Encoding Category Correlations into Bilingual Topic Modeling for Cross-Lingual Taxonomy Alignment

Tianxing Wu[1(✉)], Lei Zhang[2], Guilin Qi[1], Xuan Cui[1], and Kang Xu[1]

[1] School of Computer Science and Engineering, Southeast University, Nanjing, China
{wutianxing,gqi,xcui,kxu}@seu.edu.cn
[2] Institute AIFB, Karlsruhe Institue of Technology, Karlsruhe, Germany
l.zhang@kit.edu

Abstract. Cross-lingual taxonomy alignment (CLTA) refers to mapping each category in the source taxonomy of one language onto a ranked list of most relevant categories in the target taxonomy of another language. Recently, vector similarities depending on bilingual topic models have achieved the state-of-the-art performance on CLTA. However, these models only model the textual context of categories, but ignore explicit category correlations, such as correlations between the categories and their co-occurring words in text or correlations among the categories of ancestor-descendant relationships in a taxonomy. In this paper, we propose a unified solution to encode category correlations into bilingual topic modeling for CLTA, which brings two novel category correlation based bilingual topic models, called **CC-BiLDA** and **CC-BiBTM**. Experiments on two real-world datasets show our proposed models significantly outperform the state-of-the-art baselines on CLTA (at least **+10.9%** in each evaluation metric).

1 Introduction

Over past decades, with the dramatic growth of multilingual knowledge on the Web, aligning knowledge of different languages becomes an important way of realizing globalization of information. Taxonomies are a kind of significant knowledge, which often refers to category hierarchies used for organizing and classifying multilingual big data, and are prevalent on the Web, such as Web site directories (e.g., Dmoz.org) and product catalogues (e.g., eBay product taxonomy). Due to the different grounded languages and intentions of usage, even cross-lingual taxonomies of the same genre are highly heterogenous in linguistics, structure and contents. Hence, to facilitate knowledge sharing across languages, cross-lingual taxonomy alignment (CLTA), which maps each category in the source taxonomy of one language onto a ranked list of most relevant categories in the target taxonomy of another language, is a critical task to solve.

Previous work [2,10,15] on CLTA relies on string similarities based on a translation tool and domain-specific information, such as book instances and financial

C. d'Amato et al. (Eds.): ISWC 2017, Part I, LNCS 10587, pp. 728–744, 2017.
DOI: 10.1007/978-3-319-68288-4_43

calculation items. There are two limitations: 1) string similarities suffer from the vocabulary mismatch problem, i.e., translated texts might be semantically similar even though the specific terms used differ substantially; 2) domain-specific information is often unavailable when aligning cross-lingual and cross-domain taxonomies (e.g., Web site directories and product catalogues).

To overcome these two limitations, our previous work [18] on CLTA introduces a vector similarity based approach relying on bilingual topic models without using any domain-specific information and has achieved the state-of-the-art performance. However, the problem is that these bilingual topic models directly model textual context of categories without considering explicit category correlations. The first category correlation is **co-occurrence correlation**, which exists between the categories and their co-occurring words in text. Some studies such as [9,13] have shown that simultaneously modeling co-occurred metadata (e.g., tags and authors) and text can learn higher-quality topic vectors for many applications. Another important category correlation is **structural correlation**, which means the associations among categories of ancestor-descendant relationships in a taxonomy. The idea of using this kind of correlation is intuitive, that is, if two categories from different taxonomies have similar ancestors or descendants, they may be of high relevance. Thus, we argue that if the above two kinds of category correlations are directly neglected, the topic vector of each category generated by existing bilingual topic models is insufficient to CLTA.

In this paper, we aim to exploit the benefits from both vector similarities and explicit category correlations to deal with the problem of CLTA. Therefore, we try to encode co-occurrence correlations and structural correlations into bilingual topic modeling, which poses two challenges:

- **How to capture both co-occurrence correlations and structural correlations**?
- **How to integrate such explicit category correlations into bilingual topic modeling**?

To solve these challenges, we propose a unified solution to encode category correlations into existing bilingual topic models, i.e., Bilingual Latent Dirichlet Allocation (BiLDA) [17] and Bilingual Biterm Topic Model (BiBTM) [18]. Before applying our solution, we use the same way in [18] to acquire textual context of categories by querying each category with a search engine and constructing paired bilingual documents with a translation tool, which results in a corpus of paired bilingual documents containing all categories. Here, a *modeling object* is defined as a pair of bilingual documents composed of a set of words in BiLDA or a biterm constructed by two distinct words from a pair of bilingual documents in BiBTM. Our solution is to (1) transform the co-occurrence correlations and structural correlations into a prior category distribution of each modeling object, and (2) integrate all prior category distributions into bilingual topic modeling by designing general steps of generating a word in each modeling object. After applying our solution to BiLDA and BiBTM, we obtain two new category correlation based bilingual topic models, called CC-BiLDA and CC-BiBTM. With the topic vector of each category learned by these two models,

we compute vector similarities between the categories of different languages for CLTA.

In summary, the main contributions of this paper are as follows:

- We propose a *unified solution* to encode category correlations into bilingual topic modeling for CLTA, which *leverages the benefits from both vector similarities and explicit category correlations*.
- We design two *new category correlation based bilingual topic models*, CC-BiLDA and CC-BiBTM, by extending BiLDA and BiBTM with our solution. To the best of our knowledge, they are the *first work* on bilingual topic modeling that *simultaneously models bilingual text and its co-occurring categories* to *learn the vector representation* for each category.
- We conduct *experiments* on two real-world datasets and the results show the *effectiveness* of our bilingual topic modes for CLTA, when compared with several state-of-the-art baselines (at least +10.9% in each evaluation metric).

The rest of this paper is organized as follows. Section 2 introduces the background of this work. Section 3 presents the details of two new bilingual topic models by applying our proposed solution. Section 4 gives the experimental results. Section 5 outlines some related work and we conclude in the last section.

2 Preliminaries

In this section, we firstly provide an overview of cross-lingual taxonomy alignment (CLTA) and then discuss the existing bilingual topic models.

2.1 Cross-Lingual Taxonomy Alignment

The wide variety of Web taxonomies from different domains and languages are usually organized in a tree or a directed acyclic graph with categories as nodes. Given two independently created taxonomies of different languages, CLTA aims to map each category in the source taxonomy of one language to the most relevant category in the target taxonomy of another language. The key to CLTA is to measure the relevance between each category in the source taxonomy and its candidate matched categories in the target taxonomy.

Since categories usually do not have textual information to describe themselves, some strategies can be used for getting the textual context of categories in different languages, e.g., by utilizing Wikipedia as an intermediate source and following the interwiki links from one language to another [2] or by querying each category using a search engine and constructing paired bilingual documents by a translation tool [18]. To measure the relevance between categories for CLTA, bilingual topic models, such as BiLDA [17] and BiBTM [18], have been introduced to learn the vector representations of categories from their textual context, which will be discussed in Sect. 2.2. After obtaining the topic distribution of each category, the relevance score between one category in the source taxonomy and another one in the target taxonomy can be computed based on the topic vectors of categories in the same topic space.

2.2 Bilingual Topic Modeling

BiLDA and BiBTM are two existing bilingual topic models and a main difference between them is their modeling objects. BiLDA models paired bilingual documents, each of which is a pair of documents of similar contents but in different languages, such as two Wikipedia articles in different languages interlinked by Wikipedia's language links or a document in one language and its translated version in another language. The generation of a word in a pair of bilingual documents is defined by firstly drawing a topic from a topic distribution of this pair of bilingual documents, and then drawing a word from the topic-word distribution of some language.

BiBTM was proposed to model paired bilingual short documents because BiLDA suffers from the data sparsity problem [6] when documents are short. The modeling objects in BiBTM are biterms, which are unordered word-pairs occurring in a pair of bilingual documents. Any two distinct words in a pair of bilingual documents compose a biterm. For example, given a pair of bilingual documents (d^s, d^t), in which d^s and d^t respectively consist of n distinct words of language s and m distinct words of language t, totally $C_n^2 + C_m^2 + m \times n$ biterms will be generated, where C_m^2 and C_n^2 represent the binomial coefficients. To generate a word in each biterm, BiBTM first draws a topic from a global topic distribution of all biterms, and then draws a word from the topic-word distribution of some language.

3 Models

In this section, we first present an overview of our unified solution to encode category correlations into bilingual topic modeling for CLTA, and then discuss the details of two novel category correlation based bilingual topic models CC-BiLDA and CC-BiBTM resulting from the proposed solution.

3.1 Overview

To perform CLTA, we first learn the vector representations of all categories in the two given taxonomies of different languages using bilingual topic models, where each category can be represented as a topic vector. Then we compute the relevance between each category in the source taxonomy and its candidate matched categories in the target taxonomy using the cosine similarity between the vectors in the same topic space. Since the training of topic models needs large-scale corpus, we apply the same strategy used in [18] to query each category with a search engine to acquire its textual context (i.e., returned snippets). After translating each snippet into another language with a translation tool, each category corresponds to a set of paired bilingual documents and each pair contains at least the given category (maybe more categories) in text. This results in a corpus of paired bilingual documents containing all categories.

Based on the corpus, the previous work [18] first learns the word distribution in BiLDA or the biterm distribution in BiBTM for each topic, and then perform

an additional step of topic inference to derive the topic vector for each category. In contrast, we explicitly model each category such that it allows further encoding various category correlations into bilingual topic modeling. In this work, we mainly consider two types of correlations: (1) co-occurrence correlations between the categories and their co-occurring words in text; (2) structural correlations among the categories of ancestor-descendant relationships in a taxonomy.

To capture co-occurrence correlations, we denote each modeling object (i.e., a pair of bilingual documents in BiLDA or a biterm in BiBTM) as a mixture of categories when the words in the modeling object co-occur with these categories in paired bilingual documents. Such a mixture serves as a prior category distribution of each modeling object. Concerning structural correlations among categories, we leverage information content [14] and path length in the taxonomic structure to improve the prior category distribution (the details of computing the prior category distribution of each modeling object are given in Sect. 3.3).

With both co-occurrence correlations and structural correlations encoded in the prior category distribution of each modeling object, we then integrate them into bilingual topic modeling. Since we need to utilize the low-dimensional topic vector of each category for CLTA, connections between explicit categories and latent topics have been built by supposing there exists a probability distribution over topics for each category, i.e., each category is treated as a mixture of topics. Similar to existing methods, for each language, we represent each topic with a mixture of words in that language. Therefore, we design *general steps* of generating a word in each modeling object as follows:

(1) Drawing a category from the prior category distribution of a modeling object;
(2) Drawing a topic from the category-topic distribution;
(3) Drawing a word from the topic-word distribution of some language.

With the above solution, we obtain two novel category correlation based bilingual topic models, CC-BiLDA and CC-BiBTM, which will be discussed in detail in the following sections.

3.2 Generative Processes

Firstly, we introduce some notations and the generative processes of CC-BiLDA and CC-BiBTM.

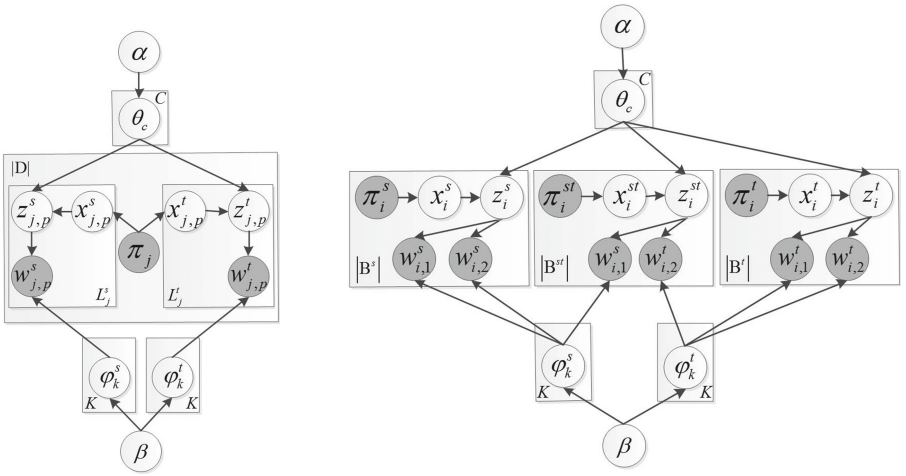
Given a corpus $\mathbb{O}$, suppose it contains $|\mathbf{D}|$ pairs of bilingual documents, $|\mathbf{B}|$ biterms and C explicit categories from two taxonomies to be aligned, which are of different languages. All paired bilingual documents are denoted by $\mathbf{D} = \{d_j\}_{j=1}^{|\mathbf{D}|} = \{(d_j^s, d_j^t)\}_{j=1}^{|\mathbf{D}|}$, where d_j represents a pair of bilingual documents composed of document d_j^s of length L_j^s in language s and document d_j^t of length L_j^t in language t, and a word in position p of d_j^s (or d_j^t) is denoted by $w_{j,p}^s$ (or $w_{j,p}^t$). All biterms are denoted by $\mathbf{B} = \mathbf{B}^s \cup \mathbf{B}^{st} \cup \mathbf{B}^t = \{b_i^s\}_{i=1}^{|\mathbf{B}^s|} \cup \{b_i^{st}\}_{i=1}^{|\mathbf{B}^{st}|} \cup \{b_i^t\}_{i=1}^{|\mathbf{B}^t|}$, where $b_i^s = (w_{i,1}^s, w_{i,2}^s)$ contains two words in language s, $b_i^{st} = (w_{i,1}^s, w_{i,2}^t)$ contains two words in different languages s and t, $b_i^t = (w_{i,1}^t, w_{i,2}^t)$ contains two words in language t.

Algorithm 1. Generative Process of CC-BiLDA

initialize: (1) set the number of topics K;
(2) set values for Dirichlet priors α and β;
foreach *topic* $k \in [1, K]$ **do**
  sample: $\boldsymbol{\varphi}_k^s$, $\boldsymbol{\varphi}_k^t \sim Dirichlet(\beta)$;
foreach *category* $c \in [1, C]$ **do**
  sample: $\boldsymbol{\theta}_c \sim Dirichlet(\alpha)$;
foreach *pair of bilingual documents* $d_j = (d_j^s, d_j^t)$ **do**
  given the prior category distribution $\boldsymbol{\pi}_j$,
  foreach *word position* $p \in d_j^s$ **do**
    sample: $x_{j,p}^s \sim Multinomial(\boldsymbol{\pi}_j)$;
    sample: $z_{j,p}^s \sim Multinomial(\theta_{x_{j,p}^s})$;
    sample: $w_{j,p}^s \sim Multinomial(\varphi_{z_{j,p}^s}^s)$;
  foreach *word position* $p \in d_j^t$ **do**
    sample: $x_{j,p}^t \sim Multinomial(\boldsymbol{\pi}_j)$;
    sample: $z_{j,p}^t \sim Multinomial(\theta_{x_{j,p}^t})$;
    sample: $w_{j,p}^t \sim Multinomial(\varphi_{z_{j,p}^t}^t)$;

Like BiLDA and BiBTM, the modeling objects in CC-BiLDA and those in CC-BiBTM are respectively paired bilingual documents and biterms. Since we define each modeling object as a mixture of categories, a pair of bilingual documents d_j is represented with a C-dimensional multinomial distribution $\boldsymbol{\pi}_j = \{\pi_{j,c}\}_{c=1}^C$ and a biterm b_i is represented with a C-dimensional multinomial distribution $\boldsymbol{\pi}_i = \{\pi_{i,c}\}_{c=1}^C$, also expressed as $\boldsymbol{\pi}_i^s$, $\boldsymbol{\pi}_i^{st}$ and $\boldsymbol{\pi}_i^t$ to distinguish three kinds of biterms. $\boldsymbol{\pi}_j$ and $\boldsymbol{\pi}_i$ serve as the prior category distributions of each pair of bilingual documents d_j in CC-BiLDA and each biterm b_i in CC-BiBTM, respectively. Let $x \in [1, C]$ be the category indicator variable, which is denoted by x^s, x^{st} and x^t respectively for biterms (or words in paired bilingual documents) in language s, biterms composed of two words in different languages s and t, and biterms (or words in paired bilingual documents) in language t. Similarly, the topic indicator variable $z \in [1, K]$ is denoted by z^s, z^{st} and z^t. Then, each category is expressed over K latent topics, which are also expressed over W^s and W^t distinct words of language s and language t, respectively. We use a K-dimensional multinomial distribution $\boldsymbol{\theta}_c = \{\theta_{c,k}\}_{k=1}^K$ to describe the topics of each category c. Regarding the word distributions of languages s and t for topic k, they are respectively represented by a W^s-dimensional multinomial distribution $\boldsymbol{\varphi}_k^s$ with entry $\varphi_{k,w^s}^s = P(w^s|z = k)$ and a W^t-dimensional multinomial distribution $\boldsymbol{\varphi}_k^t$ with entry $\varphi_{k,w^t}^t = P(w^t|z = k)$. Following the convention of bilingual topic modeling, the hyperparameters α and β are the symmetric Dirichlet priors.

With the summarized general steps of generating a word in each modeling object (introduced in Sect. 3.1), the generative processes of CC-BiLDA and CC-BiBTM are respectively given in Algorithms 1 and 2, and their graphical representations are shown in Fig. 1.



(a) CC-BiLDA (b) CC-BiBTM

Fig. 1. Graphical representations of our models

Algorithm 2. Generative Process of CC-BiBTM

initialize: (1) set the number of topics K;
(2) set values for Dirichlet priors α and β;
foreach *topic* $k \in [1, K]$ **do**
  sample: $\boldsymbol{\varphi}_k^s$, $\boldsymbol{\varphi}_k^t \sim Dirichlet(\beta)$;
foreach *category* $c \in [1, C]$ **do**
  sample: $\boldsymbol{\theta}_c \sim Dirichlet(\alpha)$;
foreach *biterm* $b_i^s \in \mathbf{B}^s$ **do**
  given the prior category distribution $\boldsymbol{\pi}_i^s$,
  sample: $x_i^s \sim Multinomial(\boldsymbol{\pi}_i^s)$, $z_i^s \sim Multinomial(\theta_{x_i^s})$;
  sample: $w_{i,1}^s$, $w_{i,2}^s \sim Multinomial(\varphi_{z_i^s}^s)$;
foreach *biterm* $b_i^{st} \in \mathbf{B}^{st}$ **do**
  given the prior category distribution $\boldsymbol{\pi}_i^{st}$,
  sample: $x_i^{st} \sim Multinomial(\boldsymbol{\pi}_i^{st})$, $z_i^{st} \sim Multinomial(\theta_{x_i^{st}})$;
  sample: $w_{i,1}^s \sim Multinomial(\varphi_{z_i^{st}}^s)$, $w_{i,2}^t \sim Multinomial(\varphi_{z_i^{st}}^t)$;
foreach *biterm* $b_i^t \in \mathbf{B}^t$ **do**
  given the prior category distribution $\boldsymbol{\pi}_i^t$,
  sample: $x_i^t \sim Multinomial(\boldsymbol{\pi}_i^t)$, $z_i^t \sim Multinomial(\theta_{x_i^t})$;
  sample: $w_{i,1}^t$, $w_{i,2}^t \sim Multinomial(\varphi_{z_i^t}^t)$;

3.3 Computing Prior Category Distribution

Now we present our method to compute the prior category distribution $\boldsymbol{\pi}$ of each modeling object by leveraging different category correlations. With the strategy resulting in the corpus of paired bilingual documents as introduced in Sect. 3.1,

each category from a taxonomy occurs in a set of paired bilingual documents. In other words, each modeling object corresponds to one or more categories, which are defined as the *co-occurring categories* of the modeling object. The category distribution over each modeling object reflects the co-occurrence correlation between all words in the modeling object and its co-occurring categories. Here, we simply assume that each co-occurring category of a modeling object has the same probability to be sampled. Given the jth modeling object R and the set of its co-occurring categories, denoted by $CC(R)$, the prior category probability $\pi_{j,c}^{CC}$ of each category $c \in [1, C]$ for R based on the co-occurrence correlation is computed as

$$\pi_{j,c}^{CC} = \begin{cases} \frac{1}{|CC(R)|}, & \text{if } c \in CC(R) \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where $|CC(R)|$ is the number of categories in $CC(R)$.

Besides the co-occurrence correlation between words and categories, we introduce two kinds of structural correlations among the categories of ancestor-descendant relationships in a taxonomy. The first structural correlation is based on information content [14]. The intuition is that co-occurring categories of a modeling object should have different importance since they may convey different amounts of information in a taxonomic structure. Similar to [12,14], we argue that the more abstract a category (i.e., more closer to the root of a taxonomy), the lower its information content, or there would be no need to further differentiate it with descendant categories. Thus, more specific co-occurring categories with more information content are more important for a modeling object. For jth modeling object R, we calculate the category probability $\pi_{j,c}$ of each category c by incorporating the intrinsic information content (IIC) measure [14] based on the set of descendants of c in the taxonomy T, denoted by $DES(c)$, as

$$\pi_{j,c} = IIC(c) \cdot \pi_{j,c}^{CC} \tag{2}$$

$$IIC(c) = 1 - \frac{\log\left(|DES(c)| + 1\right)}{\log N_T} \tag{3}$$

where $|DES(c)|$ is the number of categories in $DES(c)$ and N_T is the number of all categories in T. An imaginary root is created for each given taxonomy so as to avoid 0 IIC values of actual categories. We then normalize $\sum_c \pi_{j,c} = 1$.

The second structural correlation is based on path length. We find that the ancestors (in a taxonomy) of co-occurring categories for each modeling object might be also relevant to it. For example, if a pair of bilingual documents has category "*Computer Vision*", its ancestor categories such as "*Artificial Intelligence*" are also relevant to this pair of bilingual documents. Hence, we treat the ancestors of co-occurring categories similarly w.r.t. a modeling object and also assign prior probabilities to them. Given the jth modeling object R, the intuition is that the greater distance in the taxonomy between a co-occurring category c_c and its ancestor c_a, the lower probability of c_a being a relevant category of R. Based on that, we use the shortest path length $SPL(c_c, c_a, T)$ (counted by

Algorithm 3. Prior Category Distribution Updating

Input: the *jth* modeling object R, its category distribution $\boldsymbol{\pi}_j$, *and the set of co-occurring categories* $CC(R)$

Output: updated $\boldsymbol{\pi}_j$

Sort all categories $c_1, \cdots, c_{|CC(R)|}$ *in* $CC(R)$ *as* $c'_1, \cdots, c'_{|CC(R)|}$ *in descending order according to* $\boldsymbol{\pi}_j$;

for $i = 1, \cdots, |CC(R)|$ **do**

  foreach *ancestor* c_a *of* c'_i **do**

    if $PP(c'_i, c_a) > \pi_{j,c_a}$ **then**

      $\pi_{j,c_a} = PP(c'_i, c_a)$

Normalize $\sum_c \pi_{j,c} = 1$

edge numbers in the taxonomy T) between c_c and c_a to measure the propagation probability (PP) from c_c to c_a as

$$PP(c_c, c_a) = \pi_{j,c_c} \cdot \frac{1}{SPL(c_c, c_a, T) + 1} \tag{4}$$

where π_{j,c_c} is the prior category probability of c_c for the jth modeling object. As shown in Fig. 2, since an ancestor category *1* can get different propagation probabilities (propagated from category *2*, *4*, *6*), we decide to pick the highest one propagated from all co-occurring categories, and if this propagation probability is higher than the current prior category probability of category *1*, we will make a replacement. However, a co-occurring category *2* also gets a propagation probability (from category *4*), which may be used to replace the current prior category probability of category *2* with a higher value, thereby may lead to the change of the highest propagation probability and prior category probability for category *1*. To ensure each ancestor category can get the highest prior category probability, we first sort co-occurring categories by their current prior category probabilities in descending order, then compute all propagation probabilities and update the prior category probability for each ancestor of each co-occurring category in order. The details are given in Algorithm 3, which is used to update the prior category distribution of each modeling object.

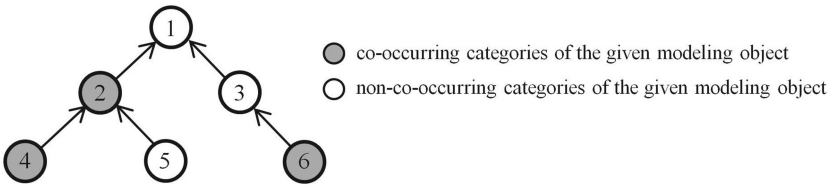


Fig. 2. An example of category locations in a taxonomy

3.4 Parameters Estimation

Since the coupled parameters $\boldsymbol{\theta}_c$, $\boldsymbol{\varphi}_k^s$ and $\boldsymbol{\varphi}_k^t$ in CC-BiLDA (or CC-BiBTM) are intractable to exactly solve, we follow BiLDA [17] and BiBTM [18] to utilize Gibbs Sampling [5] to perform approximate inference. Gibbs Sampling estimates the parameters with the samples drawn from the posterior distributions of latent variables sequentially, which are conditioned on the current values of all other variables and data. Here, we jointly sample latent variables x and z. Due to space limit, we only show the derived Gibbs Sampling formulas for CC-BiLDA and CC-BiBTM. For CC-BiLDA, given jth pair of bilingual documents $d_j = (d_j^s, d_j^t)$ in corpus $\mathbb{O}$, we sample the category c and topic k for the word in position p of document d_j^s in language s (or document d_j^t in language t) as follows:

$$P(x_{j,p}^s = c, z_{j,p}^s = k | x_{\neg(j,s,p)}, z_{\neg(j,s,p)}, \mathbb{O}) \propto \pi_{j,c} \cdot \frac{(n_{\neg(j,s,p),k|c} + \alpha)}{(n_{\neg(j,s,p),\cdot|c} + K\alpha)} \cdot \frac{(n_{\neg(j,s,p),w_{j,p}^s|k} + \beta)}{(n_{\neg(j,s,p),\cdot^s|k} + W^s\beta)} \tag{5}$$

$$P(x_{j,p}^t = c, z_{j,p}^t = k | x_{\neg(j,t,p)}, z_{\neg(j,t,p)}, \mathbb{O}) \propto \pi_{j,c} \cdot \frac{(n_{\neg(j,t,p),k|c} + \alpha)}{(n_{\neg(j,t,p),\cdot|c} + K\alpha)} \cdot \frac{(n_{\neg(j,t,p),w_{j,p}^t|k} + \beta)}{(n_{\neg(j,t,p),\cdot^t|k} + W^t\beta)} \tag{6}$$

In Eq. (5), $x_{j,p}^s$ and $z_{j,p}^s$ are respectively the category assignment and topic assignment for word $w_{j,p}^s$ in the current position. For all words in the corpus except the word in position p of document d_j^s, $x_{\neg(j,s,p)}$ is their category assignments and $z_{\neg(j,s,p)}$ is the topic assignments. $\pi_{j,c}$ means the prior probability of the pair of bilingual documents d_j assigned to category c. Also after excluding the word in position p of document d_j^s, $n_{\neg(j,s,p),k|c}$ is the number of words jointly assigned to category c and topic k, $n_{\neg(j,s,p),\cdot|c} = \sum_k n_{\neg(j,s,p),k|c}$, $n_{\neg(j,s,p),w_{j,p}^s|k}$ denotes the number of times for word $w_{j,p}^s$ assigned to topic k and $n_{\neg(j,s,p),\cdot^s|k} = \sum_{w^s} n_{\neg(j,s,p),w^s|k}$. In Eq. (6), all symbols have the same meaning as those in Eq. (5) after replacing language s with t.

With respect to CC-BiBTM, the Gibbs Sampling formulas for biterms $b_i^s \in \mathbf{B}^s$, $b_i^{st} \in \mathbf{B}^{st}$ and $b_i^t \in \mathbf{B}^t$ are as follows:

$$P(x_i^s = c, z_i^s = k | x_{\neg b_i^s}, z_{\neg b_i^s}, \mathbb{O}) \propto \pi_{i,c}^s \cdot \frac{(n_{\neg b_i^s,k|c} + \alpha)}{(n_{\neg b_i^s,\cdot|c} + K\alpha)} \cdot \frac{(n_{\neg b_i^s,w_{i,1}^s|k} + \beta)(n_{\neg b_i^s,w_{i,2}^s|k} + \beta)}{(n_{\neg b_i^s,\cdot^s|k} + W^s\beta)(n_{\neg b_i^s,\cdot^s|k} + 1 + W^s\beta)} \tag{7}$$

$$P(x_i^{st} = c, z_i^{st} = k | x_{\neg b_i^{st}}, z_{\neg b_i^{st}}, \mathbb{O}) \propto \pi_{i,c}^{st} \cdot \frac{(n_{\neg b_i^{st},k|c} + \alpha)}{(n_{\neg b_i^{st},\cdot|c} + K\alpha)} \cdot \frac{(n_{\neg b_i^{st},w_{i,1}^s|k} + \beta)(n_{\neg b_i^{st},w_{i,2}^t|k} + \beta)}{(n_{\neg b_i^{st},\cdot^s|k} + W^s\beta)(n_{\neg b_i^{st},\cdot^t|k} + W^t\beta)} \tag{8}$$

$$P(x_i^t = c, z_i^t = k | x_{\neg b_i^t}, z_{\neg b_i^t}, \mathbb{O}) \propto \pi_{i,c}^t \cdot \frac{(n_{\neg b_i^t, k|c} + \alpha)}{(n_{\neg b_i^t, \cdot|c} + K\alpha)} \cdot \frac{(n_{\neg b_i^t, w_{i,1}^t|k} + \beta)(n_{\neg b_i^t, w_{i,2}^t|k} + \beta)}{(n_{\neg b_i^t, \cdot^t|k} + W^t\beta)(n_{\neg b_i^t, \cdot^t|k} + 1 + W^t\beta)} \tag{9}$$

where x and z are respectively current category assignment and topic assignment for the given biterm. For all biterms except biterm b, $x_{\neg b}$ is their category assignments and $z_{\neg b}$ denotes the topic assignments. $\pi_{i,c}$ represents the prior category probability of ith biterm $b_i^s \in \mathbf{B}^s$ or $b_i^{st} \in \mathbf{B}^{st}$ or $b_i^t \in \mathbf{B}^t$ assigned to category c. Under the condition of excluding biterm b, $n_{\neg b,k|c}$ is the number of biterms jointly assigned to category c and topic k, $n_{\neg b,\cdot|c} = \sum_k n_{\neg b,k|c}$, $n_{\neg b,w^s|k}$ is the number of times word w^s of language s assigned to topic k, $n_{\neg b,\cdot^s|k} = \sum_{w^s} n_{\neg b,w^s|k}$, $n_{\neg b,w^t|k}$ is the number of times word w^t of language t assigned to topic k, and $n_{\neg b,\cdot^t|k} = \sum_{w^t} n_{\neg b,w^t|k}$.

After a sufficient number of sampling iterations, we can estimate the parameters in CC-BiLDA and CC-BiBTM. Instead of computing all parameters like $\boldsymbol{\varphi}_k^s$ and $\boldsymbol{\varphi}_k^t$, our solution to CLTA only needs $\boldsymbol{\theta}_c$, which is given as follows:

$$\theta_{c,k} = \frac{\alpha + n_{k|c}}{K\alpha + n_c} \tag{10}$$

where n_c is the number of words (or biterms) assigned to category c in CC-BiLDA (or CC-BiBTM), $n_{k|c}$ is the number of words (or biterms) simultaneously assigned to category c and topic k in CC-BiLDA (or CC-BiBTM).

With the topic distribution $\boldsymbol{\theta}_c$ obtained in CC-BiLDA and CC-BiBTM, we can represent categories from two taxonomies of different languages in the same topic space. The relevance score between each category in the source taxonomy and its candidate matched categories (identified with the same method in [18]) in the target taxonomy is computed as the cosine similarity between the topic vectors directly derived from $\boldsymbol{\theta}_c$.

4 Experiments

In this section, we evaluate CC-BiLDA and CC-BiBTM on two real-world datasets for CLTA. The source codes of these two models are publicly available[1].

4.1 Experiment Settings

(a) Datasets. We validated our models to CLTA on two public datasets[2] (also used in [18]), each of which consists of two cross-domain taxonomies of different languages and a set of labeled cross-lingual alignments. The taxonomies in one dataset are two product catalogues respectively extracted from JD.com (one of

[1] https://github.com/143230/CLTA.
[2] https://github.com/jxls080511/080424.

Table 1. Details of each taxonomy in each dataset

Taxonomy	JD.com	eBay.com	Chinese Dmoz.org	Yahoo! Directory
#category	7,741	7,782	2,084	2,353
#paired doc	67,594	72,979	19,277	21,467
#Chinese word	24,483	18,190	11,064	8,581
#English word	15,489	14,729	8,806	8,100

the largest Chinese B2C online retailers) and eBay.com, and those in another dataset are two Web site directories: Chinese Dmoz.org (the largest Chinese Web site directory) and Yahoo! Directory. We got a corpus of paired bilingual documents for each dataset with the strategy in [18], and processed them by word segmentation, stop words removal, stemming, etc. The details of each taxonomy and its extracted corpus of paired bilingual documents are given in Table 1.

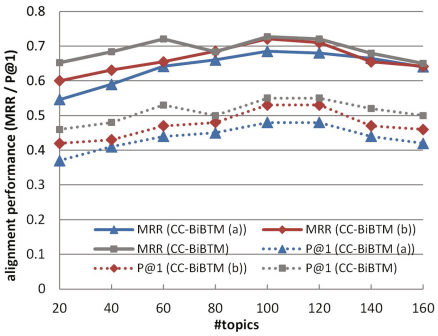
(b) Baselines. We compared our models (i.e., CC-BiLDA and CC-BiBTM) with three kinds of baselines, which are existing bilingual topic models, variants of our models and cross-lingual ontology matching systems. Note that the hyperparameters α and β of all topic models are respectively set to $50/K$ (K is number of topics) and 0.1 according to [18]. All experiments were carried out on a Linux server with Intel Xeon E5-2630 v4 2.20 GHz CPU and 256 GB memory.

- **Existing Bilingual Topic Models:** They are BiLDA and BiBTM introduced in Sect. 2.2. To our knowledge, these two models are the state-of-the-art baselines for CLTA. In BiLDA and BiBTM, we respectively set the topic number K to 80 and 120 based on [18].
- **Variants of Our Models:** The full version of CC-BiLDA and that of CC-BiBTM apply three category correlations to category distribution computation. A kind of variants (denoted as CC-BiLDA (a) and CC-BiBTM (a)) of our models only utilize co-occurrence correlations. Another kind of variants (denoted as CC-BiLDA (b) and CC-BiBTM (b)) use information content based structural correlations besides co-occurrence correlations.
- **Cross-Lingual Ontology Matching Systems:** Although CLTA and cross-lingual ontology matching are different tasks, we can treat the taxonomies as a special kind of ontologies without formally defined properties, instances, axioms, etc. Thus, we took two state-of-the-art cross-lingual ontology matching systems (i.e., AML [4] and LogMap [7]) as the baselines.

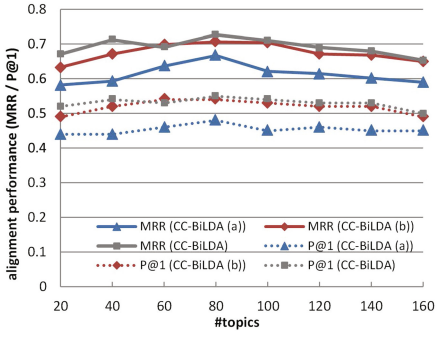
(c) Evaluation Metrics. Similar to the work [2,10,15,18], we used MRR (Mean Reciprocal Rank) [3] and P@1 (precision for the top 1 ranking result) as the evaluation metrics because CLTA is seen as a ranking problem.

4.2 Parameter Tuning

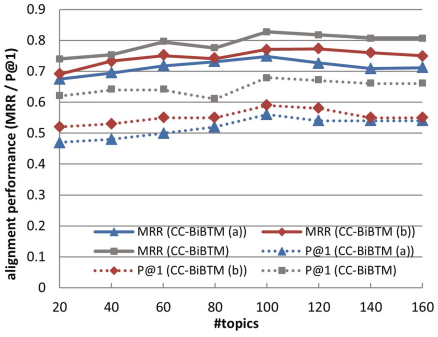
Since different number of topics may lead to different performance in CLTA, we conducted an analysis by varying the number of topics K in our models and their variants. Figure 3 gives the alignment performance of CC-BiBTM, CC-BiLDA and their corresponding variants on each dataset when using different number of topics K. For CC-BiBTM and its variants, MRR or P@1 values reach the peak when K is from 100 to 120 on each dataset (in Fig. 3(a) and (c)), so K was set to 100 in these models for efficient training. For CC-BiLDA and its variants, most of their MRR and P@1 values are the highest when $K = 80$ (in Fig. 3(b) and (d)), so K was empirically set to 80 in these models.



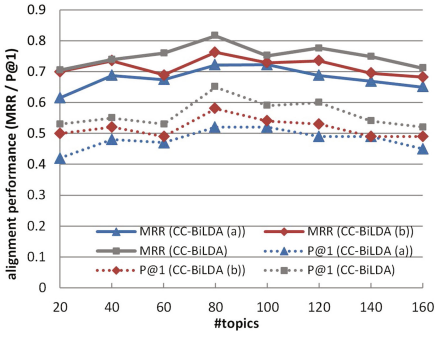
(a) performance of CC-BiBTM and its variants on product catalogues



(b) performance of CC-BiLDA and its variants on product catalogues



(c) performance of CC-BiBTM and its variants on Web site directories



(d) performance of CC-BiLDA and its variants on Web site directories

Fig. 3. Alignment performance vs. number of topics K

4.3 Result Analysis

For each dataset, we trained all topic models with 500 iterations of Gibbs Sampling to converge. Table 2 gives the overall results of our proposed models and the baselines, and we can see that:

Table 2. Overall results

Approach	Product catalogues		Web site directories	
	MRR	P@1	MRR	P@1
AML	0.102	0.100	0.314	0.270
LogMap	0.105	0.100	0.265	0.250
BiLDA	0.553	0.390	0.679	0.480
CC-BiLDA (a)	0.667	0.480	0.721	0.520
CC-BiLDA (b)	0.706	0.540	0.763	0.580
CC-BiLDA	0.720	**0.550**	0.815	0.650
BiBTM	0.597	0.440	0.719	0.520
CC-BiBTM (a)	0.685	0.480	0.748	0.560
CC-BiBTM (b)	0.721	0.530	0.771	0.590
CC-BiBTM	**0.727**	**0.550**	**0.828**	**0.680**

- Our models CC-BiBTM and CC-BiLDA outperform all baselines, especially CC-BiBTM significantly improves the CLTA performance of the state-of-the-art baseline BiBTM (at least **+10.9%** in each evaluation metric). This reflects the value of our solution for encoding correlations into bilingual topic modeling, and the remarkable effects of category correlations on CLTA.
- Cross-lingual ontology matching systems have rather poor performance. Although they are not well tuned for the task of CLTA, it still shows that they cannot work well in real-world CLTA without internal features such as properties, instances and axioms available in ontologies.
- The performance of CLTA improves each time when we encoded one more kind of the proposed category correlations into bilingual topic modeling. It means that the co-occurrence correlations, structural correlations based on information content and those based on path length are all useful to CLTA.
- The performance of CC-BiLDA is close to that of CC-BiBTM. It reveals that although the training corpus are actually paired bilingual short documents, the data sparsity problem suffered by BiLDA has been greatly alleviated via the semantic information of category correlations.

Since the proposed models CC-BiLDA and CC-BiBTM have the best performance on MRR and P@1, we further compared their efficiency of model training by the average running time (per iteration) of CC-BiLDA and CC-BiBTM on

Table 3. Efficiency comparison of CC-BiLDA and CC-BiBTM

Model	Running time (seconds) per iteration		Time complexity per iteration
	Product catalogues	Web site directories	
CC-BiLDA	15.90	10.14	$O(K_1\|D\|\overline{L}_D\overline{C}_D)$
CC-BiBTM	453.31	251.39	$O(K_2\|B\|\overline{C}_B)$

the given datasets in Table 3. We can find that the running time of CC-BiBTM is about 25 times and 29 times of CC-BiLDA on Web site directories and product catalogues, respectively. The time complexity (per iteration) of each model is also shown in Table 3, where the topic number $K_1 = 80$ and $K_2 = 100$ according to Sect. 4.2; $|D|$ is the number of paired bilingual documents, each of which averagely contains $\overline{L}_D$ words and $\overline{C}_D$ co-occurring categories; and $|B|$ is the number of biterms, each of which averagely has $\overline{C}_B$ co-occurring categories. Suppose each document in each pair of bilingual documents averagely has $\bar{l}$ words ($\bar{l} \geq 2$), i.e., $\overline{L}_D \approx 2\bar{l}$, so $|B| \approx |D| \cdot (2 \cdot \frac{\bar{l}(\bar{l}-1)}{2} + \bar{l}^2)$. A biterm may have the co-occurring categories of more than one pair of bilingual documents, so $\frac{\overline{C}_B}{\overline{C}_D} \geq 1$. Since we have $\frac{K_2|B|\overline{C}_B}{K_1|D|\overline{L}_D\overline{C}_D} \approx \frac{5}{4} \cdot \frac{\overline{C}_B}{\overline{C}_D} \cdot (\bar{l} - \frac{1}{2})$, the time complexity of CC-BiBTM is much higher than that of CC-BiLDA. However, with the strategy in [18], the bilingual documents used for CLTA were actually extracted from the snippets (i.e., short documents) returned by a search engine, so the number of words in each document is small (e.g., $\bar{l} = 10.21$ for product catalogues and $\bar{l} = 9.73$ for Web site directories), and the running time of CC-BiBTM is still acceptable.

To sum up, for CLTA, if users have a high demand on accuracy and do not care about the efficiency, we suggest to use CC-BiBTM. If users care more about the efficiency and can accept a little lower accuracy, we recommend CC-BiLDA.

5 Related Work

5.1 Cross-Lingual Schema Matching

The problem of cross-lingual schema matching has been mainly studied in the area of ontology matching and taxonomy alignment. Some approaches or systems [4,7,16] for cross-lingual ontology matching mainly use the features based on string similarities after translation. The performance is often unsatisfactory due to the problems of vocabulary mismatch and improper translations. Different to ontologies, taxonomies do not always have logically rigorous structures with formally defined properties, instances and axioms to help solve matching tasks. Thus, several approaches have been especially designed to CLTA. Some of them [2,10,15] focus on aligning domain-specific taxonomies using string similarities based on a translation tool and domain-specific information. The most relevant work [18] also tries to align cross-lingual and cross-domain taxonomies with bilingual topic models. We improved this work by encoding different explicit category correlations into bilingual topic modeling for CLTA.

5.2 Metadata Topic Models

Topic models such as Latent Dirichlet Allocation (LDA) [1] and its numerous variants are well studied generative models for analysing latent semantic topics in text. Besides bilingual topic models BiLDA and BiBTM, metadata topic models are also related to our work. To simultaneously model the text and its metadata (e.g., authors and tags), a set of metadata topic models have been proposed including Author Topic Model [13], labeled-LDA [11], Tag-Weighted Topic Model [9], Tag-Weighted Dirichlet Allocation [8], etc. They denote each metadata as a mixture of topics or words, but cannot be applied to cross-lingual text mining. Our models CC-BiLDA and CC-BiBTM are the first work of cross-lingual metadata topic models, which already show the superiority in CLTA.

6 Conclusions and Future Work

In this paper, we proposed a unified solution to encode category correlations into bilingual topic modeling for CLTA. Our solution captures different category correlations with a prior category distribution of each modeling object, and integrates such distributions into bilingual topic modeling. This brings two novel category correlation based bilingual topic models CC-BiLDA and CC-BiBTM, which significantly outperform the state-of-the-art baselines on CLTA. In the future, we will apply our models to CLTA in knowledge graphs, to benefit cross-lingual knowledge graph fusion and cross-lingual semantic search.

Acknowledgements. This work is supported in part by the National Natural Science Foundation of China (Grant No. 61672153), the 863 Program (Grant No. 2015AA015406), the Fundamental Research Funds for the Central Universities and the Research Innovation Program for College Graduates of Jiangsu Province (Grant No. KYLX16_0295).

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
2. Boldyrev, N., Spaniol, M., Weikum, G.: ACROSS: a framework for multi-cultural interlinking of web taxonomies. In: WebSci, pp. 127–136 (2016)
3. Craswell, N.: Mean reciprocal rank. In: Liu, L., Tamer Özsu, M. (eds.) Encyclopedia of Database Systems, pp. 1703–1703. Springer, New York (2009). doi:10.1007/978-0-387-39940-9_488
4. Faria, D., Pesquita, C., Santos, E., Palmonari, M., Cruz, I.F., Couto, F.M.: The AgreementMakerLight ontology matching system. In: Meersman, R., Panetto, H., Dillon, T., Eder, J., Bellahsene, Z., Ritter, N., Leenheer, P., Dou, D. (eds.) OTM 2013. LNCS, vol. 8185, pp. 527–541. Springer, Heidelberg (2013). doi:10.1007/978-3-642-41030-7_38
5. Geman, S., Geman, D.: Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. **6**, 721–741 (1984)

6. Hong, L., Davison, B.D.: Empirical study of topic modeling in Twitter. In: SOMA, pp. 80–88 (2010)
7. Jiménez-Ruiz, E., Cuenca Grau, B.: LogMap: logic-based and scalable ontology matching. In: Aroyo, L., Welty, C., Alani, H., Taylor, J., Bernstein, A., Kagal, L., Noy, N., Blomqvist, E. (eds.) ISWC 2011. LNCS, vol. 7031, pp. 273–288. Springer, Heidelberg (2011). doi:10.1007/978-3-642-25073-6_18
8. Li, S., Huang, G., Tan, R., Pan, R.: Tag-weighted Dirichlet allocation. In: ICDM, pp. 438–447 (2013)
9. Li, S., Li, J., Pan, R.: Tag-weighted topic model for mining semi-structured documents. In: IJCAI, pp. 2855–2861 (2013)
10. Prytkova, N., Weikum, G., Spaniol, M.: Aligning multi-cultural knowledge taxonomies by combinatorial optimization. In: WWW, pp. 93–94 (2015)
11. Ramage, D., Hall, D., Nallapati, R., Manning, C.D.: Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In: EMNLP, pp. 248–256 (2009)
12. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. In: IJCAI, pp. 448–453 (1995)
13. Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smyth, P.: The author-topic model for authors and documents. In: UAI, pp. 487–494 (2004)
14. Seco, N., Veale, T., Hayes, J.: An intrinsic information content metric for semantic similarity in WordNet. In: ECAI, pp. 1089–1090 (2004)
15. Spohr, D., Hollink, L., Cimiano, P.: A machine learning approach to multilingual and cross-lingual ontology matching. In: Aroyo, L., Welty, C., Alani, H., Taylor, J., Bernstein, A., Kagal, L., Noy, N., Blomqvist, E. (eds.) ISWC 2011. LNCS, vol. 7031, pp. 665–680. Springer, Heidelberg (2011). doi:10.1007/978-3-642-25073-6_42
16. Trojahn, C., Fu, B., Zamazal, O., Ritze, D.: State-of-the-art in multilingual and cross-lingual ontology matching. In: Buitelaar, P., Cimiano, P. (eds.) Towards the Multilingual Semantic Web, pp. 119–135. Springer, Heidelberg (2014). doi:10.1007/978-3-662-43585-4
17. Vulić, I., De Smet, W., Tang, J., Moens, M.F.: Probabilistic topic modeling in multilingual settings: an overview of its methodology and applications. Inf. Process. Manag. **51**(1), 111–147 (2015)
18. Wu, T., Qi, G., Wang, H., Xu, K., Cui, X.: Cross-lingual taxonomy alignment with bilingual biterm topic model. In: AAAI, pp. 287–293 (2016)