Building Comparable Corpora Based on Bilingual LDA Model

Zede Zhu
University of Science and Technology of China, Institute of Intelligent Machines Chinese Academy of Sciences
Hefei, China
zhuzede@mail.ustc.edu.cn

Miao Li, Lei Chen, Zhenxin Yang
Institute of Intelligent Machines Chinese Academy of Sciences
Hefei, China
mli@iim.ac.cn,alan.cl@163.com,
xinzyang@mail.ustc.edu.cn

Abstract

Comparable corpora are important basic resources in cross-language information processing. However, the existing methods of building comparable corpora, which use inter-translate words and relative features, cannot evaluate the topical relation between document pairs. This paper adopts the bilingual LDA model to predict the topical structures of the documents and proposes three algorithms of document similarity in different languages. Experiments show that the novel method can obtain similar documents with consistent topics own better adaptability and stability performance.

1 Introduction

Comparable corpora can be mined fine-grained translation equivalents, such as bilingual terminologies, named entities and parallel sentences, to support the bilingual lexicography, statistical machine translation and cross-language information retrieval (AbduI-Rauf et al., 2009). Comparable corpora are defined as pairs of monolingual corpora selected according to the criteria of content similarity but non-direct translation in different languages, which reduces limitation of matching source language and target language documents. Thus comparable corpora have the advantage over parallel corpora in which they are more up-to-date, abundant and accessible (Ji, 2009).

Many works, which focused on the exploitation of building comparable corpora, were proposed in the past years. Tao et al. (2005) acquired comparable corpora based on the truth that terms are inter-translation in different languages if they have similar frequency correlation at the same time periods. Talvensaari et al. (2007) extracted appropriate keywords from the source language documents and translated them into the target language, which were regarded as the query words to retrieve similar target documents. Thuy et al. (2009) analyzed document similarity based on the publication dates, linguistic independent units, bilingual dictionaries and word frequency distributions. Otero et al. (2010) took advantage of the translation equivalents inserted in Wikipedia by means of interlanguage links to extract similar articles. Bo et al. (2010) proposed a comparability measure based on the expectation of finding the translation for each word.

The above studies rely on the high coverage of the original bilingual knowledge and a specific data source together with the translation vocabularies, co-occurrence information and language links. However, the severest problem is that they cannot understand semantic information. The new studies seek to match similar documents on topic level to solve the traditional problems. Preiss (2012) transformed the source language topical model to the target language and classified probability distribution of topics in the same language, whose shortcoming is that the effect of model translation seriously hampers the comparable corpora quality. Ni et al. (2009) adapted monolingual topic model to bilingual topic model in which the documents of a concept unit in different languages were assumed to share identical topic distribution. Bilingual topic model is widely adopted to mine translation equivalents from multi-language documents (Mimno et al., 2009; Ivan et al., 2011).

Based on the bilingual topic model, this paper predicts the topical structure of documents in different languages and calculates the similarity of topics over documents to build comparable corpora. The paper concretely includes: 1) Introduce the Bilingual LDA (Latent Dirichlet Allocation) model which builds comparable corpora and improves the efficiency of matching similar documents; 2) Design a novel method of TFIDF (Topic Frequency-Inverse Document Frequency) to enhance the distinguishing ability of topics from different documents; 3) Propose a tailored

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 278–282,
Sofia, Bulgaria, August 4-9 2013.

method of conditional probability to calculate document similarity; 4) Address a language-independent study which isn't limited to a particular data source in any language.

2 Bilingual LDA Model

2.1 Standard LDA

LDA model (Blei et al., 2003) represents the latent topic of the document distribution by Dirichlet distribution with a *K*-dimensional implicit random variable, which is transformed into a complete generative model when β is exerted to Dirichlet distribution (Griffiths et al., 2004) (Shown in Fig. 1),

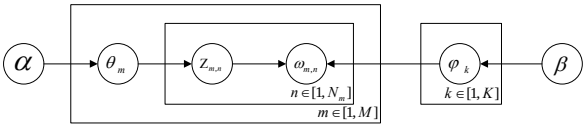

Figure 1: Standard LDA model

where α and β denote the parameters distributed by Dirichlet; *K* denotes the topic numbers; φ_k denotes the vocabulary probability distribution in the topic *k*; *M* denotes the document number; θ_m denotes the topic probability distribution in the document *m*; N_m denotes the length of *m*; $Z_{m,n}$ and $\omega_{m,n}$ denote the topic and the word in *m* respectively.

2.2 Bilingual LDA

Bilingual LDA is a bilingual extension of a standard LDA model. It takes advantage of the document alignment which shares the same topic distribution θ_m and uses different word distributions for each topic (Shown in Fig. 2), where *S* and *T* denote source language and target language respectively.

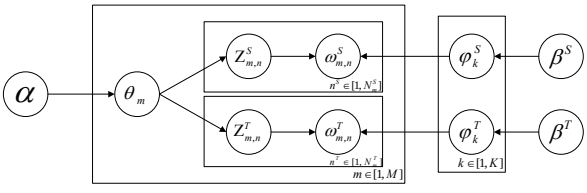

Figure 2: Bilingual LDA model

For each language l ($l \in \{S, T\}$), $Z^l_{m,n}$ and $\omega^l_{m,n}$ are drawn using $Z^l_{m,n} \sim P(Z^l_n \mid \theta_m)$ and $\omega^l_{m,n} \sim P(\omega^l_n \mid Z^l_{m,n}, \varphi^l)$.

Giving the comparable corpora *M*, the distribution $\varphi_{k,v}$ can be obtained by sampling a new token as word *v* from a topic *k*. For new collection of documents $\tilde{M}$, keeping $\varphi_{k,v}$, the distribution $\theta_{\tilde{m}^l,k}$ of sampling a topic *k* from document $\tilde{m}$ can be obtained as follows:

$$P(Z_k \mid \tilde{m}^l) = \theta_{\tilde{m}^l,k} = \frac{n^{(k)}_{\tilde{m}^l} + \alpha_k}{\sum_{k=1}^{K} (n^{(k)}_{\tilde{m}^l} + \alpha_k)}, \quad (1)$$

where $n^{(k)}_{\tilde{m}^l}$ denotes the total number of times that the document $\tilde{m}$ is assigned to the topic *k*.

3 Building comparable corpora

Based on the bilingual LDA model, building comparable corpora includes several steps to generate the bilingual topic model $\varphi_{k,v}$ from the given bilingual corpora, predict the topic distribution $\theta_{\tilde{m}^l,k}$ of the new documents, calculate the similarity of documents and select the largest similar document pairs. The key step is that the document similarity is calculated to align the source language document $\tilde{m}^S$ with relevant target language document $\tilde{m}^T$.

As one general way of expressing similarity, the Kullback-Leibler (KL) Divergence is adopted to measure the document similarity by topic distributions $\theta_{\tilde{m}^S,k}$ and $\theta_{\tilde{m}^T,k}$ as follows:

$$Sim_{KL}(\tilde{m}^S, \tilde{m}^T) = KL[P(Z \mid \tilde{m}^S), P(Z \mid \tilde{m}^T)]$$
$$= \sum_{k=1}^{K} \left[\theta_{\tilde{m}^S,k} \log\left(\theta_{\tilde{m}^S,k} / \theta_{\tilde{m}^T,k} \right) \right]. \quad (2)$$

The remainder section focuses on other two methods of calculating document similarity.

3.1 Cosine Similarity

The similarity between $\tilde{m}^S$ and $\tilde{m}^T$ can be measured by Topic Frequency-Inverse Document Frequency. It gives high weights to the topic which appears frequently in a specific document and rarely appears in other documents. Then the relation between $TFIDF_{\tilde{m}^S,Z}$ and $TFIDF_{\tilde{m}^T,Z}$ is measured by Cosine Similarity (CS).

Similar to Term Frequency-Inverse Document Frequency (Manning et al.,1999), Topic Frequency (TF) denoting frequency of topic Z for the document $\tilde{m}^l$ is denoted by $P(Z \mid \tilde{m}^l)$. Given a constant value λ , Inverse Document Frequency (IDF) is defined as the total number of documents $|\tilde{M}|$ divided by the number of documents

$\left|\tilde{m}^l : P(Z \mid \tilde{m}^l) > \lambda\right|$ containing a particular topic, and then taking the logarithm, which is calculated as follows:

$$IDF = \log \frac{\left|\tilde{M}\right|}{1+\left|\tilde{m}^l : P(Z \mid \tilde{m}^l) > \lambda\right|}. \quad (3)$$

The TFIDF is calculated as follows:

$$TFIDF = TF * IDF$$
$$= P(Z \mid \tilde{m}^l) \log \frac{\left|\tilde{M}\right|}{1+\left|\tilde{m}^l : P(Z \mid \tilde{m}^l) > \lambda\right|}. \quad (4)$$

Thus, the TFIDF score of the topic k over document $\tilde{m}^l$ is given by:

$$TFIDF_{\tilde{m}^l,k}$$
$$= P(Z_k \mid \tilde{m}^l) \log \frac{\left|\tilde{M}\right|}{1+\left|\tilde{m}^l : P(Z_k \mid \tilde{m}^l) > \lambda\right|}$$
$$= \theta_{\tilde{m}^l,k} \log \frac{\left|\tilde{M}\right|}{1+\left|\tilde{m}^l : \theta_{\tilde{m}^l,k} > \lambda\right|}. \quad (5)$$

The similarity between $\tilde{m}^S$ and $\tilde{m}^T$ is given by:

$$Sim_{CS}(\tilde{m}^S, \tilde{m}^T) = Cos(TFIDF_{\tilde{m}^S,Z}, TFIDF_{\tilde{m}^T,Z})$$
$$= \frac{\sum_{k=1}^{K} TFIDF_{\tilde{m}^S,k} TFIDF_{\tilde{m}^T,k}}{\sqrt{\sum_{k=1}^{K} TFIDF^2_{\tilde{m}^S,k}} \sqrt{\sum_{k=1}^{K} TFIDF^2_{\tilde{m}^T,k}}}. \quad (6)$$

3.2 Conditional Probability

The similarity between $\tilde{m}^S$ and $\tilde{m}^T$ is defined as the Conditional Probability (CP) of documents $P(\tilde{m}^T \mid \tilde{m}^S)$ that $\tilde{m}^T$ will be generated as a response to the cue $\tilde{m}^S$.

$P(Z)$ as prior topic distribution is assumed a uniform distribution and satisfied the condition $P(Z_k) = P(Z)$. According to the total probability formula, the document $\tilde{m}^T$ is given as:

$$P(\tilde{m}^T) = \sum_{k=1}^{K} P(\tilde{m}^T \mid Z_k) P(Z_k)$$
$$= P(Z) \sum_{k=1}^{K} P(\tilde{m}^T \mid Z_k). \quad (7)$$

Based on the Bayesian formula, the probability that a given topic Z is assigned to a particular target language document $\tilde{m}^T$ is expressed:

$$P(\tilde{m}^T \mid Z) = \left[P(Z \mid \tilde{m}^T) P(\tilde{m}^T) \right] / P(Z)$$
$$= P(Z \mid \tilde{m}^T) \sum_{k=1}^{K} P(\tilde{m}^T \mid Z_k). \quad (8)$$

The sum of all probabilities $\sum_{k=1}^{K} P(\tilde{m}^T \mid Z_k)$ that all topics Z are assigned to a particular document $\tilde{m}^T$ is a constant Ω, thus equation (8) is converted as follows:

$$P(\tilde{m}^T \mid Z) = \Omega P(Z \mid \tilde{m}^T). \quad (9)$$

According to the total probability formula, the similarity between $\tilde{m}^S$ and $\tilde{m}^T$ is given by:

$$Sim_{CP}(\tilde{m}^S, \tilde{m}^T) = P(\tilde{m}^T \mid \tilde{m}^S)$$
$$= \sum_{k=1}^{K} [P(\tilde{m}^T \mid Z_k) P(Z_k \mid \tilde{m}^S)]$$
$$= \Omega \sum_{k=1}^{K} [P(Z_k \mid \tilde{m}^T) P(Z_k \mid \tilde{m}^S)]$$
$$= \Omega \sum_{k=1}^{K} [\theta_{\tilde{m}^S,k} \theta_{\tilde{m}^T,k}]. \quad (10)$$

4 Experiments and analysis

4.1 Datasets and Evaluation

The experiments are conducted on two sets of Chinese-English comparable corpora. The first dataset is news corpora with 3254 comparable document pairs, from which 200 pairs are randomly selected as the test dataset *News-Test* and the remainder is the training dataset *News-Train*. The second dataset contains 8317 bilingual Wikipedia entry pairs, from which 200 pairs are randomly selected as the test dataset *Wiki-Test* and the remainder is the training dataset *Wiki-Train*. Then *News-Train* and *Wiki-Train* are merged into the training dataset *NW-Train*. And the hand-labeled gold standard namely *NW-Test* is composed of *News-Test* and *Wiki-Test*.

Braschler et al. (1998) used five levels of relevance to assess the alignments as follows: Same Story, Related Story, Shared Aspect, Common Terminology and Unrelated. The paper selects the documents with Same Story and Related Story as comparable corpora. Let C_p be the comparable corpora in the building result and C_l be the comparable corpora in the labeled result. The Precision (P), Recall (R) and F-measure (F) are defined as:

$$P = \frac{\left|C_p \cap C_l\right|}{\left|C_p\right|}, R = \frac{\left|C_p \cap C_l\right|}{\left|C_l\right|}, F = \frac{2PR}{P+R}. \quad (11)$$

4.2 Results and analysis

Two groups of validation experiments are set with sampling frequency of 1000, parameter α

of 50/K, parameter β of 0.01 and topic number K of 600.

Group 1: Different data source

We learn bilingual LDA models by taking different training datasets. The performance of three approaches (KL, CS and CP) is examined on different test datasets. Tab. 1 demonstrates these results with the winners for each algorithm in bold.

Train	*Test*	*KL*		*CS*		*CP*	
		P	***F***	***P***	***F***	***P***	***F***
News	***News***	0.62	0.52	0.73	0.59	0.69	0.56
News	***Wiki***	0.60	0.47	0.68	0.56	0.66	0.52
Wiki	***News***	0.61	0.48	0.71	0.58	0.68	0.55
Wiki	***Wiki***	0.63	0.50	0.75	0.60	0.71	0.59
NW	***NW***	0.66	**0.55**	0.76	**0.62**	0.73	**0.60**

Table 1: Sensitivity of Data Source

The results indicate the robustness and effectiveness of these algorithms. The performance of algorithms on *Wiki-Train* is much better than *News-Train*. The main reason is that *Wiki-Train* is an extensive snapshot of human knowledge which can cover most topics talked in *News-Train*. The probability of vocabularies among the test dataset which have not appeared in the training data is very low. And then the document topic can effectively concentrate all the vocabularies' expressions. The topic model slightly faces with the problem of knowledge migration issue, so the performance of the topic model trained by *Wiki-Train* shows a slight decline in the experiments on *News-Test*.

CS shows the strongest performance among the three algorithms to recognize the document pairs with similar topics. CP has almost equivalent performance with CS. Comparing the equation (5) and (6) with (10), we can find out that CP is similar to a simplified CS. CP can improve the operating efficiency and decrease the performance. The performance achieved by KL is the weakest and there is a large gap between KL and others. In addition, the shortage of KL is that when the exchange between the source language and the target language documents takes place, different evaluations will occur in the same document pairs.

Group 2: Existing Methods Comparison

We adopt the *NW-Train* and *NW-Test* as training set and test set respectively, and utilize the CS algorithm to calculate the document similarity to verify the excellence of methods in the study. Then we compare its performance with the existing representative approaches proposed by Thuy et al. (2009) and Preiss (2012) (Shown in Tab. 2).

Algorithm	***P***	***R***	***F***
Thuy	0.45	0.32	0.37
Preiss	0.67	0.44	0.53
CS	0.76	0.53	**0.62**

Table 2: Existing Methods Comparison

The table shows CS outperforms other algorithms, which indicates that bilingual LDA is valid to construct comparable corpora. Thuy et al. (2009) matches similar documents in the view of inter-translated vocabulary and co-occurrence information features, which cannot understand the content effectively. Preiss (2012) uses monolingual training dataset to generate topic model and translates source language topic model into target language topic model respectively. Yet the translation accuracy constrains the matching effectiveness of similar documents, and the cosine similarity is directly used to calculate document-topic similarity failing to highlight the topic contributions of different documents.

5 Conclusion

This study proposes a new method of using bilingual topic to match similar documents. When CS is used to match the documents, TFIDF is proposed to enhance the topic discrepancies among different documents. The method of CP is also addressed to measure document similarity.

Experimental results show that the matching algorithm is superior to the existing algorithms. It can utilize comprehensively large scales of document information in training set to avoid the information deficiency of the document itself and over-reliance on bilingual knowledge. The algorithm makes the document match on the basis of understanding the document. This study does not calculate similar contents existed in the monolingual documents. However, a large number of documents in the same language describe the same event. We intend to incorporate monolingual document similarity into bilingual topics analysis to match multi-documents in different languages perfectly.

Acknowledgments

The work is supported by the National Natural Science Foundation of China under No. 61070099 and the project of MSR-CNIC Windows Azure Theme.

References

AbduI-Rauf S, Schwenk H. On the use of comparable corpora to improve SMT performance[C]//Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2009: 16-23.

Ji H. Mining name translations from comparable corpora by creating bilingual information networks[C] // Proceedings of BUCC 2009. Suntec, Singapore, 2009: 34-37.

Braschler M, Schauble P. Multilingual Information Retrieval based on document alignment techniques[C] // Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries. Heraklion, Greece. 1998: 183-197.

Tao Tao, Chengxiang Zhai. Mining comparable bilingual text corpora for cross-language information integration[C] // Proceedings of ACM SIGKDD, Chicago, Illinois, USA. 2005:691-696.

Talvensaari T, Laurikkala J, Jarvelin K, et al. Creating and Exploiting a Comparable Corpus in Cross-Language Information Retrieval[J]. ACM Transactions on Information Systems. 2007, 25(1): 322-334.

Thuy Vu, Ai Ti Aw, Min Zhang. Feature-based method for document alignment in comparable news corpora[C] // Proceedings of the 12th Conference of the European Chapter of the ACL, Athens, Greece. 2009: 843-851.

Otero P G, L'opez I G. Wikipedia as Multilingual Source of Comparable Corpora[C] // Proceedings of the 3rd Workshop on BUCC, LREC2010. Malta. 2010: 21-25.

Li B, Gaussier E. Improving corpus comparability for bilingual lexicon extraction from comparable corpora[C]//Proceedings of the 23rd International Conference on Computational Linguistics. Association for Computational Linguistics, 2010: 644-652.

Judita Preiss. Identifying Comparable Corpora Using LDA[C]//2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Montre´al, Canada, June 3-8, 2012: 558-562.

Mimno D, Wallach H, Naradowsky J et al. Polylingual topic models[C]//Proceedings of the EMNLP. Singapore, 2009: 880-889.

Vulic I, De Smet W, Moens M F, et al. Identifying word translations from comparable corpora using latent topic models[C]//Proceedings of ACL. 2011: 479-484.

Ni X, Sun J T, Hu J, et al. Mining multilingual topics from wikipedia[C]//Proceedings of the 18th international conference on World wide web. ACM, 2009: 1155-1156.

Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. the Journal of machine Learning research, 2003, 3: 993-1022.

Griffiths T L, Steyvers M. Finding scientific topics[J]. Proceedings of the National academy of Sciences of the United States of America, 2004, 101: 5228-5235.

Manning C D, Schütze H. Foundations of statistical natural language processing[M]. MIT press, 1999.