS approach with the conventional query expansion approaches. Following [18], post-translation expansion is performed based on pseudo-relevance feedback (PRF) techniques. We first perform CLIR in the same way as before. Then we use the traditional PRF algorithm described in [24] to select expansion terms. In our experiments, the top 10 terms are selected to expand the original query, and the new query is used to search the collection for the second time. The new CLIR performance in terms of average precision is shown in Table 3. The 11-point P-R curves are drawn in Figure 3.

Although being enhanced by pseudo-relevance feedback, the CLIR using either machine translation or dictionary-based query translation still does not perform as well as CLQS-based approach. Statistical t-test [13] is conducted to indicate whether the CLQS-based CLIR performs significantly better. Pair-wise p-

values are shown in Table 4. Clearly, CLQS significantly outperforms MT and DT without PRF as well as DT+PRF, but its superiority over MT+PRF is not significant. However, when combined with PRF, CLQS significant outperforms all the other methods. This indicates the higher effectiveness of CLQS in related term identification by leveraging a wide spectrum of resources. Furthermore, post-translation expansion is capable of improving CLQS-based CLIR. This is due to the fact that CLQS and pseudo-relevance feedback are leveraging different categories of resources, and both approaches can be complementary.

IR System	AP without PRF	AP with PRF
Monolingual	0.266 (100%)	0.288 (100%)
MT	0.217 (81.6%)	0.222 (77.1%)
DT	0.186 (69.9%)	0.220 (76.4%)
CLQS	0.233 (87.6%)	0.259 (89.9%)

Table 3. Comparison of average precision (AP) on TREC 6 without and with post-translation expansion. (%) are the relative percentages over the monolingual IR performance
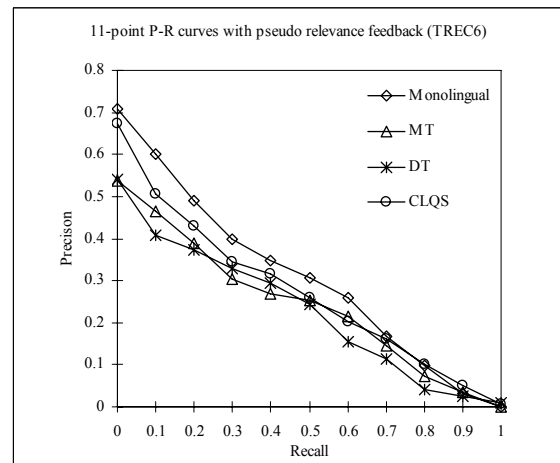

Figure 3. 11 points precision-recall on TREC6 CLIR dataset with pseudo relevance feedback

	MT	DT	MT+PRF	DT+PRF
CLQS	0.0298	3.84e-05	0.1472	0.0282
CLQS+PRF	0.0026	2.63e-05	0.0094	0.0016

Table 4. The results of pair-wise significance t-test. Here p-value < 0.05 is considered statistically significant

6. CONCLUSIONS

In this paper, we proposed a new approach to cross-lingual query suggestion by mining relevant queries in different languages from query logs. The key solution to this problem is to learn a cross-lingual query similarity measure by a discriminative model exploiting multiple monolingual and bilingual resources. The model is trained based on the principle that cross-lingual similarity should best fit the monolingual similarity between one query and the other query's translation.

The baseline CLQS system applies a typical query translation approach, using a bilingual dictionary with co-occurrence-based translation disambiguation. This approach only covers 10% of the relevant queries suggested by an MLQS system (when the exact translation of the original query is given). By leveraging additional resources such as parallel corpora, web mining and log-based monolingual query expansion, the final system is able to cover 42% of the relevant queries suggested by an MLQS system with precision as high as 79.6%.

To further test the quality of the suggested queries, CLQS system is used as a query "translation" system in CLIR tasks. Benchmarked using TREC 6 French to English CLIR task, CLQS demonstrates higher effectiveness than the traditional query translation methods using either bilingual dictionary or commercial machine translation tools.

The improvement on TREC French to English CLIR task by using CLQS demonstrates the high quality of the suggested queries. This also shows the strong correspondence between the input French queries and English queries in the log. In the future, we will build CLQS system between languages which may be more loosely correlated, e.g., English and Chinese, and study the CLQS performance change due to the less strong correspondence among queries in such languages.

7. REFERENCES

[1] Ambati, V. and Rohini., U. Using Monolingual Clickthrough Data to Build Cross-lingual Search Systems. In *Proceedings of New Directions in Multilingual Information Access Workshop of SIGIR 2006*.

[2] Ballestors, L. A. and Croft, W. B. Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval. In *Proc. SIGIR 1997*, pp. 84-91.

[3] Ballestors, L. A. and Croft, W. B. Resolving Ambiguity for Cross-Language Retrieval. In *Proc. SIGIR 1998*, pp. 64-71.

[4] Brown, P. F., Pietra, D. S. A., Pietra, D. V. J., and Mercer, R. L. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263-311, 1993.

[5] Chang, C. C. and Lin, C. LIBSVM: a Library for Support Vector Machines (Version 2.3). 2001. http://citeseer.ist.psu.edu/chang01libsvm.html

[6] Chen, H.-H., Lin, M.-S., and Wei, Y.-C. Novel Association Measures Using Web Search with Double Checking. In *Proc. COLING/ACL 2006*, pp. 1009-1016.

[7] Cheng, P.-J., Teng, J.-W., Chen, R.-C., Wang, J.-H., Lu, W.-H., and Chien, L.-F. Translating Unknown Queries with Web Corpora for Cross-Language Information Retrieval. In *Proc. SIGIR 2004*, pp. 146-153.

[8] Cui, H., Wen, J. R., Nie, J.-Y., and Ma, W. Y. Query Expansion by Mining User Logs. *IEEE Trans. on Knowledge and Data Engineering*, 15(4):829-839, 2003.

[9] Fujii A. and Ishikawa, T. Applying Machine Translation to Two-Stage Cross-Language Information Retrieval. In *Proceedings of 4th Conference of the Association for Machine Translation in the Americas*, pp. 13-24, 2000.

[10] Gao, J. F., Nie, J.-Y., Xun, E., Zhang, J., Zhou, M., and Huang, C. Improving query translation for CLIR using statistical Models. In *Proc. SIGIR 2001*, pp. 96-104.

[11] Gao, J. F., Nie, J.-Y., He, H., Chen, W., and Zhou, M. Resolving Query Translation Ambiguity using a Decaying Co-occurrence Model and Syntactic Dependence Relations. In *Proc. SIGIR 2002*, pp. 183-190.

[12] Gleich, D., and Zhukov, L. SVD Subspace Projections for Term Suggestion Ranking and Clustering. In *Technical Report,* Yahoo! Research Labs, 2004.

[13] Hull, D. Using Statistical Testing in the Evaluation of Retrieval Experiments. In *Proc. SIGIR 1993*, pp. 329-338.

[14] Jeon, J., Croft, W. B., and Lee, J. Finding Similar Questions in Large Question and Answer Archives. In *Proc. CIKM 2005*, pp. 84-90.

[15] Joachims, T. Optimizing Search Engines Using Clickthrough Data. In *Proc. SIGKDD* 2002, pp. 133-142.

[16] Lavrenko, V., Choquette, M., and Croft, W. B. Cross-Lingual Relevance Models. *In Proc. SIGIR* 2002, pp. 175-182.

[17] Lu, W.-H., Chien, L.-F., and Lee, H.-J. Anchor Text Mining for Translation Extraction of Query Terms. In *Proc. SIGIR 2001*, pp. 388-389.

[18] McNamee, P. and Mayfield, J. Comparing Cross-Language Query Expansion Techniques by Degrading Translation Resources. In *Proc. SIGIR 2002*, pp. 159-166.

[19] Monz, C. and Dorr, B. J. Iterative Translation Disambiguation for Cross-Language Information Retrieval. In *Proc. SIGIR 2005*, pp. 520-527.

[20] Nie, J.-Y., Simard, M., Isabelle, P., and Durand, R. Cross-Language Information Retrieval based on Parallel Text and Automatic Mining of Parallel Text from the Web. In *Proc. SIGIR 1999*, pp. 74-81.

[21] Och, F. J. and Ney, H. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19-51, 2003.

[22] Pirkola, A., Hedlund, T., Keshusalo, H., and Järvelin, K. Dictionary-Based Cross-Language Information Retrieval: Problems, Methods, and Research Findings. *Information Retrieval*, 4(3/4):209-230, 2001.

[23] Robertson, S. E., Walker, S., Hancock-Beaulieu, M. M., and Gatford, M. OKAPI at TREC-3. In *Proc.TREC-3*, pp. 200-225, 1995.

[24] Robertson, S. E. and Jones, K. S. Relevance Weighting of Search Terms. *Journal of the American Society of Information Science*, 27(3):129-146, 1976.

[25] Smola, A. J. and Schölkopf, B. A. Tutorial on Support Vector Regression. *Statistics and Computing,* 14(3):199-222, 2004.

[26] Wen, J. R., Nie, J.-Y., and Zhang, H. J. Query Clustering Using User Logs. *ACM Trans. Information Systems*, 20(1):59-81, 2002.

[27] Zhang, Y. and Vines, P. Using the Web for Automated Translation Extraction in Cross-Language Information Retrieval. In *Proc. SIGIR* 2004, pp. 162-169.