Cross-lingual Information Retrieval Model based on Bilingual Topic Correlation

Yuansheng LUO\textsuperscript{1,2}, Zhongjian LE\textsuperscript{2}, Mingwen WANG\textsuperscript{3,*}

\textsuperscript{1}Modern Education Technology Center, Jiangxi University of Finance and Economics, Nanchang 330013, China
\textsuperscript{2}School of Information Management, Jiangxi University of Finance and Economics, Nanchang 330013, China
\textsuperscript{3}School of Computer Information and Engineering, Jiangxi Normal University, Nanchang 330022, China

Abstract

How to construct relationship between bilingual texts is important to effectively processing multi-lingual text data and cross language barriers. Cross-lingual latent semantic indexing (CL-LSI) corpus-based does not fully take into account bilingual semantic relationship. The paper proposes a new model building semantic relationship of bilingual parallel document via partial least squares (PLS). In the model, the parallel documents are viewed as two different lingual representations for the same semantic object and a single latent semantic space for each language is created. The task of cross-lingual information retrieval (CLIR) is performed in the new bilingual latent semantic spaces. Experimental results on the Chinese-English aligned news stories collected from bilingual news website of the Wall Street Journal and Finance Times show that performance of CLIR in bilingual semantic spaces built by the presented model is over CL-LSI.

Keywords: Cross-lingual Information Retrieval; Bilingual Text Correlation; Bilingual Latent Semantic Spaces; Partial Least Squares

1 Introduction

With rapid development of the Internet and the accelerating process of globalization, information resource provided by the Internet is no longer from English and a few other languages. It is showed from website Internet World Stats (www.internetworldstats.com) that the number of the Internet users in the world is 2.01 billion as of May 31, 2011. English is ranked first. Its users are 26.8% of all Internet users in the world. The second languages of the Internet is Chinese, and the percentage of its users in the world Internet users is 24.2. It is reported from W3Techs that

\textsuperscript{*}Project supported by the National Nature Science Foundation of China (No. 60963014) and and the Science and Technology Project of Education Department of Jiangxi Province (No. GJJ10116).
\textsuperscript{*}Corresponding author.
\textit{Email address: mwwang@jxnu.edu.cn} (Mingwen WANG).
English is the largest percentages of websites using various content languages and its percentages is 55.7. Other languages are all less than 7%, including Chinese. The percentage of Chinese is 4.5. If the Internet users are not proficient non-native speakers, it is the more difficultly that they use non-native language to search information, but it is easy that they use native language to look up information and browse them by means of translation tools. The diversity of the language in Web pages and the users’ varied proficiency in their native languages and foreign languages inevitably bring barriers to their use of the Internet resources. Cross-Language Information Retrieval (CLIR) is the query and the target documents are in different languages in information retrieval.

Topic model attracts researchers’ attention in the community of machine learning, information retrieval and natural information process. It has become an efficient technique of CLIR. Given a document set, topic model builds relationship between term and document from semantically related latent topic sets. Topic is defined as the probability distribution of terms. A document is defined as bag of words generated by mixture topics. The typical topic models include Latent Semantic Indexing (LSI)[1], Probabilistic Latent Semantic Indexing (PLSI)[2] and Latent Dirichlet Allocation (LDA)[3].

The paper focuses on semantic correlation of bilingual topics extracted from bilingual parallel documents. The document pair is viewed as two languages’ representation of the same semantic context. The corpus which contains the document pairs constructs latent semantic space for each language. The task of CLIR is performed in these spaces. So we propose a model of building bilingual related topic using Partial Least Squares (PLS). The new model is called as Bilingual Partial Least Squares correlation (BiPLS). In this model each topic space is separate, semantically related and language independent. Furthermore, the model does not use machine translation or bilingual dictionary built by manual way or parallel corpus.

This paper is organized as follows. In section 2 we will review the related works of topic models in CLIR. In section 3 we will introduce our new model, followed by the experimental evaluations in section 4. Finally, the conclusion and future work will be drawn in Section 5.

2 Related Work

Topic model commonly includes probability and non-probability methods[4]. Topic is defined as probability distribution of terms and document is generated by mixture topics in the probability methods. PLSI and LDA are widely studied probabilistic topic models, which are received more attention and in-depth research in the community of machine learning and information retrieval. There are many related works in CLIR. A Polylingual Topic Model (PLTM) based on LDA constructed polylingual topics from multi-lingual corpus in Mimno et al. work[5]. Vulic et al. extracted language independent and inter-semantic bilingual topic space from bilingual corpus of document alignment[6]. Jagarlamudi and Daum III proposed JointLDA model to extract multi-lingual themes from the bilingual comparable corpus with no aligned documents. Zhang et al. presented a probabilistic cross-lingual latent semantic analysis (PCLSA) method to get latent topics shared by two language from bilingual dictionary[7]. Muramatsu and Mori incorporated probabilistic CLIR model with information of translation probability provided by PLSI from the bilingual parallel corpus[8]. Jin et al. reported a weekly-supervised probabilistic latent semantic indexing method[9].

Cross-language latent semantic indexing (LSI), non-negative matrix factorization (NMF) and
canonical correlation analysis (CCA) are non-probability topic models. Topic model is derived from latent semantic indexing, which is a commonly used method in information retrieval[1]. LSI builds a new low-dimension latent semantic space by singular value decomposition (SVD). The semantic similarities between documents, between queries, between terms are computed in the new space. In the field of CLIR, Dumais et al. presented Cross-language LSI (CL-LSI) [10, 11, 12]. In CL-LSI, each aligned bilingual document is concatenated as a dual document. The training document set is made by these dual documents. The training set is used to build bilingual multi-dimension indexing semantic space. CL-LSI can automatically retrieve and categorize cross-lingual documents in multi-lingual document without directly translation. The dimension of subspace constructed by CL-LSI is far less than the original dimension. The subspace can well pick up latent semantic structure with multi-lingual semantic information and language-independent feature. CL-LSI concatenates each pair of bilingual documents as a dual document and catches semantic relationship between two languages by exploiting the co-occurrence of bilingual terms. However, the method of mixing documents cannot incorporate with intrinsic property of each language and semantic correlation of cross-language. The mixing method would affect the performance of CLIR.

To address this problem, Littman et al. constructed latent semantic space based SVD for each language[13]. Chew et al. presented a PARAFAC2 model (a multi-way SVD model) on multi-language Bible text documents. The model built a term-document matrix for each language and retrieved cross-lingual document separately using LSI. Furthermore, NMF[14, 15], CCA[16, 17] and Kernel CCA[18, 19, 20] were applied to retrieve cross-language document. In previous work, we built bilingual inter-semantic latent space by mean of partial least squares (PLS) [21].

Our model is different from previous works. At first, BiPLS is a non-probability bilingual topic model. The new model looks two bilingual aligned documents as two views of representing the same semantic object. In the step of retrieve cross language documents or queries, BiPLS directly projects documents or queries onto bilingual topic space without integrating information of machine translation or bilingual dictionary. Compared with CL-LSI, BiPLS doesnot combine the original bilingual document spaces, but builds a single latent semantic space for each language and incorporate with semantic relationship between languages. Compared to CCA methods using in CLIR, the objective function of BiPLS maximize the covariance of the corresponding topics, but that of CCA maximize the correlation of the corresponding topics. Mathematically, BiPLS more reveals the latent semantic structure of bilingual corpus.

3 Bilingual Partial Least Squares Correlation Model

A semantic object is represented as strings (word or phrase) using different language in natural language. Essentially, the strings are some meaningful symbolic description of the object and regarded as different views of the same object. For example, datum in English is expressed as dato in Spanish and daten in German. Therefore, the bilingual documents are looked as two views of the same semantic object. The views are semantically equivalent objects and a pair of translation. A document set of language $L_1$ is denoted as an $m$ by $n$ matrix $X = \{x_i\}^m_{i=1}$, where each row vector is defined as a document and column vector is defined as a term. All terms in $X$ combine a vocabulary $V_x = \{c_1, c_2, \ldots, c_n\}$. Likely, a document set of language $L_2$, which is a document-aligned corpus of $X$, is denoted as an $m$ by $r$ matrix $Y = \{y_i\}^m_{i=1}$. Its vocabulary is $V_y = \{e_1, e_2, \ldots, e_r\}$. Our model aims to extract two separate topic spaces for two languages from
bilingual parallel corpus \(< X, Y >\). The model formulates statistical dependencies of bilingual semantic correlation to build semantic correspondence of bilingual topics. Partial Least Squares (PLS) is an effective multivariate statistical analysis for modeling statistical dependencies of \(X\) and \(Y\). It models correlation between original sample sets by latent variances\(^1\).

First, dot product of column vectors in bilingual parallel document matrices \(X\) and \(Y\) is computed. The dot product matrix \(Z\) is \(Z = X^T Y\).

Therefore, the correlation between the \(i\)th column vector of \(X\) and the \(j\)th column vector of \(Y\) is measured by their dot product between these two columns. If these column vectors are centered, the dot product gives the covariance between these two columns. If these column vectors are normalized, the dot product expresses the correlation between these two columns.

How to extract latent bilingual topic space from \(Z\)? Partial least squares is an effective way to extract the latent bilingual topic space from matrices \(X\) and \(Y\). The space is consisted of a series of latent variable pairs \(< u_i, v_i >\) \((i = 1, 2, \ldots, k)\). These variable pairs are bilingual related. They abstract the semantic information shared by bilingual documents. These variable pairs are estimated by coefficients of the linear combination (or weight vectors). The first weight vector \(w_1\) is get by the objective function as follows:

\[
\max_{w_1} w_1^T X^T Y Y^T X w_1, \text{ subject to } \|w_1\| = 1 \tag{1}
\]

It is shown from the equation (1) that \(w_1\) is an eigenvector corresponding to the largest eigenvalue of matrix \(X^T Y Y^T X\). To solve the equation (2), \(Z\) is decomposed into three matrices by SVD:

\[
Z = U \Sigma V^T = \sum_{i=1}^{k} \lambda_i u_i v_i^T \tag{2}
\]

where \(U\) is made up of the top \(k\) left eigenvectors of \(Z\), \(V\) is consisted of the top \(k\) right eigenvectors of \(Z\) and \(\lambda_i\) is correlation between bilingual eigenvectors. From a semantic perspective, \(U\) is the best semantic description of terms in language \(L_1\) and \(V\) is the best semantic description of terms in language \(L_2\). The number of basis vectors in equation (2) is \(min\{\text{Rank}(X)\text{Rank}(Y)\}\). In practice, the number of topics \(k\) is far less than the size of terms. The top \(k\) basis vectors constitute space of bilingual topic \((U, V)\). Any document can be projected onto the space. Give a document \(x \in \mathbb{R}^n\) and a document \(y \in \mathbb{R}^r\):

\[
x \rightarrow U^T x = u \in \mathbb{R}^k \tag{3}
\]

\[
y \rightarrow V^T y = v \in \mathbb{R}^k \tag{4}
\]

So we construct bilingual topic space with semantic correlation from parallel document set according to above steps. Then cross-language retrieve task is conducted by projected the original document onto the latent space of its own language as shown in equation (3) and (4). The similarity between \(x\) and \(y\) is \(\text{Sim}(x, y) = u^T v / \|u\| \|v\|\) In our previous work, we presented CLIR model based on latent inter-semantic space by exploiting Partial Least Squares Regression (PLSR) \([22, 23]\). The first pair of weight vector \(w_1\) and \(c_1\) is derived from the following objective function:

\[
\max_{w_1} w_1^T X^T Y c_1, \text{ subject to } \|w_1\| = 1, \|c_1\| = 1 \tag{5}
\]

The model is called as CL-PLSR. BiPLS is a symmetrical approach and equally treats bilingual parallel document set \(X\) and \(Y\). It can effectively capture the semantic information shared by
two languages. BiPLS’s algorithm is implemented by SVD. The algorithm could simultaneously extract all topics. CL-PLSR is an asymmetrical approach and can pick up more topics. It can build predicting analysis of the relationship between independent variable $X$ and dependent variable $Y$. However, CL-PLSR requires iteratively deflating $X$ and $Y$. It is not suitable to large-scale computing because it consumes higher memory when it runs on a large document set. On the other hand, CL-PLS difficultly determined the number of latent variable pairs.

4 Experiments

In this section, we describe the empirical researches that we conducted on Chinese-English CLIR system to evaluate the practicality of our model.

4.1 Experimental design

Data Sets: Since there is no standard bilingual parallel corpus with queries of different languages and relevance judgments for CLIR evaluation, we conduct our evaluation on Web pages which were crawled or manually downloaded from bilingual Websites articles published from January 2006 to April 2009. The Websites include the Wall Street Journal (WSJ) bilingual news (chinese.wsj.com) and Financial Times (FT) bilingual news (www.ftchinese.com). Most of WSJ and FT stories involve economy. The total number of the downloaded Web pages was 19,387. Only the titles and bodies in Web page were extracted. The texts were obtained by parsing the html files. Each document consists of Chinese article and corresponding English article. To learn the latent semantic dual-space, we randomly sample 1000, 2000 and 4000 documents without replacement. The three document sets are denoted WSJFT-1000, WSJFT-2000 and WSJFT-4000, respectively.

Experimental Setup: In the step of data preprocessing, all numbers and stop-words were filtered out, all characters were converted into lowercase, and word stemming was performed by the Porter stemmer. All Chinese texts, articles, and queries were word-segmented using the Chinese word segmentation system ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System). In order to reduce the size of the feature space and speed up the training of our model on data set, a simple feature selection method, term frequency (TF) was used to cut down the number of features. In our experiments, TF threshold is set to 3. We compute the weight vector for each document using the LTC weighting, a variant of $TF \times IDF$ weighting.

CLIR System: The system is made up of three steps. They are pretreatment of document sets, building latent space and projection of query or document. We trained WSJFT-1000 and WSJFT-2000 to construct bilingual latent space. In the phrase of projection, WSJFT-4000 is mapped onto the built spaces and WSJFT-1000 and WSJFT-2000 are projected onto their own of the latent spaces. The task of retrieval is cross-lingual mate search. Cross-language mate retrieval uses documents to find their cross-language mates. It can be thought of as treating each of Chinese or English documents as query, each one with exactly one relevant document in corresponding language - its translation (or mate).

CLIR Baselines: We compared the retrieval performance of our model (BiPLS) with two baselines: CL-LSI and CL-PLSR under the same experimental settings.
4.2 Experimental results and discussions

Table 1: Mate retrieval (1000 training documents): the accuracy rates averaged over all the training documents and over other 4000 test documents, respectively. Different numbers of eigenvectors were used and BiPLS was compared with CL-LSI and CL-PLSR.

<table>
<thead>
<tr>
<th># Variable pair</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>300</th>
<th>400</th>
<th>500</th>
<th>800</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BiPLS(C→E)</td>
<td>0.5010</td>
<td>0.9250</td>
<td>0.9930</td>
<td>0.9980</td>
<td>0.9980</td>
<td>0.9980</td>
<td>0.9980</td>
<td>0.9980</td>
<td>0.9980</td>
<td>0.9980</td>
<td>0.9980</td>
</tr>
<tr>
<td>BiPLS(E→C)</td>
<td>0.5040</td>
<td>0.9170</td>
<td>0.9850</td>
<td>0.9980</td>
<td>0.9980</td>
<td>0.9980</td>
<td>0.9980</td>
<td>0.9980</td>
<td>0.9980</td>
<td>0.9980</td>
<td>0.9980</td>
</tr>
<tr>
<td>CL-LSI(C→E)</td>
<td>0.2480</td>
<td>0.7920</td>
<td>0.9600</td>
<td>0.9960</td>
<td>0.9980</td>
<td>0.9980</td>
<td>0.9980</td>
<td>0.9980</td>
<td>0.9980</td>
<td>0.9980</td>
<td>0.9980</td>
</tr>
<tr>
<td>CL-LSI(E→C)</td>
<td>0.2640</td>
<td>0.7730</td>
<td>0.9480</td>
<td>0.9900</td>
<td>0.9970</td>
<td>0.9980</td>
<td>0.9980</td>
<td>0.9980</td>
<td>0.9980</td>
<td>0.9980</td>
<td>0.9980</td>
</tr>
<tr>
<td>CL-PLSR(C→E)</td>
<td>0.5300</td>
<td>0.9010</td>
<td>0.9880</td>
<td>0.9960</td>
<td>0.9980</td>
<td>0.9980</td>
<td>0.9980</td>
<td>0.9980</td>
<td>0.9980</td>
<td>0.9980</td>
<td>0.9980</td>
</tr>
<tr>
<td>CL-PLSR(E→C)</td>
<td>0.5230</td>
<td>0.8940</td>
<td>0.9850</td>
<td>0.9980</td>
<td>0.9980</td>
<td>0.9980</td>
<td>0.9980</td>
<td>0.9980</td>
<td>0.9980</td>
<td>0.9980</td>
<td>0.9980</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th># Variable pair</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>300</th>
<th>400</th>
<th>500</th>
<th>800</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>BiPLS(C→E)</td>
<td>0.1515</td>
<td>0.5738</td>
<td>0.8020</td>
<td>0.9200</td>
<td>0.9555</td>
<td>0.9695</td>
<td>0.9695</td>
<td>0.9710</td>
<td>0.9735</td>
<td>0.9710</td>
<td>0.9685</td>
</tr>
<tr>
<td>BiPLS(E→C)</td>
<td>0.1505</td>
<td>0.5470</td>
<td>0.7798</td>
<td>0.8980</td>
<td>0.9325</td>
<td>0.9500</td>
<td>0.9580</td>
<td>0.9585</td>
<td>0.9583</td>
<td>0.9515</td>
<td>0.9510</td>
</tr>
<tr>
<td>CL-LSI(C→E)</td>
<td>0.0635</td>
<td>0.4063</td>
<td>0.7228</td>
<td>0.8903</td>
<td>0.9345</td>
<td>0.9563</td>
<td>0.9628</td>
<td>0.9635</td>
<td>0.9645</td>
<td>0.9638</td>
<td>0.9610</td>
</tr>
<tr>
<td>CL-LSI(E→C)</td>
<td>0.0778</td>
<td>0.4173</td>
<td>0.7125</td>
<td>0.8678</td>
<td>0.9100</td>
<td>0.9368</td>
<td>0.9460</td>
<td>0.9515</td>
<td>0.9531</td>
<td>0.9480</td>
<td>0.9455</td>
</tr>
<tr>
<td>CL-PLSR(C→E)</td>
<td>0.1890</td>
<td>0.5708</td>
<td>0.8005</td>
<td>0.9265</td>
<td>0.9605</td>
<td>0.9720</td>
<td>0.9743</td>
<td>0.9740</td>
<td>0.9733</td>
<td>0.9700</td>
<td>0.9688</td>
</tr>
<tr>
<td>CL-PLSR(E→C)</td>
<td>0.1723</td>
<td>0.5433</td>
<td>0.7710</td>
<td>0.9073</td>
<td>0.9363</td>
<td>0.9553</td>
<td>0.9508</td>
<td>0.9580</td>
<td>0.9565</td>
<td>0.9563</td>
<td>0.9560</td>
</tr>
</tbody>
</table>

Table 2: Mate retrieval (2000 training documents): the accuracy rates averaged over all the training documents and over other 4000 test documents, respectively. Different numbers of eigenvectors were used and BiPLS was compared with CL-LSI and CL-PLSR.

<table>
<thead>
<tr>
<th># Variable pair</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>300</th>
<th>400</th>
<th>500</th>
<th>800</th>
<th>1500</th>
<th>2000</th>
</tr>
</thead>
<tbody>
<tr>
<td>BiPLS(C→E)</td>
<td>0.3565</td>
<td>0.8720</td>
<td>0.9735</td>
<td>0.9950</td>
<td>0.9970</td>
<td>0.9970</td>
<td>0.9975</td>
<td>0.9975</td>
<td>0.9975</td>
<td>0.9975</td>
<td>0.9975</td>
<td>0.9975</td>
</tr>
<tr>
<td>BiPLS(E→C)</td>
<td>0.3665</td>
<td>0.8620</td>
<td>0.9650</td>
<td>0.9905</td>
<td>0.9970</td>
<td>0.9975</td>
<td>0.9975</td>
<td>0.9975</td>
<td>0.9975</td>
<td>0.9975</td>
<td>0.9975</td>
<td>0.9975</td>
</tr>
<tr>
<td>CL-LSI(C→E)</td>
<td>0.1830</td>
<td>0.7060</td>
<td>0.9250</td>
<td>0.9815</td>
<td>0.9955</td>
<td>0.9960</td>
<td>0.9970</td>
<td>0.9970</td>
<td>0.9975</td>
<td>0.9975</td>
<td>0.9975</td>
<td>0.9975</td>
</tr>
<tr>
<td>CL-LSI(E→C)</td>
<td>0.1980</td>
<td>0.6890</td>
<td>0.9080</td>
<td>0.9765</td>
<td>0.9915</td>
<td>0.9965</td>
<td>0.9975</td>
<td>0.9975</td>
<td>0.9975</td>
<td>0.9975</td>
<td>0.9975</td>
<td>0.9975</td>
</tr>
<tr>
<td>CL-PLSR(C→E)</td>
<td>0.3990</td>
<td>0.8345</td>
<td>0.9710</td>
<td>0.9955</td>
<td>0.9970</td>
<td>0.9970</td>
<td>0.9970</td>
<td>0.9970</td>
<td>0.9975</td>
<td>0.9975</td>
<td>0.9975</td>
<td>0.9975</td>
</tr>
<tr>
<td>CL-PLSR(E→C)</td>
<td>0.4000</td>
<td>0.8185</td>
<td>0.9550</td>
<td>0.9675</td>
<td>0.9955</td>
<td>0.9975</td>
<td>0.9975</td>
<td>0.9975</td>
<td>0.9975</td>
<td>0.9975</td>
<td>0.9975</td>
<td>0.9975</td>
</tr>
</tbody>
</table>

Table 1 and Table 2 list the results of mate retrieval when WSJFT-1000 and WSJFT-2000 are training for building the latent semantic space, respectively. We compared the average accuracies of BiPLS, CL-LSI and CL-PLSR. The average accuracies is the percentage of mate document ranking at top one. The number of variable pairs k of the models is listed in the first row in Table 1 and Table 2. The value k ranges from 5 to 100 in Table 1 because WSJFT-1000 has only 1000 documents. The value k ranges from 5 to 2000 in Table 2. In the tables, E→C means English query to retrieve Chinese documents and C→E means Chinese document as query to search English documents.

We can see from Table 1 and Table 2 that the average accuracies of BiPLS reach 0.9 when the number of variable pairs k is set to 10 over WSJFT-1000 Training Docs as queries and when k is set to 20 over WSJFT-2000 Training Docs as queries in task of C→E and E→C. The average accuracies of BiPLS achieve 0.9 when k is set to 50 when WSJFT-4000 is tested. So
the fewer latent variable pairs can catch the semantic information shared by bilingual documents. Furthermore, the number of variable pairs is about 500 when BiPLS reaches the best performance. It is showed from Table 1 and Table 2 that the performance of BiPLS, CL-LSI and CL-PLSR are very close when WSJFT-1000 and WSJFT-2000 are tested as queries when $k$ is greater than 100. Because the training documents are used are queries, the documents provide full bilingual semantic information.

Overall, BiPLS outperforms over CL-LSI and is near to CL-PLSR in the task of retrieving English query against Chinese documents and Chinese query against English documents. It is noteworthy that the average accuracies of three models in the task of C→E is higher than those of in the task of E→C. The possible reason is that terms of Chinese is the more informative than terms of English because English words are stepped and less semantic information.

5 Conclusions

In this paper, we present a bilingual Partial Least Squares correlation (BiPLS) model based on bilingual topic correlation. The model is a non-probability bilingual topic model. It treats two bilingual aligned documents as two views of representing the same semantic object. Also, our model does not extract latent structure of combination the original bilingual document spaces, but builds a single latent semantic space for each language and incorporate with semantic relationship between languages. BiPLS aim to maximize the covariance of the corresponding topics. It more uncovers the latent semantic structure of parallel corpus, but does not synthesize information of machine translation or bilingual dictionary. We downloaded Chinese-English aligned-documents from bilingual news website of the Wall Street Journal and Finance Times to create bilingual topic space. The experimental results on these parallel documents showed that BiPLS outperforms CL-LSI. In the future, we plan to improve the effectiveness of our model using some effective strategies on the larger documents sets. For example, the model integrates with query expansion and applied to multi-language text classification and clustering. On the other hand, BiPLS model will be extended to multi-language information retrieval.

References


