Cross Lingual Entity Linking with Bilingual Topic Model

Tao Zhang, Kang Liu and Jun Zhao
Institute of Automation, Chinese Academy of Sciences
HaiDian District, Beijing, China
{tzhang, kliu, jzhao}@nlpr.ia.ac.cn

Abstract

Cross lingual entity linking means linking an entity mention in a background source document in one language with the corresponding real world entity in a knowledge base written in the other language. The key problem is to measure the similarity score between the context of the entity mention and the document of the candidate entity. This paper presents a general framework for doing cross lingual entity linking by leveraging a large scale and bilingual knowledge base, Wikipedia. We introduce a bilingual topic model that mining bilingual topic from this knowledge base with the assumption that the same Wikipedia concept documents of two different languages share the same semantic topic distribution. The extracted topics have two types of representation, with each type corresponding to one language. Thus both the context of the entity mention and the document of the candidate entity can be represented in a space using the same semantic topics. We use these topics to do cross lingual entity linking. Experimental results show that the proposed approach can obtain the competitive results compared with the state-of-art approach.

1 Introduction

The web is growing rapidly, which can be reflected by the number of Internet user and the amount of web content on the Internet. It has become a central goal for search engine company to maximize the user's satisfaction. Moreover, since a user's query useually expresses the information need of a named entity, it would be very useful if a general search engine could automatically discover information about named entities. The emergence of large scale knowledge base like Wikipedia and DBpedia has facilitated this problem. These knowledge bases are usually written in English. However, according to a report of August, 2008[1], there are 90 billion web pages on Internet and 31% of them are written in other languages. This report tells us that the percentage of web pages written in English are decreasing, which indicates that English pages are becoming less dominant compared with before. Therefore, it becomes more and more important to maintain the growing knowledge base by discovering and extracting information from all web contents written in different languages. Also, inserting new extracted knowledge derived from the information extraction system into a knowledge base inevitably needs a system to map the entity mentions in one language to their entries in a knowledge base which may be another language. This task is known as cross-lingual entity linking.

Intuitively, to resolve the cross-lingual entity linking problem, the key is how to define the similarity score between the context of entity mention and the document of candidate entity. Once the similarity score is defined, the system can select the entity with the highest score as the answer in an unsupervised way or use it as a feature in a supervised framework. However, the context of the entity mention and the document associated with the entity are described using different languages. Thus, the traditional BOW model, which is based on exact matching of terms between the query and candidate entity, is not suitable for the cross-lingual entity linking task.

To overcome this problem, many methods have been proposed. Previous studies [McNamee et al., 2011] usually use statistical machine translation tool to first translate the original document contexts into English equivalents and transform the cross-lingual problem into a monolingual entity linking problem. However, the performance of the entity linking system is highly depended on the performance of the statistical machine translation system. In order to train the statistical machine translation system, one must obtain parallel corpora in which texts in one language and their translation in other language are well aligned at word or sentence level. Such corpora are usually edited by human editors in specific domain. Obtaining such corpora is expensive. Therefore, the machine translation based method is not easy to adapt to other languages.

Other researchers develop explicit semantic model for this problem [Fahrni et al., 2011]. They regard the entity linking problem as an information retrieval problem and believe that the cross-lingual entity linking is the extreme case of the

[1] http://www.imakenews.com/lweaver/e_article001189962.cfm?x=bdhk3s9,b4wmV4K7

vocabulary mismatch problem. To overcome this, they map texts with respect to the given Wikipedia concepts using ESA method [Gabrilovich and Markovitch, 2007] and do entity linking in the space of Wikipedia concept. Multilingual knowledge base is needed for this method. However, the size of different knowledge base is different. For example, the English version Wikipedia in 2009 contains 3,000,000 articles, while the Chinese version Wikipedia only contain 300,000 articles, which is far small from the English version. This will cause the problem that large percentage of English Wikipedia concept have no corresponding Chinese concept, thus influence the performance of the method.

Inspired by the idea of using external data source mentioned above, we present a general framework for building cross lingual entity linking system with hidden topics discovered from large-scale parallel corpora. Wikipedia, which was launched in 2001, has become a very big multilingual data resource. In Wikipedia, each article describes a concept and each concept is usually described by different languages. The articles written in different language but describing the same concept are quite related in their hidden topic. It is interesting to use topic modeling algorithms to mine cross lingual topics from Wikipedia. Topic modeling has been widely used by various text applications [Phan et al., 2008; Titov and McDonald, 2008]. Typical topic modeling methods include LDA [Blei et al., 2003]. LDA can be seen as a typical probabilistic approach to latent topic computation. Each topic is represented by a distribution of words (probabilistic language model) and each word has a probability score used to measure its contribution to the topic.

In this paper, we propose a novel bilingual topic model to modeling bilingual topics from Wikipedia data in two different languages. A topic is inherently bilingual: each has two types of representation and each representation corresponds to one language. The same Wikipedia concept described by different language follows that constrain of sharing one identical topic distribution. Based on this framework, we represent the query and the candidate entity using as the distribution of semantic topic. The similarity of a query and an candidate entity is computed as the cosine similarity between their vectors in the semantic space. This is illustrated in Figure 1. Experiments on the standard cross-lingual entity linking dataset show that our approach is promising. Different from previous research work, our approach does not require additional linguistic resources like dictionaries or translation tools. In addition, once the topic model is estimated, it can be applied to other cross lingual task provided that they are consistent.

The reminder of this paper is organized as follows. In the next section we briefly discuss related work on entity linking task. Section 3 describes our bilingual topic model in detail. Cross-lingual entity linking is presented in Section 4. We then present the experimental results in section 5. Section 6 concludes the paper and presents the future work.

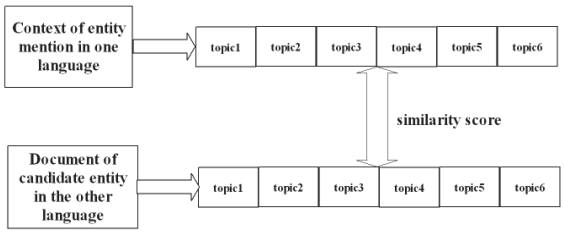

Figure 1. Illustration of our proposed method

2 Related Work

The first type of research related to our work is about entity linking. Many researchers developed entity linking systems based on the context similarity between the query and the candidate entities. Various approaches represent different methods in computing the similarity. Cucerzan [2007] developed a disambiguation approach based on the agreement between the contextual information of candidate entities and the query, as well as the agreement among the category tags associated with candidate entities. Bunescu and Pasca [2006] trained a disambiguation SVM kernel on the Wikipedia data, the similarity between query and candidate entity was computed based on the SVM kernel, the linking system thus selected the entity with the highest similarity as the answer. Han and Zhao [2009] constructed a large-scale semantic network from Wikipedia, and computed the similarity between query and candidate entity based on the constructed semantic network. Radford et al. [2010] applied a Graph-Based ranking method in their entity linking system, context entities were taken into account in order to reach a global optimized solution. Han and Sun [2011] proposed a generative probabilistic model, which can leverage the entity popularity knowledge, name knowledge, and context knowledge for the entity linking task. Other researchers developed supervised learning approach for the entity linking task. Milne and Witten [2008] proposed a machine learning approach to disambiguation that use the links found within Wikipedia articles for training. The two main features they used in their classifier were the commonness of each candidate entity and its relatedness to the surrounding context. Dredze et al. [2010] employed a learning to rank method to disambiguate candidate entities. A comprehensive feature set was used in their system. Zhang et al. [2010] proposed a method to automate generate a large-scale corpus, which were used to train a binary classifier. Zhang et al. [2011] employed a Wikipedia-LDA model to model the contexts as the probability distributions over Wikipedia categories and used this model to compute the similarity between the query and the candidate entities. Such semantic similarity was considered as an important feature in their framework. Dai et al. [2011] employed a Markov logic network in handling this task, but they only conducted experiments in biomedical domain. Shen et al. [2012] proposed a system called LINDEN that can leverage the rich semantic knowledge embedded in the Wikipedia and the taxonomy of the knowledge base. For cross-lingual entity linking task, McNamee et al. [2011] use statistical machine

translation tool to first translate the original document contexts into English equivalents and transform the cross-lingual problem into a monolingual entity linking problem. Fahrni et al. [2011] they map texts with respect to the given Wikipedia concepts and do entity linking in the space of Wikipedia concept.

The second type of research related to our work is about the use of topic model in text application. Phan et al. [2008] utilized the LDA models estimated from Wikipedia to help classify short text segment. Ni et al. [2011] utilized LDA models to extract topics from Wikipedia documents and use these topics to do cross lingual text classification. Titov and McDonald [2008] present a novel multi-grain topic model for extracting the ratable aspects of objects from online user reviews. Similay to our work in [Mimmo et al., 2009], they propose a polylingual topic model that discovers topics aligned across multiple languages and demonstrate its usefulness in supporting machine translation..

3 Bilingual Topic Modeling

As discussed in the preceding section, our goal is to provide a method for computing similarity score between the context of entity mention and the document of an entity in different languages. We use bilingual topic model method, which represent document as mixtures of latent topics, to address this problem. In this section, we will first introduce LDA briefly. And then, we describe the bilingual topic model and its estimation method.

3.1 Latent Dirichlet Allocation (LDA)

We will give a brief introduction to Latent Dirichlet Allocation (LDA). LDA is a generative graphical model as shown in Figure 2. It can be used to model and discover underlying topic structure of any kind of discrete data in which text is a typical example. In LDA, generation of a collection is started by sampling a word distribution φ from a prior Dirichlet distribution dir($\vec{\beta}$) for each latent topic. Then a document is first generated by picking a distribution over topics $\vec{\vartheta}$ from a dirichlet distribution dir($\vec{\alpha}$), which determines topic assignment for words in that document. Then the topic assignment for each word is performed by sampling a topic from topic distribution Mult($\vec{\theta}$). And finally, a particular word w is generated from the distribution of multionmial distribution.

According to the generative graphical model depicted in Figure 2, the joint distribution of all known and hidden variables can be written as:

$$p(\vec{w}_m, \vec{z}_m, \vec{\theta}_{\mathrm{m}}, \Phi|\vec{\alpha}, \vec{\beta}) = p(\Phi|\vec{\beta}) \prod_{n=1}^{N_m} p\left(w_{m,n}|\vec{\varphi}_{z_{m,n}}\right) p\left(z_{m,n}|\vec{\theta}_{\mathrm{m}}\right) p(\vec{\theta}_{\mathrm{m}}|\vec{\alpha})$$

And the marginal distribution of a document $\vec{w}_m$ can be obtained by integrating over $\vec{\theta}_{\mathrm{m}}$ and Φ, and summing over z as follows

$$p(\vec{w}_m|\vec{\alpha}, \vec{\beta}) = \iint p(\vec{\theta}_{\mathrm{m}}|\vec{\alpha}) p(\Phi|\vec{\beta}) \prod_{n=1}^{N_m} \sum_{z=1}^{K} p(w_{m,n}|z) p(z|\vec{\theta}_{\mathrm{m}})\, d\Phi d\vec{\theta}_{\mathrm{m}}$$

Finally, the probability of the whole collection W=$\{\vec{w}_1, \vec{w}_2, \ldots, \vec{w}_M\}$ can be obtained by taking the product of the likelihood of all single document as follows:

$$p(W|\vec{\alpha}, \vec{\beta}) = \prod_{m=1}^{M} p(\vec{w}_m|\vec{\alpha}, \vec{\beta})$$

The parameters of LDA can be estimated by directly and exactly maximizing the likelihood of the whole data collection. However, exact inference is intractable. The solution to this is to use approximate estimation methods such as Variational EM and Gibbs Sampling [Heinrich, 2005]. Gibbs Sampling is a special case of Markov-chain Monte Carlo (MCMC) and often yields relatively simple algorithms for approximate inference in high-dimensional models such as LDA.

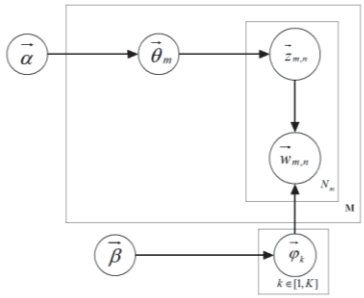

Figure 2. Graphical Model representation of LDA

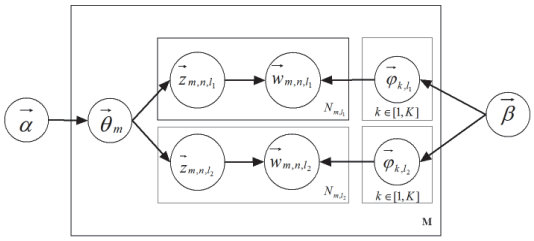

Figure 3. Graphical Model representation of BLDA

K: the number of hidden (latent) topics
M: the total number of documets
$\vec{\alpha}, \vec{\beta}$:hyper parameters for Dirichlet distribuation
$\vec{\theta}_{\mathrm{m}}$:topic distribuation for document m
N_m: the length of document m
$z_{m,n}$: topic index of the n-th word in document m
$w_{m,n}$: a particular word in document m
$\vec{\varphi}_k$: word distribuation for topic k
$\vec{w}_m$: a particular document

Table 1: Notations of LDA

3.2 Bilingual Latent Dirichlet Allocation

In this section, we introduce bilingual Latent Dirichlet Allocation model. The bilingual LDA model is an extension of latent Dirichlet allocation (LDA) for modeling bilingual document tuple. Each tuple is a pair of documents that are semantically related to each other, but written in different languages. We assume all documents in a tuple share the same topic distribution. Also, BLDA assumes that each "topic" consists of a set of discrete distributions over words, which corresponds each language. Figure 2 presents the graphical model of BLDA.

A new document is generated by first drawing a tuple-specific topic distribution from a Dirichlet prior. Then for each word in document, a latent topic distribution is drawn from multinomial distribution. Finally, the observed word is drawn using the language-specific topic multinomial distribution. The generation process of BLDA is shown in Figure 4

For each tuple-specifc topic z, z=1,2,…,K
  For each language l
    Sample word distribution $\vec{\varphi}_{k,l}$
For each tuple in a corpus
  Sample topic distribution $\vec{\theta}_{\mathrm{m}}$
  For each document $\vec{w}_{m,l}$ in tuple
    For each word $w_{m,n,l}$ in document $\vec{w}_{m,l}$
      Sample topic $z_{m,n,l}$ from distribution Mult($\vec{\theta}_{\mathrm{m}}$)
      Sample word $w_{m,n,l}$ from distribution Mult($\vec{\varphi}_{z_{m,n,l}}$)

Figure 4. Generation Process of BLDA

From the generative process in Figure 3, we can write the joint distribution of all known and hidden variables given the Dirichlet parameters for a specific language j as follows:

$$p(\vec{w}_{m,j}, \vec{z}_{m,j}, \vec{\theta}_{\mathrm{m}}, \Phi_j | \vec{\alpha}, \vec{\beta}) = p(\Phi_j|\vec{\beta}) \prod_{n=1}^{N_{m,j}} p\left(w_{m,n,j} \middle| \vec{\varphi}_{z_{m,n,j}}\right) p(z_{m,n,j}|\vec{\theta}_{\mathrm{m}}) p(\vec{\theta}_{\mathrm{m}}|\vec{\alpha})$$

The marginal distribution of a document w for a specific language j can be obtained as follows:

$$p(\vec{w}_{m,j}|\vec{\alpha}, \vec{\beta}) = \iint p(\vec{\theta}_{\mathrm{m}}|\vec{\alpha}) p(\Phi_j|\vec{\beta}) \prod_{n=1}^{N_{m,j}} p(w_{m,n,j}|\vec{\theta}_{\mathrm{m}}, \Phi_j)\, d\Phi_j\, d\vec{\theta}_{\mathrm{m}}$$

We can obtain the marginal distribution of a tuple as follows:

$$p(\vec{t}_m|\vec{\alpha}, \vec{\beta}) = \int p(\vec{\theta}_{\mathrm{m}}|\vec{\alpha})\, d\vec{\theta}_{\mathrm{m}} \prod_{\vec{w}_{m,j}, j\epsilon L} \int p(\Phi_j|\vec{\beta}) d\Phi_j \prod_{n=1}^{N_{m,j}} \sum_{z=1}^{k} p(w_{m,n,j}|\Phi_{z,j}) p(z|\vec{\theta}_{\mathrm{m}})$$

Where $\vec{t}_m$ denotes a tuple document set.

Finally, the probability of the whole document collection W = $\{\vec{t}_m\}_{\mathrm{m}=1}^{\mathrm{M}}$ is obtained as follows:

$$p(W|\vec{\alpha}, \vec{\beta}) = \prod_{m=1}^{M} p(\vec{t}_m|\vec{\alpha}, \vec{\beta})$$

3.3 BLDA Estimation

In this section, we will present a modification of Gibbs Sampling method for the estimation of BLDA.

The first use of Gibbs Sampling for estimating LDA is reported in [Griffiths and Steyvers, 2004] and a more comprehensive description of this method is from the technical report [Heinrich, 2005]. One can refer to these papers for a better understanding of this sampling technique. Here, we only show the most important formula that is used for topic sampling for words. In order to apply Gibbs Sampling, we need to compute the conditional probability $\mathrm{P}(\mathrm{z}_{\mathrm{m,i,j}} = \mathrm{k}|\vec{\mathrm{z}}_{\neg(\mathrm{m,i}),\mathrm{j}}, \vec{w}_{\mathrm{j}})$, where $\vec{w}_{\mathrm{j}}$ means the vector of all words in language j of the whole data collection D and $\vec{\mathrm{z}}_{\neg(\mathrm{m,i}),\mathrm{j}}$ denotes the vector of topic assignment except the considered word at position i in the document item written in language j of tuple m. The topic assignment for a particular word depends on the current topic assignment of all the other word positions in the same language. More specifically, the topic assignment of a particular word t of language j is sampled from the following multinomial distribution:

$$\mathrm{P}(\mathrm{z}_{\mathrm{m,i,j}} = \mathrm{k}|\vec{\mathrm{z}}_{\neg(\mathrm{m,i}),\mathrm{j}}, \vec{w}_{\mathrm{j}}) = \frac{\mathrm{n}^{\mathrm{t}}_{\mathrm{k},\neg(\mathrm{m,i}),\mathrm{j}} + \beta^{\mathrm{t}}_{\mathrm{j}}}{\sum_{\mathrm{v}=1}^{\mathrm{v_j}}(\mathrm{n}^{\mathrm{v}}_{\mathrm{k,j}} + \beta^{\mathrm{v}}_{\mathrm{j}}) - 1} \frac{\mathrm{n}^{\mathrm{k}}_{\mathrm{m},\neg(\mathrm{i})} + \alpha_{\mathrm{k}}}{\sum_{\mathrm{p}=1}^{\mathrm{k}}(\mathrm{n}^{\mathrm{p}}_{\mathrm{m}} + \alpha_{\mathrm{p}}) - 1}$$

Where $\mathrm{n}^{\mathrm{t}}_{\mathrm{k},\neg(\mathrm{m,i}),\mathrm{j}}$ is the number of times word t of language j is assigned to topic k except the current assignment; $\sum_{\mathrm{v}=1}^{\mathrm{v_j}}(\mathrm{n}^{\mathrm{v}}_{\mathrm{k,j}} + \beta^{\mathrm{v}}_{\mathrm{j}}) - 1$ is the total number of words in language j assigned to topic k except the current assignment; and $\mathrm{n}^{\mathrm{k}}_{\mathrm{m},\neg(\mathrm{i})}$ is the number of words in tuple-unit m assigned to topic k except the current assignment; and $\sum_{\mathrm{p}=1}^{\mathrm{k}} \mathrm{n}^{\mathrm{p}}_{\mathrm{m}} - 1$ is the total number of words in tuple-unit m except the current word t.

After finishing Gibbs Sampling, we can obtain the multionmil parameter sets θ and $\{\Phi_j\}_{\mathrm{j}\epsilon\mathrm{l}}$ as follows:

$$\theta_{m,k} = \frac{n_m^k + \alpha_k}{\sum_{p=1}^{k}(n_m^p + \alpha_p)}$$

$$\varphi_{\mathrm{k,t,j}} = \frac{\mathrm{n}^{\mathrm{t}}_{\mathrm{k,j}} + \beta^{\mathrm{t}}_{\mathrm{j}}}{\sum_{\mathrm{v}=1}^{\mathrm{v_j}}(\mathrm{n}^{\mathrm{v}}_{\mathrm{k,j}} + \beta^{\mathrm{v}}_{\mathrm{j}})}$$

4 Cross-lingual Entity Linking

Choosing an appropriate document-aligned comparable training data is extremely important for mining bilingual topics in bilingual LDA model. The data should large enough and should balanced distributions over words and topics, and more importantly, should deal well with the diversity of

future unseen data. In this section, we will empirically investigate the effectiveness of BLDA for mining bilingual topics from Wikipedia.

4.1 Hidden Topics from Wikipedia

Today, the growth of the web, and in particular Wikipedia, has made comparable text corpora-documents that are topically similar of one another. In our work, we first build a document-aligned comparable corpus from Wikipedia. We downloaded XML copies of all Wikipedia articles in two languages: Chinese, English. The corpus is available at this address: http://download.wikipedia.org. We preprocessed the data by removing tables, references, images and info-boxes. We dropped articles which are written in only one language. Finally, we got a total of 750M with more than 150,000 parallel Wikipedia articles. This collection covers a lot of concepts and domains, such as sport, computer science, internet, business, political and so on. It is reasonable to use it as a bilingual comparable dataset. Our BLDA model is general enough and can be easily applied on multilingual corpus without additional efforts.

We estimated our BLDA models for the Wikipedia data using JGibbLDA, which is a java implementation of LDA using Gibbs Sampling. We set the hyper parameters alpha and beta to be 50/k and 0.1 respectively, where k is the number of topics. The number of topics ranges from 10 to 150, with 10 as the step size. We estimated the model using 1000 Gibbs Sampling iterations. Figure 5 shows some example universal-topics produced by BLDA with K=100. We oberved that the outputs are impressive and satisfy our exception.

T2:	water chemical carbon gas hydrogen compound element oxygen air energy
T2	水(water) 化学(chemical) 元素(element) 气体(gas) 化合物(compound)
T14	Window system computer software user network version microsoft version file
T14	系统(system) 版本(version) 软件(software) 电脑(computer) 网络(network)
T25	nba team season race championship driver game car club player league won
T25	比赛(race) 球员(player) 冠军(championship) 球队(team) 赛季(season)
T43	Music opera instrument musical composer string symphony orchestra piano
T43	音乐(music) 歌剧(opera) 贝多芬(beethoven) 交响曲(symphony) 钢琴(piano)
T64	freud psychology martial psychoanalysi alsace patient therapy gestalt wundt
T64	心理学(psychology) 弗洛伊德(freud) 精神(spirit) 潜意识(subconsciousness)

Figure 5. Words with highest probability produced by BLDA

4.2 Topic Inference for New Documents

Given a new document in language j, we need to do topic inference using our BLDA model. Gibbs Sampling is performed for topic inference, however, the number of sampling iterations for inference is much smaller than that for the parameter estimation. We observed that 20 iterations are enough for topic inference.

Let $\vec{w}_j$ and $\vec{z}_j$ be the vectors of all words and their topic assignment of the document-aligned comparable corpus in language j. And $\vec{w}_j^d$ and $\vec{z}_j^d$ are the vectors of all words and their topic assignment of the new document. The topic assignment for a particular word in t in $\vec{w}_j^d$ depends on the current topics of all the other words in $\vec{w}_j^d$ and the topics of all words in $\vec{w}_j$ as follows:

$$P\left(z_{i,j}^d = k | \vec{z}_{\neg i,j}^d, \vec{w}_{\neg i,j}^d, \vec{z}_j, \vec{w}_j\right) = \frac{n_{k,j}^t + \beta_j^t + n_{\neg i,k,j}^{d,t}}{\sum_{v=1}^{v_j}\left(n_{k,j}^v + \beta_j^v + n_{k,j}^{d,v}\right) - 1} \frac{n_{\neg i}^{d,k} + \alpha_k}{\sum_{p=1}^{k}\left(n^{d,p} + \alpha_p\right) - 1}$$

Where $n_{\neg i,k,j}^{d,t}$ is the number of times t is assigned to topic k in d except the current assignment; $n_{\neg i}^{d,k}$ is the number of words in d assigned to topic k except the current assignment; $\sum_{v=1}^{v_j} n_{k,j}^{d,v} - 1$ is the number of words in d that are assigned to topic k except the current assignment; $\sum_{p=1}^{k} n^{d,p} - 1$ is the total number of words in d except the current word t.

After performing topic sampling, the document in any language can be represented as $\vec{\theta}^d = \{\theta_1^d, \theta_2^d, \ldots, \theta_k^d\}$, where each distribution is computed as follows:

$$\theta_k^d = \frac{n^{d,k} + \alpha_k}{\sum_{p=1}^{k}(n^{d,p} + \alpha_p)}$$

4.3 Cross Lingual Entity Linking

The first step in our cross lingual entity linking framework is to find candidate entity in a Knowledge Base for a given entity mention. The candidate entity means the entity in the knowledge base that this entity mention may refer to. We leverage the knowledge sources in Wikipedia to find candidate entity, including "entity pages", "disambiguation pages", "redirect pages" and "anchor text". After obtaining the candidate entity in one language, we can obtained its other language form using existing translated link in Wikipedia. For example, given an entity mention in Chinese "乔丹 (Jordan)", we can obtain its candidate entity in English knowledge such as "Michael Jeffrey Jordan", "Michael I. Jordan" and so on.

The second step is to define a similarity score for each candidate entity. Our system selects the candidate entity with the biggest score as the answer. Even though the document of entity and the context of the query are written in different language, we can represent them in the same semantic topic space using BLDA model. Thus, the similarity score can be computed by the inner product of their topic distribution vector.

5 Experiments and Results

In this section, we evaluate our method in the KBP dataset to demonstrate the effectiveness of our approach.

5.1 Dataset

In our experiment, the knowledge base is derived from Wikipedia pages that contain an infobox. It consists of roughly 818 thousand distinct entities. The KBP 2011 dataset is used as our test dataset. The dataset contains 2176 queries and query document come from news wire and Web pages. Micro-Averaged Accuracy is adopted as the performance

measure in our experiment. It was used in KBP 2011 [Ji et al., 2011] as the official metric to evaluate the performance of entity linking.

5.2 Parameters Turning

There are several parameters that need to be determined in our models. The first is the number of universal-topics. The second is the number of Gibbs Sampling iterations.

The first experiment is to see how entity linking accuracy change if we change the universal-topic number.We estimated many BLDA models for the wikipedia data with different number of topics (from 10 to 150). The change of the entity linking accuracy is depicted in Figure 6.We can see that the optimal results can be obtained with 100 universal-topics. Moreover, the accuracy is quite stable when the number of topic change from 100 to 150.

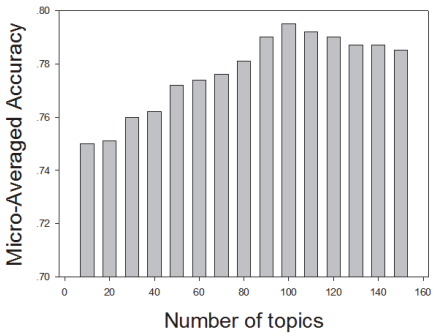

Figure 6. The effect of the number of universal-topic number

The next experiment is to see how Gibbs Sampling influences the performance of the entity linking system. We trained different BLDA models using the Wikipedia data with different numbers of topics. We ran 1000 Gibbs iterations to estimate the parameters, and saved the estimated model at every 200 iterations. At these saving point, we used the estimaed model to performe topic inference for the entity mention document and the candidate entity, and then performed cross-lingual entity linking using their topic distribuation vector. The results are shown in Figure 7. From Figure 7, we can see that the performance of the cross-lingual entity linking system is robust with respect to the number of different numbers of Gibbs Sampling iterations. Although it is hard to control the number of Gibbs Sampling iterations, we found that it yields stable results after 200 iterations.

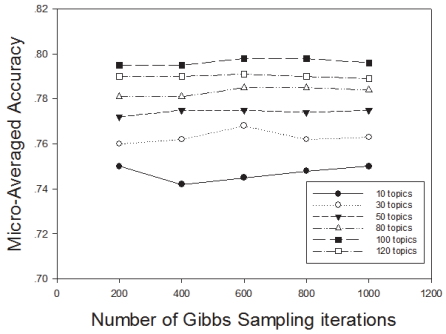

Figure 7. The effect of the number of gibbs sampling iterations

5.3 Comparison with other state-of-art method

In order to further demonstarte the effectiveness of our proposed approach, we compared it with other two other recently proposed methods in KBP 2011. The first approach [McNamee et al., 2011] is a translation based approach. To accomplish the cross-lingual entity linking task, they relied on name transliteration to match the Chinese entities to English names, and they used statistical machine translation and cross-lingual information retrieval to transform entity mention document into English equivalents. The second approach [Fahrni et al., 2011] is an explicit semantic model approach. They first build a multilingual knowledge base from Wikipeida. And then, they use the multilingual knowledege base to obtain a concpet-based language-independent representation of the entity mention document and the candidate entity.

Table 2 shows the comparison results of BLDA approach with translation based approach and explicit semantic model approach. From the table, we can see that the BLDA approach outperformes explicit semantic model approach. This indicates the effectiveness of our BLDA approach. Althrough the translation based approach gets a slighty better result than BLDA method, their method need parallel corpora which are well aligned at word or sentence level. It is expensive to obtain such parallel corpora.

	Micro-Averaged Accuracy
BLDA	0.792
Explicit semantic model	0.785
Translation based method	0.800

Table 2: comparison with other state-of-art methods

6 Conclusion and Future Work

In this paper, we propose a BLDA model to mine bilingual topics from Wikipedia to address the problem of cross-lingual entity linking. Experimental results show that the proposed method is suitable for discovering bilingual topic and obtain competitive results on the standard KBP dataset.

In the future of our work, we plan to investigate the use of category link structure in Wikipedia and combine it into our method.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 61070106, No. 61272332 and No. 61202329), the National High Technology Development 863 Program of China (No. 2012AA011102), the National Basic Research Program of China (No. 2012CB316300), the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDA06030300), We thank the anonymous reviewers for their insightful comments.

References

[Blei et al., 2003] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. Journal of Machine Learning Research, 993-1022, 2003.

[Bunescu and Pasca, 2006] R. Bunescu and M. Pasca. Using Encyclopedic Knowledge for Named Entity Disambiguation. In *proceeding of EACL*, pages 9-16, 2006.

[Cucerzan, 2007] S. Cucerzan. Large-Scale Named Entity Disambiguation Based on Wikipedia Data. In *proceeding of EMNLP*, pages 708-716, 2007.

[Dredze et al., 2010] M. Dredze, P. McNamee, D. Rao, A. Gerber and T. Finin. Entity Disambiguation for Knowledge Base Population. In *proceeding of Coling*, pages 277-285, 2010

[Dai et al., 2011] H. Dai, R.T. Tsai and W. Hsu. Entity Disambiguation Using a Markov-logic Network. In *proceeding of IJCNLP*, pages 846-855, 2011.

[Fahrni et al., 2011] Angela Fahrni, and Michael Strube. HITS' Cross-lingual Entity Linking System at TAC 2011: One Model for All Languages. *In proceeding of Text Analysis Conference*, November 14-15, 2011

[Gabrilovich and Markovitch, 2007] E. Gabrilovich and S. Markovitch. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In *proceeding of IJCAI*, pages 1606-1611.

[Griffiths and Steyvers, 2004] T. Griffiths and M. Steyvers. Finding scientific topics. The National Academy of Sciences, 101:5228–5235, 2004.

[Han and Zhao, 2009] X. Han and J. Zhao. Named Entity Disambiguation by Leveraging Wikipedia Semantic Knowledge. In *proceeding of CIKM*, pages 215-224, 2009.

[Han and Sun, 2011] X. Han and L. Sun. A Generative Entity-Mention Model for Linking Entities with Knowledge Base. In *proceeding of ACL*, pages 945-954, 2011.

[Heinrich, 2005] G. Heinrich. Parameter estimation for text analysis.Technical report, 2005.

[Ji et al., 2011] Heng Ji, Ralph Grishman, and Hoa Dang. Overview of the TAC2011 Knowledge Base Population Track. *In proceeding of Text Analysis Conference*, November 14-15, 2011

[McNamee et al., 2011] Paul McNamee, James Mayfield, Veselin Stoyanov, Douglas W. Oard, Tan Xu, Wu Ke, and David Doermann. Cross-Language Entity Linking in Maryland during a Hurricane. *In proceeding of Text Analysis Conference*, November 14-15, 2011

[Mline and Witten, 2008] D. Milne and I.H. Witten. Learning to Link with Wikipedia. In *proceeding of CIKM*, pages 509-518, 2008.

[Mimmo et al., 2009] D. Mimmo, H.M. Wallach, J. Naradowsky, D.A. Smith, A.McCallum. Polylingual Topic Models. In proceeding of ACL, 2009.

[Ni et al., 2011] X. Ni, J. Sun, J. Hu, and Z. Chen. Cross Lingual Text Classification by Mining Multilingual Topics from Wikipedia. In *proceeding of WSDM*, 2011.

[Phan et al., 2008] Xuan-Hieu Phan, Le-Minh Nguyen, and Susumu Horiguchi. Learning to Classify Short and Sparse Text & Web with Hidden Topics from Large-scale Data Collections. In *proceeding of WWW*, pages 91-100, Beijing 2008.

[Radford et al., 2010] W. Radford, B. Hachey, J. Nothman, M. Honnibal and J.R. Curran. Document-level Entity Linking: CMCRC at TAC 2010. *In proceeding of Text Analysis Conference*, 2010.

[Shen et al., 2012] W. Shen, J. Wang, P. Wang and M. Wang. LINDEN: Linking Named Entities with Knowledge Base via Semantic Knowledge. In *proceeding of WWW*, 2012.

[Titov and McDonald, 2008] Ivan Titov, and Ryan McDonald. Modeling Online Reviews with Multi-grain Topic Models. In proceeding of WWW, April 21–25, Beijing 2008.

[Zhang et al., 2010] W. Zhang, J. Su, C.L. Tan and W.T. Wang. Entity Linking Leveraging Automatically Generated Annotation. In *proceeding of Coling*, pages 1290-1298, 2010.

[Zhang et al., 2011] W. Zhang, J. Sun and C.L. Tan. A Wikipedia-LDA Model for Entity Linking with Batch Size Changing Instance Selection. In proceeding of IJCNLP, pages 562-570, 2011.