

# Study of Cross Lingual Information Retrieval Using On-line Translation Systems

Rong Jin

Dept. of Computer Science and Engineering  
Michigan State University  
East Lansing, MI 48824  
rongjin@cse.msu.edu

Joyce Y. Chai

Dept. of Computer Science and Engineering  
Michigan State University  
East Lansing, MI 48824  
jchai@cse.msu.edu

## ABSTRACT

Typical cross language retrieval requires special linguistic resources, such as bilingual dictionaries and parallel corpus. In this study, we focus on the cross lingual retrieval problem that only uses online translation systems. We compare two approaches: a translation-based approach that directly translates queries into the language of documents and then applies traditional information retrieval techniques; and a model-based approach that first learns a statistical translation model from the translations acquired from an online translation system and then applies the learned statistical model to cross lingual information retrieval. Our empirical study with ImageCLEF has shown the model-based approach performs significantly better than the translation-based approach.

## Categories and Subject Descriptors

H.3.3 [Information storage and Retrieval]: Information Search and Retrieval—*Retrieval Models*

## General Terms

Algorithms and Experimentation

## Keywords

Cross lingual information retrieval, statistical model

## 1. INTRODUCTION

Cross lingual information retrieval (CLIR) has been one of the major research areas in information retrieval during last few years [1, 2, 4-6]. Most CLIR algorithms fall into two categories: translation-based approaches and approaches based on statistical models.

Typically, a translation-based approach translates a query into the language of documents, and relevant documents are found by matching the translated queries with documents [1]. Different approaches are applied to translate queries, ranging from the simplest one based on bilingual dictionaries to the sophisticated one based on a full-scale machine translation system. Compared to the dictionary-based translation, a full-scale machine translation system has the advantage in that it can reduce the

translation ambiguity of a query using the context information. However, when a query is truly ambiguous and multiple possible translations need to be considered, a translation based CLIR approach can perform poorly. Thus, it is important for a translation system based CLIR approach to maintain the uncertainty in translating queries when queries are ambiguous.

A model-based approach usually utilizes the existing statistical machine translation models that were developed by the IBM group [3]. Given a translation model  $\theta$ , the relevance of a document  $d$  to a given query  $q$  is computed as  $p(q|d;\theta)$ , which is the likelihood of translating document  $d$  into query  $q$ . Compared to the translation-based approaches, the model-based approaches have an advantage in that the uncertainties in translating queries are maintained through the usage of translation probabilities which are learned from a parallel corpus. However, building a statistical translation model requires a large bilingual parallel corpus, which is usually expensive to acquire.

In this paper, we present an approach that first utilizes an online translation system to create a bilingual parallel corpus and then learns a statistical translation model based on the established bilingual corpus. Unlike the translation-based approaches where only the best translation is used in information retrieval, this approach maintains the uncertainty in translation and therefore will be more robust to the translation errors. On the other hand, unlike the typical model-based approaches, which require large bilingual parallel corpora, this approach automatically creates a bilingual parallel corpus by applying an online system to translate test documents into the language of queries.

## 2. LEARNING A STAT. MODEL FROM AN ONLINE TRANSLATION SYSTEM

We will first briefly review the statistical translation model and then describe how to acquire parallel corpus using online translation systems.

### 2.1 Statistical Translation Models

Let's assume that the language of queries is Chinese and the language of documents is English. Let  $\theta = \{t(w_i^e | w_j^c)\}$  be the set of translation probabilities.  $t(w_i^e | w_j^c)$  is the probability of translating a Chinese word  $w_j^c$  into an English word  $w_i^e$ . Let  $\Omega = \{(s_i^c, s_i^e)\}$  be the bilingual parallel corpus. Each  $(s_i^c, s_i^e)$  is a translation pair in which sentence  $s_i^c$  is the Chinese translation of

the English sentence  $s_i^e$ . Then,  $p(s_i^e | s_i^c; \theta)$ , i.e., the probability of translation an English sentence  $s_i^e$  into a Chinese sentence  $s_i^c$ , can be written as:

$$p(s_i^e | s_i^c; \theta) \propto \prod_j \left( \sum_k o(w_k^c, s_i^c) t(w_j^e | w_k^c) \right)^{o(w_j^e, s_i^e)}$$

where  $o(w_k^c, s_i^c)$  and  $o(w_k^e, s_i^e)$  are the term frequency of Chinese and English words in sentence  $s_i^c$  and  $s_i^e$ , respectively. Thus, to learn translation probabilities, we maximize the log-likelihood of all translation pairs in the parallel corpus, i.e.  $\theta = \arg \max_{\theta \in \Theta} \sum_{i=1}^N \log p(s_i^e | s_i^c; \theta)$ . Finally, to estimate the relevancy of a document  $d$  to a given query  $q$ , probability  $p(q | d)$  is estimated using the following expression:

$$\begin{aligned} \log p(q^c | d^e; \theta) &= \log \int dq^e p(q^c | q^e) p(q^e | d^e) \\ &\propto \sum_{i=1}^{V_e} \left( \sum_{j=1}^{V_c} o(w_j^c, q^c) p(w_i^e | w_j^c) \right) \log p(w_i^e | d^e) \end{aligned}$$

## 2.2 Acquiring Bilingual Parallel Corpus

To automatically acquire a large bilingual corpus for training statistical translation models, we applied a simple strategy. We first used an online translation system to translate English documents into Chinese. To enhance the diversity of our translation pairs, the Chinese documents that were generated by the online translation system were further translated back into English documents. The final bilingual corpus is created by aggregating the Chinese-English pairs in both translations together. The online translation system used in our study is Systran (<http://www.systransoft.com/>). With the acquired translation pairs, we can now learn translation probabilities between Chinese words and English words. Examples of learned translation probabilities are listed in Table 1. Note that all English words are stemmed using the Porter algorithm.

Chinese	English	Prob.	Chinese	English	Prob.
塔	tower	0.8692	大教堂	cathedr	0.7312
	turret	0.0200		st	0.0475
	pinnac	0.0198		iona	0.0231
	build	0.0048		dunblan	0.0161
	clock	0.0044		eli	0.0152

Table 1. Examples of translation probabilities.

## 3. EXPERIMENT

In this experiment, we examine- the CLIR approach that learns a statistical translation model from the bilingual corpus generated by an online translation system. The document collection used in this experiment comes from ImageCLEF (<http://ir.shef.ac.uk/imageclef/>). The collection consists of 28,133 English documents and each document provides content description for a historical picture. The average length of each document is about 50 words. The 25 Chinese queries provided by ImageCLEF 2004 are used in our evaluation. The baseline approach simply applies the online translation system to translate Chinese queries into English, which we refer as ‘translation-based approach’. The average precision across 11 retrieval points and

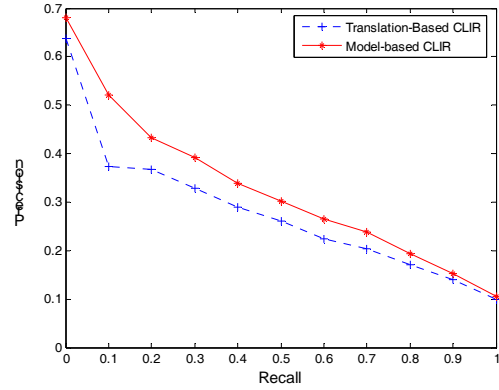


Figure 1: Precision-recall

the precisions for the top ranked documents are summarized in Table 2. Figure 1 shows the precision-recall curves. From these results, we clearly see that the model-based approach outperform the translation-based approach substantially, and the improvement is most noticeable for the top ranked documents.

Prec@	Translation-based	Model-based
5 doc	0.288	0.416
10 doc	0.260	0.344
100 doc	0.150	0.161
<b>Avg Prec.</b>	<b>0.245</b>	<b>0.293</b>

Table 2. Precision results.

## 4. CONCLUSION

In the study, we examine the CLIR approach that learns a statistical translation model from an automatically generated parallel corpus by an online translation system. Our empirical study with documents from ImageCLEF has shown that this approach is more effective than the translation-based approach that directly applies the online translation system to translate queries. Given the limitation of automatic translation systems, our plan is to include bilingual dictionaries to improve the quality of the automatically generated parallel corpus.

## 5. REFERENCES

- [1] Ballesteros, L. and W.B. Croft. Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval. in *Proc. SIGIR'97*. 1997..
- [2] Ballesteros, L. and W.B. Croft. Resolving Ambiguity for Cross-Language Retrieval. in *Proc. SIGIR'98*. 1998.
- [3] Brown, P., et al., The Mathematics of Statistical Machine Translation. *Computational Linguistics*, 1993. **19**(2): p. 263-311.
- [4] Federico, M. and N. Bertoldi. Statistical Cross-Language Information Retrieval Using N-Best Query Translations. in *Proc. SIGIR'02*. 2002.
- [5] Gao, J., et al. Improving Query Translation for Cross-Language Information Retrieval Using Statistical Models. in *Proc. SIGIR'01*. 2001.
- [6] Lavrenko, V., M. Choquette, and W.B. Croft. Cross-Lingual Relevance Model. in *Proc. SIGIR'02*. 2002