# Cross-Lingual Latent Topic Extraction

**Duo Zhang**
University of Illinois at Urbana-Champaign
dzhang22@cs.uiuc.edu

**Qiaozhu Mei**
University of Michigan
qmei@umich.edu

**ChengXiang Zhai**
University of Illinois at Urbana-Champaign
czhai@cs.uiuc.edu

## Abstract

Probabilistic latent topic models have recently enjoyed much success in extracting and analyzing latent topics in text in an unsupervised way. One common deficiency of existing topic models, though, is that they would not work well for extracting cross-lingual latent topics simply because words in different languages generally do not co-occur with each other. In this paper, we propose a way to incorporate a bilingual dictionary into a probabilistic topic model so that we can apply topic models to extract shared latent topics in text data of different languages. Specifically, we propose a new topic model called Probabilistic Cross-Lingual Latent Semantic Analysis (PCLSA) which extends the Probabilistic Latent Semantic Analysis (PLSA) model by regularizing its likelihood function with soft constraints defined based on a bilingual dictionary. Both qualitative and quantitative experimental results show that the PCLSA model can effectively extract cross-lingual latent topics from multilingual text data.

## 1 Introduction

As a robust unsupervised way to perform shallow latent semantic analysis of topics in text, probabilistic topic models (Hofmann, 1999a; Blei et al., 2003b) have recently attracted much attention. The common idea behind these models is the following. A topic is represented by a multinomial word distribution so that words characterizing a topic generally have higher probabilities than other words. We can then hypothesize the existence of multiple topics in text and define a generative model based on the hypothesized topics. By fitting the model to text data, we can obtain an estimate of all the word distributions corresponding to the latent topics as well as the topic distributions in text. Intuitively, the learned word distributions capture clusters of words that co-occur with each other probabilistically.

Although many topic models have been proposed and shown to be useful (see Section 2 for more detailed discussion of related work), most of them share a common deficiency: they are designed to work only for mono-lingual text data and would not work well for extracting cross-lingual latent topics, i.e. topics shared in text data in two different natural languages. The deficiency comes from the fact that all these models rely on co-occurrences of words forming a topical cluster, but words in different language generally do not co-occur with each other. Thus with the existing models, we can only extract topics from text in each language, but cannot extract common topics shared in multiple languages.

In this paper, we propose a novel topic model, called Probabilistic Cross-Lingual Latent Semantic Analysis (PCLSA) model, which can be used to mine shared latent topics from *unaligned* text data in different languages. PCLSA extends the Probabilistic Latent Semantic Analysis (PLSA) model by regularizing its likelihood function with soft constraints defined based on a bilingual dictionary. The dictionary-based constraints are key to bridge the gap of different languages and would force the captured co-occurrences of words in each language by PCLSA to be "synchronized" so that related words in the two languages would have similar probabilities. PCLSA can be estimated efficiently using the General Expectation-Maximization (GEM) algorithm. As a topic extraction algorithm, PCLSA would take a pair of unaligned document sets in different languages and a bilingual dictionary as input, and output a set of aligned word distributions in both languages that can characterize the shared topics in the two languages. In addition, it also outputs a topic cov-



*Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1128–1137,
Uppsala, Sweden, 11-16 July 2010. 

erage distribution for each language to indicate the relative coverage of different shared topics in each language.

To the best of our knowledge, no previous work has attempted to solve this topic extraction problem and generate the same output. The closest existing work to ours is the MuTo model proposed in (Boyd-Graber and Blei, 2009) and the JointLDA model published recently in (Jagarlamudi and Daumé III, 2010). Both used a bilingual dictionary to bridge the language gap in a topic model. However, the goals of their work are different from ours in that their models mainly focus on mining cross-lingual topics of matching word pairs and discovering the correspondence at the vocabulary level. Therefore, the topics extracted using their model cannot indicate how a common topic is covered *differently* in the two languages, because the words in each word pair share the same probability in a common topic. Our work focuses on discovering correspondence at the topic level. In our model, since we only add a soft constraint on word pairs in the dictionary, their probabilities in common topics are generally different, naturally capturing which shows the different variations of a common topic in different languages.

We use a cross-lingual news data set and a review data set to evaluate PCLSA. We also propose a "cross-collection" likelihood measure to quantitatively evaluate the quality of mined topics. Experimental results show that the PCLSA model can effectively extract cross-lingual latent topics from multilingual text data, and it outperforms a baseline approach using the standard PLSA on text data in each language.

## 2 Related Work

Many topic models have been proposed, and the two basic models are the Probabilistic Latent Semantic Analysis (PLSA) model (Hofmann, 1999a) and the Latent Dirichlet Allocation (LDA) model (Blei et al., 2003b). They and their extensions have been successfully applied to many problems, including hierarchical topic extraction (Hofmann, 1999b; Blei et al., 2003a; Li and McCallum, 2006), author-topic modeling (Steyvers et al., 2004), contextual topic analysis (Mei and Zhai, 2006), dynamic and correlated topic models (Blei and Lafferty, 2005; Blei and Lafferty, 2006), and opinion analysis (Mei et al., 2007; Branavan et al., 2008). Our work is an extension of PLSA by incorporating the knowledge of a bilingual dictionary as soft constraints. Such an extension is similar to the extension of PLSA for incorporating social network analysis (Mei et al., 2008a) but our constraint is different.

Some previous work on multilingual topic models assume documents in multiple languages are aligned either at the document level, sentence level or by time stamps (Mimno et al., 2009; Zhao and Xing, 2006; Kim and Khudanpur, 2004; Ni et al., 2009; Wang et al., 2007). However, in many applications, we need to mine topics from *unaligned* text corpus. For example, mining topics from search results in different languages can facilitate summarization of multilingual search results.

Besides all the multilingual topic modeling work discussed above, comparable corpora have also been studied extensively (e.g. (Fung, 1995; Franz et al., 1998; Masuichi et al., 2000; Sadat et al., 2003; Gliozzo and Strapparava, 2006)), but most previous work aims at acquiring word translation knowledge or cross-lingual text categorization from comparable corpora. Our work differs from this line of previous work in that our goal is to discover shared latent topics from multi-lingual text data that are *weakly* comparable (e.g. the data does not have to be aligned by time).

## 3 Problem Formulation

In general, the problem of cross-lingual topic extraction can be defined as to extract a set of common cross-lingual latent topics covered in text collections in different natural languages. A cross-lingual latent topic will be represented as a multinomial word distribution over the words in *all* the languages, i.e. a multilingual word distribution. For example, given two collections of news articles in English and Chinese, respectively, we would like to extract common topics simultaneously from the two collections. A discovered common topic, such as the terrorist attack on September 11, 2001, would be characterized by a word distribution that would assign relatively high probabilities to words related to this event in both English and Chinese (e.g. "terror", "attack", "afghanistan", "taliban", and their translations in Chinese).

As a computational problem, our input is a multi-lingual text corpus, and output is a set of cross-lingual latent topics. We now define this problem more formally.

**Definition 1 (Multi-Lingual Corpus)** A ***multi-lingual corpus*** $\mathcal{C}$ is a set of text collections $\{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_s\}$, where $\mathcal{C}_i = \{d_1^i, d_2^i, \ldots, d_{M_i}^i\}$ is a collection of documents in language $L_i$ with vocabulary $V_i = \{w_1^i, w_2^i, \ldots, w_{N_i}^i\}$. Here, $M_i$ is the total number of documents in $\mathcal{C}_i$, $N_i$ is the total number of words in $V_i$, and $d_j^i$ is a document in collection $\mathcal{C}_i$.

Following the common assumption of bag-of-words representation, we represent document $d_j^i$ with a bag of words $\{w_{j_1}^i, w_{j_2}^i, \ldots, w_{j_d}^i\}$, and use $c(w_k^i, d_j^i)$ to denote the count of word $w_k^i$ in document $d_j^i$.

**Definition 2 (Cross-Lingual Topic):** A cross-lingual topic $\theta$ is a semantically coherent multinomial distribution over all the words in the vocabularies of languages $L_1$, ..., $L_s$. That is, $p(w|\theta)$ would give the probability of a word $w$ which can be in any of the $s$ languages under consideration. $\theta$ is semantically coherent if it assigns high probabilities to words that are semantically related either in the same language or across different languages. Clearly, we have $\sum_{i=1}^{s} \sum_{w \in V_i} p(w|\theta) = 1$ for any cross-lingual topic $\theta$.

**Definition 3 (Cross-Lingual Topic Extraction)** Given a multi-lingual corpus $\mathcal{C}$, the task of cross-lingual topic extraction is to model and extract $k$ major cross-lingual topics $\{\theta_1, \theta_2, \ldots, \theta_k\}$ from $\mathcal{C}$, where $\theta_i$ is a cross-lingual topic, and $k$ is a user specified parameter.

The extracted cross-lingual topics can be directly used as a summary of the common content of the multi-lingual data set. Note that once a cross-lingual topic is extracted, we can easily obtain its representation in each language $L_i$ by "splitting" the cross-lingual topic into multiple word distributions in different languages. Formally, the word distribution of a cross-lingual topic $\theta$ in language $L_i$ is given by $p_i(w^i|\theta) = \frac{p(w^i|\theta)}{\sum_{w \in V_i} p(w|\theta)}$.

These aligned language-specific word distributions can directly review the variations of topics in different languages. They can also be used to analyze the difference of the coverage of the same topic in different languages. Moreover, they are also useful for retrieving relevant articles or passages in each language and aligning them to the same common topic, thus essentially also allowing us to integrate and align articles in multiple languages.

# 4 Probabilistic Cross-Lingual Latent Semantic Analysis

In this section, we present our probabilistic cross-lingual latent semantic analysis (PCLSA) model and discuss how it can be used to extract cross-lingual topics from multi-lingual text data.

The main reason why existing topic models can't be used for cross-lingual topic extraction is because they cannot cross the language barrier. Intuitively, in order to cross the language barrier and extract a common topic shared in articles in different languages, we must rely on some kind of linguistic knowledge. Our PCLSA model assumes the availability of bi-lingual dictionaries for at least some language pairs, which are generally available for major language pairs. Specifically, for text data in languages $L_1$, ..., $L_s$, if we represent each language as a node in a graph and connect those language pairs for which we have a bilingual dictionary, the minimum requirement is that the whole graph is connected. Thus, as a minimum, we will need $s-1$ distinct bilingual dictionaries. This is so that we can potentially cross all the language barriers.

Our key idea is to "synchronize" the extraction of monolingual "component topics" of a cross-lingual topic from individual languages by forcing a cross-lingual topic word distribution to assign similar probabilities to words that are potential translations according to a $L_i$-$L_j$ bilingual dictionary. We achieve this by adding such preferences formally to the likelihood function of a probabilistic topic model as "soft constraints" so that when we estimate the model, we would try to not only fit the text data well (which is necessary to extract coherent component topics from each language), but also satisfy our specified preferences (which would ensure the extracted component topics in different languages are semantically related). Below we present how we implement this idea in more detail.

A bilingual dictionary for languages $L_i$ and $L_j$ generally would give us a many-to-many mapping between the vocabularies of the two languages. With such a mapping, we can construct a bipartite graph $G_{ij} = (V_{ij}, E_{ij})$ between the two languages where if one word can be potentially translated into another word, the two words would be connected with an edge. An edge can be weighted based on the probability of the corresponding translation. An example graph for

Chinese-English dictionary is shown in Figure 1.

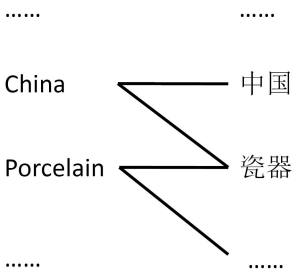


Figure 1: A Dictionary based Word Graph

With multiple bilingual dictionaries, we can merge the graphs to generate a multi-partite graph $G = (V, E)$. Based on this graph, the PCLSA model extends the standard PLSA by adding a constraint to the likelihood function to "smooth" the word distributions of topics in PLSA on the multi-partite graph so that we would encourage the words that are connected in the graph (i.e. possible translations of each other) to be given similar probabilities by every cross-lingual topic. Thus when a cross-lingual topic picks up words that co-occur in mono-lingual text, it would prefer picking up word pairs whose translations in other languages also co-occur with each other, giving us a coherent multilingual word distribution that characterizes well the content of text in different languages.

Specifically, let $\Theta = \{\theta_j\}$ $(j = 1, ..., k)$ be a set of $k$ cross-lingual topic models to be discovered from a multilingual text data set with $s$ languages such that $p(w|\theta_i)$ is the probability of word $w$ according to the topic model $\theta_i$.

If we are to use the regular PLSA to model our data, we would have the following log-likelihood and we usually use a maximum likelihood estimator to estimate parameters and discover topics.

$$L(\mathcal{C}) = \sum_{i=1}^{s} \sum_{d \in \mathcal{C}_i} \sum_{w} c(w, d) \log \sum_{j=1}^{k} p(\theta_j|d) p(w|\theta_j)$$

Our main extension is to add to $L(\mathcal{C})$ a cross-lingual constraint term $R(\mathcal{C})$ to incorporate the knowledge of bilingual dictionaries. $R(\mathcal{C})$ is defined as

$$R(\mathcal{C}) = \frac{1}{2} \sum_{\langle u,v \rangle \in E} w(u, v) \sum_{j=1}^{k} \left(\frac{p(w_u|\theta_j)}{Deg(u)} - \frac{p(w_v|\theta_j)}{Deg(v)}\right)^2$$

where $w(u, v)$ is the weight on the edge between $u$ and $v$ in the multi-partite graph $G = (V, E)$, which in our experiments is set to 1, and $Deg(u)$ is the degree of word $u$, i.e. the sum of the weights of all the edges ending with $u$.

Intuitively, $R(\mathcal{C})$ measures the difference between $p(w_u|\theta_j)$ and $p(w_v|\theta_j)$ for each pair $(u, v)$ in a bilingual dictionary; the more they differ, the larger $R(\mathcal{C})$ would be. So it can be regarded as a "loss function" to help us assess how well the "component word distributions" in multiple languages are correlated semantically. Clearly, we would like the extracted topics to have a small $R(\mathcal{C})$. We choose this specific form of loss function because it would make it convenient to solve the optimization problem of maximizing the corresponding regularized maximum likelihood (Mei et al., 2008b). The normalization with $Deg(u)$ and $Deg(v)$ can be regarded as a way to compensate for the potential ambiguity of $u$ and $v$ in their translations.

Putting $L(\mathcal{C})$ and $R(\mathcal{C})$ together, we would like to maximize the following objective function which is a regularized log-likelihood:

$$O(\mathcal{C}, G) = (1 - \lambda) L(\mathcal{C}) - \lambda R(\mathcal{C}) \quad (1)$$

where $\lambda \in (0, 1)$ is a parameter to balance the likelihood and the regularizer. When $\lambda = 0$, we recover the standard PLSA.

Specifically, we will search for a set of values for all our parameters that can maximize the objective function defined above. Our parameters include all the cross-lingual topics and the coverage distributions of the topics in all documents, which we denote by $\Psi = \{p(w|\theta_j), p(\theta_j|d)\}_{d,w,j}$ where $j = 1, ..., k$, $w$ varies over the entire vocabularies of all the languages , $d$ varies over all the documents in our collection. This optimization problem can be solved using a Generalized Expectation-Maximization (GEM) algorithm as described in (Mei et al., 2008a).

Specifically, in the **E-step** of the algorithm, the distribution of hidden variables is computed using Eq. 2.

$$z(w, d, j) = \frac{p(\theta_j|d) p(w|\theta_j)}{\sum_{j'} p(\theta_{j'}|d) p(w|\theta_{j'})} \quad (2)$$

Then in the **M-step**, we need to maximize the complete data likelihood $Q(\Psi; \Psi_n)$:

$$Q(\Psi; \Psi_n) = (1 - \lambda) L'(\mathcal{C}) - \lambda R(\mathcal{C})$$

where

$$L'(\mathcal{C}) = \sum_d \sum_w c(w,d) \sum_j z(w,d,j) \log p(\theta_j|d) p(w|\theta_j), \quad (3)$$

with the constraints that $\sum_j p(\theta_j|d) = 1$ and $\sum_w p(w|\theta_j) = 1$.

There is a closed form solution if we only want to maximize the $L'(\mathcal{C})$ part:

$$\begin{aligned} p^{(n+1)}(\theta_j|d) &= \frac{\sum_w c(w,d) z(w,d,j)}{\sum_w \sum_{j'} c(w,d) z(w,d,j')} \\ p^{(n+1)}(w|\theta_j) &= \frac{\sum_d c(w,d) z(w,d,j)}{\sum_d \sum_w' c(w',d) z(w',d,j)} \end{aligned} \quad (4)$$

However, there is no closed form solution in the M-step for the whole objective function. Fortunately, according to GEM we do not need to find the local maximum of $Q(\Psi; \Psi_n)$ in every M-step, and we only need to find a new value $\Psi_{n+1}$ to improve the complete data likelihood, i.e. to make sure $Q(\Psi_{n+1}; \Psi_n) \geq Q(\Psi_n; \Psi_n)$. So our method is to first maximize the $L'(\mathcal{C})$ part using Eq. 4 and then use Eq. 5 to gradually increase the $R(\mathcal{C})$ part.

$$\begin{aligned} p^{(t+1)}(w_u|\theta_j) &= (1-\alpha) p^{(t)}(w_u|\theta_j) \\ &+ \alpha \sum_{\langle u,v \rangle \in E} \frac{w(u,v)}{Deg(v)} p^{(t)}(w_v|\theta_j) \end{aligned} \quad (5)$$

Here, parameter $\alpha$ is the length of each smoothing step. Obviously, after each smoothing step, the sum of the probabilities of all the words in one topic is still equal to 1. We smooth the parameters until we cannot get a better parameter set $\Psi_{n+1}$. Then, we continue to the next E-step. If there is no $\Psi_{n+1}$ s.t. $Q(\Psi_{n+1}; \Psi_n) \geq Q(\Psi_n; \Psi_n)$, then we consider $\Psi_n$ to be the local maximum point of the objective function Eq. 1.

# 5 Experiment Design

## 5.1 Data Set

The data set we used in our experiment is collected from news articles of Xinhua English and Chinese newswires. The whole data set is quite big, containing around 40,000 articles in Chinese and 35,000 articles in English. For different purpose of our experiments, we randomly selected different number of documents from the whole corpus, and we will describe the concrete statistics in each experiment. To process the Chinese corpus, we use a simple segmenter[1] to split the data into Chinese phrases. Both Chinese and English stopwords are removed from our data.

The dictionary file we used for our PCLSA model is from mandarintools.com[2]. For each Chinese phrase, if it has several English meanings, we add an edge between it and each of its English translation. If one English translation is an English phrase, we add an edge between the Chinese phrase and each English word in the phrase.

## 5.2 Baseline Method

As a baseline method, we can apply the standard PLSA (Hofmann, 1999a) directly to the multilingual corpus. Since PLSA takes advantage of the word co-occurrences in the document level to find semantic topics, directly using it for a multilingual corpus will result in finding topics mainly reflecting a single language (because words in different languages would not co-occur in the same document in general). That is, the discovered topics are mostly monolingual. These monolingual topics can then be aligned based on a bilingual dictionary to suggest a possible cross-lingual topic.

# 6 Experimental Results

## 6.1 Qualitative Comparison

To qualitatively compare PCLSA with the baseline method, we compare the word distributions of topics extracted by them. The data set we used in this experiment is selected from the Xinhua News data during the period from Jun. 8th, 2001 to Jun. 15th, 2001. There are totally 1799 English articles and 1485 Chinese articles in the data set. The number of topics to be extracted is set to 10 for both methods.

Table 1 shows the experimental results. To make it easier to understand, we add an English translation to each Chinese phrase in our results. The first ten rows show sample topics of the modeling results of traditional PLSA model. We can see that it only contains mono-language topics, i.e. the topics are either in Chinese or in English. The next ten rows are the results from our PCLSA model. Compared with the baseline method, PCLSA can not only find coherent topics from the cross-lingual corpus, but it can also show the content about one topic from both two language corpora. For example, in 'Topic 2'

[1] http://www.mandarintools.com/segmenter.html
[2] http://www.mandarintools.com/cedict.html

Table 2: Synthetic Data Set from Xinhua News

| | | | |
|---|---|---|---|
| English | Shrine<br>90 | Olympic<br>101 | Championship<br>70 |
| Chinese | CPC Anniversary<br>95 | Afghan War<br>206 | Championship<br>72 |

which is about 'Israel' and 'Palestinian', the Chinese corpus mentions a lot about 'Arafat' who is the leader of 'Palestinian', while the English corpus discusses more on topics such as 'cease fire' and 'women'. Similarly, in 'Topic 9', the topic is related to Philippine, the Chinese corpus mentions some environmental situation in Philippine, while the English corpus mentions a lot about 'Abu Sayyaf'.

### 6.2 Discovering Common Topics

To demonstrate the ability of PCLSA for finding common topics in cross-lingual corpus, we use some event names, e.g. 'Shrine' and 'Olympic', as queries and randomly select a certain number of documents from the whole corpus, which are related to the queries. The number of documents for each query in the synthetic data set is shown in Table 2. In either the English corpus or the Chinese corpus, we select a smaller number of documents about topic 'Championship' combined with the other two topics in the same corpus. In this way, when we want to extract two topics from either English or Chinese corpus, the 'Championship' topic may not be easy to extract, because the other two topics have more documents in the corpus. However, when we use PCLSA to extract four topics from the two corpora together, we expect that the topic 'Championship' will be found, because now the sum of English and Chinese documents related to 'Championship' is larger than other topics. The experimental result is shown in Table 3. The first two columns are the two topics extracted from Engish corpus, the third and the forth columns are two topics from Chinese corpus, and the other four columns are the results from cross-lingual corpus. We can see that in either the Chinese sub-collection or the English sub-collection, the topic 'Championship' is not extracted as a significant topic. But, as expected, the topic 'Championship' is extracted from the cross-lingual corpus, while the topic 'Olympic' and topic 'Shrine' are merged together. This demonstrate that PCLSA is capable of extracting common topics from a cross-lingual corpus.

### 6.3 Quantitative Evaluation

We also quantitatively evaluate how well our PCLSA model can discover common topics among corpus in different languages. We propose a "cross-collection" likelihood measure for this purpose. The basic idea is: suppose we got $k$ cross-lingual topics from the whole corpus, then for each topic, we split the topic into two separate set of topics, English topics and Chinese topics, using the splitting formula described before, i.e. $p_i(w^i|\theta) = \frac{p(w^i|\theta)}{\sum_{w \in V_i} p(w|\theta)}$. Then, we use the word distribution of the Chinese topics (translating the words into English) to fit the English Corpus and use the word distribution of the English topics (translating the words into Chinese) to fit the Chinese Corpus. If the topics mined are common topics in the whole corpus, then such a "cross-collection" likelihood should be larger than those topics which are not commonly shared by the English and the Chinese corpus. To calculate the likelihood of fitness, we use the folding-in method proposed in (Hofmann, 2001). To translate topics from one language to another, e.g. Chinese to English, we look up the bilingual dictionary and do word-to-word translation. If one Chinese word has several English translations, we simply distribute its probability mass equally to each English translation.

For comparison, we use the standard PLSA model as the baseline. Basically, suppose PLSA mined $k$ semantic topics in the Chinese corpus and $k$ semantic topics in the English corpus. Then, we also use the "cross-collection" likelihood measure to see how well those $k$ semantic Chinese topics fit the English corpus and those $k$ semantic English topics fit the Chinese corpus.

We totally collect three data sets to compare the performance. For the first data set, (English_1, Chinese_1), both the Chinese and English corpus are chosen from the Xinhua News Data during the period from 2001.06.08 to 2001.06.15, which has 1799 English articles and 1485 Chinese articles. For the second data set, (English_2, Chinese_2), the Chinese corpus Chinese_2 is the same as Chinese_1, but the English corpus is chosen from 2001.06.14 to 2001.06.19 which has 1547 documents. For the third data set, (English_3, Chinese_3), the Chinese corpus is the same as in data set one, but the English corpus is chosen from 2001.10.02 to 2001.10.07 which contains 1530 documents. In other words, in the first data set,

Table 1: Qualitative Evaluation

| Topic 0 | Topic 1 | Topic 2 | Topic 3 | Topic 4 | Topic 5 | Topic 6 | Topic 7 | Topic 8 | Topic 9 |
|---|---|---|---|---|---|---|---|---|---|
| 党(party) | 犯罪(crime) | 选手(athlete) | 巴(palestine) | 合作(collaboration) | 教育(education) | israel | bt | dollar | china |
| 共产党(communist) | 农业(agriculture) | 冠军(champion) | 巴勒斯坦(palestine) | 上海(shanghai) | 球(ball) | palestinian | beat | percent | cooperate |
| 革命(revolution) | 旅游(travel) | 锦标赛(championship) | 以色列(israel) | 关系(relation) | 联赛(league) | eu | final | million | shanghai |
| 党员(party member) | 邪教(heathendom) | 底(base) | 停火(cease fire) | 两国(bilateral) | 足球(soccer) | police | championship | index | develop |
| 中央(central) | 公安(public security) | 羽毛球(badminton) | 联合国(UN) | 贸易(trade) | 分钟(minute) | report | play | stock | beije |
| 主义(ism) | 钱宏(name) | 体育(sports) | 中东(mid east) | 总统(president) | 队员(team member) | secure | champion | point | particulate |
| 干部(cadre) | 案(case) | 决赛(final) | 黎巴嫩(lebanon) | 国(country) | 教师(teacher) | kill | win | share | matter |
| 毛泽东(chairman mao) | 执法(law enforcement) | 女子(women) | 马其顿(macedon) | 友好(friendly) | 中小学(school) | europe | olympic | close | sco |
| 中共(chinese communist) | 市(city) | 象棋(chess) | 冲突(conflict) | 会晤(meet) | 球队(team) | egypt | game | 0 | invest |
| 领导(leader) | 处罚(penalize) | 健身(fitness) | 会谈(talk) | 俄罗斯(russia) | 甲(grade A) | treaty | cup | billion | project |
| 两国(bilateral) | 联赛(league) | israel | cooperate | 选手(athlete) | party | eu | invest | 0 | 吸入(absorb) |
| 合作(collaboration) | 钱宏(name) | 以色列(israel) | sco | particulate | 党(party) | khatami | 投资(investment) | dollar | II |
| 会谈(talk) | 球(ball) | bt | develop | 冠军 | communist | ireland | 亿元(billion) | percent | 阿布沙耶夫(abu) |
| 友好(friendly) | 申花(shenhua) | palestinian | country | athlete | revolution | 爱尔兰(ireland) | 教育(education) | index | I |
| 巴(palestine) | 主场(host) | ceasefire | president | champion | 主义(-ism) | elect | 环保(environ. protect.) | million | 颗粒(particle) |
| country | A | 阿拉法特(arafat) | apec | ii | 抗日(antiwar) | vote | 资金(money) | stock | philippine |
| 联合国(UN) | ball | women | shanghai | 象棋(chess) | 同志(comrade) | presidential | 中小学(school) | billion | abu |
| 领导人(leader) | 金德(jinde) | jerusalem | africa | competition | 革命(revolution) | cpc | market | point | 底(base) |
| bilateral | 赛季(season) | mideast | meet | contestant | 党组织(party) | iran | 教师(teacher) | 十亿(billion) | III |
| state | 球员(player) | lebanon | 江泽民(zemin jiang) | 体操(gymnastics) | ideology | referendum | business | share | 物(object) |

Table 3: Effectiveness of Extracting Common Topics

| English 1 | English 2 | Chinese 1 | Chinese 2 | Cross 1 | Cross 2 | Cross 3 | Cross 4 |
|---|---|---|---|---|---|---|---|
| japan | olympic | 共产党(CPC) | 阿富汗(afghan) | koizumi | 利班(taliban) | swim | 工人(worker) |
| shrine | ioc | 锦(championship) | 塔(taliban) | yasukuni | 军事(military) | 锦(championship) | party |
| visit | beije | 世(world) | 利班(taliban) | ioc | city | 自由泳(free style) | 三个(three) |
| koizumi | game | 思想(thought) | 军事(military) | japan | refugee | 跳水(diving) | 马克思(marx) |
| yasukuni | july | 理论(theory) | 打击(attack) | olympic | side | 锦标赛(championship) | communist |
| war | bid | 马克思(marx) | 美军(US army) | beije | 美军(US army) | 半决赛(semi final) | marx |
| august | swim | 游泳(swim) | 丹(laden) | shrine | 轰炸(bomb) | competition | theory |
| asia | vote | 锦标赛(championship) | 部队(army) | visit | 喀布尔(kabul) | 游泳(swim) | 建党(found party) |
| criminal | championship | 党(party) | 轰炸(bomb) | 奥运会(olympic) | 空袭(attack) | 赛纪录(record) | 共产党(CPC) |
| ii | committee | 建党(found party) | 喀布尔(kabul) | 奥林匹克(olympic) | 难民(refugee) | 罗雪娟(xuejuan luo) | revolution |

the English corpus and Chinese corpus are comparable with each other, because they cover similar events during the same period. In the second data set, the English and Chinese corpora share some common topics during the overlap period. The third data is the most tough one since the two corpora are from different periods. The purpose of using these three different data sets for evaluation is to test how well PCLSA can mine common topics from either a data set where the English corpus and the Chinese corpus are comparable or a data set where the English corpus and the Chinese corpus rarely share common topics.

The experimental results are shown in Table 4. Each row shows the "cross-collection" likelihood of using the "cross-collection" topics to fit the data set named in the first column. For example, in the first row, the values are the "cross-collection" likelihood of using Chinese topics found by different methods from the first data set to fit English_1. The last collum shows how much improvement we got from PCLSA compared with PLSA. From the results, we can see that in all the data sets, our PCLSA has higher "cross-collection" likelihood value, which means it can find better common topics compared to the baseline method. Notice that the Chinese corpora are the same in all three data sets. The results show that both PCLSA and PLSA get lower "cross-collection" likelihood for fitting the Chinese corpora when the data set becomes "tougher", i.e. less topic overlapping, but the improvement of PCLSA over PLSA does not drop much. On the other hand, the improvement of PCLSA over PLSA on the three English corpora does not show any correlation with the difficulty of the data set.

Table 4: Quantitative Evaluation of Common Topic Finding ("cross-collection" log-likelihood)

| | PCLSA | PLSA | Rel. Imprv. |
|---|---|---|---|
| English_1 | -2.86294E+06 | -3.03176E+06 | 5.6% |
| Chinese_1 | -4.69989E+06 | -4.85369E+06 | 3.2% |
| English_2 | -2.48174E+06 | -2.60805E+06 | 4.8% |
| Chinese_2 | -4.73218E+06 | -4.88906E+06 | 3.2% |
| English_3 | -2.44714E+06 | -2.60540E+06 | 6.1% |
| Chinese_3 | -4.79639E+06 | -4.94273E+06 | 3.0% |

### 6.4 Extracting from Multi-Language Corpus

In the previous experiments, we have shown the capability and effectiveness of the PCLSA model in latent topic extraction from two language corpora. In fact, the proposed model is general and capable of extracting latent topics from multi-language corpus. For example, if we have dictionaries among multiple languages, we can construct a multi-partite graph based on the correspondence between those vocabularies, and then smooth the PCLSA model with this graph.

To show the effectiveness of PCLSA in mining multiple language corpus, we first construct a simulated data set based on 1115 reviews of three brands of laptops, namely IBM (303), Apple(468) and DELL(344). To simulate a three language cor-

Table 5: Effectiveness of Latent Topic Extraction from Multi-Language Corpus

| Topic 0 | Topic 1 | Topic 2 | Topic 3 | Topic 4 | Topic 5 | Topic 6 | Topic 7 |
|---|---|---|---|---|---|---|---|
| cd(apple) | battery(dell) | mouse(dell) | print(apple) | port(ibm) | laptop(ibm) | os(apple) | port(dell) |
| port(apple) | drive(dell) | button(dell) | resolution(dell) | card(ibm) | t20(ibm) | run(apple) | 2(dell) |
| drive(apple) | 8200(dell) | touchpad(dell) | burn(apple) | modem(ibm) | thinkpad(ibm) | 1(apple) | usb(dell) |
| airport(apple) | inspiron(dell) | pad(dell) | normal(dell) | display(ibm) | battery(ibm) | ram(apple) | 1(dell) |
| firewire(apple) | system(dell) | keyboard(dell) | image(dell) | built(ibm) | notebook(ibm) | mac(apple) | 0(dell) |
| dvd(apple) | hour(dell) | point(dell) | digital(apple) | swap(ibm) | ibm(ibm) | battery(apple) | slot(dell) |
| usb(apple) | sound(dell) | stick(dell) | organize(apple) | easy(ibm) | 3(ibm) | hour(apple) | firewire(dell) |
| rw(apple) | dell(dell) | rest(dell) | cds(apple) | connector(ibm) | feel(ibm) | 12(apple) | display(dell) |
| card(apple) | service(dell) | touch(dell) | latch(apple) | feature(ibm) | hour(ibm) | operate(apple) | standard(dell) |
| mouse(apple) | life(dell) | erase(dell) | advertise(dell) | cd(ibm) | high(ibm) | word(apple) | fast(dell) |
| osx(apple) | applework(apple) | port(dell) | battery(dell) | lightest(ibm) | uxga(dell) | light(ibm) | battery(apple) |
| memory(dell) | file(apple) | port(apple) | battery(ibm) | quality(dell) | ultrasharp(dell) | ultrabay(ibm) | point(dell) |
| special(dell) | bounce(apple) | port(ibm) | battery(apple) | year(ibm) | display(dell) | connector(ibm) | touchpad(dell) |
| crucial(dell) | quit(apple) | firewire(apple) | geforce4(dell) | hassle(ibm) | organize(apple) | dvd(ibm) | button(dell) |
| memory(apple) | word(apple) | imac(apple) | 100mhz(apple) | bania(dell) | learn(apple) | nice(ibm) | hour(apple) |
| memory(ibm) | file(ibm) | firewire(dell) | 440(dell) | 800mhz(apple) | logo(apple) | modem(ibm) | battery(ibm) |
| netscape(apple) | file(dell) | firewire(ibm) | bus(apple) | trackpad(apple) | postscript(apple) | connector(dell) | battery(dell) |
| reseller(apple) | microsoft(apple) | jack(apple) | 8200(dell) | cover(ibm) | ll(apple) | light(apple) | fan(dell) |
| 10(dell) | ms(apple) | playback(dell) | 8100(dell) | workmanship(dell) | sxga(dell) | light(dell) | erase(dell) |
| special(apple) | excel(apple) | jack(dell) | chipset(dell) | section(apple) | warm(apple) | floppy(ibm) | point(apple) |
| 2000(ibm) | ram(apple) | port(dell) | itune(apple) | uxga(dell) | port(apple) | pentium(dell) | drive(ibm) |
| window(ibm) | ram(ibm) | port(apple) | applework(apple) | screen(dell) | port(ibm) | processor(dell) | drive(dell) |
| 2000(apple) | ram(dell) | port(ibm) | imovie(apple) | screen(ibm) | port(dell) | p4(dell) | drive(apple) |
| 2000(dell) | screen(apple) | 2(dell) | import(apple) | screen(apple) | usb(apple) | power(dell) | hard(ibm) |
| window(apple) | 1(apple) | 2(apple) | battery(apple) | ultrasharp(dell) | plug(apple) | pentium(apple) | osx(apple) |
| window(dell) | screen(ibm) | 2(ibm) | iphoto(apple) | 1600x1200(dell) | cord(apple) | pentium(ibm) | hard(dell) |
| portege(ibm) | screen(dell) | speak(dell) | battery(ibm) | display(dell) | usb(ibm) | keyboard(dell) | hard(apple) |
| option(ibm) | 1(ibm) | toshiba(dell) | battery(dell) | display(apple) | usb(dell) | processor(ibm) | card(ibm) |
| hassle(ibm) | 1(dell) | speak(ibm) | hour(apple) | display(ibm) | firewire(apple) | processor(apple) | dvd(ibm) |
| device(ibm) | maco(apple) | toshiba(ibm) | hour(ibm) | view(dell) | plug(ibm) | power(apple) | card(dell) |

pus, we use an 'IBM' word, an 'Apple' word, and a 'Dell' word to replace an English word in their corpus. For example, we use 'IBM10', 'Apple10', 'Dell10' to replace the word 'CD' whenever it appears in an IBM's, Apple's, or Dell's review. After the replacement, the reviews about IBM, Apple, and Dell will not share vocabularies with each other. On the other hand, for any three created words which represent the same English word, we add three edges among them, and therefore we get a simulated dictionary graph for our PCLSA model.

The experimental result is shown in Table 5, in which we try to extract 8 topics from the cross-lingual corpus. The first ten rows show the result of our PCLSA model, in which we set a very small value to the weight parameter $\lambda$ for the regularizer part. This can be used as an approximation of the result from the traditional PLSA model on this three language corpus. We can see that the extracted topics are mainly written in mono-language. As we set the value of parameter $\lambda$ larger, the extracted topics become multi-lingual, which is shown in the next ten rows. From this result, we can see the difference between the reviews of different brands about the similar topic. In addition, if we set the $\lambda$ even larger, we will get topics that are mostly made of the same words from the three different brands, which means the extracted topics are very smooth on the dictionary graph now.

## 7 Conclusion

In this paper, we study the problem of cross-lingual latent topic extraction where the task is to extract a set of common latent topics from multi-lingual text data. We propose a novel probabilistic topic model (i.e. the Probabilistic Cross-Lingual Latent Semantic Analysis (PCLSA) model) that can incorporate translation knowledge in bilingual dictionaries as a regularizer to constrain the parameter estimation so that the learned topic models would be synchronized in multiple languages. We evaluated the model using several data sets. The experimental results show that PCLSA is effective in extracting common latent topics from multilingual text data, and it outperforms the baseline method which uses the standard PLSA to fit each monolingual text data set.

Our work opens up some interesting future research directions to further explore. First, in this paper, we have only experimented with uniform weighting of edge in the bilingual graph. It should be very interesting to explore how to assign weights to the edges and study whether weighted graphs can further improve performance. Second, it would also be interesting to further extend PCLSA to accommodate discovering topics in each language that aren't well-aligned with other languages.

## 8 Acknowledgments

We sincerely thank the anonymous reviewers for their comprehensive and constructive comments. The work was supported in part by NASA grant

NNX08AC35A, by the National Science Foundation under Grant Numbers IIS-0713581, IIS-0713571, and CNS-0834709, and by a Sloan Research Fellowship.

## References

David Blei and John Lafferty. 2005. Correlated topic models. In *NIPS '05: Advances in Neural Information Processing Systems 18*.

David M. Blei and John D. Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120.

D. Blei, T. Griffiths, M. Jordan, and J. Tenenbaum. 2003a. Hierarchical topic models and the nested chinese restaurant process. In *Neural Information Processing Systems (NIPS) 16*.

D. Blei, A. Ng, and M. Jordan. 2003b. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

J. Boyd-Graber and D. Blei. 2009. Multilingual topic models for unaligned text. In *Uncertainty in Artificial Intelligence*.

S. R. K. Branavan, Harr Chen, Jacob Eisenstein, and Regina Barzilay. 2008. Learning document-level semantic properties from free-text annotations. In *Proceedings of ACL 2008*.

Martin Franz, J. Scott McCarley, and Salim Roukos. 1998. Ad hoc and multilingual information retrieval at IBM. In *Text REtrieval Conference*, pages 104–115.

Pascale Fung. 1995. A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. In *Proceedings of ACL 1995*, pages 236–243.

Alfio Gliozzo and Carlo Strapparava. 2006. Exploiting comparable corpora and bilingual dictionaries for cross-language text categorization. In *ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 553–560, Morristown, NJ, USA. Association for Computational Linguistics.

T. Hofmann. 1999a. Probabilistic latent semantic analysis. In *Proceedings of UAI 1999*, pages 289–296.

Thomas Hofmann. 1999b. The cluster-abstraction model: Unsupervised learning of topic hierarchies from text data. In *IJCAI' 99*, pages 682–687.

Thomas Hofmann. 2001. Unsupervised learning by probabilistic latent semantic analysis. *Mach. Learn.*, 42(1-2):177–196.

Jagadeesh Jagaralamudi and Hal Daumé III. 2010. Extracting multilingual topics from unaligned corpora. In *Proceedings of the European Conference on Information Retrieval (ECIR)*, Milton Keynes, United Kingdom.

Woosung Kim and Sanjeev Khudanpur. 2004. Lexical triggers and latent semantic analysis for cross-lingual language model adaptation. *ACM Transactions on Asian Language Information Processing (TALIP)*, 3(2):94–112.

Wei Li and Andrew McCallum. 2006. Pachinko allocation: Dag-structured mixture models of topic correlations. In *ICML '06: Proceedings of the 23rd international conference on Machine learning*, pages 577–584.

H. Masuichi, R. Flournoy, S. Kaufmann, and S. Peters. 2000. A bootstrapping method for extracting bilingual text pairs. In *Proc. 18th COLINC*, pages 1066–1070.

Qiaozhu Mei and ChengXiang Zhai. 2006. A mixture model for contextual text mining. In *Proceedings of KDD '06*, pages 649–655.

Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang Zhai. 2007. Topic sentiment mixture: Modeling facets and opinions in weblogs. In *Proceedings of WWW '07*.

Qiaozhu Mei, Deng Cai, Duo Zhang, and ChengXiang Zhai. 2008a. Topic modeling with network regularization. In *WWW*, pages 101–110.

Qiaozhu Mei, Duo Zhang, and ChengXiang Zhai. 2008b. A general optimization framework for smoothing language models on graph structures. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 611–618, New York, NY, USA. ACM.

David Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew Mccallum. 2009. Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 880–889, Singapore, August. Association for Computational Linguistics.

Xiaochuan Ni, Jian-Tao Sun, Jian Hu, and Zheng Chen. 2009. Mining multilingual topics from wikipedia. In *WWW '09: Proceedings of the 18th international conference on World wide web*, pages 1155–1156, New York, NY, USA. ACM.

F. Sadat, M. Yoshikawa, and S. Uemura. 2003. Bilingual terminology acquisition from comparable corpora and phrasal translation to cross-language information retrieval. In *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 141–144.

Mark Steyvers, Padhraic Smyth, Michal Rosen-Zvi, and Thomas Griffiths. 2004. Probabilistic author-topic models for information discovery. In *Proceedings of KDD'04*, pages 306–315.

Xuanhui Wang, ChengXiang Zhai, Xiao Hu, and Richard Sproat. 2007. Mining correlated bursty topic patterns from coordinated text streams. In *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 784–793, New York, NY, USA. ACM.

Bing Zhao and Eric P. Xing. 2006. Bitam: Bilingual topic admixture models for word alignment. In *In Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics*.