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. 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 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 untranslated out-of-vocabulary (OOV) query terms and translation quality of non-OOV 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 another 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], phrasal translation [17, 11], and unknown words translation [7]. Such approaches pursue the reduction of erroneous or irrelevant 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.

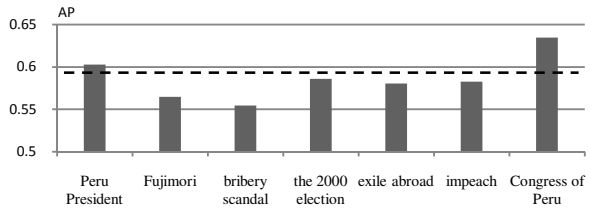

Figure 1: AP for each untranslated query term.

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 stop words removal). Its correct Chinese translations result in average precision (AP) of 0.5914 for CLIR. Figure 1 shows that if one of the query terms is not translated (x-axis), how the corresponding AP (y-axis) changes using the correct translations of the rest of terms as a query. It is observed that without correct translation of "Fujimor" or "bribery scandal", we are far from satisfying retrieval performance, compared to AP 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 AP can be even achieved. Still the missing of the translation such as "the 2000 election", "exile abroad", or "impeach" seems to be tolerable. This observation reveals that some untranslated 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.

In the above example, term "Fujimori" seems to bear more important semantics and thus should be translated. It might appear OOV terms always need perfect translations. Take into account the query from another NTCIR-5 English-Chinese topic (after stop words removal): "Chinese-American, scien-

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.

tist, Wen-Ho Lee, suspect, steal, classified information, nuclear weapon, US's Los Alamos National Laboratory". It could be found that the AP decreases 45.9% when "Wen-Ho Lee" is not translated, whereas untranslated "US's Los Alamos National Laboratory" conversely helps improve 39.6% of AP. Although missing the translation of "US's Los Alamos National Laboratory" loses some information about the query, we notice that term "實驗室" (laboratory) luckily emerges in its (post-translation) feedback documents, which alleviates the problem. Moreover, there are many possible transliterations of "Los Alamos" in Chinese such as "洛薩拉摩" 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 untranslated would probably be a reasonable choice.

Conventional approaches to query translation mostly put efforts in finding accurate translations [15] or examining how translation resources affect CLIR performance [12, 16]. Few did pose the problem of predicting CLIR performance or whether to translate a query term or not. Our most relevant work [10] presented a method to predict the performance of CLIR according to translation quality and ease of queries. Yet [10] focused merely on evaluating the performance of a whole query and did not give insight into the effect of translation for each query term. Also, [10] did not propose the issue of translation necessity which potentially helps improve retrieval performance.

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 untranslated OOV and various translations of non-OOV terms affect CLIR performance.

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, statistical, and CLIR features in source and target language corpora. Experiments on NTCIR-4 and NTCIR-5 English-Chinese CLIR tasks show that CLIR performance can be significantly improved based on our approach. Such effectiveness is consistent across different translation approaches as well as benchmarks. An in-depth analysis of how untranslated OOV terms and translation quality of non-OOV terms influence CLIR performance is also provided. We highlight that query terms needing no translation may result from intrinsical ineffectiveness in CLIR, semantic recovery by query expansion, or poor translation quality.

2. RELATED WORK

Improving translation accuracy is important for query translation. Phrasal translation approach [17, 11] was inspected for improving CLIR performance. Disambiguation of multiple-sense terms by estimating co-occurrence for each chandidate[3] has also shown evident accuracy enhancement. Some others utilized statistical properties in parallel corpus [5, 13] as well as query expansion techniques [2] in search of better translation accuracy. Machine translation techniques [18, 14] are effective for long sentences, but they are not suitable for short, context-inadequate queries. Though these works have brought significant improvement in translation accuracy, they eventually tried to translate as many terms as possible, which we believe is not always an effective approach in CLIR.

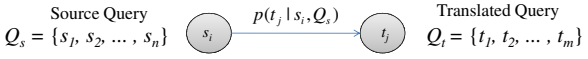

Figure 2: Basic query translation model.

Our work is also related to term selection from a query. The focus of previous works[1, 4] did key-term selection in the mono-lingual environment; however, our discovery of various causes such as pre- and post-translation query expansion would influence the preference of translation in CLIR.

Our most relevant work [10] has developed regression models for predicting the performance of CLIR. The translation quality and ease of query were taken into account. Their concern was evaluated on a whole query, whereas we think every single term has its own impact on CLIR performance. Moreover, their method focused on the goodness of their regression models, while we aim to learn the need of translation for each term and bring up CLIR performance.

3. TRANSLATE QUERY TERMS OR NOT?

3.1 Estimation of Translation Probability

Given a query topic $Q_s = \{s_1, s_2, ..., s_n\}$ in source language, conventional query translation methods endeavor to find a set of translated terms $Q_t = \{t_1, t_2, ..., t_m\}$ in target language. Various translation methodologies such as phrasal translation or sense disambiguation have brought significant improvements in CLIR. Particularly, they incorporate dictionaries, bilingual corpora, or the Web to estimate the probability of translation $p(t_j|s_i, Q_s)$. This probability shows how good it will be to translate s_i to t_j given topic Q_s, as shown in Fig. 2. $p(t_j|s_i, Q_s)$ also means the translation depends on not only s_i and t_j but the rest of terms in Q_s. For simplicity, some previous works focused on $p(t_j|s_i)$ under assumption that the probability is irrelevant from Q_s.

As illustrated in Fig. 1, the effect of query term translation may differ from one to another. Noticing this point, the goal of this paper does not focus on seeking high translation accuracy. Rather, we are interested in realizing, given a source term s_i, whether it should be translated or not. This can be casted as a classification problem by introducing a binary variable $T \in \{0, 1\}$. $T = 1$ 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 estimation of $p(t_j|s_i, Q_s)$, we get the following:

$$\begin{aligned} & p(t_j|s_i, Q_s) \\ &= \textstyle\sum_T p(T|s_i, Q_s) p(t_j|s_i, Q_s, T) \\ &= p(T=0|s_i, Q_s) p(t_j|s_i, Q_s, T=0) \\ &\quad + p(T=1|s_i, Q_s) p(t_j|s_i, Q_s, T=1) \\ &= p(T=1|s_i, Q_s) p(t_j|s_i, Q_s, T=1) \end{aligned}$$

$p(t_j|s_i, Q_s, T = 0)$ equals 0 because the probability of translation from s_i to t_j is 0 given that s_i should not be translated. The final probability is composed of two parts. $p(T = 1|s_i, Q_s)$ estimates if it is suggested that s_i should

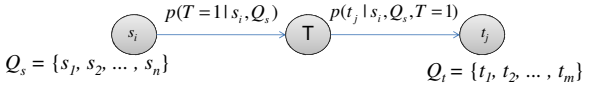

Figure 3: Extended query translation model.

be translated, whereas $p(t_j|s_i, Q_s, T = 1)$ shows how proper it is to translate s_i to some particular t_j. This point is illustrated in Fig. 3 with the newly introduced variable T, where s_i is mapped to t_j only if it is worth being translated ($T = 1$). The focus of previous work lies in generating translation equivalences based on $p(t_j|s_i, Q_s, T = 1)$ or $p(t_j|s_i, Q_s)$, since every s_i is required to be translated by default. Yet the goal of this paper is to predict the probability $p(T = 1|s_i, Q_s)$, which concerns whether to translate s_i or not.

Given $Q_s = \{s_1, s_2, ..., s_n\}$, we formulate our problem by seeking a classifier $c : Q_s \rightarrow T$, and the classification gives probabilities of 0 or 1 in the prediction process. To estimate real numbered probabilities, we resort to finding regression function $r : Q_s \rightarrow R$ which predicts a value indicating the necessity of translating each s_i. With the regressed value, say $r(s_i)$, as input of the Sigmoid function, we can obtain the probability of translating or not-translating ranged within $[0, 1]$. Mathematically,

$$p(T = 1|s_i, Q_s) = 1/(1 + e^{-r(s_i)})$$

With the probabilities (either binary or real numbered) in hand, we could use them in CLIR retrieval models by integrating $p(T = 1|s_i, Q_s)$ into $p(t_j|s_i, Q_s)$. In this paper, we simply use $p(T = 1|s_i, Q_s)$ to translate worthily-translated terms from Q_s. This approach enjoys the flexibility and extendability across various frameworks, because translating some portion of query terms is independent of what retrieval models are adopted. The probabilities based on Sigmoid function rank the source terms with a permutation $\pi : s_{\pi(1)} > s_{\pi(2)} > ... > s_{\pi(n)}$ such that

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

Based on classifier c, query terms are easily classified according to their needs of translation. Similarly, based on regression r, top k query terms $\{s_{\pi(1)}, s_{\pi(2)}, ..., s_{\pi(k)}\}$ will be selected to be translated. Four different translation strategies and various threshold k's have been examined in Sections 4 and 5. Here, we apply support vector machine (SVM) and support vector regression (SVR) [6] to do classification and regression; other alternatives can also be adopted.

Finally, we define that the regressed value $r(s_i)$ equals to $LR_{CLIR}(s_i)$ with the following formula:

$$r(s_i) = LR_{CLIR}(s_i) = \frac{\varphi_{CLIR}(Q_s) - \varphi_{CLIR}(Q_s - \{s_i\})}{\varphi_{CLIR}(Q_s)}$$

where $\varphi_{CLIR}(q)$ is the AP measure for query q in CLIR. The larger the loss ratio $LR_{CLIR}(s_i)$ is, the more importantly we translate s_i due to its better effectiveness in CLIR.

We develop the regression function $r : Q_s \rightarrow R$ 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 and will be described in Section 3.2. For classifier $c : Q_s \rightarrow T$, the form would be $<f(s_i), sign(LR_{CLIR}(s_i))>$, where $sign(x)$ converts LR_{CLIR} into non-negative and negative classes based on x.

3.2 Feature Set

We utilize linguistic (Ling), statistical (Stat), and CLIR features $f(s_i)$ of query term s_i to capture its characteristics from different aspects. We denote t_j as the corresponding translation of s_i in target language.

Table 1 shows all features adopted. Linguistic features include parts of speech (POS), named entities (NE), acronym, phrase, and size (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 binary and are labeled manually.

Statistical features are good predictors from the viewpoints of document corpora. In our experiment, both source and target language corpora are used. Here, we consider co-occurrence, context, and TFIDF features. Co-occurrence features reveal the degree of how often a term co-exists 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. Pointwise mutual information (PMI) is adopted for calculation in pre- and post-translation corpora. Given two sets of terms x and y, we measure their co-existence level by

$$pmi(x, y) = log\{p(x, y)/p(x)p(y)\}.$$

In addition, context features are helpful for low frequency query terms that yet share common contexts in search results. The context vector $v(t)$ takes term(s) t as input, and is composed of a list of search-result pairs <document ID, relevance score> returned by retrieval systems. Given two vectors x and y, we measure their cohesive level by

$$cos(x, y) = (x \cdot y)/(\|x\|\|y\|).$$

For those pairwisely computed sets in co-occurrence or context features, we extract their 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 compute TFIDF in both source and target language corpora for each term.

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 isOOV to indicate if a term exists within the coverage of dictionary.

Expansion features express if the losing information from an untranslated term can be recovered by the semantics from the rest of terms with query expansion. Query expansion in source language reserves the room for untranslated terms by including relevant terms in advance. Also, query expansion in target language recovers the semantics loss by inspecting the rest well-translated terms. Here we denote $QE(q)$ as the set of expanded terms obtained from the search results of query q by query expansion, and θ can either be $pmi()$ or $cos()$ functions. From formulas in Table 1, the measurements estimate the similarity between expanded terms derived with or without term s_i. The same calculation is re-

Table 1: All features used in experiments

Type	Feature	Description
Ling	POS	noun, verb, adjective, adverb
	NE	person name, location, organization, event, time
	Other	acronym, phrase, size
Stat	Co-occurrence	$pmi(s_i, Q_s - \{s_i\})$ $pmi(t_j,\ Q_t - \{t_j\})$ $pmi(s_i, s_p), (\forall p \neq i\ ,\ s_p \in Q_s)$ $pmi(t_j, t_q),\ (\forall q \neq j\ ,\ t_q \in Q_t)$
	Context	$cos(v(s_i), v(Q_s - \{s_i\}))$ $cos(v(t_j),\ v(Q_t - \{t_j\}))$ $cos(v(s_i), v(s_p)), (\forall p \neq i\ ,\ s_p \in Q_s)$ $cos(v(t_j), v(t_q)),\ (\forall q \neq j\ ,\ t_q \in Q_t)$
	TFIDF	Term frequency Inverse document frequency
CLIR	Translation	isOOV (binary OOV indicator) trans-size (# of translations)
	Expansion	$\theta(QE(Q_s - \{s_i\}), QE(Q_s))$ $\theta(QE(Q_t - \{t_j\}),\ QE(Q_t))$
	Replacement	$\theta(Q_s, QE(Q_s - \{s_i\}) \cup (Q_s - \{s_i\}))$ $\theta(Q_t,\ QE(Q_t - \{t_j\}) \cup (Q_t - \{t_j\}))$

peated in target language for each t_j. It is inferred that the more the two expanded sets of terms resemble each other, the more likely the loss information from untranslated s_i can be made up.

Lastly, replacement features estimate if the rest of terms, $(Q_s - \{s_i\})$, within the same topic together with its expanded terms set, $QE(Q_s - \{s_i\})$, can take the place of s_i. If the replacement intension is strong, it implies translation of only the rest of terms is sufficient for retrieval. In other words, $QE(Q_s - \{s_i\})$ replaces the position of s_i in original query Q_s, while $QE(Q_t - \{t_j\})$ substitutes the semantics of t_j in query Q_t.

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 2 (after data clean). The poorly-performing queries whose AP is below 0.02 are filtered to ensure the quality of our training data for classification and regression models. Table 3 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 vector space model (TFIDF), the language model (Indri), and the probabilistic 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 mean average precision (MAP) as performance metric evaluating over top 1000 documents retrieved. To avoid inside test, 5-fold cross validation is used through the entire experiments.

We use correct translations in the benchmarks to train the regression and classification models. The correct translations are available since NTCIR-4 and NTCIR-5 CLIR tasks provide both English and Chinese topics at the same time. Usage of correct translations shall help reveal the necessity of translation. NTCIR-4 and NTCIR-5 CLIR tasks also provide English and Chinese documents, which are used as the source and target language corpora, respectively. Note that the English and Chinese documents are not parallel texts.

[1]Lemur Project: http://www.lemurproject.org

Table 2: Data set of English topics(after data clean)

Setting		# query topics	# distinct words	# avg words per topic
NTCIR4	title	44	216	4.90
	desc	58	865	14.90
NTCIR5	title	35	198	5.65
	desc	47	623	13.20

Table 3: Numbers of OOV and non-OOV terms

Setting		# terms	# OOV	# non-OOV
NTCIR4	title	154	15 (9.8%)	139 (90.2%)
	desc	298	15 (5.0%)	283 (95.0%)
NTCIR5	title	131	27 (20.6%)	104 (79.4%)
	desc	277	36 (13.0%)	241 (87.0%)

4.2 Regression and Classification Performance

The coefficient of determination R^2 measures how well future outcomes are likely to be predicted by the statistical models. 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 4 demonstrates the R^2 results.

Averagely speaking, the best regression performance can be achieved when both pre- and post-translation corpora is used, as query expansion is often helpful in CLIR. Also, it shows that a higher R^2 can be found in post-translation corpus than in pre-translation one. This is because post-translation corpus can provide more effective expanded terms for MIR in the set of target documents. Moreover, within each corpus setting, we go into details to inspect the effectiveness using different features. Statistical features consistently achieve better R^2 than CLIR features, which are followed by linguistic features (R^2 of linguistic features is the same across different corpora since such properties remain still despite change of languages). It is caused by that statistical features reflect the underlying distribution of translated terms in the document collection, and also that CLIR features reveal the degree of translation necessity. Finally, a larger R^2 can be achieved by including more features for training. We also review the classification accuracy under the same settings, where similar results can be found. Roughly speaking, overall classification accuracy climbs up to 80.15% when all features are adopted. As linguistic, statistical and CLIR features are complementary, we use all of the features in the following experiments.

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.

Table 4: Regression performance under various feature sets and document collections

		Pre-translation Corpus				Post-translation Corpus				Pre- and Post-translation Corpora			
Model	Topic	Lin	Stat	CLIR	All	Lin	Stat	CLIR	All	Lin	Stat	CLIR	All
Indri	Title	0.0657	0.5215	0.1720	0.8848	0.0657	0.3442	0.1726	0.8623	0.0657	0.9183	0.5773	0.9878
	Desc	0.0472	0.1322	0.0417	0.5274	0.0472	0.1454	0.0887	0.5990	0.0472	0.4542	0.1793	0.9260
TFIDF	Title	0.1780	0.6872	0.2767	0.7718	0.1780	0.3379	0.3023	0.8555	0.1780	0.9611	0.4611	0.9712
	Desc	0.0879	0.2284	0.0410	0.8268	0.0879	0.3235	0.2328	0.8458	0.0879	0.8062	0.2796	0.9688
Okapi	Title	0.1163	0.6092	0.2154	0.7146	0.1163	0.4046	0.2650	0.8709	0.1163	0.8386	0.3948	0.9820
	Desc	0.0406	0.0423	0.0083	0.3193	0.0406	0.0794	0.0455	0.4604	0.0406	0.3126	0.0766	0.9100
Avg.	Title	0.1200	0.6060	0.2214	0.7904	0.1200	0.3622	0.2466	0.8629	0.1200	0.9060	0.4777	0.9803
Avg.	Desc	0.0586	0.1343	0.0303	0.5578	0.0586	0.1828	0.1223	0.6351	0.0586	0.5243	0.1785	0.9349

Table 5: CLIR performance under various translation resources, document collections, query topics, and prediction methods. T-test with p < 0.01 (**) and p< 0.05 (*) against baseline method

Okapi		Correct Trans	Google Dict Top1	Google Dict All	Google Trans	Average
Ntcir4	Title BL	0.2366	0.0902	0.0659	0.1692	0.1405
	Title UB	0.2774	0.1088	0.0874	0.1966	0.1676
	Title C	0.2475 (+4.60%)	0.1019 (+13.0%)	0.0785* (+19.2%)	0.1875 (+10.8%)	0.1539 (+9.52%)
	Title R	0.2602** (+9.98%)	0.1062* (+17.8%)	0.0775 (+14.6%)	0.1884* (+11.4%)	0.1576 (+12.2%)
	Desc BL	0.2121	0.0876	0.0671	0.1601	0.1317
	Desc UB	0.3025	0.1347	0.1319	0.2168	0.1965
	Desc C	0.2448* (+15.4%)	0.1003* (+14.5%)	0.0998** (+48.7%)	0.1803** (+12.6%)	0.1563 (+18.7%)
	Desc R	0.2493** (+17.5%)	0.1073** (+22.5%)	0.0847** (+26.2%)	0.1856** (+15.9%)	0.1567 (+19.0%)
Ntcir5	Title BL	0.3541	0.1376	0.1065	0.3089	0.2267
	Title UB	0.4253	0.1552	0.1252	0.3496	0.2638
	Title C	0.3945 (+11.4%)	0.1437 (+4.46%)	0.1136 (+6.68%)	0.3299* (+6.79%)	0.2454 (+8.22%)
	Title R	0.4059** (+14.6%)	0.1546* (+12.3%)	0.1235* (+16.0%)	0.3348* (+8.39%)	0.2547 (+12.3%)
	Desc BL	0.357	0.1841	0.0835	0.2728	0.2243
	Desc UB	0.4788	0.2464	0.1893	0.3904	0.3262
	Desc C	0.4349* (+21.8%)	0.2073** (+12.6%)	0.1484** (+77.7%)	0.3267** (+19.8%)	0.2793 (+24.5%)
	Desc R	0.4363** (+22.2%)	0.2102** (+14.2%)	0.1348** (+65.7%)	0.3394** (+24.4%)	0.2810 (+25.3%)

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. Context features are effective through inspecting retrieval results, but such features meantime suffer from higher cost of computation. Another group of useful features are CLIR features. As mentioned previously, CLIR features are crucial for estimation of semantic recovery, which is captured by expansion and replacement features. It is worth noticing that the "isOOV" feature is evidently correlated to retrieval performance. It again assures that efforts in translating OOV terms are significant for CLIR, as indicated by previous work [7]. 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). The more senses (or translations) a term contains, the more challenging correct translation can be detected.

Linguistic features such as NE are relatively important for search. It is not always the case yet is usually true. A named entity often has unshirkable responsibility in describing the information needs especially for short queries. And this is why NE is much more correlated to MAP than POS is. Overall, statistical features are more powerful than linguistic ones. Specifically, context features are more effective than co-occurrence features. CLIR features also contribute substantially in translation estimation.

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 <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.

Table 5 shows the MAP results using translated queries for search. Based on the pre-trained model, we'd like to test if we can improve the CLIR performance with 4 different translation strategies. Each strategy generates its own t_j given source term s_i. "Correct Trans" gives the standard translations in the benchmark, which is also recognized as MIR; "Google Dict top1" extracts the first translation from Google Dictionary[2]; "Google Dict all" combines all possible translations from Google Dictionary for a given term; finally "Google Trans" returns the translations from Google Translation[3]. Moreover, for each setting, we show its baseline and upper bound performance. The baseline methods (BL) suppose entire terms in Q_s need to be translated, and simply combine all the translated terms in Q_t as one query string. For each topic in <title> or <desc>, there are in total $2^m - 1$ possible subclasses of Q_t, considering that each s_i in Q_s can be translated to t_j or not. We construct the upper bounds (UB) by discovering the subclass (sub-query as well) with the highest AP. We also run the two-sample pairwise significance test against BL.

From Table 5, our classification (C) and regression (R) models consistently outperform the baseline methods using different translation strategies. The retrieval result proves

[2]Google Dictionary: http://www.google.com.tw/dictionary
[3]Google Translation: http://translate.google.com/?hl=en#

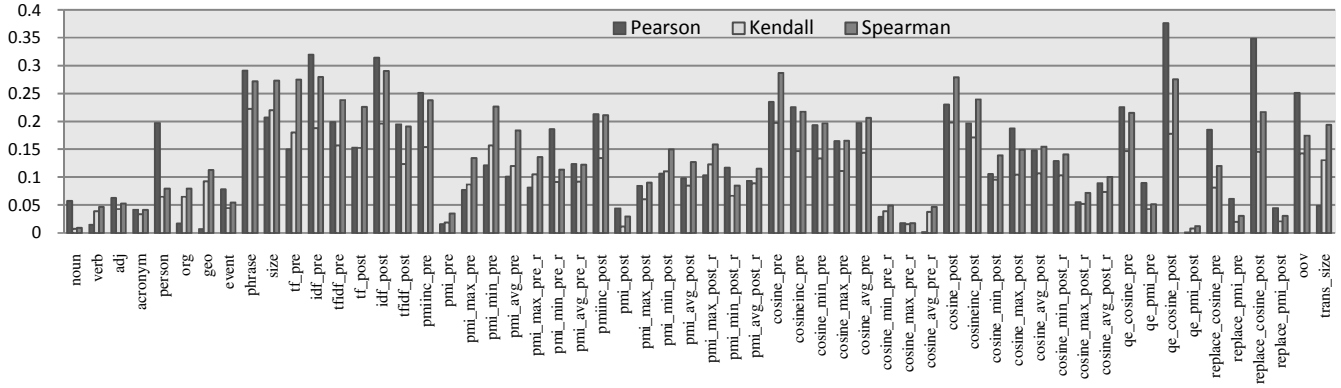

Figure 4: Absolute values of correlation using Okapi retrieval model on NTCIR-4 data set.

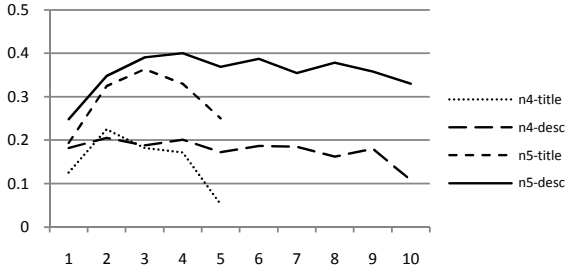

Figure 5: MAP with various k (number of top terms translated) on different dataset.

our observation that some terms are "meant to be" translated while others are not. It is also worth noticing that the improvement rate of description queries is larger than title 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 strategies, we find that "Google Dict all" leaves the most room for improvement. We attribute this to the ambiguity it involves in due to inclusion all translations in dictionary. Fig. 5 illustrates the impact of the variable k.

5. TRANSLATION ANALYSIS

In this section, we discuss the effect of translating OOV and non-OOV query terms on CLIR. We will explore what factors make a query term favorable for translation (T=1), adverse for translation (T=0), or just somewhere in between.

Given a query topic $Q_s = \{s_1, s_2, ..., s_n\}$, we denote its correct translation as $Q'_t = \{t'_1, t'_2, ..., t'_n\}$ where t'_j is the correct translation of s_j. In the following, we will stick to $LR_{MIR}(s_j)$ which indicates the necessity of translation based on s_j, and is defined as

$$LR_{MIR}(s_j) = \frac{\varphi_{MIR}(Q'_t) - \varphi_{MIR}(Q'_t - \{t'_j\})}{\varphi_{MIR}(Q'_t)},$$

where $\varphi_{MIR}(q)$ is the AP measure for query q in MIR. $LR_{MIR}(s_j)$ tells the influence of translating s_j to t'_j. Terms

Table 6: Average raking percentages (ARP)(x100%) and proportion of effective and ineffective OOV terms (N4: NTCIR-4, N5: NTCIR-5)

	Title		Desc	
LR_{MIR}	≥ 0	< 0	≥ 0	< 0
N4 ARP	0.6794	0.8750	0.2996	0.3095
N5 ARP	0.6507	0.6667	0.2705	0.5067
Prop.	83.3%	16.7%	76.5%	23.5%

Table 7: Classification accuracy with selected features (N4: NTCIR-4, N5: NTCIR-5)

	N4<title>	N4<desc>	N5<title>	N5<desc>
OOV	90.66%	91.05%	93.82%	89.78%
Non-OOV	77.33%	69.21%	78.13%	72.20%

with positive $LR_{MIR}(s_j)$ are thought intrinsically-effective in target language and had better be translated.

5.1 OOV Terms Analysis

We discuss how untranslated 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 3 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 (2/5)*100%. Following this manner, it is expected that effective terms ($LR_{MIR} > 0$), i.e., the terms need to be translated, are usually ranked in front of ineffective ones ($LR_{MIR} \leq 0$) and thus have smaller average ranking percentage. Table 6 reveals the reliability of our ranking lists. In addition, it is worth noting that for longer queries such as <desc>, we can earlier discover should-be-translated terms, as the ranking percentage in <desc> is often smaller. The result is somehow not surprising since longer queries usually contain more noises.

By calculating the proportions of effective terms and ineffective terms, Table 6 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%).

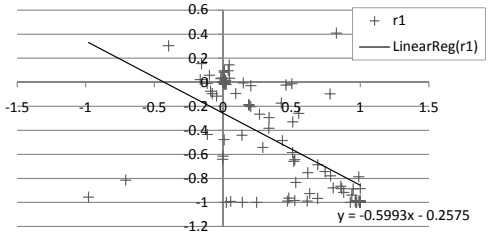

Figure 6: LR_{MIR} (x-axis) versus $r1$ (y-axis).

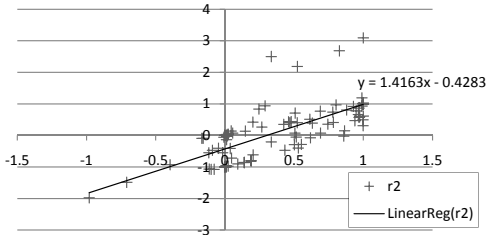

Figure 7: LR_{MIR} (x-axis) versus $r2$ (y-axis).

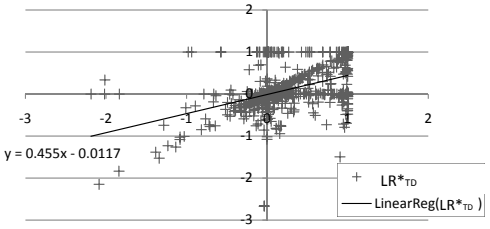

Figure 8: LR_{MIR} (x-axis) versus LR_{TD*} (y-axis).

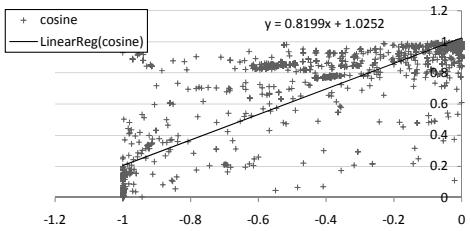

Figure 9: LR_{TD} (x-axis) versus *cosine* (y-axis).

The result is consistent with [7]. The untranslated OOV terms in short queries, especially for title queries or Web queries, crucially influence retrieval performance.

Moreover, we show what factors (features) distinguish effective and ineffective OOV terms. We collect all OOV instances from the entire dataset to train a classifier. The upper part of Table 7 shows the cross-validation 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, expansion and TFIDF features. We propose the following formula to reveal the capability of these features in predicting the need of translation for OOV terms,

$$r1 = \frac{\varphi_{MIR}(QE(Q'_t - \{t'_j\}) \cup Q'_t - \{t'_j\}) - \varphi_{MIR}(Q'_t)}{\varphi_{MIR}(Q'_t)}$$

$$r2 = \frac{\varphi_{MIR}(t'_j) - \varphi_{MIR}(Q'_t - \{t'_j\})}{\varphi_{MIR}(Q'_t)}$$

Measure $r1$ estimates how possible the loss semantics caused by untranslated s_j can be recovered by other terms together with its post-translation expanded terms (expansion and replacement features). Figure 6 shows the relations between LR_{MIR} and $r1$ for each s_j. Interestingly, a negative correlation exists between the two variables. If OOV term s_j is slightly effective (LR_{MIR} is positive but close to 0) and cannot be translated, the semantics it carries may be rescued by expansion of the rest terms. An extremely-effective OOV term s_j ($LR_{MIR} \gg 0$) is the term whose semantics cannot be recovered well ($r1 \ll 0$). For those ineffective OOV terms ($LR_{MIR} < 0$), not-translating such terms is beneficial to CLIR performance.

Measure $r2$ captures the relevance of OOV term s_j to the rest of the terms. Figure 7 shows a positive correlation between LR_{MIR} and $r2$. Reasonably, an effective OOV term often has higher distinguishing power (TFIDF feature) in locating 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 even if its correct translation can be obtained. Also, for those which are effective for search, some terms are not necessary to appear in the query.

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:

$$LR_{TD}(s_j) = \frac{\varphi_{MIR}(Q'_t \cup \{t^k_{D(s_j)}\} - \{t'_j\}) - \varphi_{MIR}(Q'_t)}{\varphi_{MIR}(Q'_t)}$$

where $t^k_{D(s_j)}$ denotes the k-th translation given in translation resource D (Google Dictionary in our case). $LR_{TD}(s_j)$ tells the influence of translating s_j to $t^k_{D(s_j)}$ in CLIR. Translations with non-negative LR_{TD} are regarded having good translation quality, as they perform as well as or better than correct translation in the benchmarks.

We first focus on the most effective translation $t^*_{D(s_j)}$ that has the highest LR_{TD} (i.e., LR^*_{TD}) for each s_j. From Fig. 8, LR_{MIR} and LR^*_{TD} exhibit a positive relation for each s_j. We can see that an effective term ($LR_{MIR} > 0$) usually has good translation quality ($LR^*_{TD} > 0$) if one is able to find the best translation $t^*_{D(s_j)}$ in D, and this is why these terms need to be translated. It also means that we will gain more performance boost ($LR^*_{TD} > 0$) if we translate the should-be-translated terms ($LR_{MIR} > 0$).

In addition, we apply feature selection to non-OOV terms as what we do in the OOV analysis, and we find that statistical features are the most important, including context and co-occurrence features. The lower part of Table 7 demonstrates the classification accuracy for all non-OOV terms. Figure 9 shows the relations between LR_{TD} and $\cos(Q'_t \cup \{t^k_{D(j)}\} - \{t'_j\}, Q'_t)$ for each s_j (context feature). Assuming that translation $t^k_{D(s_j)}$ is extremely ineffective ($LR_{TD} \ll 0$) in CLIR, it can be inferred that $t^k_{D(s_j)}$ would be irrelevant to $Q'_t - \{t'_j\}$. Hence, $t^k_{D(s_j)}$ and $(Q'_t - \{t'_j\})$ together are dissimilar from Q'_t. We can see that a worse translation usually comes along with a weaker similarity to the original topic.

Context features indeed help distinguish diverse translation qualities and help in prediction of T.

6. CONCLUSIONS AND DISCUSSIONS

In this paper, we propose an approach to estimate the translation probability of query term s_i, $p(T = 1|s_i, Q_s)$, indicating if s_i should be translated or not. One of our merits is that we consider comprehensive factors including linguistic, statistical, and CLIR aspects to predict T. It shows that T is influenced by intrinsic ineffectiveness, semantic recovery by query expansion, or poor translation quality. We also verify that translating should-be-translated terms indeed helps improve CLIR performance across various translation methods, retrieval models, and benchmarks.

Realizing what factors determine translation necessity is important. For an OOV term, we discover that it does not always need translation, as sometimes translations in target corpus are ineffective or are irrelevant to original query. Specifically, leaving s_i untranslated could be a wise choice if its semantics could be recovered by pre- or post-translation expansion. For a non-OOV term, we show that if there exists an effective translation in dictionaries, it is suggested that translating s_i would help CLIR performance. Context features are useful for predicting translation quality. If the translations tend to deviate original intention, we'd better leave s_i untranslated. It is not worth taking a risk to translate a term if the term probably perform poorly in CLIR.

In brief sum, "to-translate-or-not-to-translate" is influenced by various and complicated causes. Some should-not-be-translated terms inherently suffer from their ineffectiveness in CLIR. Still others are affected by the translation quality obtained. These findings are consistent with [10]. Our approach could minimize the efforts of translation by selecting terms that really need it. This is especially helpful under condition that lexicon coverage or time constraint is limited, we can translate terms according to the ranking lists. Further, our approach can be easily extended to predict the effect of translating whole queries on CLIR as in [10].

We plan 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. Despite the difficulty of automatic determination of k, it turns out that a fixed value 2 in <title> and 4 in <desc> work acceptably in our experiments. Finally, we need the Web corpus for calculating statistical features before applying our method to Web applications. We leave these limitations as our future work.

7. ACKNOWLEDGMENTS

This work was supported by the National Science Council, Taiwan, under contract NSC97-2221-E-002-222-MY2.

8. REFERENCES

[1] J. Allan, J. Callan, W. B. Croft, L. Ballesteros, J. Broglio, J. Xu, and H. Shu. Inquery at trec-5. In *Proc. of the Fifth Text Retrieval Conference TREC-5*, pages 119–132, 1997.

[2] L. Ballesteros and W. B. Croft. Dictionary methods for cross-lingual information retrieval. In *Database and Expert Systems Applications*, pages 791–801, 1996.

[3] L. Ballesteros and W. B. Croft. Resolving ambiguity for cross-language retrieval. In *Proc. of ACM-SIGIR '98*, pages 64–71, 1998.

[4] M. Bendersky and W. B. Croft. Discovering key concepts in verbose queries. In *Proc. of ACM-SIGIR '08*, 2008.

[5] J. Carbonell, Y. Yang, R. Frederking, R. Brown, Y. Geng, and D. Lee. Translingual information retrieval: A comparative evaluation. In *Proc. of IJCAI*, pages 708–715, 1997.

[6] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[7] P.-J. Cheng, J.-W. Teng, R.-C. Chen, J.-H. Wang, W.-H. Lu, and L.-F. Chien. Translating unknown queries with web corpora for cross-language information retrieval. In *Proc. of ACM-SIGIR '04*, pages 146–153, 2004.

[8] M. Federico and N. Bertoldi. Statistical cross-language information retrieval using n-best query translations. In *Proc. of ACM-SIGIR '02*, pages 167–174, 2002.

[9] J. Gao, J.-Y. Nie, E. Xun, J. Zhang, M. Zhou, and C. Huang. Improving query translation for cross language information retrieval using statistical models. In *Proc. of ACM-SIGIR '01*, pages 96–104, 2001.

[10] K. Kishida. Prediction of performance of cross-language information retrieval using automatic evaluation of translation. *Library & Information Science Research*, 30(2):138–144, 2008.

[11] J. Kupiec. An algorithm for finding noun phrase correspondences in bilingual corpora. In *Proc. of ACL*, pages 17–22. Association for Computational Linguistics, 1993.

[12] P. McNamee and J. Mayfield. Comparing cross-language query expansion techniques by degrading translation resources. In *Proc. of ACM-SIGIR '02*, pages 159–166, 2002.

[13] J.-Y. Nie, M. Simard, P. Isabelle, and R. Durand. Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web. In *Proc. of ACM-SIGIR '99*, pages 74–81, 1999.

[14] D. Oard. A comparative study of query and document translation for cross language information retrieval. *Machine Translation and the Information Soup*, pages 472–483, 1998.

[15] D. Oard and A. Diekema. Cross-language information retrieval. *Anne Diekema*, page 5, 1998.

[16] A. Pirkola. The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In *Proc. of ACM-SIGIR '98*, pages 55–63, 1998.

[17] F. Smadja, K. McKeown, and V. Hatzivassiloglou. Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, 22(1):1–38, 1996.

[18] J. Zhu and H. Wang. The effect of translation quality in mt-based cross-language information retrieval. In *Proc. of ACL*, pages 593–600. Association for Computational Linguistics, 2006.