Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)

Cross-Lingual Taxonomy Alignment with Bilingual Biterm Topic Model

Tianxing Wu[1], **Guilin Qi**[1], **Haofen Wang**[2], **Kang Xu**[1] **and Xuan Cui**[1]
[1] Key Laboratory of Computer Network and Information Integration of State Education Ministry,
School of Computer Science and Engineering, Southeast University, China
{wutianxing, gqi, kxu, xcui}@seu.edu.cn
[2] East China University of Science & Technology, China
whfcarter@ecust.edu.cn

Abstract

As more and more multilingual knowledge becomes available on the Web, knowledge sharing across languages has become an important task to benefit many applications. One of the most crucial kinds of knowledge on the Web is taxonomy, which is used to organize and classify the Web data. To facilitate knowledge sharing across languages, we need to deal with the problem of cross-lingual taxonomy alignment, which discovers the most relevant category in the target taxonomy of one language for each category in the source taxonomy of another language. Current approaches for aligning cross-lingual taxonomies strongly rely on domain-specific information and the features based on string similarities. In this paper, we present a new approach to deal with the problem of cross-lingual taxonomy alignment without using any domain-specific information. We first identify the candidate matched categories in the target taxonomy for each category in the source taxonomy using the cross-lingual string similarity. We then propose a novel bilingual topic model, called Bilingual Biterm Topic Model (BiBTM), to perform exact matching. BiBTM is trained by the textual contexts extracted from the Web. We conduct experiments on two kinds of real world datasets. The experimental results show that our approach significantly outperforms the designed state-of-the-art comparison methods.

Introduction

Nowadays, as the advent of more and more multilingual resources, the Web has become a global information space. Thus, sharing knowledge across languages has become an important and challenging task. One of the most crucial kinds of knowledge is taxonomy, which refers to a hierarchy of categories that entities are classified to (Prytkova, Weikum, and Spaniol 2015). Different kinds of taxonomies are everywhere on the Web, such as Web site directory (e.g. Yahoo Directory and Dmoz.org) and product catalogue (e.g. eBay.com and Google Product Taxonomy). To facilitate knowledge sharing across languages, we need to deal with the problem of cross-lingual taxonomy alignment, which is the task of discovering the most relevant category in the target taxonomy of one language for each category in the source taxonomy of another language. Cross-lingual taxonomy alignment not only contributes to globalize knowledge sharing, but also benefits many applications, such as cross-lingual information retrieval (Potthast, Stein, and Anderka 2008; Nguyen et al. 2009) and multilingual knowledge base construction (Lehmann et al. 2014; Mahdisoltani, Biega, and Suchanek 2014).

The key step of aligning cross-lingual taxonomies is to measure the relevance between one category in the source taxonomy and another one in the target taxonomy. Once all the relevance scores have been determined, we can obtain the most relevant category in the target taxonomy for each category in the source taxonomy in an unsupervised way. However, since categories are described in different languages, traditional monolingual similarity metrics are not suitable in cross-lingual scenarios.

In order to overcome this problem, several approaches have been proposed. The work given in (Spohr, Hollink, and Cimiano 2011) first translates cross-lingual taxonomies into monolingual taxonomies, and then captures the linguistic features and structural features that rely on string similarities to predict the relevance score between categories. However, the translated label of a category in the source taxonomy may be dissimilar to its matched category in the target taxonomy. For example, category "锟斤拷锟斤拷/锟剿讹拷锟斤拷" in JD.com can be translated to "*Outdoor/Sportswear*" by Google Translate [1], but the translated string is totally different from that of its matched category "*Athletic Apparel*" in eBay.com. Thus, the features that rely on string similarities are insufficient to decide the relevance score between two categories of different languages, due to different language habits and improper translations.

Another work (Prytkova, Weikum, and Spaniol 2015) tries to solve this problem by referring to Wikipedia. It strongly relies on the domain-specific information (i.e. book intances) to map original categories in book domain onto Wikipedia categories. Categories of different languages can be directly compared using interwiki links. However, this approach cannot be easily extended to the other kinds of taxonomies, because instance information is often unavailable.

In this paper, we study the problem of cross-lingual taxonomy alignment. The problem is non-trivial and poses the

[1] http://translate.google.com/

following challenges.

- **Feature**. When the domain-specific information is unavailable, the existing approach (Spohr, Hollink, and Cimiano 2011) only depends on string similarities to capture different kinds of features, resulting in a rather poor performance. Since vector similarities have achieved great success in natural language processing tasks (Denhière and Lemaire 2004; Gabrilovich and Markovitch 2007; Li, Ji, and Yan 2015), can we introduce a new powerful feature which relies on vector similarities?
- **Representation**. Vector similarities are based on rich textual information, but categories do not contain such information. Can we find a way to enrich the representation of categories with textual information? Can this new representation reveal the real meaning of each category, especially ambiguous categories?
- **Approach**. The features that depend on string similarities do not work well in cross-lingual taxonomy alignment, but they still have positive impacts. Can we design an approach using both the features that rely on vector similarities and the ones that depend on string similarities?

To solve the above challenges, we propose a new approach to solve the problem of cross-lingual taxonomy alignment without using any domain-specific information. Firstly, we identify the candidate matched categories in the target taxonomy for each category in the source taxonomy using a new linguistic feature, i.e. cross-lingual string similarity. Then, we propose a novel bilingual topic model, called Bilingual Biterm Topic Model (BiBTM), to obtain the topic vector of the context for each category. BiBTM is trained by the textual contexts extracted from the Web. Finally, the relevance score between each category in the source taxonomy and its candidate matched categories is computed as the cosine similarity between topic vectors. The experiments on two real world datasets show that our approach significantly outperforms the designed state-of-the-art comparison methods.

The rest of this paper is organized as follows. Section 2 outlines some related work. Section 3 introduces the proposed approach in detail. Section 4 presents the experimental results and finally Section 5 concludes this work and describes the future work.

Related Work

In this section, we review some related work on schema matching and multilingual knowledge alignment.

Schema Matching

Schema matching aims at identifying semantic correspondences between two schemas includes database schemas and ontologies (Do and Rahm 2007). (Rahm and Bernstein 2001; Berlin and Motro 2002; Do, Melnik, and Rahm 2003) introduce different methods for matching database schemas. However, the setting of matching database schemas is quite different from aligning heterogeneous taxonomies due to the difference in size and structure.

Ontology matching (Shvaiko and Euzenat 2006) solves the problem of finding relationships (e.g. equivalence, subsumption) between discrete entities of ontologies, including classes, properties, etc. There exists plenty of work (Euzenat and Shvaiko 2007) on matching different kinds of ontologies. Several systems (Jiménez-Ruiz et al. 2014; Faria et al. 2014) can match multilingual ontologies with the features that only rely on the string similarities after using machine translation, but the performance is not good. The difference between our work and ontology matching is that we focus on aligning general taxonomies rather than standard ontologties. Ontologies are usually of more internal information, such as properties, functions, axioms, .etc. However, the taxonomies handled in this work do not contain such information to assist the matching operation.

Another kind of related work is catalog integration (Agrawal and Srikant 2001; Ichise, Takeda, and Honiden 2003; Wang et al. 2014), but they do not aligning the taxonomies in cross-lingual scenarios.

Multilingual Knowledge Alignment

There exists some work on multilingual knowledge alignment. (Wang et al. 2012) proposed a linkage factor graph model to link articles from English Wikipedia to those in Baidu Baike. They further proposed a concept annotation method and a regression-based learning model to iteratively predict new cross-lingual links (Wang, Li, and Tang 2013). X-LiSA (Zhang and Rettinger 2014) is a semantic annotation system , which can annotate text documents and web pages in different languages using resources from Wikipedia and Linked Open Data. Different from our work, all of the above work only focuses on aligning cross-lingual data level knowledge, i.e. cross-lingual entity linking.

The most relevant work is (Spohr, Hollink, and Cimiano 2011; Prytkova, Weikum, and Spaniol 2015). Both of them depend on domain-specific information and the features based on string similarities to align cross-lingual taxonomies in specific domains. Here, we focus on aligning more general cross-lingual and cross-domain taxonomies without domain specific information, such as product catalogues and Web site directories.

The Proposed Approach

In this section, we present our proposed approach in detail, which consists of two main steps: candidates identification and exact matching.

Candidates Identification

To avoid unnecessary comparisons of the categories between two given taxonomies, we aim to obtain all the possible matched categories in the target taxonomy for each category in the source taxonomy. The output of this step is taken as the input of exact matching.

The simplest way to represent a category is using its category label. However, this may not be desirable for category matching since synonymous categories may own totally different labels. For example, "*Sports Clothing*" and "*Athletic Apparel*" do not share any word, let alone two synonymous categories of different languages. Besides, directly comparing the translated category labels of the same language also

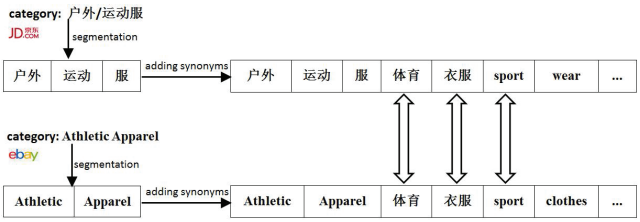


Figure 1: An Example of Candidates Identification

has limitations due to different language habits and improper translations.

To avoid the above problems, we capture cross-lingual string similarities between categories in word level by using BabelNet (Navigli and Ponzetto 2010), a Web-scale multilingual synonym thesaurus. The key idea of candidates identification is that "*two categories of different languages may be relevant if they share the same or synonymous words*". Given a category c^s in the source taxonomy of language s and a category c^t in the target taxonomy of language t, each of them is segmented into a set of words. After removing stop words, c^s and c^t contain a set of words $W_{c^s} = \{w_i^s\}_{i=1}^m$ and $W_{c^t} = \{w_j^t\}_{j=1}^n$, respectively. For each word w_i^s, we get its synonymous words of language s and t by BabelNet and add them to W_{c^s}. The same process is also applied for each word w_j^t and we get the new W_{c^t}. The cross-lingual string similarity between c^s and c^t is defined as:

$$CSS(c^s, c^t) = \begin{cases} 1, & \text{if } W_{c^s} \cap W_{c^t} \neq \emptyset \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

If $CSS(c^s, c^t)$ equals 1, c^t will be taken as one of the candidate matched category of c^s. Figure 1 illustrates why "Athletic Apparel" in eBay.com is taken as the candidate matched category of " 锟斤拷锟斤拷/锟剿讹拷锟斤拷" in JD.com.

Textual Context Extraction

Since categories do not have textual information to describe themselves, vector similarities cannot be applied to category matching until we find a way to enrich the representation of categories with textual information. There might exist some Web pages which contain the textual information associated with a given category, but manually finding appropriate Web pages is unrealistic. Therefore, we choose to acquire textual information by querying the Web with the search engine Google.

Categories of the same label in different structures may have different meanings. For example, category "*Sports*" occurs twice in Yahoo Directory. One is the child of category "*Shopping and Services*" which means sports goods, and the other is the child of category "*Recreation*" representing kinds of physical activities. However, the results (i.e. titles, snippets and URLs) returned only by submitting their own labels to Google are the same. To accurately get the relevant textual information returned by Google for each category, the labels of the given category c and its parent category pc are jointly submitted to Google. For example, in Yahoo Directory, the label of category "*Sports*" representing sports goods is submitted to Google jointly with its parent category label "*Shopping and Services*". In each returned snippet, we extract the words co-occurred with c in the same sentence except pc, because pc is part of the query, thus it occurs quite a lot of times. These extracted words are taken as the textual context to better reveal the meaning of the given category. Note that the root categories do not have parent categories, but they are usually unambiguous, otherwise users will be easily confused when exploring the taxonomy in a top-down manner. Thus, we simply submit the label of each root category to Google to get its textual context.

Exact Matching

After utilizing a linguistic feature (i.e. the cross-lingual similarity) for candidates identification, we aim to perform exact matching by determining the relevance score between each category in the source taxonomy and its candidate matched categories in the target taxonomy with one or more features based on vector similarities. The bag-of-words (BOW) model is the most common method to model text. However, the textual context of each category is extracted from the snippets that vary a lot in wording styles, because the snippet may be a tweet or a piece of news with more formal language expressions. For each category, it may be the case that the words extracted from different snippets are totally different, which means quite a lot of the words are of low frequency. Therefore, the BOW model may not work well in this scenario.

To address the above problem, we try to discover the topics of the extracted textual contexts with a bilingual topic model. The textual context of each category is actually a set of short text documents extracted from the snippets. Given a short text document d^s of language s, we first translate it into the document d^t of language t with Google Translate, and then construct **a pair of bilingual documents** (d^s, d^t). After applying the same process for all the documents of language s, **a paired bilingual document corpus** $\{(d_i^s, d_i^t)\}_{i=1}^{N_d}$ will be generated. Then we can directly apply a widely used bilingual topic model, i.e. Bilingual Latent Dirichlet Allocation (BiLDA) model (Vulić et al. 2015) to model the corpus as a generation process (see Figure 3 (a)). However, this model will suffer from the data sparsity problem in short text documents (Hong and Davison 2010). Hence, we propose a new bilingual topic model, called Bilingual Biterm Topic Model (BiBTM) to explicitly model the word co-occurrence in each pair of bilingual short text documents. BiBTM can not only avoid the problems caused by applying the BOW model or BiLDA, but also better uncover the topics of textual contexts for exact matching.

a) Bilingual Biterm Topic Model BiBTM is an extension of Biterm Topic Model (BTM) (Yan et al. 2013; Cheng et al. 2014) for modeling the generation of biterms. The key idea is that if two words co-occur more frequently, they are more likely to belong to a same topic. Different from BTM, a biterm used in BiBTM denotes an unordered word-pair co-occurring in a pair of bilingual documents. Any two distinct words in a pair of bilingual documents construct a biterm. For example, given a pair of bilingual documents (d^s, d^t),

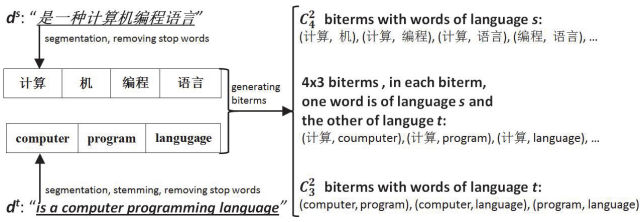


Figure 2: An Example of Biterm Generation

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. Figure 2 gives an example of biterm generation. BiBTM assumes that all biterms extracted from the whole corpus share the same topic distribution, and each topic consists of two discrete distributions over words of different languages.

Given a paired bilingual document corpus, suppose it contains $|\mathbf{B}|$ biterms $\mathbf{B} = \mathbf{B}^s \cup \mathbf{B}^{st} \cup \mathbf{B}^t = \{b_i^s\}_{i=1}^{|\mathbf{B}^s|} \cup \{b_i^{st}\}_{i=1}^{|\mathbf{B}^{st}|} \cup \{b_i^t\}_{i=1}^{|\mathbf{B}^t|}$ with $b_i^s = (w_{i,1}^s, w_{i,2}^s)$ where each word is in language s, $b_i^{st} = (w_{i,1}^s, w_{i,2}^t)$ where two words are in different languages and $b_i^t = (w_{i,1}^t, w_{i,2}^t)$ where each word is in language t, as well as K topics expressed over W^s and W^t distinct words of language s and language t respectively. The topic indicator variable $z \in [1, K]$ can be denoted as z^s, z^{st} and z^t for the three kinds of biterms. We represent the topics in the corpus by a K-dimensional multinomial distribution $\boldsymbol{\theta} = \{\theta_k\}_{k=1}^K$ with $\theta_k = P(z = k)$. The word distribution of language s and language t are respectively represented by a $K \times W^s$ matrix $\boldsymbol{\varphi}^s$ and a $K \times W^t$ matrix $\boldsymbol{\varphi}^t$, where the kth row φ_k^s and φ_k^t are respective a W^s-dimensional multinomial distribution with entry $\varphi_{k,w^s}^s = P(w^s|z = k)$ and a W^t-dimensional multinomial distribution with entry $\varphi_{k,w^t}^t = P(w^t|z = k)$.

Following the convention of BTM, the hyperparameters α and β are the symmetric Dirichlet priors. Figure 3 (b) shows the graphical representation of BiBTM and its generative process is described in Algorithm 1. Using BiBTM, the probability of generating the whole corpus given hyperparameters α and β can be expressed as:

$$P(\mathbf{B}|\alpha,\beta) = \prod_{i=1}^{|\mathbf{B}^s|} \int\int \sum_{k=1}^{K} \theta_k \varphi_{k,w_{i,1}^s}^s \varphi_{k,w_{i,2}^s}^s \mathrm{d}\boldsymbol{\theta}\mathrm{d}\boldsymbol{\varphi}^s \times \prod_{i=1}^{|\mathbf{B}^{st}|} \int\int\int \sum_{k=1}^{K} \theta_k \varphi_{k,w_{i,1}^s}^s \varphi_{k,w_{i,2}^t}^t \mathrm{d}\boldsymbol{\theta}\mathrm{d}\boldsymbol{\varphi}^s\mathrm{d}\boldsymbol{\varphi}^t \times \prod_{i=1}^{|\mathbf{B}^t|} \int\int \sum_{k=1}^{K} \theta_k \varphi_{k,w_{i,1}^t}^t \varphi_{k,w_{i,2}^t}^t \mathrm{d}\boldsymbol{\theta}\mathrm{d}\boldsymbol{\varphi}^t \quad (2)$$

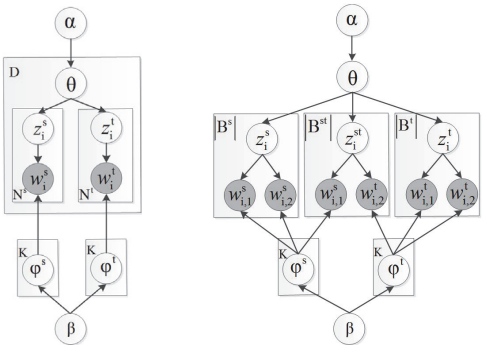
b) Parameters Estimation Since it is intractable to exactly solve the coupled parameters $\boldsymbol{\theta}$, $\boldsymbol{\varphi}^s$ and $\boldsymbol{\varphi}^t$ by maximizing the likelihood in Eq. (2), we adopt collapsed Gibbs Sampling (Liu 1994) to resolve this problem. $\boldsymbol{\theta}$, $\boldsymbol{\varphi}^s$ and $\boldsymbol{\varphi}^t$ can be integrated out due to the use of conjugate priors. Thus, we only need to sample the topic of each biterm. Due to space limit, we only show the derived Gibbs sampling formulas for $b_i^s \in \mathbf{B}^s$, $b_i^{st} \in \mathbf{B}^{st}$ and $b_i^t \in \mathbf{B}^t$ as follows,

$$P(z_i^s = k|z_{\neg b_i^s}, \mathbf{B}) \propto (n_{\neg b_i^s,k} + \alpha) \times \frac{(n_{\neg b_i^s,w_{i,1}^s|k} + \beta)(n_{\neg b_i^s,w_{i,2}^s|k} + \beta)}{(n_{\neg b_i^s,\cdot^s|k} + 1 + W^s\beta)(n_{\neg b_i^s,\cdot^s|k} + W^s\beta)} \quad (3)$$

$$P(z_i^{st} = k|z_{\neg b_i^{st}}, \mathbf{B}) \propto (n_{\neg b_i^{st},k} + \alpha) \times \frac{(n_{\neg b_i^{st},w_{i,1}^s|k} + \beta)(n_{\neg b_i^{st},w_{i,2}^t|k} + \beta)}{(n_{\neg b_i^{st},\cdot^s|k} + W^s\beta)(n_{\neg b_i^{st},\cdot^t|k} + W^t\beta)} \quad (4)$$

$$P(z_i^t = k|z_{\neg b_i^t}, \mathbf{B}) \propto (n_{\neg b_i^t,k} + \alpha) \times \frac{(n_{\neg b_i^t,w_{i,1}^t|k} + \beta)(n_{\neg b_i^t,w_{i,2}^t|k} + \beta)}{(n_{\neg b_i^t,\cdot^t|k} + 1 + W^t\beta)(n_{\neg b_i^t,\cdot^t|k} + W^t\beta)} \quad (5)$$

where $z_{\neg b}$ denotes the topic assignments for all biterms except the biterm b, $n_{\neg b,k}$ is the number of biterms assigned to topic k excluding b, $n_{\neg b,w^s|k}$ is the number of times word w of language s assigned to topic k excluding b, $n_{\neg b,w^t|k}$ is the number of times word w of language t assigned to topic k excluding b, and $n_{\neg b,\cdot^s|k} = \sum_{w^s} n_{\neg b,w^s|k}$ as well as $n_{\neg b,\cdot^t|k} = \sum_{w^t} n_{\neg b,w^t|k}$.



(a) BiLDA (b) BiBTM

Figure 3: Graphical representation of BiLDA and BiBTM

Algorithm 1: Generative Process of BiBTM

initialize: (1) set the number of topics K;
(2) set values for Dirichlet priors α and β;
sample: K times $\boldsymbol{\varphi}^s \sim Dir(\beta)$;
sample: K times $\boldsymbol{\varphi}^t \sim Dir(\beta)$;
sample: $\boldsymbol{\theta} \sim Dir(\alpha)$ for all biterms;
foreach *biterm* $b_i^s \in \mathbf{B}^s$ **do**
  sample: $z_i^s \sim Multi(\boldsymbol{\theta})$;
  sample: $w_{i,1}^s, w_{i,2}^s \sim Multi(\varphi_{z_i^s}^s)$
foreach *biterm* $b_i^{st} \in \mathbf{B}^{st}$ **do**
  sample: $z_i^{st} \sim Multi(\boldsymbol{\theta})$;
  sample: $w_{i,1}^s \sim Multi(\varphi_{z_i^{st}}^s)$, $w_{i,2}^t \sim Multi(\varphi_{z_i^{st}}^t)$
foreach *biterm* $b_i^t \in \mathbf{B}^t$ **do**
  sample: $z_i^t \sim Multi(\boldsymbol{\theta})$;
  sample: $w_{i,1}^t, w_{i,2}^t \sim Multi(\varphi_{z_i^t}^t)$

After a sufficient number of iterations, we can estimate the global topic distribution $\boldsymbol{\theta}$ and topic-word distributions $\boldsymbol{\varphi}^s$, $\boldsymbol{\varphi}^t$ by

$$\theta_k = \frac{\alpha + n_k}{K\alpha + |\mathbf{B}|} \tag{6}$$

$$\varphi^s_{k,w^s} = \frac{\beta + n_{w^s|k}}{W^s\beta + n_{\cdot^s|k}} \tag{7}$$

$$\varphi^t_{k,w^t} = \frac{\beta + n_{w^t|k}}{W^t\beta + n_{\cdot^t|k}} \tag{8}$$

where n_k is the number of biterms assigned to topic k, $n_{w^s|k}$ is the number of times word w of language s assigned to topic k, $n_{w^t|k}$ is the number of times word w of language t assigned to topic k, and and $n_{\cdot^s|k} = \sum_{w^s} n_{w^s|k}$ as well as $n_{\cdot^t|k} = \sum_{w^t} n_{w^t|k}$.

c) Context Topics Inference To perform exact matching, we need to know the topic distribution of the context for each category. Given a category c, suppose it contains N_c biterms $\{b_j\}_{j=1}^{N_c}$, which are extracted from all the pairs of bilingual documents of c. We utilize the following formula (Yan et al. 2013) to infer the topic distribution of the context for c.

$$P(z|c) = \sum_{j=1}^{N_c} P(z = k|b_j)P(b_j|c) \tag{9}$$

In Eq. (9), $P(b_j|c)$ is estimated by empirical distribution:

$$P(b_j|c) = \frac{n(b_j)}{\sum_{j=1}^{N_c} n(b_j)} \tag{10}$$

where $n(b_j)$ is the frequency of biterm b_j in all the pairs of bilingual documents of c. Meanwhile, $P(z = k|b_j)$ can be computed via Bayes' formula based on the parameters learned in BiBTM:

$$P(z = k|b_j) = \begin{cases} \frac{\theta_k \cdot \varphi^s_{k,w^s_{j,1}} \cdot \varphi^s_{k,w^s_{j,2}}}{\sum_{k'=1}^{K} \theta_{k'} \cdot \varphi^s_{k',w^s_{j,1}} \cdot \varphi^s_{k',w^s_{j,2}}}, & \text{if } b_j \in \mathbf{B}^s \\ \frac{\theta_k \cdot \varphi^s_{k,w^s_{j,1}} \cdot \varphi^t_{k,w^t_{j,2}}}{\sum_{k'=1}^{K} \theta_{k'} \cdot \varphi^s_{k',w^s_{j,1}} \cdot \varphi^t_{k',w^t_{j,2}}}, & \text{if } b_j \in \mathbf{B}^{st} \\ \frac{\theta_k \cdot \varphi^t_{k,w^t_{j,1}} \cdot \varphi^t_{k,w^t_{j,2}}}{\sum_{k'=1}^{K} \theta_{k'} \cdot \varphi^t_{k',w^t_{j,1}} \cdot \varphi^t_{k',w^t_{j,2}}}, & \text{if } b_j \in \mathbf{B}^t \end{cases} \tag{11}$$

After obtaining the topic distribution of the context for each category, we can represent categories of different languages in the same topic space. The final relevance score between each category in the source taxonomy and its candidate matched categories is computed as the cosine similarity between topic vectors.

Experiments

To facilitate knowledge sharing across languages on the Web, we evaluated our proposed approach on two different kinds of real world datasets, which are publicly available[2].

[2]https://github.com/jxls080511/080424

Experiment Settings

a) Tasks and Data Sets Two kinds of cross-lingual and cross-domain taxonomies on the Web (i.e. product catalogue and Web site directory) were used to validate the proposed approach. The details of the tasks and datasets are given as follows:

- **Cross-lingual Product Catalogue Alignment**. In this task, given a category in JD.com (i.e. one of the largest Chinese B2C online retailers), we aim to find the most relevant category in eBay.com. We collected 7,741 Chinese categories in JD.com and 7,782 English categories in eBay.com.
- **Cross-lingual Web Site Directory Alignment**. In this task, given a category in Chinese Dmoz.org (i.e. the largest Chinese Web site directory), we intend to find the most relevant category in Yahoo Directory. We collected 2,084 Chinese categories in Chinese Dmoz.org and 2,353 English categories in Yahoo Directory.

To generate the ground truth data, given a pair of taxonomies, five annotators labelled the most relevant category in the target taxonomy (eBay.com or Yahoo Directory) for each of the 100 randomly selected categories in the source taxonomy (JD.com or Chinese Dmoz.org). The labelled results are based on majority voting.

b) Evaluation Metrics Similar to (Prytkova, Weikum, and Spaniol 2015; Spohr, Hollink, and Cimiano 2011), we take cross-lingual taxonomy alignment as a ranking problem in the experiments. For each category in the source taxonomy, we ranked all categories in the target taxonomy according to the relevance score predicted by our approach and the designed comparison methods. Thus, we evaluated the ranking results in terms of MRR (Mean Reciprocal Rank) (Craswell 2009).

c) Comparison methods We compared our approach with the following methods.

- ***RSVM:*** The ranking SVM (RSVM) model is used in (Spohr, Hollink, and Cimiano 2011) for cross-lingual taxonomy alignment. After removing some domain-specific features, there still exist 20 linguistic features and 8 structural features for training the model. These features depends on string similarities after using machine translation.
- ***BiBTM:*** This is the proposed model for exact matching in our approach. Here, we only use BiBTM to align corsslingual taxonomies without the step of candidates identification. In BiBTM, we set $\alpha = 50/K$, $\beta = 0.1$ and $K = 120$ (the empirical tuning results will be presented in Section 4.2.2).
- ***CSS+RSVM:*** This approach first uses our proposed cross-lingual string similarity (CSS) for candidates identification, and then utilizes the ranking SVM model trained in RSVM for exact matching.
- ***CSS+BOW:*** This approach also uses CSS for candidates identification at first, and then applies the traditional bag-of-words (BOW) model to exact matching. Given a category c, after merging all pairs of bilingual short text documents of c into one document d, each word in d is weighed with TF-IDF (Baeza-Yates and Ribeiro-Neto 1999).

Table 1: Overall results: MRR values

Approach	JD.com→ eBay.com	Chinese Dmoz.org→ Yahoo Directory
RSVM	0.195	0.261
BiBTM	0.199	0.246
CSS+RSVM	0.210	0.301
CSS+BOW	0.423	0.489
CSS+BiLDA	0.553	0.679
CSS+BiBTM	**0.597**	**0.719**

- ***CSS+BiLDA:*** After utilizing CSS for candidates identification, this approach leverages BiLDA (Vulić et al. 2015) to exact matching. In BiLDA, we set $\alpha = 50/K$, $\beta = 0.1$ and $K = 80$ (the empirical tuning results will be presented in Section 4.2.2).

Result Analysis

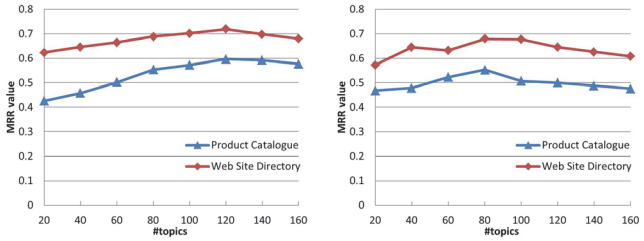
a) Overall Performance In our proposed approach (i.e. CSS+BiBTM) and all designed comparison methods except RSVM and CSS+RSVM, we need to extract the textual context of each category from the web. For each category, we extracted the snippets of top 20 results returned by Google. In each snippet, we only kept the words that co-occur with the given category in the same sentence except its parent category (mentioned in Section 3.2). Each processed Chinese (English) snippet was translated into English (Chinese) by Google Translate to construct a pair of bilingual short text documents. We further processed the documents via the following normalization steps:

- **Processing Chinese Documents**. 1) Segmenting words with FudanNLP (Qiu, Zhang, and Huang 2013) and removing stop words; 2) removing words with document frequency less than 10; 3) filtering out documents with length less than 2.
- **Processing English Documents**. 1) Removing non-Latin characters and stop words; 2) converting letters into lower case and stemming each word; 3) removing words with document frequency less than 10; 4) filtering out documents with length less than 2.

At last, for the textual contexts of categories in product catalogues (i.e. JD.com and eBay.com), we got 11,473 distinct Chinese words and 9,093 distinct English words. For the textual contexts of categories in Web site directories (i.e. Chinese Dmoz.org and Yahoo Directory), we got 26,473 distinct Chinese words and 16,852 distinct English words.

For each task in our experiments, we trained a BiBTM and a BiLDA model. For each model, we ran 500 iterations of Gibbs sampling to converge. Table 1 gives the overall results of our approach and the designed comparison methods, and we can see that:

- RSVM, BiBTM and CSS+RSVM are of the worst performance. It means that the approaches only using the features that rely on string similarities or the ones that depend on vector similarities can not work well in aligning the real-world cross-lingual taxonomies.
- CSS+BOW and CSS+BiLDA are of the same framework (i.e. candidate identification with string similarities and exact matching with vector similarities) as our approach (i.e. CSS+BiBTM) has. These three approaches significantly outperform others, which shows the superiority of the framework of our proposed approach.
- After performing candidates identification with CSS, bilingual topic models do much better than the BOW model in exact matching. It demonstrates that compared with the word vector generated by the BOW model, representing the textual context of each category as a topic vector is more suitable for exact matching.
- Compared with CSS+BiLDA, CSS+BiBTM (i.e. our approach) achieves about a 4% MRR improvement on both of the tasks, which shows that BiBTM can better discover the topics of textual contexts for exact matching.



(a) BiBTM (b) BiLDA

Figure 4: MRR value vs. number of topics K

b) Parameter Tuning One important parameter in BiBTM and BiLDA is the number of topics K. Different number of topics may lead to different performance in cross-lingual taxonomy alignment. Thus, we performed an analysis by varying the number of topics in the BiBTM and BiLDA model. Figure 4 (a) shows the performance of BiBTM with different number of topics K on two given datasets. The performance improves by increasing K when $K < 120$. Figure 4 (b) shows the performance of BiLDA with different number of topics K. The MRR value is the highest when $K = 80$ for both of the datasets. It shows that when aligning cross-lingual and cross-domain taxonomies, as well as the size of categories is large enough, K can be empirically set to 120 and 80 in BiBTM and BiLDA, respectively.

Conclusions and Future Work

In this paper, we present a new approach to address the problem of cross-lingual taxonomy alignment. We first proposed the cross-lingual string similarity for candidates identification. We then proposed a novel bilingual topic model to obtain the topic vector of the extracted textual context for each category. Finally, we obtained the alignment result by using the cosine similarity between topic vectors. We evaluated our approach on two kinds of real world taxonomies. The experimental results showed that our approach significantly outperforms the designed state-of-the-art comparison methods. Specifically, compared with the methods that combine CSS and other models, our approach got the best performance, which validates the advantage of our new bilingual topic model.

As for the future work, we will validate our approach

on some domain-specific taxonomies, such as the datasets in OAEI [3] Multifarm track. We also plan to utilize the structured information in knowledge bases to enhance our approach for cross-lingual taxonomy alignment.

Acknowledgements

This work is supported in part by the National Natural Science Foundation of China (NSFC) under Grant No. 61272378, No. 61402173, the 863 Program under Grant No. 2015AA015406 and the Fundamental Research Funds for the Central Universities under Grant No. 22A201514045.

References

Agrawal, R., and Srikant, R. 2001. On integrating catalogs. In *Proc. of WWW*, 603–612.

Baeza-Yates, R., and Ribeiro-Neto, B. 1999. *Modern information retrieval*, volume 463. ACM press New York.

Berlin, J., and Motro, A. 2002. Database schema matching using machine learning with feature selection. In *Proc. of CAiSE*, 452–466.

Cheng, X.; Yan, X.; Lan, Y.; and Guo, J. 2014. Btm: Topic modeling over short texts. *IEEE Transactions on Knowledge and Data Engineering* 26(12):2928–2941.

Craswell, N. 2009. Mean reciprocal rank. In *Encyclopedia of Database Systems*. Springer. 1703–1703.

Denhière, G., and Lemaire, B. 2004. A computational model of children's semantic memory. In *Proc. of CogSci*, 297–302.

Do, H.-H., and Rahm, E. 2007. Matching large schemas: Approaches and evaluation. *Information Systems* 32(6):857–885.

Do, H.-H.; Melnik, S.; and Rahm, E. 2003. Comparison of schema matching evaluations. In *Web, Web-Services, and Database Systems*. Springer. 221–237.

Euzenat, J., and Shvaiko, P. 2007. *Ontology matching*, volume 333. Springer.

Faria, D.; Martins, C.; Nanavaty, A.; Taheri, A.; Pesquita, C.; Santos, E.; F. Cruz, I.; and M. Couto, F. 2014. Agreementmakerlight results for oaei 2014. In *Proc. of OM*, 105–112.

Gabrilovich, E., and Markovitch, S. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proc. of IJCAI*, volume 7, 1606–1611.

Hong, L., and Davison, B. D. 2010. Empirical study of topic modeling in twitter. In *Proc. of SOMA*, 80–88.

Ichise, R.; Takeda, H.; and Honiden, S. 2003. Integrating multiple internet directories by instance-based learning. In *Proc. of IJCAI*, volume 3, 22–28.

Jiménez-Ruiz, E.; Grau, B. C.; Xia, W.; Solimando, A.; Chen, X.; Cross, V.; Gong, Y.; Zhang, S.; and Chennai-Thiagarajan, A. 2014. Logmap family results for oaei 2014. In *Proc. of OM*, volume 20, 126–134.

Lehmann, J.; Isele, R.; Jakob, M.; Jentzsch, A.; Kontokostas, D.; Mendes, P. N.; Hellmann, S.; Morsey, M.; van Kleef, P.; Auer, S.; et al. 2014. Dbpedia-a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal* 5:1–29.

Li, C.; Ji, L.; and Yan, J. 2015. Acronym disambiguation using word embedding. In *Proc. of AAAI*, 4178–4179.

Liu, J. S. 1994. The collapsed gibbs sampler in bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association* 89(427):958–966.

Mahdisoltani, F.; Biega, J.; and Suchanek, F. 2014. Yago3: A knowledge base from multilingual wikipedias. In *Proc. of CIDR*.

Navigli, R., and Ponzetto, S. P. 2010. Babelnet: Building a very large multilingual semantic network. In *Proc. of ACL*, 216–225.

Nguyen, D.; Overwijk, A.; Hauff, C.; Trieschnigg, D. R.; Hiemstra, D.; and De Jong, F. 2009. Wikitranslate: query translation for cross-lingual information retrieval using only wikipedia. In *Evaluating Systems for Multilingual and Multimodal Information Access*. Springer. 58–65.

Potthast, M.; Stein, B.; and Anderka, M. 2008. A wikipedia-based multilingual retrieval model. In *Proc. of ECIR*, 522–530.

Prytkova, N.; Weikum, G.; and Spaniol, M. 2015. Aligning multi-cultural knowledge taxonomies by combinatorial optimization. In *Proc. of WWW*, 93–94.

Qiu, X.; Zhang, Q.; and Huang, X. 2013. Fudannlp: A toolkit for chinese natural language processing. In *Proc. of ACL*, 49–54.

Rahm, E., and Bernstein, P. A. 2001. A survey of approaches to automatic schema matching. *the VLDB Journal* 10(4):334–350.

Shvaiko, P., and Euzenat, J. 2006. Tutorial on ontology matching. In *Proc. of SWAP*, 26–228.

Spohr, D.; Hollink, L.; and Cimiano, P. 2011. A machine learning approach to multilingual and cross-lingual ontology matching. In *Proc. of ISWC*, 665–680.

Vulić, I.; De Smet, W.; Tang, J.; and Moens, M.-F. 2015. Probabilistic topic modeling in multilingual settings: An overview of its methodology and applications. *Information Processing & Management* 51(1):111–147.

Wang, Z.; Li, J.; Wang, Z.; and Tang, J. 2012. Cross-lingual knowledge linking across wiki knowledge bases. In *Proc. of WWW*, 459–468.

Wang, H.; Wu, T.; Qi, G.; and Ruan, T. 2014. On publishing chinese linked open schema. In *Proc. of ISWC*, 293–308.

Wang, Z.; Li, J.; and Tang, J. 2013. Boosting cross-lingual knowledge linking via concept annotation. In *Proc. of IJCAI*, 2733–2739.

Yan, X.; Guo, J.; Lan, Y.; and Cheng, X. 2013. A biterm topic model for short texts. In *Proc. of WWW*, 1445–1456.

Zhang, L., and Rettinger, A. 2014. X-lisa: cross-lingual semantic annotation. In *Proc. of VLDB*, 1693–1696.

[3] http://oaei.ontologymatching.org/