Identifying Word Translations in Scientific Literature based on Labeled Bilingual Topic Model and Co-occurrence Features

Mingjie Tian, Yahui Zhao and Rongyi Cui*

Intelligent Information Processing Lab.,
Department of Computer Science and Technology, Yanbian University,
Yanji 133002, China
iipmjtian@qq.com
{yhzhao, cuirongyi}@ybu.edu.cn

Abstract. Aiming at the increasingly rich multi language information resources and multi-label data in scientific literature, in order to mining the relevance and correlation in languages, this paper proposed the labeled bilingual topic model and co-occurrence feature based similarity metric which could be adopted to the word translation identifying task. First of all, it could assume that the keywords in the scientific literature are relevant to the abstract in the same article, then extracted the keywords and regard it as labels, labels with topics are assigned and the "latent" topic was instantiated. Secondly, the abstracts in article were trained by the labeled bilingual topic model and got the word representation on the topic distribution. Finally, the most similar word between both languages was matched with similarity metric proposed in this paper. The experiment result shows that the labeled bilingual topic model reaches better precision than "latent" topic model based bilingual model, and co-occurrence features enhance the attractiveness of the bilingual word pairs to improve the identifying effects.

Keywords: Topic Model, Label, Co-occurrence Features, Word Translations.

1 Introduction

Without any doubt, the Web is growing rapidly, which can be reflected by the amount of online Web content. One challenging but very desirable task accompanying the Web growth is to organize information written in different languages, to make them easily accessible for all users. Recently there has been a new trend that from news to scientific literature, a significant proportion of the world's textual data is labeled with multiple human-provided tags. This trend allows us to understand the content of the document with a more detailed dimension.

The diversity of language has enriched information resources, but differences between languages have inevitably hindered users to use them. Take word translation task

* corresponding author

as an example, it would be very helpful to leverage the word w_1 in source language for finding the translation w_2 in target language. This work is called word alignment and many research works has been conducted to solve this problem. [1, 2, 3] acquired translation candidates based on methods that need an initial lexicon of translations, cognates or similar words to extend additional translations from the context. Some attempts of obtaining translations using cross-lingual topic models have been made in the last few years. [4] builded topics as distributions over bilingual matchings where matching priors may come from different initial evidences such as machine readable dictionary. The main shortcoming of method above is that it introduces external knowledge. [5, 6, 7] proposed cross-lingual topic models to solve cross-lingual text classification, machine translation and event detection task, [8] made use of these cross-lingual topic models to obtain word translation in bilingual corpus and proposed word similarity standard *TF-ITF* (term frequency – inverse topic frequency) to improve the matching effects.

We combine the multi-label in scientific literature with the application of topic models word translation task, and propose the Labeled Bilingual Topic Model (LBTM) to solve the implicit meaning of the topics in "latent" topic models, make the topic has defined semantic and be better explained. And we propose a new similarity metric based on co-occurrence features, combine word's feature in topic and corpus level to enhance the translation word's relevance.

2 Methodology

2.1 Latent Dirichlet Allocation

The LDA topic model is a generative model for documents proposed by Blei [9]. It is also known as a three levels Bayesian probability model, which contains three levels of word, topic and document. The LDA topic model introduce Dirichlet prior parameters [10] on the distribution of document-topic and topic-word, which solves the over-fitting problem that occurs when processing the large-scale corpus.

The LDA topic model is a directed probabilistic graphical model [11], as is shown in Fig. 1.

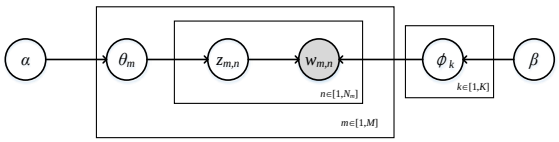

Fig. 1. Probabilistic Graphical Model of LDA

Given document dataset D contains number of M documents, the m-th document contains number of N_m words, assuming that the number of topics in document dataset D is K totally, the generative process can be described by the following steps:

1. For each document, get the length $N \sim Poission(\xi)$
   a. Choose the distribution over topics $\theta \sim Dir(\alpha)$

2. For each position n (n = 1 to N_m) in document
   a. Choose topic $z_n \sim Multinomial(\theta)$
   b. Choose word w_n with probability $p(w_n \mid z_n, \beta)$

Here, $Poission(\cdot)$, $Dir(\cdot)$ and $Multinomial(\cdot)$ represent possion, dirichlet and multinomial distributions respectively.

The parameters should be estimated during the modelling of the dataset through LDA topic model, the commonly used methods is Variational Bayesian Inference [12], EM algorithm [13] or Collapsed Gibbs Sampling [14, 15]. The method based on Collapsed Gibbs Sampling can effectively sample topics from large-scale document dataset. The parameter estimation process can be considered as the inverse process of the document generation, the parameters are estimated with the documents distribution have been known. The conditional probability of the topic sequence under the known word sequence is calculated as follow.

$$\begin{aligned} & p(z_i = k \mid \overrightarrow{z_{\neg i}}, \vec{w}) \\ & = \frac{p(\vec{w}, \vec{z})}{p(\vec{w}, \overrightarrow{z_{\neg i}})} = \frac{n_{k,\neg i}^{t} + \beta_t}{\sum_{v=1}^{V} (n_k^v + \beta_v) - 1} \cdot \frac{n_{m,\neg i}^{k} + \alpha_k}{\sum_{l=1}^{L} (n_m^l + \alpha_l) - 1} \end{aligned} \tag{1}$$

2.2 Labeled Bilingual Topic Model

The LDA topic model replace the high-dimensional term by the low-dimensional "latent" topic to capture the semantic information of the document. However, each "latent" topic is implicit defined and lack of explanations. We make flexible use of the multi-label data in scientific literature or news such as keywords in the literature to improve the LDA topic model, and propose the Labeled Bilingual Topic Model. Regard the labels in the literature or news as the "topic", make the topics can be explained, instantiate the "latent" topics and assign an explicit meaning. After modelling the document dataset by the model proposed by us, the documents are represented by the meaningful topics. Each word in the document has a probability distribution on the topics, and can be represented as a vector in the vector space to implement word translations identifying.

The model

Assume that the document dataset consists of M documents, the content of each document is described by two languages L_1 and L_2, the content described by each language is the same. The model has a set of language-independent "common" topics to describe the two languages documents, each "common" topic has two different representations, each corresponding to the each language. The probabilistic graphical model of the Labeled Bilingual Topic Model is shown in Fig. 2.

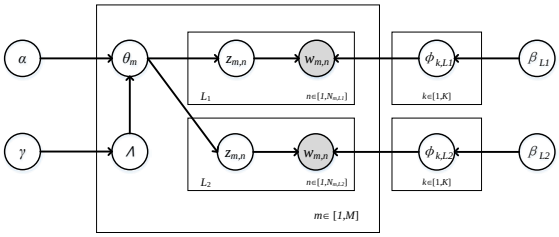

Fig. 2. Probabilistic Graphic Model of LBTM

where α is the hyper parameter of Dirichlet distribution between document and topics, β_{Lj} is the hyper parameter of Dirichlet distribution between topic and words in language L_j (j = 1 , 2), Λ is the constraint relationship between document and topics, and the relationship between document and topics are specific , γ is the hyper parameter of the Bernoulli distribution constrained by the document and topics, Λ and γ are given empirically. $w_{m,n}$ represents the n-th word in the m-th document, $z_{m,n}$ is the topic corresponding to the word $w_{m,n}$, the parameter θ_m is the distribution of the m-th document on the topic, the parameter $\phi_{k,Lj}$ is the distribution of the k-th topic on the language L_j's word; Given document dataset D contains number of M documents, the m-th document's language L_j's part contains number of N_m, L_j words, assuming that the number of topics in document dataset D is K totally, the generative process of Labeled Bilingual Topic Model can be described by the following steps:

1. For each "common" topic z, z=1,2,…,K
   a. For each language L_j, j = 1 or 2
      (1) Choose topic's distribution on words $\phi_{z,L_j} \sim Dir(\beta_{L_j})$
2. For the document m in document dataset
   a. Choose document-topic constraint $\Lambda_{m,k} \in \{0,1\} \sim bernoulli(\gamma)$
   b. Choose the distribution over "common" topics $\theta_m \sim Dir(\alpha) \mid \Lambda$
3. For each position n (n = 1 to N_m) in document
   a. Choose topic $z_{m,n} \sim Multinomial(\theta_m)$
   b. Choose word $w_{m,n,L_j} \sim Multinomial(\phi_{m,n,L_j})$

Estimation

In parameter estimation phase, the Gibbs Sampling method should be modified to fit the bilingual and labeled features. The conditional probability of the topic sequence under the word sequence is extended from monolingual to bilingual. The conditional probability can be calculated as follow.

$$p(z_{i,L_j} = k \mid \vec{z}_{\neg i,L_j}, \vec{w}_{L_j})$$
$$= \frac{n^t_{k,\neg i,L_j} + \beta^t_{L_j}}{\sum_{v=1}^{V_{L_j}} (n^v_{k,L_j} + \beta^v_{L_j}) - 1} \cdot \frac{\sum_{j=1}^{2} n^k_{m,\neg i,L_j} + \alpha_k}{\sum_{l=1}^{L} (\sum_{j=1}^{2} n^l_{m,L_j} + \alpha_l) - 1} \tag{2}$$

Where $n^t_{k,\neg i,L_j}$ is the number of times word t of language L_j assigned to topic k except t's current assignment, $\sum_{v=1}^{V_{L_j}} n^v_{k,L_j} - 1$ is the total number of words in language L_j assigned to topic k except t's current assignment, V_{Lj} is the vocabulary of language L_j, $n^k_{m,\neg i,L_j}$ is the number of words in language L_j in document m assigned to topic k except t's current assignment, $\sum_{l=1}^{L}\sum_{j=1}^{2} n^l_{m,L_j} - 1$ is the total number of words in all languages in document m except the current word t. Finally, we can obtain word's distribution on the topics φ as follow.

$$\varphi_{t,k,Lj} = \frac{n^t_{k,L_j} + \beta^t_{L_j}}{\sum_{v=1}^{V_{L_j}} (n^v_{k,L_j} + \beta^v_{L_j})} \tag{3}$$

Differences with "Latent" Topic Model

Compare to the multi-lingual models [5, 6, 7] derived from "latent" topic model [9], Labeled Bilingual Topic Model takes advantage of the multi-label data in the documents, instantiate the "latent" topics so that the meaning of the topic is no longer "implicit", but "explicit". The differences with the method derived from "latent" topic model are as follows:

1. Topic size K: The determination of the number of topics is one of the difficulties in the LDA topic model, the value of K needs to be choose from the experimental results. The number of topics in LBTM is determined, that is, the number of unique labels in the dataset;
2. Vector representation of the word and document: The topic sampling range for each word in each document at "latent" topic model is 1~K, documents are represented by the assignment of word and topic in the document, therefore, the value of each topic component in the vector representation of the document and word may not be 0. In LBTM, each document has a constraint with the fixed labels, and the words in the document are also have a constraint with the labels, the topic component value that is constrained to the document which contains the word is not 0, and the rest must be 0.
3. Range of topic sampling: The conditional probability between each word in each document and all topics need to be calculated in each iteration. Due to no constraint between the document and the topic in "latent" topic model, the range of topic sampling for each word in each document is K. In LBTM, there is a fixed constraint

relationship between each document and topics, the range of topic sampling of word in a document is a collection of topics (labels) that have a constraint with the document;

4. Sampling computational complexity: Due to the range of sampling, the "latent" topic model needs to calculate the conditional probability between each word and all the topics during each iteration. In LBTM, it is only necessary to calculate the conditional probabilities each word with topics that have fixed constraint with the document. The model proposed by us has advantages in the computational efficiency at sampling process.

3 Similarity Metrics

3.1 Cosine Similarity

The cosine similarity is a measure difference by using the cosine of the angle between two vectors, the smaller the angle between vectors, the more similar the two vectors. Comparing to Euclidean distance, the cosine similarity focus on the difference in the directions of the two vectors, and the cosine similarity has better robustness to the stretching transformation of the vector. The cosine similarity of words representation on topics are as follow.

$$cos(w_{i,L1}, w_{j,L2}) = \frac{\varphi_{i,L1} \cdot \varphi_{j,L2}}{||\varphi_{i,L1}|| \ \ ||\varphi_{j,L2}||} \tag{4}$$

Where $w_{i,L1}$ is *i*-th word in language L_1 and $w_{i,L2}$ is *j*-th word in language L_2, and $\varphi_{i,L1}$ is the distribution of *i*-th word in language L_1 on topics and $\varphi_{j,L2}$ is the distribution of *j*-th word in language L_2 on topics.

3.2 TF-ITF

Author in [8] borrowed an idea from information retrieval and constructs word vectors over a shared latent topic space. Values within vectors are the *TF-ITF*(term frequency – inverse topic frequency) scores which are calculated in a completely analogical manners as the *TF-IDF* scores for the original word-document space[16]. Given *i*-th word in language L_j , $n^i_{k,Lj}$ denotes the number of times the *i*-th word is associated with a topic *k*. *TF-ITF* score for the *i*-th word in language L_j and topic *k* is calculated as follow.

$$TF-ITF_{i,k} = \frac{n^i_{k,Lj}}{\sum_{v=1}^{V_{Lj}} n^v_{k,Lj}} \cdot log \frac{K}{1+|k : n^i_{k,Lj} > 0|} \tag{5}$$

After words in both languages are represented by *TF-ITF*, the standard cosine similarity metric is then used to find the most similar word vectors from the target vocabulary for a source word vector.

3.3 Co-occurrence Features

In addition to the LBTM to improve the word translation task, we also introduce a similarity metric based on co-occurrence features.

The corpus-based co-occurrence word acquisition method is based on the distributed hypothesis [17], it is based on large-scale corpus and represents the distribution of words in each document as vectors. Finally co-occurrence words is selected by calculating the correlation between the vectors.

If two words in corpus are usually occurred in the same document, the two words can be considered semantically related to each other. The concept of co-occurrence words is mostly applied to the query expansion of information retrieval, when a document is related to the query requirements but does not contain query terms, the query can be expanded by the query co-occurrence word as related information [18]. Method in [19] applied the co-occurrence words to calculate the similarity of cross-lingual documents.

We combined traditional similarity metrics with word co-occurrence to maximize the similarity between the word translation pairs. The similarity combine with co-occurrence features between words in different languages is as follow:

$$sim(w_{i,L1}, w_{j,L2,}) = v_{i,L1} \cdot v_{j,L2} + \lambda \log(m^{i,j} + 1) \tag{6}$$

Where $w_{i,L1}$ denotes *i*-th word in language L_1 and $w_{j,L2}$ denotes *j*-th word in language L_2, $v_{j,L2}$ is the vector representation of word $w_{j,L2}$ and could be topic distributions or *TF-ITF*. $m^{i,j}$ denotes the number of documents that the words $w_{i,L1}$ and $w_{j,L2}$ co-occurred.

The similarity of the previous item to the right of the equation is the distribution of the word in the topic level, and value of the latter item is the co-occurrence degree between the words at the corpus level. We believe that merging the similarity of different level will take both advantages to improve the word translation precision and it will be proved in the next experiment.

4 Experiments

To verify the validity and feasibility of LBTM and co-occurrence based similarity metric, we carried out word translation identifying experiment. We compare alignment approach with the method proposed in [8], the set of approach contains bilingual topic models and similarity metrics.

4.1 Dataset

The bilingual corpus used in the experiment is the parallel corpus of Chinese-Korean scientific literature. The dataset contains keywords and Abstracts of 2427 Aerospace domain Chinese-Korean parallel scientific literature with the sentence level aligned. In order to reduce data sparsity, we keep only lemmatized nouns and verbs forms for further analysis. Our Chinese vocabulary consists of 20608 terms, while our Korean vocabulary contains 18391 terms. The subset of the 1867 most frequent word translation pairs was used for testing. The sample of parallel corpus is shown in Fig.3.

Title-一种二元定几何混压式超声速进气道流场控制概念研究
Abstract-针对二元定几何混压式超声速进气道低马赫数时流量系数低加速性能差的问题,提出了一种新的泄流槽流场控制概念,并通过数值仿真,揭示了泄流槽控制激波结构机理及其主要几何参数对进气道性能的影响规律.研究结果表明:采用该流场控制方案可通过泄流槽入口处的波系结构使进气道在低于设计马赫数时的出口总压恢复系数和流量系数相对于原型方案均得到明显提高,而在设计点关闭泄流槽后进气道的性能与原型进气道基本相当,这对改善冲压发动机在低马赫数转级后的加速性能是有利的.
Keywords-航空、航天推进系统; 冲压发动机; 二元超声速进气道; 流场控制; 泄流槽

Title- 2 차원 고정 형상 혼합-압력식 초음속 흡입구의 유동장 제어에 관한 연구
Abstract-본 논문에서는 2 차원 고정 형상 혼합-압력식 초음속 흡입구가 저 마하수(Mach number)에서 흐름 계수의가속 성능이 낮은 문제점을 해결하기 위하여 바이패스 출구로 유동장(flow field)을 제어하는 새로운 기법을 제안하였다.수치해석을 통하여 바이패스 출구가 충격파를 제어하는 메커니즘과 기하학적 파라미터가 흡입구 성능에 주는 영향을 연구하였다.연구 결과를 통하여,제안한 유동장 제어 기법은 바이패스 출구 입구의 파형 구조를 통하여 흡입구의 설계 마하수보다 낮은 유동 속도에서 출구 절대 압력 회복 계수와 흐름 유량 계수가 현저하게 증가하는 것을 확인하였다.바이패스 출구가 닫힌 후, 흡입구의 성능은 원형(Prototype) 흡입구 성능과 같은 특성을 보였으며, 이는 램제트 엔진의 저 마하수 가속 성능에 유리하다는 것을 보여준다.
Keywords-항공 우주 추진 시스템; 램제트 엔진; 2 차원 초음속 흡입구; 유동장 제어; 바이패스 출구

Fig. 3. Chinese-Korean Parallel Scientific Literature

4.2 Parameters Setting

The size of topics K, Dirichlet hyper parameters α and β, training and test iterations need to be determined in advance. The topic numbers of LBTM is fixed, and we choose four kinds of topic size to the comparative models. The parameters of LBTM and the "latent" topic model based bilingual model are shown in Table 1.

Table 1. Comparative Experiment Parameters Setting

Parameters	**Comparative Model**	
	LBTM	***Bi-LDA model***
K	8718	200/400/600/800
α	50/K	50/K
β	0.01	0.01

Parameters	Comparative Model	
	LBTM	***Bi-LDA model***
Training iterations	1000	1000
Test iterations	100	100

4.3 Similarity Metrics

We divide the similarity metrics into two types, single kind of similarity method and hybrid method.

The single similarity method includes cosine similarity, *TF-ITF* proposed in [8] and co-occurrence between words of both language that independent with the topic model's result.

In order to verify that similarity metric combined with co-occurrence features has advantages over the single similarity method, we combined the cos and *TF-ITF* with co-occurrence features, namely cos + co-occurrence and *TF-ITF* + co-occurrence. At the same time, we also compare our metrics with the *TF-ITF* + *TF-IDF* method which obtained the highest result in [8].

4.4 Results and Analysis

Table 2 shows some example of topics produced by LBTM and bilingual"Latent" topic models with K=200,400,600 and 800. Each topic has two representations: first corresponds with the distribution of Chinese words and the second line is associated with Korean words distribution. Words on each line are ranked by probability score in decreasing order. In LBTM, we select topic "涡扇发动机" (turbofan engine), as a fair comparison, we extract the two topics that Chinese word "涡扇发动机" and Korean word "터보팬엔진" (turbofan engine)most frequently appear in each bilingual "Latent" topic models. It can be found that LBTM could catch the words related to the topic "涡扇发动机", even the word "涡扇发动机", 터보팬(turbofan) and 엔진(engine).

Table 2. Sample of Topics

Model	Topic	Words assigned with Topic
LBTM	涡扇发动机 터보팬엔진 (TurboFan Engine)	涡 扇 发 动 机 (turbofan engine) 发 动 机 (engine) 压 气 机 (compressor) 转速(rotating speed) 高压(high pressure)
		터보팬(turbofan) turbofan(turbofan) engine(engine) 압축（compression) 엔진(engine)
Bi-LDA Topic 200	86th Topic	湍流(turbulent flow) 湍流模型(turbulent flow model) 对比(compare) 流动(flow) 机匣(casing)
		난류(turbulence) 모델(model) 결과(result) 평균(average) 비교(compare)
	42th Topic	工艺(technics) 焊缝(weld seam) 接头(connect) 对接(butt) 焊接接头(welded joint)
		용접(welding) 접합(joint) 이음(connection) 결과(result) 더블(double)

Bi-LDA Topic 400	213th Topic	要求(demand) 满足(satisfaction) 需求(requirement) 使用(use) 能够(can)
		요구(demand) 만족(satisfaction) 대한(about) 충족(satisfy) 요구사항(requirement)
	361th Topic	变(change) 高(high) 简单(simple) 调节(adjust) 可变(variable)
		가변(variable) variable(variable) 롤링(rolling) 탐색(quest) rolling(rolling)
Bi-LDA Topic 600	372th Topic	吸气(sunction) 对转压气机(Counter rotating compressor) Rotor(rotor) 量(quantity) 效(effect)
		독립(independent) 이중반전(double inversion) 압축기(compressor) Rotor(Rotor) 샘플(sample)
	87th Topic	发动机(engine) 航空(aviation) 型(type) 工作(working) 加力(afterburner)
		엔진(engine) 항공기(aircraft) engine(engine) 터보팬(turbofan) 추진력(propulsion)
Bi-LDA Topic 800	771th Topic	压缩(compression) 压缩强度(compressive strength) 记忆合金环(memory alloy ring) 压缩性能(compression performance) 下降(descent)
		압축(compression) compression(compression) 파이프(pipe) 실효성(effectiveness) compressive(compressive)
	562th Topic	加力(afterburner) 起动(start-up) 飞行包线(flight envelope) 调节规律(regulation law) 加速（Acceleration)
		토대(foundation) 조정(adjust) 비행영역선도(flight area line) off(off) 배기(exhaust)

Table 3 shows the Precision@1 scores (the percentage of words where the first word from the list of translation is the correct one) for all similarity metrics, for different number of topics K in bilingual-LDA and LBTM.

Table 3. Precision@1 Scores for Experiment

Model	Single metrics			Hybrid Metrics			Time Cost (hours)
	cos	***TF-ITF***	***co-occur***	***TF-ITF + TF-IDF***	***TF-ITF + co-occur***	***cos + co-occur***	
Topic200	24.64%	24.42%	55.60%	36.31%	36.58%	36.80%	16.62
Topic400	22.28%	22.12%		33.05%	35.99%	37.06%	34.74
Topic600	24.26%	23.03%		34.33%	37.65%	38.35%	50.82
Topic800	25.01%	23.94%		35.62%	38.73%	36.69%	68.64
LBTM	**55.17%**	**50.99%**		**64.69%**	**65.13%**	**68.61%**	**6.06**

LBTM achieves the highest percentage 68.61% in cos + co-occurrence similarity metrics. Compared with the method in [8], LBTM has a higher precision@1 than it in all similarity metrics. We instantiated the “latent” topic means that the topic has a fixed constraint with the documents, it also means that the sampling range of topics assign to the word in the document is fixed. There is a certain relationships between words and topics at the whole corpus level, in other words, the word will must assigned to the one of the certain topic collections. In the vector space model with the topics as the dimensions, the value of each word under these topics are greatly increased, also the probability between word translation pair is increased.

Our model took 6.06 hours to train model, with the method in [8] took 16.62 hours at least. The reason for saving so much time is that at the training phase, the range of topic sampling in LBTM is given, which is generally the number of keywords (4~6) in

per scientific literature, but in comparative experiment, there is no constraint between the document and the topics, the conditional probabilities with all topics should be calculated in whole phase, which increases the computational complexity.

Whether it is LBTM or "latent" topic based bilingual topic model, comparing with the single metrics, the hybrid metrics are greatly improved at least 24%. There are great improvement when cosine similarity or *TF-ITF* combine with the co-occurrence features. Comparing the promotion between co-occurrence and *TF-IDF* to the *TF-ITF*, it could be found the effects are almost the same. Finally, it could be concluded that when the features of the word at corpus level (*TF-IDF* or co-occurrence) are combined with the features of the topic level (Cosine similarity or *TF-ITF*), the combination could play a great synergy. In addition, in the same environment, we can find that whether in single metrics or in hybrid metrics, cosine similarity has better matching result than *TF-ITF*.

5 Conclusion

In this paper, firstly, we combine the multi-label in scientific literature with the application of topic models in word translation identifying, propose the Labeled Bilingual Topic Model (LBTM). Compared with the "latent" topic model, we instantiate the "latent" topics so that the meaning of the topic is no longer "implicit" but "explicit". In the phase of training parameters, due to the given range of topics sampling, in terms of efficiency, LBTM is superior to the "latent" topic model based bilingual LDA model. The result in word translation task indicate that LBTM could reach higher efficiency and precision than "latent" topic model based bilingual LDA model. Secondly, we propose a new similarity metrics that combine bilingual word's co-occurrence features with traditional similarity metrics. Compare with single metrics, the hybrid metrics improved the precision of word translation identifying at least 24%. Limited by the corpus, it can only be applied to the bilingual topic modelling at this stage. If there are multilingual multi-labeled parallel corpus in more than two language, we will extend our model to the multi-lingual.

6 Acknowledgement

This research was financially supported by State Language Commission of China under Grant No. YB135-76.

References

1. Diab, M. T., Finch, S.: A statistical translation model using comparable corpora. In: Proceedings of the 2000 Conference on Content-Based Multi-media Information Access, pp. 1500–1508. (2000).
2. Koehn, P., Knight, K.: Learning a translation lexicon from monolingual corpora. In: Proceedings of the ACL-02 Workshop on Unsupervised Lexical Acquisition, vol. 9, pp. 9–16. ACL, Stroudsburg (2002).

3. Gaussier, E., Renders, J. M., Matveeva, I., Goutte, C., Déjean, H.: A geometric view on bilingual lexicon extraction from comparable corpora. In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, pp. 526–533. ACL, Stroudsburg (2004).
4. Boyd-Graber, J., Blei, D. M.: Multilingual topic models for unaligned text. In: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, pp. 75–82. AUAI Press, Arlington (2009).
5. Ni, X., Sun, J. T., Hu, J., Chen, Z.: Mining multilingual topics from Wikipedia. In: Proceedings of the 18th International World Wide Web Conference, pp. 1155–1156. ACM, New York (2009).
6. Mimno, D., Wallach, H. M., Naradowsky, J., Smith, D. A., McCallum, A.: Polylingual topic models. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pp. 880–889. ACL, Stroudsburg (2009).
7. De Smet, W., Moens, M. F.: Cross language linking of news stories on the web using interlingual topic modelling. In: Proceedings of the 2nd ACM Workshop on Social Web Search and Mining, pp. 57–64. ACM, New York (2009).
8. Vulić, I., De Smet, W., Moens, M. F.: Identifying word translations from comparable corpora using latent topic models. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers, vol. 2, pp. 479–484. ACL, Stroudsburg (2011).
9. Blei, D. M., Ng, A. Y., Jordan, M. I.: Latent dirichlet allocation. Journal of Machine Learning Research 3(Jan), pp.993–1022 (2003).
10. Qian, X. U., Zhou, J., Chen, J.: Dirichlet Process and Its Applications in Natural Language Processing. Journal of Chinese Information Processing 23(5), 25–33 (2009).
11. Xu, G., Wang, H. F.: The development of topic models in natural language processing. Chinese Journal of Computers 34(8), 1423-1436 (2011).
12. Fang A., Macdonald C., Ounis I., Habel P., Yang X.: Exploring Time-Sensitive Variational Bayesian Inference LDA for Social Media Data. In: Jose J. et al. (eds) Advances in Information Retrieval. ECIR 2017, LNCS, vol. 10193, pp.252–265. Springer, Cham (2017).
13. Aiping, W., Gongying, Z., Fang, L.: Research and application of EM algorithm. Computer Technology and Development 19(9), 108-110 (2009).
14. Heinrich, G.: Parameter estimation for text analysis. Technical Report, (2008).
15. Yerebakan, H. Z., Dundar, M.: Partially collapsed parallel Gibbs sampler for Dirichlet process mixture models. Pattern Recognition Letters 90, 22–27 (2017).
16. Manning, C. D., Schutze, H.: Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA, USA. (1999).
17. Goodstein, R. L., Harris, Z.: Mathematical structures of language. Mathematical Gazette 54(388), 173 (1970).
18. Bajpai, P., Verma, P.: Improved Query Translation for English to Hindi Cross Language Information Retrieval. Indonesian Journal of Electrical Engineering and Informatics 4(2), 134–140 (2016).
19. Liu, J., Cui R. Y., Zhao, Y. H.: Cross-lingual Similar Documents Retrieval Based on Co-occurrence Projection. In: Proceedings of the 6th International Conference on Computer Science and Network Technology, pp. 11–15. IEEE (2017).