Knowledge Transfer across Multilingual Corpora via Latent Topics

Wim De Smet[1], Jie Tang[2], and Marie-Francine Moens[1]

[1] K.U.Leuven, Leuven, Belgium
[2] Tsinghua University, Beijing, China
{wim.desmet,marie-francine.moens}@cs.kuleuven.be,
jie.tang@tsinghua.cn.edu

Abstract. This paper explores bridging the content of two different languages via latent topics. Specifically, we propose a unified probabilistic model to simultaneously model latent topics from bilingual corpora that discuss comparable content and use the topics as features in a cross-lingual, dictionary-less text categorization ask. Experimental results on multilingual Wikipedia data show that the proposed topic model effectively discover the topic information from the bilingual corpora, and the learned topics successfully transfer classification knowledge to other languages, for which no labeled training data are available.

1 Introduction

Cross-lingual text mining aims to transfer knowledge across different languages to help applications such as cross-lingual information retrieval, summarization and categorization in cases where translation and class-labeled resources are scarce. A specific challenge is to build models that capture content of comparable corpora, i.e. texts that to a varying degree contain shared and non-shared content. Recently, with the rapid development of online social networks, such as Facebook, MySpace, Ning, and Twitter, users have generated a huge volume of valuable multilingual language resources. However, content is seldom well-paired across the languages, forming cross-lingual parallel corpora. For example, a user may write a lot in English about her recent travel in a blog, but write only a few sentences in French. Indeed, even in well-organized Wikipedia documents, one cannot find an exact translation of a page across the different languages. However, there are many so-called "comparable" corpora available, where the documents discuss similar content in different language, but the content in one language is not an exact translation of the content in the other language. Extracting knowledge from these comparable corpora is very valuable for cross-lingual tasks.

Current research for text categorization on bilingual corpora often focuses on learning the classification model from monolingual labeled documents and their cross-lingual pairing. Few efforts are made to directly build cross-lingual models, with which the classification knowledge can be transferred across different languages.

In this paper, we address the problem of knowledge transfer across multilingual and comparable corpora. We propose a unified probabilistic model to simultaneously extract latent topics from the multilingual corpora. The learned topic

models capture the common knowledge across the different languages which can be used in many applications. We apply the topic models to bilingual document classification. Our experimental results on multilingual Wikipedia data (written in English, Spanish, French and Italian) show that the proposed Bilingual Latent Dirichlet Allocation (Bi-LDA) model can effectively represent a target language without the use of any information from translation dictionaries. Additionally, the learned bilingual topic models yield a very strong categorization performance in target languages, for which no labeled training examples are available.

The rest of the paper is organized as follows: Section 2 formally formulates the problem. Section 3 explains the proposed model. Section **??** gives experimental results that validate the effectiveness of our methodology. Finally, Section **??** discusses related work and Section **??** concludes.

2   Problem Definition

In this section, we present several necessary definitions and then define the sub-problems of knowledge transfer across multilingual documents addressed in this paper.

Symbol	description
K	number of topics
$\mathcal{C}$	bilingual corpus
$\mathbf{M}$	number of paired documents in $\mathcal{C}$
N_d	number of words in document d
$\mathbf{w}_d$	vector form of words in document d
θ_d	multinomial distribution over topics specific to document d
ϕ_i	multinomial distribution over words for topic z_i for the source language
ψ_i	multinomial distribution over words for topic z_i for the target language
$\mathcal{L}^{\mathcal{S}}$	labeled data in the source language S
$\mathcal{U}^{\mathcal{T}}$	unlabeled data in the target language T

Table 1. Definition of symbols

A paired bilingual document corpus can be defined as $\mathcal{C} = \{(d_1^S, d_1^T), \cdots, (d_M^S, d_M^T)\}$, where (d_j^S, d_j^T) (briefly d_j for simplicity) is a pair of documents in the source language S and the target language T, respectively. Each document may describe a number of topics. For example, the Wikipedia page of a city may describe topics related to history, culture, and tourism. Before formulating the problem, we first give the definition of a probabilistic topic model.

Definition 1. *Topic model of documents: A topic model ϕ of a document collection D is a multinomial distribution of words $p(w|\phi_i)$ over a vocabulary V, for each ϕ_i represented in θ_d where $d \in D$. The document collection is considered a mixture of multiple topics θ.*

The underlying assumption of a topic model is that words in the document are sampled following word distributions corresponding to each topic, i.e. $p(w|\phi_i)$ and $p(\phi_i|\theta_d)$. Therefore, words with the highest probability in the distribution would the semantics represented by the topic. For example, the words *travel*, *tourist*, and *hotel* would represent the topic "Travel". Given this, we can define the problem of bilingual topic modeling.

Problem 1. **Bilingual topic modeling.** Given a collection of paired documents in language S and T, i.e. $\mathcal{C} = \{(d_1^S, d_1^T), \cdots, (d_M^S, d_M^T)\}$, the goal of bilingual topic modeling is to learn for every document pair a set of K topics θ, each of which defines an associated set of words in S and in T.

The topic model θ bridges knowledge across documents in two languages, which creates many potential applications. In this work, we consider a general application, i.e. bilingual document classification. In particular, we consider how to take advantage of the topic models to transfer knowledge from one language to help document classification in another language. More specifically, let $\mathcal{L}^S$ be a labeled document collection in the source language, in which each document d_j^S is annotated with a class label $c \in \mathcal{Y}^S$, where $\mathcal{Y}^S = \{c_1, \cdots, c_{S_p}\}$ denotes the label space and S_p is the number of class labels in the source language. Let $\mathcal{U}^T$ be a unlabeled document collection in the target language. Formally, we can define the problem of bilingual document classification as follows.

Problem 2. **Bilingual document classification.** Given a labeled document collection $\mathcal{L}^S$ in the source language S, an unlabeled document collection $\mathcal{U}^T$ in the target language T, and the learned bilingual topic models, the goal is to transfer the labeled supervision information from the source language to predict the label of the documents in the target language.

Please note that although we only give the definition of bilingual document modeling and classification, the problem can easily be extended to multilingual corpora.

3 Our Approach

For bilingual topic modeling, we can simply consider a general topic model as a baseline method, i.e. Latent Dirichlet Allocation (LDA) [?], to model the topic information of all bilingual documents. We propose a probabilistic topic model, referred to as Bi-LDA, to simultaneously model bilingual documents within a unified model. The model describes each pair of bilingual documents (d_j^S, d_j^T) using a common mixture θ_j, thus knowledge can be transferred across different languages via the common mixture model.
In the remainder of this section, we will first briefly review Latent Dirichlet Allocation, and then describe our proposed approach in detail.

3.1 Latent Dirichlet Allocation

Recently, probabilistic topic models attracted considerable interest and have been successfully applied to text mining tasks such as information retrieval, recommendation, and text analysis [?,?]. Latent Dirichlet Allocation (LDA) [?] is a

three-level Bayesian network, which models documents using a latent topic layer. In LDA, for each document d_j in the corpus, a multinomial distribution θ_j over topics is first sampled from a Dirichlet distribution with parameter α. Second, for each word w_{ji}, a topic z_{ji} is chosen from θ_j and the word w_{ji} is generated from a topic-specific multinomial $\phi_{z_{ji}}$. Thus, the generating probability of word w from document d_j is:

$$P(w|d_j, \theta, \phi) = \sum_{z=1}^{K} P(w|z, \phi_z) \cdot P(z|\theta_j) \qquad (1)$$

In other words, it uses the topic distribution θ_j of the document and the probability ϕ_z^w of the topic generating word w to calculate the word's probability. Directly applying LDA to our bilingual problem means that the contents of both documents of a pair have to be combined together as one big document, learning a topic model were the two languages are mixed. This approach will serve as our baseline.

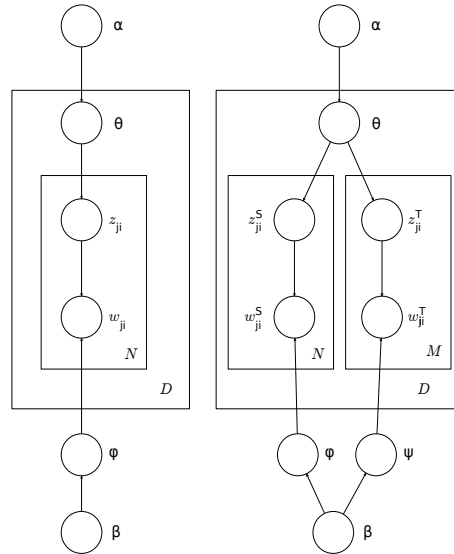### 3.2   Bilingual Latent Dirichlet Allocation

Fig. 1. Graphical representation of LDA and Bi-LDA.

Algorithm 3.1: Bilingual LDA()

for each document pair d_j

do $\begin{cases} \text{sample } \theta \sim Dir(\alpha) \\ \textbf{for each} \text{ word position } i \in d_j^S \\ \quad \textbf{do} \begin{cases} \text{sample } z_{ji}^S \sim Mult(\theta_j) \\ \text{sample } w_{ji}^S \sim Mult(\phi, z_{ji}^S) \end{cases} \\ \textbf{for each} \text{ word position } i \in d_j^T \\ \quad \textbf{do} \begin{cases} \text{sample } z_{ji}^T \sim Mult(\theta_j) \\ \text{sample } w_{ji}^T \sim Mult(\psi, z_{ji}^T) \end{cases} \end{cases}$

Figure **??** shows the graphical representation of the proposed model, Bilingual Latent Dirichlet Allocation (Bi-LDA). In Algorithm **??** we present its generative story. For each document pair d_j, a topic distribution θ is sampled from a K-dimensional Dirichlet distribution with parameter α. θ defines a distribution common to both languages (S and T). Then, each word $w_{ji}^\S$ in the source language is generated from a multinomial distribution $\phi_{z_{ji}^S}$, specific to a chosen topic z_{ji}^S. Similarly, each word w_{ji}^T of the target language is also sampled with a same procedure. We see that there is a one common θ for both languages, which implies that all topics in θ are shared across the bilingual documents.

To train this model, we used the Gibbs sampling approach. This requires two sets of formulas to converge to correct distributions: one for each topic z_{ji}^S and one for each topic z_{ji}^T. For the first, the updating formula becomes:

$$p(z_{ji}^S = k | \mathbf{w}_{d_j}^S, \mathbf{w}_{d_j}^T, \mathbf{z}_{\neg ji}^S, \mathbf{z}^T) = \frac{n_{j,k,\neg i}^S + n_{j,k}^T + \alpha}{n_{j,\cdot,\neg i}^S + n_{j,\cdot}^T + K \cdot \alpha} \cdot \frac{v_{k,w_{ji}^S,\neg}^S + \beta}{v_{k,\cdot,\neg}^S + W^S \cdot \beta} \tag{2}$$

where $n_{j,k}^S$ is the number of times that topic k has been sampled from the multinomial distribution specific to document d_j^S. $n_{j,k,\neg i}^S$ is the same except the current z_{ji}^S is not counted. $v_{k,\cdot,\neg}^S$ is the number of times that word w_{ji}^S in language S has been generated by topic k minus one (not including the currently associated word w_{ji}^S). In these counters a dot ($\cdot$) means summation over all values of this variable, i.e. all topics in case of $n_{j,\cdot}^S$ and all words in $v_{k,\cdot,\neg}^S$.

For z_{ji}^T, the change in formula **??** is trivial.

3.3 Cross-lingual Document Classification

For cross-lingual document classification, we are given a set of labeled documents in the source language and no labeled documents in the target language. The objective is to learn a classification model from the labeled documents of the source language and apply to the classification of documents in the target language. The task obviously cannot be achieved by the traditional method that only uses words as features, as there is minimal word overlap between the two languages.

Our idea for cross-lingual document classification is to take advantage of the common topics learned by the proposed topic models, to transfer the knowledge from the source language to the target language. Basically we first learn the topic models (either LDA or Bi-LDA) on a general bilingual corpus (e.g., Wikipedia). Then given a bilingual document classification task, i.e., a $\mathcal{L}^S$ in language S and an unlabeled document collections $\mathcal{U}^T$ in language T, we use the learned topic model to infer the topic distribution of each document in $\mathcal{L}^S$ and $\mathcal{U}^T$.

Each document is then taken as a data instance in the classification model and the features are defined as the inferred topic distribution. The value of each feature of an instance (e.g., document d_j^S) is the probability of the corresponding topic k in the document, i.e. $p(z = k|d_j^S)$. For the classification model, one can use any classifier such as Naive Bayes, Perceptron, Maximum Entropy, and Support Vector Machine (SVM). In this paper, we use SVM.

4 Experimental Results

4.1 Experimental Setup

Datasets: We conduct the experiments on two datasets, one (called Topic) for training and evaluating the topic models, and one (called Classification) for evaluating the transfer classification. Both datasets are downloaded from the online encyclopedia Wikipedia, from the English, Spanish, French and Italian language sections. All documents were acquired from Wikipedia *dumps*, which mirror the entire online encyclopedia, are regularly updated and can be downloaded freely. For the Topic dataset we collect three bilingual corpora with paired documents by following "language links" between articles. Table **??** shows statistics of the Topic dataset. Note that not every article appears in each language, resulting in a different content for each bilingual dataset.

Table 2. Statistics of the Topic dataset with S = English and different values for T

	W^S	W^T	#Pair-docs
T= Spanish	$29,201$	$27,745$	$18,672$
T= French	$27,033$	$20,860$	$18,911$
T= Italian	$23,346$	$31,388$	$18,898$

The Classification dataset, which is different from the Topic dataset, is collected by exploiting the category-labels of Wikipedia. Specifically, we first select 5 high level categories: *books ("book")*, *films ("film")*, *programming languages ("prog")*, *sports* ("sport") and *videogames ("video")*. Then for each category, we extract up to 1,000 articles which are annotated with the category label. To acquire a workable set of documents, the categories were chosen to have examples of very broad and more specific classes. A Wikipedia article have more than one label, and these labels can be very specific. Using the hierarchy that Wikipedia provides, we extract all subcategories for the above base classes up to three sublevels. Articles are then crawled that belonged to any one of these subcategories. Since not all Wikipedia's have as large a collection of articles, we sometimes collected fewer than thousand articles for Spanish, French and Italian. Table **??** shows the size of the Classification dataset.

Comparison Methods: We compare the following methods for bilingual classification:

- *SVM+LDA*. This will be our baseline method. It combines each bilingual pair of documents into a single document, mixing the words from both languages, and employs LDA to learn the topic model from the Topic data set

Table 3. # documents of the Classification dataset.

	books	films	program	sport	video
English	1,000	1,000	1,000	1,000	1,000
Spanish	1,000	1,000	263	1,000	1,000
French	1,000	1,000	592	1,000	1,000
Italian	1,000	1,000	290	1,000	764

and then to infer the the topic distribution of the Classification data set. The two languages thus share only one common topic space. The learned topic distribution is used as the feature for SVM to learn the classification model. For SVM, we employ SVM-light.[3]

– *SVM+Bi-LDA*. This method uses the proposed Bi-LDA to learn the topic distribution from the bilingual corpora and then uses the inferred topic distribution as the features for SVM.

The code for this process is implemented in C++ and will be publicly available along with the data sets.

4.2 Perplexity

Perplexity measures a topic model's capability of predicting previously unseen documents. For a collection of these new documents (in our case the articles from the Classification dataset) C_u, it is calculated as:

$$Perp(C_u) = \exp\left(-\frac{\sum_{d \in C_u} \log\left(\prod_{w \in d} p(w)\right)}{\sum_{d \in C_u} N^d}\right) \qquad (3)$$

A lower perplexity score means that the model has assigned a higher probability to the documents. Theoretically, if a model is good, i.e. has a low perplexity, it will be well adapted to the new documents and therefore yield a good representation. Although trained on paired bilingual documents, inference of each of these models has to happen on one language at a time (as we do not have any *a priori* information of the test document's content). Therefore we present the perplexity for both of the languages separately.

Table **??** lists the perplexity of each model for all three language pairs, averaged over a number of different parameter settings for the model. These settings are: (named settings will be used in further experiments)

– for LDA and Bi-LDA: ranging the number of topics from 10 to 200 in steps of 10.

Table 4. Average perplexity of the different models, for each language pair.

LDA	English - French		English - Italian		English - Spanish	
Bi-LDA	1040.3	1082.2	1139.3	1523.6	1111.9	1277.3

The difference between the perplexity of LDA and Bi-LDA can be explained easily: since the vocabulary sizes doubles by merging the two languages in the

---

[3] http://svmlight.joachims.org/

LDA model, it follows that the probability of each word halves, which then again results in a doubling of the perplexity.

4.3 Classification Accuracy

Table 5. Average F1-score for the SVM classification, for each model and language pair.

		book	film	prog	sport	video
English -	LDA	5.7	12.0	48.5	27.7	42.6
Spanish	Bi-LDA	**80.5**	**60.0**	**74.3**	**59.4**	**59.6**
English -	LDA	3.9	1.4	67.9	42.6	**64.5**
French	Bi-LDA	**53.2**	**64.0**	**86.5**	**85.4**	34.8
English -	LDA	1.8	3.4	64.7	62.0	17.6
Italian	Bi-LDA	**52.1**	**46.7**	**84.1**	**79.3**	**76.9**

The use of knowledge transfer in a cross-lingual text categorization task is to our knowledge not studied in the literature. As a baseline we use LDA performed on concatenated texts (SVM + LDA) where the two vocabularies are mixed. Table **??** summarizes the performance of the models for each of our chosen classes, in each language pair. The F1 score is again averaged, over the same ranges of number of topics used for Table **??**. It can be seen that the Bi-LDA model realizes an acceptable transfer of categorization knowledge. Figure
In order to better assess the capabilities of the proposed topic models for cross-lingually transferring knowledge, we perform a number of additional tests.

4.4 Topic Smoothing

In this subsection, we first analyze how the bilingual document classification is influenced by different parameters (i.e., number of topics and hyperparameters), and then present a new method, called topic smoothing, to further improve the performance of bilingual document classification.

Effect of #topics on the categorization. We train the Bi-LDA model with the number of topics varied, and each time we apply the learned model to the bilingual document classification. Figure **??** plots the F1-score of the classification results using Bi-LDA, when choosing a different number of topics *a priori*. We see on some categories the classification results are not very sensitive to the number of topics, except when the number is very small. But on several tasks (e.g., the classification on Italian), the results vary largely with the different number of topics.

Topic smoothing To deal with the above sensitivity problem, we present a strategy of topic smoothing. Basically, for Bi-LDA the smoothing method is to combine topic models that are learned with a different number of topics for bilingual document classification. We train ten topic models, with the number of topics ranging from 10 to 200 with a 20 topic step. Then we apply the topic models to inference the topic distribution of the documents in the test data set, and concatenate all the different topic distributions as features of the SVM to train the
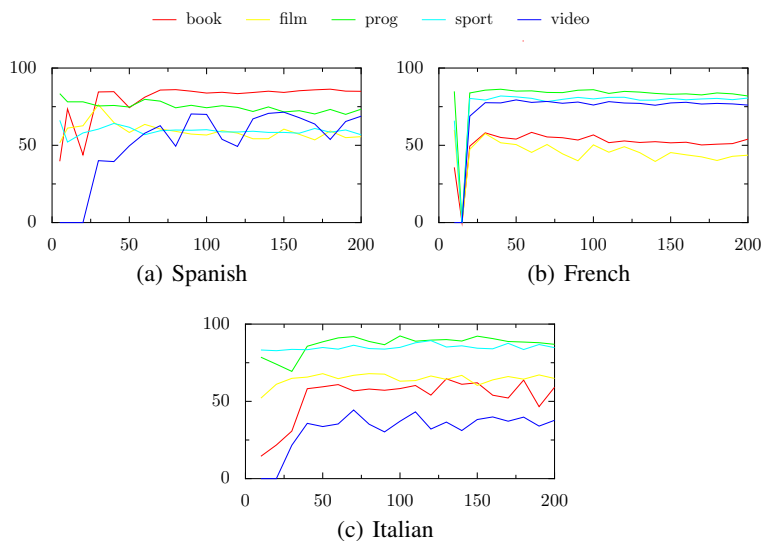
Fig. 2. Bi-LDA's performance for several values of K, language combinations and categories.

classification model. Specifically, different topic models are trained using different hyperparameters, and then topic distributions from the different topic models are concatenated as features to train the classification model. Table **??** shows the comparison of Bi-LDA and sBi-LDA. It can be seen that the smoothing method can efficiently improve (averagely +8.5%) the classification performance. As a trade-off of course, it requires substantially more computation.

Table 6. Average F1-score for the smoothed LDA compared to the average over the unsmoothed models (SBi-LDA − smoothed Bi-LDA).

		book	film	prog	sport	video
English - Spanish	Bi-LDA	80.5	60.0	74.3	59.4	59.6
	SBi-LDA	**83.6**	**60.3**	**77.7**	**61.3**	**72.5**
English - French	Bi-LDA	53.2	64.0	86.5	85.4	34.8
	SBi-LDA	**65.5**	**72.1**	**92.3**	**87.0**	**49.9**
English - Italian	Bi-LDA	52.1	46.7	84.1	79.3	76.9
	SBi-LDA	**58.9**	**53.9**	**88.1**	**83.3**	**79.1**

5 Related Work

Cross-lingual text mining is a popular research topic and quite a few research works have been conducted for cross-lingual information retrieval (e.g.,

[**?,?,?,?**]); however, cross-lingual classification is up till now rarely studied. Existing methods often rely on a translation tool to bridge documents of different languages and thus transform the problem into a monolingual classification [**?,?**]. Cross-lingual sentiment classification in text has recently drawn the attention [**?**] relying on translation dictionaries. Some efforts have also been made to reduce the number of labeled examples in both languages using techniques such as co-training [**?,?**]. However, these methods still rely on a well-organized parallel multilingual corpus [**?**] or on a translation tool. In the work presented here we train and test on comparable corpora and do not make use of any external translation resources.

Recently, how to model the multilingual corpora so as to discover the common and differential topics among the different languages becomes an interesting research topic and many methods have been proposed. A basic idea in these methods is to use the Latent Semantic Analysis (LSA) [**?,?,?**], document clustering [**?**], word alignment and machine translation [**?**], and parallel LDA model [**?,?,?**] to find the correlation between different languages. While much progress has been made, two fundamental problems have been largely ignored. First, the bilingual corpora may be quite unbalanced, i.e., documents of the source language may be not comparable with the target language. The cross-lingual documents on the Web, in particular the Web 2.0, are freely authored by the users, thus the contents would be very different. Second, it is unclear how to model the bilingual documents simultaneously. Directly employing LSA, clustering, or LDA can only model the bilingual in a common space, but cannot differentiate the topics specific to each language.

Another related work is transfer learning, which aims to transfer knowledge from a source domain to a related target domain. Two fundamental issues in transfer learning are "what to transfer" and "when to transfer". Many approaches have been proposed by reweighting instances in source domain for the use in the target domain [**?**]. [**?**] propose a locally weighted ensemble framework which can utilize different models for transferring labeled information from multiple training domains. Also many works rely on new feature representations [**?,?**]. For example, [**?**] propose a method to learn a shared low-dimensional representation for multiple related tasks and the task functions simultaneously. [**?**] propose to use a large amount of unlabeled data in the source domain to improve the performance on the target domain in which there are only few labeled data. They don't assume that the two domains share the class labels or topic distributions. [**?**] proposed a structural correspondence learning approach to induce correspondences among features from source and target domains. There are also other approaches which transfer information by shared parameters [**?**] or relational knowledge. Transfer learning techniques are widely used in classification, regression, clustering and dimensionality reduction problems.

The use of bilingual topic models for transferring the category knowledge across languages is completely new. Moreover, our topic models are trained on comparable corpora which are abundantly available.

6   Conclusion

In this paper we investigate knowledge transfer across multilingual corpora via latent topics. We formalize the major tasks and propose a probabilistic approach to

solve the tasks. We study and compare several strategies for simultaneously modeling the content of bilingual documents. One is the bilingual topic model based on LDA. We present a sampling algorithm to learn the model. Experimental results for categorizing Wikipedia documents according to their labels demonstrate the effectiveness of the proposed transfer learning approach.

There are several directions for future work. It would be interesting to develop new algorithms to automatically find the number of topics, and detect topics that are not shared between comparable documents. As the information in different languages might be very unbalanced, the numbers of topics in the documents of different languages could also be different. Another potential issue is to apply the proposed approach to other applications (e.g., recommendation, and link prediction across multilingual documents, cross-lingual information retrieval or cross-lingual summarization) to further validate its effectiveness.

Acknowledgments

References

1. Amini, M.R., Goutte, C.: A co-classification approach to learning from multilingual corpora. Mach. Learn. 79(1-2), 105–121 (2010)
2. Argyriou, A., Evgeniou, T., Pontil, M.: Multi-task feature learning. In: NIPS'06. pp. 41–48 (2006)
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. JMLR 3, 993–1022 (2003)
4. Blitzer, J., McDonald, R., Pereira, F.: Domain adaptation with structural correspondence learning. In: EMNLP'06. pp. 120–128 (2006)
5. Bonilla, E., Chai, K.M., ChrisWilliams: Multi-task gaussian process prediction. In: NIPS'08. pp. 153–160 (2008)
6. Chew, P.A., Bader, B.W., Kolda, T.G., Abdelali, A.: Cross-language information retrieval using parafac2. In: KDD'07. pp. 143–152 (2007)
7. Dai, W., Yang, Q., Xue, G.R., Yu, Y.: Boosting for transfer learning. In: ICML'07. pp. 193–200 (2007)
8. Fortuna, B., Shawe-Taylor, J.: The use of machine translation tools for cross-lingual text mining. In: ICML'05 Workshop (KCCA) (2005)
9. Gao, J., Fan, W., Jian, J., Han, J.: Knowledge transfer via multiple model local structure mapping. In: KDD'08. pp. 283–291 (2008)
10. Gliozzo, A., Strapparava, C.: Exploiting comparable corpora and bilingual dictionaries for cross-language text categorization. In: ACL'06. pp. 553–560 (2006)
11. Grefenstette, G.: Cross-Language Information Retrieval. Kluwer Academic Publishers, Norwell, MA, USA (1998)
12. Hofmann, T.: Probabilistic latent semantic indexing. In: SIGIR'99. pp. 50–57 (1999)
13. Jebara, T.: Multi-task feature and kernel selection for svms. In: ICML'04 (July 2004)

14. Mathieu, B., Besançon, R., Fluhr, C.: Multilingual document clusters discovery. In: RIAO. pp. 116–125 (2004)
15. Mihalcea, R., Banea, C., Wiebe, J.: Learning multilingual subjective language via cross-lingual projections. In: ACL'07 (2007)
16. Mimno, D., Wallach, H.M., Naradowsky, J., Smith, D.A., McCallum, A.: Polylingual topic models. In: EMNLP'09. pp. 880–889 (2009)
17. Muramatsu, T., Mori, T.: Integration of plsa into probabilistic clir model. In: Proceedings of NTCIR-04 (2004)
18. Ni, X., Sun, J.T., Hu, J., Chen, Z.: Mining multilingual topics from wikipedia. In: WWW'09. pp. 1155–1155 (April 2009)
19. Nie, J.Y., Simard, M., Isabelle, P., Durand, R.: Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web. In: SIGIR'99. pp. 74–81 (1999)
20. Olsson, J.S., Oard, D.W., Hajič, J.: Cross-language text classification. In: SIGIR'05. pp. 645–646 (2005)
21. Raina, R., Battle, A., Lee, H., Packer, B., Ng, A.Y.: Self-taught learning: Transfer learning from unlabeled data. In: ICML'07. pp. 759–766 (June 2007)
22. Savoy, J.: Combining multiple strategies for effective monolingual and cross-language retrieval. Inf. Retr. 7(1-2), 121–148 (2004)
23. Smet, W.D., Moens, M.F.: Cross-language linking of news stories on the web using interlingual topic modelling. In: CIKM-SWSM. pp. 57–64 (2009)
24. Wan, X.: Co-training for cross-lingual sentiment classification. In: ACL-IJCNLP'09. pp. 235–243 (2009)
25. Xue, G.R., Dai, W., Yang, Q., Yu, Y.: Topic-bridged plsa for cross-domain text classification. In: SIGIR'08. pp. 627–634. New York, NY, USA (2008)
26. Yang, Y., Carbonell, J.G., Brown, R.D., Frederking, R.E.: Translingual information retrieval: Learning from bilingual corpora. Artif. Intell. 103(1-2), 323–345 (1998)
27. Zhao, B., Xing, E.P.: Bitam: Bilingual topic admixture models for word alignment. In: ACL'06 (2006)