To Translate or Not to Translate?

Chia-Jung Lee, Chin-Hui Chen, Shao-Hang Kao and Pu-Jen Cheng
Department of Computer Science and Information Engineering
National Taiwan University, Taiwan
(cjlee1010, chchen.johnson, denehs)@gmail.com, pjcheng@csie.ntu.edu.tw

ABSTRACT
Query translation is an important task in cross-language information retrieval (CLIR) aiming to translate queries into languages used in documents. Previous work focused mainly on generating translation equivalences of query terms. The purpose of this paper is to investigate the necessity of translating query terms, which might differ from one term to another. Some untranslated terms cause irreparable performance drop while others do not. We propose an approach to estimate the translation probability of a query term, which helps decide if it should be translated or not. The approach learns regression and classification models based on a rich set of linguistic and statistical properties of the term. Experiments on the NTCIR-4 and NTCIR-5 English-Chinese CLIR tasks demonstrate that the proposed approach can significantly improve CLIR performance. An in-depth analysis is also provided for discussing the impact of out-of-vocabulary and wrongly-translated query terms on CLIR performance.

Categories and Subject Descriptors
H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval.

General Terms
Algorithm, Experimentation, Performance.

Keywords
Query Translation, Translation Quality, Query Term Performance, Cross-language Information Retrieval.

1. INTRODUCTION
Query translation, which aims to translate queries in one language into a different language used in documents, has been widely adopted in CLIR. Conventional approaches to query translation have focused mainly on correctly translating as many query terms as possible, including translation disambiguation [3,8,9,10,15,20], phrasal translation [1,2], and unknown words translation [5,6,22]. Such approaches pursue the reduction of erroneous or non-relevant translations in hope that the CLIR performance could approach to that of monolingual information retrieval (MIR). However, the accuracy of query translation is not always perfect. Each query term has a risk of being translated incorrectly. Some incorrect translations can be remedied in the process of MIR but others may cause irreparable retrieval performance drop. In other words, query translation may cause deterioration of CLIR performance. This phenomenon motivates us to explore whether a query term should be translated or not.

Consider the query: “Peru President, Fujimori, bribery scandal, the 2000 election, exile abroad, impeach, Congress of Peru”, which is obtained based on the description field from a NTCIR-5 English-Chinese CLIR topic (after removing stop words). Its correct Chinese translations result in mean average precision (MAP) of 0.5914 for CLIR. Figure 1 shows that if one of the query terms is not translated (x-axis), how the corresponding MAP (y-axis) changes using the correct translations of the rest of the terms as a query. It is observed that without correct translation of “Fujimori” or “bribery scandal”, we are far from satisfying retrieval performance, compared to MAP of 0.5914 (dash line). However, on the other hand, we find interestingly that if the (correct) translation of “Peru President” or “Congress of Peru” is ignored, a better MAP can be even achieved. Still the missing of the translation of “the 2000 election”, “exile abroad”, or “impeach” seems to be tolerable. This observation reveals that some not-translated terms cause irreparable performance drop while others do not. That is to say, the query terms are not equally important for translation, and it is not always the case that all translations are required.

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, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SIGIR '10, July 19-23, Geneva, Switzerland
Copyright 2004 ACM 1-58113-000-0/00/0004...$5.00.

Figure 1. MAP value for a not-translated query term.

In the above example, term “Fujimori” seems to bear more important semantics and thus should be translated. It might appear out-of-vocabulary (OOV) terms always need perfect translations. Take into account the query from another NTCIR-5 English-Chinese topic (after stop words removal): “Chinese-American, scientist, Wen-Ho Lee, suspect, steal, classified information, nuclear weapon, US's Los Alamos National Laboratory”. It could be found that the MAP decreases 45.9% when “Wen-Ho Lee” is not translated, whereas not-translated “US's Los Alamos National Laboratory” conversely helps to improve 39.6% of MAP. Although missing the translation of “US's Los Alamos National Laboratory” loses some information about the query, we notice that term “Laboratory” luckily emerges in its (pre-translation) query expansion set, which alleviates the problem. Moreover,
there are many possible transliterations of “Los Alamos” in Chinese such as “洛萨拉摩” and “洛斯阿拉莫斯”，which introduce a further mismatching problem in MIR and are harmful to the retrieval. This example illustrates that sometimes leaving an OOV term not-translated would probably be a reasonable choice.

Conventional approaches to query translation mostly put efforts in improving translation quality of queries [18] or examining how translation resources affect CLIR performance [4,16]. Generally, the overall MAP increased in the benchmarks when better translation accuracy or translation coverage was achieved. These works did not carefully analyze the effect of translation for each individual query term on CLIR. Few did pose the problem of whether to translate a query term or not. [13,14] is the one most relevant to this work. [13] presented a method to predict the performance of CLIR according to translation quality and difficulty of queries. If a query’s retrieval accuracy was expected to be low, then the query should not be translated. Yet [13] focused merely on evaluating the performance of a whole query and did not give insight into the effect of translation for each query term. Moreover, in [14] translation quality is estimated by manually-defined formulas, instead of being automatically learned in this paper.

The purpose of this paper is to investigate the necessity of translating query terms, which might differ from one term to another. We are interested in realizing (1) the possibility of predicting a query term to be translated or not; (2) whether the prediction can effectively improve CLIR performance; and (3) how not-translated OOV and wrongly-translated non-OOV terms affect CLIR performance, respectively.

We propose an approach to estimate the translation probability of a query term according to its effect on CLIR. The translation probability serves as a basis for the decision to translate the query term or not. The proposed approach learns classification and regression models, where comprehensive factors that are essential in determination of CLIR performance are considered, inclusive of linguistic and statistical features, as well as a rich set of CLIR features in source and target languages corpora. To determine the performance of the proposed translation probability when applied to CLIR, we have conducted extensive experiments on the NTCIR-4 and NTCIR-5 English-Chinese CLIR tasks. We examine various dictionary-based translation strategies and find that CLIR performance can be significantly improved compared to original queries given in the benchmarks. An in-depth analysis is also provided for discussing the impact of not-translated OOV and wrongly-translated non-OOV terms on CLIR performance.

We highlight that query terms needing no translation may result from intrinsically ineffective property or semantic recovery by their post-translation expansion sets.

In the rest of this paper, we first make a brief review on related work in Section 2. The description of our approach is elaborated in Section 3. Section 4 presents the experimental results, and Section 5 scrutinizes details for the need of translation. Finally, in Section 6, we give our discussions and conclusions.

2. RELATED WORK

Translation quality. A great number of researchers focused on improving translation quality, as it is tractable to use query translation technique. One common way to improve translation quality is to disambiguate multiple-sense terms by heuristically selecting the most frequent translation in dictionary. Some advanced works dig into parts of speech information [2,8] in seeking good translations. Still others utilize statistical properties in parallel corpus [3,17] as well as query expansion techniques [1,16] to gain a better chance in search of translation quality. Phrasal translation approach [1,2] is also inspected for enhancing CLIR performance, as “a phrase” is usually more semantically important than “a word”. Though these works have brought significant improvement in translation quality, they eventually tried to translate as many terms as possible, which is not always an effective approach.

Translation strategy. The key to success of translation-based approaches is the coverage of vocabularies. In particular, [16,19,21] showed continuous performance variation by gradually downsizing or selecting different amount of translation resources. Moreover, though machine translation techniques [11,12,23] are effective for long sentences, it however is not suitable for short, context-inadequate queries. Recently, people [5,6,22] started to translate especially OOV terms by “crowd knowledge” from WWW. Nevertheless, it comes along with unavoidable noises and intensive computational time. Again, our purpose here is not pursuing breakthrough of translation quality. Rather, given any kind of translation technique or resource which has its own pros and cons, we want to realize whether a term should be translated.

CLIR performance prediction. [13,14] developed regression models for predicting the performance of CLIR. The translation quality and ease of query were taken into account. Their concern was evaluated at the unit of a whole query, whereas we think every single term has its own impact on CLIR performance. Moreover, their regression model focused on the degree of explainable variation with few CLIR performance verifications, while we are interested in learning the need of translation for each term, and eventually bring up CLIR performance.

3. The Effect of Query-Term Translation

3.1 Estimation of Translation Probability

Given a query topic \(Q = \{s_1, s_2, \ldots, s_n\}\) in source language, conventional query translation methods endeavor to find a set of translated terms \(\mathcal{Q} = \{t_1, t_2, \ldots, t_m\}\) in target language. Particularly, they incorporate translation dictionaries, domain-specific bilingual corpora, or the Web to estimate the probability of translation from source term \(s_i \in \mathcal{Q}\) to target term \(t_j \in \mathcal{Q}\), given source topic \(Q\), \(p(t_j|s_i, Q)\), as shown in Fig. 2. \(p(t_j|s_i, Q)\) means the translation depends on not only \(s_i\) and \(t_j\) but the rest of terms in \(Q\). For simplicity, some previous work ignores \(Q\), i.e., \(p(t_j|s_i)\).

As illustrated in Fig. 1, the effect of query-term translation may differ from one to another. We introduce a binary variable \(T \in \{0,1\}\), which is used to determine the need of translation. \(T=1\) and \(T=0\) represent should-be-translated and should-not-be-translated, respectively, w.r.t. a given source term. When bringing variable \(T\) in the estimation of \(p(t_j|s_i, Q)\), we get the following:

\[
p(t_j | s_i, Q) = \begin{cases} p(T = 1 | s_i, Q) p(t_j | s_i, Q, T = 1) \\
\sum_T p(T | s_i, Q) p(t_j | s_i, Q, T) \\
= p(T = 0 | s_i, Q) p(t_j | s_i, Q, T = 0) \\
+ p(T = 1 | s_i, Q) p(t_j | s_i, Q, T = 1) \\
= p(T = 1 | s_i, Q) p(t_j | s_i, Q, T = 1)
\end{cases}
\]

where we set \(p(t_j|s_i, Q, T=0)\) to be 0 because the probability of translation to \(t_j\) is 0 given that source term \(s_i\) should not be translated. Finally, \(p(t_j|s_i, Q)\) is determined by the probability that
source term $s_i$ should be translated, i.e., $p(T=1|s_i, Q_s)$, and the probability of translation to $t_j$, given that source term $s_i$ should be translated, i.e., $p(t_j|s_i, Q_s, T=1)$.

Figure 3 shows the newly introduced variable T, where source term $s_i$ is mapped to target term $t_j$ only if it is worth being translated ($T=1$). Note that the focus of previous work [2,3,16,22] lies in generating translation equivalences based on $p(t_j|s_i, Q_s, T=1)$, or $p(s_i|t_j, Q_s)$ since every term $s_i$ is required to be translated by default, while the goal of this paper is to predict the probability $p(T=1|s_i, Q_s)$, which concerns whether to translate or not.

$$Q_t = \{s_1, s_2, \ldots, s_n\} \quad \text{Translated Query} \quad Q_t = \{t_1, t_2, \ldots, t_m\}$$

**Figure 2. Basic translation model.**

$$Q_t = \{s_1, s_2, \ldots, s_n\} \quad \text{Translated Query} \quad Q_t = \{t_1, t_2, \ldots, t_m\}$$

**Figure 3. Extended query translation model.**

Given $Q_s = \{s_1, s_2, \ldots, s_n\}$, we formulate our problem by seeking a classifier $c$: $S$→$T$, which predicts a binary class label but does not provide any estimate of the underlying probability. Hence, we resort to finding a ranking function $r$: $S$→$R$, which ranks $\{s_1, s_2, \ldots, s_n\}$ according to their necessities for translation. In such a case, we use regression techniques to rank the terms with a permutation $\pi$: $s_{\pi(1)} > s_{\pi(2)} > \cdots > s_{\pi(n)}$ such that

$$p(T=1|s_{\pi(1)}, Q_s) > p(T=1|s_{\pi(2)}, Q_s) > \cdots > p(T=1|s_{\pi(n)}, Q_s).$$

Based on classifier $c$, query terms are easily classified according to its need for translation. Similarly, based on regression $r$, top $k$ query terms $\{t_{\pi(1)}, t_{\pi(2)}, \ldots, t_{\pi(k)}\}$ will be selected to be translated. Four different translation strategies and various threshold $k$’s have been examined in Sections 4 and 5. In this paper, we apply support vector machine (SVM) and support vector regression (SVR) [7] to do classification and regression, other alternatives can also be adopted for calculation.

We develop a regression function $r$: $S$→$R$ by learning examples in the form of $<f(s_i), LR_{CLIR}(s_i)>$, where $f(s_i)$ is the set of features for $s_i$, which will be described in Section 3.2, and

$$LR_{CLIR}(s_i) = (\varphi_{CLIR}(Q_s) - \varphi_{CLIR}(Q_s - \{s_i\})) / \varphi_{CLIR}(Q_s),$$

where $\varphi_{CLIR}(q)$ is the MAP measure for query $q$ in CLIR. The larger the loss ratio value $LR_{CLIR}(s_i)$ is, the more importantly we translate $s_i$ due to its better effectiveness in CLIR. For classifier $c$: $S$→$T$, the form would be $<f(s_i), sign(LR_{CLIR}(s_i))>$, where $sign(x)$ converts $LR_{CLIR}$ into non-positive and positive classes based on $x$.

### 3.2 Feature Set

To understand the effect of query translation, we utilize linguistic, statistical, and CLIR features $f(s_i)$ of query term $s_i$ to capture its characteristics from different aspects.

**Linguistic features.** Linguistic features used in this paper include parts of speech (POS), named entities (NE), acronym, phrase, and size (i.e., the number of words in a term). More precisely, the POS features contain noun, verb, adjective, and adverb, while the NE features comprise person names, locations, organizations, event, and time. POS and NE in our experiments are labeled manually.

**Statistical features.** Statistical features are good predictors from the viewpoints of documents corpus compared to users’. In our experiment, we use both source and target language document corpora. Particularly, we consider co-occurrence, context, and TFIDF features for estimation.

Co-occurrence features reveal the degree of how often a term tends to co-exist with others, and hence the degree of semantic substitutions by them. The more a term can be replaced by others, the less likely it needs to be exactly translated. Point-wise mutual information (PMI) is adopted for measurement over a variety of settings. In the pre-translation phase, for each $s_i$, we pairwise compute PMI between $s_i$ and $s_p$ ($\forall p \neq i$ and $s_p \in Q_s$), as well as $s_i$ and $Q_s - \{s_i\}$. Similarly, in post-translation where $t_j$ stands for the translation of term $s_i$, PMI between all pairs of $t_j$ and $t_q$ ($\forall q \neq j$ and $t_q \in Q_t$), as well as $t_j$ and $Q_t - \{t_j\}$ are calculated.

Context features are helpful for low frequency query terms that yet share common contexts in search results. The context vector is composed of a list of pairs <document ID, relevance score>, which can be obtained from the search results returned by IR systems. On obtaining the context vectors, we can estimate the degree of resemblance between any two objects by directly computing cosine similarity. Similar to co-occurrence features, we extract context vectors from various search results. Specifically, in pre-translation, for each $s_i$, we pairwise compute cosine similarity between $s_i$ and $s_p$ ($\forall p \neq i$ and $s_p \in Q_s$), as well as $s_i$ and $Q_s - \{s_i\}$. In post-translation phase, cosine similarity between all pairs of $t_j$ and $t_q$ ($\forall q \neq j$ and $t_q \in Q_t$), as well as $t_j$ and $Q_t - \{t_j\}$ are calculated. For those pairwise computed sets in co-occurrence or context similarity, we extract its maximal, minimal, and average values as the features for the corresponding term.

TFIDF features show a term’s capability of distinguishing relevant documents from irrelevant ones. We adopt its conventional definitions and compute TFIDF in both source and target language corpora for each term.

**CLIR features.** CLIR features are the key to learning what characteristics make a term favorable or adverse for translation. We define translation, expansion, and replacement features.

Translation features such as the number of translations a term encompasses measure the degree of ambiguity according to dictionary knowledge. Also, we use binary feature OOV to indicate if a term exists within the coverage of dictionary.

Expansion features express whether or not the losing information from an untranslatable term can be recovered by the semantics from the rest of terms with query expansion. In particular, query expansion in source language reserves the room for untranslatable terms by including relevant terms in advance. Also, query expansion in target language recovers the semantics loss from the noisy translation channel by inspecting the rest well-translated terms. Here we denote $QE(\cdot)$ as the query expansion set. We aim to measure the following.

$$\theta(QE((Q_s - s_i)), QE(Q_s))$$
$$\theta(QE((Q_t - t_j)), QE(Q_t))$$
\( \theta \) can either be PMI or cosine similarity. These measurements estimate the similarity between terms in the expansion sets derived with or without term \( s_i \). The same calculation is repeated in target language corpus for each translated \( t_j \). It is inferred that the more the two expansion sets resemble each other, the more likely the loss information from untranslatable \( s_i \) can be made up.

Lastly, replacement features estimate whether or not the rest of terms within the same topic together with its expansion set can take the place of \( s_i \). Hence, we resort to the similarity between the following sets of terms,

\[
\theta(Q_s, QE(Q_s - s_i) \cup \{Q_s - s_i\}) \\
\theta(Q_t, QE(Q_t - t_j) \cup \{Q_t - t_j\})
\]

If the replacement intensity is strong, it implies translation of only the rest of terms is sufficient for document retrieval. To put it differently, in source language corpus, \( QE(Q_s - s_i) \) replaces the position of \( s_i \) in original query \( Q_s \), while \( QE(Q_t - t_j) \) substitutes the semantics of \( t_j \) in original query \( Q_t \) in target language corpus.

4. EXPERIMENTS

4.1 Experimental Data

The data used in the experiments includes NTCIR-4 and NTCIR-5 English-Chinese CLIR tasks, whose statistics in the title and description fields of English topics can be found in Table 1 (after data clean). The poorly-performing queries whose MAP is below 0.02 are filtered to ensure the quality of our training data for classification and regression models. Table 2 shows the numbers of OOV and non-OOV terms in detail for each task. Note “term” refers to manual segmentation on original topic words after stop words removal, which forms a set of semantic-rich building blocks. We construct the probabilistic retrieval model (Okapi) using the Lemur Toolkit \(^1\). Both queries and documents are stemmed with Porter stemmer and filtered with standard stop words lists. We use MAP as performance metric evaluating over top 1000 documents retrieved. To avoid inside test, 5-fold cross validation is used through the entire experiments.

<table>
<thead>
<tr>
<th>Setting</th>
<th># terms</th>
<th># OOV</th>
<th># non-OOV</th>
</tr>
</thead>
<tbody>
<tr>
<td>NTCIR4</td>
<td>title</td>
<td>154</td>
<td>139</td>
</tr>
<tr>
<td></td>
<td>desc</td>
<td>298</td>
<td>283</td>
</tr>
<tr>
<td>NTCIR5</td>
<td>title</td>
<td>131</td>
<td>104</td>
</tr>
<tr>
<td></td>
<td>desc</td>
<td>277</td>
<td>241</td>
</tr>
</tbody>
</table>

4.2 Regression & Classification Performance

The coefficient of determination \( R^2 \) measures how well future outcomes are likely to be predicted by the statistical models. Particularly, the \( R^2 \) statistics (\( R^2 \in [0, 1] \)) is used to evaluate the variation between prediction result \( \hat{y}_i \) and observed ground-truth \( y_i \), wherein we denote the mean of \( y_i \) of as \( \bar{y} \). Mathematically, \( R^2 \) is defined as one minus the ratio of the residual sum of squares and the total sum of squares:

\[
R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}
\]

A higher \( R^2 \) gives us more confidence in prediction.

We train and test the regression models under a variety of features and document collections, Table 3 demonstrates the results. Averagely speaking, best regression performance can be achieved when both pre- and post-translation corpora is used, as the statistical importance of query expansion properties is entirely captured. Also, as post-translation corpus can more correctly extract effective expanded terms for MIR in target document sets (note that in NTCIR-4 and NTCIR-5, English and Chinese documents are not parallel texts), it shows that a higher \( R^2 \) can be found in post-translation corpus than in pre-translation one. Moreover, within each corpus setting, we go to details for inspection of the effectiveness using different features. The statistical features consistently achieve better \( R^2 \) values than the CLIR features, which are later followed by linguistic features. It is caused by that statistical features reflect the underlying distribution of translated terms in the document collection, also that CLIR features reveal the degree of translation quality. Finally, a larger \( R^2 \) can be achieved by including more features for training.

4.3 Feature Analysis

By inspecting correlation between the features and MAP, we may have better understanding of the effectiveness of our features. Three standard measurements inclusive of Pearson's product-moment, Kendall's tau and Spearman's rho are adopted.

Figure 4 depicts a wholesome picture of all features, where the absolute value of correlation using Okapi on NTCIR-4 data is shown. Clearly, classic TFIDF features show its discriminative power in identifying terms that need translation. The context features are more effective through inspecting retrieval results, but such features meantime suffer from higher cost of computation. Another group of efficacious features are the CLIR features. As mentioned previously, the CLIR features are crucial for estimation of semantic recovery, which is captured by query expansion and replacement features. It is worth noticing that the “oov” feature is evidently correlated to retrieval performance. It again assures that efforts in translating OOV terms are significant for CLIR, as indicated by many previous works. Lastly, “trans_size”, which records the number of translations for each term, is negatively correlated to MAP (positive in Fig. 4 because of absolute value). Truly, the more senses (translations) a term contains, the more challenging correct translation can be detected.

4.4 CLIR Performance

In this section, we show the effectiveness of our approach for CLIR. We use NTCIR-4 and NTCIR-5 English-Chinese tasks for evaluation and consider both of the <title> and <desc> fields as queries. We use 5-fold cross-validation and ensure that a test instance would not appear in the training set.

---

\(^1\) Lemur Project: http://www.lemurproject.org/
Table 3. Regression performance under various feature sets and document collections,

<table>
<thead>
<tr>
<th>Model</th>
<th>Topic</th>
<th>Pre-translation Corpus</th>
<th>Post-translation Corpus</th>
<th>Pre- and Post-translation Corpora</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>lin</td>
<td>stat</td>
<td>CLIR</td>
</tr>
<tr>
<td>Indri</td>
<td>Title</td>
<td>0.0657</td>
<td>0.5215</td>
<td>0.1720</td>
</tr>
<tr>
<td></td>
<td>Desc</td>
<td>0.0472</td>
<td>0.1322</td>
<td>0.0417</td>
</tr>
<tr>
<td>TFIDF</td>
<td>Title</td>
<td>0.1780</td>
<td>0.6872</td>
<td>0.2767</td>
</tr>
<tr>
<td></td>
<td>Desc</td>
<td>0.0879</td>
<td>0.2284</td>
<td>0.0410</td>
</tr>
<tr>
<td>Okapi</td>
<td>Title</td>
<td>0.1165</td>
<td>0.6092</td>
<td>0.2154</td>
</tr>
<tr>
<td></td>
<td>Desc</td>
<td>0.0406</td>
<td>0.0423</td>
<td>0.0083</td>
</tr>
<tr>
<td>Avg. Title</td>
<td>0.1200</td>
<td>0.6060</td>
<td>0.2214</td>
<td>0.7904</td>
</tr>
<tr>
<td>Avg. Desc</td>
<td>0.0586</td>
<td>0.1343</td>
<td>0.0303</td>
<td>0.5578</td>
</tr>
</tbody>
</table>

Table 4. CLIR performance under various translation resources, document collections, query topics, and prediction methods. Test with p < 0.01 (**) and p< 0.05 (*) against baseline method

<table>
<thead>
<tr>
<th>Okapi</th>
<th>Correct Trans</th>
<th>Google Dict Top1</th>
<th>Google Dict All</th>
<th>Google Trans</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ntcir4</td>
<td>Title BL.</td>
<td>0.2366</td>
<td>0.0902</td>
<td>0.0659</td>
<td>0.1692</td>
</tr>
<tr>
<td></td>
<td>Title UB</td>
<td>0.2774</td>
<td>0.1088</td>
<td>0.0874</td>
<td>0.1966</td>
</tr>
<tr>
<td></td>
<td>Title C</td>
<td>0.2475 (+4.60%)</td>
<td>0.1019 (+13.0%)</td>
<td>0.0785 (+19.2%)</td>
<td>0.1875 (+10.8%)</td>
</tr>
<tr>
<td></td>
<td>Title R</td>
<td>0.2602** (+9.98%)</td>
<td>0.1062** (+17.8%)</td>
<td>0.0775 (+14.6%)</td>
<td>0.1884* (+11.4%)</td>
</tr>
<tr>
<td></td>
<td>Desc BL</td>
<td>0.2121</td>
<td>0.0876</td>
<td>0.0671</td>
<td>0.1601</td>
</tr>
<tr>
<td></td>
<td>Desc UB</td>
<td>0.3025</td>
<td>0.1347</td>
<td>0.1319</td>
<td>0.2168</td>
</tr>
<tr>
<td></td>
<td>Desc C</td>
<td>0.2448* (+15.4%)</td>
<td>0.1003* (+14.5%)</td>
<td>0.0998** (+48.7%)</td>
<td>0.1803** (+12.6%)</td>
</tr>
<tr>
<td></td>
<td>Desc R</td>
<td>0.2493** (+17.5%)</td>
<td>0.1073** (+22.5%)</td>
<td>0.0847** (+26.2%)</td>
<td>0.1856** (+15.9%)</td>
</tr>
<tr>
<td>Ntcir5</td>
<td>Title BL</td>
<td>0.3541</td>
<td>0.1376</td>
<td>0.1065</td>
<td>0.3089</td>
</tr>
<tr>
<td></td>
<td>Title UB</td>
<td>0.4253</td>
<td>0.1552</td>
<td>0.1252</td>
<td>0.3496</td>
</tr>
<tr>
<td></td>
<td>Title C</td>
<td>0.3945 (+11.4%)</td>
<td>0.1437 (+4.46%)</td>
<td>0.1136 (+6.68%)</td>
<td>0.3299** (+6.79%)</td>
</tr>
<tr>
<td></td>
<td>Title R</td>
<td>0.4059** (+14.6%)</td>
<td>0.1546* (+12.3%)</td>
<td>0.1235* (+16.0%)</td>
<td>0.3348* (+8.39%)</td>
</tr>
<tr>
<td></td>
<td>Desc BL</td>
<td>0.357</td>
<td>0.1841</td>
<td>0.0835</td>
<td>0.2728</td>
</tr>
<tr>
<td></td>
<td>Desc UB</td>
<td>0.4788</td>
<td>0.2464</td>
<td>0.1893</td>
<td>0.3904</td>
</tr>
<tr>
<td></td>
<td>Desc C</td>
<td>0.4349* (+21.8%)</td>
<td>0.2073* (+12.6%)</td>
<td>0.1484* (+77.7%)</td>
<td>0.3267** (+19.8%)</td>
</tr>
<tr>
<td></td>
<td>Desc R</td>
<td>0.4363** (+22.2%)</td>
<td>0.2102** (+14.2)</td>
<td>0.1348** (+61.4%)</td>
<td>0.3394** (+24.4%)</td>
</tr>
</tbody>
</table>

Figure 4. Absolute values of correlation using Okapi retrieval model on NTCIR-4 data set.
Table 4 shows the MAP results using translated queries for search. Several techniques using different translation resources have been inspected: “Correct Trans” gives the standard translation in the benchmark; “Google Dict top1” extracts the first translation from Google Dictionary; “Google Dict all” combines all possible translations from Google Dictionary for a given term; finally the “Google Trans” returns translations from Google Translation. Moreover, for each setting, we show its baseline and upper bound performance. The baseline methods (BL) simply select all the translated terms in \( Q_i \) as one query string. For each topic in <title> or <desc>, we permute all sub queries and discover the sub-query with the highest MAP value to generate the upper bounds (UB). We also run the two-sample pairwise significance test against BL.

From Table 4, our classification (C) and regression (R) models consistently outperform the baseline methods using different translation resources. The retrieval result proves our assumption that regardless of translation quality, some terms are “meant to be” translated while others are not. It is also worth noticing that the improvement rate of title queries is larger than description queries. As longer queries have more chances to encompass noisy terms, we can thereby improve retrieval performance by not translating them. Short queries such as Web queries, on the other hand, lose a great amount of information if a term cannot be well translated. Further, comparing the improvement rate between different translation resources, we find that “Google Dict all” leaves the most room for improvement. We attribute this to the ambiguity it involves in by including as many translations as possible. Fig. 5 illustrates the impact of the variable \( k \).

![Figure 5. MAP with various k values on different dataset.](image)

### 5. TRANSLATION ANALYSIS

In this section, we discuss the effect of the translation of OOV and non-OOV query terms on CLIR performance. We first explore what factors make a query term favorable for translation (\( T=1 \)) or not (\( T=0 \)). Based on [14], we assume that whether a query term should be translated or not depends on its intrinsic effectiveness in locating relevant documents or its translation quality. To focus only on the translation problem, we should filter intrinsically-ineffective query terms, which perform poorly even their correct translations are obtained. Given a query topic \( Q_s = \{ s_1, s_2, \ldots, s_6 \} \) in source language, suppose its correct translation is \( Q_t = \{ t_1, t_2, \ldots, t_n \} \), which are available in our experiments because NTCIR-4 and NTCIR-5 CLIR tasks provide both of English and Chinese topics at the same time. \( t_j \) is the correct translation of \( s_j \).

We define a MIR loss ratio as follows:

\[
L_{MIR}(s_j) = \left( \varphi_{MIR}(Q_{t_j} - t_j) - \varphi_{MIR}(Q_{t_j}) \right) / \varphi_{MIR}(Q_{t_j}),
\]

where \( \varphi_{MIR}(q) \) is the MAP measure for query \( q \) in MIR. \( L_{MIR}(s_j) \) tells the influence of translating \( s_j \) to \( t_j \) in CLIR. The larger the value is, the more we are unwilling to translate \( s_j \). In the following, we are more interested in intrinsically-effective query terms (\( L_{MIR} <= 0 \)).

#### 5.1 OOV Terms Analysis

We discuss how not-translated OOV terms affect CLIR performance, and why some OOV terms are not required to be perfectly translated. All of the OOV terms appearing in <title> and <desc> from both NTCIR-4 and NTCIR-5 are collected. Table 2 shows the numbers in detail.

Firstly, based on the term ranking lists generated by regression function \( r \), we calculate the ranking percentage for each OOV term. For example, if an OOV term is ranked at top 2 in a list of size 5, its ranking percentage equals to \( 2/5 \times 100\% \). Following this manner, we expect that effective terms (\( L_{MIR} <= 0 \)), i.e., the terms need to be translated, are usually ranked in front of ineffective ones (\( L_{MIR} > 0 \)) and thus have smaller average ranking percentage. Table 5 reveals the reliability of our ranking lists. In addition, it is worth noting that for longer queries such as <desc>, we have a better chance to determine whether to translate a term or not, as the ranking percentage in <desc> is often smaller. The result is somehow not surprising since longer queries usually contain more noises.

<table>
<thead>
<tr>
<th>Title</th>
<th>Desc</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR &lt;=0</td>
<td>0 &lt; LR</td>
</tr>
<tr>
<td>Prop.</td>
<td>83.3%</td>
</tr>
<tr>
<td>N4 Perc.</td>
<td>0.6794</td>
</tr>
<tr>
<td>N5 Perc.</td>
<td>0.6507</td>
</tr>
</tbody>
</table>

By calculating the proportion of effective terms to ineffective terms, Table 5 tells that the less number of terms a query such as <title> includes, the more effective each query term is for retrieval (83.3%:16.7% vs 76.5%:23.5%). The result is consistent with [6]. The not-translated OOV terms in short queries, especially for title queries or Web queries, crucially determines retrieval performance.

Moreover, to investigate what makes difference between effective and ineffective OOV terms, we collect all of the OOV instances from the entire dataset to train our classifier \( c \). The upper part of Table 6 shows the classification accuracy for OOV terms. When applying the greedy-hill-climbing-based method to decide the best feature set, we get top-ranked features, including replacement, query expansion and TFIDF features, which again show their capability of predicting the need of translation for OOV terms. We seek a closer examination on how the features differentiate between effective and extremely-effective for source term \( s_j \) by proposing the following measurements.

\[
r_1 = \varphi_{MIR}(Q_E(\{Q_t - t_j\}) \cup \{Q_t - t_j\}) - \varphi_{MIR}(Q_t) / \varphi_{MIR}(Q_t),
\]

\[
r_2 = \varphi_{MIR}(t_j) - \varphi_{MIR}(Q_t - t_j) / \varphi_{MIR}(Q_t)
\]
where $\varphi_{\text{MIR}}(q)$ is the MAP measure in MIR and $\text{QE}(q)$ is the set of expanded terms obtained from query $q$. Measure $r_1$ estimates how possible the loss semantics caused by not-translated $s_j$ can be recovered by other terms together with its post-translation expansion set (QE and replacement features). Figure 6 shows the relations between $\text{LR}_{\text{MIR}}$ and $r_1$ for each $s_j$. Interestingly, a positive correlation exists between the two variables. If OOV term $s_j$ is slightly effective ($\text{LR}_{\text{MIR}}$ is negative but close to 0) and cannot be translated, the semantics it carries may be rescued by the expansion set of the rest terms. An extremely-effective OOV term $s_j$ ($\text{LR}_{\text{MIR}} << 0$) is the term whose semantics cannot be recovered well ($r_1 << 0$). For those ineffective OOV terms ($\text{LR} > 0$), not-translating such terms is beneficial to CLIR performance.

Measure $r_2$ captures the relevance of OOV term $s_j$ to the rest of the terms. Figure 7 shows a negative correlation between $\text{LR}_{\text{MIR}}$ and $r_2$. Reasonably, an effective OOV term often has higher distinguishing power (TFIDF feature) in finding relevant documents compared to others. Consequently, our features explain why some OOV terms need to be translated while others do not. Especially, a difficult query term with low distinguishing power had better not be translated despite how good its translation quality is. Also, for those which are effective for search, some terms are not necessary to appear in the query.

Like OOV terms analysis, we apply feature selection to non-OOV terms, and the statistical features are most important, including context similarity and co-occurrence features. The lower part of Table 6 demonstrates the classification accuracy for all non-OOV terms. Figure 8 shows the relations between $\text{LR}_{\text{TD}}$ and $\text{cosine} (Q_i \cup \{t_{\text{D}}^{k}\} - \{t_j\}, Q_i)$ for each $s_j$ (context similarity feature). Suppose translation $t_{\text{D}}^{k}$ is extremely ineffective ($\text{LR}_{\text{TD}} << 0$) in CLIR, it can be inferred that $t_{\text{D}}^{k}$ would be irrelevant to $Q_i - \{t_j\}$. Hence, $t_{\text{D}}^{k}$ and $(Q_i - \{t_j\})$ together are dissimilar from $Q_i$. A worse translation usually comes along with a weaker similarity to the original topic.

**Figure 6. LR_{MIR} versus r1.**

**Figure 7. LR_{MIR} versus r2.**

**Table 6. Classification accuracy with selected features.**

<table>
<thead>
<tr>
<th></th>
<th>N4&lt;title&gt;</th>
<th>N4&lt;desc&gt;</th>
<th>N5&lt;title&gt;</th>
<th>N5&lt;desc&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>OOV</td>
<td>90.66%</td>
<td>91.05%</td>
<td>93.82%</td>
<td>89.78%</td>
</tr>
<tr>
<td>Non-OOV</td>
<td>77.33%</td>
<td>69.21%</td>
<td>78.13%</td>
<td>72.20%</td>
</tr>
</tbody>
</table>

5.2 Non-OOV Terms Analysis

To understand the impact of different translations for non-OOV term $s_j$, we define a translation loss ratio as follows:

$$\text{LR}_{\text{TD}}(s_j) = \left( \varphi_{\text{MIR}}(Q_i \cup \{t_{\text{D}}^{k}\} - \{t_j\}) - \varphi_{\text{MIR}}(Q_i) \right) / \varphi_{\text{MIR}}(Q_i) .$$

where $\varphi_{\text{MIR}}(Q)$ is the MAP measure and $t_{\text{D}}^{k}$ denotes the $k$-th translation given in translation resource $D$ (Google Dictionary in our case). $\text{LR}_{\text{TD}}(s_j)$ tells the influence of translating $s_j$ to $t_{\text{D}}^{k}$ in CLIR.

We collect all translations from $D$ for each non-OOV term. A good translation may have positive $\text{LR}_{\text{TD}}$ value, as it outperforms correct translation in the benchmarks. Figure 8 shows the relation between $\text{LR}_{\text{MIR}}$ and $\text{LR}_{\text{TD}}$ for each $s_j$. We can see that an effective term ($\text{LR}_{\text{MIR}}<<0$) usually has good translation quality ($\text{LR}_{\text{TD}}>0$), and this is why we think these terms had better being translated.

**6. DISCUSSIONS AND CONCLUSIONS**

We have proposed an approach to estimate the translation probability of a query term $p(s_j|t_i, Q_i)$, which measures if term $s_j$ should be translated or not.

An important aspect tells that the translations of certain query terms are crucial for high retrieval performance, while some others had better not be translated. The rationale lies in that some terms “inherently” suffer from its ineffectiveness in CLIR, as indicated by MIR loss ratio $\text{LR}_{\text{MIR}}$ in Section 5. A key point is that it is not worth taking a risk to translate a term especially if it is inherently ineffective. Also, the translation quality critically
affects our preference for translation. Specifically, the more number of senses or translations a term has, the more probable a wrongly translated translation could be adopted.

A macroscopic experiment in Section 4.4 shows the feasibility of our approach, where both mono- and cross-lingual retrieval performance can be significantly improved. We further analyze why our approach really works in Section 5. For an OOV term, we discover that it is not always needed for translation, as sometimes the translations in target corpus are ineffective, or are quite irrelevant to the original query topic. Even for those that are effective for search, pre- or post-translation query expansion largely influences the necessity of translation by recovering primordial intention or not. Most importantly, these are all well captured by the defined features, given in Section 3.2. On the other hand, for non-OOV terms, the context features are effective for predicting the translation quality. A wrong translation could also hurt retrieval performance due to its co-occurrence with other terms, which should be carefully dealt with in the future.

Our approach could minimize the efforts of translation by selecting terms that really need it. This is especially helpful under the condition that lexicon coverage is limited or time constraint is restricted, we can translate terms according to the priority embedded in the ranking lists. Further, our approach can be easily extended to predict the effect of translating whole queries on CLIR such as [13].

We are planning to train different models for OOV and non-OOV terms instead of a universal one, as they are intrinsically different from each other. We also want to explore how to automatically choose the best value for parameter $k$, which is anticipated to optimize the retrieval performance for each query topic in our algorithms. Despite the difficulty of automatic determination of $k$, it turns out that a fixed value 2 in $\langle$title$\rangle$ and 4 in $\langle$desc$\rangle$ still work acceptably on all of the tested retrieval models in our experiments. Finally, like all other training-based algorithms, we have to obtain the Web corpus for calculating the statistical features before applying our method to Web applications. We leave these limitations as our future work.

7. REFERENCES


