

Translation Enhancement: A New Relevance Feedback Method for Cross-Language Information Retrieval

Daqing He

School of Information Sciences
University of Pittsburgh, Pittsburgh PA 15260, USA
dah44@pitt.edu

Dan Wu

School of Information Management
Wuhan University, Wuhan 430072, China
woodan@whu.edu.cn

ABSTRACT

As an effective technique for improving retrieval effectiveness, relevance feedback (RF) has been widely studied in both monolingual and cross-language information retrieval (CLIR) settings. The studies of RF in CLIR have been focused on query expansion (QE), in which queries are reformulated before and/or after they are translated. However, RF in CLIR actually not only can help select better query terms, but also can enhance query translation by adjusting translation probabilities and even resolve some out-of-vocabulary terms. In this paper, we propose a novel RF method called translation enhancement (TE), which uses the extracted translation relationships from relevant documents to revise the translation probabilities of query terms and to identify extra translation alternatives if available so that the translated queries are more tuned to the current search. We studied TE using pseudo relevance feedback (PRF) and interactive relevance feedback (IRF). Our results show that TE can significantly improve CLIR with both types of RF methods, and that the improvement is comparable to that of QE. More importantly, the effects of TE and QE are complementary. Their integration can produce further improvement, and makes CLIR more robust for a variety of queries.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Relevance feedback

General Terms

Algorithms, Languages, Experimentation, Performance

Keywords

Cross-Language Information Retrieval, Relevance Feedback, Translation Enhancement, Query Expansion

1. INTRODUCTION

One of the most commonly used approaches for Cross-Language Information Retrieval (CLIR) is to utilize translation to cross the language barriers between a query and the documents [18]. Because of its flexibility and effectiveness, query

translation-based approach has been the dominant method for CLIR [18]. In this approach, it is the query -- not the documents -- that is translated so that the retrieval can be performed. However, the problems of not knowing which translation alternatives are appropriate for the current query (called translation ambiguities) are much more pervasive in query translation. Researchers have developed various methods to handle translation ambiguities. These include methods for reducing the impact of translation ambiguities such as the structured query method [21] and the probabilistic structured query method [5], as well as methods for disambiguating translations by using phrase translation [3], or resources such as parallel corpus [30].

Researchers also borrow tested techniques from monolingual IR to improve CLIR effectiveness. One example is query expansion (QE) based on relevance feedback (RF). In CLIR, there are pre-translation QE that performed before translating the query with the help of an extra document collection at the query language side, post-translation QE that performed after the query is translated, and the combination of pre and post-translation QE.

As discussed in iCLEF experiments (such as in [19]), and as demonstrated by Google's recently launched CL search engine inside Google Translate¹, it is too-narrow a view to think that CLIR ends with a ranked list of documents in a language that the user cannot understand. Such view is reasonable for testing retrieval algorithms in TREC-like testing frameworks. However, a CLIR system that is truly useful to the user has to consider how the user can select, examine or even use the returned documents. To help the user, some form of the translations of returned documents has to be provided. The translations can be the outputs from machine translations (MT) [10], from noun phrase based translations [16], or from word-by-word gloss translations [27]. Consequently, there are two translation processes in a truly useful query translation-based CLIR -- one is for translating queries, and the other is for translating returned documents. It is possible for CLIR systems to use the same translation resources for both translation processes, but most CLIR systems rely on an existing MT system to provide the translation of returned documents, so the translation resources used in the two translation processes are often different. Therefore, there is an advantage for certain translation information to be exchanged between the two translation processes.

Although QE has shown to be an effective method for improving CLIR performance [17], it is not the only possible method for using RF information in CLIR. We believe that users'

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CIKM'08, October 26–30, 2008, Napa Valley California, USA.
Copyright 2008 ACM 978-1-59593-991-3/08/10...\$5.00.

¹ http://translate.google.com/translate_s?hl=en

feedback on relevant documents can be viewed as the confirmation of the translation relationships expressed by the terms in the original returned relevant documents and their translations in the translated documents. Therefore, naturally, these translation relationships can be applied back to the query translation process to remove translation ambiguities, just like using translation relationships obtained from parallel corpus for the same purpose. The difference here is that the translation relationships obtained through RF have been extracted from relevant document set. In addition, the extracted translation relationships can sometimes produce better quality or even new translations that do not exist in the original dictionary. We call this new usage of RF information “translation enhancement” (TE).

Our goal in this paper is to examine the usage of TE as a RF technique in CLIR. Our research here has three objectives:

- First, through studying several methods of extracting translation relationships based on RF information, we want to examine whether TE is an effective method for using pseudo RF (PRF) information in CLIR.
- Second, because TE and QE are actually working on different parts of the CLIR process, it is important to know whether TE and QE can be integrated for further improvement.
- Third, we will study whether TE is still an effective technique for dealing with interactive RF (IRF) information.

In the remainder of this paper, we will first review the related work on RF and QE in monolingual and CLIR in section 2; then in section 3, we talk about TE in detail with the focus on our current methods of implementing TE in CLIR process. Then, we will talk about a group of experiments conducted in order to obtain answers to the research objectives. Finally, we will conclude with discussions about TE, the relationship between TE and QE, and future works.

2. RELATED WORK

Using RF to improve retrieval effectiveness has a long history. It was initially introduced in the mid 1960s, and some of the early important work was achieved in vector space modeling using the SMART system [24, 25]. Since then, RF has also been explored in probabilistic modeling [9] and statistical language modeling [22] too. Relevance-based language model [15] represents a recent attempt to directly model relevance based on RF information. There are two ways to obtain RF information. Interactive RF relies on users’ relevance judgments on the retrieved results [25], whereas pseudo RF assumes that the top N returned documents are all relevant so that users do not have to make judgments [4].

In monolingual IR, it has been well studied about the issues of how exactly RF information can be extracted, and how such information should be applied in query reformulation. There are basically two methods of applying RF information in query reformulation. Term reweighting adjusts the relative importance of the query terms (e.g., their associated weights) based on RF information (including using Rocchio’s methods [24] and Ide’s improvements [13]). Query Expansion (QE), on the other hand, adds more terms into the query with the hope that the users’ information needs are more closely represented [8]. Between the two, QE is more effectiveness and received well deserved

attentions. For example, Harman examined different algorithms for QE [8], Voorhees used lexical semantic information for QE [26], and Xu studied local and global context information for QE [29].

Term mismatch is a difficult problem in monolingual IR, but it is a significantly more serious problem in CLIR. This is because CLIR users have to identify the right query terms whose translations rather than the terms themselves would match to the critical words used by the authors of the relevant documents. Naturally, QE plays an important role in using RF in CLIR. Depending upon whether it has happened before or after the query translation, QE in CLIR is classified into pre-, post-translation QE, or the combination of the two. Ballesteros and Croft demonstrated that post-translation QE is more effective in improving CLIR performance [3], whereas McNamee and Mayfield found that the combination of the two actually performs better particularly when the translation resources are not of a high quality [17]. Orenge et al. [20] confirmed that CLIR users can provide reliable RF information based on translations of original documents. Hiemstra et al. [12] provided a formula for applying RF back to a unigram statistical language model for CLIR. The cross-language relevance models proposed by Lavrenko et al. [14] use either parallel corpora or a bilingual lexicon to estimate a relevance model between two languages, and the model is used for QE and disambiguation.

Except for when Hiemstra et al. [12] touched in the statistical language modeling setting, no study in the CLIR literature has focused on examining whether and how RF can affect the quality of query translation, and whether and how translation enhancement can be integrated with query expansion.

3. TRANSLATION ENHANCEMENT: A NEW RF APPROACH IN CLIR

As stated in Section 1, in addition to translating its queries to the document language side, a query translation-based CLIR system also needs to translate the returned documents back to the query language side in order to present them to the users. This allows the users to seamlessly search and utilize documents in foreign languages. But more importantly, the relevance feedback on these returned documents should be viewed as performed on the translated documents (i.e., at the query language side) rather than on the original documents (i.e., at the document language side). Therefore, our TE modeling makes the following two assumptions:

1. RF information obtained in these situations not only tells us which query terms can be useful for QE, but also informs us about the users’ intended translation relationship between a word in the translated relevant documents and the corresponding term in the original relevant documents.
2. Regardless of the translation resources used in the translation of both the queries and the returned documents, it still makes sense to apply the translation relationships obtained through RF to the query translation in order to improve retrieval effectiveness.

Therefore, the core issues related to TE include: 1) how to extract the intended translation relationships from a set of relevant

document pairs; and 2) how to apply such relationships to query translation.

3.1 Extracting Intended Translation Relationships from Relevant Documents Pairs

When obtaining relevance feedback from users, the relevance judgments can be performed on various granularities of documents. Researchers in the HARD track of TREC have explored RF on terms, named entities, passages, and documents, and find that they each have their own advantages and limitations [1]. Because feedback on documents involves the least effort from users, it is the most commonly used RF approach. In this study, all our RF was performed at the document level. However, the developed TE techniques can be easily applied to other RF granularities.

The purpose of extracting translation relationships is to improve the translation quality of the query; therefore, in our current design of TE, we concentrate on identifying the intended translation relationships to the query terms. Future studies can aim for extracting various other types of translation relationships, such as important named entities, potential QE terms, and out-of-vocabulary terms, etc.

Since the set of relevant documents and their translations can be viewed as a small scale parallel corpus, it is natural to borrow ideas from parallel corpus based text processing. The first step in our extraction of indented translation relationships is to establish sentence alignment within the document pairs. This is a relatively easy and accurate step to perform, but it gives us a much narrower searching space to look for intended translation relationships, which helps to achieve better extraction accuracy.

Based on the sentence alignment of document pairs, we designed four methods to extract intended translation relationships (see below). In all four methods, we extended the back-off translation approach [23] to maximize the search for instances of query terms and their translations in the relevant document set. The first three methods assume that we only have the sentence aligned document pairs. The fourth one (TWA) assumes that word alignment information is either available or can be easily obtained.

Keep All Translations (KAT): This approach only requires a set of document pairs aligned at sentence level, and a dictionary that is used for translating the query. It firstly searches for all the instances of the query terms inside the documents at the query language side (i.e., those MT documents). Then, it utilizes the sentence alignment information to search inside the documents at the document language side (i.e., those original documents) for all the instances of the translation alternatives of the query term. All the found instances of the query terms and their translation alternatives in the same sentence pairs are treated as the intended translation relationships.

Keep One Best Translation (KIT): The fact that the KAT approach views all found translation alternatives in a sentence pair as the intended translation relationships may introduce lots of noise, especially when the dictionary is automatically generated, in which many translations are noise themselves. Following one of the heuristics often used in word sense disambiguation community, the KIT approach assumes that the translation that has the highest translation probability in the dictionary is probably the most plausible meaning in the relevant feedback. Therefore, it favors the one translation alternative that has the highest

translation probability in the dictionary among all of the translation alternatives returned by the KAT approach.

Keep Most Frequent Translations (KFT): Like KIT, this approach is again based on KAT, but with a different assumption. Extending the heuristics used in word sense disambiguation [31], KFT assumes that the correct translation of a query term is consistent in a discourse (say in a relevant document) so that that translation would appear in the highest frequency in the discourse among all the translations. Therefore, KFT only keeps the translation that has the highest frequency in a relevant document.

Translations based on Word Alignment (TWA): Regardless of using KAT, KIT or KFT, the search space is the whole sentence, where lots of noise may be introduced. To obtain more accurate identification of translation relationships, we used word alignment tools that have been widely used in a parallel corpus setting for extracting translation relationships. The tool we used was GIZA++². After obtaining word level alignments for each sentence pair in the relevant document set, it is straightforward to extract the translation relationships. All the found instances of the query terms and their translation alternatives identified by word alignment in the relevant document set are treated as the intended translation relationships for translation enhancement later.

3.2 Enhancing Query Translation with Extracted Translation Relationships

Translation probabilities have been demonstrated to be very important in CLIR [7]. This is why we think that, by enhancing the translation probabilities used in query translation, the CLIR effectiveness can be improved. However, we do not think that the extracted translation relationships can replace the translation probabilities in the dictionary because the feedback from the users could sometimes be incomplete or even inaccurate. Therefore, our approach is to firstly convert the extracted translation relationships into translation probabilities, then to integrate these probabilities with the corresponding translation probabilities in the dictionary to obtain the final enhanced translation probabilities.

To estimate the translation probability of a query term based on the extracted translation relationships, we use the frequency of a particular translation relationship over the total count of the extracted translation relationships for a given query term. Therefore, the extracted translation probability for a translation alternative j as the translation of a term i in a relevant document set is calculated as in formula (1):

$$P_{i,j}(Rel) = \frac{\sum_{k \in N} w_k * tf_{j,k}}{\sum_{a \in M_i} \sum_{b \in N} w_b * tf_{a,b}} \quad (1)$$

where $P_{i,j}(Rel)$ is the probability of translation alternative j being the translation of term i based on a relevant document set; $tf_{j,k}$ and $tf_{a,b}$ are the frequency of translation alternative j or a in document k or b ; N is the relevant document set; M_i is the set of unique translation alternatives identified in the document set N for term i ; w_k is the relevance weight of document k , which takes 1 (relevant)

²<http://www-i6.informatik.rwth-aachen.de/web/Software/>

or 0 (irrelevant) for binary relevance judgments, but w_k can take different values for graded relevance assessments.

We then use weighted linear interpolation to integrate the extracted translation probabilities with their corresponding original translation probabilities in the dictionary (see formula (2)):

$$P_{i,j} = \lambda * P_{i,j}(Rel) + (1 - \lambda) * P_{i,j}(Dic) \quad (1)$$

where $P_{i,j}$ is the final probability of translation alternative j being the translation of term i after translation enhancement; $P_{i,j}(Dic)$ is the original translation probability of j being the translation of term i in the dictionary; λ is the parameter to adjust different weights between the extracted probabilities and the original ones.

Finally, suppose M_i is still the set of unique translation alternatives identified from the relevant document set for term i , L_i is the set of unique translation alternatives in the dictionary for term i , so $M_i \cup L_i$ is the union of the two sets. We then normalize the enhanced probabilities of each translation alternative j of term i so that the sum of the normalized probabilities $P'_{i,j}$ is 1 (see formula (3)):

$$\sum_{\forall j \in M_i \cup L_i} P'_{i,j} = 1 \quad (2)$$

3.3 Translation Enhancement and Query Expansion

Both translation enhancement and query expansion are relevance feedback techniques. QE improves retrieval effectiveness by introducing new terms into the query, with the hope that the terms would provide highly content related information. Pre-translation QE expands the original query at the query language side, whereas post-translation QE adds terms to the translated query at the document language side. TE, on the other hand, mainly concentrates on improving the translation resources used for query translation without necessarily introducing new query terms at both language sides. The improvement of TE often come from better or tailored translation probabilities, or newly identified translation alternatives that did not exist in the translation resources before the feedback. Therefore, as illustrated in Figure 1, TE and QE occur at different stages of the CLIR process. An interesting research question is whether or not it is beneficial to combine TE and QE in CLIR.

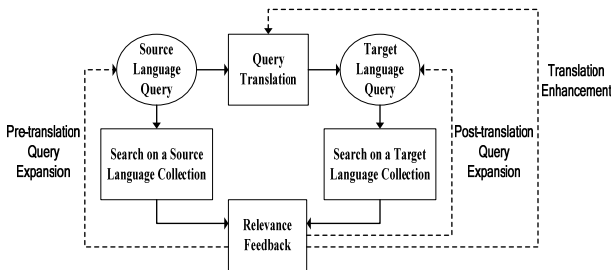


Figure 1: Translation Enhancement and Query Expansion (pre and post-translation) in CLIR

Because QE alone is not the focus of this paper, we directly adopted a state of the art QE implementation that is the adaptation of Lavrenko's relevance model [15] and is implemented inside the Indri search engine. Formula (4) summarizes this implementation, where I is the original query, r is a term to be considered for query expansion, and D is a document. The details of this implementation can be found online at <http://ciir.cs.umass.edu/~metzler/indriretmodel.html>.

$$P(r|I) = \frac{\sum P(r|D)P(I|D)P(D)}{P(I)} \quad (3)$$

4. EXPERIMENTS AND RESULT ANALYSIS

The goal of our experiments is to examine the effectiveness of the proposed approaches for translation enhancement. In particular, we are interested in the following three research questions:

1. Is TE an effective Pseudo RF technique in CLIR? Here the effectiveness is measured by comparing to the non-RF CLIR and monolingual baselines. We are also interested in the effectiveness difference among the four translation relationship extraction methods: KAT, K1T, KFT, and TWA.
2. Is it beneficial to combine TE with QE in Pseudo RF? Here, we want to compare TE and QE, and examine how TE can be integrated with QE.
3. Can TE still be effective in interactive RF? Here, we want to examine the extent to which TE can perform when RF is gathered from users.

4.1 Experiment Setup

All our experiments were performed on English-Chinese CLIR, where English queries are used to retrieve Chinese documents. The document collections used in the experiments were TDT4 and TDT5 Multilingual News Text corpora issued by Linguistic Data Consortium (LDC)³. The English and Chinese news articles were collected daily from 20 news sources in two time periods: Oct. 2000-Jan. 2001 and Apr.-Sept. 2003. In total there are 83,627 Chinese documents and 306,498 English documents. For each original Chinese document, there is an output from the MT system developed at the Information Sciences Institute (ISI) at the University of Southern California.

The reasons that we use TDT4&5 collections rather than the collections from CLEF or NTCIR are: 1) TDT4&5 collections are standard CLIR collections too. They have been used in multilingual Topic Detection and Tracking (TDT) experiments from 2002 to 2005 [2]; 2) for our TE experiments, we need English translations for Chinese documents. Then for our study of pre-translation QE, it is better that we have a comparable English collection besides the Chinese target collection. As stated above, the TDT collections not only have English translations for the Chinese documents, but also have the English comparable documents from the same time period. Therefore, the TDT collections are perfect for our current experiments. Of course, we acknowledge that it is possible to translate a CLEF or NTCIR

³ <http://www ldc.upenn.edu/>

collection using a machine translation system. However, this is much more expensive and subject to the availability of a reliable machine translation system/service for us to use.

```
<num> Number: 41012
<E-title> Trouble in the Ivory Coast

<E-desc> Description:
Presidential election; Laurent Gbagbo, Alassane Ouattara, Ivory Coast voters; Ivory Coast; October 25, 2000

<E-narr> Narrative:
On October 25, Laurent Gbagbo, head of the Ivorian Popular Front, declared himself president, as early polls showed him in the lead. Alassane Ouattara called the election unfair, but then conceded, though tens of thousands of his supporters took to the streets. A recent history of power struggle that led to the current election. Disputes concerning the election including violence by the opposition groups.
```

Figure 2: An Example of the modified TDT Topic

We selected 44 English topics which have more than 20 relevant documents each from TDT4&5 corpora, and manually translated them into Chinese to obtain topic statements at Chinese side. We reorganized TDT topics into the TREC topic like style with a title, a description and a narrative field (see Figure 2). This helped us to generate three types of queries based on these topics: short length queries that contain titles only (T query), medium length queries with title and description (TD query), and long length queries with all the three fields (TDN query). The average lengths of queries were: T queries (4 terms), TD queries (27 terms), and TDN queries (127 terms).

The bilingual dictionary we used was an English-Chinese lexicon compiled by training GIZA++ on multiple sources including the Foreign Broadcast Information Service (FBIS) corpus, HK News and HK Law, UN corpus, and Sinorama, et al [28]. The dictionary contains 126,320 English entries with translation probabilities for each Chinese translation alternative. The translation probabilities are obtained based on the normalized frequency of an English word and a Chinese word being linked together by word alignment. All Chinese texts and queries were segmented using the Stanford Chinese word segmenter⁴. The Porter stemmer⁵ was used to stem English texts, queries and the dictionary when necessary. Stop words were removed using a Chinese stopword list⁶ and an English stopword list⁷. We used our own sentence breaker to identify sentence boundaries in order to achieve sentence alignment. The breaker uses punctuation, such as periods, question marks, exclamation points, and ellipses. It also compares the sentence length line by line between the sentence pairs based on the assumption that the length of a translation sentence (English in our study) should be roughly the same as that of the original sentence (Chinese). If the difference exceeds a

certain amount (6 terms⁸ in our study), we combined the shorter sentence with its next one.

For query translation in the baseline, to remove low probability translations which often are noise, we took a fixed threshold called Cumulative Probability Threshold (CPT) to select translations from the dictionary. This is done by ranking the translations in decreasing order of their normalized probabilities, then iteratively selecting the translations top-down until the cumulative probability of the selected translations is firstly reached or exceeds the threshold. A threshold of 0 thus corresponds to using the single most probable translation (a well-studied baseline) and a threshold of 1 corresponds to the use of all translation alternatives in the dictionary. Once the queries were translated, we used Indri v.2.4⁹ to perform the document retrieval.

The main evaluation metric is mean average precision (MAP). Statistical significance tests were two tailed paired samples t-test.

4.2 TE Experiment with Pseudo RF

In total, we performed seven runs in this experiment. Pseudo RF was performed on the top 20 documents from the original ranked lists.

Monolingual Baseline (MONO-BASE): a run of retrieving Chinese documents using the manually translated Chinese queries.

CLIR Baseline (CLIR-BASE): a run using English queries to retrieve Chinese documents without using any RF techniques. In order to fully cover the dictionary, we adopted a four-stage back-off translation strategy [23]. First, the surface form of an input English term is matched to the surface forms of the English terms in the dictionary. If it fails, the input English term is stemmed, and then is matched to the surface forms of the English terms in the dictionary again. If this still fails, the dictionary is stemmed, so that the surface form of the input term is matched to the stems of the terms in the dictionary. If all these fail, match the stem of the input term to the stems of the terms in the dictionary.

Higher CLIR Baseline (CLIR-QE): We performed pre-translation QE, post-translation QE, and the combination of the two over CLIR-BASE. Top 20 expanded terms were selected from the top 20 returned documents. We selected the best performance of the three QE methods as the higher CLIR baseline CLIR-QE.

CLIR Translation Enhancement using all translations (CLIR-KAT): a TE run using KAT approach as stated in Section 3.1.

CLIR Translation Enhancement using one-best translation (CLIR-K1T): a TE run using K1T approach as stated in Section 3.1.

CLIR Translation Enhancement using most-frequent translation (CLIR-KFT): a TE run using KFT approach as stated in Section 3.1.

CLIR Translation Enhancement using word alignment (CLIR-TWA): a TE run using TWA extraction approach as stated in Section 3.1.

⁴ <http://nlp.stanford.edu/software/segmenter.shtml>

⁵ <http://tartarus.org/~martin/PorterStemmer/>

⁶ <http://www.unine.ch/info/clef/englishST.txt>

⁷ <http://bbs.ir-lab.org/cgi-bin/topic.cgi?forum=3&topic=127>

⁸ The parameter is within 5-7 terms based on prior experience. Here we fix it at 6 in our study after testing.

⁹ <http://sourceforge.net/projects/lemur/>

To get a better idea of the effectiveness of translation enhancement, we compared the results from all seven runs at different CPT from 0.0 to 1.0 with an increment of 0.1 at each time. Table 1 shows the best CPT results of these seven runs.

As shown in Table 1, all four TE approaches performed better than the lower CLIR baseline CLIR-BASE. The smallest improvement came from CLIR-KAT, and CLIR-TWA performed the best among the four. However, only CLIR-TWA significantly outperformed CLIR-BASE with all three types of queries. Only at the short (T) and medium query (TD) conditions, did CLIR-KFT perform significantly better than CLIR-BASE. Only at the long query (TDN) condition, did CLIR-KIT perform significantly better than CLIR-BASE. All these demonstrate that TE is a valid and effective RF technique for improving CLIR performance.

Table 1: Comparison of four TE methods with monolingual and CLIR baselines (* indicates that the improvement is statistically significant)

		MAP(Perc. of MONO-BASE)	Impr. over CLIR-BASE
T	MONO-BASE	0.4739(100%)	+42.06%*
	CLIR-BASE	0.3336(70.39%)	-
	CLIR-QE	0.4415(93.16%)	+32.34%*
	Pre-Trans QE	0.3714	
	Post-Trans QE	0.4118	
	Combined QE	0.4415	
	CLIR-KAT	0.3510(74.07%)	+5.22%
	CLIR-KIT	0.3599(75.94%)	+7.88%
	CLIR-KFT	0.3552(74.95%)	+6.48%*
CLIR-TWA	0.3992(84.24%)	+19.66%*	
TD	MONO-BASE	0.5817(100%)	+36.84%*
	CLIR-BASE	0.4251(73.08%)	-
	CLIR-QE	0.5080(87.33%)	+19.50%*
	Pre-Trans QE	0.4377	
	Post-Trans QE	0.5080	
	Combined QE	0.5007	
	CLIR-KAT	0.4410(75.81%)	+3.74%
	CLIR-KIT	0.4602(79.11%)	+8.26%
	CLIR-KFT	0.4584(78.80%)	+7.83%*
CLIR-TWA	0.5340(91.80%)	+25.62%*	
TDN	MONO-BASE	0.6215(100%)	+32.21%*
	CLIR-BASE	0.4701(75.64%)	-
	CLIR-QE	0.5182(83.38%)	+10.23%*
	Pre-Trans QE	0.4477	
	Post-Trans QE	0.5182	
	Combined QE	0.5170	
	CLIR-KAT	0.4885(78.60%)	+3.91%
	CLIR-KIT	0.5016(80.71%)	+6.70%*
	CLIR-KFT	0.4936(79.42%)	+5.00%
CLIR-TWA	0.5818(93.61%)	+23.76%*	

Compared to the higher CLIR baseline CLIR-QE, the TE approach CLIR-TWA outperformed CLIR-QE when the queries were TD and TDN. However, only the improvement obtained with TDN queries was statistically significant. When we further examine the specific QE methods (i.e., pre-translation, post-translation and combined, see Table 1), all four TE methods obtained better results than the pre-translation QE approach with the TD and TDN queries. It should be noted that the TE methods are mostly comparable to post-translation QE and the combined

QE methods except that CLIR-TWA with TD and TDN queries are better than the QE methods. Besides the difference between QE and TE, our results also confirm the finding in the literature that post-translation QE and combined QE are generally better than the pre-translation QE method [17].

With TE alone, the CLIR performance cannot outperform the corresponding monolingual baselines. However, all four TE methods met at least 75% of the monolingual effectiveness, with CLIR-TWA matching 93.61% of the effectiveness of MONO-BASE with the TDN queries.

The quality of the extracted translation relationship can affect the TE performance significantly. With extra resources such as word alignment, TWA method can extract much better translation relationships, and better extracted translation relationships helped CLIR-TWA to outperform the other three runs consistently. The differences between CLIR-TWA and CLIR-KAT were statistically significant for all three query types. The difference between CLIR-TWA and the remaining two runs (CLIR-KIT and CLIR-KFT) were significant for TD and TDN queries.

Besides extracting much better translation relationships than other TE methods, TWA can sometimes identify the translations for some out-of-vocabulary (OOV) terms with the help of word alignment information. Table 2 shows 11 OOV terms and their translations found by TWA method. Almost all these terms are named entities (NEs) of people names, locations, etc. This is consistent with previous findings that many OOV terms are NEs because the translations of NEs often do not exist in the dictionary [6]. Therefore, it becomes an important advantage for TWA (thus for TE) to be able to identify translations for some OOV NEs. Of course, as shown in Table 2, some of the found translations are wrong (such as the translations for No 3 and 5) which is the result of word alignment errors. However, the fact that majority found translations are correct indicates that it is reasonable reliable to use TWA method for resolving some OOV terms.

Table 2: OOV terms and their translations found by TWA. # after the No means that the translation is wrong

No	Topic ID	OOV Term	Translations found by TWA
1	55087	Bingol	宾格省
2	55087	diyarbakir	迪亚巴克尔
3#	40007	Garner	还
4	55087	kandilli	坎迪利
5#	55029	karolinska	推动/科技
6	55179/55127	Kumba	昆巴
7	41025	montesinos	蒙特西诺斯
8	40037	morariu	莫拉留
9	41012	ouattara	瓦塔拉
10	55181	Qurei	库赖
11	41025	vladimiro	弗拉迪米罗

There could be some concern about using word alignment for TE. It might be infeasible to conduct word alignment on the fly when users are waiting for their search results, or it may be difficult to obtain a word alignment tool at all. However, there are solutions to the concerns. For example, word alignment can be

done offline on the whole collection as soon as the translations of the documents are ready. Or, we can rely on modern statistical machine translation systems to generate hypothesized translation relationships at the word or phrase level. These translation relationships can then be transformed into the word alignment information needed in our TWA based TE approach. Therefore, it is reasonable to rely on word alignment information for TE.

The three less effective TE methods all have their weakness. KAT is unstable particularly when the translations in the dictionary are generated automatically, which contains a great deal of noise. KIT pays much attention to the information in the dictionary, and could be too restricted in selecting translations. If such selection fails, KIT has no way to recover. Although KFT selects translations in the scope of the individual relevant document, which has a reasonable large context in which to make selections, it still selects just one translation. Like KIT, this could be too restricted.

In both monolingual and cross-language retrieval, query length is a factor that affects the performance of RF techniques. For example, as shown in Table 1, QE worked the best with short T queries, and its effectiveness decreased with TD and TDN queries. When the query is a long query, pre-translation QE even hurts the results. TE is affected by query length as well. For example, CLIR-TWA performed the best with long TDN queries, and was significantly better than with the short T queries. However, whatever the query length is, all the four TE methods get positive results. Therefore, it seems that TE and QE are most effective with different query length, which is another motivation to combine these two in one CLIR.

Table 3: MAP of the two new runs with short, medium, and long queries. MAP of MONO-BASE, CLIR-QE and CLIR-TWA are copied from Table 1 for comparison purpose

Run ID	MAP		
	T	TD	TDN
MONO-BASE	0.4739	0.5817	0.6215
CLIR-QE	0.4415	0.5080	0.5182
CLIR-TWA	0.3992	0.5340	0.5818
CLIR-TEQE	0.4748	0.5905	0.5972
MONO-QE	0.5575	0.6255	0.6364

4.3 Experiments on Integrating TE and QE with Pseudo RF

In this section our goal is to study whether it is beneficial to integrate TE and QE in CLIR. We will examine this question with pseudo RF. For this part of the experiments, we conducted another two runs with the same resources stated in Section 4.1:

Integrated TE and QE run (CLIR-TEQE): a run combined the best method of TE, which is the word alignment approach, with the best QE method, which is post-translation QE.

Higher Monolingual Baseline (MONO-QE): a monolingual baseline run with query expansion.

As shown in Table 2, and then further elaborated in Table 3, it makes sense to combine TE and QE in CLIR. First, the combined run CLIR-TEQE outperformed CLIR-BASE with significant improvement over all three types of queries. This is the evidence that the combination works. More importantly, CLIR-TEQE also

outperformed runs with TE or QE individually (i.e., CLIR-TWA and CLIR-QE). Again, some of these improvements were statistically significant (with T and TD over TE and with TD and TDN over QE).

The combination of TE and QE (CLIR-TEQE) achieved comparable results to the monolingual baseline MONO-BASE for all three types of queries. In the case of TD queries, CLIR-TEQE even exceeded the monolingual run. Of course, these runs still cannot outperform the higher monolingual baseline MONO-QE.

Another interesting point is that CLIR-TEQE showed relatively stable performance with all three types of queries. Different to TE that works better with long queries than short queries, and to QE that works well with short queries but is losing performance with long queries, the combined run performed consistently comparable to the monolingual run with all three types of queries. It seems that the combination helped to use one’s advantages to overcome the limitations of the other. Therefore, we can conclude that it is beneficial and effective to combine TE with QE.

There could be questions about the improvement obtained by TE over the simple CLIR baseline. For example, one criticism could be that the reason that TE obtained significant improvement is because the CLIR baseline CLIR-BASE is rather low. We acknowledge that CLIR-BASE is a simple baseline that adopted CPT as a threshold to control the translation noises and used probabilistic structured query method. Without any other performance enhancement method, its performance is lower than the state of the art of CLIR performance. However, we have also established several higher baselines (such as CLIR-QE and MONO-BASE). The performance of TE has been comparable to CLIR-QE, and close to MONO-BASE even it is based on the simple CLIR baseline. More importantly, our study of integrating TE and QE clearly demonstrates that the benefits of TE is not to use it in isolation, but rather to integrate it with other RF techniques such as QE. The combinations of TE and QE have achieved the state of the art CLIR performance, which is around 95% of monolingual retrieval. In addition, it is important to know that TE can help to resolve OOV terms too.

Table 4: Comparison of combining translation enhancement with query expansion (CLIR-TEQE) to monolingual baseline, CLIR baseline and CLIR runs with TE or QE alone (* indicates that the improvement is statistically significant)

CLIR-TEQE	MAP			
	Perc. Of MONO-BASE	Impr. over CLIR-BASE	Impr. over CLIR-TWA	Impr. over CLIR-QE
T	100.19%	+42.33%*	+18.94%*	+7.54%
TD	101.51%	+38.91%*	+10.58%*	+16.24%*
TDN	96.09%	+27.04%*	+2.65%	+15.25%*

4.4 TE Experiment with Interactive RF

Through Sections 4.2 and 4.3, we have established that TE is a valid and effective RF method for CLIR. However, all those experiments were conducted using pseudo-relevance feedback (PRF) information. We do not know whether it would make any difference if the relevance feedback information is from users directly, that is with the interactive RF condition. We therefore

tested the TE methods with IRF information. Because we are still interested in TE’s effectiveness, and based the evaluation ideas tried in the previous TREC HARD experiment [1], we used the evaluation measures for judging ranked retrieval results rather than the measures that are more user-oriented.

The test collections and the CLIR system used in this experiment were the same as in the previous PRF experiments. To elicit relevance judgments from users, we again followed the HARD experiment design. Instead of asking users to issues queries to obtained initial search results, we asked users to perform relevance judgments on the top 20 documents generated by our system. To more closely resemble the real search process where users’ queries are short, we only used the results from short T queries. We also only examined the TE method based on word alignment. We had to reduce the number of topics to 40 because the remaining 4 topics had zero relevant documents in the top 20 returned documents. It makes no sense to ask users to select relevant documents from those results.

Eight native English speakers from University of Pittsburgh were recruited to participate in the experiment. Six of them were master students, one was a PhD student, and one was an undergraduate. Five were female, and the average age was 24.6 years old. Five of them have taken IR course before. The average time they spend on online searching was 2 hours per day. The average degree of how confident they feel in being able to locate specific information was 4 with 5 as the highest confidence value and 1 as the lowest.

The whole experiment lasted for 120 minutes for each person. Each participant was given 10 topics to judge. Their tasks were to go through the 20 retrieved documents denoting one of the three possible judgments: “not relevant”, “somewhat relevant”, and “highly relevant”. The default value is “not judged”. They had a maximum of 10 minutes to finish a topic. Topic allocation was carefully designed so that each topic was judged by 2 participants. There was also time allocated for training, breaks, and completion of questionnaires. Our logging ability included using the screen capture tool Camtasia Studio¹⁰.

Our relevance judgment interface (see Figure 3) provided adequate cues to the participants: (1) topic descriptions were printed out for users to consult during judgments; (2) the query used to generate the results were displayed at the top of the screen; (3) English surrogate of each returned document is displayed too. The surrogate consists of three sentences that contain the most query terms extracted from MT document, and the sentences were displayed according to their original sequences in the documents. Query terms inside the surrogates were highlighted; (4) if the user requested it, the full text of the returned documents and their corresponding MT version were displayed in a pop-up window. Query terms and their translations were highlighted in the pop-up window too.

In our handling of IRF information for TE, we took advantage of the facts that we had 2 relevance judgments on each document for a given topic, and that the judgment had three levels of relevance degree. Formula (5) shows how we assigned weight w_k to a relevant document k .

$$w_k = \frac{\sum_{j \in \{high, some, none\}} N_j * S_j}{\sum_{j \in \{high, some, none\}} N_j} \quad (4)$$

where N_{high} is the number of users who judged “highly relevant” for document k ; N_{some} is the number of users who judged “somewhat relevant”; N_{none} is the number of users who judged “not relevant”; S_{high} is the score (set as 4 here) given to a highly relevant document; S_{some} is the score (set as 1) given to a somewhat relevant document; S_{none} is the score (set as 0) given to a non-relevant document.

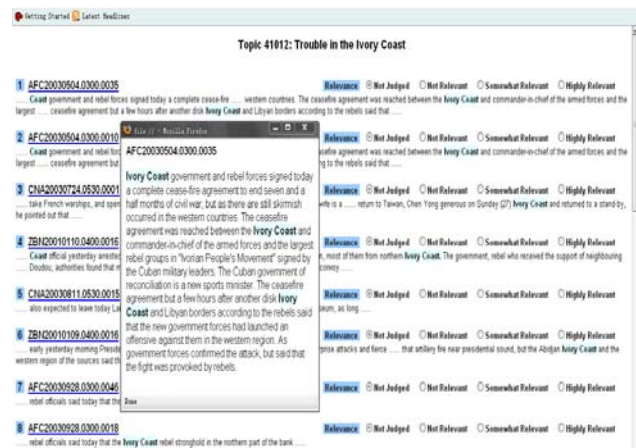


Figure 3: The interface for user making relevance judgments on search results

To examine how well the interactive RF information worked, we calculated the precision (P) and recall (R) of the judgments across all the topics and users. Here precision is the ratio of the number of correctly identified “relevant” documents over the total number of identified documents which is 20 in this experiment. Recall is the ratio of the number of correctly identified “relevant” documents over the actual number of relevant documents based on the ground truth in the top 20 documents. When calculating the numbers of “relevant” documents, we followed the “strict” and “loose” relevance judgment approach of converting three levels of relevance judgments into two levels [11]: “strict” relevance judgment treats “somewhat relevant” documents as the part of the “not relevant” documents, whereas “loose” relevance judgment treats “somewhat relevant” documents as the part of the “relevant” documents. As shown in Table 7, because the calculations of both the strict and loose relevance methods generate precision and recall values greater than 0.7, it is reasonable to state that the RF from users is reliable.

Before we performed TE on IRF data, we further removed another 3 topics because none of the participants marked any relevant document for them. Therefore, the reports presented below are from the remaining 37 topics. For this part, three runs were performed:

CLIR Baseline with IRF (USER-CLIR-BASE): an IRF run used 37 English queries to retrieve Chinese documents without using any QE or TE technique.

¹⁰ <http://www.techsmith.com/camtasia.asp>

CLIR TE with IRF (USER-CLIR-TE): an IRF run with TE using word alignment.

CLIR TE with PRF (BLIND-CLIR-TE): a PRF run that is identical to CLIR-TWA but the results reported here are based on the 37 topics.

Table 5: Performance of user’s relevance judgments

	P	R
Strict Relevance	0.8411	0.7397
Loose Relevance	0.7327	0.9274

Table 6: Comparison for the runs with IRF and PRF (* indicates that the improvement is statistically significant)

Run ID	MAP	Impr. over USER-CLIR-BASE	Impr. over BLIND-CLIR-TE
USER-CLIR-BASE	0.3945		
USER-CLIR-TE	0.4533	+14.91%*	+2.65%
BLIND-CLIR-TE	0.4416	+11.94%*	

As shown in Table 6, similar to TE based on PRF, the TE run based on IRF (USER-CLIR-TE) outperformed the corresponding CLIR baseline (USER-CLIR-BASE), and the improvement is statistically significant. Comparing to the TE run using PRF information (BLIND-CLIR-TE), USER-CLIR-TE also earned a level of 2.65% relative improvement, although the difference is not statistically significant. This demonstrates that TE is a valid and effective method for interactive RF too, and that the users’ relevance judgments can be a better RF source than PRF.

5. CONCLUSION AND FUTURE WORK

In this paper, translation enhancement is proposed as a novel RF technique for CLIR. Based on the experiment results, all four TE approaches are found to be effective. However, only the TWA approach achieved competitive performance as compared to QE, the most commonly used RF technique in CLIR. The limited capability of KAT, KIT, and KFT methods to find translation relationships accurately in relevant documents is probably the reason that these three methods showed relatively small improvements. TWA relies on word alignment to provide more accurate translation relationships, which increased its effectiveness. The effectiveness of TE was examined using both interactive RF and pseudo RF information, and the improvement of TE is actually higher with interactive RF.

Although both TE and QE are RF techniques, they actually improve CLIR performance at different retrieval stages. In addition, our studies show that TE is more suitable for long queries, whereas QE is more useful for short queries. And TE is more stable than QE when different length queries are processed. Therefore, the integration of TE and QE is beneficial not only because more RF information is used, but also because the retrieval system is more robust for a variety of queries.

Our future work on TE is in three areas. First, by knowing that the quality of the extracted translation relationships from the

relevant documents is critical, we would like to explore methods for better extracting the relationships without many extra resources. Second, we want to explore an integrated model for both QE and TE because they are complementary RF techniques. Third, we are in the process of implementing TE into a fully interactive CLIR system so that we can examine the effectiveness of TE in a more realistic setting.

6. REFERENCES

- [1] Allan, J. HARD Track Overview in TREC 2003 High Accuracy Retrieval from Documents. In *12th Text Retrieval Conference*. pages 24-37. 2003.
- [2] Allan, J. Introduction to Topic Detection and Tracking. *Topic Detection and Tracking, Event-based Information Organization*. J. Allan (Eds). Kluwer Academic Publishers. pp:1-16. 2004.
- [3] Ballesteros, L. and Croft, W. B. Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval. In *Proceedings of the 20th International ACM SIGIR Conference on Research and Development in Information Retrieval*. pages 84-91. 1997.
- [4] Buckley, C., Salton, G., Allan, J. and Singhal, A. Automatic Query Expansion Using SMART: TREC 3. In *The Third Text REtrieval Conference (TREC-3)*. 1994.
- [5] Darwish, K. and Oard, D. W. Probabilistic Structured Query Methods. In *Proceedings of the 26th International ACM SIGIR Conference on Research and Development in Information Retrieval*. pages 338-344. 2003.
- [6] Demner-Fushman, D. and Oard, D. W. The Effect of Bilingual Term List Size on Dictionary-Based Cross-Language Information Retrieval. In *36th Annual Hawaii International Conference on System Sciences (HICSS'03) - Track 4*. 2003.
- [7] Gey, F. C. and Chen, A. TREC-9 Cross-Language Information Retrieval (English - Chinese) Overview. In *TREC 2001*. pages 1-10. 2001.
- [8] Harman, D. K. Relevance feedback revisited. In *ACM-SIGIR 92*. pages 1-10. 1992.
- [9] Harper, D. J. *Relevance Feedback in document retrieval systems: an evaluation of probabilistic strategies*. Jesus College, Cambridge University, 1980.
- [10] He, D., Oard, D. W., Wang, J., Luo, J., Demner-Fushman, D., Darwish, K., Resnik, P., Khudanpur, S., Nossal, M., Subotin, M. and Leuski, A. Making MIRACLES: Interactive Translingual Search for Cebuano and Hindi. *ACM Transactions on Asian Language Information Processing*, 2(3): 219-244.2003
- [11] He, D., Wang, J., Oard, D. W. and Nossal, M. Comparing user-assisted and Automatic Query Translation. In *Proceedings of CLEF'02*. pages 400-415. 2002.
- [12] Hiemstra, D., Kraaij, W., Pohlmann, R. and Westerveld, T. Translation resources, merging strategies, and relevance feedback for cross-language information retrieval. In *Proceedings of the 1st Cross-Language Evaluation Forum (CLEF 2001)*. pages 102-115. 2001.

- [13] Ide, E. New experiments in relevance feedback. *The SMART retrieval system, Experiments in automatic document processing*. G. Salton (Eds). Prentice-Hall. pp:337-354. 1971.
- [14] Lavrenko, V., Choquette, M. and Croft, W. B. Cross-Lingual Relevance Models. *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pages 175-182. 2002
- [15] Lavrenko, V. and Croft, W. B. Relevance-based Language Models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pages 120-127. 2001.
- [16] López-Ostenero, F., Gonzalo, J., Penas, A. and Verdejo, F. Noun Phrase Translation for Cross-Language Document Selection. In *Proceedings of CLEF 2001*. pages 1639-1650. 2001.
- [17] McNamee, P. and Mayfield, J. 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*. pages 159-166. 2002.
- [18] Oard, D. W. and Diekema, A. R. Cross-Language Information Retrieval. *Annual Review of Information Science and Technology*. B. Cronin (Eds). American Society for Information Science. 33 pp:223-256. 1998.
- [19] Oard, D. W. and Gonzalo, J. The CLEF2001 Interactive Track. In *the Cross-Language Evaluation Forum (CLEF) 2001* pages 176. 2001.
- [20] Orengo, V. M. and Huyck, C. Relevance feedback and cross-language information retrieval. *Information Processing and Management*, 42(5): 1203-1217. 2006
- [21] Pirkola, A. The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pages 55-63. 1998
- [22] Ponte, J. M. Language models for relevance feedback. *Advances in Information retrieval: Recent Research from the Center for Intelligent Information Retrieval*. W. B. Croft (Eds). Kluwer Academic Publishers. pp:73-95. 2000.
- [23] Resnik, P., Oard, D. and Levow, G. Improved Cross-Language Retrieval using Backoff Translation. In *First International Conference on Human Language Technologies*. pages 1-3. 2001.
- [24] Rocchio, J. J. Relevance Feedback in Information Retrieval. *the SMART Retrieval System*. G. Salton (Eds). Prentice Hall, Inc. pp:313-323. 1971.
- [25] Salton, G. and Buckley, C. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41(4): 288-297. 1990
- [26] Voorhees, E. M. Query Expansion Using Lexical-Semantic Relations. In *Proceedings of SIGIR '94*. pages 61-69. 1994.
- [27] Wang, J. and Oard, D. W. iCLEF 2001 at Maryland: Comparing Word-for-Word Gloss and MT. In *2nd Workshop of the Cross-Language Evaluation Forum, CLEF 2001*. pages 336-354. 2001.
- [28] Wang, J. and Oard, D. W. Combining Bidirectional Translation and Synonymy for Cross-language Information Retrieval. In *Proceedings of the ACM SIGIR 2006*. pages 202-209. 2006.
- [29] Xu, J. and Croft, W. B. Improving the effectiveness of information retrieval with local context analysis. *ACM Transactions on Information Systems*, 18(1): 79-112. 2000
- [30] Xu, J., Fraser, A. and Weischedel, R. TREC 2001 Cross-lingual Retrieval at BBN. In *Proceeding of TREC01*. 2001.
- [31] Yarowsky, D. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods In *Proceeding of Meeting of the Association for Computational Linguistics (ACL-95)*. pages 189-196. 1995.