Information Processing and Management 51 (2015) 111–147

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Probabilistic topic modeling in multilingual settings: An overview of its methodology and applications

Ivan Vulić [a,*], Wim De Smet [a], Jie Tang [b], Marie-Francine Moens [a]

[a] Department of Computer Science, KU Leuven, Celestijnenlaan 200A, 3001 Heverlee, Belgium
[b] Department of Computer Science and Technology, Tsinghua University, 100084 Beijing, China

ARTICLE INFO

Article history:
Received 4 February 2013
Received in revised form 11 August 2014
Accepted 18 August 2014
Available online 7 October 2014

Keywords:
Multilingual probabilistic topic models
Cross-lingual text mining
Cross-lingual knowledge transfer
Cross-lingual information retrieval
Language-independent data representation
Non-parallel data

ABSTRACT

Probabilistic topic models are unsupervised generative models which model document content as a two-step generation process, that is, documents are observed as mixtures of latent concepts or topics, while topics are probability distributions over vocabulary words. Recently, a significant research effort has been invested into transferring the probabilistic topic modeling concept from monolingual to multilingual settings. Novel topic models have been designed to work with parallel and comparable texts. We define multilingual probabilistic topic modeling (MuPTM) and present the first full overview of the current research, methodology, advantages and limitations in MuPTM. As a representative example, we choose a natural extension of the omnipresent LDA model to multilingual settings called bilingual LDA (BiLDA). We provide a thorough overview of this representative multilingual model from its high-level modeling assumptions down to its mathematical foundations. We demonstrate how to use the data representation by means of output sets of (i) per-topic word distributions and (ii) per-document topic distributions coming from a multilingual probabilistic topic model in various real-life cross-lingual tasks involving different languages, without any external language pair dependent translation resource: (1) cross-lingual event-centered news clustering, (2) cross-lingual document classification, (3) cross-lingual semantic similarity, and (4) cross-lingual information retrieval. We also briefly review several other applications present in the relevant literature, and introduce and illustrate two related modeling concepts: topic smoothing and topic pruning. In summary, this article encompasses the current research in multilingual probabilistic topic modeling. By presenting a series of potential applications, we reveal the importance of the language-independent and language pair independent data representations by means of MuPTM. We provide clear directions for future research in the field by providing a systematic overview of how to link and transfer aspect knowledge across corpora written in different languages via the shared space of latent cross-lingual topics, that is, how to effectively employ learned per-topic word distributions and per-document topic distributions of any multilingual probabilistic topic model in various cross-lingual applications.

* Corresponding author. Tel.: +32 16 32 87 14.
E-mail addresses: ivan.vulic@cs.kuleuven.be (I. Vulić), wdesmet@gmail.com (W. De Smet), jie.tang@tsinghua.cn.edu (J. Tang), marie-francine.moens@cs.kuleuven.be (M.-F. Moens).

http://dx.doi.org/10.1016/j.ipm.2014.08.003
0306-4573/

1. Introduction

Probabilistic latent topic models such as probabilistic Latent Semantic Analysis (pLSA) (Hofmann, 1999b, 1999a) and Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003b) along with their numerous variants are well studied generative models for representing the content of documents in large document collections. They provide a robust and unsupervised framework for performing shallow latent semantic analysis of themes (or topics) discussed in text. The families of these probabilistic latent topic models are all based upon the idea that there exist latent variables, that is, *topics*, which determine how words in documents have been generated. Fitting such a generative model actually denotes finding the best set of those latent variables in order to explain the observed data. With respect to that generative process, documents are seen as mixtures of latent topics, while topics are simply probability distributions over vocabulary words. A topic representation of a document constitutes a high-level language-independent view of its content, unhindered by a specific word choice, and it improves on text representations that contain synonymous or polysemous words (Griffiths, Steyvers, & Tenenbaum, 2007).

Probabilistic topic modeling constitutes a very general framework for unsupervised topic mining, and over the years it has been employed in miscellaneous tasks in a wide variety of research domains, e.g., for object recognition in computer vision (e.g., Li & Perona, 2005; Russell, Freeman, Efros, Sivic, & Zisserman, 2006; Wang & Grimson, 2007), dialogue segmentation (e.g., Purver, Körding, Griffiths, & Tenenbaum, 2006), video analysis (e.g., Wang, Ma, & Grimson, 2009), automatic harmonic analysis in music (e.g., Arenas-Garca et al., 2007; Hu & Saul, 2009), genetics (e.g., Blei, Franks, Jordan, & Mian, 2006), and others.

Being originally proposed for textual data, probabilistic topic models have also organically found many applications in natural language processing (NLP). Discovered distributions of words over topics (further *per-topic word distributions*) and distributions of topics over documents (further *per-document topic distributions*) can be directly employed to detect main themes[1] discussed in texts, and to provide gists or summaries for large text collections (see, e.g., Hofmann, 1999b; Blei et al., 2003b; Griffiths & Steyvers, 2004; Griffiths et al., 2007). Per-document topic distributions for each document might be observed as a low-dimensional latent semantic representation of text in a new topic-document space, potentially better than the original word-based representation in some applications. In an analogous manner, since the number of topics is usually much lower than the number of documents in a collection, per-topic word distributions also model a sort of dimensionality reduction, as the original word-document space is transferred to a lower-dimensional word-topic space. Apart from the straightforward utilization of probabilistic topic models as direct summaries of large document collections, these two sets of probability distributions have been utilized in a myriad of NLP tasks, e.g., for inferring captions for images (Blei & Jordan, 2003), sentiment analysis (e.g., Mei, Ling, Wondra, Su, & Zhai, 2007; Titov & McDonald, 2008), analyzing topic trends for different time intervals in scientific literature, social networks and e-mails (e.g., Wang & McCallum, 2006; McCallum, Wang, & Corrada-Emmanuel, 2007; Hall, Jurafsky, & Manning, 2008), language modeling in information retrieval (e.g., Wei & Croft, 2006; Yi & Allan, 2009), document classification (e.g., Blei et al., 2003b; Lacoste-Julien, Sha, & Jordan, 2008), word sense disambiguation (e.g., Boyd-Graber, Blei, & Zhu, 2007), modeling distributional similarity of terms (e.g., Ritter, Mausam, & Etzioni, 2010; Dinu & Lapata, 2010), etc. Lu, Mei, and Zhai, 2011 examine task performance of pLSA and LDA as representative monolingual topic models in typical tasks of document clustering, text categorization and ad hoc information retrieval. Data representation, i.e., representations of words and documents in all applications presented in this article will be based on those per-topic word distributions and per-document topic distributions.

However, all these models have been designed to work with monolingual data, and they have been applied in monolingual contexts only. Following the ongoing growth of the World Wide Web and its omnipresence in today's increasingly connected world, users tend to abandon English as the *lingua franca* of the global network, since more and more content becomes available in their native languages or even dialects and different community languages (e.g., the idiomatic usage of the same language typically differs between scientists, social media consumers or the legislative domain). It is difficult to determine the exact number of languages in the world, but the estimations vary between 6000 and 7000 languages and almost 40,000 unofficial languages and dialects.[2] It is extremely time-consuming and labor-intensive to build quality *translation resources* and *parallel corpora* for each single language/dialect pair. Therefore, we observe an increasing interest in language-independent unsupervised corpus-based cross-lingual text mining from non-parallel corpora without any additional translation resources. High-quality parallel corpora where documents are sentence-aligned exact translations of each other (such as Europarl (Koehn, 2005)) are available only for a restricted number of languages and domains. There has been a recent interest to build parallel corpora from the Web (e.g., Resnik & Smith, 2003; Munteanu & Marcu, 2005, 2006), but the obtained parallel data still typically remain of limited size and scope as well as domain-restricted (e.g., parliamentary proceedings).

With the rapid development of Wikipedia and online social networks such as Facebook or Twitter, users have generated a huge volume of multilingual text resources. The user-generated data are often noisy and unstructured, and seldom well-paired across languages. However, unlike parallel corpora, such *comparable corpora*, where texts in one language are paired with texts in another language discussing the same themes or subjects, are abundant in various online sources (e.g., Wikipedia or news sites). Documents from comparable corpora do not necessarily share all their themes with their counterparts in the other language, but, for instance, Wikipedia articles discussing the same subject, or news stories discussing the

[1] To avoid confusion, we talk about *themes* when we address the true content of a document, while we talk about topics when we address the probability distributions constituting a topic model.

[2] Source: http://www.ethnologue.com.

same event contain a significant thematic overlap. We could say that such documents in different languages, although inherently non-parallel, are *theme-aligned*.

Multilingual probabilistic topic models (MuPTM-s) have recently emerged as a group of unsupervised, language-independent generative machine learning models that can be efficiently utilized on such large-volume non-parallel theme-aligned multilingual data and effectively deal with uncertainty in such data collections. Due to its generic language-independent nature and the power of inference on unseen documents, these models have found many interesting applications. The knowledge from learned MuPTM-s has been used in many different cross-lingual tasks such as cross-lingual event clustering (DeSmet & Moens, 2009), cross-lingual document classification (De Smet, Tang, & Moens, 2011; Ni, Sun, Hu, & Chen, 2011), cross-lingual semantic similarity of words (Mimno, Wallach, Naradowsky, Smith, & McCallum, 2009; Vulić, DeSmet, & Moens, 2011a; Vulić & Moens, 2012), cross-lingual information retrieval (Vulić, Smet, & Moens, 2011b, 2013; Vulić & Moens, 2013; Ganguly, Leveling, & Jones, 2012) and others.

The main goal of this work is to provide an overview of the recently developed multilingual probabilistic topic modeling concept. It aims to model topic discovery from multilingual data in a conceptually sound way, taking into account thematic alignment between documents in document collections given in different languages. We have decided to provide a thorough analysis of the framework of multilingual probabilistic topic modeling because we feel that the current relevant literature lacks a systematic and complete overview of the subject. Moreover, during our tutorials at ECIR 2013 and WSDM 2014 on the subject we realized even more that, after being provided with feedback on our tutorials, it would be extremely beneficial for the IR community to have an extended written overview of the whole subject, along with its formalisms, definitions and modeling perspectives (both conceptual and mathematical), relevant state-of-the-art, a broad relevant references list, and also with standard evaluation procedures, an overview of applications, and suggestions for future work.

As a representative example, we choose bilingual LDA, which has been designed as a basic and natural extension of the standard omnipresent LDA model in the multilingual settings where document-aligned articles in different languages are available (e.g., Wikipedia articles about the same subject in multiple languages). We provide a complete and comprehensive overview of that model all the way up from the conceptual and modeling level down to its core mathematical foundations as it could serve as a valuable starting point for other researchers in the field of multilingual probabilistic topic modeling and cross-lingual text mining. Alternative multilingual probabilistic topic models that build upon the idea of the standard pLSA and LDA models are presented in a nutshell. These models differ in the specific assumptions they make in their generative processes as well as in knowledge that is presupposed before training (e.g., document alignment, prior word matchings or bilingual dictionaries), but all these models have the ability to discover latent cross-lingual topics from comparable data such as Wikipedia or news. Additionally, all these models output the same basic sets of probability distributions, that is, per-topic word distributions and per-document topic distributions. Finally, we also demonstrate how to utilize the high-level structured text representations by means of per-topic word distributions and per-document topic distributions from any multilingual probabilistic topic model, and establish knowledge transfer across different languages via the shared space of latent language-independent concepts, that is, cross-lingual topics in several real-life NLP/IR tasks: (1) cross-lingual event-centered news clustering, (2) cross-lingual document classification, (3) cross-lingual semantic similarity, and (4) cross-lingual information retrieval. In this article, we show the results obtained by BiLDA, but the presented solutions are completely topic model-independent.

The results reported across all these tasks show the validity of multilingual comparable data as training data, as well as the superiority of MuPTM over monolingual probabilistic topic modeling (MoPTM) and other data-driven modeling paradigms which do not rely on any expensive translation resources. We also expose and discuss an issue present in all current multilingual topic models – a need to set the number of topics in advance before training. Different applications reach their optimal results with different number of topics set a priori and it is often difficult to accurately predict that application-dependent number of topics in advance. We also report on a mismatch between the standard intrinsic evaluation measure of perplexity and the extrinsic evaluation in terms of final scores in the cross-lingual tasks.

2. Multilingual probabilistic topic modeling

This section presents and defines the basic concepts and modeling assumptions related to multilingual probabilistic topic modeling, with a special focus on learning from comparable theme-aligned corpora. We also draw an analogy to the broader paradigm of *latent cross-lingual concepts*, and their relation to latent cross-lingual topics. Following that, the representative bilingual LDA (BiLDA) is presented in its entirety, which includes its generative story, the explanation of the Gibbs sampling training procedure for the model, the output of the model in terms of per-topic word distributions and per-document topic distributions, the procedure to infer the trained model on unseen data, and a qualitative analysis of its output in terms of the top N most important words for some selected topics. We also define several evaluation metrics utilized to compare different topic models. At the very end of the section, alternative multilingual topic models are also presented.

2.1. Definitions and assumptions

Definition 1 (*Multilingual theme-aligned corpus*). In the most general definition, a *multilingual theme-aligned corpus* $\mathcal{C}$ of $l = |\mathcal{L}|$ languages, where $\mathcal{L} = \{L_1, L_2, \ldots, L_l\}$ is the set of languages, is a set of corresponding text collections $\{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_l\}$. Each $\mathcal{C}_i = \{d_1^i, d_2^i, \ldots, d_{dn_i}^i\}$ is a collection of documents in language L_i with vocabulary $V^i = \{w_1^i, w_2^i, \ldots, w_{wn_i}^i\}$. Collections $\{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_l\}$ are said to be *theme-aligned* if they discuss at least a portion of similar themes. Here, dn_i denotes the total

number of documents in document collection $\mathcal{C}_i$, while wn_i is the total number of words in V^i. Moreover, d_j^i denotes the j-th document in document collection $\mathcal{C}_i$, and w_j^i denotes the j-th word in vocabulary V^i associated with document collection $\mathcal{C}_i$.

Definition 2. (*Multilingual probabilistic topic model*). A *multilingual probabilistic topic model* of a theme-aligned multilingual corpus $\mathcal{C}$ is a set of semantically coherent multinomial distributions of words with values $P_i\left(w_j^i \mid z_k\right), i = 1, \ldots, l$, for each vocabulary $V^1, \ldots, V^i, \ldots, V^l$ associated with text collections $\mathcal{C}_1, \ldots, \mathcal{C}_i, \ldots, \mathcal{C}_l \in \mathcal{C}$ given in languages $L_1, \ldots, L_i, \ldots, L_l$. $P_i\left(w_j^i \mid z_k\right)$ is calculated for each $w_j^i \in V^i$. The probabilities $P_i\left(w_j^i \mid z_k\right)$ build *per-topic word distributions* (denoted by ϕ_i), and they constitute a language-specific representation (e.g., a probability value is assigned only for words from V^i) of a language-independent latent cross-lingual concept – topic $z_k \in \mathcal{Z}$. $\mathcal{Z} = \{z_1, \ldots, z_K\}$ represents the set of all K latent cross-lingual topics present in the multilingual corpus. Each document in the multilingual corpus is thus considered a mixture of K latent cross-lingual topics from the set $\mathcal{Z}$. This mixture for some document $d_j^i \in \mathcal{C}_i$ is modeled by the probabilities $P_i\left(z_k \mid d_j^i\right)$ that altogether build *per-document topic distributions* (denoted by θ). In summary, each language-independent latent cross-lingual topic z_k has some probability to be found in a particular document (modeled by per-document topic distributions), and each such topic has a language-specific representation in each language (modeled by language-specific per-topic word distributions).

We can interpret Definition 2 in the following way: each cross-lingual topic from the set $\mathcal{Z}$ can be observed as a latent language-independent concept present in the multilingual corpus, but each language in the corpus uses only words from its own vocabulary to describe the content of that concept (see Fig. 1 for an illustrative example). In other words, we could observe each latent cross-lingual topic as a set of discrete distributions over words, one for each language. For instance, having a multilingual collection in English, Italian and Dutch and discovering a topic on *Soccer*, that cross-lingual topic would be represented by words (actually probabilities over words) {*player*, *goal*, *scorer*, ...} in English, {*squadra (team), calcio (soccer), allenatore (coach)*, ...} in Italian, and {*doelpunt (goal), voetballer (soccer player), elftal (soccer team)*, ...} in Dutch. We have

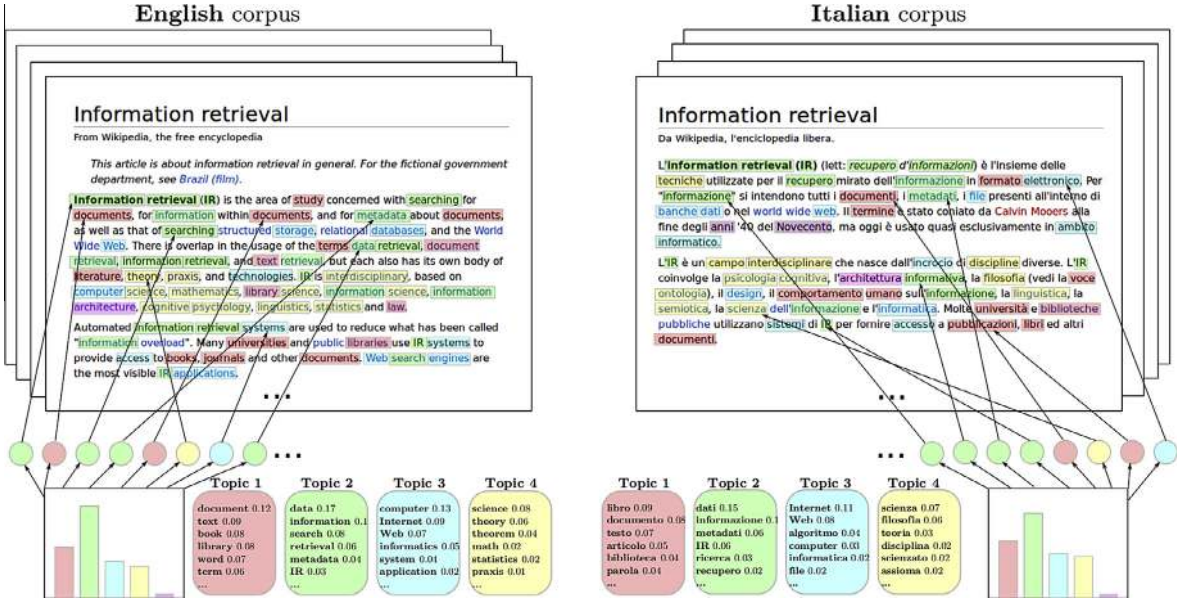


Fig. 1. An illustrative overview of the intuitions behind multilingual probabilistic topic modeling. Each document is represented as a mixture of latent cross-lingual topics (*per-document topic distributions*, presented by histograms), where some latent cross-lingual topics are more important for the particular document. These cross-lingual topics are language-independent concepts, but each language provides a language-specific interface to each cross-lingual topic. In other words, each cross-lingual topic is modeled as a distribution over vocabulary words in each language (*per-topic word distributions*, presented by rounded rectangles). Each document is then generated as follows. First, choose the per-document topic distribution and, according to the distribution, for each word position choose a topic assignment (the colored circles). Following that, according to per-topic word distributions in that language, choose the specific word from the corresponding latent cross-lingual topic that will occur at that word position. Documents that discuss similar themes tend to have similar distributions over cross-lingual topics (the colored bars), but when we operate in the multilingual setting, different per topic-word distributions (the rounded rectangles) are used to generate the observed words in the documents. The generative process does not make any assumptions about syntax, grammar and word order in general, as it assumes that each word is *independently and identically distributed (iid)*, that is, drawn independently from the same distribution (the *bag-of-words assumption)*. Extending the models beyond the bag-of-words assumption (or rather restriction) is possible, but it will not be covered in this article. The figure represents an illustrative toy example, and it is not based on real data. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

$\sum_{w_j^i \in V^i} P_i\left(w_j^i \mid z_k\right) = 1$, for each vocabulary V^i representing language L_i, and for each topic $z_k \in \mathcal{Z}$. We say that a latent cross-lingual topic is *semantically coherent* if it assigns high probabilities to words that are semantically related. Definition 2 is predominant in the MuPTM literature (e.g., see DeSmet & Moens, 2009; Mimno et al., 2009; Platt, Toutanova, & Yih, 2010).

Zhang, Mei, and Zhai (2010) provide an alternative, more general definition of a multilingual topic model, but we will show that their definition may be brought down to Definition 2 after a partition over the languages is performed. Namely, the whole multilingual corpus is observed as a mixture of latent cross-lingual topics from $\mathcal{Z}$. They then define a latent cross-lingual topic $z_k \in \mathcal{Z}$ as a semantically coherent multinomial distribution over all the words in all the vocabularies of languages $L_1, \ldots, L_i, \ldots, L_l$, and $P(w_j \mid z_k)$ gives the probability of any word $w_j \in \{V^1, \ldots, V^i, \ldots, V^l\}$ to be generated by topic z_k. In this case, we have $\sum_{i=1}^{l} \sum_{w_j \in V^i} P(w_j \mid z_k) = 1$. The language-specific representation for language L_i of topic z_k is then obtained by retaining only probabilities for words which are present in its own vocabulary V^i, and normalizing those distributions. For a word $w_j^i \in V^i$, we have $P_i\left(w_j^i \mid z_k\right) = \frac{P\left(w_j^i \mid z_k\right)}{\sum_{w_j \in V^i} P\left(w_j \mid z_k\right)}$. After the partition over languages and normalizations are performed, this definition is effectively equivalent to Definition 2. However, note that their original definition is more general than Definition 2, but it is also unbalanced over the languages from $\mathcal{L}$ present in $\mathcal{C}$, that is, words from the languages that are more present in the original corpus $\mathcal{C}$ might dominate the multinomial per-topic word distributions. By performing the partition and normalization over the languages, that imbalance is effectively removed.

Definition 3 (*Multilingual probabilistic topic modeling*). Given a theme-aligned multilingual corpus $\mathcal{C}$, the goal of *multilingual probabilistic topic modeling* or *latent cross-lingual topic extraction* is to learn and extract a set $\mathcal{Z}$ of K latent language-independent concepts, that is, *latent cross-lingual topics* $\mathcal{Z} = \{z_1, \ldots, z_K\}$ that optimally describe the observed data, that is, the multilingual corpus $\mathcal{C}$. Extracting latent cross-lingual topics actually implies learning *per-document topic distributions* for each document in the corpus, and discovering language-specific representations of these topics given by *per-topic word distributions* in each language (see Definition 2).

This shared and language-independent set of latent cross-lingual topics $\mathcal{Z}$ serves as the core of unsupervised *cross-lingual text mining* and *cross-lingual knowledge linking and transfer* by means of multilingual probabilistic topic models. It is the cross-lingual connection that bridges the gap across documents in different languages and transfers knowledge across languages in case when translation resources and labeled instances are scarce or missing. The trained multilingual probabilistic topic model may be further inferred on unseen documents.

Definition 4 (*Inference of a multilingual probabilistic topic model*). Given an unseen document collection $\mathcal{C}_u$, the inference of a multilingual topic model on the collection $\mathcal{C}_u$ denotes learning topical representations of the unseen documents $d_u \in \mathcal{C}_u$, that is, acquiring per-document topic distributions for the new documents based on the previous output of the model.

Definition 5 (*Cross-lingual knowledge transfer*). Knowledge transfer in general refers to transferring knowledge learned from one corpus to another corpus, which was unavailable during the learning procedure. *Cross-lingual knowledge transfer* is characterized by the fact that corpora are present in more than one language.

Additionally, following the assumptions and general definitions provided in this section, *monolingual probabilistic topic models* such as pLSA (Hofmann, 1999b, 1999a) and LDA (Blei et al., 2003b) could be interpreted as a degenerate special case of multilingual probabilistic topic models where only one language is involved, and all the definitions and assumptions remain the same (see later the discussion in Section 2.3).

2.2. A more general framework: latent cross-lingual concepts (intermezzo)

The latent cross-lingual topics presented in Section 2.1 constitute only one possibility when the aim is to detect and induce a *latent semantic structure* from multilingual data, that is, to extract *latent cross-lingual concepts* that are hidden within the data. Latent cross-lingual concepts may be interpreted as language-independent semantic concepts present in a multilingual corpus (e.g., document-aligned Wikipedia articles in English and Spanish) that have their language-specific representations in different languages. To repeat, for instance, having a multilingual collection in English, Spanish and Croatian, and discovering a latent semantic concept on *Basketball*, that concept would be represented by words (actually probabilities over words) {*player, ball, coach, ...*} in English, {*pelota (ball), jugador (player), partido (match), ...*} in Spanish, and {*trener (coach), razigravač (playmaker), doigravanje (playoff), ...*} in Croatian.

These K semantic concepts span a latent cross-lingual semantic space. Each word w may be represented in that latent semantic space as a K-dimensional vector, where each vector component is a *conditional concept probability score* $P(z_k \mid w)$. In other words, each word is actually represented as a multinomial probability distribution over the induced latent cross-lingual semantic concepts. Moreover, each document d, regardless of its actual language, may be represented as a multinomial probability distribution, a mixture over the same induced latent cross-lingual semantic concepts $P(z_k \mid d)$.

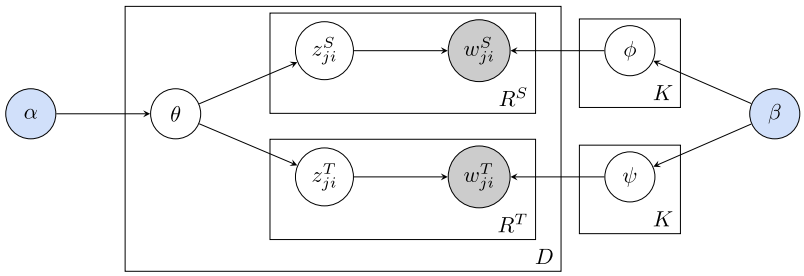


Fig. 2. Graphical representation of the bilingual LDA (BiLDA) model in plate notation. R^S and R^T denote lengths of the source document and the target document in terms of word tokens for each aligned document pair.

The description in this article relies on the multilingual probabilistic topic modeling framework, but we emphasize that all the work described in this article is independent of the actual method used to induce the latent cross-lingual concepts. The reader has to be aware of the fact that the description of how to utilize this latent knowledge in applications is generic and model-independent as they allow the usage of all other models that compute probability scores $P(z_k \mid w)$ and $P(z_k \mid d)$ (obtained from per-topic word distributions and per-document topic distributions). Besides MuPTM, a number of other models may be employed to induce the latent cross-lingual concepts. For instance, one could use cross-lingual Latent Semantic Indexing (Dumais, Landauer, & Littman, 1996), probabilistic principal component analysis (Tipping & Bishop, 1999), LDA (Blei et al., 2003b), or a probabilistic interpretation of non-negative matrix factorization (Lee & Seung, 1999; Gaussier & Goutte, 2005; Ding, Li, & Peng, 2008) on concatenated documents in aligned document pairs. Other more recent models include matching canonical correlation analysis (Haghighi, Liang, Berg-Kirkpatrick, & Klein, 2008; Daumé III & Jagarlamudi, 2011) or other families of multilingual topic models (Fukumasu, Eguchi, & Xing, 2012).

2.3. A representative example: Bilingual Latent Dirichlet Allocation (BiLDA)

Without loss of generality, from now on we deal with *bilingual probabilistic topic modeling*. We work with a *bilingual corpus* and present cross-lingual tasks in the *bilingual setting*. For bilingual corpora we introduce the source language L_S (further with indices S) and the target language L_T (further with indices T). We will show that all the definitions and assumptions may be easily generalized to a setting where more than two languages are available.

2.3.1. Main modeling assumptions

Bilingual Latent Dirichlet Allocation (BiLDA) is a bilingual extension of the standard LDA model (Blei et al., 2003b), tailored for modeling parallel or, even more importantly, comparable theme-aligned bilingual document collections. An example of such a document collection is Wikipedia in 2 languages with paired articles. BiLDA has been independently designed by several researchers (Ni, Sun, Hu, & Chen, 2009; DeSmet & Moens, 2009; Mimno et al., 2009; Platt et al., 2010). Unlike LDA, where each document is assumed to possess its own document-specific distribution over topics, the generative process for BiLDA assumes that each *aligned document pair* shares the same distribution of topics. Therefore, the model assumes that we already possess *document alignments* in a corpus, that is, links between paired documents in different languages in a bilingual (or a multilingual) corpus. This assumption is certainly valid for multilingual Wikipedia data, where document alignment is established via cross-lingual links between articles written in different languages. These links are provided by the nature of the Wikipedia structure. Cross-lingual document alignment for news crawled from the Web is also a well-studied problem. Since the establishing of cross-lingual links between similar documents is not the focus of the research reported here, these algorithms are not elaborated in the article, but we refer the curious reader to the literature (see, e.g., Utiyama & Isahara, 2003; Resnik & Smith, 2003; Tao & Zhai, 2005; Munteanu & Marcu, 2006; Vu, Aw, & Zhang, 2009).

Definition 6 (*Paired bilingual corpus*). A *paired bilingual document corpus* is defined as $\mathcal{C} = \{d_1, d_2, \ldots, d_r\} = \{(d_1^S, d_1^T), (d_2^S, d_2^T), \ldots, (d_r^S, d_r^T)\}$, where $d_j = (d_j^S, d_j^T)$ denotes the j-th pair of linked documents in the source language L_S and the target language L_T, respectively. The goal of bilingual probabilistic topic modeling is to learn for a (paired or non-paired) bilingual corpus a set of K latent cross-lingual topics $\mathcal{Z}$, each of which defines an associated set of words in both L_S and L_T.

2.3.2. The model

BiLDA can be observed as a three-level Bayesian network that models document pairs using a latent layer of shared topics. Fig. 2 shows the graphical representation of the BiLDA model in plate notation, while Algorithm 1 presents its generative story.

Algorithm 1. Generative story for BiLDA

initialize: (1) set the number of topics K;
(2) set values for Dirichlet priors α and β;
sample: K times $\phi \sim Dirichlet(\beta)$;
sample: K times $\psi \sim Dirichlet(\beta)$;
foreach *aligned document pair* $d_j = \{d_j^S, d_j^T\}$ **do**
  sample $\theta_j \sim Dirichlet(\alpha)$
  foreach *word position* $i \in d_j^S$ **do**
    sample $z_{ji}^S \sim Multinomial(\theta)$
    sample $w_{ji}^S \sim Multinomial(\phi, z_{ji}^S)$
  end
  foreach *word position* $i \in d_j^T$ **do**
    sample $z_{ji}^T \sim Multinomial(\theta)$
    sample $w_{ji}^T \sim Multinomial(\psi, z_{ji}^T)$
  end
end

BiLDA takes advantage of the assumed topical alignment at the level of linked documents by introducing a single variable θ (see Section 2.1) shared by both documents. θ_j denotes the distribution of latent cross-lingual topics over each document pair d_j. For each document pair d_j, a per-document topic distribution[3] θ_j is sampled from a conjugate Dirichlet prior[4] with K parameters $\alpha_1, \ldots, \alpha_K$. Then, with respect to θ_j, a cross-lingual topic z_{ji}^S is sampled. Each word w_{ji}^S at the position i in the source document of the current document pair d_j is then generated from a multinomial distribution $\phi_{z_{ji}^S}$. Similarly, each word w_{ji}^T of the target language[5] is also sampled following the same procedure. Note that words at the same positions in source and target documents in a document pair need not be sampled from the same latent cross-lingual topic. The only constraint imposed by the model is that the overall distributions of topics over documents in a document pair modeled by θ_j have to be the same. The validity of this assumption/constraint is dependent on the actual degree of thematic alignment of two coupled documents, as well as on the chosen topic granularity (e.g., two Wikipedia articles about the same subject may have the same focus and share global topics, but at a finer scale, they might exhibit different sub-focuses and do not share a subset of other more local topics).

2.3.3. Hyper-parameters

According to Griffiths et al. (2007), each hyper-parameter α_k could be interpreted as a prior observation count for the number of times topic z_k is sampled in a document (or document pair) before having observed any actual words. If one is in possession of a certain prior or external knowledge (e.g., document metadata, main themes of a document collections) about the topic importance and the likelihood of its presence in the data, introducing asymmetric priors gives more preference to a subset of the most important topics, which could in the end lead to a better estimated set of output distributions (Mimno & McCallum, 2008; Jagarlamudi, DauméIII, & Udupa, 2012). However, it is often the case that we do not possess any prior knowledge about themes in a text collection, and then it is reasonable to assume that all topics are a priori equally likely. Therefore, it is convenient to use a symmetric Dirichlet distribution with a single hyper-parameter α such that $\alpha_1 = \cdots = \alpha_K = \alpha$. Similarly, a symmetric Dirichlet prior is placed on ϕ and ψ with a single hyper-parameter β. β may be interpreted as a prior observation count of the number of times words in each language are sampled from a topic before any observations of actual words. Placing these Dirichlet prior distributions on multinomial distributions θ, ϕ and ψ results in smoothed per-topic word and per-document topic distributions, where the values for α and β determine the degree of smoothing. The influence of these hyper-parameters on the quality of learned latent topics is a well-studied problem in monolingual settings (Asuncion, Welling, Smyth, & Teh, 2009; Lu et al., 2011) and it can be generalized to multilingual settings.

2.3.4. Extending BiLDA to more languages

A natural extension of BiLDA that operates with more than two languages, called *polylingual topic model* (PolyLDA) has been presented by Mimno et al. (2009). A similar model has also been proposed by Ni et al. (2009, 2011). Instead of document pairs, they deal with aligned *document tuples* (where links between documents in a tuple are given), but the assumptions made by their model remain the same. Fig. 3 shows the graphical representation in plate notation of the BiLDA model generalized to l languages, $l \geqslant 2$, with document tuples $d_j = \{d_j^1, \ldots, d_j^l\}$ and a discrete set of l language-specific per-topic word distributions $\{\phi_1, \ldots, \phi_l\}$ (see Section 2.1).

[3] The correct term here should be per-pair topic distribution for BiLDA and per-tuple topic distribution in case when more than 2 languages are involved, but we have decided to retain the original name of the distribution in order to draw a direct comparison with standard monolingual LDA.

[4] For an introduction to conjugate distributions, priors and Bayesian inference, we refer the curious reader to the excellent Heinrich's overview (Heinrich, 2008).

[5] Both words (w-s) and topics (z-s) are annotated with a corresponding superscript S or T to denote which language they are used in.

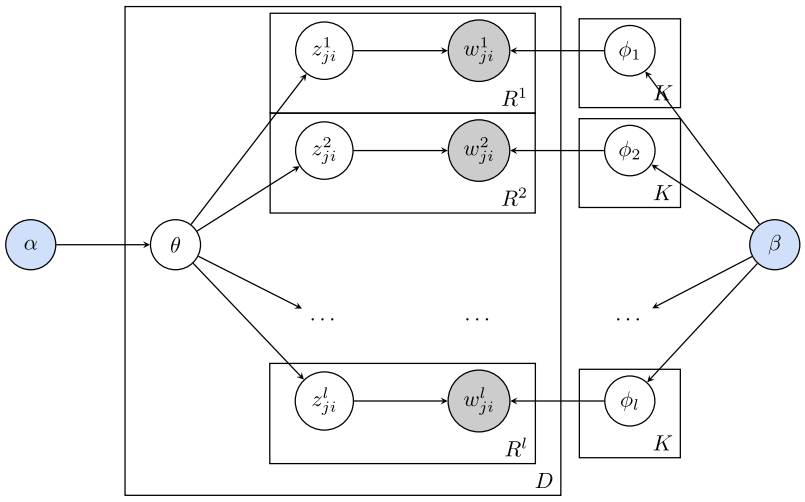


Fig. 3. Polylingual topic model: The generalization of the BiLDA model which operates with l languages, $l \geqslant 2$.

On the other hand, when operating with only one language, BiLDA or (more generally) PolyLDA is effectively reduced to the standard monolingual LDA model (see Fig. 4 and compare it with Fig. 2 or Fig. 3) (Blei et al., 2003b), that is, the monolingual LDA model is only a degenerate special case of BiLDA and PolyLDA (see also Section 2.1).

2.3.5. Training: estimating the BiLDA model

The goal of training the BiLDA model is to discover the layer of latent cross-lingual topics that describe observed data, that is, a given bilingual document collection in an optimal way. It means that the most likely values for θ, ϕ and ψ have to be found by the training procedure. In simple words, we need to detect and learn which words are important for a particular topic in each language (that is reflected in per-topic word distributions ϕ and ψ), and which topics are important for a particular document pair (as reflected in per-document topic distribution θ). Similarly to the LDA model, the topic discovery for BiLDA is complex and cannot be solved by an exact learning procedure. There exist a few approximative training techniques which aim at converging to the correct distributions. Variational estimation for the monolingual LDA was used as the estimation technique in the seminal paper by Blei et al. (2003b). Other estimation techniques for the monolingual case include Gibbs sampling (Geman & Geman, 1984; Steyvers & Griffiths, 2007), and expectation propagation (Minka & Lafferty, 2002; Griffiths & Steyvers, 2004).

An extension of the variational method to multilingual settings and its complete formulation for BiLDA was proposed and described by the authors (DeSmet & Moens, 2009). Due to its prevalent use in topic modeling literature in both monolingual and multilingual contexts (Boyd-Graber & Blei, 2009; Mimno et al., 2009; Jagarlamudi & Daumé III, 2010; Vulić et al., 2011a), we opt for Gibbs sampling as the estimation technique for the BiLDA models in all applications described in this article. Therefore, we here provide an overview of Gibbs sampling for BiLDA.

Gibbs sampling is a Monte Carlo Markov chain (MCMC) estimation technique. MCMC is a random walk over a Markov chain where each state represents a sample from a specific joint distribution. Starting from a random initial state, the next state is repeatedly sampled randomly from the transition probabilities, and this is repeated until the equilibrium state is reached, in which case states are samples from the joint probability distribution. Gibbs sampling considers each word token in a text collection in turn and then samples a topic for that word token, where the probability of generating the current word by each topic is calculated conditioned given all other variables (including all other topics). For BiLDA in specific, the Gibbs sampling procedure follows the steps presented in Algorithm 2.

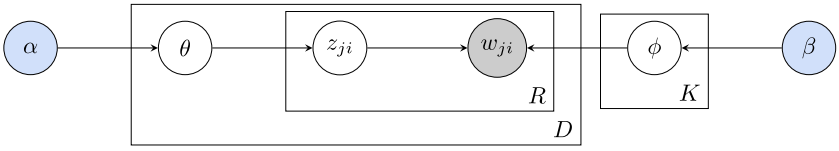


Fig. 4. Standard monolingual LDA model from Blei et al. (2003b).

Algorithm 2. GIBBS SAMPLING FOR BILDA: AN OVERVIEW

repeat
  consider each word token in each document in the collection **in turn**;
  update/estimate the probability to assign the word token to one of the cross-lingual topics conditioned on all other variables (including all other topic assignments);
  sample the actual topic assignment for the word token according to the estimated probabilities;
until *convergence/the equilibrium state* ;

After the convergence or the equilibrium state is reached, a standard practice is to provide estimates of the output distributions as averages over several samples taken in the equilibrium state.

BiLDA requires two sets of formulas to converge to correct distributions: (i) one for each topic assignment z^S_{ji} (a topic assigned to a word position i that generated word w^S_{ji} in a document pair d_j) and (ii) one for each topic assignment z^T_{ji}. θ, ψ and ϕ are not calculated directly, but estimated afterwards. Therefore, they are integrated out of all the calculations, which actually leaves topic assignments for each word position, z^S_{ji}-s and z^T_{ji}-s as the only hidden variables. For the source part S of each document pair d_j and each word position i, the probability is calculated that z^S_{ji} assumes, as its new values, one of the K possible topic indices (from a set of K topics), as indicated by variable z_k:

$$\text{sample } z^S_{ji} \sim P\left(z^S_{ji} = z_k \mid \mathbf{z}^S_{\neg ji}, \mathbf{z}^T_j, \mathbf{w}^S, \mathbf{w}^T, \alpha, \beta\right) \sim \int_{\theta_j}\int_{\phi} P\left(z^S_{ji} = z_k \mid, \mathbf{z}^S_{\neg ji}, \mathbf{z}^T_j, \mathbf{w}^S, \mathbf{w}^T, \alpha, \beta, \theta_j, \phi\right) d\phi \; d\theta_j \tag{1}$$

In this formula, $\mathbf{z}^T_j$ refers to all target topic indices for document pair d_j, and $\mathbf{z}^S_{\neg ji}$ denotes all source topic indices in d_j excluding z^S_{ji}. $\mathbf{w}^S$ denotes all source word tokens in the corpus, $\mathbf{w}^T$ all target words. Sampling for the target side (indices T) is performed in an analogous manner:

$$\text{sample } z^T_{ji} \sim P\left(z^T_{ji} = z_k \mid \mathbf{z}^T_{\neg ji}, \mathbf{z}^S_j, \mathbf{w}^S, \mathbf{w}^T, \alpha, \beta\right) \sim \int_{\theta_j}\int_{\psi} P\left(z^T_{ji} = z_k \mid, \mathbf{z}^T_{\neg ji}, \mathbf{z}^S_j, \mathbf{w}^S, \mathbf{w}^T, \alpha, \beta, \theta_j, \psi\right) d\psi \; d\theta_j \tag{2}$$

We further show the derivation of the Gibbs sampler for BiLDA and explain the notation only for the source side of a bilingual corpus and the source language L_S with indices S, since the derivation for the target side (with indices T) follows in a completely analogous manner. Starting from Eq. (1), we can further write:

$$\begin{aligned}\text{sample } z^S_{ji} &\propto \int_{\theta_j}\int_{\phi} P\left(z^S_{ji} = z_k \mid \mathbf{z}^S_{\neg ji}, \mathbf{z}^T_j, \theta, \alpha\right) \cdot P\left(w^S_{ji} \mid z^S_{ji} = z_k, \mathbf{z}^S_{\neg ji}, \mathbf{w}^S_{\neg ji}, \phi, \beta\right) d\phi \; d\theta \\ &\propto \int_{\theta_j} P\left(z^S_{ji} = z_k \mid \theta_j\right) \cdot P\left(\theta_j \mid \mathbf{z}^S_{\neg ji}, \mathbf{z}^T_j, \alpha\right) d\theta_j \cdot \int_{\phi_k} P\left(w^S_{ji} \mid z^S_{ji} = z_k, \phi_k\right) \cdot P\left(\phi_k | \mathbf{z}^S_{\neg ji}, \mathbf{w}^S_{\neg ji}, \beta\right) d\phi_k\end{aligned}$$

Both θ and ϕ have a prior Dirichlet distribution and their posterior distributions are updated with the counter variable n (which counts the number of assigned topics in a document) and the counter variable v (which counts the number of assigned topics in the corpus) respectively (see the explanations of the symbols after the derivation). The expected values $(\int xf(x)dx)$ for θ and ϕ become:

$$= E_{Dirichlet(n^S_{j,k,\neg i} + n^T_{j,k} + \alpha)}[\theta_{j,k}] \cdot E_{Dirichlet(v^S_{k,w^S_{ji},\neg} + \beta)}[\phi^{w_{ji}}_k] \tag{3}$$

Following Eq. (3), the final updating formulas for both source and target language for the BiLDA Gibbs sampler are as follows:

$$P\left(z^S_{ji} = z_k | \mathbf{z}^S_{\neg ji}, \mathbf{z}^T_j, \mathbf{w}^S, \mathbf{w}^T, \alpha, \beta\right) \propto \frac{n^S_{j,k,\neg i} + n^T_{j,k} + \alpha}{n^S_{j,\cdot,\neg i} + n^T_{j,\cdot} + K\alpha} \cdot \frac{v^S_{k,w^S_{ji},\neg} + \beta}{v^S_{k,\cdot,\neg} + |V^S|\beta} \tag{4}$$

$$P\left(z^T_{ji} = z_k | \mathbf{z}^T_{\neg ji}, \mathbf{z}^S_j, \mathbf{w}^S, \mathbf{w}^T, \alpha, \beta\right) \propto \frac{n^T_{j,k,\neg i} + n^S_{j,k} + \alpha}{n^T_{j,\cdot,\neg i} + n^S_{j,\cdot} + K\alpha} \cdot \frac{v^T_{k,w^T_{ji},\neg} + \beta}{v^T_{k,\cdot,\neg} + |V^T|\beta} \tag{5}$$

The counter variable $n^S_{j,k}$ denotes the number of times source words in the source document d^S_j of a document pair d_j are assigned to a latent cross-lingual topic z_k (with index k), while $n^S_{j,k,\neg i}$ has the same meaning, but not counting the current w^S_{ji} at position i (i.e., it is $n^S_{j,k} - 1$). The same is true for the target side and the T indices. When a "$\cdot$" occurs in the subscript of a counter variable, this means that the counts range over all values of the variable whose index the "$\cdot$" takes. So, while $n^S_{j,k}$ counts the number of assignments of words w^S_{ji} to one latent topic z_k in $d^S_j, n^S_{j,\cdot}$ does so over all K topics in d^S_j.

The second counter variable, $v^S_{k,w_{ji},\neg}$ is the number of times a word type whose token appears at position i (w^S_{ji}) gets assigned a latent cross-lingual topic z_k in the source side of the entire document collection, but not counting the current w^S_{ji} (i.e., it is $v^S_{k,w_{ji}} - 1$). Additionally, $\mathbf{z}^S_j$ denotes all latent topic assignments for the source side of the document pair d_j, $\mathbf{z}^S_{\neg ji}$ denotes all topic assignments for the source side of d_j but excluding w^S_{ji}. $v^S_{k,\cdot,\neg}$ counts the total number of occurrences of source language words from V^S associated with the topic z_k in the whole corpus, as it is the sum over all possible source language words (a "·" appears instead of the w^S_{ji}). Again, because of the $\neg$ symbol in the superscript, the current w^S_{ji} is not counted (i.e., the count is then $v^S_{k,\cdot} - 1$). Finally, $|V^S|$ and $|V^T|$ are vocabulary sizes for the source and the target language, respectively.

As can be seen from the first term of Eqs. (4) and (5), the document pairs are linked by the counter variables n^S_j and n^T_j, as both sets of assignments: z^S_{ji} and z^T_{ji} are drawn from the same θ_j (see the first term which is exactly the same in the updating formulas described by Eqs. (4) and (5)). The vocabulary counter variables operate only within the language of the word token currently being considered.

2.3.6. Output: per-document topic and per-topic word distributions

With formulas (4) and (5). each z^S_{ji} and z^T_{ji} of each document pair is sampled and cyclically updated. After a random initialization, usually using a uniform distribution, the sampled values will converge to samples taken from the real joint distribution of θ, ϕ and ψ, after a time called the *burn-in* period. From a set of complete *burned-in Gibbs samples* of the whole document collection, the final output probability distributions, that is, per-topic word distributions and per-document topic distributions are estimated as averages over these samples.

Language-independent *per-document topic distributions* provide distributions of latent cross-lingual topics for each document in a collection. They reveal how important each topic is for a particular document. We need to establish the exact formula for per-document topic distributions for documents in an aligned document pair using Eqs. (4) and (5):

$$P(z_k|d_j) = \theta_{j,k} = \frac{n^S_{j,k} + n^T_{j,k} + \alpha}{\sum_{k^*=1}^{K} n^S_{j,k^*} + \sum_{k^*=1}^{K} n^T_{j,k^*} + K\alpha} \tag{6}$$

The representation $Rep(d_j)_{\mathcal{Z}}$ of the document d_j by means of latent cross-lingual topics $\mathcal{Z}$ is then a K-dimensional vector where the k-th dimension of the vector is exactly the probability of the latent cross-lingual topic z_k in the document d_j:

$$Rep(d_j)_{\mathcal{Z}} = [P(z_1|d_j), P(z_2|d_j), \ldots, P(z_K|d_j)] \tag{7}$$

We may detect from Eq. (6) that two documents from an aligned document pair are enforced to have exactly the same topical representations, that is, the two documents discussing the same themes will be presented as exactly the same mixtures over the induced latent cross-lingual topics. This property is achieved by making the computation of $P(z_k \mid d_j)$ for both documents in the pair explicitly dependent on the topic assignments counts from the source language document ($n^S_{j,k}$) as well as the target language document ($n^T_{j,k}$). Previous standard approaches to multilingual probabilistic topic modeling (see later Section 2.5) typically trained monolingual LDA on concatenated documents from an aligned document pair, where this property was not taken into account. In other words, unlike BiLDA, monolingual LDA builds a single topical representation for the artificially created concatenated document, and does not enforce this property at all. Comparisons of monolingual LDA trained on concatenated documents forming aligned document pairs (further *MixLDA*) and bilingual LDA reveal the superiority of the true multilingual approach modeled by BiLDA.

Language-specific *per-topic word distributions* measure the importance of each word in each language for a particular latent cross-lingual topic z_k. Given the source language with vocabulary V^S, and the target language with vocabulary V^T, and following Eq. (4), a probability that some word $w^S_i \in V^S$ will be generated by the cross-lingual topic z_k is given by:

$$P(w^S_i|z_k) = \phi_{k,i} = \frac{v^S_{k,w^S_i} + \beta}{\sum_{i^*=1}^{|V^S|} v^S_{k,w^S_{i^*}} + |V^S|\beta} \tag{8}$$

The same formula, but now derived from Eq. (5) is used for the per-topic word distributions (ψ) for the target language:

$$P(w^T_i|z_k) = \psi_{k,i} = \frac{v^T_{k,w^T_i} + \beta}{\sum_{i^*=1}^{|V^T|} v^T_{k,w^T_{i^*}} + |V^T|\beta} \tag{9}$$

In summary, these per-document topic distributions and per-topic word distributions are in fact mathematical realizations of the high-level intuitions and modeling premises clearly demonstrated in Fig. 1. For a better understanding of these core concepts, we refer the reader to study that figure again.

2.3.7. Inference or "What with New Documents?"

Since the model possesses a fully generative semantics, it is possible to train the model on one multilingual corpus (e.g., multilingual Wikipedia) and then infer it on some other, previously unseen corpus. Inferring a model on a new corpus means calculating per-document topic distributions for all the unseen documents in the unseen corpus based on the output of the trained model (i.e., we effectively learn the MuPTM-based representation of an unseen document, see Definition 4 and Eq.

(7)). Inference on the unseen documents is performed only one language at a time, e.g., if we train on English-Dutch Wikipedia, we can use the trained BiLDA model to learn document representations, that is, per-document topic distributions for Dutch news stories, and then separately for English news.

In short, we again randomly sample and then iteratively update topic assignments for each word position in an unseen document, but now start from the fixed v counters learned in training, and then cyclically update the probability distributions from which the topic assignments are sampled. Since the inference is performed monolingually, dependencies on the topic assignments from another language are removed from the updating formulas. Hence, similar to Eq. (4), the updating formula for the source language L_S is:

$$P(z_{ji}^S = z_k | \mathbf{z}_{\neg ji}^S, \mathbf{w}^S, \alpha, \beta) \propto \frac{n_{j,k,\neg i}^S + \alpha}{n_{j,\cdot,\neg i}^S + K\alpha} \cdot \frac{v_{k,w_{ji}^S}^S + \beta}{v_{k,\cdot}^S + |V^S|\beta} \tag{10}$$

Learning a multilingual topic model on one multilingual corpus and then inferring that model on previously unseen data constitutes the key concept of cross-lingual *knowledge transfer* by means of multilingual probabilistic topic models and that property is extensively utilized in various cross-lingual applications.

2.4. Evaluation of multilingual probabilistic topic models

A simple way of looking at the output quality of a topic model is by simply inspecting top words associated with a particular topic learned during training. We say that a latent topic is *semantically coherent* if it assigns high probability scores to words that are semantically related (Gliozzo, Pennacchiotti, & Pantel, 2007; Newman, Lau, Grieser, & Baldwin, 2010; Mimno, Wallach, Talley, Leenders, & McCallum, 2011; Stevens, Kegelmeyer, Andrzejewski, & Buttler, 2012; Aletras & Stevenson, 2013; Deveaud, SanJuan, & Bellot, 2013). It is much easier for humans to judge semantic coherence of cross-lingual topics and their alignment across languages when observing the actual words constituting a topic. These words provide a shallow qualitative representation of the latent topic space, and could be seen as direct and comprehensive word-based summaries of a large document collection. In other words, humans can get the first clue "what all this text is about in the first place".

The desirable property of semantic coherence comprises both a strong *intra-semantic coherence*, that is, words from the same vocabulary grouped together in the same topic are closely semantically related, as well as a strong *inter-semantic coherence*, that is, words across languages used to represent the same cross-lingual topic are also closely semantically related. Samples of cross-lingual topics extracted by BiLDA trained on aligned Wikipedia articles are provided in Table 1. We may consider this visual inspection of the top words associated with each topic as an initial *qualitative evaluation*, suitable for human judges.

Besides this shallow qualitative analysis relying on the top words, there are other, theoretically well-founded evaluation metrics for *quantitative* analysis and comparison of different models. In the literature, latent topic models are often evaluated by their perplexity, where the perplexity or "confusion" of a model is a measure of its ability to explain a collection $\mathcal{C}_u$ of unseen documents. The perplexity of a probabilistic topic model is expressed as follows:

$$perp(\mathcal{C}_u) = exp\left(-\frac{\sum_{d \in \mathcal{C}_u} \log\left(\prod_{w \in d} P(w)\right)}{\sum_{d \in \mathcal{C}_u} N^d}\right) \tag{11}$$

where N^d is defined as the number of words in a document d, $P(w)$ is word w's marginal probability according to a specific model, calculated as $\sum_k P(w \mid z_k, \Upsilon)$, where k ranges over all K topics in the model, and Υ is the set of the corpus independent parameters of the model. For BiLDA, the parameter set is $\Upsilon = \{\alpha, \beta, \phi, \psi, K\}$. A lower perplexity score means less confusion of the model in explaining the unseen data, and, theoretically, a better model. A good model with a low perplexity score should be well adapted to new documents and yield a good representation of those previously unseen documents. Since the perplexity measure defines the quality of a topic model independently of any application, it is considered an *intrinsic* or in vitro evaluation metric.

Another intrinsic evaluation metric for multilingual probabilistic topic models, named *cross-collection likelihood*, was proposed recently in (Zhang et al., 2010), but that measure also presupposes an existing bilingual dictionary as a critical resource. Additionally, a number of intrinsic quantitative evaluation methods (but for the monolingual settings) are proposed in (Wallach, Murray, Salakhutdinov, & Mimno, 2009). Other studies for the monolingual setting focused more on automatic evaluation of semantic coherence (e.g., Chang, Boyd-Graber, Gerrish, Wang, & Blei, 2009; Newman et al., 2010; Mimno et al., 2011). However, perplexity still remains the dominant quantitative in vitro evaluation method that is predominantly found in the literature.

Finally, the best way to evaluate multilingual probabilistic topic models is to test how well they perform in practice for different real-life tasks (e.g., document classification, information retrieval), that is, to carry out an *extrinsic ex vivo* evaluation. We later investigate whether there exists a mismatch between the intrinsic and extrinsic evaluation in information retrieval (see Section 6).

Table 1
Randomly selected examples of latent cross-lingual topics represented by top 10 words based on their counts after Gibbs sampling. Topics are obtained by BiLDA trained on Wikipedia for various language pairs: French–English (FR–EN), Dutch–English (NL–EN), Italian–English (IT–EN), and Spanish–English (ES–EN). For non-English words we have provided corresponding English translations. $K = 100$ for all models.

FR–EN Topic 17	NL–EN Topic 55	IT–EN Topic 73	ES–EN Topic 52
moteur (engine)	gebouw (building)	rete (network)	dinero (money)
voiture (vehicle)	eeuw (century)	chiave (key)	mercado (market)
automobile (car)	meter (meter)	protocollo (protocol)	precio (price)
vitesse (speed)	kasteel (castle)	server (server)	bienes (goods)
constructeur (constructor)	bisschop (bishop)	messaggio (message)	valor (value)
roue (wheel)	stad (city)	connessione (connection)	cantidad (amount)
vapeur (steam)	gebouwd (built)	client (client)	oferta (offer)
puissance (power)	theater (theater)	servizion (service)	pago (payment)
diesel (diesel)	museum (museum)	indirizzo (address)	impuesto (tax)
cylindre (cylinder)	tuin (garden)	sicurezza (security)	empresa (company)
engine	building	link	economic
car	court	network	price
vehicle	built	display	money
fuel	garden	calendar	market
speed	museum	client	capital
power	palace	key	tax
production	construction	server	goods
design	theater	protocol	interest
diesel	tower	address	demand
drive	castle	packet	inflation

2.5. A short overview of other multilingual probabilistic topic models

Similarly to LDA in the monolingual setting (for which we have already shown that it is only a special case of BiLDA operating with only one language), we believe that bilingual LDA can be considered the basic building block of this general framework of multilingual probabilistic topic modeling. It serves as a firm baseline for future advances in multilingual probabilistic topic modeling. Although MuPTM is a quite novel concept, several other models have emerged over the last years. All current state-of-the-art multilingual probabilistic topic models build upon the idea of standard monolingual pLSA and LDA and closely resemble the described BiLDA model, but they differ in the assumptions they make in their generative processes, and in knowledge that is presupposed before training (e.g., document alignments, prior word matchings or bilingual dictionaries). However, *they all share the same concepts defined in Section 2.1, that is, the sets of output distributions and the set of latent cross-lingual topics that has to be discovered in a multilingual text collection.*

The early approaches (see, e.g., Dumais et al., 1996; Carbonell et al., 1997) tried to mine topical structure from multilingual texts using an algebraic model, that is, Latent Semantic Analysis (LSA) and then use the discovered latent topical structure in cross-lingual information retrieval. Artificial "cross-lingual" documents were formed by concatenating aligned parallel documents in two different languages, and then LSA on a word-by-document matrix of these newly built documents was used to learn the lower dimensional document representation. Documents across languages are then compared in that lower-dimensional space.

Another line of work Zhao and Xing (2006, 2007) focused on building topic models suitable for word alignment and statistical machine translation operations. Again inspired by monolingual LDA, they have designed several variants of topic models that operate on parallel corpora aligned at sentence level. The topical structure at the level of aligned sentences or word pairs is used to re-estimate word translation probabilities and force alignments of words and phrases generated by the same topic.

However, the growth of the global network and increasing amounts of comparable theme-aligned texts have formed a need for constructing more generic models that are applicable to such large-volume, but less-structured text collections. Standard monolingual probabilistic topic models coming from the families of pLSA and LDA cannot capture and accurately represent the structure of such theme-aligned multilingual text data in a form of joint latent cross-lingual topics. That inability comes from the fact that topic models rely on word co-occurrence information to group similar words into a single topic. In case of multilingual corpora (e.g., Wikipedia articles in English and Dutch) two related words in different languages will seldom co-occur in a monolingual text, and therefore these models are unable to group such pairs of words into a single coherent topic (for examples see, e.g., Boyd-Graber & Blei, 2009; Jagarlamudi & Daumé III, 2010). In order to anticipate that issue, there have been some efforts that trained monolingual probabilistic topic models on concatenated document pairs in two languages (e.g., Dumais et al., 1996; Littman, Dumais, & Landauer, 1998; Carbonell et al., 1997; Chew, Bader, Kolda, & Abdelali, 2007; Xue, Dai, Yang, & Yu, 2008; Cimiano, Schultz, Sizov, Sorg, & Staab, 2009; Roth & Klakow, 2010), but such approaches also fail to build a shared latent cross-lingual topical space where the boundary between the topic representations with words in two languages is firmly established. In other words, when training on concatenated English and Spanish Wikipedia articles, the learned topics contain both English and Spanish words. However, we would like to learn latent

cross-lingual topics for which their representation in English is completely language-specific and differs from their representation in Spanish.

Recently, several novel models have been proposed that remove such deficiency. These models are trained on the individual documents in different languages and their output are joint latent cross-lingual topics in an aligned latent cross-lingual topical space. The utility of such new topic representations is clearly displayed further in this article (see Sections 3–6). These models require alignments at document level a priori before training, which is easily obtained for Wikipedia or news articles. These document alignments provide hard links between topic-aligned semantically similar documents across languages.

Recently, there has been a growing interest in multilingual topic modeling from unaligned text, again inspired by monolingual LDA. The MuTo model (Boyd-Graber & Blei, 2009) operates with *matchings* instead of words, where matchings consist of pairs of words that link words from the source vocabulary to words from the target vocabulary. These matchings are induced by the matching canonical correlation analysis (MCCA) (Haghighi et al., 2008; Daumé III & Jagarlamudi, 2011) which ties together words with similar meanings across languages, where similarity is based on different features. Matchings are induced based on pointwise mutual information (PMI) from parallel texts, machine-readable dictionaries and orthographic features captured by, for instance, edit distance. A stochastic expectation–maximization (EM) algorithm is used for training, and their evaluation has been performed on a parallel corpus. A similar idea of using matchings has been investigated in (Jagarlamudi & Daumé III, 2010). In their JointLDA model, they also observe each cross-lingual topic as a mixture over these matchings (or *word concepts*, as they name them), where the matchings are acquired directly from a machine-readable bilingual dictionary. JointLDA uses Gibbs sampling for training and it is trained on Wikipedia data. Although these two models claim that they have removed the need for document alignment and are fit to mine latent cross-lingual topics from unaligned multilingual text data, they have introduced bilingual dictionaries as a new critical resource. These machine-readable dictionaries have to be compiled from parallel data or hand-crafted, which is typically more expensive and time-consuming than obtaining alignments for Wikipedia or news data.

Another work that aims to extract latent cross-lingual topics from unaligned datasets is presented by Zhang et al. (2010). Their Probabilistic Cross-lingual Latent Semantic Analysis (PCLSA) extends the standard pLSA model (Hofmann, 1999b) by regularizing its likelihood function with soft constraints defined by an external machine-readable bilingual dictionary. They use the generalized expectation maximization (GEM) algorithm (Mei, Cai, Zhang, & Zhai, 2008) for training. Similar to MuTo and JointLDA, a bilingual dictionary is presupposed before training and it is a critical resource for PCLSA. The dictionary-based constraints are the key to bridge the gap between languages by pushing related words in different vocabularies to occur in the same cross-lingual topics. The same relationship between pLSA and LDA (Girolami & Kabán, 2003) in the monolingual setting is also reflected between their multilingual extensions, PCLSA and BiLDA.

In this article, we will present a subset of cross-lingual applications in which any multilingual probabilistic topic model may be utilized. In specific, we show the results obtained by BiLDA and provide an overview of its task performance. The goal of this article is however not to provide a direct comparison of different multilingual probabilistic topic models in various cross-lingual tasks, but to provide a comprehensive and didactic description of a general model-independent framework for building systems that rely on such multilingual probabilistic topic models and MuPTM-based representations of words and documents, and do not exploit any external expensive knowledge resource (e.g., parallel corpora, machine-readable dictionaries, extensive human annotations). Such data-driven unsupervised systems which exploit only internal evidence are essential for languages and language pairs with limited resources. We acknowledge that there exist numerous different techniques proposed for solving the presented tasks. However, our main focus is not to detect the best technique for each cross-lingual task, but to give a "cookbook" on how to exploit the latent cross-lingual topical knowledge as one source of evidence when dealing with these tasks in an unsupervised, language-independent and language pair independent manner.

In addition, the reader has to be aware that significant portions of Application I (Section 3), Application II (Section 4), and Application IV (Section 6) contain already published work (DeSmet & Moens, 2009; De Smet et al., 2011; Vulić, DeSmet, & Moens, 2013), but we have decided to retain the essence and have rewritten the previously published work in a systematic and didactic manner in order to better stress the general applicability of text representations by means of latent cross-lingual topics in a variety of cross-lingual tasks, and to provide some new insights from observing all the applications together. Moreover, a significant portion of Application III (Section 5) is novel and previously unpublished.

3. Application I: Cross-lingual event-centered news clustering

3.1. Task description

The first task we have chosen to present is cross-lingual *event-centered news clustering*. In general, event-centered news clustering may be considered an information retrieval task in which it is necessary to group news stories into coherent clusters, where each item (i.e., each news story) in one cluster should report on the same event. A special case is *cross-lingual event-centered news clustering* where one has to perform the clustering of news stories now written in different languages into groups of stories that describe the same event. Implicitly, that also defines a method for linking news

stories across languages. Due to the dynamic and ever-changing nature of news, one needs an unsupervised tool that can coherently capture such dynamics and provide a structured representation of news stories irrespective to their actual language. Such event-centered cross-lingual clustering of related news stories is highly desirable in systems for browsing, categorizing and summarizing large news archives given in multiple different languages (Chen et al., 2000; Pouliquen, Steinberger, Ignat, Käsper, & Temnikova, 2004; Evans, Klavans, & McKeown, 2004; Kabadjov, Atkinson, Steinberger, Steinberger, & der Goot, 2010). Cross-lingual event-centered news clustering may be observed as a special case of the cross-lingual document clustering task (e.g., Montalvo, Martínez-Unanue, Casillas, & Fresno, 2006; Wu & Lu, 2007; Tang, Xia, Zhang, Li, & Zheng, 2011), with an extra constraint which specifies that documents – news stories should be clustered together if and only if they cover the same event.

An *event* is defined as a well-specified happening at a certain moment in time (e.g., a single day or a short period) and space which deals with a certain set of news themes (e.g., a flood and a shortage of drinking water) and involves some named entities. Those named entities are actors of the events (e.g., persons or companies) or locations where the events occurred. Each news story typically reports on a single event, and since different sources can produce several stories on the same event in different languages, cross-lingual event-related clustering of these stories is required.

An event can be observed as a mixture of different themes, where some themes are dominant, while others are only marginally present. That phenomenon can be captured by probabilistic topic models – *per-document topic distributions* will be higher for topics closely related to the themes prominent in a news story. Two news stories s_i and s_j are considered similar and are most likely discussing the same event if their per-document topic distributions are similar, that is, if the values $P(z_k \mid s_i)$ and $P(z_k \mid s_j)$ are similar for all $z_k \in \mathcal{Z}$, where $\mathcal{Z}$ is the set of K latent cross-lingual topics (see Section 2.1). Note that by utilizing the language-independent set $\mathcal{Z}$ and per-document topic distributions as the news story representation, we are able to perform the cross-lingual event-centered news clustering, i.e., the clustering of stories written in different languages, regardless of the actual language in the story. Previous systems for cross-lingual news clustering either relied on a readily available machine translation system (Montalvo, Martínez-Unanue, Casillas, & Fresno, 2007; VanGael & Zhu, 2007) for feature or document translation, or on the knowledge of shared or cognate named entities (Montalvo et al., 2007). As already proven by DeSmet and Moens (2009), here we stress that the cross-lingual topical knowledge and the representations by means of per-document topic distributions also prove beneficial for this task.

3.2. Methodology

Additionally, news stories reporting on same events usually involve many shared named entities. We can therefore represent each news story with two disjunct sets of information or *aspects*: (1) words that describe generally applicable subjects captured by our cross-lingual topics and (2) shared named entities.[6] For instance, in case of any earthquake story it is highly likely to detect general terms such as *earthquake*, *damage*, and *casualties* and the only information that unambiguously separates one event from another lies in the named entities. When a topic model is trained on the documents' full texts, the named entities are part of the output MuPTM distributions, and this has the undesirable property that entities that were not apparent in the training set are unable to influence the topic inference of a new event. Therefore, we have decided to explicitly split each news story into two different aspect representations as: (1) a mixture of latent cross-lingual topics, that is, a MuPTM-based representation by means of per-document topic distributions and (2) a vector of shared named entities occurring in the news story. To achieve this: (1) a topic model is trained on news stories with named entities removed and (2) a score for a named entity in a news story is the probability of the entity obtained by smoothed maximum likelihood estimation. Here, since we develop a completely data-driven approach to cross-lingual news clustering, we do not use any translation resource to link named entities across languages, and follow the prior work from Montalvo et al. (2007), who also represented news stories across languages as vectors over shared and cognate named entities.

News stories are now represented by two probability distributions: (1) a probability distribution over cross-lingual topics (a per-document topic distribution) and (2) a probability distribution over shared named entities. In order to cluster the news stories based on the event they discuss, we need to choose a dissimilarity function. We use the symmetric Kullback–Leibler (KL) divergence of the O-dimensional probability distributions (where O equals K when dealing with topical representations, and some other constant when dealing with representations by means of shared named entities) x_i and x_j, defined as:

$$KL(x_i, x_j) = \frac{1}{2}\left(\sum_{o=1}^{O} x_i^o \log\left(\frac{x_i^o}{x_j^o}\right) + \sum_{o=1}^{O} x_j^o \log\left(\frac{x_j^o}{x_i^o}\right)\right) \tag{12}$$

In order to obtain a final dissimilarity function between two news stories, the dissimilarities for each of the two aspects (i.e., the topical aspect and the shared named entities aspect) obtained by Eq. (12) are combined by the maximum function, which proved to yield the best result in monolingual event-centered news clustering (DeSmet & Moens, 2013). The maximum function ensures that two stories are dissimilar when at least one of the aspects has dissimilar distributions, that is, if different events and locations are detected, we assume that we deal with different events. Vice versa, events that cover different

[6] A named entity is considered shared if it is present in the vocabularies of both languages, e.g., *Angela Merkel* will occur in the same spelling in English, Dutch, German of even Finnish news stories.

themes (represented by different cross-lingual topics) that happen at the same location or performed by the same actors are also treated as different events. The final dissimilarity function is as follows: $dis(s_i, s_j) = \max_a dis\left(A^a_{s_i}, A^a_{s_j}\right), a = 1 \rightarrow |A|$, where $|A|$ denotes the number of aspects a news story is split into (i.e., $|A| = 2$ in this case), and $A^a_{s_i}$ is the a-th aspect representation of story s_i. We have also tried splitting news stories into more fine-grained named entity representations based on their semantic class ($A > 2$, person–location–organization), but it has not improved the clustering performance.

The final story (dis) similarity $dis(s_i, s_j)$ is used in a clustering algorithm. We opt for a hierarchical agglomerative clustering with complete linkage (Voorhees, 1986). This algorithm does not require the number of clusters to be chosen in advance. Its adapting ability is a very important property in the dynamic news environment. The algorithm iteratively merges clusters until a certain criterion is reached. To create a natural, unsupervised stopping criterion, a fitness-condition on the clustering is used. The consequence is that the data dictates the optimal number of clusters. For each story s_i in the corpus, its fitness in cluster CL_i is calculated as the normalized difference between the distance of s_i to the second best cluster CL_j, and the average distance of s_i to the other stories in $CL_i : f(s_i) = \frac{h(s_i) - g(s_i)}{\max\{g(s_i), h(s_i)\}}$, where $g(s_i) = \frac{1}{|CL_i|-1} \sum_{s_j \in CL_i} dis(s_i, s_j)$ and $h(s_i) = \arg\min_{CL_j} \frac{1}{|CL_j|} \sum_{s_j \in CL_j} dis(s_i, s_j)$. If CL_i is a singleton cluster (containing only s_i), we set $f(s_i) = 0$. We search for the clustering that maximizes the average of f over all stories, over all possible stops in the hierarchy.

3.3. Experimental setup

3.3.1. Datasets

In all our applications, our *training corpora* are Wikipedia datasets for four language pairs, aligned through the cross-lingual links provided by the Wikipedia metadata.[7] There are 7612 aligned English–Dutch Wikipedia articles, 18,898 English–Italian, 18,911 English–French and 13,696 English–Spanish articles.[8] Since it is not true that every Wikipedia article appears in each language, the statistics of the datasets in terms of the number of tokens and vocabulary words vary over different language pairs. We don't use all training datasets in all applications. Additionally, since the aims and evaluation procedures for each application differ, we use different benchmarking datasets for testing in each application. Those datasets will be explicitly mentioned.

For the event-centered news clustering, we *train* on the English–Dutch Wikipedia articles. The Dutch articles are usually shorter and of a lesser quality (84 words tokens on average) than the English articles (986 word tokens on average). We have trained BiLDA with $K = 100$ topics, and hyper-parameters were set to $\alpha = 50/K$ and $\beta = 0.01$ (Griffiths & Steyvers, 2004; Steyvers & Griffiths, 2007). Unless noted otherwise, this is our default setting for the hyper-parameters in all further experiments. We have also trained standard LDA with $K = 100$ in the monolingual context for both languages to measure the difference between the monolingual and multilingual environment. In another experiment, we have trained BiLDA with $K = 20$, 50, 100, 200, 300 topics and compare with LDA trained on concatenated documents (MixLDA) trained on the same data with the same number of topics K.

Following VanGael and Zhu (2007), we have created our *test* set by compiling 18 events from Google news, forming 18 clusters of English and Dutch documents. In these 18 clusters, there are in total 50 English and 60 Dutch documents. A multilingual probabilistic topic model trained on Wikipedia is inferred on this test collection. Therefore, we perform knowledge transfer via the latent space of cross-lingual topics.

3.3.2. Evaluation metrics

Evaluation is done using the B-Cubed metric (Bagga & Baldwin, 1998). Let CL_i be the cluster that story s_i gets clustered in, and G_i its correct cluster from the ground truth. The B-Cubed metric then calculates $Precision = \frac{|CL_i \cap G_i|}{|CL_i|}$ and $Recall = \frac{|CL_i \cap G_i|}{|G_i|}$. The total precision and recall of the clustering are taken as the average of the precision and recall scores over all stories. The B-Cubed metric rewards a singleton clustering with a precision score of 100%, as no story is clustered together with an unrelated one, but recall is very low in case of singleton clustering. Therefore, we present our results in terms of the F_1 measure that balances between the two: $F_1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$.

3.4. Results and discussion

Table 2 shows the results of clustering the news stories according to the event discussed in the story. In the monolingual setting, the event-centered news clustering is quite accurate, and we can observe two interesting phenomena: (i) results for English are significantly higher than results for Dutch and (ii) there is a huge difference in favor of English

[7] All Wikipedia articles for all experiments are downloaded from Wikipedia *dumps:* http://dumps.wikimedia.org/.
[8] These datasets are available online: http://people.cs.kuleuven.be/~ivan.vulic/software/.

between the results that rely on per-document topic distributions only. Wikipedia articles in Dutch are in general shorter and contain much less content, which disturbs the correct topic estimation, and effectively leads to learning per-document topic distributions of a lesser quality for Dutch news stories. Moreover, we can see that representations relying on named entities score better for Dutch than for English. That difference is attributed to the fact that Dutch news stories are on average more than 10 times shorter and contain less variation in wording.

As expected, when transferring the problem to the multilingual setting, results decrease, but it has to be noted that the results are higher than the results obtained by a baseline system (56.5%(36)) similar to the one reported in (VanGael & Zhu, 2007), which uses a translation resource (*Google Translate*) to perform automatic machine translation of Dutch stories to English, and then the cosine metric to measure distances between the news stories. The F_1 score of 63.2% obtained by per-document topic distributions only is especially promising because it does not rely on any external knowledge, and it still relies on a training corpus of limited size. It is also very interesting to note the cross-lingual news clustering relying only on the topical representation outscores the representation which relies on shared named entities which was previously used in (Montalvo et al., 2007).

In another experiment, we have varied the number of topics K which also influences the dimension of the representation of the test news articles. We have made a comparison of the BiLDA-based topical representation against the representation obtained by training the standard monolingual LDA model on concatenated documents from aligned document pairs and then inferring this model on test news articles (*MixLDA*). The results and the comparison are provided in Table 3. We may observe that after a sufficient granularity of representation is used (given by the number of topics/dimensions K), the clustering results are comparable over different K-s (from 50 to 300). On the other hand, the 20-dimensional representation is too coarse-grained to produce comparable clustering results. The results also reveal that training BiLDA on separate documents in aligned document pairs yields better topical representations than concatenating these documents into one artificial document as with MixLDA. By training on separate documents and imposing the constraint that two documents discussing the same themes need to have the same topical representations by means of per-document topic distributions, we are able to effectively able to remove an imbalance which occurs when two documents are mixed together into a single large document. This imbalance might occur because a document in one language may be much longer and more informative than its counterpart, and hence "dominate" the sampling of topic assignments and the final representation of its counterpart. The main strength of BiLDA lies in enforcing the topics of the two coupled documents to be sampled from the same distribution θ (see Section 2.3 again).

In summary, the application of multilingual probabilistic topic models to cross-lingual event-centered news clustering provides first evidence that utilizing true multilingual probabilistic topic modeling (as with BiLDA) is more beneficial than pseudo-multilingual probabilistic modeling (as with MixLDA). These small experiments have shown (i) the validity of topical knowledge in cross-lingual event-centered news clustering and (ii) the better performance of an BiLDA-based clustering model over the clustering model which relies on topical representations obtained by standard LDA trained on artificially created concatenated documents. The work on cross-lingual event-centered news clustering follows the previous work in the monolingual event-centered news clustering (DeSmet & Moens, 2013), where it was already demonstrated that the topical representation of news stories by means of per-document topic distributions leads to improved clustering scores. Here, we have briefly described a similar cross-lingual framework where similarities between texts/news stories written in different languages can also be obtained by measuring similarities of cross-lingual topics' distributions over those texts.

The goal of this section was to give the reader the first insight on how to exploit language-independent representations of documents by means of latent cross-lingual topics. This cross-lingual news clustering framework provides ample room for future work and building more advanced cross-lingual news clustering models. For instance, one might try to build event-centered clusters separately for each language, and then design a method to align these monolingual clusters and merge them into larger cross-lingual clusters discussing the same events. As another line of future work it is also worth investigating whether the topical knowledge might prove beneficial when combined with models that rely on readily available bilingual dictionaries or machine translation tools.

Table 2
Results in terms of F_1 measure (B-Cubed) when using per-document topic distributions only, shared named entity distributions only and their combination for monolingual and cross-lingual event-centered news clustering. $K = 100$ for all models. The number of found clusters is given in parentheses.

	Monolingual (English)	Monolingual (Dutch)	Cross-lingual
Per-document topic distributions only	0.912 (17)	0.595 (12)	**0.632** (23)
Named entity distributions only	0.801 (20)	**0.877** (23)	0.567 (37)
Combined	**0.941** (23)	0.853 (23)	0.569 (48)

Bold values denote the best results for that column.

Table 3
A comparison between MixLDA (trained on concatenated documents from aligned document pairs) and BiLDA in the task of event-centered news clustering for different *K*-s. Per-document topic distributions are used to represent news articles. The number of found clusters is given in parentheses. *BEST* refers to the overall best result over all possible cuts in the whole dendrogram, *GTRUTH* refers to the result with the number of clusters fixed to the number of clusters from the ground truth, while *SIL* refers to the result with the automatically detected number of clusters using the stopping criterion from Section 3.2.

Topic model	MixLDA			BiLDA		
K	*BEST*	*GTRUTH*	*SIL*	*BEST*	*GTRUTH*	*SIL*
20	0.428 (25)	0.404 (18)	0.416 (29)	0.484 (23)	0.447 (18)	0.441 (35)
50	0.517 (30)	0.424 (18)	0.489 (47)	0.635 (30)	0.593 (18)	0.615 (32)
100	0.476 (38)	0.409 (18)	0.464 (33)	0.654 (27)	0.607 (18)	0.632 (23)
200	0.476 (42)	0.400 (18)	0.475 (38)	0.622 (24)	0.599 (18)	0.619 (27)
300	0.487 (35)	0.436 (18)	0.465 (47)	0.645 (25)	0.625 (18)	0.631 (30)

4. Application II: Cross-lingual document classification

4.1. Task description

Another task where representations of documents written in different languages as mixtures of cross-lingual topics (as represented by *per-document topic distributions*) prove to be useful is cross-lingual *document classification*. Cross-lingual document classification starts from a set of labeled documents in the source language L_S and unlabeled documents in the target language L_T. The objective is to learn a classification model from the labeled documents in the source language and then apply it to the classification of documents in the target language. The task obviously cannot be achieved by a method that only uses words from the labeled documents as features, since there is a minimal or no word overlap between the two languages. Hence, we have to find another solution. Here, we again deal with cross-lingual knowledge transfer, where the knowledge that is transferred across languages are *text categories*, that is, high-level labels that describe the content of a text.

Unlike in the previous application, where the similarity between two news stories or (more generally) documents has been established directly according to their respective per-document topic distributions, here we can observe each document as a data instance and use the probabilities $P(z_k \mid d_j)$ from their per-document topic distributions as classification features. Again, by having the language independent set $\mathcal{Z}$ of K cross-lingual topics, we can operate in the same shared cross-lingual feature space regardless of the actual languages in which documents were written.

4.2. Methodology

A multilingual probabilistic topic model is first learned on a general multilingual corpus (e.g., Wikipedia). Then, given a cross-lingual document classification task, that is, a labeled document collection $\mathcal{L}^S$ in L_S, and an unlabeled document collection $\mathcal{U}^T$ in L_T, the learned cross-lingual topics are used to infer their per-document topic distributions on each document in $\mathcal{L}^S$ and $\mathcal{U}^T$. The methodology was proposed and described by De Smet et al. (2011).

Each document from $\mathcal{L}^S$ and $\mathcal{U}^T$ is then taken as a data instance in the classification model, where its features are the inferred per-document topic distributions. The exact value of each classification feature of an instance, e.g., of document d_i^S is exactly the probability $P(z_k \mid d_i^S)$, for all $k = 1, \ldots, K$. The same is valid for some target document d_j^T. Since the documents in both languages are represented by the language-independent features (the distributions of cross-lingual topics over the documents), supervision in only one language is needed and labels from the documents in the source language are then propagated to the unlabeled documents in the target language according to the learned classification, that is, we again perform cross-lingual knowledge transfer. For the classification model, one can choose any existing classifier such as Naive Bayes, Perceptron, Maximum Entropy or Support Vector Machines (SVM). This choice is beyond the research reported in this article, and we present the classification results obtained by SVM on the aforementioned feature vectors comprising inferred per-document topic distributions as dimensions.[9]

4.3. Experimental setup

4.3.1. Datasets

Training has been conducted on English–Italian, English–French and English–Spanish Wikipedia (see Section 3.3). English is considered the source language L_S in all classification experiments.

[9] For SVM, we employ the SVM-Light package (Joachims, 1999, chap. 11): http://svmlight.joachims.org/ with default parameter settings.

Table 4
Average perplexity and F_1 scores for the cross-lingual classification with features based on per-document topic distributions of a multilingual probabilistic model (BiLDA) over all categories for each language pair.

Language pair	*Perplexity (EN)*	*Perplexity (L_T)*	*Book* (%)	*Film* (%)	*Prog* (%)	*Sport* (%)	*Video* (%)
English–Spanish	459.2	661.3	79.6	59.0	75.1	59.3	59.5
English–French	435.8	501.3	52.7	64.5	87.0	85.0	35.7
English–Italian	393.9	794.0	52.5	46.8	84.3	79.5	76.9

A *test* dataset, also obtained from Wikipedia, but different from the training corpus, is used for classification. It is collected by exploiting the category labels of Wikipedia. Specifically, five high level categories were first collected: *books ("book")*, *films ("film")*, *programming languages ("prog")*, *sports ("sport")* and *video games ("video")*, and then for each category up to 1000 articles annotated with the category label were extracted.[10] Since Wikipedia variants differ in the number of articles, sometimes fewer than 1000 articles were collected for Italian, French and Spanish. We have extracted 1000 articles in the classification dataset for each category, except for the following: (i) there are 263 Spanish, 592 French and 290 Italian articles respectively, labeled with *"prog"* and (ii) there are 764 Italian articles labeled with *"video"*. The BiLDA model has been trained on the three bilingual corpora, ranging the number of topics from 10 to 200 in steps of 20. All results, except when noted, are obtained as an average over classification results with each model.

4.3.2. Evaluation metrics

Again, we calculate the precision and recall scores and then combine them into the balanced F_1 score. Precision is now the number of correctly labeled documents divided by the total number of documents labeled that way, and recall is the number of correctly labeled documents divided by the actual number of documents with that label as found in the ground truth.

4.4. Results and discussion

Table 4 displays the performance of the models in terms of their average F_1 scores for each of our chosen classes and for each language pair. Average perplexity scores after inferring the models on classification datasets are also presented in Table 4. We observe that the results vary over language pairs and over categories for each language pair. With respect to the fact that the approach is completely unsupervised and without any additional resource (e.g., a machine translation system or a bilingual dictionary), the obtained results are reasonably high for all language pairs. Perplexity scores reveal that the English side of our training corpora is of a higher quality for all language pairs. Additional inspection of perplexity also exposes the mismatch between those scores and the actual classification results across categories for different languages. A lower perplexity of a model does not necessarily imply a better classification score (e.g., see the results over the last three categories for English–Italian and English–Spanish).

4.4.1. A comparison with a MixLDA-based classification model

As already mentioned in Section 3, instead of utilizing a multilingual topic model, a common approach when dealing with cross-lingual tasks is to combine each bilingual pair of documents from training into a single document, mixing the words from both languages, and then to employ a monolingual topic model (e.g., LDA) to extract the shared latent topic space (MixLDA). The monolingual topic model may then again inferred on the test dataset. The two languages again share only one common topic space, and the learned per-document topic distributions are again used as the features for SVM to learn the classification model. In order to test the utility of MuPTM, we have compared the classification results of our BiLDA-based classification models with the classification results which rely on LDA trained on concatenated document pairs. Based on the results from Fig. 6, one may conclude that the cross-lingual knowledge transfer by means of BiLDA leads to much better classification results than the one relying on LDA. The difference in results in favor of the BiLDA-based classification models is reported for all three language pairs and the majority of classification categories. The main causes of such a difference in results between BiLDA and MixLDA have already been discussed in Section 3.4.

4.4.2. Effect of K on the classification results

An additional experimental study tests how the classification results change related to the number of topics K set before training. Fig. 5 shows the F_1 scores of the classification results depending on K. Again, similar to the finding reported for the task of cross-lingual event-centered news clustering (see Section 3.4 and Table 3), we can observe that in general the classification results for all languages and almost all categories are not very sensitive to the number of topics, except when K is very small. Once K is set large enough to produce a finer-grained representation of documents based on cross-lingual topics, per-document topic distributions as features will be distinctive enough to group this document with similar documents in a hyperplane and produce the correct label for the document.

[10] Note that a Wikipedia article may have more than one label, and these labels can be very specific. In order to obtain a workable set with enough documents for classification, we have opted for the broad categories. In cases when we have collected more than 1000 articles for a category, we have randomly sampled 1000 articles to be included in the final test set.

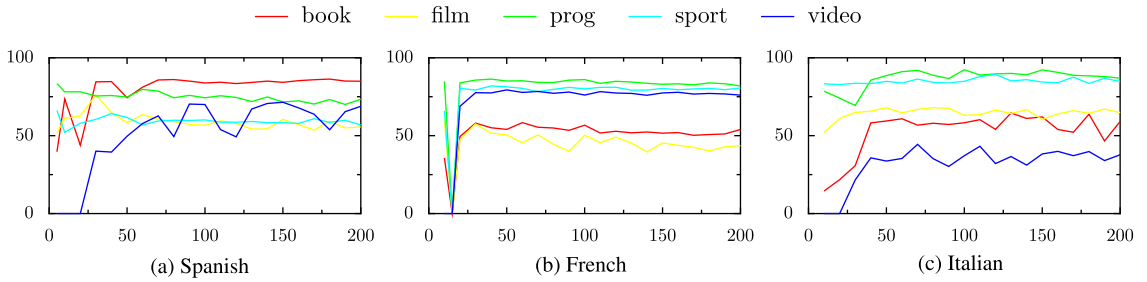


Fig. 5. Performance of the BiLDA model for several values of *K*, for all language combinations and categories. *x*-axes refer to *K*.

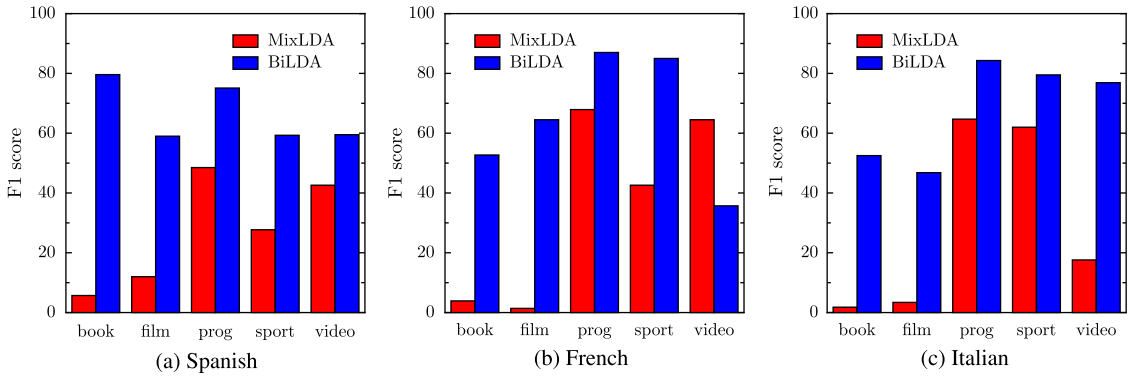


Fig. 6. Comparison of performances of the BiLDA model and the MixLDA model over all 5 categories for all 3 classification setups: (a) English–Spanish, (b) English–French, and (c) English–Italian.

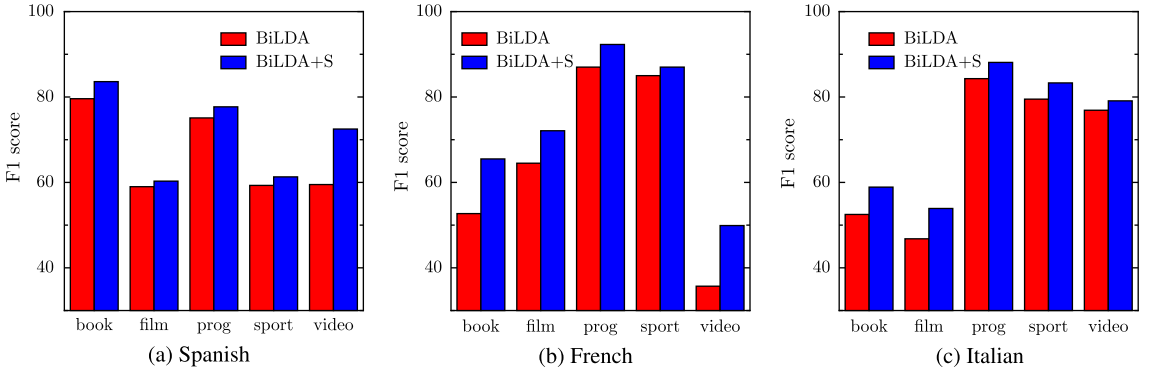


Fig. 7. Performance of the BiLDA model over all 5 categories for all 3 classification setups without (BiLDA) and with topic smoothing (BiLDA + S): (a) English–Spanish, (b) English–French, and (c) English–Italian.

The described classification framework also allows us to combine per-document topic distributions from different models as features for classification. Therefore, instead of averaging over scores obtained by different topic models, we could use all their per-document topic distributions and enrich the feature set. Per-document topic distributions from all 10 trained BiLDA models from the previous experiment are used as features. As the comparison of classification scores displayed in Fig. 7 reveals, we can observe a significant improvement in the classification results (+8.5% on average) by employing this procedure called *topic smoothing*.

In summary, we have shown how to utilize per-document topic distributions of a multilingual probabilistic topic model in the cross-lingual document classification across languages where no labeled instances are given a priori. In this case we deal with a low number of categories, and topic models with smaller values for *K* (e.g., 50 or higher) already provide a representation of documents that is fine-grained enough to help us decide which category label to assign to documents.

5. Application III: Cross-lingual semantic similarity

5.1. Task description

So far, we have only used the per-document topic distributions in the previous applications. But, what about per-topic word distributions that provide language-specific representations of each latent cross-lingual topic in each language? Since we have already detected that there ideally should exist a strong intra semantic and inter semantic coherence within cross-lingual topics, we could use the *per-topic word distributions* for mining semantically similar words across languages, that is, providing a list of target language words that are semantically similar to a given source language word w_1^S. For each source word w_1^S, we can build a *ranked list* $RL(w_1^S)$ which consists of all words $w_j^T \in V^T$ ranked according to their respective similarity scores $sim(w_1^S, w_j^T)$. In the similar fashion, we can build a ranked list $RL(w_2^T)$, for each target word w_2^T. We call the top M best scoring target words w_j^T for some source word w_1^S its M *nearest neighbors*. The ranked list for w_1^S comprising only its M nearest neighbors is called *pruned ranked list* (i.e., the ranked list is effectively pruned at position M), and we denote it as $RL_M(w_1^S)$. In the monolingual setting, the nearest neighbor in the list is usually a direct synonym of the word. The single nearest cross-lingual neighbor for w_1^S is called its *translation candidate*. By retaining only the first target candidate from the list, we can build a one-to-one word bilingual lexicon. Additionally, by retaining the list of a few first word candidates in the target language along with their scores, we actually build a probabilistic lexicon that could be used, for instance, in query expansion techniques for cross-lingual information retrieval (Vulić et al., 2013). The question is, however, how to obtain those lists of semantically similar words across languages by exploiting the knowledge present in per-topic word distributions of a multilingual topic model. The following section provides the answer. It introduces the complete framework for modeling cross-lingual semantic similarity by means of MuPTM.

5.2. Methodology

A straightforward approach to building one-to-one bilingual lexicons based on the knowledge from per-topic word distributions was presented by Mimno et al. (2009). For each topic z_k, they select a small number M' of the most probable words in the source and the target language representations of that topic, according to the respective per-topic word distributions. Following that, they add the Cartesian product of these two sets to a final set of candidate translation pairs $\mathcal{T}$. In a similar fashion, Boyd-Graber and Blei (2009) use learned matchings as one-to-one bilingual word lexicon entries (see Section 2.5). These approaches suffer from several issues: (1) by observing only the top candidates for each topic, the size of the set of candidate translation pairs is limited to only the best scoring words, and the size heavily depends on the chosen number of topics, (2) expanding M' increases the size of $\mathcal{T}$, but it introduces many incorrect one-to-one translation pairs, and (3) they do not exploit the fact that each word can be important for more than only one latent cross-lingual topic, i.e., they do not exploit the actual probability distributions of words over topics to mine semantically similar words across languages.

In order to fully exploit those per-topic word distributions, we make a connection between latent cross-lingual topics and an idea known as the *distributional hypothesis* (Harris, 1954). It states that words with similar meanings are likely to appear in similar contexts across languages. Here, cross-lingual topics provide a sound mathematical representation of this context and may be regarded as *context features*. In other words, we consider that word w_1^S in L_S and word w_2^T in L_T are semantically similar if they are often present and almost equally important in the same latent cross-lingual topics, and not observed or marginally present in other latent cross-lingual topics, that is, the word w_2^T is semantically similar to w_1^S if the distribution of w_2^T over latent cross-lingual topics (extracted from per-topic word distributions for L_T) is similar to the probability distribution of w_1^S over the same set of latent topics (extracted from per-topic word distributions for L_S). In this article, we present and evaluate a series of models of semantic similarity which rely on this essential hypothesis and significantly extend our earlier work (Vulić et al., 2011a).

5.2.1. Conditional topic distributions

After training, a multilingual topic model outputs per-topic word distributions with probability scores $P(w_1^S \mid z_k)$ and $P(w_2^T \mid z_k)$, for each $w_1^S \in V^S, w_2^T \in V^T$ and $z_k \in \mathcal{Z}$. It holds that $\sum_{k=1}^{K} P(w_1^S \mid z_k) = 1$ and $\sum_{k=1}^{K} P(w_2^T \mid z_k) = 1$, since each language has its own language-specific distribution over vocabulary words (see Sections 2.1 and 2.2).

In order to quantify the similarity between two words $w_1^S \in V^S$ and $w_2^T \in V^T$, we may employ the same trick that has been used for obtaining the degree of similarity between two documents in the monolingual setting (Steyvers & Griffiths, 2007) and the cross-lingual setting (Ni et al., 2011) (see also Eq. (7) in Section 2.3). Since each document d_j is represented as a mixture of topics by means of per-document topic distributions given by the probability scores $P(z_k|d_j)$, the similarity between two documents can be established by measuring the similarity of these probability distributions. When dealing with the similarity of words w_1^S and w_2^T, we need to measure the similarity of their respective *conditional topic distributions*, given by the probability scores $P(z_k|w_1^S)$ and $P(z_k|w_2^T)$, for each $z_k \in \mathcal{Z}$. Each word, regardless of its actual language, is then represented by its K-dimensional vector $vec(w_i)$ (where features are latent cross-lingual topics) as a point in a K-dimensional latent semantic space. In other words, each word, irrespective to the language, is represented as a distribution over the K latent topics/concepts, where the K-dimensional vector representation of $w_1^S \in V^S$ (similar for $w_2^T \in V^T$) is:

$$vec(w_1^S) = [P(z_1|w_1^S), \ldots, P(z_k|w_1^S), \ldots, P(z_K|w_1^S)] \tag{13}$$

Using Bayes' rule, we can compute these probability scores:

$$P(z_k|w_1^S) = \frac{P(w_1^S|z_k)P(z_k)}{P(w_1^S)} = \frac{P(w_1^S|z_k)P(z_k)}{\sum_{k^*=1}^{K} P(w_1^S|z_{k^*})P(z_{k^*})} \tag{14}$$

where $P(w_1^S|z_k)$ is known directly from the per-topic word distributions. $P(z_k)$ is the prior topic distribution which can be used to assign higher a priori importance to some cross-lingual topics from the set $\mathcal{Z}$ (Jagarlamudi et al., 2012). However, in a typical setting where we do not possess any prior knowledge about the corpus and the likelihood of finding specific latent topics in that corpus, we assume the uniform prior over latent cross-lingual concepts/topics (Griffiths et al., 2007) (i.e., that all topics/concepts are equally likely before we observe any training data). The probability scores $P(z_k|w_1^S)$ from Eq. (14) for conditional topic distributions in that case may be further simplified:

$$P(z_k|w_1^S) = \frac{P(w_1^S|z_k)}{\sum_{k^*=1}^{K} P(w_1^S|z_{k^*})} = \frac{\phi_{k,1}^S}{\sum_{k^*=1}^{K} \phi_{k^*,1}^S} = \frac{\phi_{k,1}^S}{Norm_{\phi_{\cdot,1}^S}} \tag{15}$$

where we denote the normalization factor $\sum_{k^*=1}^{K} \phi_{l,1}^S$ as $Norm_{\phi_{\cdot,1}^S}$. A similar derivation follows for each $w_2^T \in V^T$ and the similarity between two words may then be computed as the similarity between their conditional topic distributions as given by Eq. (14) or Eq. (15). We will use this property extensively in our models of cross-lingual similarity.

5.2.2. KL model and JS model

Now, once the conditional topic distributions are computed, any similarity metric may be used as a similarity function (SF) between word vectors to quantify the degree of similarity between the representations of words by means of these conditional topic distributions. We present and evaluate a series of models which employ the most popular SF-s reported in the relevant literature. Each SF in fact gives rise to a new model of cross-lingual semantic similarity!

The first model relies on the Kullback–Leibler (KL) divergence which is a common measure of (dis) similarity between two probability distributions (Lin, 1991). The KL divergence of conditional topic distributions for two words w_1^S and w_2^T is an asymmetric measure computed as follows (our *KL model*):

$$sim(w_1^S, w_2^T) = KL(vec(w_1^S), vec(w_2^T)) = \sum_{k=1}^{K} P(z_k|w_1^S) \log \frac{P(z_k|w_1^S)}{P(z_k|w_2^T)} = \sum_{k=1}^{K} \frac{\phi_{k,1}^S}{Norm_{\phi_{\cdot,1}^S}} \log \frac{\phi_{k,1}^S \cdot Norm_{\psi_{\cdot,2}^T}}{\psi_{k,2}^T \cdot Norm_{\phi_{\cdot,1}^S}} \tag{16}$$

The Jensen–Shannon (JS) divergence (Lin, 1991; Dagan, Pereira, & Lee, 1994; Dagan, Lee, & Pereira, 1997) is a symmetric (dis) similarity measure closely related to the KL divergence, defined as the average of the KL divergence of each of two distributions to their average distribution. The dissimilarity of conditional topic distributions is computed as follows (our *JS model*):

$$sim(w_1^S, w_2^T) = JS(vec(w_1^S), vec(w_2^T)) = \frac{1}{2}\left(\sum_{k=1}^{K} P(z_k|w_1^S) \log \frac{P(z_k|w_1^S)}{P_{avg}(z_k|w_1^S)} + \sum_{k=1}^{K} P(z_k|w_2^T) \log \frac{P(z_k|w_2^T)}{P_{avg}(z_k|w_2^T)}\right) \tag{17}$$

where $P_{avg}(z_k \mid w_1^S) = P_{avg}(z_k \mid w_2^T) = \frac{P(z_k|w_1^S)+P(z_k|w_2^T)}{2}$, for each $z_k \in \mathcal{Z}$. Both KL and JS output non-negative scores, where a lower output score implies a lower divergence between two words, and therefore, a closer semantic similarity. Both KL and JS are defined only if they deal with real probability distributions, that is, if probability scores sum up to 1, Additionally, $P(z_k \mid w_1) > 0$ and $P(z_k \mid w_2) > 0$ have to hold for each z_k. Conditional topic distributions satisfy all these conditions. A ranked list $RL(w_1^S)$ may be obtained by sorting words $w_2^T \in V^T$ in ascending order based on their respective similarity/divergence scores computed by Eq. (16) (KL) or Eq. (17) (JS).

5.2.3. TCos model

The next distance measure is the cosine similarity which is one of the most popular choices for SF in distributional semantics (Fung & Yee, 1998; Bullinaria & Levy, 2007; Turney & Pantel, 2010). Cosine similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them. The cosine similarity of conditional topic distributions (our *TCos model*) is computed as follows:

$$sim(w_1^S, w_2^T) = TCos(vec(w_1^S), vec(w_2^T)) = \frac{\sum_{k=1}^{K} P(z_k|w_1^S)P(z_k|w_2^T)}{\sqrt{\sum_{k=1}^{K} P(z_k|w_1^S)^2} \cdot \sqrt{\sum_{k=1}^{K} P(z_k|w_2^T)^2}} \tag{18}$$

The higher the score in the range $[0, 1]$ (all dimensions of our vectors are positive numbers and therefore the lower similarity bound is 0 instead of -1), the higher the similarity between two words. A ranked list $RL(w_1^S)$ may be obtained by sorting words $w_2^T \in V^T$ in descending order based on their respective scores computed by Eq. (18).

5.2.4. BC model

Another similarity measure is the Bhattacharyya coefficient (BC) (Bhattacharyya, 1943; Kazama, Saeger, Kuroda, Murata, & Torisawa, 2010). The similarity of two words based on this similarity measure is defined as follows (our *BC model*):

$$sim(w_1^S, w_2^T) = BC(vec(w_1^S), vec(w_2^T)) = \sum_{k=1}^{K} \sqrt{P(z_k|w_1^S)P(z_k|w_2^T)} \tag{19}$$

In general, it measures the amount of overlap between two statistical samples which, unlike for KL and JS, do not have to be described by proper probability distributions. A higher score again implies a stronger semantic similarity between two words. The utility of the BC measure has not been investigated well in the literature on distributional semantics. Our experiments will reveal its potential in identifying semantically similar words across languages.

5.2.5. Cue model

Another way of utilizing per-topic word distributions is to directly model the probability $P(w_2^T|w_1^S)$, where semantically most similar target words should have the highest probability to be generated as a response to a cue source word. The probability $P(w_2^T \mid w_1^S)$ emphasizes the (cross-lingual) associative relation between words (Griffiths et al., 2007). Again, under the assumption of uniform topic prior, we can decompose the probability $P(w_2^T \mid w_1^S)$ as follows (our *Cue model*):

$$sim(w_1^S, w_2^T) = Cue(vec(w_1^S), vec(w_2^T)) = \sum_{k=1}^{K} P(w_2^T|z_k)P(z_k|w_1^S) \tag{20}$$

The probability value directly provides the degree of semantic similarity and a ranked list $RL(w_1^S)$ may be obtained by sorting words $w_2^T \in V^T$ in descending order based on their respective probability scores computed by Eq. (20).

5.2.6. TI model

The next model moves away from utilizing conditional topic distributions explicitly and aims to exploit latent cross-lingual topics in a different way. It builds a context vector $vec(w_1^S)$ as follows:

$$vec(w_1^S) = [\text{TTF} - \text{ITF}(w_1^S, z_1), \ldots, \text{TTF} - \text{ITF}(w_1^S, z_K)] \tag{21}$$

TTF-ITF (*term-topic frequency – inverse topic frequency*) is a novel weighting scheme which is analogous to and directly inspired by the TF-IDF (*term frequency – inverse document frequency*) weighting scheme in IR (SparckJones, 1973; Salton, Wong, & Yang, 1975; Manning & Schütze, 1999; Manning, Raghavan, & Schütze, 2008). Instead of conditional topic probability scores $P(z_k \mid w_1^S)$ the context features $sc_1(c_k)$ are now TTF-ITF scores. In our TTF-ITF weighting scheme, the $\text{TTF}(w_1^S, z_k)$ part of the complete score $\text{TTF-ITF}(w_1^S, z_k)$ measures importance of w_1^S for the particular topic z_k. It denotes the number of assignments of the latent cross-lingual topic z_k to the occurrences of w_1^S in the whole training corpus (i.e., that number is exactly the Gibbs count variable v^S_{k,w_1^S} which is one of the variables utilized to obtain the output per-topic word distributions in Section 2.3). The $\text{ITF}(w_1^S)$ score measures global importance of w_1^S across all latent cross-lingual topics. Words that are prominent for only a small subset of topics from $\mathcal{Z}$ are given higher importance for these topics as such words are in general more descriptive for these specific topics than high-frequency words that occur frequently over all topics. The inverse topic frequency for the word w_1^S across the set of cross-lingual topics is computed as $ITF(w_1^S) = \log \frac{K}{1+|z_k: v^S_{k,w_1^S} > 0|}$. The final $\text{TTF-ITF}(w_1^S, z_k)$ score for the source language word w_1^S and the topic z_k is then calculated as $\text{TTF-ITF}(w_1^S, z_k) = \text{TTF}(w_1^S, z_k) \cdot \text{ITF}(w_1^S)$. Once the same vector representation for $w_2^T \in V^T$ has been obtained, the similarity between words w_1^S and w_2^T may again be computed by means of their K-dimensional vector representations from Eq. (21) using the cosine similarity (or any other SF) as follows (our *TI* model):

$$sim(w_1^S, w_2^T) = TI(vec(w_1^S), vec(w_2^T)) = \frac{\sum_{k=1}^{K} \text{TTF} - \text{ITF}(w_1^S, z_k) \cdot \text{TTF} - \text{ITF}(w_2^T, z_k)}{\sqrt{\sum_{k=1}^{K} \text{TTF} - \text{ITF}(w_1^S, z_k)} \cdot \sqrt{\sum_{k=1}^{K} \text{TTF} - \text{ITF}(w_1^S, z_k)}} \tag{22}$$

5.2.7. TI + Cue model

In the original paper (Vulić et al., 2011a) we have discussed that the Cue model and the TI model interpret and exploit the shared set of latent cross-lingual topics in different ways. Therefore, by combining the two models and capturing different evidences of similarity, we should be able to boost the quality of obtained ranked lists. As in (Vulić et al., 2011a), we present a linear combination of the two models (with γ as the interpolation parameter), where the overall score is computed as follows (our *TI + Cue model*):

$$sim(w_1^S, w_2^T) = \gamma sim_{TI}(w_1^S, w_2^T) + (1 - \gamma) sim_{Cue}(w_1^S, w_2^T) \tag{23}$$

Following Vulić et al. (2011a), the parameter γ is set to 0.1.

5.2.8. Topic pruning

All these models of similarity have a straightforward theoretical explanation – they assign high similarity scores for pairs of words that assign similar importance to the same latent cross-lingual topics/concepts, that is, the same axes in the shared semantic space. In the core of all models of similarity are point-wise additive formulas, i.e., the models perform calculations over each cross-lingual topic $z_k \in \mathcal{Z}$ and each calculation contributes to the overall sum. However, each word is usually important for only a limited number of topics/concepts. For models that make use of conditional topic distributions it means that words exhibit high conditional topic probability scores for only a small subset of cross-lingual topics. In a typical setting

for mining semantically similar words using latent topic models in both monolingual (Griffiths et al., 2007; Dinu & Lapata, 2010) and cross-lingual settings (Vulić et al., 2011a), the best results are obtained with the number of topics set to a few thousands (≈ 2000). Therefore, the procedure of *topic pruning* might lead to improvements in terms of overall quality and the speed of calculation.

For the sake of simplicity, the description is valid for models that utilize conditional topic distributions with scores $P(z_k|w_1^S)$. Since $P(z_k|w_1^S) > 0$ for each $z_k \in \mathcal{Z}$, a lot of probability mass is assigned to latent topics that are not relevant to the given word. Reducing the dimensionality of the semantic representation a posteriori to only a smaller number of the most informative semantic axes in the latent space should decrease the effects of that statistical noise, and even more firmly emphasize the latent correlation among words. Therefore, given two words w_1^S and w_2^T, we prune the representation of these words in the shared latent semantic space spanned by cross-lingual topics as summarized in Algorithm 3.

Algorithm 3. Topic pruning

1: **Obtain** a subset $\mathcal{Z}_{K'} \subseteq \mathcal{Z}$ of $K' \leqslant K$ latent cross-lingual topics with the highest values $P(z_k \mid w_1^S)$.
Calculating the similarity score $sim(w_1^S, w_2^T)$ may be interpreted as: "Given a word w_1^S detect how similar another word w_2^T is to the word w_1^S." Therefore, when calculating $sim(w_1^S, w_2^T)$, even when dealing with symmetric similarity functions such as JS or BC, we always consider only the ranking of scores $P(z_k \mid w_1^S)$ for pruning. The subset $\mathcal{Z}_{K'}$ is then $\{z'_1, \ldots, z'_{K'}\}$.
2: **Retain** conditional topic probability scores $P(z'_k \mid w_1^S)$ for the word w_1^S only over topics $z'_k \in \mathcal{Z}_{K'}$.
3: **Retain** conditional topic probability scores $P(z'_k \mid w_2^T)$ for w_2^T over the same cross-lingual topics $z'_k \in \mathcal{Z}_{K'}$.

Both w_1^S and w_2^T are now represented by their K'-dimensional context vectors: for w_1^S the pruned vector is $\overrightarrow{vec}(w_1^S) = [P(z'_1 \mid w_1^S), \ldots, P(z'_{K'} \mid w_1^S)]$, where context features are now the semantically most relevant cross-lingual topics for the word w_1^S. We may again employ any SF (e.g., JS, BC, TCos) on these reduced representations, that is, pruned feature vectors with the adjusted conditional topic probability scores to calculate similarity.

5.3. Experimental setup

5.3.1. Datasets

We *train* on English-Italian Wikipedia. Following a common practice in relevant related work (Koehn & Knight, 2002; Haghighi et al., 2008; Boyd-Graber & Blei, 2009; Prochasson & Fung, 2011), we focus only on translation of nouns. Again, following related work, we use TreeTagger (Schmid, 1994) for POS-tagging and lemmatization of the corpora, and then retain only nouns that occur at least 5 times in the corpus. We record the lemmatized form when available, and the original form otherwise. Therefore, our final vocabularies consist of 7160 Italian nouns and 9166 English nouns.

We designed a set of ground truth one-to-one translation pairs to measure the ability of these models of similarity to extract one-to-one word translations from the data (the bilingual lexicon extraction task). We randomly sampled a set of 1000 Italian nouns from our Wikipedia corpora (i.e., we conduct our experiments in the Italian-to-English direction) which are to be regarded as our *test words*. Following that, we used the *Google Translate* tool plus an additional annotator to translate those words to English. The annotator manually revised the lists and retained only words that have their corresponding translation in the English vocabulary. Additionally, only one possible translation was annotated as correct. When more than one translation is possible (e.g., when dealing with polysemous words), the annotator marked as correct the translation that occurs more frequently in the English part of our Wikipedia data.[11]

For the multilingual topic model training, we have varied the number of topics K for BiLDA from 200 to 3500 with steps of 300 or 400 to measure the influence of the parameter K on the overall scores, that is, to test how the granularity of the shared topical space influences the quality of our models of similarity.

5.3.2. Evaluation metrics

The first, stricter evaluation metric calculates Acc_1 scores (Gaussier, Renders, Matveeva, Goutte, & Déjean, 2004; Tamura, Watanabe, & Sumita, 2012), that is, the percentage of words where the first word from the ranked list is actually the correct translation according to the ground truth). This metric directly measures the quality of one-to-one non-probabilistic bilingual lexicons. In a more lenient evaluation setting, we also measure Acc_{10} scores (i.e., the correct translation should appear among top 10 best scoring words in the ranked list for each word), and we also employ the *mean reciprocal rank* (*MRR*) (Voorhees, 1999). Here, we retain the entire ranked list, and *MRR* rewards if the correct translation is found higher in the list.

[11] The test set is available online: http://people.cs.kuleuven.be/~ivan.vulic/software/.

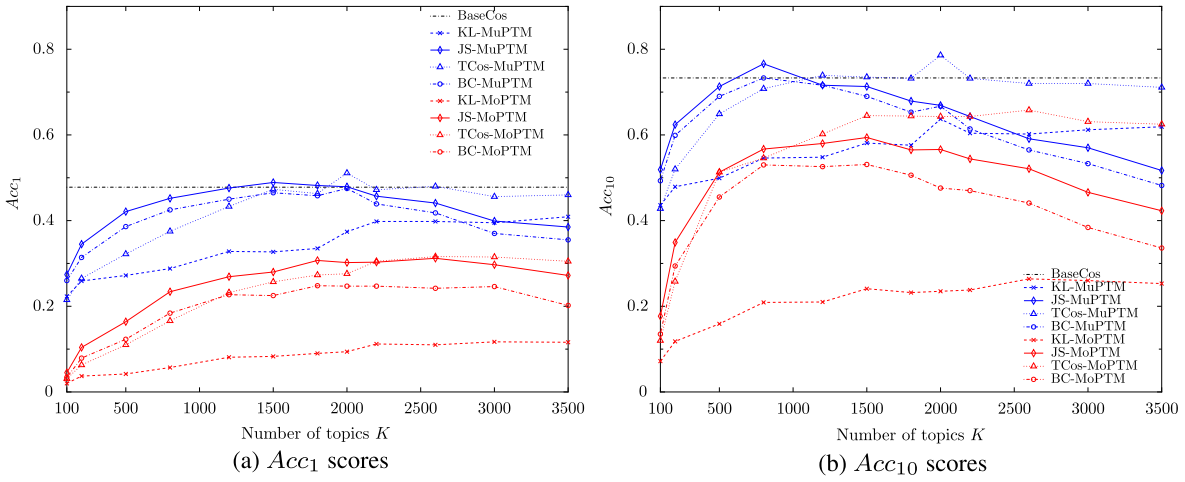


Fig. 8. Acc_1 and Acc_{10} scores for KL-MuPTM, JS-MuPTM, TCos-MuPTM, BC-MuPTM, and comparisons with baseline models: KL-MoPTM, JS-MoPTM, TCos-MoPTM, BC-MoPTM, and BaseCos.

5.3.3. Models for comparison

We evaluate all our MuPTM-based models of similarity in the BLE task (their codes are *-*MuPTM*, e.g., KL-MuPTM or JS-MuPTM). We compare them against baseline models which also exploit document alignments when mining semantically similar words and translation candidates from comparable corpora:

(1) The first set of models utilizes standard monolingual LDA (Blei et al., 2003b) (see Section 2.3) on concatenated aligned Wikipedia articles (MixLDA, as in Sections 3 and 4). The LDA models are trained with exactly the same parameter setup as our BiLDA models. As mentioned before, by the concatenation of aligned Wikipedia articles given in the source and the target language, we effectively remove the gap between the languages and train on the obtained set of "merged" documents and acquire a shared set of latent topics represented by words in both languages. However, we may still distinguish between source language words and target language words, and use only a subset of all words comprising all target language words in final ranked lists. The goal of this comparison is again to test whether it is useful to provide separate topic representations in two languages by jointly training on separate documents (see Section 2.5), and how it affects the final word representations (as given by Eq. (13)). We again test all models from the previous sections as with MuPTM. Since the obtained set of models effectively relies on a monolingual topic model, their codes are *-*MoPTM*, e.g., KL-MoPTM, JS-MoPTM.

(2) Another baseline model is conceptually similar to our TI model. The model constructs feature vectors, but now in the original word-document space (instead of the lower-dimensional word-topic space) using the TF-IDF weighting scheme (which is completely analogous to the TTF-ITF weighting scheme) and the cosine similarity on the obtained word vectors. This comparison serves to test whether we gain some extra contextual information by translating

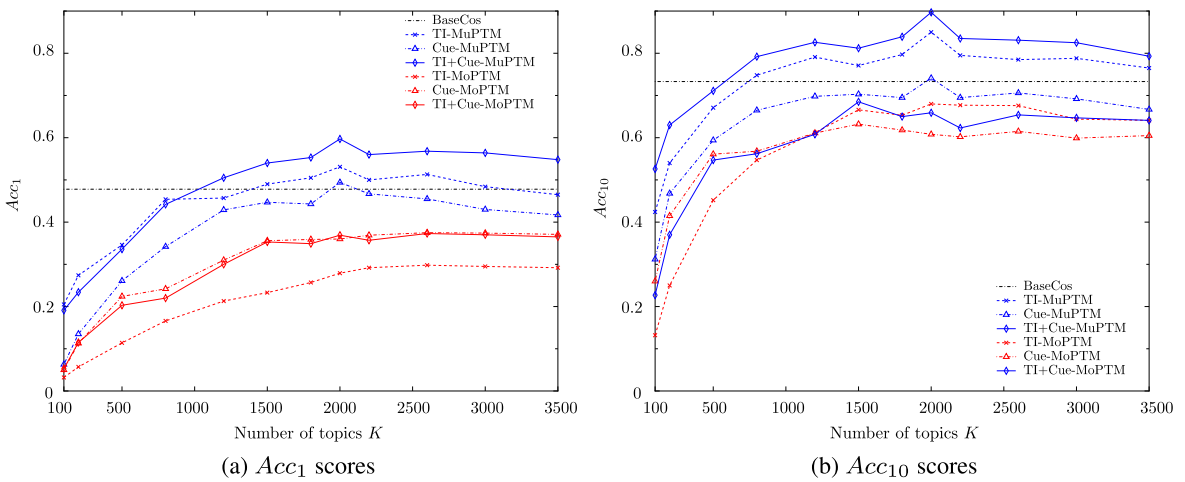


Fig. 9. Acc_1 and Acc_{10} scores for Cue-MuPTM, TI-MuPTM, TI + Cue-MuPTM, and comparisons with baseline models: Cue-MoPTM, TI-MoPTM, TI + Cue-MoPTM, and BaseCos.

Table 5
Best Acc_1 and Acc_{10} scores over all values of K (in parentheses after each result) for all compared models.

Topic model	-MuPTM		-MoPTM	
Similarity model	Acc_1 (K)	Acc_{10} (K)	Acc_1 (K)	Acc_{10} (K)
KL	0.409 (3500)	0.619 (3500)	0.117 (3000)	0.264 (2600)
JS	0.489 (1500)	0.766 (800)	0.312 (2600)	0.594 (1500)
TCos	0.511 (2000)	0.786 (2000)	0.316 (2600)	0.658 (2600)
BC	0.475 (2000)	0.733 (1200)	0.248 (1800)	0.531 (1500)
Cue	0.494 (2000)	0.741 (2000)	0.375 (2600)	0.615 (2600)
TI	0.531 (2000)	0.850 (2000)	0.298 (2600)	0.680 (2000)
TI + Cue	0.597 (2000)	0.897 (2000)	0.373 (2600)	0.685 (1500)
BaseCos	0.478 (–)	0.733 (–)	–	–

Table 6
Lists of the top 5 semantically similar words (Italian to English), where the correct translation candidate is not found (column 1), lies hidden lower in the pruned ranked list (2), and is retrieved as the most similar words (3). All three lists are obtained with TI + Cue-MuPTM.

(1) Romanzo (novel)	(2) Paesaggio (landscape)	(3) Cavallo (horse)
writer	tourist	*horse*
novella	painting	stud
novelette	*landscape*	horseback
humorist	local	hoof
novelist	visitor	breed

our problem from the word-document to the word-topic space, besides the obvious fact that we produce a sort of *dimensionality reduction* which in turn speeds up computations. In other words, we test whether topic models have the ability to build clusters of words which might not always co-occur together in the same textual units and therefore add extra information of similarity besides a direct co-occurrence captured by this baseline model (*BaseCos*).

5.4. Results and discussion

We conduct two different batches of experiments: (1) We compare all our proposed models against the baseline models, and measure the influence of the number of topics K on the overall results in the BLE task and (2) we test and report the effect of topic pruning for a selection of models.

5.4.1. Experiment I: Comparison of all models

Acc_1 and Acc_{10} scores in the BLE task for all models relying directly on the similarity function operating on context vectors with conditional topic probability scores (KL-*, JS-*, BC-*, TCos-* models) are displayed in Fig. 8a and b respectively. Acc_1 and Acc_{10} for all other models are displayed in Fig. 9a and b respectively. Additionally, Table 5 lists the best results for all models along with the optimal number of topics K with which these results have been obtained. Based on all these results, we may derive a series of important conclusions:

(i) A comparison of all *-MuPTM models (blue lines) and all *-MoPTM models (red lines) clearly reveals the utility of training BiLDA on separate documents in place of training standard LDA on concatenated documents. All MuPTM-based models significantly outscore their MoPTM-based variants with LDA trained with exactly the same parameters as BiLDA. By training LDA on concatenated documents, we inherently introduce imbalance in the model, since one of the languages might clearly dominate the latent topic estimation (e.g., in cases when, for instance, English data is of higher quality than Italian data).
(ii) The choice of a similarity function matters. If we compare strictly SF-s operating with exactly the same representations (context vectors comprising conditional topic distributions) as given in Fig. 8a and b, we may observe that the KL model is significantly outperformed by the related JS model (which effectively performs a sort of symmetrization) and two other novel similarity models (the TCos model and the BC model) operating with exactly the same word representations.
(iii) Based on these initial results, the TI model which relies on the representation with the new TTF-ITF weighting scheme is the best scoring basic model of similarity. However, we will later show that by pruning the topic space, other basic models such as JS and BC may outperform the TI model. Moreover, it is very interesting to note that the TI model, which is effectively the same model as BaseCos, but with a shift from the original word-document space to the newly induced word-topic space, outscores the BaseCos model. All other MuPTM-based models (except for KL-MuPTM) are at least on a par with BaseCos. Additionally, for large document collections, the methods such as BaseCos which operate

Table 7
Topic space pruning: Acc_1, MRR, and Acc_{10} scores for JS-MuPTM, TCos-MuPTM and BC-MuPTM which rely on word representations by means of conditional topic distributions over different values of pruning parameter K'. BiLDA. $K = 2000$.

K'	JS-MuPTM			TCos-MuPTM			BC-MuPTM		
	Acc_1	MRR	Acc_{10}	Acc_1	MRR	Acc_{20}	Acc_1	MRR	Acc_{10}
1	0.380	0.488	0.704	0.0	0.0	0.0	0.477	0.575	0.760
2	0.454	0.543	0.717	0.113	0.142	0.195	0.497	0.601	0.792
5	0.508	0.593	0.754	0.191	0.232	0.301	0.543	0.635	0.817
10	0.543	0.622	0.761	0.210	0.243	0.303	0.554	0.648	0.824
20	0.546	0.626	0.772	0.207	0.245	0.314	0.566	0.661	0.831
30	0.539	0.621	0.772	0.220	0.260	0.329	0.572	0.667	0.834
50	0.542	0.624	0.774	0.267	0.318	0.427	0.574	0.668	0.835
70	0.541	0.625	0.771	0.315	0.375	0.502	0.575	**0.669**	**0.844**
100	0.545	0.626	0.769	0.357	0.425	0.572	**0.578**	0.667	0.834
150	0.539	0.620	0.775	0.394	0.471	0.639	0.575	0.666	0.833
200	0.533	0.612	0.769	0.408	0.490	0.670	0.575	0.665	0.831
500	0.517	0.590	0.732	0.471	0.559	0.739	0.569	0.641	0.781
800	0.491	0.555	0.683	0.483	0.573	0.748	0.528	0.600	0.740
1000	0.477	0.536	0.653	0.497	0.586	0.757	0.516	0.584	0.714
1500	0.463	0.573	0.692	0.512	0.598	0.769	0.486	0.550	0.678
2000	0.508	0.571	0.699	0.511	0.605	0.786	0.475	0.538	0.667

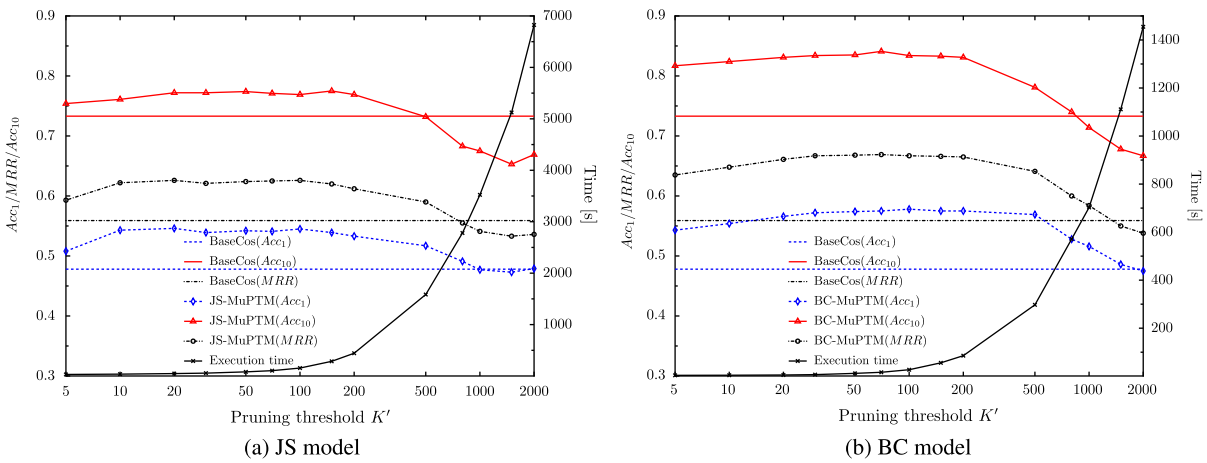
Bold values denote the highest overall Acc_1, MRR, and ACC_{10} scores.

in the original word-document space might become computationally infeasible. This insight shows that reducing the dimensionality of the feature space in word representations might lead both to more effective and computationally tractable models of similarity.

(iv) We may observe that by combining the TI-MuPTM model and the Cue-MuPTM model, we are able to boost the overall performance. The results of the combined TI + Cue-MuPTM model outperform the results obtained by using any of the component models alone. The combined TI + Cue model displays the best overall performance across all compared models of similarity.

(v) Our models of similarity reach their optimal performances with larger values of K (e.g., around the 2000 topics mark). While the tasks that required only coarse categorizations, such as event-centered news clustering (Section 3) or document classification (Section 4) typically used a lower number of topics (in the [50–300] interval), cross-lingual semantic word similarity and bilingual lexicon extraction require a set of fine-grained latent cross-lingual topics which consequently leads to finer-grained topical representations of words. Based on these results, in all further experiments, we will set $K = 2000$, unless noted otherwise.

(vi) Additionally, it has been noted for both monolingual (Turney & Pantel, 2010) and cross-lingual settings (Peirsman & Padó, 2011) that for distributional models synonymy is not the only semantic relation detected within the (pruned) ranked lists of words. The same is true for our distributional models relying on topical knowledge. For instance, besides direct cross-lingual synonymy, that is, the actual translational equivalence, we observe other semantic relations with words ranked highly in the lists (in top ten candidate words): near-synonymy (e.g., *incidente (accident) – crash*), antonymy (e.g., *guerra (war) – peace*), hyponymy (e.g., *particella (particle) – boson*), hypernymy (e.g., *ragazzo (boy) – child*),



(a) JS model (b) BC model

Fig. 10. Acc_1, Acc_{10}, MRR scores for JS-MuPTM and BC-MuPTM along with their execution times. The horizontal axis is in log scale.

meronymy (e.g., *soldato (soldier) – troop*), holonymy (e.g., *mazzo (deck) – card*) and other, uncategorized semantic relations (e.g., *vescovo (bishop) – episcopate*). The quantitative analysis (as performed in (Peirsman & Padó, 2011)) of the semantic relations detected by the models is beyond the scope of this work and will not be further investigated. Ranked lists of semantically similar words provide comprehensible and useful contextual information in the target language given a source word, even when the correct translation candidate is missing, as might be seen in Table 6. This finding may be exploited when building information retrieval models with query expansion (Vulić et al., 2013).

5.4.2. Experiment II: Analysis of topic pruning

In the next set of experiments, we analyze the influence of topic pruning on the behavior of our MuPTM-based models of cross-lingual similarity. All models use output per-topic word distributions from the BiLDA model trained with $K = 2000$ topics. Table 7 displays the results over different values for the pruning parameter K'. For the sake of clear presentation, we omit the results for: (1) the KL model whose behavior resembles the behavior of the JS model, only with lower overall scores, (2) the Cue model where we have not detected any major influence on the overall scores (i.e., the pruning is useful since it reduces execution time, but it does not lead to any improvements in scores), and (3) TI and TI + Cue models which rely on the cosine similarity and whose behavior resembles the behavior of the TCos model. Additionally, Fig. 10a and b display the change in overall scores for JS and BC over different values of K' along with execution times for all pruned JS and BC models. These time-related experiments were conducted on an Intel(R) Xeon(R) CPU E5-2667 2.9 GHz processor. We may notice several interesting phenomena:

(i) Topic pruning helps to obtain higher results in the BLE task for the JS model (the increase is 7.5% even when probability scores do not sum up to 1) and the BC model (the increase is 21.7%). Using only a small subset of possible features (e.g., $K' = 10$), we are able to acquire bilingual lexicons of higher quality with these models as reflected in Acc_1 scores. The improvement in Acc_{10} and *MRR* scores is also prominent. Moreover, as Fig. 10a and b reveal, the utility of topic pruning is especially visible when we compare execution times needed to retrieve ranked lists for test words. For instance, while the BC model needs 1454.3 s to obtain ranked lists when we operate with full K-dimensional representations, the execution time is only 4.2 s with $K' = 10$, and we even obtain better results.

(ii) The reader might wonder why it is useful to induce a fine-grained latent topic space with a large number of topics, and then to perform pruning of the space afterwards, that is, to select only a small subset of the most informative features to represent words. The answer is as follows: While we desire to have a semantic space in which a large number of linguistic phenomena and topics are covered (a large set $\mathcal{Z}$), only a small subset of these topics is relevant to a particular word (topic space pruning). For instance, although we require that our semantic space is expressive enough to present topics and words related to *marine biology* or *radioactive isotopes*, these topics are completely irrelevant when we build a representation for a word *playmaker*.

(iii) We have detected that topic space pruning for similarity models relying on the cosine similarity (TCos, TI, TI + Cue) negatively affects the performance. In the cosine similarity, since the normalization of an inner product is performed (unlike in BC), the absolute weights/probability scores are neglected and the direction (instead of the absolute score) in the semantic space dominates the similarity function (Jones & Furnas, 1987). With a limited number of dimensions, the semantic space is not expressive enough and simply assigns high scores for many irrelevant target language words whose vectors are proportionally similar to the given source language word. For instance, take an extreme case where only one feature is left after pruning. The cosine similarity will assign a perfect similarity score of 1.0 for each target language word regardless of the actual absolute weights, while BC will not produce equal similarity scores for all words.

(iv) The best overall results are obtained with the BC model with $K' = 100$, and we may observe a major improvement over the baseline BaseCos model. A similar behavior is observed with the JS model. Moreover, the BC model is also the fastest model of all proposed models (e.g., 26.9 s compared to 51.5 s of TCos and 155.3 s of JS with $K' = 100$). In summary, by performing topic pruning, we are able to improve our models of cross-lingual similarity both in terms of accuracy and speed.

6. Application IV: Cross-lingual information retrieval

6.1. Task description

Finally, we show how to employ both the *per-topic word distributions* and *per-document topic distributions* of a multilingual probabilistic topic model together, and how to exploit its inference power on an unseen collection in a language modeling (LM) approach to the task of cross-lingual information retrieval (CLIR). CLIR deals with retrieval of documents that are written in a language different from the language of the user's query. In a typical CLIR setting, queries (or rarely documents) are translated using a machine-readable dictionary or a machine translation system, and then a myriad of techniques for the monolingual retrieval may be applied (e.g., Ponte & Croft, 1998; Berger & Lafferty, 1999; Lavrenko & Croft, 2001). In case when readily available translation resources are unavailable, multilingual probabilistic topic models may serve as a valid tool to build a CLIR system that does not rely on any external translation resource and can be trained on general-domain non-parallel data (e.g., Wikipedia) and later inferred and used on in-domain data (e.g., news corpora). Therefore, we again deal with the cross-lingual knowledge transfer.

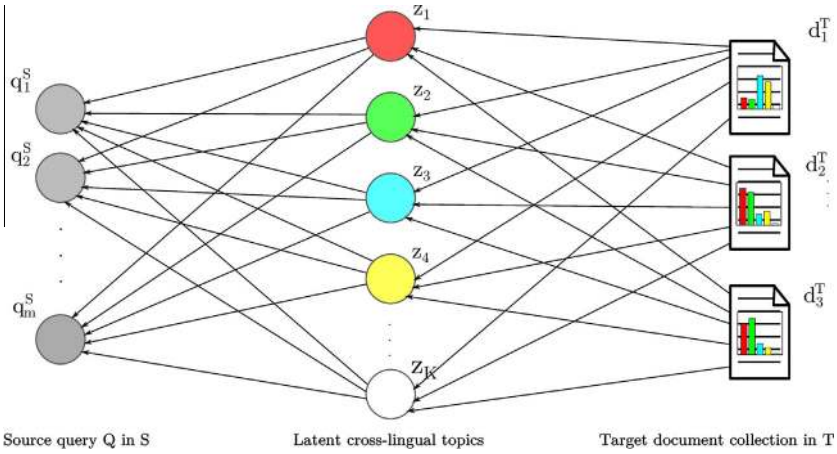


Fig. 11. Graphical representation of the basic probabilistic cross-lingual information retrieval model that relies on the latent layer of cross-lingual topics obtained by a multilingual probabilistic topic model.

Each target document d_j^T can again be presented as a mixture over cross-lingual topics from the set $\mathcal{Z}$ (see Section 2.1) as given by per-document topic distributions (with the values $P\left(z_k|d_j^T\right)$). Additionally, the values $P(w_i^S|z_k)$ from per-topic word distributions may be used to calculate the probability that cross-lingual topic z_k will generate some source word w_i^S. If that word w_i^S is actually a word from the user's query written in the source language, the cross-lingual topics again serve as a bridge that links semantics of the query in the source language with semantics of the document written in the target language.

6.2. Methodology

Given a monolingual setting with only one language L_S, the basic approach for using language models in information retrieval is the query likelihood model, where each document is scored by the likelihood of its model d_j^S to generate a query Q^S of length $m : P(Q^S|d_j^S) = \prod_{i=1}^{m} P(q_i^S|d_j^S)$, where q_i^S denotes a query term from Q^S. We assume the independence between the query terms, i.e., the unigram language model. Since we here deal with cross-lingual information retrieval, where documents are in the target language L_T and query terms are in the source language L_S, we need to establish the cross-lingual connection between them by means of a multilingual probabilistic topic model. The basic MuPTM-based language model for CLIR (Vulić et al., 2011b) that uses only knowledge from the trained multilingual probabilistic topic model follows these steps:

1. Train the model on a (usually general-domain) training corpus and learn per-topic word distributions ϕ and ψ, and per-document topic distributions.
2. Infer the trained model on the target collection given in the target language L_T and obtain per-document topic distributions θ^T for all documents in the collection.
3. For each term $q_i^S \in Q^S$ do: (a) Obtain probabilities $\phi_{k,i} = P(q_i^S|z_k)$ from per-topic word distributions for L_S, for all $k = 1, \ldots, K$; (b) Obtain probabilities $\theta_{j,k}^T = P(z_k|d_j^T)$ for a document d_j^T from the target collection for all $k = 1, \ldots, K$; and (c) Combine the probabilities to obtain the final probability that a source term q_i^S is generated by a document model d_j^T via the latent layer of cross-lingual topics: $P_{muptm}(q_i^S|d_j^T) = \sum_{k=1}^{K} P(q_i^S|z_k)P(z_k|d_j^T)$.
4. Compute the final query likelihood for the entire query:

$$P_{muptm}\left(Q^S|d_j^T\right) = \prod_{i=1}^{m} P_{muptm}\left(q_i^S|d_j^T\right) = \prod_{i=1}^{m}\sum_{k=1}^{K} P(q_i^S|z_k)P\left(z_k|d_j^T\right) = \prod_{i=1}^{m}\sum_{k=1}^{K} \phi_{k,i}\theta_{j,k}^T \tag{24}$$

Eq. (24) provides a query likelihood score for one document, so it has to be repeated for all documents in the collection. Finally, documents are ranked according to their respective query likelihood scores. The intuitive graphical representation of this CLIR technique that connects documents given in target language L_T with query words given in source language L_S is displayed in Fig. 11. There, each target document is represented as a mixture of latent language-independent cross-lingual topics (colored bars that denote per-document topic distributions) and assigns a probability value $P\left(z_k|d_j^T\right)$ for each $z_k \in \mathcal{Z}$ (edges between documents and topics). Moreover, each cross-lingual topic may generate each query word by the probability $P(q_i^S|z_k)$ given by per-topic word distributions (edges between topics and query words). We will call this model the *MuPTM-Basic* model.

Table 8
Statistics of the experimental setup.

Collection	Contents	# Docs
(a) Statistics of test collections		
LAT	LA Times 94 (EN)	110,861
LAT + GH	LA Times 94 (EN) Glasgow Herald 95 (EN)	166,753
NC + AD Algemeen	NRC Handelsblad 94–95 (NL) Dagblad 94–95 (NL)	190,604
CLEF themes (Year: theme nr.)	# Queries	Used for
(b) Statistics of used queries		
NL 2001: 41–90	47	LAT
NL 2002: 91–140	42	LAT
NL 2003: 141–200	53	LAT + GH
EN 2001: 41–90	50	NC + AD
EN 2002: 91–140	50	NC + AD
EN 2003: 141–200	56	NC + AD

This basic model can be efficiently combined with other models that capture additional evidence for estimating $P\left(q_i^S|d_j^T\right)$. When dealing with monolingual retrieval, Wei and Croft (2006) have detected that their model that relies on knowledge from a probabilistic topic model (e.g., LDA) is too coarse to be used as the only representation for retrieval and to produce quality retrieval results. Therefore, they have linearly combined it with the original document model in the monolingual context and observed a major improvement in their results. We can follow the same principle in the cross-lingual setting, but with limited efficacy, since there is sometimes a minimum word overlap between languages. However, that model proves to be useful for languages from the same family, since, for instance, many named entities do not change across languages.[12] The *MuPTM-Unigram* CLIR model combines the representation by means of a multilingual probabilistic topic model with the knowledge of the shared words across languages within the unified language modeling framework with the Jelinek–Mercer and Dirichlet smoothing (Zhai & Lafferty, 2004). The model is as follows:

$$P\left(q_i^S|d_j^T\right) = \lambda\left(\frac{N_{d_j^T}}{N_{d_j^T}+\mu}P_{mle}\left(q_i^S|d_j^T\right) + \left(1-\frac{N_{d_j^T}}{N_{d_j^T}+\mu}\right)P_{mle}\left(q_i^S|Coll^T\right)\right) + (1-\lambda)P_{muptm}\left(q_i^S|d_j^T\right) \quad (25)$$

$N_{d_j^T}$ is the length in words of the document d_j^T, $P_{mle}\left(q_i^S \mid d_j^T\right)$ is the maximum likelihood estimate of the source query term q_i^S in the target document d_j^T, $P_{mle}\left(q_i^S \mid Coll^T\right)$ is the maximum likelihood estimate of q_i^S in the entire target collection, μ is the Dirichlet coefficient (Zhai & Lafferty, 2004), λ is the interpolation parameter, and $P_{muptm}(q_i^S \mid d_j^T)$ is given by Eq. (24).

In this overview, we only report results obtained by these two basic MuPTM-based CLIR models. The language modeling framework for IR allows combining more various evidences in the query likelihood model. The framework is also topic model-independent and it allows experimentations and comparisons of different multilingual probabilistic topic models. More complex LM CLIR models that rely on the representation by means of multilingual topic models are described and evaluated by Vulić et al. (2013), and we refer the interested reader to that article. For instance, there we use bilingual MuPTM-based probabilistic dictionaries obtained by the methods from Section 5 instead of the shared words across languages and perform query expansion. The LM framework also allows blending knowledge from the external resources (e.g., machine-readable dictionaries) with the topical representation, but that research is beyond the scope of this work.

6.3. Experimental setup

6.3.1. Datasets

For *training*, we use English–Dutch Wikipedia articles, but, to reduce data sparsity, we augment that dataset with 6206 Europarl (Koehn, 2005) English–Dutch document pairs.[13] We train the BiLDA model with $K = 400$, 1000, and 2200 topics.

Experiments were conducted on three *test* datasets taken from the CLEF 2001–2003 CLIR campaigns: the LA Times 1994 (*LAT*), the LA Times 1994 plus the Glasgow Herald 1995 (*LAT + GH*) in English, and the NRC Handelsblad 1994–1995 plus the Algemeen Dagblad 1994–1995 (*NC + AD*) in Dutch. Queries were extracted from the *title* and *description* fields of all CLEF themes for each year and queries without relevant documents were removed from the query sets. The overall statistics are provided in Table 8a and b (Vulić et al., 2013). We set $\lambda = 0.3$ to assign more weight to topic models, and $\mu = 2000$.

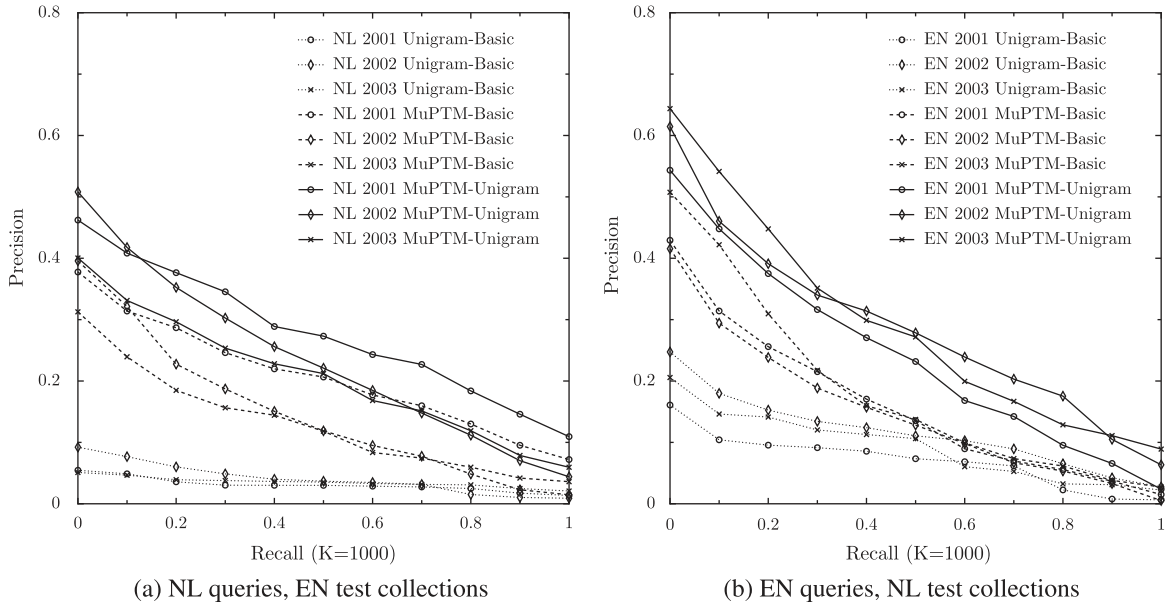
[12] If the user is searching for the volcano *Mauna Loa* in Croatian or Dutch, there is a fair chance that relevant documents in English, German or even Hungarian and Finnish (which do not come from the same phylum) may be retrieved since the term *Mauna Loa* remains unchanged across all these languages.

[13] We do not take advantage of the structure of the parallel documents and do not make us of the sentence-level alignments to improve the scores. We use the exact same approach that we apply for the Wikipedia documents and presuppose only document alignment before training.

Table 9
Perplexity scores on all CLEF test collections and MAP scores for all campaigns for the *MuPTM-Basic* and the *MuPTM-Unigram* CLIR models. Multilingual probabilistic topic model used in the experiments is BiLDA.

Test coll.	$K = 400$		$K = 1000$		$K = 2200$
(a) Perplexity scores					
LAT	**111.12**		215.98		437.11
LAT + GH	**107.91**		210.15		432.76
NC + AD	**110.85**		219.45		527.43
	MuPTM-Basic			*MuPTM-Unigram*	
Queries\K	400	1000	2200	400	1000	2200
(b) MAP scores						
NL 2001	0.1777	0.1969	0.2028	0.2330	0.2673	**0.2813**
NL 2002	0.1117	0.1396	0.1371	0.2093	**0.2253**	0.2206
NL 2003	0.0781	0.1227	0.0784	0.1608	**0.1990**	0.1658
EN 2001	0.1270	0.1453	0.1624	0.2204	0.2275	**0.2398**
EN 2002	0.0932	0.1374	0.1412	0.2455	**0.2683**	0.2665
EN 2003	0.0984	0.1713	0.1529	0.2393	**0.2783**	0.2450

Bold values denote the lowest (i.e., the best) perplexity scores after the inference on each of the test collections.



(a) NL queries, EN test collections (b) EN queries, NL test collections

Fig. 12. Comparison of *MuPTM-Basic*, *Unigram-Basic*, and *MuPTM-Unigram* as their combination. BiLDA with $K = 1000$.

6.3.2. Evaluation metrics

The performance of the CLIR models is reported in the mean average precision (MAP) scores and/or 11-pt precision-recall diagrams for all experiments.

6.4. Results and discussion

Table 9a shows the perplexity scores after the inference of the BiLDA model on the CLEF test collections, while Table 9b displays the MAP scores for all campaigns in both retrieval directions (i.e., English queries – Dutch documents, and vice versa) for the two models presented in Section 6.2. We can observe that, similar to the monolingual setting, using only topical representation seems to be too coarse to produce quality retrieval models. The topical knowledge is, however, useful as an extra portion of information that can be easily embedded within the language modeling framework. The synergy between different evidences when performing retrieval leads to better unsupervised models. Due to a high percentage of shared words between English and Dutch, there is a possibility that the *MuPTM-Unigram* model draws its performance mainly from the part specified by the "non-MuPTM" part of the model (see Eq. (25)). This model called *Unigram-Basic* which can be obtained by setting $\lambda = 1$ in Eq. (25) does not use any topical knowledge and relies only on the shared words. However, Fig. 12a and b clearly show that the final combined *MuPTM-Unigram* model clearly works as a positive synergy between the two simpler basic models, outperforming both of them. However, *MuPTM-Basic* provides higher scores than *Unigram-*

Basic and is therefore more important for the overall performance of the combined model. As in Sections 4 and 5, we have also experimented with monolingual LDA trained on merged document pairs (MixLDA), but the reported retrieval results with the MoPTM-Basic retrieval model are in the range [0.01–0.03] in terms of MAP scores which is extremely lower than the MAP scores reported for our *MuPTM-Basic* model in Table 9b.

The reported results also show the potential of training the multilingual probabilistic topic model on huge-volume out-of-domain data and its inference and usage on another collection, which is not necessarily completely theme-aligned to the training corpus. It is also again difficult to predict the "sweet spot" for K for which the best retrieval results are expected.

Since the goal in this section is to give the reader a basic insight on how to utilize per-topic word distributions and per-document topic distributions of a multilingual probabilistic topic model in CLIR, we do not provide a thorough discussion here. That, along with many more experiments and comparisons (e.g., a comparison where it was shown that these results with MuPTM-s are comparable to results of the models which rely on translation tools such as *Google Translate*) might again be found in (Vulić et al., 2013), and we encourage the interested reader to study that article in more detail. One important aspect lacking in that article is the reported mismatch between intrinsic perplexity scores and extrinsic evaluation results, that is, MAP scores for CLIR. The comparison of the results in Table 9a and b clearly shows that the theoretical in vitro measure of perplexity, often used to compare the quality of probabilistic topic models, does not guarantee a better in vivo performance in actual applications such as CLIR. The same general conclusion for language models in information retrieval is also drawn by Azzopardi, Girolami, and van Rijsbergen (2003). Additionally, these findings strengthen the claims from Section 4.4 where we have also stated that better perplexity scores of a probabilistic topic model do not necessarily reflect in a better "real-life" classification task performance.

7. Conclusions and future perspectives in multilingual probabilistic topic modeling

7.1. Final discussion

In this article, we have conducted the first systematic and thorough overview of the current advances in multilingual probabilistic topic modeling, with a special focus on unsupervised learning from non-parallel data easily obtainable from the Web, and cross-lingual knowledge transfer across corpora written in different languages by means of the multilingual probabilistic topic models. We have provided precise formal definitions of this modeling concept and drew analogies with its monolingual variants and a broader concept of inducing latent cross-lingual concepts from multilingual data. Additionally, we have described the importance of obtaining a shared latent topical space in the form of latent cross-lingual topics. Multilingual probabilistic topic models which induce such a language-independent cross-lingual topical space in general comprise two sets of probability distributions: (1) per-document topic distributions that define topic importance in a document and (2) per-topic word distributions that define importance of vocabulary words in each language for each cross-lingual topic. As a case study, we have shown how to utilize these sets of distributions by providing a short survey of various cross-lingual tasks and the utility of the distributions in these tasks. An insight into these applications reveals several interesting phenomena.

In this article, we have shown that monolingual probabilistic topic models are only a special, degenerate case of multilingual topic models (e.g., LDA is only a special case of BiLDA which operates with only one language). As a consequence, all frameworks presented in this paper which tackle cross-lingual tasks are easily adapted to and functional in the monolingual settings.

A standard approach to induce topical knowledge when handling cross-lingual tasks in cases when an external translation resource is absent is to train a monolingual topic model such as LDA on merged/concatenated documents from an aligned document pair with documents in two different languages). However, our comparisons across different cross-lingual tasks (see Sections 3–6) clearly indicate that training a multilingual topic model on separate documents from a document pair instead of a monolingual topic model on the "merged" artificially created document leads to significantly higher (sometimes even to extremely higher) scores in all these applications.

We also observe that the optimal setting of a priori parameters (e.g., the number of topics) is heavily application-dependent, and it is not apparent how to detect the "sweet spot", that is, the optimal number of topics K. The tasks that require only coarse categorizations, such as event-centered news clustering or document classification typically also require coarse-grained text representations, and use a lower number of topics. It seems that such coarse semantic representations are sufficient to successfully learn the correct category matchings. In Section 4, we have also shown that the technique of *topic smoothing* which combines representations obtained by different K-s may lead to a more robust final model. On the other hand, tasks like cross-lingual semantic similarity of words and information retrieval needed finer representations, and, even with the high number of topics topical representations lack sufficient detail.

Another advantage of multilingual probabilistic modeling lies in the fact that their output sets of distributions implicitly provide dimensionality reduction similar to other latent semantic models (e.g., LSA). By learning per-document topic distributions, the document representation is translated from the original high-dimensional word-document space to a lower-dimensional topic-document space. In a similar fashion, by learning per-topic word distributions, the word representation is effectively translated from the original high-dimensional word-document space to a lower-dimensional word-topic space. In Section 5, where a complete framework for modeling cross-lingual semantic similarity by means of MuPTM has been introduced and described, we have also introduced the paradigm of topic pruning, which selects only a subset of highly

relevant topics to represent a word or a document. By selecting only a subset of reduced features/topics, we are able to further decrease the dimensionality of the representation and obtain such *pruned representations*, which lead to final improvements in both the quality of results and the speed of computations.

The results across these applications also reveal that the lower-dimensional representations (i.e., word-topic space and topic-document space) of the original word-document space alone might not be discriminative enough for the tasks that require fine granularity matchings (e.g., cross-lingual information retrieval in Section 6). However, it seems that combining the latent semantic lower-dimensional representations with the original higher-dimensional representations leads to more effective and robust models. For instance, the *MuPTM-Unigram* model that combines unigrams shared across languages with topical knowledge and topical representation of documents leads to much better retrieval scores than *MuPTM-Basic* that uses only topical knowledge as the only representation for retrieval.

Finally, since all topic models in this article have been trained on comparable Wikipedia data, we have also implicitly shown the validity of training on such high-volume easily obtainable comparable datasets in a wide spectrum of NLP/IR tasks. The presented MuPTM framework is unsupervised and language pair independent in its design (since it does not rely on any external translation resource and induces knowledge directly from the given multilingual data). Consequently, that makes it potentially applicable to many language pairs.

7.2. Other applications

We have demonstrated how to make use of the output per-document topic distributions and per-topic word distributions in four fundamental cross-lingual tasks. Besides cross-lingual event-centered news clustering (which is only a special case of cross-lingual document clustering), cross-lingual document classification, cross-lingual semantic similarity and cross-lingual information retrieval that have been tackled and perused here, MuPTM has been applied to a series of other NLP/IR cross-lingual tasks which we briefly list here:

- **Cross-lingual keyword recommendation** (Takasu, 2010). Both the text and the keywords are mapped into the same latent cross-lingual topical space (induced from parallel data), and a ranked list of keywords is provided based on the learned per-topic word distributions and per-document topic distributions.
- **Cross-lingual document matching** (Platt et al., 2010; Zhu, Li, Chen, & Yang, 2013). The goal of the task is to retrieve the most similar document in the target language given a document in the source language. The knowledge of tightly coupled documents may be used to build high quality aligned multilingual datasets. Document representations by means of per-document topic distributions are again used to provide language independent representations of documents and allow their comparison in the shared latent cross-lingual space.
- **Cross-lingual sentiment analysis** (Boyd-Graber & Resnik, 2010). A multilingual topic model is used to discover a consistent, unified picture of sentiment across multiple languages.
- **Transliteration mining** (Richardson, Nakazawa, & Kurohashi, 2013). A framework for cross-lingual semantic similarity described in Section 5 (i.e., the knowledge from per-topic word distributions) has been used to provide information about semantic similarity between potential transliteration pairs.
- **Cross-lingual entity linking** (Zhang, Liu, & Zhao, 2013). Given an entity mention, the goal of the task is to link the mention of an entity to some given knowledge base (e.g., Wikipedia). The output of the BiLDA model is again used to compute the similarity between the context of the entity mention and the appropriate Wikipedia page to which the mention should be linked.
- **Cross-lingual word sense disambiguation** (Tan & Bond, 2013). Given a sentence along with a polysemous word, the goal is to provide a correct sense for the polysemous word. A multilingual topic model may be used in this task to match the query sentence to a list of sentences in the other language based on the most probable topics of these sentences. The correct sense is then extracted from the topically most similar sentence in the other language.
- **Bilingual lexicon extraction** (Vulić & Moens, 2012; Liu, Duh, & Matsumoto, 2013; Chu, Nakazawa, & Kurohashi, 2013). Models of MuPTM-based semantic similarity from Section 5 and the knowledge coming from language-specific per-topic word distributions may be used in various ways to extract word translation pairs from multilingual data.
- **More advanced CLIR models** (Ganguly et al., 2012; Vulić & Moens, 2013). The topical knowledge obtained by a multilingual topic model may be embedded into the relevance modeling framework for cross-lingual information retrieval. This combination results in more effective and more robust models of cross-lingual IR.

All these tasks can be accomplished by means of the MuPTM-based representations of words and documents, that is, by the output per-document topic distributions and the per-topic word distributions.

7.3. Future work

A straightforward line of future work leads to investigating more applications of the MuPTM framework (e.g., cross-lingual document summarization, keyword and keyphrase extraction).

In order to improve the quality of the lower-dimensional topical representations of documents in the multilingual domain, there is a huge number of paths that could be followed. In the same manner as for the natural "LDA to BiLDA"

extension, other more sophisticated and application-oriented probabilistic topic models developed for the monolingual setting could be ported into the multilingual setting (e.g., Wallach, 2006; Blei & McAuliffe, 2007; Gormley, Dredze, Durme, & Eisner, 2012). These extensions include, among others, the use of multi-word expressions and collocations along with single words (Wallach, 2006), the use of sentence information or word ordering (using Hidden Markov Models) to yield more coherent topic distributions over documents (e.g., Griffiths, Steyvers, Blei, & Tenenbaum, 2004; Boyd-Graber & Blei, 2008). The use of hierarchical topics (general super-topics connected with more focused sub-topics, see, e.g., Blei, Griffiths, Jordan, & Tenenbaum, 2003a; Mimno, Li, & McCallum, 2007) is another interesting field of research in the multilingual setting. Moreover, there is a need to develop multilingual probabilistic topic models that fit data which is less comparable and more divergent and unstructured than Wikipedia or news stories, where only a subset of latent cross-lingual topics overlaps across documents written in different languages. Additionally, the more data-driven topic models should be able to learn the optimal number of topics dynamically according to the properties of training data itself (the so-called *non-parametrized models*, Li, Blei, & McCallum, 2007; Zavitsanos, Paliouras, & Vouros, 2011), and clearly distinguish between shared and non-shared topics in a multilingual corpus.

Additionally, besides being a multilingual environment, the Web and the world of information are also locales for multiple idioms of the same language. For instance, the "language" of the social media consumers or typical end-users differs from the language of Wikipedia entries, online shops or legal terminology. Different domains also display different usage of language. Therefore, one line of future research also lies in studying and applying the models initially tailored for the multilingual setting within this *multi-idiomatic* setting, or building new *multi-idiomatic topic models*. The first step in that direction has already been made as it was proven that the topical knowledge combined with the (CL) IR framework from Section 6 is extremely useful in a new task of linking Pinterest users to relevant online shops based on the content the users post on their personal pages (Zoghbi, Vulić, & Moens, 2013b, 2013a; Vulić, Zoghbi, & Moens, 2014).

Finally, as the probabilistic topic models have proven to work in the multilingual settings with comparable corpora, we feel that it is time to expand the current research and carry it into the multimodal domain. Since there exists a significant semantic gap between the "visual words" and textual words, the comparability in the multimodal setting is inherent, and the multilingual probabilistic topic models that operate with comparable data should be transferred and adapted to the multimodal setting, with some preliminary steps already made in that direction (Feng & Lapata, 2010; Bruni, Uijlings, Baroni, & Sebe, 2012; Roller & Schulteim Walde, 2013). These multimodal models should be capable of dealing with the inherent comparable nature of any dataset consisting of text and images or video. However, some initial studies have revealed that the multimodal context serves as a more complex setting, and additional expansions of the multimodal topic models are needed to effectively handle the existing gap between different modalities.

Acknowledgments

The research presented in this article has been carried out in the framework of several research projects. It has been partially supported by the following projects: **CLASS** (EU FP6-027978) financed by the EU Sixth Framework Programme ICT, **AMASS++** (SBO-060051) financed by Instituut voor de Aanmoediging van Innovatie door Wetenschap en Technologie in Vlaanderen (IWT), **WebInsight** (BIL/08/08), and **TermWise** (IOF-KP/09/001) financed by the Flemish government.

References

Aletras, N., & Stevenson, M. (2013). Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th international conference on computational semantics (IWCS)* (pp. 13–22).

Arenas-Garca, J., Meng, A., Petersen, K. B., Schiøler, T. L., Hansen, L. K., & Larsen, J. (2007). Unveiling music structure via PLSA similarity fusion. In *Proceedings of the IEEE international workshop on machine learning for signal processing* (pp. 419–424).

Asuncion, A., Welling, M., Smyth, P., & Teh, Y. W. (2009). On smoothing and inference for topic models. In *Proceedings of the 25th conference on uncertainty in artificial intelligence (UAI)* (pp. 27–34).

Azzopardi, L., Girolami, M., & van Rijsbergen, C. J. (2003). Investigating the relationship between language model perplexity and IR precision–recall measures. In *Proceedings of the 26th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR)* (pp. 369–370).

Bagga, A., & Baldwin, B. (1998). Algorithms for scoring coreference chains. In *Proceedings of the 1st international conference on language resources and evaluation workshop on linguistics coreference* (pp. 563–566).

Berger, A., & Lafferty, J. (1999). Information retrieval as statistical translation. In *Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval (SIGIR)*, pp. 222–229.

Bhattacharyya, A. (1943). On a measure of divergence two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society, 35*, 199–209.

Blei, D. M. & Jordan, M. I. (2003). Modeling annotated data. In *Proceedings of the 26th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR)* (pp. 127–134).

Blei, D. M. & McAuliffe, J. D. (2007). Supervised topic models. In *Proceedings of the 21st annual conference on advances in neural information processing systems (NIPS)* (pp. 121–128).

Blei, D., Griffiths, T. L., Jordan, M. I., & Tenenbaum, J.B. (2003a). Hierarchical topic models and the nested Chinese restaurant process. In *Proceedings of the 16th annual conference on advances in neural information processing systems (NIPS)*.

Blei, D. M., Franks, K., Jordan, M. I., & Mian, I. S. (2006). Statistical modeling of biomedical corpora: Mining the caenorhabditis genetic center bibliography for genes related to life span. *BMC Bioinformatics, 7*, 250.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003b). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Boyd-Graber, J. L., & Blei, D. (2008). Syntactic topic models. In *Proceedings of the 22nd annual conference on advances in neural information processing systems (NIPS)* (pp. 185–192).

Boyd-Graber, J., & Blei, D. M. (2009). Multilingual topic models for unaligned text. In *Proceedings of the 25th conference on uncertainty in artificial intelligence (UAI)* (pp. 75–82).

Boyd-Graber, J. L., & Resnik, P. (2010). Holistic sentiment analysis across languages: Multilingual supervised latent Dirichlet allocation. In *Proceedings of the 2010 conference on empirical methods in natural language processing (EMNLP)* (pp. 45–55).
Boyd-Graber, J. L., Blei, D. M., & Zhu, X. (2007). A topic model for word sense disambiguation. In *Proceedings of the joint 2007 conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)* (pp. 1024–1033).
Bruni, E., Uijlings, J. R. R., Baroni, M., & Sebe, N. (2012). Distributional semantics with eyes: Using image analysis to improve computational representations of word meaning. In *Proceedings of the 20th ACM multimedia conference (MM)* (pp. 1219–1228).
Bullinaria, J. A., & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods, 39*(3), 510–526.
Carbonell, J. G., Yang, J. G., Frederking, R. E., Brown, R. D., Geng, Y., Lee, D., et al. (1997). Translingual information retrieval: A comparative evaluation. In *Proceedings of the 15th international joint conference on artificial intelligence (IJCAI)* (pp. 708–714).
Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., & Blei, D. (2009). Reading tea leaves: How humans interpret topic models. In *Proceedings of the 23rd annual conference on advances in neural information processing systems (NIPS)* (pp. 288–296).
Chen, H. -H., & Lin, C. -J. (2000). A multilingual news summarizer. In *Proceedings of the 18th conference on computational linguistics (COLING)* (pp. 159–165).
Chew, P. A., Bader, B. W., Kolda, T. G., & Abdelali, A. (2007). Cross-language information retrieval using PARAFAC2. In *Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining (KDD)* (pp. 143–152).
Chu, C., Nakazawa, T., & Kurohashi, S. (2013). Iterative bilingual lexicon extraction from comparable corpora with topical and contextual knowledge. In *Proceedings of the 15th international conference on intelligent text processing and computational linguistics (CICLING)* (pp. 296–309).
Cimiano, P., Schultz, A., Sizov, S., Sorg, P., & Staab, S. (2009). Explicit versus latent concept models for cross-language information retrieval. In *Proceedings of the 21st international joint conference on artifical intelligence (IJCAI)*, pp. 1513–1518.
Dagan, I., Pereira, F. C. N., & Lee, L. (1994). Similarity-based estimation of word cooccurrence probabilities. In *Proceedings of the 32nd annual meeting of the association for computational linguistics (ACL)* (pp. 272–278).
Dagan, I., Lee, L., & Pereira, F. C. N. (1997). Similarity-based methods for word sense disambiguation. In *Proceedings of the 35th annual meeting of the association for computational linguistics (ACL)* (pp. 56–63).
Daumé III, H., & Jagarlamudi, J. (2011). Domain adaptation for machine translation by mining unseen words. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies (ACL–HLT)* (pp. 407–412).
DeSmet, W., & Moens, M. -F. (2009). Cross-language linking of news stories on the Web using interlingual topic modeling. In *Proceedings of the CIKM 2009 workshop on social web search and mining (SWSM@CIKM)*, pp. 57–64.
De Smet, W., Tang, J., & Moens, M. -F. (2011). Knowledge transfer across multilingual corpora via latent topics. In *Proceedings of the 15th Pacific-Asia conference on knowledge discovery and data mining (PAKDD)* (pp. 549–560).
DeSmet, W., & Moens, M.-F. (2013). Representations for multi-document event clustering. *Data Mining and Knowledge Discovery, 26*(3), 533–558.
Deveaud, R., SanJuan, E., & Bellot, P. (2013). Are semantically coherent topic models useful for ad hoc information retrieval? In *Proceedings of the 51st annual meeting of the association for computational linguistics (ACL)* (pp. 148–152).
Ding, C. H. Q., Li, T., & Peng, W. (2008). On the equivalence between non-negative matrix factorization and probabilistic Latent Semantic Indexing. *Computational Statistics & Data Analysis, 52*(8), 3913–3927.
Dinu, G., & Lapata, M. (2010). Topic models for meaning similarity in context. In *Proceedings of the 23rd international conference on computational linguistics (COLING)* (pp. 250–258).
Dumais, S. T., Landauer, T. K., & Littman, M. (1996). Automatic cross-linguistic information retrieval using Latent Semantic Indexing. In *Proceedings of the SIGIR workshop on cross-linguistic information retrieval* (pp. 16–23).
Evans, D. K., Klavans, J. L., & McKeown, K. R. (2004). Columbia newsblaster: Multilingual news summarization on the Web. In *Proceedings of the 6th meeting of the North American chapter of the association for computational linguistics: Human language technologies (NAACL–HLT)* (pp. 1–4).
Feng, Y., & Lapata, M. (2010). Topic models for image annotation and text illustration. In *Proceedings of the 11th meeting of the North American chapter of the association for computational linguistics: Human language technologies (NAACL–HLT)* (pp. 831–839).
Fukumasu, K., Eguchi, K., & Xing, E. P. (2012). Symmetric correspondence topic models for multilingual text analysis. In *Proceedings of the 25th annual conference on advances in neural information processing systems (NIPS)* (pp. 1295–1303).
Fung, P., & Yee, L. Y. (1998). An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 36th annual meeting of the association for computational linguistics and the 17th international conference on computational linguistics (ACL–COLING)* (pp. 414–420).
Ganguly, D., Leveling, J., & Jones, G. (2012). Cross-lingual topical relevance models. In *Proceedings of the 24th international conference on computational linguistics (COLING)* (pp. 927–942).
Gaussier, É., & Goutte, C. (2005). Relation between PLSA and NMF and implications. In *Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR)*, pp. 601–602.
Gaussier, É., Renders, J. -M., Matveeva, I., Goutte, C., & Déjean, H. (2004). A geometric view on bilingual lexicon extraction from comparable corpora. In *Proceedings of the 42nd annual meeting of the association for computational linguistics (ACL)* (pp. 526–533).
Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 6*(6), 721–741.
Girolami, M., & Kabán, A. (2003). On an equivalence between PLSI and LDA. In *Proceedings of the 26th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR)* (pp. 433–434).
Gliozzo, A. M., Pennacchiotti, M., & Pantel, P. (2007). The domain restriction hypothesis: Relating term similarity and semantic consistency. In *Proceedings of the 9th meeting of the North American chapter of the association for computational linguistics: Human language technologies (NAACL-HLT)* (pp. 131–138).
Gormley, M. R., Dredze, M., Durme, B. V., & Eisner, J. (2012). Shared components topic models. In *Proceedings of the annual conference of the North-American chapter of the association for computational linguistics: Human language technologies (NAACL–HLT)* (pp. 783–792).
Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. In *Proceedings of the national academy of sciences of the United States of America (PNAS)* (pp. 5228–5235).
Griffiths, T. L., Steyvers, M., Blei, D. M., & Tenenbaum, J. B. (2004). Integrating topics and syntax. In *Proceedings of the 17th annual conference on neural information processing systems (NIPS)* (pp. 537–544).
Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review, 114*(2), 211–244.
Haghighi, A., Liang, P., Berg-Kirkpatrick, T., & Klein, D. (2008). Learning bilingual lexicons from monolingual corpora. In *Proceedings of the 46th annual meeting of the association for computational linguistics: Human language technologies (ACL–HLT)* (pp. 771–779).
Hall, D., Jurafsky, D., & Manning, C. D. (2008). Studying the history of ideas using topic models. In *Proceedings of the 2008 conference on empirical methods in natural language processing (EMNLP)* (pp. 363–371).
Harris, Z. S. (1954). Distributional structure. *Word, 10*(23), 146–162.
Heinrich, G. (2008). *Parameter estimation for text analysis*. Technical report.
Hofmann, T. (1999a). Probabilistic latent semantic analysis. In *Proceedings of the 15th conference on uncertainty in artificial intelligence (UAI)* (pp. 289–296).
Hofmann, T. (1999b). Probabilistic Latent Semantic Indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval (SIGIR)* (pp. 50–57).
Hu, D., & Saul, L. K. (2009). A probabilistic topic model for unsupervised learning of musical key-profiles. In *Proceedings of the 10th international society for music information retrieval conference (ISMIR)* (pp. 441–446).
Jagarlamudi, J., & Daumé III, H. (2010). Extracting multilingual topics from unaligned comparable corpora. In *Proceedings of the 32nd annual European conference on advances in information retrieval (ECIR)*, pp. 444–456.

Jagarlamudi, J., DauméIII, H., & Udupa, R. (2012). Incorporating lexical priors into topic models. In *Proceedings of the 13th conference of the European chapter of the association for computational linguistics (EACL)* (pp. 204–213).
Joachims, T. (1999). Making large-scale SVM learning practical. In B. Schölkopf, C. Burges, & A. Smola (Eds.), *Advances in kernel methods – support vector learning* (pp. 169–184). MIT Press.
Jones, W. P., & Furnas, G. W. (1987). Pictures of relevance: A geometric analysis of similarity measures. *Journal of the American Society for Information Science, 38*(6), 420–442.
Kabadjov, M. A., Atkinson, M., Steinberger, J., Steinberger, R., & der Goot, E. V. (2010). NewsGist: A multilingual statistical news summarizer. In *Proceedings of the European conference on machine learning and knowledge discovery in databases (ECML–PKDD)* (pp. 591–594).
Kazama, J., Saeger, S. D., Kuroda, K., Murata, M., & Torisawa, K. (2010). A Bayesian method for robust estimation of distributional similarities. In *Proceedings of the 48th annual meeting of the association for computational linguistics (ACL)* (pp. 247–256).
Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th machine translation summit (MT SUMMIT)* (pp. 79–86).
Koehn, P., & Knight, K. (2002). Learning a translation lexicon from monolingual corpora. In *Proceedings of the ACL workshop on unsupervised lexical acquisition (ULA)* (pp. 9–16).
Lacoste-Julien, S., Sha, F., & Jordan, M. I. (2008). DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Proceedings of the 21nd annual conference on advances in neural information processing systems (NIPS)* (pp. 897–904).
Lavrenko, V., & Croft, W. B. (2001). Relevance-based language models. In *Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR)* (pp. 120–127).
Lee, D. D., & Seung, H. S. (1999). Algorithms for non-negative matrix factorization. In *Proceedings of the 12th conference on advances in neural information processing systems (NIPS)* (pp. 556–562).
Li, F. -F., & Perona, P. (2005). A Bayesian hierarchical model for learning natural scene categories. In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition (CVPR)* (pp. 524–531).
Li, W., Blei, D. M., & McCallum, A. (2007). Nonparametric Bayes Pachinko allocation. In *Proceedings of the 23rd conference on uncertainty in artificial intelligence (UAI)* (pp. 243–250).
Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory, 37*(1), 145–151.
Littman, M., Dumais, S. T., & Landauer, T. K. (1998). Automatic cross-language information retrieval using Latent Semantic Indexing. In *Chapter 5 of Cross-Language Information Retrieval* (pp. 51–62). Kluwer Academic Publishers.
Liu, X., Duh, K., & Matsumoto, Y. (2013). Topic models + word alignment = a flexible framework for extracting bilingual dictionary from comparable corpus. In *Proceedings of the 17th conference on computational natural language learning (CoNLL)* (pp. 212–221).
Lu, Y., Mei, Q., & Zhai, C. (2011). Investigating task performance of probabilistic topic models: An empirical study of PLSA and LDA. *Information Retrieval, 14*(2), 178–203.
Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.
Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press.
McCallum, A., Wang, X., & Corrada-Emmanuel, A. (2007). Topic and role discovery in social networks with experiments on Enron and academic e-mail. *Journal of Artificial Intelligence Research, 30*(1), 249–272.
Mei, Q., Ling, X., Wondra, M., Su, H., & Zhai, C. (2007). Topic sentiment mixture: Modeling facets and opinions in weblogs. In *Proceedings of the 16th international conference on world wide web (WWW)* (pp. 171–180).
Mei, Q., Cai, D., Zhang, D., & Zhai, C. (2008). Topic modeling with network regularization. In *Proceedings of the 17th international conference on world wide web (WWW)* (pp. 101–110).
Mimno, D., & McCallum, A. (2008). Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. In *Proceedings of the 24th conference in uncertainty in artificial intelligence (UAI)* (pp. 411–418).
Mimno, D., Li, W., & McCallum, A. (2007). Mixtures of hierarchical topics with Pachinko allocation. In *Proceedings of the 24th international conference on machine learning (ICML)* (pp. 633–640).
Mimno, D., Wallach, H., Naradowsky, J., Smith, D. A., & McCallum, A. (2009). Polylingual topic models. In *Proceedings of the 2009 conference on empirical methods in natural language processing (EMNLP)* (pp. 880–889).
Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the 2011 conference on empirical methods in natural language processing (EMNLP)* (pp. 262–272).
Minka, T. P., & Lafferty, J. D. (2002). Expectation–propagation for the generative aspect model. In *Proceedings of the 18th conference in uncertainty in artificial intelligence (UAI)* (pp. 352–359).
Montalvo, S., Martínez-Unanue, R., Casillas, A., & Fresno, V. (2006). Multilingual document clustering: An heuristic approach based on cognate named entities. In *Proceedings of the 44th annual meeting of the association for computational linguistics and the 21st international conference on computational linguistics (ACL-COLING)*.
Montalvo, S., Martínez-Unanue, R., Casillas, A., & Fresno, V. (2007). Multilingual news clustering: Feature translation vs. identification of cognate named entities. *Pattern Recognition Letters, 28*(16), 2305–2311.
Munteanu, D. S., & Marcu, D. (2006). Extracting parallel sub-sentential fragments from non-parallel corpora. In *Proceedings of the 44th annual meeting of the association for computational linguistics and the 21st international conference on computational linguistics (ACL-COLING)* (pp. 81–88).
Munteanu, D. S., & Marcu, D. (2005). Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics, 31*(4), 477–504.
Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010). Automatic evaluation of topic coherence. In *Proceedings of the 10th conference of the North-American chapter of the association for computational linguistics: Human language technologies (NAACL–HLT)* (pp. 100–108).
Ni, X., Sun, J. -T., Hu, J., & Chen, Z. (2009). Mining multilingual topics from Wikipedia. In *Proceedings of the 18th international world wide web conference (WWW)* (pp. 1155–1156).
Ni, X., Sun, J. -T., Hu, J., & Chen, Z. (2011). Cross lingual text classification by mining multilingual topics from Wikipedia. In *Proceedings of the 4th international conference on web search and web data mining (WSDM)* (pp. 375–384).
Peirsman, Y., & Padó, S. (2011). Semantic relations in bilingual lexicons. *ACM Transactions on Speech and Language Processing, 8*(2) (article 3).
Platt, J. C., Toutanova, K., & Yih, W. -T. (2010). Translingual document representations from discriminative projections. In *Proceedings of the 2010 conference on empirical methods in natural language processing (EMNLP)* (pp. 251–261).
Ponte, J. M., & Croft, W. B. (1998). A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval (SIGIR)* (pp. 275–281).
Pouliquen, B., Steinberger, R., Ignat, C., Käsper, E., & Temnikova, I. (2004). Multilingual and cross-lingual news topic tracking. In *Proceedings of the 20th international conference on computational linguistics (COLING)*.
Prochasson, E., & Fung, P. (2011). Rare word translation extraction from aligned comparable documents. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies (ACL–HLT)* (pp. 1327–1335).
Purver, M., Körding, K., Griffiths, T., & Tenenbaum, J. (2006). Unsupervised topic modelling for multi-party spoken discourse. In *Proceedings of the 44th annual meeting of the association for computational linguistics and the 21st international conference on computational linguistics (ACL-COLING)* (pp. 17–24).
Resnik, P., & Smith, N. A. (2003). The Web as a parallel corpus. *Computational Linguistics, 29*(3), 349–380.
Richardson, J., Nakazawa, T., & Kurohashi, S. (2013). Robust transliteration mining from comparable corpora with bilingual topic models. In *Proceedings of the 6th international joint conference on natural language processing (IJCNLP)* (pp. 261–269).

Ritter, A., Mausam, & Etzioni, O. (2010). A latent Dirichlet allocation method for selectional preferences. In *Proceedings of the 48th annual meeting of the association for computational linguistics (ACL)* (pp. 424–434).
Roller, S., & Schulteim Walde, S. (2013). A multimodal LDA model integrating textual, cognitive and visual modalities. In *Proceedings of the 2013 conference on empirical methods in natural language processing (EMNLP)* (pp. 1146–1157).
Roth, B., & Klakow, D. (2010). Combining Wikipedia-based concept models for cross-language retrieval. In *Proceedings of the 1st information retrieval facility conference (IRFC)* (pp. 47–59).
Russell, B. C., Freeman, W. T., Efros, A. A., Sivic, J., & Zisserman, A. (2006). Using multiple segmentations to discover objects and their extent in image collections. In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition (CVPR)* (pp. 1605–1614).
Salton, G., Wong, A., & Yang, C.-S. (1975). A vector space model for automatic indexing. *Communication of the ACM, 18*(11), 613–620.
Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the international conference on new methods in language processing*.
SparckJones, K. (1973). Index term weighting. *Information Storage and Retrieval, 9*(11), 619–633.
Stevens, K., Kegelmeyer, W. P., Andrzejewski, D., & Buttler, D. (2012). Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP–CoNLL)* (pp. 952–961).
Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. *Handbook of Latent Semantic Analysis, 427*(7), 424–440.
Takasu, A. (2010). Cross-lingual keyword recommendation using latent topics. In *Proceedings of the 1st international workshop on information heterogeneity and fusion in recommender systems* (pp. 52–56).
Tamura, A., Watanabe, T., & Sumita, E. (2012). Bilingual lexicon extraction from comparable corpora using label propagation. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP–CoNLL)* (pp. 24–36).
Tan, L., & Bond, F. (2013). XLING: Matching query sentences to a parallel corpus using topic models for WSD. In *Proceedings of the 7th international workshop on semantic evaluation (SEMEVAL)* (pp. 167–170).
Tang, G., Xia, Y., Zhang, M., Li, H., & Zheng, F. (2011). CLGVSM: Adapting generalized vector space model to cross-lingual document clustering. In *Proceedings of the 5th international joint conference on natural language processing (IJCNLP)* (pp. 580–588).
Tao, T., & Zhai, C. (2005). Mining comparable bilingual text corpora for cross-language information integration. In *Proceedings of the 11th ACM SIGKDD international conference on knowledge discovery and data mining (KDD)* (pp. 691–696).
Tipping, M. E., & Bishop, C. M. (1999). Mixtures of probabilistic principal component analysers. *Neural Computation, 11*(2), 443–482.
Titov, I., & McDonald, R. T. (2008). A joint model of text and aspect ratings for sentiment summarization. In *Proceedings of the 46th annual meeting of the association for computational linguistics: Human language technologies (ACL–HLT)* (pp. 308–316).
Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research, 37*(1), 141–188.
Utiyama, M., & Isahara, H. (2003). Reliable measures for aligning Japanese–English news articles and sentences. In *Proceedings of the 41st annual meeting of the association for computational linguistics (ACL)* (pp. 72–79).
VanGael, J., & Zhu, X. (2007). Correlation clustering for crosslingual link detection. In *Proceedings of the 20th international joint conference on artificial intelligence (IJCAI)* (pp. 1744–1749).
Voorhees, E. M. (1986). Implementing agglomerative hierarchic clustering algorithms for use in document retrieval. *Information Processing and Management, 22*(6), 465–476.
Voorhees, E. M. (1999). The TREC-8 question answering track report. In *Proceedings of the 8th text retrieval conference (TREC)*.
Vu, T., Aw, A. T., & Zhang, M. (2009). Feature-based method for document alignment in comparable news corpora. In *Proceedings of the 12th conference of the European chapter of the association for computational linguistics (EACL)* (pp. 843–851).
Vulić, I., & Moens, M. -F. (2012). Detecting highly confident word translations from comparable corpora without any prior knowledge. In *Proceedings of the 13th conference of the European chapter of the association for computational linguistics (EACL)* (pp. 449–459).
Vulić, I., & Moens, M. -F. (2013). A unified framework for monolingual and cross-lingual relevance modeling based on probabilistic topic models. In *Proceedings of the 35th annual European conference on advances in information retrieval (ECIR)* (pp. 98–109).
Vulić, I., DeSmet, W., & Moens, M.-F. (2013). Cross-language information retrieval models based on latent topic models trained with document-aligned comparable corpora. *Information Retrieval, 16*(3), 331–368.
Vulić, I., DeSmet, W., & Moens, M. -F. (2011a). Identifying word translations from comparable corpora using latent topic models. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies (ACL–HLT)* (pp. 479–484).
Vulić, I., Smet, W. D., & Moens, M. -F. (2011b). Cross-language information retrieval with latent topic models trained on a comparable corpus. In *Proceedings of the 7th Asia information retrieval societies conference (AIRS)* (pp. 37–48).
Vulić, I., Zoghbi, S., & Moens, M. -F. (2014). Learning to bridge colloquial and formal language applied to linking and search of e-commerce data. In *Proceedings of the 37th international ACM SIGIR conference on research and development in information retrieval (SIGIR)* (pp. 1195–1198).
Wallach, H. (2006). Topic modeling: Beyond bag-of-words. In *Proceedings of the 23rd international conference on machine learning (ICML)* (pp. 977–984).
Wallach, H., Murray, I., Salakhutdinov, R., & Mimno, D. (2009). Evaluation methods for topic models. In *Proceedings of the 26th annual international conference on machine learning (ICML)* (pp. 1105–1112).
Wang, X., & Grimson, E. (2007). Spatial Latent Dirichlet Allocation. In *Proceedings of the 20th annual conference on advances in neural information processing systems (NIPS)*.
Wang, X., & McCallum, A. (2006). Topics over time: A non-Markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining (CIKM)* (pp. 424–433).
Wang, X., Ma, X., & Grimson, W. E. L. (2009). Unsupervised activity perception in crowded and complicated scenes using hierarchical Bayesian models. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31*(3), 539–555.
Wei, X., & Croft, W. B. (2006). LDA-based document models for ad-hoc retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR)* (pp. 178–185).
Wu, K., & Lu, B. -L. (2007). Cross-lingual document clustering. In *Proceedings of the 11th Pacific-Asia conference on advances in knowledge discovery and data mining (PAKDD)* (pp. 956–963).
Xue, G. -R., Dai, W., Yang, Q., & Yu, Y. (2008). Topic-bridged pLSA for cross-domain text classification. In *Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval (SIGIR)* (pp. 627–634).
Yi, X., & Allan, J. (2009). A comparative study of utilizing topic models for information retrieval. In *Proceedings of the 31th European conference on advances in information retrieval (ECIR)* (pp. 29–41).
Zavitsanos, E., Paliouras, G., & Vouros, G. A. (2011). Non-parametric estimation of topic hierarchies from texts with hierarchical Dirichlet processes. *Journal of Machine Learning Research, 12*, 2749–2775.
Zhai, C., & Lafferty, J. (2004). A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems, 22*(2), 179–214.
Zhang, D., Mei, Q., & Zhai, C. (2010). Cross-lingual latent topic extraction. In *Proceedings of the 48th annual meeting of the association for computational linguistics (ACL)* (pp. 1128–1137).
Zhang, T., Liu, K., & Zhao, J. (2013). Cross lingual entity linking with bilingual topic model. In *Proceedings of the 23rd international joint conference on artificial intelligence (IJCAI)* (pp. 2218–2224).
Zhao, B., & Xing, E. P. (2006). BiTAM: Bilingual topic admixture models for word alignment. In *Proceedings of the 44th annual meeting of the association for computational linguisticsthe and the 21st international conference on computational linguistics and (ACL-COLING)* (pp. 969–976).
Zhao, B., & Xing, E. P. (2007). HM-BiTAM: Bilingual topic exploration, word alignment, and translation. In *Proceedings of the 21st annual conference on advances in neural information processing systems (NIPS)* (pp. 1689–1696).

Zhu, Z., Li, M., Chen, L., & Yang, Z. (2013). Building comparable corpora based on bilingual LDA model. In *Proceedings of the 51st annual meeting of the association for computational linguistics (ACL)* (pp. 278–282).

Zoghbi, S., Vulić, I., & Moens, M. -F. (2013a). Are words enough? A study on text-based representations and retrieval models for linking pins to online shops. In *Proceedings of the 2013 international CIKM workshop on mining unstructured big data using natural language processing (UnstructureNLP@CIKM 2013)* (pp. 45–52).

Zoghbi, S., Vulić, I., & Moens, M. -F. (2013b). I pinned it. Where can I buy one like it? Automatically linking Pinterest pins to online webshops. In *Proceedings of the 2013 international CIKM workshop on data-driven user behavioral modeling and mining from social media (DUBMOD@CIKM 2013)* (pp. 9–12).