# An Improvement in Cross-Language Document Retrieval Based on Statistical Models

Long-Yue WANG
Department of Computer and Information Science
University of Macau
vincentwang0229@hotmail.com

Derek F. WONG
Department of Computer and Information Science
University of Macau
derekfw@umac.mo

Lidia S. CHAO
Department of Computer and Information Science
University of Macau
lidiasc@umac.mo

## Abstract

This paper presents a proposed method integrated with three statistical models including **T**ranslation model, **Q**uery generation model and **D**ocument retrieval model for cross-language document retrieval. Given a certain document in the source language, it will be translated into the target language of statistical machine translation model. The query generation model then selects the most relevant words in the translated version of the document as a query. Finally, all the documents in the target language are scored by the document searching model, which mainly computes the similarities between query and document. This method can efficiently solve the problem of translation ambiguity and query expansion for disambiguation, which are critical in Cross-Language Information Retrieval. In addition, the proposed model has been extensively evaluated to the retrieval of documents that: 1) texts are long which, as a result, may cause the model to over generate the queries; and 2) texts are of similar contents under the same topic which is hard to be distinguished by the retrieval model. After comparing different strategies, the experimental results show a significant performance of the method with the average precision close to 100%. It is of a great significance to both cross-language searching on the Internet and the parallel corpus producing for statistical machine translation systems.

Keywords: Cross-Language Document Retrieval, Statistical Machine Translation, TF-IDF, Document Translation-Based.

## 1. Introduction

With the flourishing development of the Internet, the amount of information from a variety of domains is rising dramatically. Although the researchers have done a lot to develop high performance and effective monolingual Information Retrieval (IR), the diversity of information source and the explosive growth of information in different languages drove a great need for IR systems that could cross language boundaries [1].

Cross-Language Information Retrieval (CLIR) has become more important for people to access the information resources written in various languages. Besides, it is of a great significance to alignment documents in multiple languages for Statistical Machine Translation (SMT) systems, of which quality is heavily dependent upon the amount of parallel sentences used in constructing the system.

In this paper, we focus on the problems of translation ambiguity, query generation and searching score which are keys to the retrieval performance. First of all, in order to increase the probability that the best translation can be selected from multiple ones, which occurs in the target documents, the context and the most likely probability of the whole sentence should be considered. So we apply document translation approach using SMT model instead of query translation, although the latter one may require fewer computational resources. After the source documents are translated into the target language, the problem is transformed from bilingual environment to monolingual one, where conventional IR techniques can be used for document retrieval. Secondly, some terms in a certain document will be selected as query, which can distinguish the document from others. However, some of the words occur too frequently to be useful, which cannot distinguish target documents. This mostly includes two types, one is that the word frequency is high both in the current and the whole document set, which is usually classified as stop word; the other is that the frequency is moderate in several documents (not the whole document set). This type of words gives low discrimination power to the document, and is known as low discrimination word. Thus, the query generation model should filter the words which are of these types and pick the words that occur more frequently in a certain document while less frequently in the whole document set. Finally, the document searching model scores each document according to the similarity between generated query and the document. This model should give a higher mark to the target document which covers the most relevant words in the given query.

There are two cases to be considered when we investigated the method. In one case, both the source and target documents are long text, which are hard to extract exact query from the large amounts of information. In the other case, the contents of the documents are very similar, which are not easy to distinguish for retrieval. The results of experiments reveal that the proposed model shows a very good performance in dealing with both cases.

The paper is organized as follows. The related works are reviewed and discussed in Section 2. The proposed CLIR approach based on statistical models is described in Section 3. The resources and configurations of experiments for evaluating the system are detailed in Section 4. Results, discussion and comparison between different strategies are given in Section 5 followed by a conclusion and future improvements to end the paper.

## 2. Related Work

CLIR is the circumstance in which a user tries to search a set of documents written in one language for a query in another language [2]. The issues of CLIR have been discussed from different perspectives for several decades. In this section, we briefly describe some related methods.

On the matching strategies for CLIR, query translation is most widely used method due to its tractability. However, it is relatively difficult to resolve the problem of term ambiguity because “queries are often short and short queries provide little context for disambiguation” [3]. Hence, some researchers have used document translation method as the opposite strategies to improve translation quality, since more varied context within each document is available for translation [4, 5].

However, another problem introduced based on this approach is word (term) disambiguation, because a word may have multiple possible translations [3]. Significant efforts have been devoted to this problem. Davis and Ogden [6] applied a part-of-speech (POS) method which requires POS tagging software for both languages. Marcello et al. presented a novel statistical method to score and rank the target documents by integrating probabilities computed by query-translation model and query-document model [7]. However, this approach cannot aim at describing how users actually create queries which have a key effect on the retrieval performance. Due to the availability of parallel corpora in multiple languages, some authors have tried to extract beneficial information for CLIR by using SMT techniques. Sánchez-Martínez et al. [8] applied SMT technology to generate and translate queries in order to retrieve long documents.

Some researchers like Marcello, Sánchez-Martínez et al. have attempted to estimate translation probability from a parallel corpus according to a well-known algorithm developed by IBM [9]. The algorithm can automatically generate a bilingual term list with a set of probabilities that a term is translated into equivalents in another language from a set of sentence alignments included in a parallel corpus. The IBM Model 1 is the simplest among the five models and often used for CLIR. The fundamental idea of the Model 1 is to estimate each translation probability so that the probability represented is maximized

$$P(t \mid s) = \frac{\varepsilon}{(l+1)^m} \prod_{j=1}^{m} \sum_{i=0}^{l} P(t_j \mid s_i) \tag{1}$$

where $t$ is a sequence of terms $t_1$, ..., $t_m$ in the target language, $s$ is a sequence of terms $s_1$, ..., $s_l$ in the source language, $P(t_j|s_i)$ is the translation probability, and $\mathcal{E}$ is a parameter ($\mathcal{E}$ $=P(m|e)$), where $e$ is target language and $m$ is the length of source language). Eq. (1) tries to balance the probability of translation, and the query selection, in which problem still exists: it tends to select the terms consisting of more words as query because of its less frequency, while cutting the length of terms may affect the quality of translation. Besides, the IBM model 1 only proposes translations word-by-word and ignores the context words in the query. This observation suggests that a disambiguation process can be added to select the correct translation words [3]. However, in our method, the conflict can be resolved through contexts.

## 3. Proposed Model

The approach relies on three models: translation model which generates the most probable translation of source documents; query generation model which determines what words in a document might be more favorable to use in a query; and document searching model, which evaluates the similarity between a given query and each document in the target document set. The workflow of the approach for CLIR is shown in Fig. 1.

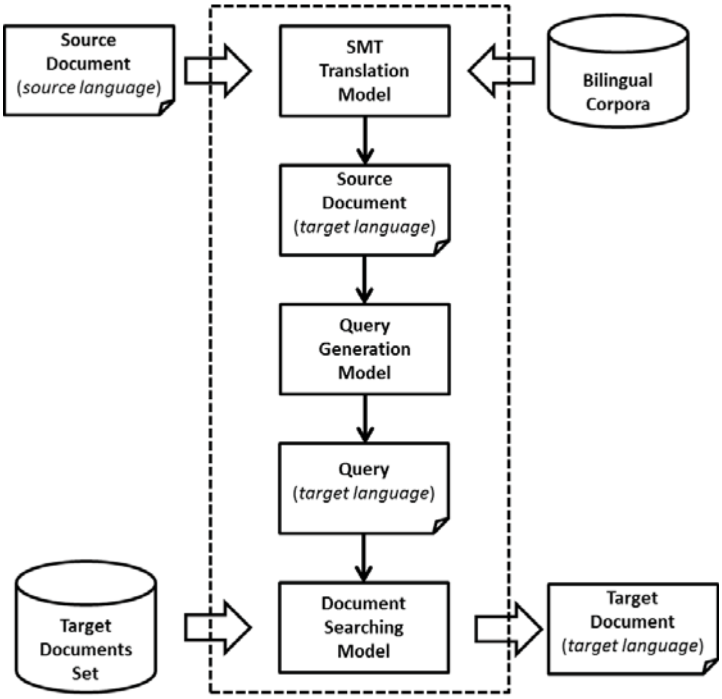

Figure 1. Approach for CLIR

### 3.1. Translation Model

Currently, the good performing statistical machine translation systems are based on phrase-based models which translate small word sequences at a time. Generally speaking, translation model is common for contiguous sequences of words to translate as a whole. Phrasal translation is certainly significant for CLIR [10], as stated in Section 1. It can do a good job in dealing with term disambiguation.

In this work, documents are translated using the translation model provided by Moses, where the log-linear model is considered for training the phrase-based system models [11], and is represented as:

$$p(e_1^I \mid f_1^J) = \frac{\exp(\sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J))}{\sum_{e'^I_1} \exp(\sum_{m=1}^{M} \lambda_m h_m(e'^I_1, f_1^J))} \tag{2}$$

where $h_m$ indicates a set of different models, $\lambda_m$ means the scaling factors, and the denominator can be ignored during the maximization process. The most important models in Eq. (2) normally are phrase-based models which are carried out in source to target and target to source directions. The source document will maximize the equation to generate the translation including the words most likely to occur in the target document set.

### *3.2. Query Generation Model*

After translating the source document into the target language of the translation model, the system should select a certain amount of words as a query for searching instead of using the whole translated text. It is for two reasons, one is computational cost, and the other is that the unimportant words will degrade the similarity score. This is also the reason why it often responses nothing from the search engines on the Internet when we choose a whole text as a query.

In this paper, we apply a classical algorithm which is commonly used by the search engines as a central tool in scoring and ranking relevance of a document given a user query. Term Frequency–Inverse Document Frequency (TF-IDF) calculates the values for each word in a document through an inverse proportion of the frequency of the word in a particular document to the percentage of documents where the word appears [12]. Given a document collection $D$, a word $w$, and an individual document $d \in D$, we calculate

$$P(w,d) = f(w,d) \times \log \frac{|D|}{f(w,D)} \tag{3}$$

where $f(w, d)$ denotes the number of times $w$ that appears in $d$, $|D|$ is the size of the corpus, and $f(w,D)$ indicates the number of documents in which $w$ appears in $D$ [13].

In implementation, if $w$ is an Out-of-Vocabulary term (OOV), the denominator $f(w,D)$ becomes zero, and will be problematic (divided by zero). Thus, our model makes $log$ ($|D|/f(w,D)$)=1 ($IDF$=1) when this situation occurs. Additionally, a list of stop-words in the target language are also used in query generation to remove the words which are high frequency but less discrimination power. Numbers are also treated as useful terms in our model, which also play an important role in distinguishing the documents. Finally, after evaluating and ranking all the words in a document by their scores, we take a portion of the ($n$-best) words for constructing the query and are guided by:

$$Size_q = [\lambda_{percent} \times Len_d] \tag{4}$$

$Size_q$ is the number of terms. $\lambda_{percent}$ is the percentage and is manually defined, which determines the $Size_q$ according to $Len_d$, the length of the document. The model uses the first $Size_q$-th words as the query. In another word, the larger document, the more words are selected as the query.

### *3.3. Document Retrieval Model*

In order to use the generated query for retrieving documents, the core algorithm of the document retrieval model is derived from the Vector Space Model (VSM). Our system takes this model to calculate the similarity of each indexed document according to the input query. The final scoring formula is given by:

$$Score(q,d) = coord(q,d) \sum_{t_{in} q} tf(t,d) \times idf(t) \times bst \times norm(t,d) \tag{5}$$

where $tf(t,d)$ is the term frequency factor for term $t$ in document $d$, $idf(t)$ is the inverse document frequency of term $t$, while $coord(q,d)$ is frequency of all the terms in query occur in a document. $bst$ is a weight for each term in the query. $Norm(t,d)$ encapsulates a few (indexing time) boost and length factors, for instance, weights for each document and field.

As a summary, many factors that could affect the overall score are taken into account in this model.

## 4. Model Evaluation

### *4.1. Datasets*

In order to evaluate the retrieval performance of the proposed model on text of cross languages, we use the Europarl corpus which is the collection of parallel texts in 11 languages from the proceedings of the European Parliament [13]. The corpus is commonly used for the construction and evaluation of statistical machine translation[1]. The corpus consists of spoken records held at the European Parliament and are labeled with corresponding IDs (e.g. <CHAPTER *id*>, <SPEAKER *id*>). The corpus is quite suitable for use in training the proposed probabilistic models between different language pairs (e.g. English-Spanish, English-French, English-German, etc.), as well as for evaluating retrieval performance of the system.

Among the existing CLIR approaches, the work of Sánchez-Martínez et al. [8] based on SMT techniques and IBM Model 1 is very closed to our approach proposed in this paper. We take it as the benchmark and compare our model against this standard. In order to be able to compare with their results, we used the same datasets (training and testing data) for this evaluation. The chapters from April 1998 to October 2006 were used as a training set for model construction, both for training the Language Model (LM) and Translation Model (TM). While the chapters from April 1996 to March 1998 were considered as the testing set for evaluating the performance of the model.

We split the test set into two parts: (1) TestSet1, where each chapter (split by <CHAPTER *id*> label) is treated as a document, for tackling the large amount of information in long texts. (2) TestSet2, where each paragraph (split by <SPEAKER *id*> label) is treated as a document, for dealing with the low discrimination power. The analytical data of the corpus are presented in Table 1. There are 1,022 documents in TestSet1, which is the number chapter that the data contains. The average document length of this dataset is 5,612 words. In TestSet2, after processing, the data contain 23,342 documents (<SPEAKER *id*> level) which are the splitting 1,022 chapters (<CHAPTER *id*> level) from TestSet1. 22 out of 100 documents are in the same topic (<CHAPTER *id*> level). Table 1 summarizes the number of documents, sentences, words and the average word number of each document.

Table 1. Analytical Data of Corpus

| **Dataset** | **Size of corpus** | | | |
|---|---|---|---|---|
| | **Documents** | **Sentences** | **Words** | **Ave. words in document** |
| Training Set | 2,900 | 1,902,050 | 23,411,545 | 50 |
| TestSet1 | 1,022 | 80,000 | 5,735,464 | 6,612 |
| TestSet2 | 23,342 | 80,000 | 7,217,827 | 309 |

### *4.2. Experimental Setup*

In order to evaluate our proposed model, the following tools have been used.

[1] Available online at http://www.statmt.org/europarl/.

The probabilistic LMs are constructed on monolingual corpora by using the SRILM [15]. We use GIZA++ [16] to train the word alignment models for different pairs of languages of the Europarl corpus, and the phrase pairs that are consistent with the word alignment are extracted. For constructing the phrase-based statistical machine translation model, we use the open source Moses [17] toolkit, and the translation model is trained based on the log-linear model, as given in Eq. (2). The workflow of constructing the translation model is illustrated in Fig. 2 and it consists of the following main steps[2]:

(1) Preparation of aligned parallel corpus.

(2) Preprocessing of training data: tokenization, case conversion, and sentences filtering where sentences with length greater than fifty words are removed from the corpus in order to comply with the requirement of Moses.

(3) A 5-gram LM is trained on Spanish data with the SRILM toolkits.

(4) The phrased-based STM model is therefore trained on the prepared parallel corpus (English-Spanish) based on log-linear model of by using the nine-steps suggested in Moses.

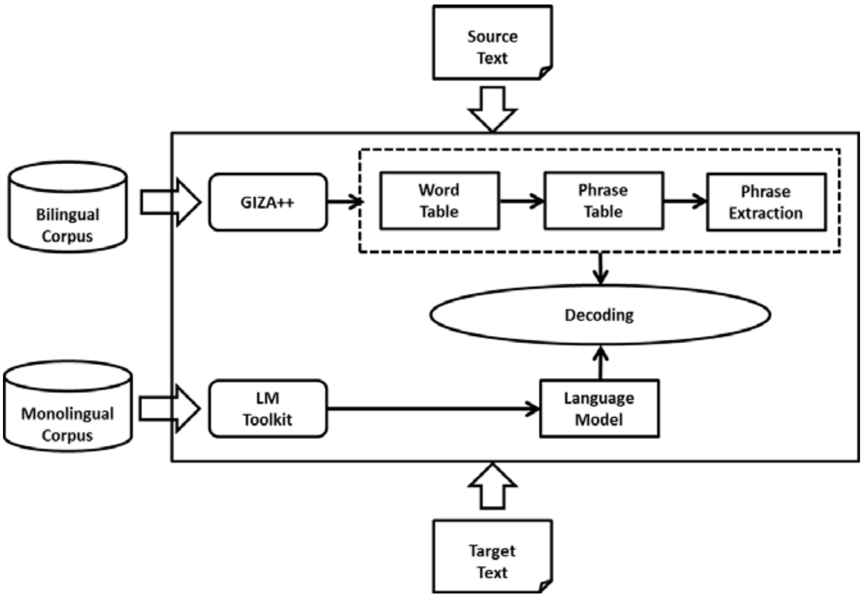

Figure 2. Main workflow of training phase

Once LM and TM have been obtained, we evaluate the proposed method with the following steps:

(1) The source documents are first translated into target language using the constructed translation model.

(2) The words candidates are computed and ranked based on a TF - IDF algorithm and the n-best words candidates then are selected to form the query based on Eq. (3) and (4).

[2] See http://www.statmt.org/wmt09/baseline.html for a detailed description of MOSES training options.

(3) All the target documents are stored and indexed using Apache Lucene[3] as our default search engine.

(4) In retrieval, target documents are scored and ranked by using the document retrieval model to return the list of most related documents with Eq. (5).

## 5. Results and Discussion

A number of experiments have been performed to investigate our proposed method on different settings. In order to evaluate the performance of the three independent models, we also conducted experiments to test them respectively before whole the CLIR experiment. The performance of the method is evaluated in terms of the average precision, that is, how often the target document is included within the first N-best candidate documents when retrieved.

Table 2. The average precision in Monolingual Environment

| **Retrieved Documents (*N*-Best)** | **Query Size (*Size*$_q$ in %)** | | | | | | |
|---|---|---|---|---|---|---|---|
| | **2** | **4** | **8** | **10** | **14** | **18** | **20** |
| 1 | 0.794 | 0.910 | 0.993 | 0.989 | 0.986 | 1.000 | 0.989 |
| 5 | 0.921 | 0.964 | 1.000 | 1.000 | 1.000 | 1.000 | 0.996 |
| 10 | 0.942 | 0.971 | 1.000 | 1.000 | 1.000 | 1.000 | 0.996 |
| 20 | 0.946 | 0.978 | 1.000 | 1.000 | 1.000 | 1.000 | 0.996 |

### *5.1. Monolingual Environment Information Retrieval*

In this experiment, we want to evaluate the performance of the proposed system to retrieve documents (monolingual environment) given the query. It supposes that the translations of source documents are available, and the step to obtain the translation for the input document can therefore be neglected. Under such assumptions, the CLIR problem can be treated as normal IR in monolingual environment. In conducting the experiment, we used all of the source documents of TestSet1. The steps are similar to that of the testing phase as described in Section 4.2, excluding the translation step. The empirical results based on different configurations are presented in Table 2, where the first column gives the number of documents returned against the number of words/terms used as the query.

The results show that the proposed method gives very high retrieval accuracy, with precision of 100%, when the top 18% of the words are used as the query. In case of taking the top 5 candidates of documents, the approach can always achieve a 100% of retrieval accuracy with query sizes between 8% and 18%. This fully illustrates the effectiveness of the retrieval model.

### *5.2. Translation Quality*

The overall retrieval performance of the system will be affected by the quality of translation. In order to have an idea the performance of the translation model we built, we employ the commonly used evaluation metric, BLEU, for such measure. The BLEU (Bilingual Evaluation Understudy) is a classical automatic evaluation method for the translation quality of an MT system [18]. In this evaluation, the translation model is created using the parallel corpus, as described in Section 4. We use another 5,000 sentences from the TestSet1 for

[3] Available at http://lucene.apache.org.

evaluation[4].

The BLEU value, we obtained, is **32.08**. The result is higher than that of the results reported by Koehn in his work [14], of which the BLEU score is **30.1** for the same language pair we used in Europarl corpora. Although we did not use exactly the same data for constructing the translation model, the value of **30.1** was presented as a baseline of the English-Spanish translation quality in Europarl corpora.

The BLEU score shows that our translation model performs very well, due to the large number of the training data we used and the pre-processing tasks we designed for cleaning the data. On the other hand, it reveals that the translation quality of our model is good.

### *5.3. Evaluation of CLIR Model*

In this section, the proposed CLIR model is compared against the approach proposed by Sánchez-Martínez et al. Table 3 presents the retrieval results given by his model. As illustrated, the best precision of the model can achieve up to 97% in precision, counting that the desired document is returned as the most relevant document among the candidates. In his method, both the probability of the translations and the relevance of the terms are taken into account in the retrieval model. The model is created based on IBM Model 1, Eq. (1), however, it still has a problem as we stated in Section 2.

Table 3. The average precision of Sánchez-Martínez et al.

| **Retrieved Documents (*N*-Best)** | **Query size (Num. of word in query)** | | | |
|---|---|---|---|---|
| | **1** | **2** | **5** | **10** |
| 1 | 0.32 | 0.51 | 0.84 | **0.97** |
| 2 | 0.43 | 0.63 | 0.90 | 0.98 |
| 5 | 0.51 | 0.73 | **0.95** | 0.99 |
| 10 | 0.55 | 0.77 | 0.97 | 1.00 |
| 20 | 0.56 | 0.80 | 0.98 | 1.00 |

Table 4. The retrieval results on TestSet1

| **Retrieved Documents (*N*-Best)** | **Query Size (*Size*$_q$ in %)** | | | | | | |
|---|---|---|---|---|---|---|---|
| | **1.0** | **1.4** | **1.8** | **2.0** | **3.0** | **6.0** | **10.0** |
| 1 | 0.90 | 0.93 | 0.95 | 0.97 | 0.99 | **1.00** | 0.99 |
| 5 | 0.98 | 0.98 | 0.99 | 0.99 | 0.99 | **1.00** | 0.99 |
| 10 | 0.98 | 0.98 | 0.99 | 0.99 | 1.00 | 1.00 | 1.00 |
| 20 | 0.98 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |

In order to obtain a higher retrieval precision, our model has been improved from different points. First, we only use individual words, instead of phrases, as well as numbers as query, which can alleviate the scarcity of tending to select long phrases that are less occurred in the training data. Secondly, our method can do better in dealing with the problem of term disambiguation because of the phrase-based SMT system, which takes a wider context of sentence in producing considers the translation. Last but not least, we did not use a fixed number of query words, instead portion of most relevant words is considered for different input of the document, Eq. (4). In other words, the longer the document, the more words will

[4] See http://www.statmt.org/wmt09/baseline.html for a detailed description of MOSES evaluation options.

be used for retrieval of the target documents. So the $Size_q$ is considered as a hidden variable in our document retrieval model.

What still needs to be explained is that the metrics in Table 3 and 4 are different. One experiment selected *static number* of words for a query, so all the queries have the same size; while the other one considers the *percentage* of the document length as its corresponding query size. Although it is hard to compare with their performances from corresponding columns, the improvements can be seen clearly when the desired document is among the first *N* (*N*=1, 2, 5, 10, 20) documents retrieved. Reviewing the experimental results presented in Tables 3 and 4, it shows that our model is able to give an improvement of 2% in precision and achieves 99% of success rate, in the case that the desired candidate is ranked in the first place. Moreover, the success rates achieved by our proposed model in different levels in all tests are above 90%.

As expected, the more the words we used to generate the query, the more the documents returned, and the higher the rate that the target document is retrieved within the candidates list.

However, the documents in TestSet1 are too large to align sentences from document level for further work, because a large document includes more sentences, which not only need more computational cost but also lead to higher error rate during sentence alignment. One way to solve this problem is to further split the large document and to retrieve it in a smaller document size. The problem in this case is that word overlap between a query and a wrong document is more probable when the document and the query are expressed in the same language. Furthermore, similar documents may include the same translation of words in the query, because the document retrieval model does not consider the weight of each word in the query which results in using more words to distinguish. This is the reason why different query size is used in Table 4 and 5, in order to guarantee the comparable retrieval performance on different types of documents.

Table 5. The retrieval results on TestSet2

| **Retrieved Documents (*N*-Best)** | **Query Size (*Size*$_q$ in %)** | | | | | | |
|---|---|---|---|---|---|---|---|
| | **10** | **15** | **20** | **25** | **30** | **35** | **40** |
| 1 | 0.884 | 0.936 | 0.964 | 0.972 | 0.983 | 0.987 | 0.990 |
| 5 | 0.944 | 0.970 | 0.984 | 0.989 | 0.992 | 0.993 | 0.995 |
| 10 | 0.955 | 0.977 | 0.987 | 0.991 | 0.993 | 0.994 | 0.996 |
| 20 | 0.966 | 0.984 | 0.991 | 0.992 | 0.994 | 0.994 | 0.997 |

As we stated in Section 4.1, TestSet2 is another concern. The results obtained are presented in Table 5. On average, the success rate is normally above 90% (in precision) by using a larger query size. It can even achieve 99.5% when the 5-best candidates are considered in the retrieval results. This result indicates that the reliable estimation of the profanities is more important than the plausibility of the probabilistic models. This fully illustrates the discrimination power of the proposed method.

## 6. Conclusion

This article presents a TQD statistical approach for CLIR which has been explored for both large and similar documents retrieval. Different from the traditional parallel corpora-based model which relies on IBM algorithm, we divided our CLIR model into three independent parts but all work together to deal with the term disambiguation, query generation and document retrieval. The performances showed that this method can do a good job of CLIR for not only large documents but also the similar documents.

The speed efficiency may be another big issue in our approach as some researchers have stated[2]. However, with the increasing of computing ability in hardware and software, there will be no difference in speed efficiency between query and document translation-based CLIR. Besides, our system only translates a certain amount of the source document to be retrieved instead of all the indexed target documents.

## Acknowledgement

This work was partially supported by the Research Committee of University of Macau under grant UL019B/09-Y3/EEE/LYP01/FST, and also supported by Science and Technology Development Fund of Macau under grant 057/2009/A2.

## References

[1] L. Ballesteros and W. B. Croft, “Statistical methods for cross-language information retrieval,” *Cross-language information retrieval*, pp. 23-40, 1998.

[2] K. Kishida, “Technical issues of cross-language information retrieval: a review,” *Information Processing & Management*, pp. 433-455, 41, 3 2005.

[3] D. W. Oard and A. R. Diekema, “Cross-language information retrieval,” *Annual review of Information science*, 33, pp. 223–256, 1998.

[4] M. Braschler and P. Schauble, “Experiments with the eurospider retrieval system for clef 2000,” *Cross-Language Information Retrieval and Evaluation*, pp. 140-148, 2001.

[5] M. Franz et al, “Ad hoc, cross-language and spoken document information retrieval at IBM,” NIST Special Publication: *The 8th Text Retrieval Conference,* TREC-8, 1999.

[6] M. W. Davis and W. C. Ogden, “Quilt: Implementing a large-scale cross-language text retrieval system,” *ACM SIGIR Forum*, pp. 92-98, 31, SI 1997.

[7] M. Federico and N. Bertoldi, “Statistical cross-language information retrieval using n-best query translations,” *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 167-174, 2002.

[8] F. Sanchez-Martinez and R. C. Carrasco, “Document translation retrieval based on statistical machine translation techniques,” *Applied Artificial Intelligence*, pp. 329-340, 25, 5 2011.

[9] P. F. Brown et al, “The mathematics of statistical machine translation: Parameter estimation,” *Computational linguistics*, pp. 263-311, 19, 2 1993.

[10] L. Ballesteros and W. B. Croft, “Phrasal translation and query expansion techniques for cross-language information retrieval,” *ACM SIGIR Forum*, pp. 84-91. 31, SI 1997.

[11] F. J. Och, H. Ney, “Discriminative Training and Maximum Entropy Models for Statistical Machine Translation,” In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 295–302, Philadelphia, PA, July

(2002)

[12] J. Ramos, "Using tf-idf to determine word relevance in document queries," *Proceedings of the First Instructional Conference on Machine Learning*, 2003.

[13] A. Berger et al, "Bridging the lexical chasm: statistical approaches to answer-finding," *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 192-199, 2000.

[14] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," *MT summit*, 5, 2005.

[15] A. Stolcke, "SRILM-an extensible language modeling toolkit," *Seventh International Conference on Spoken Language Processing*, 2002.

[16] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Computational linguistics*. pp. 19-51, 29, 1 2003.

[17] P. Koehn et al, "Moses: Open source toolkit for statistical machine translation," *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pp. 177-180, 2007.

[18] K. Papineni et al, "BLEU: a method for automatic evaluation of machine translation," *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 311-318, 2002.