1-2010

Exploiting query logs for cross-lingual query suggestions.

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

Cheng NIU

Jian-Yun NIE

Ming ZHOU

Kam-Fai WONG

See next page for additional authors

Citation

GAO, Wei; NIU, Cheng; NIE, Jian-Yun; ZHOU, Ming; WONG, Kam-Fai; and HON, Hsiao-Wuen. Exploiting query logs for cross-lingual query suggestions.. (2010). *ACM Transactions on Information Systems*. 28, (2), 1-33. Research Collection School Of Information Systems.
Available at: https://ink.library.smu.edu.sg/sis_research/4552

Exploiting Query Logs for Cross-Lingual Query Suggestion

6

WEI GAO
The Chinese University of Hong Kong
CHENG NIU
Microsoft Research Asia
JIAN-YUN NIE
Université de Montréal
MING ZHOU
Microsoft Research Asia
KAM-FAI WONG
The Chinese University of Hong Kong
and
HSIAO-WUEN HON
Microsoft Research Asia

Query suggestion aims to suggest relevant queries for a given query, which helps users better specify their information needs. Previous work on query suggestion has been limited to the same language. In this article, 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 the scenarios of cross-language information retrieval (CLIR) and other related cross-lingual applications. 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, and so on, are used to estimate the cross-lingual query similarity

The research described in this article partially appeared as Gao et al. [2007].
This work was substantially conducted while W. Gao was visiting Microsoft Research Asia.
This research is sponsored in part by the Hong Kong Innovation and Technology Fund (ITS/182/09) and the CUHK direct grant (2050443), and is partially affiliated with the Microsoft-CUHK Joint laboratory for Human-Centric Computing and Interface Technologies.
Authors' addresses: W. Gao and K.-F. Wong, Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong, China; email: {wgao, kfwong}@se.cuhk.edu.hk; C. Niu, M. Zhou, and H.-W. Hon, Microsoft Research Asia, No. 49, Zhichun Road, Beijing 100190, China; email: {chengniu, mingzhou, hon}@microsoft.com; J.-Y. Nie, Université de Montréal, Montréal, Canada; email: nie@iro.umontreal.ca.

DOI 10.1145/1740592.1740594 http://doi.acm.org/10.1145/1740592.1740594

with a discriminative model. Benchmarks show that the resulting CLQS system significantly outperforms a baseline system that uses dictionary-based query translation. Besides, we evaluate CLQS with French-English and Chinese-English CLIR tasks on TREC-6 and NTCIR-4 collections, respectively. The CLIR experiments using typical retrieval models demonstrate that the CLQS-based approach has significantly higher effectiveness than several traditional query translation methods. We find that when combined with pseudo-relevance feedback, the effectiveness of CLIR using CLQS is enhanced for different pairs of languages.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Query formulation; Search process*

General Terms: Algorithms, Performance, Experimentation, Theory

Additional Key Words and Phrases: Cross-language information retrieval, query expansion, query log, query suggestion, query translation

ACM Reference Format:

Gao, W., Niu, C., Nie, J.-Y., Zhou, M., Wong, K.-F., and Hon, H.-W. 2010. Exploiting query logs for cross-lingual query suggestion. ACM Trans. Inf. Syst. 28, 2, Article 6 (May 2010), 33 pages.
DOI 10.1145/1740592.1740594. http://doi.acm.org/10.1145/1740592.1740594

1. INTRODUCTION

Query suggestion is a designed to help users of a search engine better specify their information needs. This is accomplished by narrowing down or expanding the scope of a search with synonymous queries and relevant queries, or by suggesting related queries that have been frequently used by other users. Popular search engines, such as Google,[1] Yahoo!,[2] Live Search,[3] Ask.com,[4] do provide query suggestion functionality as a valuable addition to their core search technology. Moreover, the same approach has been applied to recommend bidding terms to online advertisers in the pay-for-performance search market [Gleich and Zhukov 2004].

Query suggestion is related to query expansion, which extends the original query with new search terms to narrow down 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. Therefore, it is expected to play an alternative or complementary role to query expansion in information retrieval applications.

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 [Gleich and Zhukov 2004; Jeon et al. 2005; Wen et al. 2002]. By suggesting the related and frequently used formulations, it is expected that the new queries can cover more relevant documents.

To our knowledge, all existing studies only deal with monolingual query suggestion. There is no research on cross-lingual query suggestion (CLQS)

[1]http://www.google.com
[2]http://search.yahoo.com
[3]http://www.live.com
[4]http://www.ask.com

by exploiting query logs. CLQS aims to suggest related queries in a different language. It has important applications on the World Wide Web such as cross-language search or suggesting relevant bidding terms in e-advertising.

CLQS can be approached as a query translation problem—to formulate the queries that are translations of the original query. Dictionaries, large-size parallel corpora, and existing commercial machine translation (MT) systems can be used for translation. However, these kinds of approaches usually rely on static knowledge and data that cannot effectively reflect the quickly shifting interests of Web users. As a consequence, even though the terms can be reasonable translations of the original terms in the source-language query, the suggested queries may not be the most reasonable and popular formulations in the target language. For example, the French query "aliment biologique" is translated into "biologic food" by Google's machine translation tool.[5] At the term level, the translation seems reasonable. However, the correct formulation should be "organic food." Similarly, the Chinese query "动物复制" is translated literally as "animal reproduction," but in fact it is widely expressed as "animal cloning" in English. There are many such mismatch cases between the translated terms and those actually used in the target language. Such mismatches render the translated queries ineffective in finding relevant documents in the target language.

A natural way of solving this mismatch problem is to exploit query logs in the target language to select the most popular query formulations corresponding to the original query in the source language. Ideally, the selection would be most effective if one has a query log with aligned queries between the source and target languages. However, such a resource does not exist. In practice, we only have separate query logs in source and target languages for the same period of time. Such resources are still very useful to us. We can assume that the two separate query logs cover many common search interests. Therefore, it can be expected that for many queries in the source language we can find their corresponding or similar queries in the target-language query log, especially for popular queries.

The query logs can be used in the following way for CLQS: when a source-language query is submitted, we try to determine the most similar query in the target-language query log. In addition to considering the translation relation between the source-language query and the target-language suggestions, we also leverage the following two effects from the target-language query log.

(1) The suggested queries from the target-language query log are complete queries, which correspond to the normal ways users formulate queries in the target language. In this way, compared to the translation approach, more natural formulation of queries can be obtained.
(2) The suggested queries from the target-language query log can not only be the translation of the original query, but also strongly related ones. Therefore, we can more naturally achieve the desired effect of query expansion to reflect users' needs.

[5]http://www.google.com/translate_t

A key issue to arrive at reasonable cross-lingual query suggestion is the estimation of cross-lingual query similarity. In this article, we propose a new method for calculating this similarity 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, and so on. A discriminative model is used to learn the calculation of 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.

The resulting CLQS system is evaluated as an independent module as well as a new means of query translation for French-English and Chinese-English CLIR tasks using prevalent retrieval models based on TREC-6 and NTCIR-4 data collections, respectively. It is then compared with several traditional query translation methods including a dictionary-based translation approach using co-occurrence-based translation disambiguation, a phrase-based statistical machine translation (SMT) system, and an automated translation extraction technique by mining unknown query translations from Web corpora. The results show that this new translation method is more effective than the other approaches. Furthermore, we show that when combined with pseudo-relevance feedback (PRF), CLIR effectiveness is further improved.

The remainder of this article 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 results; finally, we conclude in Section 6 and give future directions in Section 7.

2. RELATED WORK

Most approaches for CLIR are achieved by query translation followed by monolingual IR. Typically, queries are translated using a bilingual dictionary [Pirkola et al. 2001], a machine translation system [Fuji and Ishikawa 2000], a parallel [Nie et al. 1999], or comparable corpus [López-Ostenero et al. 2005].

Despite the various types of resources used, out-of-vocabulary (OOV) words and translation disambiguation are the two major bottlenecks for CLIR [Nie et al. 1999]. In Cheng et al. [2004] and Zhang and Vines [2004], OOV term translations were mined from the Web using a search engine. In Lu et al. [2001], bilingual knowledge was acquired based on anchor text analysis. In addition, word co-occurrence statistics in the target language have been applied for translation disambiguation [Ballesteros and Croft 1998; Gao et al. 2001, 2002; Monz and Dorr 2005].

For query translation employed for CLIR, Kwok et al. [2005] utilized translation results from different MT tools and translation resources. The system achieved better CLIR effectiveness than the single translation approach. Although we also resort to various translation resources, our CLQS approach is different from Kwok's in that we employ the resources for finding relevant candidate queries in the query log rather than for acquiring accurate translations.

It is arguable that accurate query translation may neither be necessary, nor sufficient, for CLIR. In many cases, it is helpful to introduce words that are not direct translations of any query word, but are closely related to the meaning of the query. From a translation point of view, such a translation is certainly imperfect. However, several experiments have shown that such a translation could perform better than a high-quality MT result [Kraaij et al. 2003], and even better than a professional manual translation for CLIR purpose [Gao et al. 2001]. This observation has led to the development of cross-lingual query expansion (CLQE) techniques [Ballesteros and Croft 1997; Lavrenko et al. 2002; McNamee and Mayfield 2002]. Ballesteros and Croft [1997] reported the enhancement on CLIR by post-translation expansion. Lavrenko et al. [2002] developed a cross-lingual relevancy model by using the cross-lingual co-occurrence statistics in parallel texts. McNamee and Mayfield [2002] compared the performance of multiple CLQE techniques, including pre- and post-translation expansions. However, a unified framework to combine the wide range of resources and Web mining techniques for CLQE is yet unavailable.

López-Ostenero et al. [2005] proposed a method for cross-language search by accurate translation of the noun phrases in a query, followed by a blind expansion with frequent phrases. Their bilingual phrase alignment dictionary was built on a comparable corpus, in which query refinement is fulfilled by using the phrase-based summary of document content. This technique could be considered as a noun-phrase-based CLQE.

CLQS is different from CLQE in that it aims to suggest full queries that have been formulated by users in another language. Our CLQS approach 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 of 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, especially in monolingual query suggestions [Gleich and Zhukov 2004] and in relating semantically relevant terms for query expansion [Cui et al. 2003; Joachims 2002]. In Ambati and Rohini [2006], the target-language query log has been exploited for query translation in CLIR. White et al. [2007] compared the similarity of refined queries using query logs and PRF in Web search. Using a BM25 retrieval model [Robertson et al. 1995], our recent work [Gao et al. 2007] showed that in the French-English CLIR task, a CLQS-based approach outperformed a dictionary-based method and an online MT tool from Google for query translation. The combination of CLQS and PRF could be complementary and improve CLIR effectiveness.

Nevertheless, several important issues remain unclear and unexplored in our previous study: (1) When queries are translated using online MT software such as Google, it is difficult to compare it with CLQS because the translation quality frequently changes due to product updates made by the service provider. In addition, the techniques and data resources used for constructing the MT system are unknown to us; (2) It is unclear how CLQS-based CLIR performs compared to query translation under different IR frameworks, especially when PRF is introduced. Since PRF techniques vary with the underlying retrieval

models, it is uncertain whether PRF could consistently complement CLQS; (3) It is unknown if high-quality queries could be suggested using query logs across linguistically less correspondent languages, such as Chinese-English. It is interesting to investigate the effectiveness of CLQS for such a language pair, where the correspondence between users' search interests might not be strong. In this article, we will examine all of these issues.

3. ESTIMATING CROSS-LINGUAL QUERY SIMILARITY

A search engine has a query log containing user queries with time stamps. In addition, click-through information is also recorded. Therefore, we know which documents have been selected by users for each query. A search engine is used simultaneously by users in different languages, or more precisely, each version of the search engine is used by users of a language group (and locale). We then have a query log for each language (or locale) at the same time period. The simultaneous query logs are the key resources that we exploit in this study. Given a query in the source language, our CLQS task is to determine one or more 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 proposed for monolingual terms [Cui et al. 2003; Wen et al. 2002], most of them were based on term co-occurrence statistics, and could hardly be applied directly to cross-lingual settings since terms of different languages are not so likely to co-occur as monolingual terms.

In order to define a similarity measure across languages, one has to use at least one translation tool or resource. As a result, the measure will be based on both translation relation and monolingual similarity. In this work, we aim to provide up-to-date query similarity measures; static translation resources alone are not sufficient. Therefore, we also integrate a method to mine possible translations from 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 study, we propose a discriminative model to learn the appropriate similarity measure. 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. We then use a discriminative model to learn the cross-language similarity function that best fits these examples.

In the following sections, the details of the discriminative model for cross-lingual query similarity estimation are described. We then introduce all the features (monolingual and cross-lingual information) that will be used in the model.

3.1 Discriminative Model for Estimating Cross-Lingual Query Similarity

We first assume a reasonable monolingual query similarity measure as the target in the discriminative training. For a pair of queries in different languages,

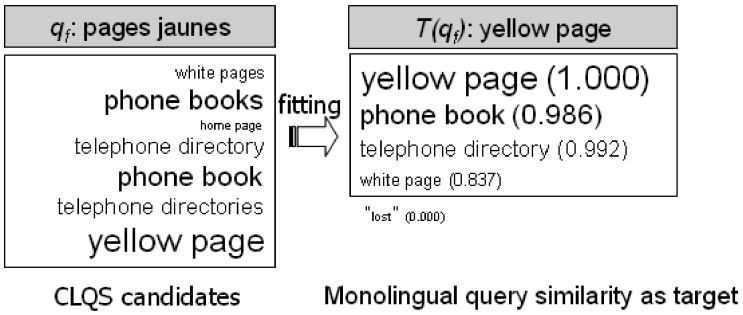

Fig. 1. An illustration of the principle to transpose cross-lingual query similarity to monolingual query similarity for CLQS candidates to fit as target values. Note that the matched queries are displayed with the characters in the same size.

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" ("yellow pages" in English) and English query "telephone directory" should be equal to the monolingual similarity between the English translation of the French query "yellow page" and "telephone directory". Figure 1 shows an illustration of our principle based on this example.

Compared to a query translation approach, this approach has several advantages.

(1) Monolingual query similarity can be estimated more accurately than cross-lingual query similarity; there are many methods and resources available for it. Using our approach, we can take advantage of monolingual similarity to deduce a way to estimate cross-lingual query similarity.
(2) Cross-lingual query suggestion is not limited to query translation. Similar queries in the target language can also be suggested, even though they are not direct translations. For example, "telephone directory" can be suggested for the French query "pages jaunes". This will naturally lead to the effect of query expansion.
(3) The suggested queries in the target language are those that appeared frequently in the query logs in the target language. Thus, we can also take into account the way that queries are formulated by users in the target language. For example, if the query "organic food" is submitted much more often than the query "biologic food" in English, then the former would be suggested for the French query "nourriture biologique."

The target monolingual query similarity can be determined in various ways, for example, using term co-occurrence based mutual information [Jiang et al. 1999] and chi-square [Cheng et al. 2004]. 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 described in the following.

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,

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

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 compiled by human experts and a monolingual query similarity function. An existing monolingual query suggestion system can then be used to automatically produce similar queries 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 work, the support vector machine (SVM) regression algorithm [Smola and Scholkopf 2004] is used to learn the cross-lingual term similarity function. Given $\mathbf{f}$, a vector of feature functions with respect to q_f and q_e, $sim_{CL}(q_f, q_e)$ is represented as an inner product between a weight vector and the feature vector in a kernel space.

$$sim_{CL}(q_f, q_e) = \mathbf{w} \cdot \phi(\mathbf{f}(q_f, q_e)), \tag{2}$$

where $\phi(.)$ is the mapping from the input feature space onto the kernel space, and $\mathbf{w}$ is the weight vector in the kernel space that will be learned by the SVM regression training. Once the weight vector is learned, Equation 2 can be used to estimate the similarity between queries of different languages.

It is noteworthy that instead of regression, one can simplify the learning process as a binary or ordinal classification task, in which case CLQS can be categorized according to discontinuous class labels, for example, relevant and irrelevant, or a series of levels of relevancies, for example, 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.

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) are presented.

3.2 Monolingual Query Similarity Measure Based on Click-Through Information

Any monolingual term similarity measure can be used as the regression target. We adopt the monolingual query similarity measure presented in Wen et al. [2002], which used search users' click-through information in query logs and performed effectively in monolingual query suggestion. The reason to choose this monolingual similarity measure is that it is defined in a context similar to ours, that is, according to a user log that reflects users' intentions and behavior. Therefore, we can expect that the cross-lingual query similarity learned from it can also reflect users' intentions and expectations.

Following Wen et al. [2002], 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,

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

where $kn(x)$ is the number of keywords in query x and $KN(p,q)$ is the number of common keywords in the two queries. Secondly, the click-through-based similarity is defined as,

$$similarity_{click\text{-}through}(p,q) = \frac{RD(p,q)}{\max(rd(p), rd(q))}, \quad (4)$$

where $rd(x)$ is the number of clicked URLs for query x, and $RD(p,q)$ is the number of common URLs clicked for the two queries. These two similarity measures represent different points of view. The content-based measure captures queries with the same or similar terms without considering semantic relatedness, such as "Barack Obama", "Obama Barack", "Senator Barack Obama", and so on, and the click-through-based measure captures queries semantically related to the same or similar topics, such as "Illinois Senator", "Obama 2004 democratic national convention", "Michelle Obama", and so on.

However, the user's need for information need may only be partially captured by either of the measures. In order to take advantage of both strategies, the similarity between two queries can be formulated as a linear combination of the two similarities,

$$sim_{ML}(p,q) = \delta * similarity_{content}(p,q) + (1-\delta) * similarity_{click\text{-}through}(p,q), \quad (5)$$

where δ is the relative importance of the content-based similarity. In this work, we set $\delta = 0.4$ empirically. If a query p has a similarity measure higher than a certain threshold with another query q, q will be regarded as a relevant monolingual query suggestion (MLQS). The threshold is empirically set as 0.9. Note that Wen et al. [2002] described details about parameter tuning and the impact of the threshold on MLQS.

3.3 Features Used for Learning Cross-Lingual Query Similarity Measurement

This section presents the extraction of candidate-relevant queries from the log with the assistance of various monolingual and bilingual resources. Also, 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 widely used for query translation in previous work [Ballesteros and Croft 1998; Gao et al. 2001; McNamee and Mayfield 2002; Nie et al. 1999; Pirkola et al. 2001]. But note that we employ them for a different purpose—for finding relevant candidates in the log rather than for acquiring accurate translations.

3.3.1 *Bilingual Dictionary.* In this subsection, we present how a bilingual dictionary is used to retrieve candidate queries from a query log. Since multiple translations may be associated with each source word, co-occurrence-based translation disambiguation [Ballesteros and Croft 1998; Gao et al. 2001; Gao et al. 2002] is performed and described.

Given an input query $q_f = w_{f_1} w_{f_2} \ldots w_{f_M}$ in the source language, for each query term w_{f_i}, a set of unique translations provided by the bilingual dictionary is denoted as $T_i : D(w_{f_i}) = \{t_{i_1}, t_{i_2}, \ldots, t_{i_m}\}$. We then try to determine a measure of cohesion between the translations of different query words w_{f_i} and w_{f_k} $(i, k = 1, 2, \ldots, M$ and $i \neq k)$. A cohesive query is the one that has a high likelihood to be formed in the target language. Here, we define the cohesion between the translation terms of two query terms: $t_{i_j} \in T_i$ and $t_{k_l} \in T_k$ (where $T_k : D(w_{f_k}) = \{t_{k_1}, t_{k_2}, \ldots, t_{k_n}\}$ is a set of translations of term w_{f_k}), according to the following mutual information (MI) formula:

$$MI(t_{i_j}, t_{k_l}) = P(t_{i_j}, t_{k_l}) log \frac{P(t_{i_j}, t_{k_l})}{P(t_{i_j})P(t_{k_l})}, \tag{6}$$

where $P(t_{i_j}, t_{k_l}) = \frac{C(t_{i_j}, t_{k_l})}{N}$ and $P(t) = \frac{C(t)}{N}$. Here $C(x, y)$ is the number of queries in the log containing both terms x and y, $C(x)$ is the number of queries containing term x, and N is the total number of queries in the log. The MI value indicates how likely two translation terms co-occur in the queries of the target-language log.

Based on the term-term cohesion defined in Equation 6, the optimal set of query translations can be approximated using the greedy algorithm in Gao et al. [2001]. The algorithm selects the translation word t_{i_j} in each T_i that has the highest degree of cohesion with the words $\{t_{k_l}\}$ in all other translation sets $\{T_k\}$ $(k \neq i)$. The set of best words from each translation set forms our query translation T'_{q_f}, whose similarity is measured by the summation of the term-term cohesion values of the selected terms.

$$S_{dict}(T'_{q_f}) = \sum_i \max_{i_j} \sum_{k, k \neq i} \max_{k_l} MI(t_{i_j}, t_{k_l}). \tag{7}$$

The process then iterates to find the next set of best translation words by gradually excluding those words already selected. All the generated query translations are added into the set $\{T'_{q_f}\}$ and ranked by the $S_{dict}(T'_{q_f})$ score.

For each query translation $T \in \{T'_{q_f}\}$, we retrieve all the queries containing the same keywords as T from the target-language query log. The retrieved queries are candidate target queries, and are assigned $S_{dict}(T)$ as the value of the feature *Dictionary-Based Translation Score*. By trial and error on different numbers of candidates, we empirically select four best candidate target queries ranked by the $S_{dict}(T)$ score for the suggestion, which yield nearly optimal training performance. The number of candidates is also determined in a similar way for candidate extraction using parallel corpora and Web mining.

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 that is useful in acquiring dominant query translations.

A parallel corpus is first aligned at sentence level. Word alignments can then be derived by training an IBM translation model-1 [Brown et al. 1993] using *GIZA++* [Och and Ney 2003]. The learned bilingual knowledge is used to extract candidate queries from the query log.

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,

$$S_{model\text{-}1}(q_f, q_e) = \sqrt{P_{model\text{-}1}(q_f|q_e) \times P_{model\text{-}1}(q_e|q_f)}, \tag{8}$$

where $P_{model\text{-}1}(y|x)$ is the word sequence translation probability given by IBM model-1, which has the form,

$$P_{model\text{-}1}(y|x) = \frac{1}{(|x|+1)^{|y|}} \prod_{j=1}^{|y|} \sum_{i=0}^{|x|} P(y_j|x_i), \tag{9}$$

where $P(y_j|x_i)$ 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 enhance the translation probability of the most specific translation candidates.

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 $S_{model\text{-}1}(q_f, q_e)$ in Equation 8 is assigned as the value for the feature, *Bi-Directional Translation Score*.

3.3.3 *Web Mining for Related Queries.* The translation of unknown words or Out-Of-Vocabulary (OOV) words is a major knowledge bottleneck for query translation and CLIR. To overcome this predicament, Web mining has been exploited in Cheng et al. [2004] and Zhang and Vines [2004] to acquire English-Chinese term translations. The proposed methods are based on the observation that Chinese terms may co-occur with their English translations, for example, "...皇家马德里 (Real Madrid)..." in the same Chinese Web page. This approach works well for foreign proper names that occur frequently in Web pages. Our goal is broader. We are not limited to mining translations of unknown words; we are also interested in mining strongly related terms. For example, we expect the queries relevant to "贝克汉姆" (David Beckham) to be mined as well for this example as this proper name is very likely to occur within the context of the Web pages and/or query logs related to Real Madrid. In this section, we describe a variant of this approach to acquire both translations and 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 (e.g., Google) 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 a 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 $\wedge$ (AND) and $\vee$ (OR) operators into a single Boolean query. For example, for a given query $q = abc$, where the set of translation entries in the dictionary for word 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 top 700 returned snippets of each constructed Boolean query, query translations are first identified using the *SCPCD* (Symmetric Conditional Probability with Context Dependency) measure from all word n-grams in the target language. *SCPCD* combines the symmetric conditional probability (*SCP*) with the context dependency (*CD*) for n-grams, and is used as an association measure for determining an n-gram as a well-formed phrase (see Cheng et al. [2004] for details). The most frequent 10 candidate queries are then retrieved from the query log using the extracted phrases and are associated with the features of *Frequency in the Snippets*.

Furthermore, we use Co-Occurrence Double-Check (*CODC*) measure to weight the relatedness between the source and target queries. The *CODC* measure is proposed in Chen et al. [2006] as an association measure based on snippet analysis, referred to as the Web Search with Double Checking (*WSDC*) model. In the *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) in the 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 counts the frequency of a in the top snippets of query b, denoted as $freq(a@b)$. The *CODC* association score is defined as,

$$S_{CODC}(q_f, q_e) = \begin{cases} 0, \text{ if } freq(q_e@q_f) \cdot freq(q_f@q_e) = 0; \\ \exp\left\{\log_{10}\left[\frac{freq(q_e@q_f)}{freq(q_f)} \times \frac{freq(q_f@q_e)}{freq(q_e)}\right]^{\epsilon}\right\}, \text{ otherwise.} \end{cases} \tag{10}$$

Note that a *CODC* value ranges between 0 and 1. In one extreme case, where $freq(q_e@q_f) = 0$ or $freq(q_f@q_e) = 0$, q_e and q_f have no association; in the other extreme case, where $freq(q_e@q_f) = freq(q_f)$ and $freq(q_f@q_e) = freq(q_e)$, they have the strongest association. In our experiment, ϵ is set at1 0.15, following Chen et al. [2006].

In addition to this frequency feature, any mined query q_e 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 a dictionary (see Section 3.3.1), a parallel corpus (see Section 3.3.2) and Web mining (see Section 3.3.3), the monolingual query suggestion system (see Section 3.2) is invoked to produce more related queries in the target language. For each monolingually suggested query q_e in target language, its monolingual source query $SQ_{ML}(q_e)$ is defined as the query in Q_0 having the highest monolingual similarity with q_e, which is given as:

$$SQ_{ML}(q_e) = argmax_{q'_e \in Q_0} sim_{ML}(q_e, q'_e). \tag{11}$$

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

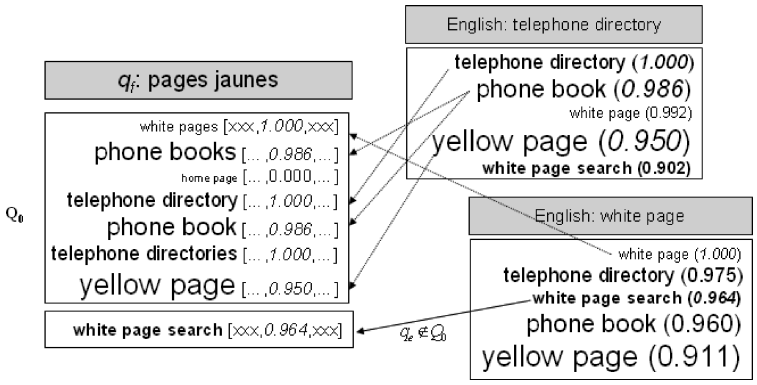

Fig. 2. An illustration on how the CLQS candidate set, Q_0, of the French query "pages jaunes" can be updated or replenished by the monolingual query suggestions of the candidates "telephone directory" and "white page". Note that the queries are normalized, and plurals and non-plurals are treated as the same.

its *Monolingual Query Suggestion Feature* is set to 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 $SQ_{ML}(q_e)$.

Following the French query example "pages jaunes" in Figure 1, we use Figure 2 to illustrate how the CLQS candidate set Q_0 can be replenished by the monolingual query suggestions of the candidates in Q_0 and how their feature values can be set. Suppose Q_0 is initially constructed as shown in the left hand side of Figure 1. As shown in Figure 2, the query "white page search" is added to Q_0, and its monolingual query suggestion feature value is set to 0.964, which is the highest with its monolingual source query "white page"; the other feature values of "white page search" are set to the same values as those of "white page".

3.4 Learning Cross-Lingual Query Similarity Measurement

In summary, four categories of features are used to learn the cross-lingual query similarity. The SVM regression algorithm [Smola and Scholkopf 2004] is adopted to learn the weights in Equation 2. In this study, the LibSVM[6] toolkit [Chang and Lin 2001] is employed for the regression training.

In the prediction stage, the candidate queries are ranked using the cross-lingual query similarity score computed using $sim_{CL}(q_f, q_e) = \mathbf{w} \cdot \phi(\mathbf{f}(q_f, q_e))$; the queries with similarity score lower than a threshold will be regarded as non-relevant. The threshold is learned using a development dataset by fitting MLQS's output. We first divided the CLQS candidates into two categories: relevant if a CLQS is in the set of MLQS, and nonrelevant otherwise. A binary classification model was then trained. The relevancy threshold on the predicted cross-lingual query similarity is determined as the decision boundary of the classifier.

[6]http://www.csie.ntu.edu.tw/ cjlin/libsvm/

4. CLIR BASED ON CROSS-LINGUAL QUERY SUGGESTION

In Section 3, we presented a discriminative model for cross-lingual query suggestion. For benchmarking purposes, we compare the effectiveness of CLQS with query translation for CLIR tasks. The resulting good performance of CLIR presumably 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 is recommended using the cross-lingual query suggestion system. The suggested queries in $\{q_e\}$ are concatenated to form a long query to retrieve documents. The advantage of this method over the retrieve-then-combine approach is that one can naturally regard the suggested queries as the user's information needs as a whole. This resembles the way in which query expansions works by considering feedback terms as natural extensions of the original query. For retrieval purposes, three different and widely used IR models are employed in our CLIR experiments: the BM25 probabilistic model [Robertson et al. 1995], the language modeling-based IR model [Ponte and Croft 1998; Zhai and Lafferty 2001b], and the vector space model [Salton and Buckley 1988].

5. PERFORMANCE EVALUATION

We benchmark the cross-lingual query suggestion system by comparing its effectiveness with monolingual query suggestion. We study the contribution of different information sources, and test their effectiveness in CLIR tasks. The language pairs concerned are French-English and Chinese-English. Such selection is due to the fact that large-scale query logs are readily available for these two language pairs. Moreover, English is considered as correlated with French more strongly than with Chinese. Thus, we can assume that there is stronger correspondence between the input French queries and the English queries in the query log and that the correspondence between Chinese and English queries is weaker. This enables us to study the effect of different language pairs in CLQS-based CLIR. Note that French-English (Chinese-English) denotes using French (Chinese) as the source language and English as the target language.

5.1 Data Resources

5.1.1 *English Query Log.* We used a one-month English query log of the *MSN* search engine (now *Live Search*) in the year 2005 as the target-language log. The log contained over 7.01 million unique English queries with occurrence frequency more than 10. A monolingual query suggestion system was built based on the method described in Section 3.2. For all the French-English and Chinese-English experiments, we used the same English query log for mining CLQS candidates.

5.1.2 *French-English Data.* In addition to the English query log, we obtained a French query log containing over 3 million queries, from which we selected a set of source queries to build a corpus for learning the CLQS model. First, we randomly selected 20,000 French queries from the French log to form

a query pool, and automatically translated them into English with Google's machine translation tool. We found that 42.17% (8,433) of the French queries had corresponding translations in the English query log. Among these French-English query pairs, professional translators then manually selected 4,171 pairs of correct translations. Only these selected query pairs were adopted for learning. Among them 70% were used for cross-lingual query similarity training, 10% were used as the development data to determine the relevancy threshold, and 20% were 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* parallel corpus [Koehn 2005] (with about 1 million French-English parallel sentences from the proceedings of the European Parliament) were also used.

In addition to benchmarking CLQS as an independent system, the CLQS system was also evaluated as a query translation system for CLIR tasks. The goal was to measure the quality of CLQS in terms of its effectiveness for CLIR. The TREC-6 CLIR dataset (AP88-90 English newswire, 750MB) and the officially-provided 25 short French-English query pairs (CL1-CL25) [Schauble and Sheridan 2000] were used for benchmarking. The average length of the title queries in the set is 3.3 words, which matches the Web queries used to train the CLQS model.

5.1.3 *Chinese-English Data.* We obtained a small Chinese query log from the same period of time with 32,730 queries. From that we selected source queries. First, machine translation was applied to translate the queries into English. We found 21.41% (7,008) Chinese queries, which had corresponding translations in the English query log. We then manually checked these translations and selected 3,767 correct Chinese-English query pairs that were used for CLQS model training (70%), testing (20%), and development (10%).

To retrieve CLQS candidates, we employed a built-in-house Chinese-English bilingual lexicon containing 940,000 unique entries and the LDC's Hong Kong parallel corpus (*Catalog No. LDC2004T08*) with about 3 million parallel sentences.

In CLIR experiments, we performed the NTCIR-4's Chinese-English CLIR task [Kishida et al. 2004]. The English documents were three subsets of the test collection, including the news of 1998-99 from Mainichi Daily News, Korea Times, and Xinhua News Agency. There were about 240,490 documents. There were 60 search topics (001-060) that were provided with their translations, and the title field of each topic was selected as the query for retrieval. The average length of the Chinese title queries was 4.4 words, a little longer than the TREC-6 queries, but it was still close to the length of Web queries. NTCIR provides two kinds of relevance judgment: "Relaxed" relevance and "Rigid" relevance. We based our evaluation on the "Rigid" judgment files.

Before translation, a Chinese query must be appropriately segmented into a sequence of meaningful words. *MSRSeg*[Gao et al. 2005], a state-of-the-art Chinese word segmenter, was used for this purpose. *MSRSeg* provides a pragmatic mathematical framework to unify five sets of fundamental features of word-level Chinese language processing: lexicon word processing, morphological

Table I. Main Data Resources Employed in Our Experiments. Both CLQS and CLQS-Based CLIR Experiments Used the CLQS Model Trained on 70% of the Query Translation Pairs Compiled by Human Experts to Generate Cross-Lingual Query Suggestions

	French-English	Chinese-English
# queries in target-language log	7.01 million	7.01 million
# translation pairs by expert	4,171	3,767
% of pairs for CLQS training	70% of 4,171 (2,920)	70% of 3,767 (2,637)
% of pairs for CLQS development	10% of 4,171 (417)	10% of 3,767 (377)
% of pairs for CLQS testing	20% of 4,171 (834)	20% of 3,767 (753)
Size of bilingual dictionary	120,000 entries	940,000 entries
Size of parallel corpus	1 million sentences (Europarl corpus)	3 million sentences (LDC HK parallel corpus)
# CLIR query pairs	25 (TREC-6)	60 (NTCIR-4)
CLIR document collection	AP news (1988-90)	Mainichi Daily News, Korea Times, Xinhua News (1998-99)

analysis, factoid detection, named entity recognition, and new word identification.

Table I summarizes these data resources.

5.2 Performance of Cross-Lingual Query Suggestion

5.2.1 *Performance Measure.* Mean-square-error (*MSE*) was used to measure the regression error; it is defined as:

$$MSE = \frac{1}{l}\sum_{i,j}\left[sim_{CL}\left(q_f^i, q_e^{ij}\right) - sim_{ML}\left(T_{q_f^i}, q_e^{ij}\right)\right]^2, \tag{12}$$

where i is the index of the i-th source query in the testing data, j is the index of the suggested queries of the i-th query, and and l is the number of cross-lingual query pairs.

A relevancy threshold was learned using the development data (see Section 3.4). Only CLQS with similarity value above the threshold was regarded as relevant to the input query. In this way, CLQS was evaluated as a classification task using precision (P) and recall (R), which are defined as:

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

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

5.2.2 *CLQS Performance.* The French-English and Chinese-English CLQS results with different feature configurations are shown in Table II and Table III, respectively.

The baseline system (DD) used a conventional query translation approach, a bilingual dictionary for co-occurrence-based translation disambiguation (see Section 3.3.1). For French-English CLQS in Table II, the baseline system only covered less than 10% of the suggestions made by MLQS (see recall). Using

Table II. French-English 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)

Features	Regression	Classification	
	MSE	Precision	Recall
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+MLQS	0.174	0.796	0.421

Table III. Chinese-English CLQS Performance with Different Feature Settings

Features	Regression	Classification	
	MSE	Precision	Recall
DD	0.236	0.854	0.149
DD+PC	0.236	0.892	0.212
DD+PC+Web	0.202	0.824	0.261
DD+PC+Web+MLQS	0.166	0.883	0.442

additional features enabled CLQS to generate more relevant queries. The most significant improvement on recall was achieved by exploiting MLQS. The final CLQS system generated 42.1% of the queries suggested by MLQS. There was no significant difference in precision among all the feature combinations. The performance of Chinese-English CLQS in Table III showed a similar trend as Table II. This indicated that our method could improve recall without loss of accuracy by effectively leveraging different information sources.

The regression performance was improved with additional features and was consistently reflected by the decrease in regression error (*MSE*). This was because the CLQS system increasingly enhanced the cross-lingual query similarity estimation by better fitting the monolingual query similarity with the help of additional information sources.

Chinese-English CLQS performed unexpectedly well. Compared to French-English performance, the high recall values of Chinese-English CLQS were likely the result of the large size of the bilingual dictionary and parallel corpus.

In addition to comparing CLQS output with the MLQS output, 200 French queries were randomly selected from the pool of 20,000 French queries. They were double-checked to make sure that they were not in the CLQS training corpus. The CLQS system was then used to suggest relevant English queries. On average, for each French query, 8.7 English queries were suggested. A total of 1,740 suggested English queries were manually cross-validated by two professional translators. Among the 1,740 suggested queries, 1,407 queries were deemed as relevant to their original counterparts, hence the accuracy was 80.9%. Figure 3 shows an example of CLQS of the French query “terrorisme international” (“international terrorism” in English), among which the queries suggested for the English translation “international terrorism” by MLQS are displayed in bold.

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)	

Fig. 3. An example of CLQS of the French query "terrorisme international", where the queries suggested by MLQS are shown in bold.

nba michael jordan retired (0.988);	**nba michael jordan retirement** (0.987);
michael and jordan and retired (0.980);	**michael jordan retirement ceremonies** (0.911);
jordan michael (0.843);	**michael jordan** (0.817);
nba jordan retirement (0.799);	nba jordan retired (0.799);
life of michael jordan (0.697);	chicago bulls (0.694)

Fig. 4. An example of CLQS of the Chinese query "NBA麦可 乔丹 退休", where the queries suggested by MLQS are shown in bold.

We then conducted a similar human evaluation for 60 Chinese queries. On average, there were 14.8 English queries suggested for each Chinese query by the system. The total number of suggested queries was 885, among which 748 queries were considered relevant. Therefore, the accuracy of Chinese-English CLQS was 84.5%. Figure 4 shows an example of CLQS of the Chinese query "NBA麦可乔丹退休" ("NBA Michael Jordan retirement").

5.3 CLIR Performance

CLQS was evaluated for French-English (F2E) and Chinese-English (C2E) CLIR tasks. We conducted F2E and C2E experiments using the TREC-6 and NTCIR-4 CLIR datasets (see Section 5.1), respectively.

CLIR was performed using a query translation system followed by a monolingual IR module based on Lemur's toolkit.[7] Three typical retrieval models were studied separately: BM25 [Robertson et al. 1995], language modeling-based IR (LM) [Ponte and Croft 1998; Zhai and Lafferty 2001b], and TFIDF vector space model (TFIDF) [Salton and Buckley 1988]. The following three systems were used to perform query translation.

(1) *CLQS. Our CLQS systems*. The F2E and C2E CLQS models were trained on the 70% of human expert compiled French-English and Chinese-English query translation pairs (see Section 5.1.2 and 5.1.3) with all the features (see Section 3.3) configured.

(2) For F2E, we used the Moses translation engine [Koehn et al. 2007], a phrase-based SMT system based on the source-channel formalism [Och 2002; Koehn et al. 2003], denoted as *SMT (Moses)*. For C2E, we used a built-in-house SMT system [Li et al. 2007; Zhang et al. 2008], denoted as *SMT (MSRA)*, which also adopted a phrase-based translation model. The two systems represented the state-of-the-art SMT tools for French-English

[7]http://www.lemurproject.org/

Table IV. Average Precision of French-English CLIR on TREC-6 Dataset (Monolingual: Monolingual IR System; DT: CLIR Based on Dictionary Translation; SMT (Moses): CLIR Based on Moses Statistical Machine Translation Engine; CLQS: CLQS-Based CLIR). IR Models are Tuned to Nearly Their Optimal Performance—BM25: $k_1 = 1.2$, $b = 0.75$, $k_3 = 7$; LM: Language Modeling with Jelinek-Mercer (Interpolate) Smoothing; TFIDF: Query Term TF Weighting Method—Raw-TF, Document Term TF Weighting Method—Log-TF

CLIR systems	BM25		LM		TFIDF	
	Average Precision	% of monolingual	Average Precision	% of monolingual	Average Precision	% of monolingual
Monolingual	0.2954	100%	0.2844	100%	0.2739	100%
DT	0.2130	72.11%	0.2115	74.37%	0.1958	71.49%
SMT (Moses)	0.2545	86.15%	0.2412	84.81%	0.2448	89.38%
CLQS	0.2916	98.71%	0.2698	94.87%	0.2585	94.38%

and Chinese-English translation, and were trained on the corresponding sets of parallel corpora used by our CLQS systems (*Europarl* for F2E and *LDC's Hong Kong corpus* for C2E).

(3) *DT*. A dictionary-based query translation system using co-occurrence statistics for translation disambiguation [Ballesteros and Croft 1998; Gao et al. 2001] was applied to the query log (see Section 3.3.1). Especially for C2E CLIR, we implemented the approach in Zhang and Vines [2004] to automatically extract OOV translations for Chinese queries from the Web, denoted as *DT (Web)*. This represented the state-of-the-art Web mining approach for dictionary-based query translation.

The monolingual IR performance using the standard target language queries was also reported as a reference.

5.3.1 *F2E CLIR Performance.* The average precision of the three F2E CLIR and the monolingual IR systems are reported in Table IV using different retrieval models.

The result on BM25 retrieval showed that using CLQS as a query translation tool outperformed CLIR based on dictionary translation, by 36.9% (relative improvement, $(0.2916 - 0.213)/0.213$), and machine translation by 14.58%. It achieved 98.71% of the monolingual IR performance. Consistent results were obtained using language modeling and the TFIDF vector space model for retrieval. Using language-modeling-based retrieval with Jelinek-Mercer (interpolate) smoothing, CLQS outperformed dictionary-based query translation by 27.57%, as well as machine translation by 11.86%, and achieved 94.87% of the monolingual IR performance. Using the TFIDF vector space model, CLQS outperformed the dictionary-based method by 32.02%, as well as machine translation by 5.6%, and achieved 94.38% of monolingual IR performance. This showed a consistent advantage of CLQS-based CLIR over the other traditional query translation approaches. We further conducted tests for significance (two-tailed pairwise student's t-test) [Hull 1993] on the results of different approaches. The p-values shown in Table V suggest that the performance of CLQS-based CLIR was significantly better at 95% confidence level.

The effectiveness of CLQS lies in its ability to suggest closely related queries rather than accurate translations. For example, consider the query CL14

Table V. The p-Values Results from Pair-Wise Significance t-Tests for Different French-English CLIR Systems. The Confidence Level is Set as 95% ($p < 0.05$ are Considered Statistically Significant)

	BM25		LM		TFIDF	
	DT	SMT (Moses)	DT	SMT (Moses)	DT	SMT(Moses)
CLQS	0.018	0.039	0.028	0.042	0.023	0.047

"terrorisme international" ("international terrorism"). Although MT translated the query correctly, the CLQS system still achieved a higher score by recommending many additional related terms such as "global terrorism", "world terrorism", and so on. (see Figure 3). For another example, consider the query CL6 "La pollution causée par l'automobile" ("air pollution due to automobile"). The Moses SMT provided the translation "the pollution caused by cars", but the CLQS system enumerated all possible synonyms of "car", and suggested the queries "car pollution", "auto pollution", "automobile pollution". In addition, other related queries such as "global warming" were also suggested, resulting in an analogous effect of query expansion. For the query CL12 "la culture écologique" ("organic farming"), Moses translated it as "ecological culture", which was not the term used in English. Thus it failed to generate the correct translation and to find the relevant documents. Although the correct translation was nor in our French-English dictionary either, the CLQS system generated "organic farm" as a relevant query due to successful Web mining.

5.3.2 *F2E CLIR Performance with Pseudo-Relevance Feedback.* These experiments demonstrated the effectiveness of using CLQS to suggest relevant queries for CLIR enhancement. A related study was to adopt query expansion to enhance CLIR effectiveness [Ballesteros and Croft 1997; McNamee and Mayfield 2002]. Pseudo-relevance feedback (PRF) is widely used to obtain more alternative query expressions from retrieved documents. Practically, our approach aims to obtain similar effects. Thus, we compared the CLQS approach with the conventional query expansion approaches. Following McNamee and Mayfield [2002], post-translation expansion was performed based on PRF techniques. We first performed CLIR in the same way as before, using different retrieval models. We then applied the traditional PRF algorithms corresponding to the different retrieval models to perform post-translation expansion. Table VI shows the corresponding feedback models with respect to different retrieval models.

For the BM25 model, we used the method described in Robertson [1990] to select expansion terms. In our experiments, the top 10 to 200 terms were selected based on RSV (see Table VI) from the top 30 feedback documents, to expand the original query for the comparison between CLQS and the baseline approaches. For the language modeling approach, PRF was done by using a mixture feedback model described in Zhai and Lafferty [2001a]. Unlike the PRF of BM25, the mixture model updates the query's language model instead of query terms using feedback documents. In addition to varying the number of feedback terms (which is the threshold to truncate the feedback model to no more than the given number of terms), we also examined the influence of the feedback model by changing the coefficient α, which controlled the extent

Table VI. The Representative Relevance Feedback Formulations Corresponding to the Three Typical Retrieval Models: BM25, Language-Modeling-Based Retrieval (LM), and TFIDF Vector Space Model

IR Model	Relevance Feedback Model	Reference
BM25	$$RSV_i = w_i \cdot r_i / R \quad (13)$$ $$w_i = \log \frac{(r_i + 0.5)/(R - r_i + 0.5)}{(n_i - r_i + 0.5)/(N - n_i - R - r_i + 0.5)},$$ where RSV_i is the Robertson Selection Value (RSV) for term i; w_i is the Robertson-Sparck Jones relevance weight [Robertson and Jones 1976] of the term; r_i is the number of relevant documents for the query containing the term; R is the total number of relevant documents for the query; n_i is the number of documents in the collection containing the term; N is the number of indexed documents in the collection.	[Robertson 1990]
LM	$$\hat{\theta}_{Q'} = (1-\alpha)\hat{\theta}_Q + \alpha\hat{\theta}_F \quad (14)$$ $$\hat{\theta}_F \propto \log p(F|\theta) = \sum_i \sum_w c(w; d_i) \log((1-\lambda)p(w|\theta) + \lambda p(w|C)),$$ where $\hat{\theta}_{Q'}$ is the updated query model based on the original query model $\hat{\theta}_Q$ and feedback model $\hat{\theta}_F$; α is the coefficient controlling the influence of the feedback model; F is the set of feedback documents; $p(F\|\theta)$ is a mixture model used to estimate the feedback model; λ is the parameter controlling the influence of background noise when generating a feedback document.	[Zhai and Lafferty 2001a]
TFIDF	$$Q_1 = Q_0 + \beta \sum_{k=1}^{n_1} \frac{R_k}{n_1} - \gamma \sum_{k=1}^{n_2} \frac{S_k}{n_2}, \quad (15)$$ where Q_1 is the new query vector, Q_0 is the initial query vector, R_k (S_k) is the vector for relevant (non-relevant) document k, n_1 (n_2) is the number of relevant (non-relevant) documents, and β (γ) is the parameter that controls the relative contribution of relevant (non-relevant) documents.	[Rocchio 1971]

of inclusion of the feedback model. For the TFIDF vector space model, we expanded the queries using the traditional Rocchio's algorithm [Rocchio 1971] associated with the vector space model (for pseudo feedback, β was set to 1 and γ to 0). Through this manual tuning, the three PRF approaches were tuned to their best possible performance. CLIR performances with PRF in terms of average precision using different IR models are shown in Figures 5–8.

These results showed that the CLQS-based CLIR consistently outperformed the other methods when PRF was incorporated. Even though PRF was not added to CLQS-based CLIR (with zero feedback term), it still performed better than the other two translation approaches plus PRF (with 10+ feedback terms)

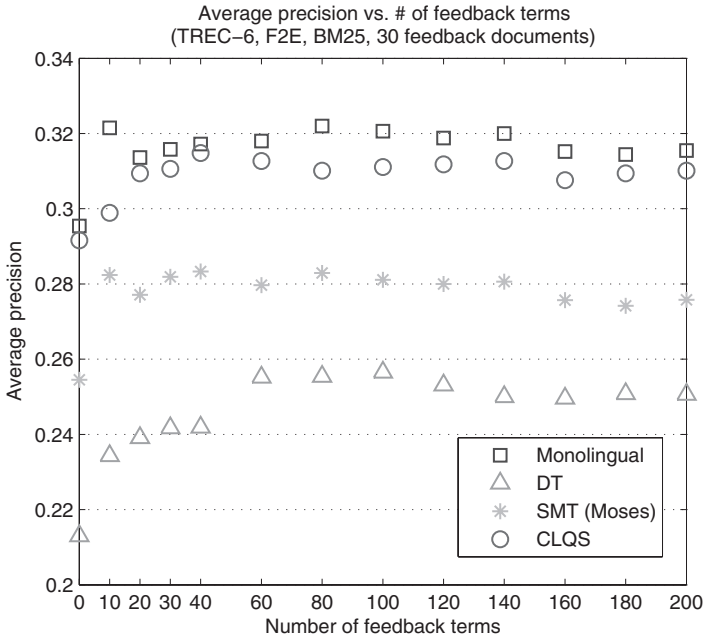

Fig. 5. Average precision of post-translation expansion using PRF varies with the number of expansion terms on the TREC-6 French-English dataset (BM25).

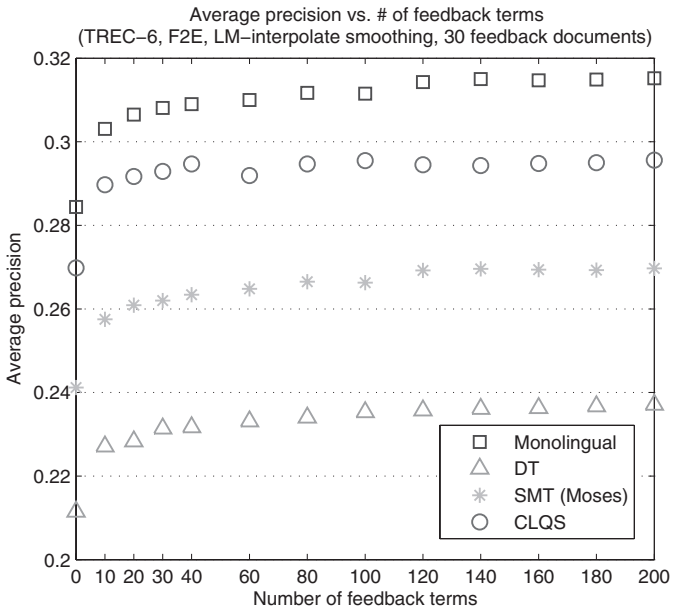

Fig. 6. Average precision of post-translation expansion using PRF changes with the number of feedback terms on the TREC-6 French-English dataset (LM with interpolate smoothing, $\alpha = 0.5$, $\lambda = 0.7$).

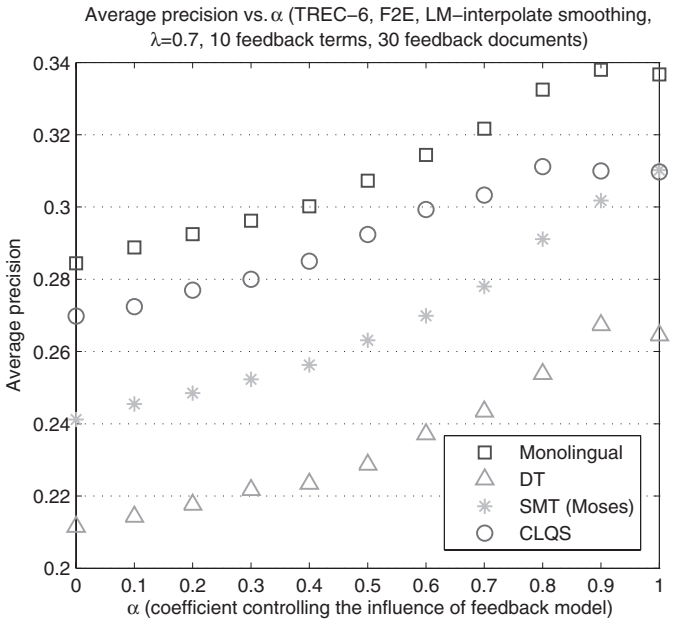

Fig. 7. Average precision of post-translation expansion using PRF changes with the feedback coefficient α on the TREC-6 French-English dataset (LM with interpolate smoothing).

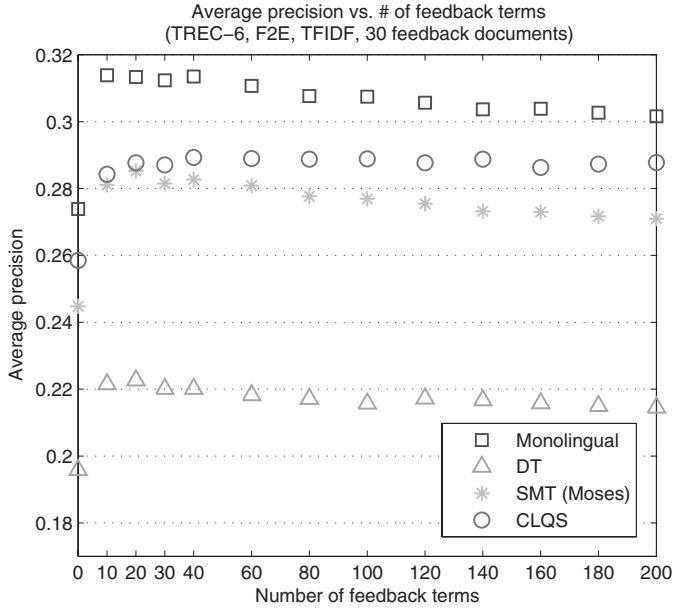

Fig. 8. Average precision of post-translation expansion using PRF changes with the number of expansion terms on the TREC-6 French-English dataset (TFIDF vector space model).

when using BM25 (see Figure 5) and language modeling (see Figure 6). In this regard, however, the performance gain was not shown as significant by a t-test. We then conducted t-tests with PRF added to CLQS-based retrieval. We found that CLQS-based CLIR with PRF was significantly better than DT-based CLIR with PRF for all the examined number of feedback terms ($p < 0.05$), and was also significantly better than SMT-based retrieval with PRF in most cases, except for BM25 (see Figure 5) using 10 feedback terms ($p = 0.095$) and TFIDF (see Figure 8) using less than 60 feedback terms (p varies from 0.112 to 0.073).

The results indicated the higher effectiveness of CLQS in related query identification by leveraging a wide range of resources. Post-translation expansion was capable of improving CLQS-based CLIR. This is due to the fact that CLQS and PRF leverage different categories of resources, and both approaches can be complementary. However, the t-test showed that CLQS-based CLIR with PRF was not significantly better than using CLQS alone, and was not always significantly better than other query translation approaches plus PRF especially when only a small number of feedback terms was involved. This may reflect the fact that the related query terms suggested by CLQS from the query log overlapped with the feedback terms from the retrieved documents, and other approaches did not. Thus, introducing a small number of feedback terms was not as helpful to CLQS-based retrieval as to the CLIR based on other query translation approaches. On the other hand, because the queries suggested by CLQS were closely related to the original query, the concatenated long query updated by PRF tended to be more robust to the noise introduced by the feedback process than other query translation approaches. This effect can be observed where the number of feedback terms increases and no drop in average precision was seen for CLQS (see Figures 5 and 8).

Using language-modeling-based IR (see Figure 6), however, CLQS was significantly better than other translation approaches regardless of the number of feedback terms. Note that the number of feedback terms in the language modeling approach was used to truncate the feedback model (see Equation 14) to no more than the given length, instead of the number of terms to add to the original query. It seemed that interpolating the query model with the feedback model improved the effectiveness of CLQS-based CLIR and other query translation approaches to a similar extent, given the same truncating threshold. We leave the specific reason to future study. In addition, average precision stopped increasing for all approaches after a certain number of feedback terms was used. This is because the feedback model was truncated when the sum of the probability of the included words reached the default threshold of 1.

We also found that CLQS-based CLIR did not need to heavily rely on a feedback model to boost retrieval performance. This is reflected in Figure 7 where the performance of CLQS-based CLIR began to decrease when the influence factor of the feedback model got to $\alpha = 0.9$. This implies that PRF was less useful to CLQS-based CLIR than to other query translation approaches since a certain amount of performance gain was due to the suggested queries themselves.

Table VII. Average Precision of Chinese-English CLIR (Rigid Test) on NTCIR-4 Dataset (Monolingual: Monolingual IR System; DT: CLIR Based on Dictionary Translation; DT (Web): CLIR Based on Dictionary Translation with OOV Query Translations Mined from Web; SMT (MSRA): CLIR Based on MSRA Statistical Machine Translation Engine; CLQS: CLQS-Based CLIR). IR Models are Tuned to Nearly their Best Performance—BM25: $k_1 = 1.2$, $b = 0.75$, $k_3 = 7$; LM: Language Modeling with Jelinek-Mercer (Interpolate) Smoothing; TFIDF: Query Term TF Weighting Method—Raw-TF, Document Term TF Weighting Method—log-TF

CLIR systems	BM25		LM		TFIDF	
	Average Precision	% of Monolingual	Average Precision	% of Monolingual	Average Precision	% of Monolingual
Monolingual	0.1857	100%	0.1729	100%	0.1733	100%
DT	0.1416	76.25%	0.1302	75.30%	0.1314	75.82%
DT (Web)	0.1564	84.22%	0.1448	83.75%	0.1453	83.84%
SMT (MSRA)	0.1545	83.20%	0.1438	83.17%	0.1389	80.15%
CLQS	0.1720	92.62%	0.1680	97.17%	0.1652	95.33%

Table VIII. The p-Values Result from Pairwise Significance t-Tests for Different Chinese-English CLIR Systems. The Confidence Level is Set as 95% ($p < 0.05$ are Considered Statistically Significant)

	BM25		LM		TFIDF	
	DT (Web)	SMT (MSRA)	DT (Web)	SMT (MSRA)	DT (Web)	SMT (MSRA)
CLQS	0.012	0.027	0.0014	0.0006	0.0004	0.0013

5.3.3 *C2E CLIR Performance.* The average precision of the four C2E CLIR and the monolingual IR systems are reported in Table VII in terms of different retrieval models.

Consistent with F2E CLIR (see Section 5.3.1), the higher effectiveness of C2E CLIR based on CLQS shed more light on the advantage of CLQS over the other traditional query translation approaches. When using BM25, CLQS-based CLIR outperformed dictionary-based query translation by 21.47%, dictionary method with OOV translation mining by 9.97%, as well as SMT-based query translation by 11.33%; and achieved 92.62% of the monolingual IR performance. When using the language modeling, CLQS-based CLIR outperformed the dictionary-based query translation by 29.03%, dictionary-based query translation plus OOV translation mining by 16.02%, as well as SMT-based query translation by 16.83%; and achieved 97.17% of the monolingual IR performance. When using the TFIDF vector space model, CLQS-based CLIR outperformed the dictionary-based method by 25.72%, the dictionary-based method with OOV translation mining by 13.7%, as well as SMT-based query translation by 18.93%; and achieved 95.33% of monolingual IR performance.

In addition, dictionary-based query translation performed better than machine translation when the OOV translations mined from the Web were added to the dictionary. The machine translation method, however, was constrained by the coverage of the parallel corpus, and could not deal with OOV translations effectively. CLQS leveraged different resources including Web mining of OOV translations to find relevant queries from the query log, and covered more relevant information than accurate query translation did. The t-test results shown in Table VIII demonstrate that the high effectiveness of CLQS-based CLIR was statistically significant.

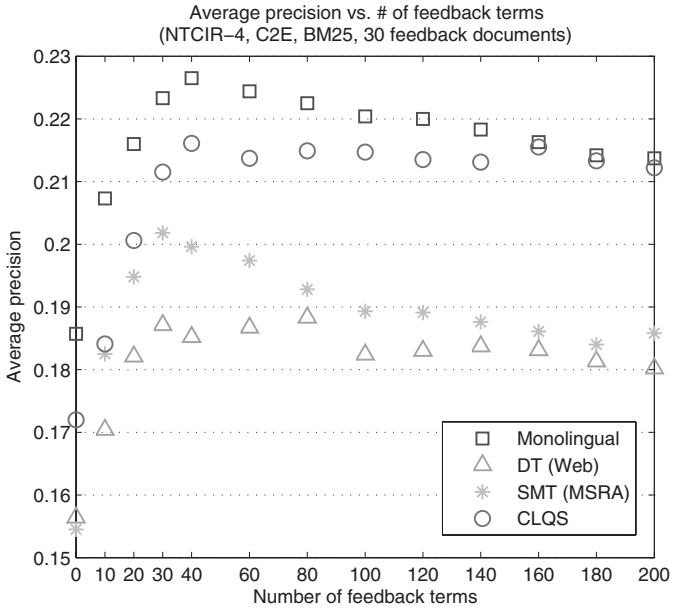

Fig. 9. Average precision of post-translation expansion using PRF changes with the number of expansion terms on NTCIR-4 Chinese-English (rigid test) dataset (BM25).

For more illustrations, we show some examples from NTCIR-4's query set. For query 005 "戴奥辛 人体 影响 威胁" ("dioxin human body effect threat") where "戴奥辛" ("dioxin") is an OOV term. Neither DT nor SMT (MSRA) correctly translated "戴奥辛" as "dioxin"; but both DT (Web) and CLQS did, as they identified the translation pair from the Web corpora. CLQS further suggested related queries in addition to the translated query, such as "how drugs affect the body", "estimated human body burdens dioxin-like chemicals", and "food chain", and so on. For query 030 "动物复制技术" ("animal cloning technique"), none of the methods, except for CLQS, generated queries with the term "clone" because "clone", was not a translation entry of "复制" ("reproduction") in our bilingual resources, nor did they co-occur frequently on the Web (what co-occurs more often is "克隆" and "clone"). CLQS correctly suggested "animal cloning technology" as it had a high similarity with "animal reproduction technology clone" in the query log, and MLQS successfully retrieved it from the query log by using "animal reproduction technology", the exact translation of the original query.

5.3.4 *C2E CLIR Performance with Pseudo-Relevance Feedback.* Under similar settings (see Section 5.3.2), we compared the average precisions of these different C2E CLIR systems with PRF added. The results are shown in Figures 9–12.

The results demonstrated that when PRF was performed, CLQS-based CLIR was consistently better than the other approaches for a different language pair. In particular, when PRF was not used in CLQS-based CLIR (with zero feedback term), it still outperformed other query translation approaches plus

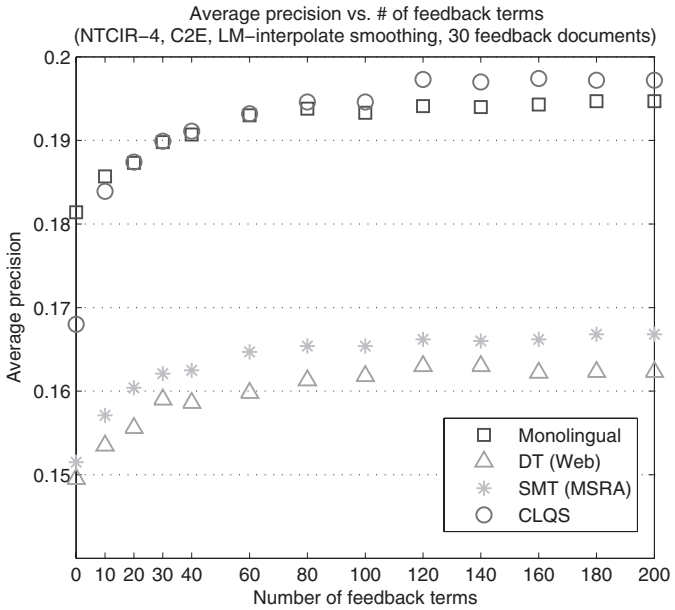

Fig. 10. Average precision of post-translation expansion using PRF changes with the number of expansion terms on NTCIR-4 Chinese-English (rigid test) dataset (LM with interpolate smoothing, $\alpha = 0.5$, $\lambda = 0.7$).

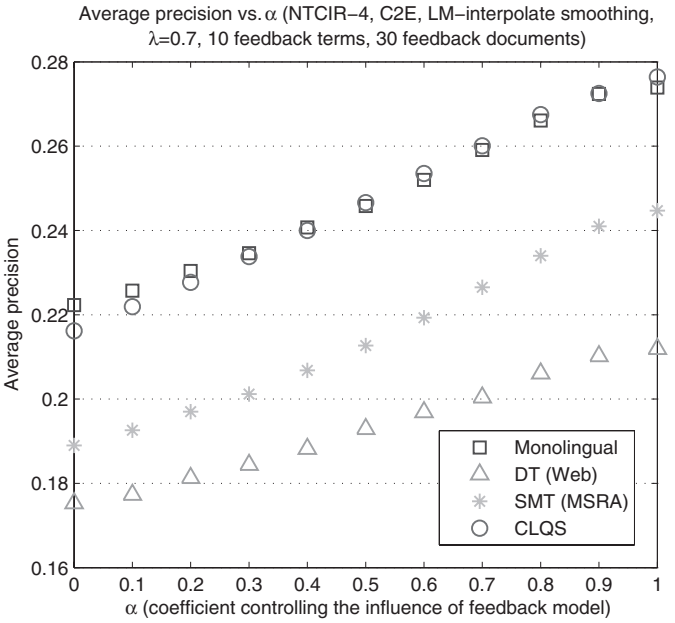

Fig. 11. Average precision of post-translation expansion using PRF changes with the feedback coefficient on NTCIR-4 Chinese-English (rigid test) dataset (LM with interpolate smoothing).

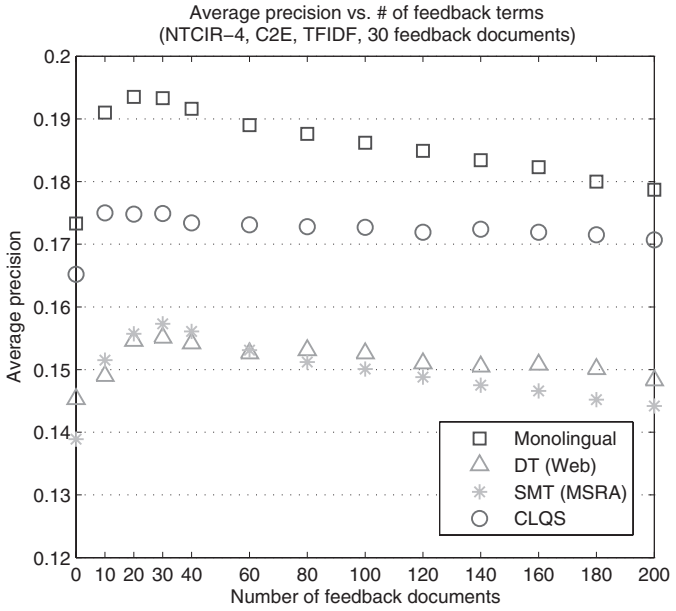

Fig. 12. Average precision of post-translation expansion using PRF changes with the number of expansion terms on NTCIR-4 Chinese-English (rigid test) dataset (TFIDF vector space model).

PRF (with 10+ feedback terms) for language modeling (see Figure 10) and TFIDF (see Figure 12) except for BM25 (see Figure 9). Similarly, t-tests did not show significant performance gain in this regard, but when adding PRF on all retrieval models, CLQS-based CLIR performed significantly better than DT (Web) using any number of feedback terms, and also significantly better than SMT (MSRA) in most cases (p varied from 0.012 to 0.035), except for BM25 using less than 40 feedback terms.

As distinct from F2E, a t-test between CLQS-based CLIR with and without PRF showed that PRF was not only useful to the CLQS-based approach, but also performed significantly better provided that the appropriate number of feedback terms was used. For example, when more than 20 terms were introduced in the case of BM25 with PRF, the average precision was significantly higher than that of CLQS alone ($p < 0.003$). Such significant improvement was also observed in language modeling as well as in TFIDF with more than 10 feedback terms. This was because C2E CLQS, although effective, could not suggest closely related queries as effectively as its F2E counterpart. Unlike the French queries, the Chinese queries corresponded less strongly to the queries in the English query log due to the wider linguistic gap and the fewer common search interests of users between the two locales. Thus it was generally harder to find the correspondences of a Chinese query from the English query log than in the F2E case. This observation was reflected by the estimated proportions of Chinese and French queries having corresponding translations in the English query log: 21.41% vs. 42.17% (see Sections 5.1.2 and 5.1.3). Therefore, the role of PRF was more important in C2E than in F2E for improving CLIR effectiveness.

Our conjecture was proven by the results shown in Figure 11 (compared to Figure 7), where the performance of CLQS-based CLIR increased monotonically with the increasing involvement of the feedback model. This implies that the performance gain increasingly came from the complementary effect of PRF.

We also noticed that the average precisions dropped after reaching peaks with around 30–40 feedback terms for BM25 and TFIDF models (see Figures 9 and 12). The trend of the drop-off was more noticeable than in the case of F2E. This was due to the errors made when segmenting the Chinese texts, which further resulted in many more noisy feedback terms in the expansion when the number of feedback terms used was large. It seemed that language-modeling-based retrieval was more robust to this kind of noise (see Figure 10). We tried to explain this distinct observation by the factors used to truncate the feedback model: the constraints on the number of terms used and the sum of the probability thresholds of these terms. Another possible reason was that the Kullback-Leibler divergence was not sensitive to the changes in the query model made by the noisy terms since their probability masses were tiny. The discussion of this problem is beyond our scope and is left to future work.

6. CONCLUSIONS

In this article, we proposed a new approach to cross-lingual query suggestion by mining relevant queries in different languages from query logs. Compared to query translation, our method can suggest not only better formulated queries, but also similar queries. The key to this approach is to learn a cross-lingual query similarity measure between the original query and the suggestion candidates. We proposed a discriminative model to determine such similarity by exploiting different types of monolingual and bilingual information. 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.

Our method differs from existing approaches for query suggestion and for query translation in the following aspects.

—We extended monolingual query suggestion to cross-lingual query suggestion. To our knowledge, this is the first attempt in this direction.
—We leveraged on the target-language query log to suggest more cohesive complete queries than by using a query translation approach.
—We proposed a discriminative method to learn to estimate cross-lingual query similarity instead of manually defining such a measure. This allowed us to not only obtain a more suitable similarity measure, but also to more easily adapt the approach to different language pairs.

In our experiments, we compared our approach with several baseline methods. The baseline CLQS system applied a typical query translation approach, using a bilingual dictionary with co-occurrence-based translation disambiguation. Benchmarked under French-English and Chinese-English settings, this baseline approach only covered 10–15% of the relevant queries suggested by a monolingual query suggestion system (when the exact translation of the

original query was given). By leveraging additional resources such as parallel corpora, Web mining, and query log-based monolingual query suggestion, the final system covered 42–44% of the relevant queries suggested by a monolingual query suggestion system with precision as high as 79.6% and 89.2% for French-English and Chinese-English tests, respectively.

To further evaluate the quality of the suggested queries, the CLQS system was used as a query translation system in the CLIR tasks. Using the TREC-6 French-English and NTCIR-4 Chinese-English CLIR tasks as benchmarks, CLQS consistently demonstrated higher effectiveness than traditional query translation methods using either bilingual dictionary or state-of-the-art statistical machine translation approaches. Three traditional information retrieval models: BM25, language modeling, and TFIDF vector space model, were adopted in the experiments.

The improvement on the TREC-6 French-English CLIR task by using CLQS demonstrated the high quality of the suggested queries. This also implied a strong correspondence between the input French queries and the English queries in the log. For queries of Chinese and English, which bear weaker correspondence in the log, CLQS performed surprisingly well due to the comprehensive bilingual data resources and the satisfactory coverage of the query logs.

Pseudo-relevance feedback (PRF) and CLQS both expanded the original query for improving CLIR performance. But they exploited different types of resources and distinctive mechanisms, and therefore could be complementary to each other. Interestingly, for French-English CLIR, the complementary effect from the pseudo feedback to CLQS was relatively smaller than that for Chinese-English CLIR. This was because French-English CLQS could suggest closely related queries more effectively from the English query log than in the Chinese-English case, due to the stronger correspondence between the search interests of users in French and English.

Our experimental results provided positive answers to the three aforementioned unresolved issues. (1) Compared with SMT-based query translation systems that were trained on the same sets of parallel corpora used by CLQS, our CLQS-based approach achieved superior CLIR performance; (2) PRF was consistently complementary to CLQS-based CLIR regardless of the underlying retrieval models adopted; (3) Across Chinese and English, two linguistically less correspondent languages, high-quality queries were suggested from the query logs. This is evidenced by the significant improvement in CLIR effectiveness.

7. FUTURE WORK

In this work, we have exploited several types of monolingual and cross-lingual information in cross-lingual query suggestion. However, more types of information can be integrated into the general framework for the estimation of cross-lingual query similarity. This is an interesting improvement for our future work. Improvements can also be made in the method for determining similar queries. For example, query popularity or click counts of queries can be explicitly taken into consideration so that the most popular (thus usual) query formulations can be suggested.

One of the key advantages of query logs is that they are up-to-date in terms of user needs and vocabulary. Our method also works well on standard text collections that are not necessarily aligned with the timeframe of query logs. This may be because our query log is newer (all queries were issued in the year of 2005) than the collections of news, which fell into 1988–90 and 1998–99. Our log is characterized with good backward compatibility with news that previously occurred. We found that nearly all the topics that the test queries can correlate to some entries in the English query log. On the other hand, our log also contains the queries that later turned out to become very popular. For example, although far from as popular as nowadays, queries on "Barack Obama" still frequently appear in this query log of early days. This suggests that query logs may have large intemporal value as a lexical resource. We would like to specifically study the temporal issues of exploiting query logs for query suggestion in future work.

ACKNOWLEDGMENTS

We would like to thank Jian Hu and Dongdong Zhang at Microsoft Research Asia for their generous help. Jian Hu provided us with an implementation of monolingual query suggestion based on the work of Wen et al. [2002]. Dongdong Zhang supported us by providing the Chinese-English SMT system for our comparative experiments. We also thank the anonymous reviewers for their valuable comments.

REFERENCES

AMBATI, V. AND ROHINI, U. 2006. Using monolingual clickthrough data to build cross-lingual search systems. In *Proceedings of ACM SIGIR Workshop on New Directions in Multilingual Information Access*.

BALLESTEROS, L. A. AND CROFT, W. B. 1997. Phrasal translation and query expansion techniques for cross-language information retrieval. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 84–91.

BALLESTEROS, L. A. AND CROFT, W. B. 1998. Resolving ambiguity for cross-language retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 64–71.

BROWN, P. F., PIETRA, D. S. A., PIETRA, D. V. J., AND MERCER, R. L. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computat. Ling. 19*, 2, 263–311.

CHANG, C. C. AND LIN, C. 2001. LIBSVM: a library for support vector machines (version 2.3). http://www.csie.ntu.edu.tw/~cjlin/libsvm/

CHEN, H.-H., LIN, M.-S., AND WEI, Y.-C. 2006. Novel association measures using Web search with double checking. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*. 1009–1016.

CHENG, P.-J., TENG, J.-W., CHEN, R.-C., WANG, J.-H., LU, W.-H., AND CHIEN, L.-F. 2004. Translating unknown queries with Web corpora for cross-language information retrieval. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 146–153.

CUI, H., WEN, J. R., NIE, J.-Y., AND MA, W.-Y. 2003. Query expansion by mining user logs. *IEEE Trans. Knowl. Data Engin. 15*, 4, 829–839.

FUJI, A. AND ISHIKAWA, T. 2000. Applying machine translation to two-stage cross-language information retrieval. In *Proceedings of 4th Conference of the Association for Machine Translation in the Americas (AMTA)*. 13–24.

GAO, J. F., LI, M., WU, A., AND HUANG, C.-N. 2005. Chinese word segmentation and named entity recognition: A pragmatic approach. *Computat. Ling. 31*, 4, 531–574.

GAO, J. F., NIE, J.-Y., HE, H., CHEN, W., AND ZHOU, M. 2002. Resolving query translation ambiguity using a decaying co-occurrence model and syntactic dependence relations. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 183–190.

GAO, J. F., NIE, J.-Y., XUN, E., ZHANG, J., ZHOU, M., AND HUANG, C. 2001. Improving query translation for CLIR using statistical models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 96–104.

GAO, W., NIU, C., NIE, J.-Y., ZHOU, M., HU, J., WONG, K.-F., AND HON, H.-W. 2007. Cross-lingual query suggestion using query logs of different languages. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 463–470.

GLEICH, D. AND ZHUKOV, L. 2004. SVD subspace projections for term suggestion ranking and clustering. Tech. rep. Yahoo! Research Labs.

HULL, D. 1993. Using statistical testing in the evaluation of retrieval experiments. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 329–338.

JEON, J., CROFT, W. B., AND LEE, J. 2005. Finding similar questions in large question and answer archives. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*. 84–90.

JIANG, M.-G., MYAENG, S. H., AND PARK, S. Y. 1999. Using mutual information to resolve query translation ambiguities and query term weighting. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*. 223–229.

JOACHIMS, T. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 133–142.

KISHIDA, K., CHEN, K.-H., LEE, S., KURIYAMA, K., KANDO, N., CHEN, H.-H., MYAENG, S. H., AND EGUCHI, K. 2004. Overview of CLIR task at the fourth NTCIR workshop. In *Proceedings of 4th NTCIR Workshop Meeting on Evaluation of Information Access Technologies*. 1–59.

KOEHN, P. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th Machine Translation Summit*. 79–86.

KOEHN, P., HOANG, H., BIRCH, A., CALLISON-BURCH, C., FEDERICO, M., BERTOLDI, N., COWAN, B., SHEN, W., MORAN, C., ZENS, R., DYER, C., BOJAR, O., CONSTANTIN, A., AND HERBST, E. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (demo)*. 177–180.

KOEHN, P., OCH, F. J., AND MARCU, D. 2003. Statistical phrase based translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*. 48–54.

KRAAIJ, W., NIE, J.-Y., AND SIMARD, M. 2003. Embedding Web-based statistical translation models in cross-language information retrieval. *Computat. Ling. 29*, 3, 381–419.

KWOK, K. L., CHOI, S., AND DINSTL, N. 2005. Rich results from poor resources: NTCIR-4 monolingual and cross-lingual retrieval of Korean texts using Chinese and English. *ACM Trans. Asian Lang. Inform. Proc. 4*, 2, 136–162.

LAVRENKO, V., CHOQUETTE, M., AND CROFT, W. B. 2002. Cross-lingual relevance models. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 175–182.

LI, C.-H., ZHANG, D., LI, M., ZHOU, M., LI, M., AND GUAN, Y. 2007. A probabilistic approach to syntax-based reordering for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. 720–727.

LÓPEZ-OSTENERO, F., GONZALO, J., AND VERDEJO, F. 2005. Noun phrases as building blocks for cross-language search assistance. *Inform. Proc. Manag. 41*, 549–568.

LU, W.-H., CHIEN, L.-F., AND LEE, H.-J. 2001. Anchor text mining for translation extraction of query terms. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 388–389.

McNamee, P. and Mayfield, J. 2002. Comparing cross-language query expansion techniques by degrading translation resources. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 159–166.

Monz, C. and Dorr, B. J. 2005. Iterative translation disambiguation for cross-language information retrieval. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 520–527.

Nie, J.-Y., Simard, M., Isabelle, P., and Durand, R. 1999. Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 74–81.

Och, F. J. 2002. Statistical machine translation: From single-word models to alignment templates. Ph.D. thesis, RWTH Aachen, Germany.

Och, F. J. and Ney, H. 2003. A systematic comparison of various statistical alignment models. *Computat. Ling. 29*, 1, 19–51.

Pirkola, A., Hedlund, T., Keshusalo, H., and Jarvelin, K. 2001. Dictionary-based cross-language information retrieval: Problems, methods, and research findings. *Inform. Retrieval 4*, 3/4, 209–230.

Ponte, J. and Croft, W. B. 1998. A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 275–281.

Robertson, S. E. 1990. On term selection for query expansion. *J. Docum. 46*, 359–364.

Robertson, S. E. and Jones, K. S. 1976. Relevance weighting of search terms. *J. Amer. Soc. Inform. Sci. 27*, 3, 129–146.

Robertson, S. E., Walker, S., Hancock-Beaulieu, M. M., and Gatford, M. 1995. Okapi at TREC-3. In *Proceedings of the 3rd Text REtrieval Conference*. 200–225.

Rocchio, J. J. 1971. Relevance feedback information retrieval. In *The Smart Retrieval System—Experiments in Automatic Document Processing*, G. Salton, Ed. Prentice-Hall, Englewood Cliffs, N.J., 313–323.

Salton, G. and Buckley, C. 1988. Term-weighting approaches in automatic text retrieval. *Inform. Proc. Manag. 24*, 5, 513–523.

Schauble, P. and Sheridan, P. 2000. Cross-language information retrieval (CLIR) track overview. In *Proceedings of the 6th Text REtrieval Conference*. 31–44.

Smola, A. J. and Scholkopf, B. A. 2004. Tutorial on support vector regression. *Statis. Comput. 14*, 3, 199–222.

Wen, J. R., Nie, J.-Y., and Zhang, H. J. 2002. Query clustering using user logs. *ACM Trans. Inform. Syst. 20*, 1, 59–81.

White, R. W., Clarke, C. L. A., and Cucerzan, S. 2007. Comparing query logs and pseudo-relevance feedback for Web-search query refinement. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 831–832.

Zhai, C. X. and Lafferty, J. 2001a. Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of the 10th International Conference on Information and Knowledge Management*.

Zhai, C. X. and Lafferty, J. 2001b. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Zhang, D., Li, M., Duan, N., Li, C.-H., and Zhou, M. 2008. Measure word generation for English-Chinese SMT systems. In *Proceedings of the 46th Annual Meeting of the Association of Computational Linguistics*. 89–96.

Zhang, Y. and Vines, P. 2004. Using the Web for automated translation extraction in cross-language information retrieval. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 162–169.

Received September 2008; revised March 2009; accepted May 2009