J Inf Process Syst, Vol.13, No.4, pp.805~817, August 2017
https://doi.org/10.3745/JIPS.04.0039

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

A Method of Chinese and Thai Cross-Lingual Query Expansion Based on Comparable Corpus

Peili Tang*, Jing Zhao**, Zhengtao Yu*, Zhuo Wang*, and Yantuan Xian*

Abstract

Cross-lingual query expansion is usually based on the relationship among monolingual words. Bilingual comparable corpus contains relationships among bilingual words. Therefore, this paper proposes a method based on these relationships to conduct query expansion. First, the word vectors which characterize the bilingual words are trained using Chinese and Thai bilingual comparable corpus. Then, the correlation between Chinese query words and Thai words are computed based on these word vectors, followed with selecting the Thai candidate expansion terms via the correlative value. Then, multi-group Thai query expansion sentences are built by the Thai candidate expansion words based on Chinese query sentence. Finally, we can get the optimal sentence using the Chinese and Thai query expansion method, and perform the Thai query expansion. Experiment results show that the cross-lingual query expansion method we proposed can effectively improve the accuracy of Chinese and Thai cross-language information retrieval.

Keywords

Comparable Corpus, Cross-Language Query Expansion, Cross-Language Information Retrieval, Words Relationship

1. Introduction

Since the establishment of diplomatic relations between China and Thailand, the relationship between two countries have developed rapidly in various aspects—political, diplomatic, economics, culture, and other exchange corporations. However, in view of the fact that Chinese and Thai belong to two different languages, Internet information sharing and communication is a serious problem due to the language barriers. The application of cross-language information retrieval (CLIR) technology provides an effective way to solve this problem [1-3]. The task of CLIR attempts to bridge the mismatch between the source and target languages using the approaches such as query and the document translation. Some researchers have also done some research on cross-language machine learning [4,5]. CLIR allows a user to retrieve the documents in language A with a query in language B [6].

As we all know, cross-language query expansion (CLQE) is an efficient way to enhance the precision and recall rate of CLIR system. The goal of query expansion in this regard is by increasing recall,

Manuscript received December 17, 2015; first revision February 24, 2017; accepted March 26, 2017.

Corresponding Author: Zhengtao Yu (ztyu@hotmail.com)

* Intelligent Information Processing Laboratory of Yunnan Province, Key Laboratory in Regional University of Yunnan Province, Kunming University of Science and Technology, Yunnan, China (778304073@qq.com, ztyu@hotmail.com, {393572445, 195426286}@qq.com)

** National University of Defense Technology, Hunan, China (407271455@qq.com)

precision can potentially increase (rather than decrease as mathematically equated), by including in the result set pages which are more relevant [7]. In the query expansion, related words are added to user's original query for the purpose of forming a longer and more precise query to express user's retrieval intentions. Moreover, the key step of the query expansion is how to find the words related to original query. Different approach has different sources for finding related words [8]. Typical sources of query expansion terms are pseudo relevant documents [4]or external static resources, such as click through data [9,10], Wikipedia [11,12]or ConceptNet [8,13] WordNet [14,15], HowNet [7]. For example, Dalton et al. [5] did query expansion using entity names, aliases and categories with several methods of linking entities to the query. Xiong and Callan [16] presented a simple and effective method of using one such knowledge base, Freebase, to improve query expansion, a classic and widely studied information retrieval task. Boer de Boer et al. [17] did query expansion using common knowledge bases ConceptNet and Wikipedia is compared to an expert description of the topic applied to content-based information retrieval of complex events. Some researchers obtain query expansion words from the pseudo relevance feedback documents, a classical model based on pseudo-relevant documents was proposed by Rocchio for the SMART retrieval system [18]. Sordoni et al. [19] target at learning semantic representations of single terms and bigrams as a way to encode valuable semantic relationships for expanding a user query. Corpus dependent methods generally employ statistical information extracted from the corpus [20]. Techniques such as stemming, clustering, and term co-occurrence are used for obtaining the query context from the document collection [21]. In a recent study, Roy et al. [22] proposed a framework for Automatic Query Expansion (AQE) by using distributed neural language model word2vec. Vaidyanathan et al. [23] obtained expansion terms by combining pseudo relevance feedback and equi-frequency partition with TF-IDF scoring technique to query expansion. Singh and Sharan [24] did query expansion by combining query term co-occurrence and query terms contextual information based on corpus of top retrieved feedback documents in first pass.

However, there is no corresponding Chinese and Thai knowledge base, and no a large number of Chinese and Thai bilingual parallel corpus either. Therefore, it is difficult to study Chinese and Thai cross language query expansion.

Bilingual comparable corpus are the collections of text documents in different languages that are about the same or similar topics. For example, the news published in the same time period tend to report the same important international events in various topics. Chinese and Thai comparable corpus (C-T CC) is more accessible than Chinese and Thai parallel corpus. Thus, this paper proposed a novel method of cross-lingual query expansion based on Chinese and Thai comparable corpus. The relationship between Chinese words and Thai words is extracted by taking advantage of the characteristics of words co-occurrence. Then, the Thai candidate query words which correspond to the Chinese query words are calculated. Finally, Thai query expansion sentence can be obtained by the use of the Chinese and Thai cross-lingual query expansion method.

2. The Method of Cross-Lingual Query Expansion Based on Chinese and Thai Comparable Corpus

In this section, we present the details of our proposed approach conducting CLQE when the available resource is comparable corpora. Fig. 1 shows a sketch of our approach. The idea is to use the word correlations, the correlations of frequency distributions of two terms, which are mined from comparable corpora for constructing the word translation probabilities between word pairs. Having a query in one language, we can select the query expansion words using these estimated probabilities and obtain the Thai query candidate words. Then, the optimal combination of Thai words can be selected with the novel method proposed, which uses the correlations between bilingual words and monolingual words. Finally, Thai query expansion sentence can be obtained.

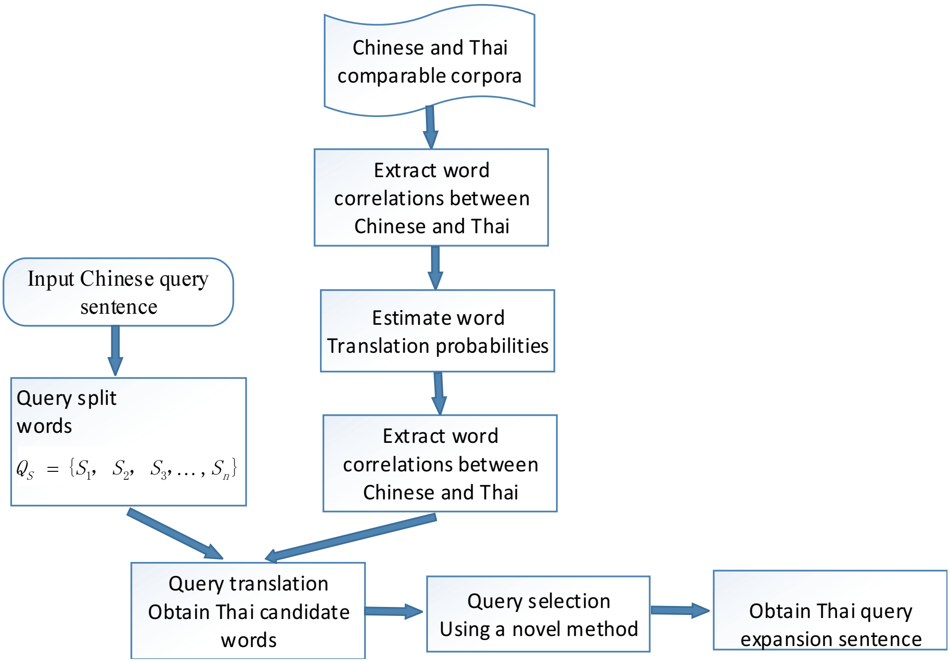

Chinese and Thai comparable corpora
Extract word correlations between Chinese and Thai
Estimate word Translation probabilities
Extract word correlations between Chinese and Thai
Input Chinese query sentence
Query split words
$Q_S = \{S_1, S_2, S_3, \ldots, S_n\}$
Query translation Obtain Thai candidate words
Query selection Using a novel method
Obtain Thai query expansion sentence

Fig. 1. The cross-lingual query expansion steps.

2.1 Comparable Corpus Construction and Word Vector Generation

The primary issue to build a comparable corpus is the alignment of the text. Wikipedia, from the beginning, is devoted to the construction of multilingual encyclopedia, and thus its entries are natural text alignment. Wikipedia has been used as a source of multilingual data for a range of monolingual and cross-language NLP and IR tasks, such as named entity, recognition, query translation for CLIR, word-sense disambiguation and statistical machine translation. Wikipedia editions have been created in more than 287 languages, with some languages more evolved than others. A proportion of topics appear in multiple Wikipedia (in multiple languages), resembling a comparable corpus.

As long as there are the corresponding versions of other languages, Wikipedia will have a corresponding link on the page. Such structure provides a great convenience for extracting bilingual comparable corpus. The comparable corpora could be constructed by using Wikipedia specific cross-lingual links. We can use the HtmlUnit technology to obtain the content on Chinese web page of each entry. At the same time, the corresponding Thai page links can be accessed and the content on Thai page can be obtained giving the Chinese and Thai alignment text documents (on the same topic).

Having Chinese and Thai aligned comparable corpus, in order to calculate the relationship between

Chinese words and Thai words, comparable corpus words should be trained into words vector. The generation of word vector is a statistical term frequency of each word which appears in Chinese and Thai alignment text. For example, let *a* be a word in Chinese and *b* be a word in Thai. The normalized frequency vectors for *a* and *b* will be $\vec{a} = (a_1, a_2, a_3, \cdots, a_n)$ and $\vec{b} = (b_1, b_2, b_3, \cdots, b_n)$ respectively, where a_i means the frequency of Chinese words which appears in the *i*th Chinese document, where b_i means the frequency of Thai words which appears in the *i*th Thailand document.

2.2 Thai Candidate Query Word Acquisition

In order to obtain the query extension word of the target language (i.e., Thai), Thai candidate words according to Chinese query words need to be found. This process can be transformed into computing the correlation problems of the Chinese words and Thai words. Rahimi et al. [25,26] used Pearson correlation coefficient to quantify the correlation between bilingual words. This method is also applied in this article. Let $\vec{a} = (a_1, a_2, a_3, \cdots, a_n)$ be a vector of Chinese word *a*, and let $\vec{b} = (b_1, b_2, b_3, \cdots, b_n)$ be a vector of Thai word *b*. Word vector is normalized by formula (1).

$$a_i = \frac{a_i}{\sum_{j=0}^{n} a_j} \qquad b_i = \frac{b_i}{\sum_{j=0}^{n} b_j} \tag{1}$$

where $\sum_{j=0}^{n} a_j$ is the total count of word *a* in Chinese documents set, and $\sum_{j=0}^{n} b_j$ is the total count of word *b* in Thailand documents set. The similarity of these two words is computed with the Pearson correlation coefficient:

$$r(a,b) = \frac{\sum_{i=1}^{n} a_i b_i - \frac{1}{n}\sum_{i=1}^{n} a_i \sum_{i=1}^{n} b_i}{\sqrt{\left(\sum_{i=1}^{n} a_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} a_i\right)^2\right)\left(\sum_{i=1}^{n} b_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} b_i\right)^2\right)}} \tag{2}$$

Having the correlations between words in different languages, in the next step we estimate the translation probabilities. A natural baseline method is to use the normalized correlation scores as translation probabilities. Formally, let *w* be a word in Chinese and $u_1, u_2, u_3, \cdots u_k$ be the top *k* correlated Thai words with correlation scores $r_1, r_2, r_3, \cdots r_k$, respectively, $r_i = (w, u_i)$, $1 < i < k$, the correlation scores of word *w* and word u_i, We construct the probabilities by normalizing these raw correlation scores:

$$p(u_i|\mathrm{w}) = \frac{r_i}{\sum_{j=1}^{m} r_j} \tag{3}$$

where $p(u_i \mid w)$ is the probability of u_i being the translation of Thai word *w*. Table 1 shows a sample set of top Chinese-Thai word pairs extracted from the Chinese and Thai comparable corpus in our experiments.

Table 1. Sample of extracted Chinese-Thai word pairs

Chinese word (*w*)	**Thailand words (*u*$_i$)**	$p(u_i \mid w)$
泰拳 (Muay Thai)	มวยไทย	0.3211
	ลายมวยไทย	0.1743
签证 (visa)	วีซ่า	0.3514
	การขอวีซ่า	0.2907
盟国 (ally)	พันธมิตร	0.2941
	กับฝ่ายพันธมิตร	0.2714
比赛 (match)	แข่งขัน	0.3122
	การแข่งขัน	0.1541

2.3 The Relationship between Monolingual Words

Query is not a simple word, but usually a sentence which is combined by several terms. However, the query terms are related. When it is extended to the target language, the relationship between target query expansion words also has the relevance. Therefore, we need to analyze and calculate the relationship between monolingual languages.

To compute the association score between monolingual words, we employ Jansen–Shannon divergence, also known as total divergence to the mean (TDM), which has been shown to be a useful measure for the term association in other applications [27,28]. TDM measures the symmetrized Kullback–Leibler divergence to the mean of the two vectors.

$$\alpha\left(\overrightarrow{w_1},\overrightarrow{w_2}\right)=2\log 2+\sum\nolimits_{y\in both}\left\{\overrightarrow{w_1(y)}\log\frac{\overrightarrow{w_1(y)}}{\overrightarrow{w_1(y)}+\overrightarrow{w_2(y)}}+\overrightarrow{w_2(y)}\log\frac{\overrightarrow{w_2(y)}}{\overrightarrow{w_1(y)}+\overrightarrow{w_2(y)}}\right\} \tag{4}$$

where $y \in both$ means that both *y*th element values of vectors $\overrightarrow{w_1}$ and $\overrightarrow{w_2}$ are not 0. Note that the sum of element values of each vector must be 1 in this equation. Vector *w*$_i$ is defined as follows:

$$\overrightarrow{w_i}=\left(wt_{i1},wt_{i2},\cdots,wt_{in}\right) \tag{5}$$

$$wt_{ij}=\frac{p\left(w_i \mid d_j\right)}{\sum\limits_{k=1}^{n}p\left(w_i \mid d_k\right)}\quad p\left(w_i \mid d_j\right)=\frac{tf_{ij}}{dl_j} \tag{6}$$

where vector $\overrightarrow{w_i}$ is the term-distribution vector for the *i*th term *w*$_i$, *wt*$_{ij}$ is the weight of $\overrightarrow{w_i}$ in the *j*th document, *n* is the number of documents in some collection that is used to calculate the term distribution, $p\left(w_i \mid d_k\right)$ is the probability that the term *w*$_i$ occurs in the *j*th document *d*$_j$; *tf*$_{ij}$ is the term frequency of *w*$_i$ in *d*$_j$, and *dl*$_j$ is the length of document *d*$_j$. $\sum\nolimits_{k=1}^{n}p\left(w_i \mid d_k\right)$ is the normalization factor required to ensure that $\sum\nolimits_{k=1}^{n}wt_{ij}=1$.

Suppose that the *j*+1st term s_{j+1} of source query Q_s has *m* translations $\{t_{j+1,1}, t_{j+1,2}, \cdots, t_{j+1,m}\}$, and $c_{ij} = t_{j1}$ and $c_{ij+1} = t_{j1+1,1}$; c_{ij} and c_{ij+1} are the *j*th and *j*+1st terms of candidate target-language query c_i, respectively. In this case, the transition probabilities, $p(c_{ij+1} \mid c_{ij})$, are estimated by normalizing the association scores as follows:

$$p(c_{ij+1} = t_{j+1,1} \mid c_{ij} = t_{j1}) = \frac{\alpha(t_{j1}, t_{j+1,1})}{\sum_{k}^{m} \alpha(t_{j1}, t_{j+1,k})} \tag{7}$$

where $\alpha(t_{j1}, t_{j+1,1})$ is the association score between the translations t_{j1} and $t_{j+1,1}$.

2.4 The Model of Thai Query Expansion

The Thai query expansion needs to select the Thai candidate words, since the Chinese query sentence contains multiple query terms which correspond to multi-group Thai candidate words. In order to obtain the optimal query expansion sentence, Thai Candidate words are combined and selected with the model of Thai query expansion.

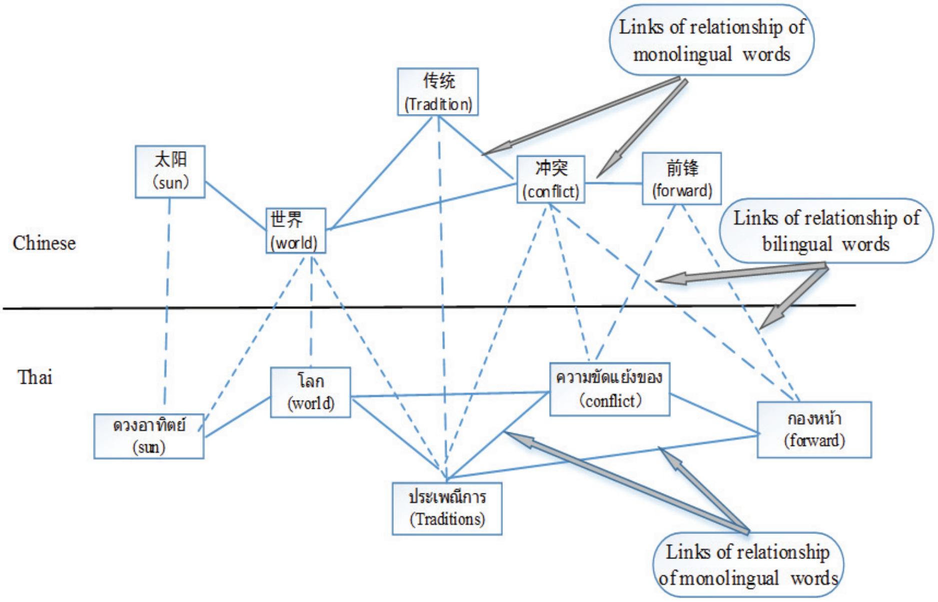


Fig. 2. The association of monolingual and bilingual words.

In the alignment of the comparable corpus, we use the term co-occurrence relationship to extract the relationship between each word. Fig. 2 shows the association of monolingual and bilingual words base on the term co-occurrence relationship. We only consider those words in Thai that are highly correlated to the query words in Chinese as the candidate words in the translated query and use the translation probabilities to construct the query expansion model.

Firstly, we use Chinese Lexical Analysis tool (ICTCLAS) for Chinese query sentence segmentation [29]. After getting the Chinese query terms $Q_s = \{s_1, s_2, s_3, \cdots, s_n\}$, the correlation of vectors between Chinese words and Thai words can be quantified with the Pearson correlation coefficient. Each Chinese word corresponds to top k high similarity Thai candidate words. According to Chinese query, the multiple Thai query expansion sentences could be constructed. Kim et al. [30] proposed a novel query expansion method that generates all the candidate target-language queries for the source-language query Q_s, assigns a score to each, ranks the candidate target queries by that score, and then chooses the candidate target-language query with the top score.

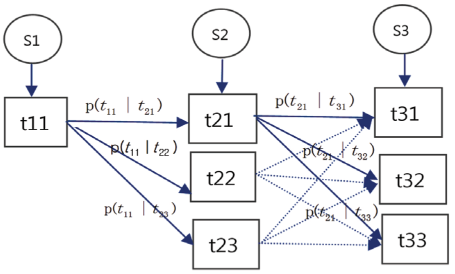
$$\phi(c_i) = p(t_{ki} \mid s_k) \prod_{j=1}^{n-1} p(t_{ij+1} \mid t_{ij}) p(t_{ij+1} \mid s_{j+1}) \tag{8}$$

$$Q_t = \arg\max \phi(c_i) \tag{9}$$

where s_k is the kth term of the source-language (Chinese) query Q_s, $\phi(c_i)$ is the score of the ith candidate target-language (Thai) query c_i. Q_t is the optimal target-language query. In Fig. 2, there are 9 candidate Thai queries from c_1 to c_9. c_{ij} is the jth term of the candidate target query c_i. Possible translations of each query term are combined for generating Thai Candidate queries. Fig. 3 illustrates the calculation process of the scores $\phi(c_i)$ for some candidate Thai queries. We assume that there are three Chinese query terms $Q_s = \{s_1, s_2, s_3\}$. s_1 could be translated into Thai word t_{11}. s_2, s_3 could be translated into Thai word $\{t_{21}, t_{22}, t_{23}\}, \{t_{31}, t_{32}, t_{33}\}$. There are 9 candidate target-language queries denoted by

$$c_1 = \{t_{11}, t_{21}, t_{31}\}, c_2 = \{t_{11}, t_{21}, t_{32}\}, c_3 = \{t_{11}, t_{21}, t_{33}\}, c_4 = \{t_{11}, t_{22}, t_{31}\}, c_5 = \{t_{11}, t_{22}, t_{32}\}, c_6 = \{t_{11}, t_{22}, t_{33}\},$$
$$c_7 = \{t_{11}, t_{23}, t_{31}\}, c_8 = \{t_{11}, t_{23}, t_{32}\} c_9 = \{t_{11}, t_{23}, t_{33}\}$$

and their scores are calculated by Eq. (8).



$$\phi(c_1) = p(t_{11}|s_1)p(t_{21}|t_{11})p(t_{21}|s_2)p(t_{31}|t_{21})p(t_{31}|s_3)$$
$$\phi(c_2) = p(t_{11}|s_1)p(t_{21}|t_{11})p(t_{21}|s_2)p(t_{32}|t_{21})p(t_{32}|s_3)$$
$$\phi(c_3) = p(t_{11}|s_1)p(t_{21}|t_{11})p(t_{21}|s_2)p(t_{33}|t_{21})p(t_{33}|s_3)$$
$$\phi(c_4) = p(t_{11}|s_1)p(t_{22}|t_{11})p(t_{22}|s_2)p(t_{31}|t_{21})p(t_{31}|s_3)$$
$$\cdots\cdots$$
$$\phi(c_9) = p(t_{11}|s_1)p(t_{23}|t_{11})p(t_{23}|s_2)p(t_{33}|t_{21})p(t_{33}|s_3)$$

Fig. 3. Example of computing process of the scores $\phi(c_i)$ for some Thai candidate queries.

We not only consider the correlation between words in different languages, but also the relationship in monolingual words. Then, the more precise query expansion can be achieved.

3. Experiments and Result

Aligned text of the same entry can be got from the Wikipedia page for building Chinese and Thai comparable corpora, and it will be regarded as an experimental training data. We also get a total of 154,300 articles from the Thai government official website, the Thai national newspaper, Thailand Daily News and other Thailand websites, These articles act as the test document set, ranging from political, economic, news and other aspects. The document set will be used for information retrieval.

3.1 The Experimental Steps

(1) The pretreatment aligned text of Chinese and Thai comparable corpus, such as segmentation and remove stop words, are trained into Chinese and Thai frequency word vectors.

(2) The pretreatment Chinese query sentences (segmentation and remove stop words) are split into many words. We use top *k* correlated words as the translations of each query word.

(3) Top-*k* correlated Thai words are used for constructing multi-group Thai query expansion sentences, Formula (8) and (9) are used to choose the optimal Thai query expansion sentences, and this expansion sentence query will be applied to CLIR.

3.2 Evaluation Measures

We report two evaluation measures: mean average precision (MAP) and recall at cutoff rank *k*(*R@k*). Recall defines the number of relevant documents that are retrieved in relation to the total.

number of relevant documents. Recall at a specified cutoff rank *k* is defined by only considering the top-*k* results. MAP is another standard measure used in IR that is also sensitive to the rank of relevant documents. It averages the precision measured at the rank of each relevant document [31]. MAP is defined as follows:

$$\mathrm{MAP}=\frac{1}{|Q|}\sum_{q\in Q}\frac{\sum_{k=1}^{n}P@k\mathrm{relevant}\left(\mathrm{d}_{\mathrm{q,k}},q\right)}{\left|\left\{d\in D\middle|relevant(d,q)=1\right\}\right|} \tag{10}$$

where $d_{q,k}$ is the retrieved document given query *q* at rank *k* and *relevant*(*d*, *q*) is the binary relevance function of document *d* given query *q*. $p@k$ means precision at cutoff rank *k*.

3.3 Evaluation Result

Lucene is currently a very popular, free Java information search (IR) library. It is a full text search engine architecture. Complete query engine, index engine and parts of the text analysis engine are provided. In this paper, we regard Lucene search technology as the basic platform [32]. Lucene inverted index of the Thai document which get from Thailand website is set up, and we can obtain the Thai index file. There are three comparative experiment methods.

(1) Chinese and Thai CLQE based on comparable corpus (CC CLQE). The pretreatment Chinese query words using the word vectors training by Chinese and Thai comparable corpus calculated a top-*k* higher correlated Thai query candidate words.

(2) Chinese and Thailand CLQE based on bilingual dictionary and comparable corpus. We combine comparable corpus with the dictionary in two different ways. In the first approach called Dic-CC CLQE, the query words are first translated with dictionary. The terms that are not found in dictionary are then translated with comparable corpus. In the second approach. Dic&CC CLQE, queries are translated with both dictionary and comparable corpus.

We use three different translation methods to select *k*(1–10) Thai candidates words; Formulas (8) and (9) are used for selecting the optimal Thai query expansion phrase. Finally, the Lucene search method is used for searching the Thai indexed document collection in the information retrieval system, and the documents which associated with the Chinese query sentence can be returned. Fig. 4 shows the mean average precision of CLIR based on different number of translation words *k*(1–10) for each Chinese query word. As illustrated in Fig. 4, the Dic&CC CLQE methods which combine two translation resources as described, appear to be our best performing methods, since its highest MAP value can reach 0.307.

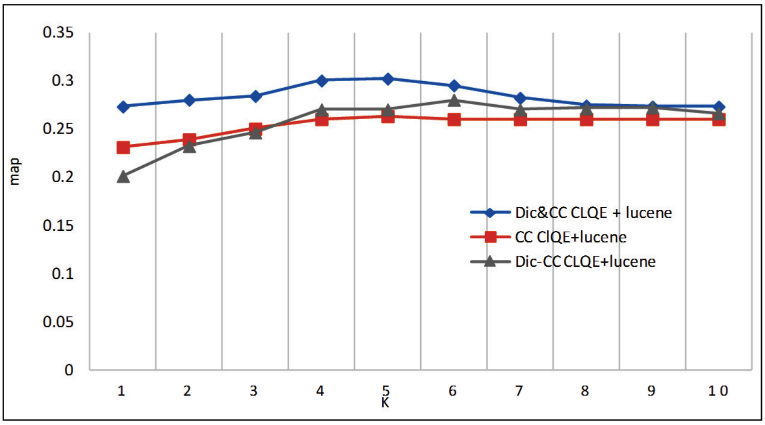


Fig. 4. The MAP of the different number of *k*.

Fig. 4 shows the MAP for different number of translation words, with varying the *k* value number from 1to 10. In our experiment, there are three methods CC CLQE, Dic&CC ClQE and Dic-CC CLQE adopted for comparison, and every line of map value was trend to steady. The MAP of Dic&CC ClQE shows the best performance and raises when using the top0–top5 translation words. The MAP value was raising, and when we use top5 translation words, the Map value reach the Maximum when using the top5 translation words. However, the MAP value diminishes and trend to be the same value when choosing the top5–top10 translation words. In general, the top4–top6 translation words may be a better selection for our experiment.

In order to more accurately evaluate the results of CLIR, there are other evaluation indicators for these three methods. In our experiments, we observe that the improvements of precision at top5 and top10 documents are more considerable than the mean average precision. Intuitively, when the precision at top documents are high, Dic&CC CLQE can contribute to improve the mean average precision. We use top4 translations of all query terms for the experiments. Table 2 shows the MAP, Precision at 5 documents (Prec@5) and Precision at 10 documents (Prec@10) of results of CLIR with our query expansion methods.

Table 2. Query expansion with three methods

Method	MAP	Prec@5	Prec@10
CC CLQE + Lucene	0.271	0.387	0.365
Dic&CC CLQE + Lucene	0.305	0.443	0.427
Dic-CC CLQE + Lucene	0.280	0.392	0.371

4. Conclusions and Future Work

The results reported in the current work make many contributions to the CLIR research. To begin with, we have shown that using a machine-readable dictionary and a comparable corpus together can yield better results from any one source alone. Then, we have shown that how prior work on the application of Pearson correlation coefficient to perform translation selection in a query translation architecture can be extended to perform the translation weighting in ways that can yield improved ranked retrieval effectiveness. Moreover, we have shown that the algorithm which association of monolingual words combine with the association of bilingual words, can be used as a basis for effective post-translation query expansion.

In this work, we presented a novel method to mine translation knowledge from comparable corpora. The most notable presented method was based on Chinese and Thai comparable corpora which consists of correlations of terms. Experiments show that the method can effectively improve the accuracy of Chinese and Thai CLIR.

In our future work, we plan to use the post-translation query expansion. It will use a random walk on the Wikipedia link graph which yield more consistent improvements, even for a language like Thai where the number of Wikipedia pages is far smaller than it is for Chinese. The result can be improved not only when using a machine-readable dictionary, but also when the coverage is augmented using translation, synonymy and polysemy relations extracted from Wikipedia. And the next step, we will focus on making the full use of the pseudo relevance feedback to improve the quality of the Chinese and Thai CLIR.

Acknowledgement

This work was supported by National Nature Science Foundation (Grant Nos.61472168,61672271), Emerging Industry Technology Development Project of Yunnan Province (Grant No. 201606), and Science and Technology Innovation Talents Fund Projects of Ministry of Science and Technology (Grant No. 2014HE001), Key Laboratory in Regional University Projects of Yunnan Province, Innovation Team Projects of Yunnan Province (Grant No. 2014HC012).

References

[1] T. Talvensaari, J. Laurikkala, K. Jarvelin, M. Juhola, and H. Keskustalo, “Creating and exploiting a comparable corpus in cross-language information retrieval,” *ACM Transactions on Information Systems (TOIS)*, vol. 25, no. 1, article no. 4, 2007.

[2] T. Talvensaari, A. Pirkola, K. Jarvelin, M. Juhola, and J. Laurikkala, “Focused web crawling in the acquisition of comparable corpora,” *Information Retrieval*, vol. 11, no. 5, pp. 427-445, 2008.

[3] O. Batarfi, R. Elshawi, A. Fayoumi, A. Barnawi, and S. Sakr, “A distributed query execution engine of big attributed graphs,” *SpringerPlus*, vol. 5, no. 1, article no. 665, 2016.

[4] H. U. Weiping and R. Wang, “Vehicle license plate recognition method based on threshold segmentation and region growing,” *Journal of Guangxi Academy of Sciences*, vol. 2016, no. 1, pp. 54-58.2016.

[5] J. Dalton, L. Dietz, and J. Allan, “Entity query feature expansion using knowledge base links,” in *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, Gold Coast, Australia, 2014, pp. 365-374.

[6] C. Carpineto and G. Romano, “A survey of automatic query expansion in information retrieval,” *ACM Computing Surveys (CSUR)*, vol. 44, no. 1, article no. 1, 2012.

[7] P. Gupta, K. Bali, R. E. Banchs, M. Choudhury, and P. Rosso, “Query expansion for mixed-script information retrieval,” in *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, Gold Coast, Australia, 2014, pp. 677-686.

[8] L. Li and H. Wang, “Multi-strategy query expansion method based on semantics,” *Journal of Digital Information Management*, vol. 12, no. 3, pp. 183-191, 2014.

[9] J. Gao, X. He, and J. Y. Nie, “Clickthrough-based translation models for web search: from word models to phrase models,” in *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, Toronto, Canada, 2010, pp. 1139-1148.

[10] J. Gao and J. Y. Nie, “Towards concept-based translation models using search logs for query expansion,” in *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, Maui, HI, 2012.

[11] J. Arguello, J. L. Elsas, J. Callan, and J. S. Carbonell, “Document representation and query expansion models for blog recommendation,” in *Proceedings of the 2nd International Conference on Weblogs and Social Media (ICWSM)*, Seattle, WA, 2008.

[12] A. Samoilenko, F. Karimi, D. Edler, J. Kunegis, and M. Strohmaier, “Linguistic neighbourhoods: explaining cultural borders on Wikipedia through multilingual co-editing activity,” *EPJ Data Science*, vol. 5, no. 1, article no. 9, 2016.

[13] A. Kotov and C. Zhai, “Tapping into knowledge base for concept feedback: leveraging conceptnet to improve search results for difficult queries. In *Proceedings of the 5th ACM International Conference on Web Search and Data Mining*, Seattle, WA, 2012, pp. 403-412.

[14] M. Lu, X. Sun, S. Wang, D. Lo, and Y. Duan, “Query expansion via WordNet for effective code search,” in *Proceedings of 2015 IEEE 22nd International Conference on Software Analysis, Evolution and Reengineering (SANER),* Montreal, Canada, 2015, pp. 545-549.

[15] V. Klyuev and A. Yokoyama, “Web query expansion: a strategy utilising Japanese WordNet,” *JoC*, vol. 1, no. 1, pp. 23-28, 2010.

[16] C. Xiong and J. Callan, “Query expansion with freebase,” in *Proceedings of the 2015 International Conference on the Theory of Information Retrieval*, Northampton, MA, 2015, pp. 111-120.

[17] M. de Boer, K. Schutte, and W. Kraaij, “Knowledge based query expansion in complex multimedia event detection,” *Multimedia Tools and Applications*, vol. 75, no. 15, pp. 9025-9043, 2016.

[18] F. Colace, M. De Santo, L. Greco, and P. Napoletano, "Improving relevance feedback-based query expansion by the use of a weighted word pairs approach," *Journal of the Association for Information Science and Technology*, vol. 66, no. 11, pp. 2223-2234, 2015.

[19] A. Sordoni, Y. Bengio, and J. Y. Nie, "Learning concept embeddings for query expansion by quantum entropy minimization," in *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, Quebec, Canada, 2014, pp. 1586-1592.

[20] H. B. Hashemi and A. Shakery, "Mining a Persian–English comparable corpus for cross-language information retrieval," *Information Processing & Management*, vol. 50, no. 2, pp. 384-398, 2014.

[21] J. Bhogal, A. MacFarlane, and P. Smith, "A review of ontology based query expansion," *Information Processing & Management*, vol. 43, no. 4, pp. 866-886, 2007.

[22] D. Roy, D. Paul, M. Mitra, and U. Garain, "Using word embeddings for automatic query expansion," 2016 [Online]. Available: https://arxiv.org/abs/1606.07608.

[23] R. Vaidyanathan, S. Das, and N. Srivastava, "Query expansion strategy based on pseudo relevance feedback and term weight scheme for monolingual retrieval," *International Journal of Computer Applications*, vol. 105, no. 8, pp. 1-6, 2014.

[24] J. Singh and A. Sharan, "Context window based co-occurrence approach for improving feedback based query expansion in information retrieval," *International Journal of Information Retrieval Research (IJIRR)*, vol. 5, no. 4, pp. 31-45, 2015.

[25] R. Rahimi, A. Shakery, and I. King, "Extracting translations from comparable corpora for Cross-Language Information Retrieval using the language modeling framework," *Information Processing & Management*, vol. 52, no. 2, pp. 299-318, 2016.

[26] A. Shakery and C. Zhai, "Leveraging comparable corpora for cross-lingual information retrieval in resource-lean language pairs," *Information Retrieval*, vol. 16, no. 1, pp. 1-29, 2013.

[27] J. Yu, C. Li, W. Hong, S. Li, and D. Mei, "A new approach of rules extraction for word sense disambiguation by features of attributes," *Applied Soft Computing*, vol. 27, pp. 411-419, 2015.

[28] J. Lafferty and C. Zhai, "Document language models, query models, and risk minimization for information retrieval," in *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New Orleans, LA, 2001, pp. 111-119.

[29] L. Quesada, F. Berzal, and F. J. Cortijo, "A lexical analysis tool with ambiguity support," 2012 [Online]. Available: https://arxiv.org/abs/1202.6583.

[30] S. Kim, Y. Ko, and D. W. Oard, "Combining lexical and statistical translation evidence for cross-language information retrieval," *Journal of the Association for Information Science and Technology*, vol. 66, no. 1, pp. 23-39, 2015.

[31] P. Sorg and P. Cimiano, "Exploiting Wikipedia for cross-lingual and multilingual information retrieval," *Data & Knowledge Engineering*, vol. 74, pp. 26-45, 2012.

[32] H. E. Ping and L. I. Fan, "Design and implementation of a Lucene-based full-text retrieval management system," *Journal of Yangtze University (Natural Science Edition)*, vol. 2014, no. 22, pp. 35-38, 2014.

Peili Tang

She is current a postgraduate in the Kunming University of Science and Technology, Kunming, China. She focus on nature language processing and information retrieval.

Jing Zhao

She graduated with a master's degree in Kunming University of Science and Technology Kunming, China. Now, she is a teacher in National University of Defense Technology. She focus on nature language processing and information retrieval.

Zhengtao Yu

He is the corresponding author. He is currently a professor and Ph.D. supervisor at School of Information Engineering and Automation, and the chairman of Key Laboratory of Intelligent Information Processing, Kunming University of Science and Technology, Kunming, China. He received the Ph.D. degree in Computer Application Technology from Beijing Institute of Technology, Beijing, China, in 2005. His main research interests include natural language processing, machine translation and information retrieval.

Zhuo Wang

He is current a postgraduate in the Kunming University of Science and Technology, Kunming, China. She focus on nature language processing and information retrieval.

Yantuan Xian

He is currently a Ph.D. candidate at Kunming University of Science and Technology, Kunming, China. He received his M.S. degree in pattern recognition and intelligent system from Shenyang Institutof Automation (SIA), Chinese Academy of Sciences, in 2006. His major research interests are machine translation and information retrieval.