Computational Linguistics and Chinese Language Processing
Vol. 17, No. 4, December 2012, pp.15-32

TQDL: Integrated Models for Cross-Language Document Retrieval

Long-Yue WANG*, Derek F. WONG*, and Lidia S. CHAO*

Abstract

This paper proposed an integrated approach for Cross-Language Information Retrieval (CLIR), which integrated with four statistical models: **T**ranslation model, **Q**uery generation model, **D**ocument retrieval model and **L**ength Filter model. Given a certain document in the source language, it will be translated into the target language of the statistical machine translation model. The query generation model then selects the most relevant words in the translated version of the document as a query. Instead of retrieving all the target documents with the query, the length-based model can help to filter out a large amount of irrelevant candidates according to their length information. Finally, the left documents in the target language are scored by the document searching model, which mainly computes the similarities between query and document.

Different from the traditional parallel corpora-based model which relies on IBM algorithm, we divided our CLIR model into four independent parts but all work together to deal with the term disambiguation, query generation and document retrieval. Besides, the TQDL method can efficiently solve the problem of translation ambiguity and query expansion for disambiguation, which are the big issues in Cross-Language Information Retrieval. Another contribution is the length filter, which are trained from a parallel corpus according to the ratio of length between two languages. This can not only improve the recall value due to filtering out lots of useless documents dynamically, but also increase the efficiency in a smaller search space. Therefore, the precision can be improved but not at the cost of recall.

* Natural Language Processing & Portuguese-Chinese Machine Translation Laboratory, Department of Computer and Information Science, University of Macau, Macau S. A. R., China
E-mail: vincentwang0229@hotmail.com
The author for correspondence is Long-Yue Wang.

In order to evaluate the retrieval performance of the proposed model on cross-languages document retrieval, a number of experiments have been conducted on different settings. Firstly, the Europarl corpus which is the collection of parallel texts in 11 languages from the proceedings of the European Parliament was used for evaluation. And we tested the models extensively to the case that: the lengths of texts are uneven and some of them may have similar contents under the same topic, because it is hard to be distinguished and make full use of the resources.

After comparing different strategies, the experimental results show a significant performance of the method. The precision is normally above 90% by using a larger query size. The length-based filter plays a very important role in improving the F-measure and optimizing efficiency.

This fully illustrates the discrimination power of the proposed method. It is of a great significance to both cross-language searching on the Internet and the parallel corpus producing for statistical machine translation systems. In the future work, the TQDL system will be evaluated for Chinese language, which is a big changing and more meaningful to CLIR.

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

1. Introduction

With the flourishing development of the Internet, the amount of information from a variety of domains is rising dramatically. Especially after the advent of the World Wide Web (WWW) in the 1900s, the amount of online information from the government, scientific and business communities has risen dramatically. Although much word has been done to develop effective and efficient retrieval systems for monolingual resources, the diversity and the explosive growth of information in different languages drove a great need for information retrieval that could cross language boundaries (Ballesteros *et al.*, 1988).

The issues of CLIR have been discussed for several decades. Its task addresses a situation in which a user tries to search a set of documents written in one language using a query in a different language (Kishida, 2005). It is of great significance, allowing people access information resources written in non-native languages and aligning documents 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 cases: one is that the word frequency is high in all the documents of a set, which is usually classified as stop word; the other one is that the frequency is moderate in several documents of a set. These words are poor in the ability of distinguishing documents. Thus, the query generation model should pick the words that occur more frequently in a certain document while less frequently in other documents. Finally, the document searching model evaluates the similarity between the query and each document. This model should give a higher score to the target document which covers the most relevant words in the given query. However, another problem 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. For example, Document *A* is larger and contains another smaller document *B*. So the retrieval system would be confused with a query including the information of *B*. In order to solve this problem, the length ratio of a language pair is considered. As the search space is reduced, both the speed efficiency and the recall value will be improved clearly.

There are two cases to be considered when we investigated the method. In one case, the lengths of documents are uneven, which are hard to balance the scores between large and small documents. 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

The issues of CLIR have been discussed from different perspectives for several decades. In this section, we briefly describe some related methods.

From a statistical perspective, the CLIR problem can be treated as document alignment. Given a set of parallel documents, the alignment that maximizes the probability over all

possible alignments is retrieved (Gale & Church, 1991) as follows:

$$\arg\max \Pr(A \mid D_s, D_t) \approx \arg\max_{A} \prod_{(L_s \Leftrightarrow L_t) \in A} \Pr(L_s \leftrightarrow L_t \mid L_s L_t) \tag{1}$$

where A is an alignment, D_s and D_t are the source and target documents, respectively L_1 and L_2 are the documents of two languages, $L_s \leftrightarrow L_t$ is an individual aligned pairs, an alignment A is a set consisting of $L_s \leftrightarrow L_t$ pairs.

On the matching strategies for CLIR, query translation is most widely used method due to its tractability (Gao *et al*., 2001). 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" (Oard & Diekema, 1998). 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 (Braschler & Schauble, 2001; Franz *et al*., 1999).

However, another problem introduced based on this approach is word (term) disambiguation, because a word may have multiple possible translations (Oard & Diekema, 1998). Significant efforts have been devoted to this problem. Davis and Ogden (1997) 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 (Federico & Bertoldi, 2002). 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*. (Sánchez-Martínez & Carrasco, 2011) 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 (Brown *et al*., 1993). 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

$$P(t \mid s) = \frac{\varepsilon}{(l+1)^m} \prod_{j=1}^{m} \sum_{i=0}^{l} P(t_j \mid s_i) \tag{2}$$

where t is a sequence of terms $t_1, \ldots, t_m$ in the target language, s is a sequence of terms $s_1, \ldots, s_l$ in the source language, $P(t_j|s_i)$ is the translation probability, and ε is a parameter ($\varepsilon = P(m|e)$),

where *e* is target language and *m* is the length of source language). Eq. (2) 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 (Oard & Diekema, 1998). However, in our method, the conflict can be resolved through contexts.

If translated sentences share cognates, then the character lengths of those cognates are correlated (Yang & Li, 2004). Brown *et al*. (1991) and Gale and Church (1991) have developed the models based on relationship between the lengths of sentences that are mutual translations. Although it has been suggested that length-based methods are language-independent (Gale & Church, 1991), they really rely on length correlations arising from the historical relationships of the languages being aligned.

The length-based model assumes that each term in L_s is responsible for generating some number of terms in L_t. This leads to a further approximation that encapsulates the dependence to a single parameter δ. $\delta(l_s, l_t)$ is function of l_s and l_t, which can be designed according to different language pairs. The length-based method is developed based on the following approximation to Eq. (3):

$$\Pr(L_s \leftrightarrow L_t \mid L_s, L_t) \approx \Pr(L_s \leftrightarrow L_t \mid \delta(l_s, l_t)) \tag{3}$$

3. Proposed Models

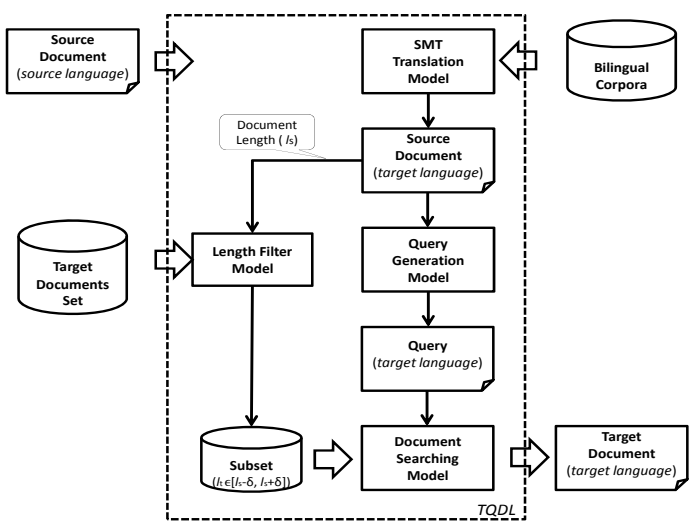

Figure 1. The proposed approach for CLIR

The approach relies on four 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; length filter model dynamically create a subset of candidates for retrieval according to the length information; 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.

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 (Ballesteros & Croft, 1997), 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 (Och & Ney, 2002), and is represented as:

$$p(e_1^I \mid f_1^J) = \frac{\exp(\sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J))}{\sum_{e'^I_1} \exp(\sum_{m=1}^{M} \lambda_m h_m(e'^I_1, f_1^J))} \tag{4}$$

where h_m indicates a set of different models, λ_m means the scaling factors, and the denominator can be ignored during the maximization process. The most important models in Eq. (4) 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 (Ramos, 2003). Given a document collection D, a word w, and an individual document $d \in D$, we calculate

$$P(w,d) = f(w,d) \times \log \frac{|D|}{f(w,D)} \tag{5}$$

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 (Berger *et al*., 2000).

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 is 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:

$$Size_q = [\lambda_{percent} \times Len_d] \tag{6}$$

$Size_q$ 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:

$$Score(q,d) = coord(q,d) \sum_{t_{in}q} tf(t,d) \times idf(t) \times bst \times norm(t,d) \tag{7}$$

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.

3.4 Length Filter Model

In order to obtain a suitable filter, we firstly analyzed the golden data[1] of ACL Workshop on SMT 2011, which includes Spanish, English, French, German and Czech 5 languages and 10 language pairs. English-Spanish language pair was used for analyzing and the data of the corpus are summarizes in Table 1.

Table 1. Analytical Data of Corpus of ACL Workshop on SMT 2011

Dataset	Size of corpus		
	No. of Sentences	***No. of Characters***	***Ave. No. Characters***
English	3,003	74,753	25
Spanish	3,003	79,426	26

Fig. 2 plots the distribution of word number in each aligned sentences. l_t is the length of English sentence while l_s is the length of sentence in Spanish. So the expectation is c= E (l_t/l_s) =1.0073, with the correlation R^2 = 0.9157. This shows that the data points are not substantially scatter in the plot and many data points are along with the regression line. Therefore, it is suitable to design a filter based on length ratio.

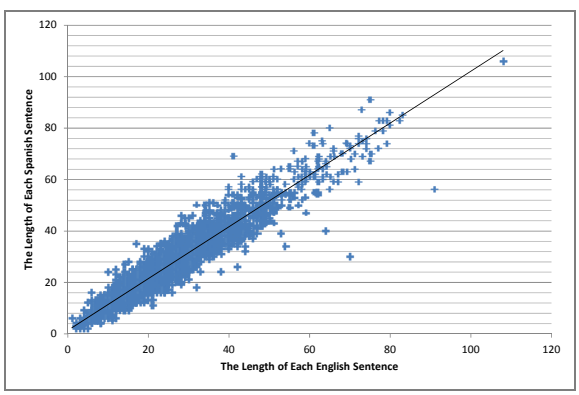

Figure 2. The length ratio of Spanish-English sentences.

To obtain an estimated length-threshold (δ) for filter model, the function δ (l_s, l_t) can be designed as follows:

$$\delta(l_s, l_t) = \frac{|l_t - l_s|}{l_s} \tag{8}$$

where l_s and l_t respectively stand for the length of a certain aligned sentence in the corpus we used. Finally, we got the average δ of around **0.15**. In implementation, we choose 4δ instead of δ to avoid some unnormal cases, where the right document would be discarded by the filter.

[1] It can be download from http://www.statmt.org/wmt11/

Filter F describes the relation between bilingual sentences based on the length ratio. Since western languages are similar in terms of word representation, the length ratio can be simply estimated as a 1:1. Given a certain document in source language, F can collect a subset for retrieval according to the average length ratio. So F is designed as follows:

$$F = \begin{cases} 1, length_t \in C \\ 0, length_t \notin C \end{cases}, C = [length_s - \delta, length_s + \delta] \tag{9}$$

where $length_s$ is the length of source document, and $length_t$ is the length of target document. δ is an average threshold obtained through Eq. (8), C is a confidence interval. If $length_t$ is included in C, F is 1, which has a chance to be retrieved, otherwise set as 0, which will be skipped during searching.

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[2] which is the collection of parallel texts in 11 languages from the proceedings of the European Parliament (Koehn, 2005). The corpus is commonly used for the construction and evaluation of statistical machine translation. 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.

Table 2. 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
TestSet	23,342	80,000	7,217,827	309

The datasets (training and test set) are collected 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. Besides, 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

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

in Table 2. The TestSet contains 23,342 documents, of which length is 309 in average. Actually 30% of documents are much more or less than the average number. Table 1 summarizes the number of documents, sentences, words and the average word number of each document.

4.2 Evaluation Metrics

The most frequent and basic evaluation metrics for information retrieval are precision and recall, which are defined as follows (Manning *et al*., 2008):

$$\mathrm{Pr}ecison = \frac{\#(relevant\quad items\quad retrieved)}{\#(retrieved\quad items)} \tag{10}$$

$$\mathrm{Re}call = \frac{\#(relevant\quad items\quad retrieved)}{\#(relevant\quad items)} \tag{11}$$

For reporting the evaluation of our method, we used the *F*1 measure, the recall and the precision values. *F*1-measure (*F*) is formulated by Van Rijsbergen as a combination of recall (*R*) and precision (*P*) with an equal weight in the following form:

$$F = \frac{2PR}{P+R} \tag{12}$$

4.3 Experimental Setup

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

The probabilistic LMs are constructed on monolingual corpora by using the SRILM (Stolcke *et al*., 2002). We use GIZA++ (Och & Ney, 2003) 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 (Koehn *et al*., 2007) toolkit, and the translation model is trained based on the log-linear model, as given in Eq. (4). The workflow of constructing the translation model is illustrated in Fig. 3 and it consists of the following main steps[3]:

(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.

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

(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.

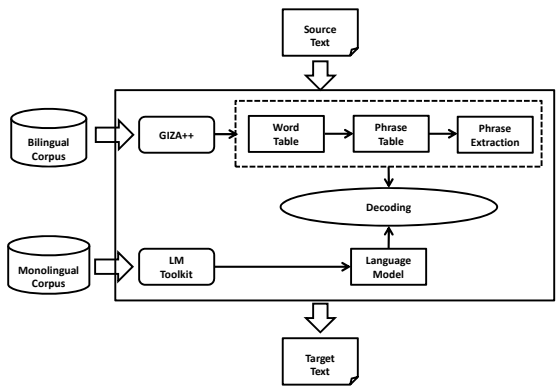

Figure 3. Main workflow of training phase

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. (5) and (6).

(3) All the target documents are stored and indexed using Apache Lucene[4] 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. (7).

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 firstly conducted experiments to test them respectively before whole the TQDL platform. 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.

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

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 TestSet. 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 3, where the first column gives the number of documents returned against the number of words/terms used as the query.

Table 3. 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

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 (Papineni *et al*., 2002). 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 evaluation[5].

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

The BLEU value, we obtained, is **32.08**. The result is higher than that of the results reported by Koehn in his work (Koehn, 2005), 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 TQDL without Filter for CLIR

In this section, the proposed model without length filter model is tested. Table 4 presents the F-measure given by TQDL system without length filter model. As illustrated, the it can only achieve up to 94.7%, counting that the desired document is returned as the most relevant document among the candidates. Although it has achieved a very good performance in the experiments, the 6.6% of documents have been discarded in the pre-processing.

Table 4. The F-measure of our system without length filter model

Retrieved Documents (*N*-Best)	**Query Size ($Size_q$ in %)**				
	2.0	**4.0**	**6.0**	**8.0**	**10.0**
1	0.905	0.943	0.942	***0.947***	0.941
2	0.922	0.949	0.949	0.953	0.950
5	0.932	0.950	0.953	0.963	0.960
10	0.936	0.954	0.960	0.968	0.971
20	0.941	0.958	0.974	0.979	0.981

To investigate the changes of the performance with removing abnormal documents (too lager or too small), query size $Size_q$ was set as a constant value (8.0%)**,** which can achieve the best precision as shown in Table 4. We believed that the abnormal document is the main obstacle to develop the performance of the system. Therefore, we removed the documents, of which length are out of a certain threshold.

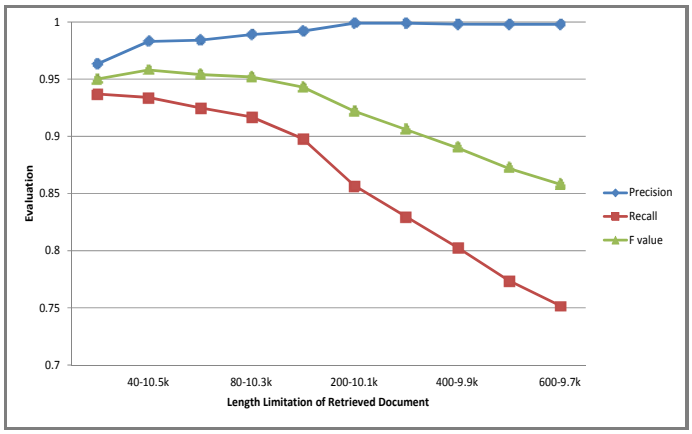

Figure 4. The changes of evaluation when removing data

Fig. 4 plots the variations of *P*, *R* and *F* with the length scope increasing. As we expected, the precision increase when the more abnormal documents are discarded from the dataset. However, the recall declines sharply, which also lead to the falling of *F*-measure. When the precision is closed to **100%**, nearly **15%** documents are removed from the dataset. So the high precision is often at the cost of reducing the recall rate. *F*-measure is only 95% at its top, so it is hard to improve the performance of CLIR using traditional methods.

5.4 TQDL with Filter for CLIR

In order to obtain a higher retrieval rate, our model has been improved from different points. Firstly, we generate the query with dynamic size, which can do better in dealing with the problem of similar documents both in length and content. In another words, the longer the document, the more words will be used for retrieval of the target documents. So the $Size_q$ is considered as a hidden variable in our document retrieval model. Besides, all the indexed documents can be filtered with *F* formula in Eq. (9), and it can alleviate the scarcity of tending to select longer documents when occurring the word overlap between shorter and longer documents, because a certain source document are only searched in a subset defined by its length. It can improve the precision without discard any so-called “abnormal” documents from dataset, so the *P*, *R* and *F* values will always be the same. Table 5 presents the *F* values given by TQDL with length filter model.

Table 5. The F-measure of our system with length filter model

Retrieved Documents (*N*-Best)	Query Size ($Size_q$ in %)				
	2.0	**4.0**	**6.0**	**8.0**	**10.0**
1	0.958	0.975	0.983	0.990	0.992
2	0.967	0.979	0.986	0.993	0.996
5	0.971	0.982	0.987	0.993	0.996
10	0.974	0.983	0.988	0.995	0.996
20	0.974	0.983	0.990	0.995	0.996

Compared with the results presented in Tables 4 and 5, it shows that the length filter model is able to give a high improvement by 4.5% in F-measure and achieve more than 99% of successful rate, in the case that the desired candidate is ranked in the first place. Above all, there is no documents waste in the dataset.

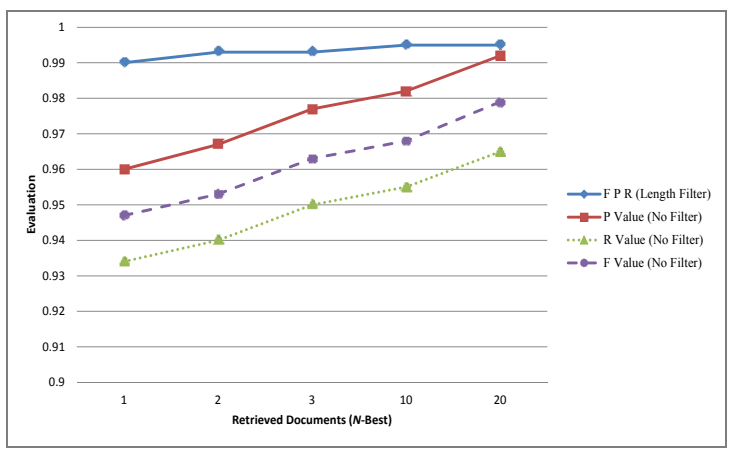

Figure 5. The changes of evaluation with N-Best

Fig. 5 presents an ideal distribution of evaluation, of which *P* and *R* should be closed to the *F* line. In this comparison, query size $Size_q$ was still set as a constant value (8.0%). With the increasing of *N*, evaluations without filter are in a low level, while the one with this filter can achieve a good and stable performance. Finally, the precision and recall values are closed to F measure, which can all keep in a high level (99%-100%).

6. Conclusion

This article presents a TQDL 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 four 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. This fully illustrates the discrimination power of the proposed method. It is of a great significance to both cross-language searching on the Internet and the parallel corpus producing for statistical machine translation systems. In the future work, the TQDL system will be evaluated for Chinese language, which is a big changing and more meaningful to CLIR. In the further work, we plan to make better use of the proposed models between significantly different languages such as Portuguese-Chinese.

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

Ballesteros, L., & Croft, W. B. (1988). Statistical methods for cross-language information retrieval. *Cross-language information retrieval*, 23-40.

Ballesteros, L. & Croft, W. B. (1997). Phrasal translation and query expansion techniques for cross-language information retrieval. *ACM SIGIR Forum*, 31(SI), 84-91.

Braschler, M., & Schauble, P. (2001). Experiments with the eurospider retrieval system for clef 2000. *Cross-Language Information Retrieval and Evaluation*, 140-148.

Brown, P. F., Lai, J. C. & Mercer, R. L. (1991). Aligning sentences in parallel corpora. In *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, 169-176.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D. & Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2), 263-311. MIT Press.

Berger, A., Caruana, R., Cohn, D., Freitag, D. & Mittal, V. (2000). Bridging the lexical chasm: statistical approaches to answer-finding. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, 192-199.

Davis, M. W. & Ogden, W. C. (1997). Quilt: Implementing a large-scale cross-language text retrieval system. *ACM SIGIR Forum*, 31(SI), 92-98.

Federico, M. & Bertoldi, N. (2002). Statistical cross-language information retrieval using n-best query translations. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, 167-174.

Franz, M., McCarley, J. S. & Ward, R. T. (1999). Ad hoc, cross-language and spoken document information retrieval at IBM. NIST Special Publication: *The 8th Text Retrieval Conference* (*TREC-8*).

Gale, W. A. & Church, K. W. (1991). Identifying word correspondences in parallel texts. In *Proceedings of the workshop on Speech and Natural Language*, 152-157.

Gao, J., Nie, J. Y., Xun, E., Zhang, J., Zhou, M., & Huang, C. (2001). Improving query translation for cross-language information retrieval using statistical models.In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, 96-104.

Kishida, K. (2005). Technical issues of cross-language information retrieval: a review. *Information processing & management*, 41(3), 433-455.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. *MT summit*, 5.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., *et al*. (2007). Moses: Open source toolkit for statistical machine translation. *Annual meeting-association for computational linguistics*, 45(2), 2.

Manning, C. D., Raghavan, P. & Schütze, H. (2008). Introduction to information retrieval (Vol. 1). *Cambridge University Press Cambridge*, 140-159.

Oard, D. W., & Diekema, A. R. (1998). Cross-language information retrieval. *Annual review of Information science*, 33, 223-256.

Och, F. J. & Ney, H. (2002). Discriminative training and maximum entropy models for statistical machine translation.In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 295-302.

Och, F. J. & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1), 19-51. MIT Press.

Papineni, K., Roukos, S., Ward, T. & Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, 311-318.

Ramos, J. (2003). Using tf-idf to determine word relevance in document queries. In *Proceedings of the First Instructional Conference on Machine Learning*.

Sánchez-Martínez, F. & Carrasco, R. C. (2011). Document translation retrieval based on statistical machine translation techniques. *Applied Artificial Intelligence*, 25(5), 329-340.

Stolcke, A. & others. (2002). SRILM-an extensible language modeling toolkit. In *Proceedings of the international conference on spoken language processing*, 2, 901-904.

Yang, C. C. & Wing Li, K. (2004). Building parallel corpora by automatic title alignment using length-based and text-based approaches. *Information processing & management*, 40(6), 939-955.