Inf Retrieval (2013) 16:331–368
DOI 10.1007/s10791-012-9200-5

Cross-language information retrieval models based on latent topic models trained with document-aligned comparable corpora

Ivan Vulić · Wim De Smet · Marie-Francine Moens

Received: 24 June 2011 / Accepted: 19 April 2012 / Published online: 5 May 2012

Abstract In this paper, we study different applications of cross-language latent topic models trained on comparable corpora. The first focus lies on the task of cross-language information retrieval (CLIR). The Bilingual Latent Dirichlet allocation model (BiLDA) allows us to create an interlingual, language-independent representation of both queries and documents. We construct several BiLDA-based document models for CLIR, where no additional translation resources are used. The second focus lies on the methods for extracting translation candidates and semantically related words using only per-topic word distributions of the cross-language latent topic model. As the main contribution, we combine the two former steps, blending the evidences from the per-document topic distributions and the per-topic word distributions of the topic model with the knowledge from the extracted lexicon. We design and evaluate the novel evidence-rich statistical model for CLIR, and prove that such a model, which combines various (only internal) evidences, obtains the best scores for experiments performed on the standard test collections of the CLEF 2001–2003 campaigns. We confirm these findings in an alternative evaluation, where we automatically generate queries and perform the known-item search on a test subset of Wikipedia articles. The main importance of this work lies in the fact that we train translation resources from comparable document-aligned corpora and provide novel CLIR statistical models that exhaustively exploit as many cross-lingual clues as possible in the quest for better CLIR results, without use of any additional external resources such as parallel corpora or machine-readable dictionaries.

Keywords Cross-language information retrieval · Unsupervised cross-language lexicon extraction · Probabilistic latent topic models · Evidence-rich retrieval models

I. Vulić (✉) · W. De Smet · M.-F. Moens
Department of Computer Science, KU Leuven, Leuven, Belgium
e-mail: ivan.vulic@cs.kuleuven.be

W. De Smet
e-mail: wim.desmet@cs.kuleuven.be

M.-F. Moens
e-mail: marie-francine.moens@cs.kuleuven.be

1 Introduction

The ongoing growth of the World Wide Web and its ubiquity lead to its further localization, where users tend to abandon English as the *lingua franca* of the global network, since more and more content is available in their native languages. However, the availability of content in a huge spectrum of different natural languages, many of which with scarce translation resources, creates a need to efficiently bridge the gap between the languages, using language-independent and generic approaches for miscellaneous problems involving multilingualism.

Machine-readable translation dictionaries do not exist for all language pairs or all domains, as they are usually trained on large parallel corpora or are hand-built. Compiling such lexicons manually is often an expensive and time-consuming task, whereas the methods for mining the lexicons from parallel corpora are not applicable for language pairs and domains where such corpora is unavailable or missing. In a parallel corpus, each text in the source language has an exact translation in the target language. In a comparable corpus on the other hand, documents in different languages are paired when they contain partially overlapping or similar content. Thus, it is much easier to build a high-volume comparable corpus.[1] For instance, news stories and encyclopedia entries often discuss the same event or topics in different languages, but with different focuses. A representative example of such a comparable text collection is Wikipedia, where one may observe articles discussing the same topic, but strongly varying in style, length and even vocabulary, while still sharing a certain amount of main concepts (or topics).

We tackle and combine two different problems: (1) cross-language information retrieval and (2) cross-language lexicon extraction, both in a hard setting, where no external knowledge source is available in the form of a translation dictionary[2], sentence-aligned parallel corpora are absent, and only document alignments for comparable training corpora are known. We accomplish both tasks utilizing the potential of a cross-language generative model, i.e., bilingual Latent Dirichlet allocation (BiLDA), which is an extension of the standard LDA model (Blei et al. 2003), operating in a cross-language setting. Probabilistic topic models are a powerful tool for discovering and analyzing topic patterns in text. They are based upon the idea that there exist latent variables which determine how words in documents might be generated. Fitting a generative model means finding the best set of those latent variables in order to explain the observed data. Within that setting, documents are observed as mixtures of latent topics, where topics are probability distributions over words. In a cross-language setting, we assume that the topic patterns are shared across languages. Another important assumption is that by collecting as many various comparable data (e.g. Wikipedia articles) as possible, the BiLDA model will be able to correctly learn many important topics shared between languages. Such a broad coverage of topics, together with word distributions over topics, will serve as a sound foundation for solving various cross-language problems.

Cross-language information retrieval (CLIR) deals with documents written in a language different from the language of the user's query. At the time of retrieval the query in the source language is typically translated into the target language of the documents with the help of a machine-readable dictionary or a machine translation system. Once a user has

[1] Comparable corpora are much easier to build and obtain than parallel corpora for many language pairs, but that does not necessarily imply that such corpora exist for every language pair. For instance, there is no quality comparable corpus of medieval German and modern German, or Marathi and Hungarian.

[2] We do not utilize any seed translation lexicon for cross-language lexicon extraction either.

retrieved relevant documents for a particular query, they can be translated to the language of the user, possibly by means of manual translation in case resources for automatic translation are unavailable.

This paper addresses the question whether suitable cross-language retrieval models can be built in case machine-readable translation dictionaries or systems that are hand-built or extracted from large parallel sentence-aligned corpora are absent. A number of words might appear with the same meaning in different languages (especially when dealing with languages from the same family). However, when only using a monolingual retrieval model for CLIR, we will miss many relevant documents. Moreover, a word might exhibit the same orthography in different languages, but actually mean something different. Consequently, we need some kind of translation resource, preferably built automatically from comparable corpora.

The transfer of the query into other languages can be accomplished by means of a cross-language probabilistic latent topic model. The language models for retrieval have a sound statistical foundation and can leverage statistical estimation to optimize the retrieval parameters. They can be easily adapted to complex retrieval tasks and have already shown their value in cross-language retrieval settings, incorporating translation probabilities obtained from a translation dictionary in the retrieval model. Our attempt is to exploit the probability distributions over interlingual topics as a translation resource, since they provide an interlingual content representation of the documents.

Cross-language lexicon extraction (CLE) is the task of automatically acquiring translation candidates from parallel, comparable or unrelated texts. We focus on the lexicon extraction from comparable, document-aligned texts, where no seed lexicons are available[3]. Our goal is to model and test the capability of probabilistic topic models to identify potential translations from document-aligned comparable text collections such as Wikipedia. State-of-the-art generative models, most commonly used for obtaining word translation probabilities from parallel corpora, such as IBM Models (Och and Ney 2003) are unusable and computationally intractable within this setting, and cannot be employed on comparable corpora aligned only at the document level.

We try to establish a connection between cross-language latent topics and an idea known as the *distributional hypothesis* (Harris 1954)—words with a similar meaning are often used in similar contexts. Besides the obvious context of direct co-occurrence, we believe that topic models constitute an additional source of knowledge which might be used to improve results in the quest for translation candidates extracted without the availability of a translation dictionary and linguistic knowledge. We designed several methods, all derived from the core idea of using word distributions over topics as an extra source of contextual knowledge. Two words are potential translation candidates if they are often present in the same cross-language topics and not observed in other cross-language topics. In short, a word w_2 from a target language is a potential translation candidate for a word w_1 from a source language, if the distribution of w_2 over the target language topics is similar to the distribution of w_1 over the source language topics.

In the remainder of the paper, the two problems (**CLIR** and **CLE**) will be first observed, described and evaluated independently, still sharing the latent topic model underpinning them. The most important step of this work will combine the two in a coherent evidence-rich document model for CLIR which blends translation probabilities from the translation

[3] Seed lexicons are often constructed by using cognates and words shared across language pairs (e.g., some personal names). However, one cannot rely on such seed lexicons for distant language pairs, while our methods are still fully applicable.

lexicon extracted from per-topic word distributions learned by the BiLDA model with our retrieval model that relies solely on per-document topic distributions and per-topic word distributions connected through the shared space of interlingual topics.

The contributions of the paper are as follows. First, we show the validity and the potential of training bilingual LDA model on bilingual comparable corpora that are available in abundance (e.g. Wikipedia, news). Second, per-topic word distributions learned during training may be used for automatic cross-language lexicon extraction which can be utilized in other cross-language applications. We demonstrate the applicability and usefulness of the BiLDA-induced lexicon in the novel framework of cross-language information retrieval. Third, we successfully integrate the knowledge from the lexicons and the knowledge from probability distributions of the BiLDA model into a novel evidence-rich cross-language statistical retrieval model which uses only internal evidence, and perform a full-fledged evaluation and comparison of all our retrieval models for: (1) the simpler task of English-Dutch and Dutch-English known-item search performed on Wikipedia articles, and (2) English-Dutch and Dutch-English CLIR on the standard CLEF test collections. We also show that the results obtained by our retrieval models, which do not exploit any linguistic knowledge from an external translation dictionary, but exploit all the evidences from probability distributions of the BiLDA model to the fullest, are competitive with and sometimes display a better performance than dictionary-based models for CLIR.

The paper is structured as follows. Section 2 describes related work in cross-language information retrieval, drawing parallels with LDA-based methods for the monolingual setting and listing several approaches which have been trying to develop retrieval models based on latent classes and concepts. This section also discusses different methods of automatic cross-language lexicon extraction, focusing mainly on previous attempts to use topic models to recognize potential translations. Section 3 provides an overview of the cross-language BiLDA model used in all our experiments. Section 4 describes the first set of BiLDA-based statistical models for CLIR, while Sect. 5 gives a complete insight in the methods we used for generating a general lexicon from topical knowledge. We continue the development of retrieval models in Sect. 6, with a model that uses only entries from the obtained lexicons, and another model which combines lexicon entries with the *LDA-only* model from Sect. 4 . In Sect. 7, we present our experimental setup, training and test collections, and used queries. In Sect. 8 we test and evaluate our CLIR models on different collections and within different settings, and discuss the obtained results. Finally, Sect. 9 lists conclusions and future work.

2 Related work

Probabilistic topic models such as probabilistic Latent Semantic Indexing (pLSI) (Hofmann 1999) and Latent Dirichlet allocation (LDA) (Blei et al. 2003) are a popular means to represent the content of a document. Although designed as generative models for the monolingual setting, their extension to multilingual domains follows naturally. Platt et al. (2010) propose several variant models to project documents from multiple languages into a single interlingual vector space, based on the pLSI and LDA models. They use discriminative training for projections creation and evaluate the models on the tasks of parallel document retrieval for Wikipedia and Europarl documents, and cross-lingual text classification on Reuters. Cimiano et al. (2009) use standard monolingual LDA, but trained on concatenated parallel and comparable documents in a document comparison task. Roth and Klakow (2010) try to use standard monolingual LDA trained on concatenated Wikipedia articles for cross-language information retrieval, but they do not obtain decent results without additional usage of a machine translation system. They use the standard Moses machine translation toolkit (Koehn et al. 2007) trained on a parallel sentence-aligned corpus to translate queries and perform monolingual retrieval afterwards.

Recently, the bilingual or multilingual LDA model was independently proposed by different authors (Ni et al. 2009; Mimno et al. 2009; De Smet and Moens 2009; Boyd-Graber and Blei 2009) who identify interlingual topics of different languages. These authors train the bilingual LDA model on a parallel corpus. Jagarlamudi and Daumé III (2010) extract interlingual topics from comparable corpora, but use information from existing hand-built translation dictionaries. Their work follows the opposite direction; while we utilize learned cross-language topics to mine potential translations, they employ knowledge from a dictionary to learn cross-language topics. None of these works apply the bilingual LDA model in a cross-language information retrieval setting.

Cross-language information retrieval is a broad and well-studied research topic (e.g., Grefenstette 1998; Nie et al. 1999; Savoy 2004; Nie 2010). As mentioned, existing methods rely on a translation dictionary to bridge documents of different languages. In a typical setting, cross-language information is learned based on parallel corpora and correlations found in the paired documents (Mathieu et al. 2004), or are based on Latent Semantic Analysis (LSA) applied on a parallel corpus. In the latter case, a singular value decomposition is applied on the term-by-document matrix, where a document is composed of the concatenated text in the two languages, and after rank reduction, the document and the query are projected in a lower dimensional space (Dumais et al. 1996; Littman et al. 1998; Chew et al. 2007; Xue et al. 2008). The term-by-document matrix formed by concatenated parallel documents was used to generate probabilistic term translations with a standard pLSI model and used in cross-language information retrieval (Muramatsu and Mori 2004). Our work follows this line of thinking, but uses generative LDA models trained on a comparable document-aligned corpus, which might be different from the document collection used for retrieval. In addition, our models are trained on the individual documents in different languages, but paired by their joint interlingual topics and, due to that fact, we expect our models to lead to better results than CLIR relying on the cross-language LSI model. Relevance models (Lavrenko et al. 2002) have also been applied for CLIR, but they still need either a parallel corpus or a translation dictionary for estimation. LDA-based monolingual retrieval has been described by Wei and Croft (2006).

Transfer learning techniques, where knowledge is transfered from one source to another, are also used in the frame of cross-language text classification and clustering. Transfer learning bridged by probabilistic topics obtained via pLSI was proposed by Xue et al. (2008) for the task of cross-domain text categorization. Recently, knowledge transfer for cross-domain learning to rank the answer list of a retrieval task was described by Chen et al. (2010). Takasu (2010) proposes cross-language keyword recommendation using latent topics. Cross-language text clustering and categorization based on the multilingual LDA model was recently proposed by the authors (De Smet and Moens 2009; De Smet et al. 2011). Except for Wang et al. (2009) where the evaluation is vague and unsatisfactory, and relies solely on 30 documents and 7 queries, none of the above works use LDA-based cross-language topics in novel cross-language retrieval models.

The idea to acquire translation candidates based on comparable and unrelated corpora is first tackled in (Rapp 1995, 1999). Over the years, other similar approaches have emerged (Fung and Yee 1998; Diab and Finch 2000; Déjean et al. 2002; Chiao and Zweigenbaum 2002; Gaussier et al. 2004; Fung and Cheung 2004; Morin et al. 2007; Shezaf and Rappoport 2010; Laroche and Langlais 2010). All these methods have examined different representations of word contexts and different methods for matching words across languages, but they all need an initial lexicon of translations, cognates or similar words which are then used to acquire additional translations of the context words, and a bootstrapping procedure is often employed. In contrast, our cross-language extraction methods do not bootstrap on language pairs that share morphology, cognates or similar words and do not use any seed lexicon at all. Haghighi et al. (2008) try to learn bilingual lexicons from unrelated corpora using a generative model with latent concept spaces and orthographic and contextual features. However, orthographic features imply that their method works better for closer language pairs, and a seed lexicon to translate context vectors is still required.

Some attempts of obtaining translations using cross-language topic models have been made in the last few years, but they are model-dependent and do not provide a general environment to adapt and apply other topic models for the task of finding translation correspondences. Ni et al. (2009) have designed a probabilistic topic model that fits Wikipedia data, but they did not use their models to obtain potential translations. Mimno et al. (2009) retrieve a list of potential translations simply by selecting a small number N of the most probable words for topics in both languages and then add the Cartesian product of these sets for every topic to a set of candidate translations. This approach for CLE is straightforward, but it does not catch the structure of the latent topic space completely and is unable to provide semantically related words for low-frequency words. Another model, proposed by Boyd-Graber and Blei (2009), builds topics as distributions over bilingual matchings where matching priors may come from different initial evidences such as a machine-readable dictionary, the edit distance, or the point-wise mutual information statistic scores from available parallel corpora. The main shortcoming is that it introduces external knowledge for matching priors and uses a restricted vocabulary. The authors also admit that their multilingual model suffers from overfitting, since it tends to learn matchings between unrelated words.

None of these works apply the extracted lexicons in a *real-life* problem such as CLIR. To our knowledge, our work is the first real application of any lexicon derived from a latent topic model, and we show its usefulness for CLIR. The usage of multiple semantically related words from the lexicon entries may be observed as a query expansion technique, constructed to improve the effectiveness of a CLIR model. Query expansion techniques relying on a statistical similarity measure among terms stored in an automatically generated thesaurus/lexicon are described by Adriani and Rijsbergen (1999) and Sheridan and Ballerini (1996), but their work differs from ours in both construction of the lexicon and its usage in the CLIR model.

3 Bilingual LDA

3.1 Description of the model

The topic model we use is a bilingual extension of a standard LDA model, called *bilingual LDA* (BiLDA) (Ni et al. 2009; Mimno et al. 2009; De Smet and Moens 2009; Boyd-Graber and Blei 2009). As the name suggests, it is an extension of the basic LDA model, taking into account bilingualism and initially designed for parallel document pairs. We test its performance on a collection of comparable texts where related documents are paired, and therefore share their topics to some extent. Unlike, for instance, Wikipedia articles, where document alignment is established via interlingual links, in some cases it is necessary to

Algorithm 3.1 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 $\psi \sim Dirichlet(\beta)$
for each document pair $d_j = \{d_{jS}, d_{jT}\}$

do
- sample $\theta_j \sim Dirichlet(\alpha)$
- **for each** word position $i \in d_{jS}$
  - **do** sample $z_{ji}^S \sim Multinomial(\theta)$; sample $w_{ji}^S \sim Multinomial(\phi, z_{ji}^S)$
- **for each** word position $i \in d_{jT}$
  - **do** sample $z_{ji}^T \sim Multinomial(\theta)$; sample $w_{ji}^T \sim Multinomial(\psi, z_{ji}^T)$

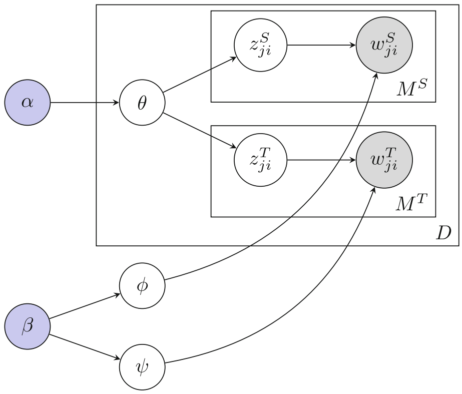

Fig. 1 The standard bilingual LDA (BiLDA) model. M^S and M^T denote lengths of the source document and the target document for each aligned document pair

perform document alignment as the initial step. Our work mainly focuses on Wikipedia data and, by the nature of the Wikipedia structure, articles about the same subject, but in different languages are linked. The cross-lingual pairing of documents is not the subject of the research reported here, but we refer the interested reader to (Utiyama and Isahara 2003; Resnik and Smith 2003; Vu et al. 2009).

BiLDA takes advantage of the document alignment by using a single variable that contains the topic distribution θ. This variable is language-independent, because it is shared by each of the paired bilingual comparable documents. Topics for each document are sampled from θ, from which the words are sampled in conjugation with the vocabulary distribution ϕ (for language S) and ψ (for language T). α and β are the parameters of the uniform conjugate Dirichlet priors[4] for the per-document topic distribution θ and the per-topic word distributions ϕ and ψ, respectively.[5] Algorithm 3.1 summarizes the generative story, while Fig. 1 shows the plate model.

[4] The interested reader may find a very nice introduction to conjugate distributions, conjugate priors and Bayesian networks with application on standard monolingual LDA in (Heinrich 2008).

[5] Following Griffiths et al. (2007), the hyper-parameter α can be interpreted as a prior observation for the number of times a topic is sampled in a document, before having observed any actual words from that document. In an analogical manner, the hyper-parameter β can be interpreted as the prior observation count on the number of times words are sampled from a topic before any observation of actual words. By placing those Dirichlet priors, results are smoothed per-topic word and per-document topic distributions.

Having one common θ for both of the related documents implies parallelism between the texts, which might not always be the case. Still, we later show that the BiLDA model can provide satisfactory results when trained on a comparable corpus such as Wikipedia.

As in the standard monolingual LDA model, we need to set K, i.e., the number of cross-language topics a priori, before training with actual data takes place.

3.2 Output of the model

The described BiLDA model serves as a framework for modeling our statistical retrieval models and methods for automatic lexicon extraction. After the training using Gibbs sampling (Geman and Geman 1984; Steyvers and Griffiths 2007), two sets of probability distributions are obtained for each of the languages. One set consists of per-topic word probability distributions and another set consists of per-document topic probability distributions. For the per-topic word probability distributions ϕ, associated with a source vocabulary W^S, the probability of sampling a new source token $w_i \in W^S$ for an interlingual topic z_k from K different topics can be obtained as follows:

$$P(w_i|z_k) = \phi_{k,i} = \frac{n_k^{(w_i)} + \beta}{\sum_{j=1}^{|W^S|} n_k^{(w_j)} + |W^S|\beta}, \tag{1}$$

where for the word w_i and the topic z_k, $n_k^{(w_i)}$ denotes the total number of times that the topic z_k is assigned to the word w_i from the source vocabulary W^S. The sum $\sum_{j=1}^{|W^S|} n_k^{(w_j)}$ is the total number of words assigned to the topic z_k, and $|W^S|$ is the total number of distinct words in the source vocabulary. The formula for a set of per-topic word probability distributions ψ for the target side of a corpus is computed in an analogical manner.

The second set of probability distributions gives us the distribution of topics for a document. For the BiLDA model, it can be calculated as follows:

$$P(z_k|D_J) = \theta_{J,k} = \frac{n_J^{(k)} + \alpha}{\sum_{j=1}^{K} n_J^{(j)} + K\alpha}, \tag{2}$$

where for a document D_J and a topic z_k, $n_J^{(k)}$ denotes the number of times a word in the document D_J is assigned to the topic z_k.

Due to the fact that the model possesses a fully generative semantics, it is possible to train it on any document-aligned comparable corpus, and later infer it on previously unseen corpora, where inferring a model means calculating the per-document topic distributions θ for the new documents. Once trained on a document-aligned corpus, it is possible to use the topic models for CLIR on any monolingual test collection, even on those which do not have a document-aligned counterpart in the query language.

4 LDA-based Cross-language information retrieval (Act I)

This section provides a theoretical insight into statistical cross-language information retrieval models relying on per-topic word distributions and per-document topic distributions from Sect. 3.2. More retrieval models, which additionally make use of our BiLDA-induced bilingual lexicon, are described in Sect. 6.

4.1 LDA-only CLIR model

Given the set $\{D_1, D_2, \ldots, D_L\}$ of documents in a target language T, and a query Q in a source language S, the task is to rank the documents according to their relevance to the query. We follow the basic approach for using language models in monolingual information retrieval. In the query likelihood model, the *bag-of-words* assumption holds, that is, the terms are independent given the documents and the score of each document is the likelihood of its model generating the query. The probability $P(Q|D_J)$ that the query Q is generated from the document model D_J, is calculated based on the unigram language model:

$$P(Q|D_J) = P(q_1, \ldots, q_m|D_J) = \prod_{i=1}^{m} P(q_i|D_J). \tag{3}$$

The main difference between monolingual IR and CLIR is that documents are not in the same language as the query. Thus, one needs to find a way to efficiently bridge the gap between languages. The common approach is to apply machine-readable translation dictionaries, translate the query and perform monolingual retrieval on the translated query. If a translation resource is absent, one needs to find another solution. In lack of any translation resource, we propose to use sets of per-topic word distributions and per-document topic distributions, assuming the shared space of latent topics. Combining (1) and (2), we can now rewrite Eq. (3) by calculating the probability $P(q_i|D_J)$ in terms of the two BiLDA-related probability distributions:

$$\begin{aligned} P(q_i|D_J) &= (1-\delta_1)\sum_{k=1}^{K} \overbrace{P(q_i|z_k^S)}^{Source\, z_k} \underbrace{P(z_k^T|D_J)}_{Target\, z_k} + \delta_1 P(q_i|Ref) \\ &= (1-\delta_1)\sum_{k=1}^{K} \phi_{k,i}^S \theta_{J,k}^T + \delta_1 P(q_i|Ref) \end{aligned} \tag{4}$$

δ_1 is the interpolation parameter, while $P(q_i|Ref)$ is the maximum likelihood estimate of the query word q_i in a monolingual source language reference collection *Ref*. It gives a non-zero probability for words unobserved during the training of the topic model in case it occurs in the query. Here, we use the observation that latent topics constitute a language-independent space shared between the languages. If that observation holds, it is justified to use the per-topic word distributions for the source language to predict the probability that the word q_i from the query Q will be sampled from the topic z_k^S, and it is justified to use the per-document topic distributions for the target language to predict the probability that the same topic z_k^T (but now in the other language[6]) is assigned to a token in the target document D_J. As mentioned, we may infer the model (learn per-document topic distributions) on any monolingual collection in the source or the target language.

We can now merge all the steps into one coherent process to calculate the probability $P(Q = q_1, q_2, \ldots, q_m|D_J)$, where Q denotes a query in the source language, and D_J denotes a document in the target language. We name this model the **LDA-only model**:

[6] z_k^S and z_k^T basically refer to the same cross-language topic z_k, but one might observe z_k^S as a representation of the cross-language topic given by source language words, and z_k^T a representation given by target language words

1. Infer the trained model on a test corpus in the target language to learn $P(z_k^T|D_J)$ for all target language topics, $k = 1, \ldots, K$, and for all documents in the test corpus.
2. For each word $q_1 \ldots q_m$ in the query, do:
   (a) Compute $P(q_i|z_k^S)$ for all source language topics, $k = 1, \ldots, K$ using (1).
   (b) Sum the products of per-topic word and per-document topic probabilities:

$$P(q_i|D_J) = \sum_{k=1}^{K} P(q_i|z_k^S)P(z_k^T|D_J)$$

3. Compute the whole probability score for the given query and the current document D_J:

$$P(Q|D_J) = \prod_{i=1}^{m} P(q_i|D_J) = \prod_{i=1}^{m} \left((1-\delta_1) \sum_{k=1}^{K} \phi_{k,i}^S \theta_{J,k}^T + \delta_1 P(q_i|Ref) \right) \tag{5}$$

This gives the score for one target language document D_J. Finally, documents are ranked based on their respective scores. If we train a bilingual (or a multilingual) model and wish to reverse the language of queries and the language of documents, the retrieval is performed in an analogical manner after the model is inferred on a desired corpus.

4.2 LDA-unigram CLIR model

The *LDA-only* CLIR model from Sect. 4.1 can be efficiently combined with other models for estimating $P(w|D)$. If we assume that a certain amount of words from the query does not change across languages (e.g. some personal names) and thus could be used as an evidence for cross-language retrieval, the probability $P(q_i|D_J)$ from (3) (where q_i is a query word in the source language, and D_J a document model for a target language document) may be specified by a document model with the Dirichlet smoothing. We adopt smoothing techniques according to evaluations and findings from Zhai and Lafferty (2004). The Dirichlet smoothing acts as a length normalization parameter and penalizes long documents. The model is then:

$$P_{lex}(q_i|D_J) = (1-\delta_2)\left(\frac{N_d}{N_d+\mu}P_{mle}(q_i|D_J) + (1-\frac{N_d}{N_d+\mu})P_{mle}(q_i|Coll)\right) + \delta_2 P(q_i|Ref) \tag{6}$$

where $P_{mle}(q_i|D_J)$ denotes the maximum likelihood estimate of the word q_i in the document D_J, $P_{mle}(q_i|Coll)$ the maximum likelihood estimate in the entire collection in the target language, μ is the Dirichlet prior, and N_d the number of words in the document D_J. δ_2 is another interpolation parameter, and $P(q_i|Ref)$ is the background probability of q_i, calculated over a large corpus. It gives a non-zero probability for words that have zero occurrences in test collections. We name this model the **simple unigram model**.

We can now combine this document model with the *LDA-only* model using linear interpolation and the Jelinek-Mercer smoothing:

$$P(q_i|D_J) = \lambda P_{lex}(q_i|D_J) + (1-\lambda)P_{lda}(q_i|D_J) \tag{7}$$

$$P(q_i|D_j) = \lambda\Bigg((1-\delta_2)\bigg(\frac{N_d}{N_d+\mu}P_{mle}(q_i|D_J) + (1-\frac{N_d}{N_d+\mu})P_{mle}(q_i|Coll)\bigg) + \delta_2 P(q_i|Ref)\Bigg) + (1-\lambda)P_{lda}(q_i|D_J) \quad (8)$$

where P_{lda} is the *LDA-only* model given by (4), P_{lex} the *simple unigram* model given by (6), and λ is the interpolation parameter. We call this model the **LDA-unigram model**.

The combined model presented here is straightforward, since it directly uses words shared across a language pair. One might also use cognates (orthographically similar words) identified, for instance, with the *edit distance* (Navarro 2001) instead of the shared words only. However, both approaches improve retrieval results only for closely related language pairs, where enough shared words and cognates are observed. In the absence of such words and any translation resources, we need to turn to a more general and language-independent method. The next section presents and discusses different methods for automatic cross-language lexicon extraction from the already obtained ϕ and ψ per-topic word distributions, which can further improve the retrieval results, especially for distant language pairs. The lexicon entries will be used to remodel the $P_{lex}(q_i|D_J)$ part of Eq. (7).

5 LDA-based Cross-language lexicon extraction (Intermezzo)

This chapter shows the potential of a cross-language latent topic model to successfully identify translation candidates and semantically related words based on language-specific per-topic word distributions (Eq. 1) and a shared topic space learned during training of the model. The models for identifying translation candidates from probability distributions of a cross-language latent topic model are thoroughly presented and evaluated by Vulić et al. (2011). Based on those results, for the cross-language lexicon extraction within the CLIR task, we opt for the combined **TI+Cue** method, which is computationally feasible since it uses a limited topic space while extracting lexicon entries. The core methods underlying the combined method will be presented shortly.

5.1 Cue method

A straightforward approach (called the **Cue** method) tries to express similarity between two words emphasizing the associative relation between the two words in a natural way. It models the probability $P(w_2|w_1)$, i.e. the probability that a target word w_2 will be generated as a response to a cue source word w_1. For the BiLDA model we can write:

$$Sim(w_1, w_2) = P(w_2|w_1) = \sum_{j=1}^{K} P(w_2|z_j)P(z_j|w_1) \quad (9)$$

Probability $P(w_2|z_j)$ follows directly from the per-topic word distributions, while we still need to find a way to compute conditional topic distributions $P(z_j|w_1)$, which describe a probability that a given word is assigned to a particular topic. If we apply Bayes' rule, we get $P(Z|w) = \frac{P(w|Z)P(Z)}{P(w)}$, where $P(Z)$ and $P(w)$ are prior distributions for topics and words respectively. We assume $P(Z)$ to be a uniform distribution (Griffiths et al. 2007)[7]. $P(w)$ is given by $P(w) = \sum_Z P(w|Z)P(Z)$. For the BiLDA model, we can further write:

[7] By using here a uniform distribution for topics, our model is more general, and is not biased towards the topical distribution of the training corpus.

$$P(z_j|w_1) \propto \frac{P(w_1|z_j)}{Norm_\phi} = \frac{\phi_{j,i}}{Norm_\phi} \tag{10}$$

where $Norm_\phi$ denotes the normalization factor $\sum_{j=1}^{K} P(w_1|z_j)$, the sum of all probabilities ϕ for the currently observed source language word w_i.

We can then calculate the similarity between two words w_1 and w_2 as follows:

$$Sim(w_1, w_2) = P(w_2|w_1) = \sum_{j=1}^{K} \psi_{j,2} \frac{\phi_{j,1}}{Norm_\phi} \tag{11}$$

The conditioning from Eq. (9) automatically compromises between word frequency and semantic relatedness (Griffiths et al. 2007), since higher frequency words tend to have higher probabilities across all topics, but the distribution over topics $P(z_j|w_1)$ ensures that semantically related topics dominate the sum.

5.2 TI method

Another approach borrows an idea from information retrieval and constructs word vectors over a shared latent topic space. Values within vectors are the *TF-ITF* (term frequency–inverse topic frequency) scores which are calculated in a completely analogical manner as the *TF-IDF* scores for the original word-document space (Manning and Schütze 1999). If we are given a source word w_i, $n_{k,S}^{(w_i)}$ denotes the number of times the word w_i is associated with a source topic z_k and refers to the absolute non-smoothed counts after the Gibbs sampling (see "Appendix").

Term frequency (TF) of the source word w_i for the source topic z_k is given as:

$$TF_{i,k} = \frac{n_{k,S}^{(w_i)}}{\sum_{w_j \in W^S} n_{k,S}^{(w_j)}} \tag{12}$$

Inverse topical frequency (ITF) measures the general importance of the source word w_i across all source topics. Rare words are given a higher importance and thus they tend to be more descriptive for a specific topic. The inverse topical frequency for the source word w_i is calculated as[8]:

$$ITF_i = \log \frac{K}{1 + |k : n_{k,S}^{(w_i)} > 0|} \tag{13}$$

The final *TF-ITF* score for the source word w_i and the topic z_k is given by $TF - ITF_{i,k} = TF_{i,k} \cdot ITF_i$. We calculate the *TF-ITF* scores for target words associated with target topics in an analogical manner. Source and target words share the same *K*-dimensional topical space, where *K*-dimensional vectors consisting of the *TF-ITF* scores are built for all words. The standard cosine similarity metric is then used to find the most similar word vectors from the target vocabulary for a source word vector. We name this method the **TI** method. For instance, given a source word w_1 represented by a *K*-dimensional vector SV^1 and a target word w_2 represented by a *K*-dimensional vector TV^2, the similarity between the two words is calculated as follows:

[8] Stronger association with a topic is modeled by setting a higher *threshold* value in $n_{k,S}^{(w_i)} > threshold$, where we have chosen 0.

Table 1 Lists of the top 10 translation candidates (Dutch to English), where the correct translation is not found (column 1), lies hidden lower in the list (2), and is retrieved as the first candidate (3)

(1) **Vlucht** (flight)	(2) **Reclame** (advertisement)	(3) **Munt** (currency)
Airlines	Advertising	Currency
Airline	Advertisements	Currencies
Carriers	Placement	Parities
Overbooked	Advertisers	Fluctuation
Easyjet	Advertisement	Devaluations
Frills	Stereotyping	Euro
Flights	Billboards	Devaluation
Booking	Adverts	Overvalued
Booked	Advert	Peseta
Ryanair	Advertise	Fluctuations

Obtained with the **TI+Cue** method

$$Sim(w_1, w_2) = \cos(w_1, w_2) = \frac{\sum_{k=1}^{K} SV_k^1 \cdot TV_k^2}{\sqrt{\sum_{k=1}^{K} (SV_k^1)^2} \cdot \sqrt{\sum_{k=1}^{K} (TV_k^2)^2}} \tag{14}$$

5.3 Properties of the methods

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 potential relatedness besides a direct co-occurrence. Vulić et al. (2011) have detected that these two methods for automatic extraction of a cross-language lexicon interpret and exploit per-topic word distributions in different ways. Hence, by combining the methods and capturing different evidences, we are able to boost overall scores. The two methods are linearly combined, where the overall score is given by:

$$Sim_{Comb}(w1, w2) = \gamma Sim_{TI}(w1, w2) + (1 - \gamma) Sim_{Cue}(w1, w2) \tag{15}$$

The value of γ is empirically set to 0.1. We have used this combined **TI+Cue** method in all our experiments.

When parallel corpora are not available, standard models for lexicon extraction (i.e., learning translation probabilities) from parallel corpora such as IBM Models (Och and Ney 2003) are then unusable. In that case, extracting the lexicon from comparable corpora using topic models proves to be very useful, since we have detected (Vulić et al. 2011) that our *TI+Cue* method for CLE significantly outperforms similarity-based methods such as cosine similarity with TF-IDF word vectors.[9]

The proposed methods from Sects. 5.1 and 5.2 (and their combination) possess another desirable property. They all generate lists of semantically related words, where synonymy is not the only semantic relation observed. Such lists provide comprehensible and useful contextual information in the target language for the source word, even when the correct translation candidate is missing, as presented in Table 1. We believe that such cross-

[9] As an illustration, on our English-Italian test set, our *TI+Cue* method with 2,000 topics achieves precision of 0.6077 and MRR of 0.6616, while the similarity-based method with TF-IDF vectors and cosine similarity achieves precision of 0.5031 and MRR of 0.5890. For the complete evaluation and results, we refer to Vulić et al. (2011).

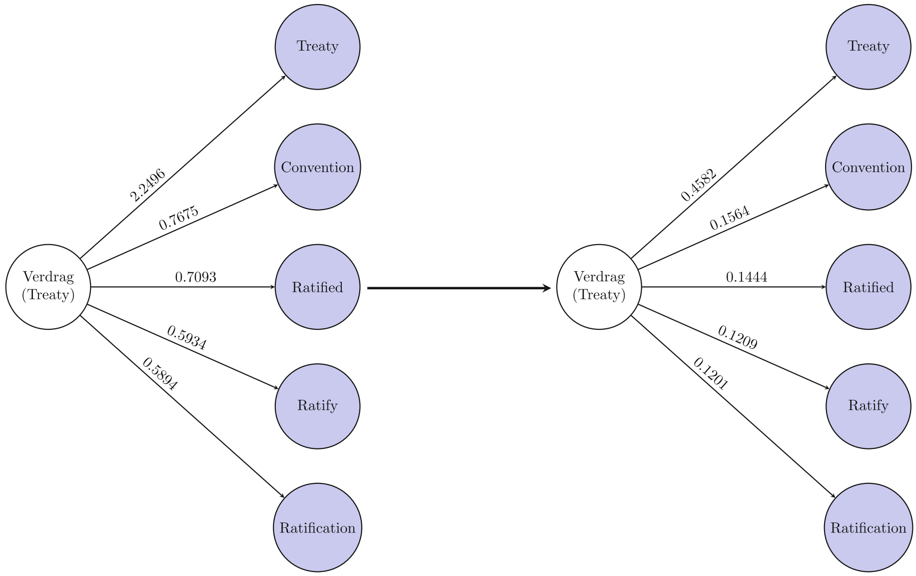

Fig. 2 A lexicon entry example from a Dutch-English lexicon showing top $V = 5$ words from the ranked list. The scores on the edges on the left side are unnormalized TI+Cue scores (the better the score, the closer the semantic relation). The scores on the edges on the left side present normalized probabilities $P(w_i|e_j)$ after Eq. (16) is employed. If we used other values for V, the scores on the edges on the right side would change, again according to Eq. (16)

language semantic relatedness might serve as a useful aid for CLIR, even when the exact translation is absent.

Since all the methods provide ranked lists with scores that measure the strength of cross-language similarity between two words, it is straightforward to convert these ranked lists into a probabilistic lexicon. Each item in the ranked list holds its **TI+Cue** score as shown in Fig. 2. In order to reduce the complexity of calculations and storage, we decided to use only the subset of top V words from the complete ranked list. Probability $P(w_i|e_j)$, which models the degree of association between a source word w_i and a target word e_j found in the subset of the top V words from the ranked list of its lexicon entry, is calculated as follows:

$$P(w_i|e_j) = \frac{Sim_{Comb}(w_i, e_j)}{\sum_{v=1}^{V} Sim_{Comb}(w_i, e_v)} \tag{16}$$

Probabilities $P(w_i|e_j)$ might change depending on the value of V.

6 LDA-based cross-language information retrieval (Act II)

This chapter describes several retrieval models that additionally exploit the knowledge from a BiLDA-induced cross-language lexicon. The construction of the lexicon is described in Sect. 5.

Recall that one lexicon entry for a word w_1 in the source language is in fact a ranked list of words in the target language with their associated scores. The ranked list is a collection of words that are semantically related to the word w_1, based on their respective distributions over cross-lingual topics. While the quality of the lexicon might prove to be inferior for direct translation tasks, it will still be useful for the CLIR task. Since the lexicons are unrelated to features of test collections, we evaluate all the lexicons we obtain (with **TI+Cue**) after we train the topic models with different parameters, and simply use the best ones for all the experiments.

Each lexicon entry is simply transformed into a probabilistic lexicon by normalizing the scores (Eq. 16).[10]

6.1 Lex-only model

The simplest model which uses the knowledge from the lexicon relies on Eq. (6). In case a source word q_i exists in the target language vocabulary[11], Eq. (6) is applied directly. If the word q_i does not exist in the target vocabulary, we need to reach out for the probabilistic lexicon. We closely follow the translation model as presented in Berger and Lafferty (1999) and Xu et al. (2001). If top V words from the lexicon entry are taken into account for retrieval, the probability $P_{lex}(q_i|D_J)$ is then given by:

$$P_{lex}(q_i|D_J) = (1 - \delta_3)\sum_{v=1}^{V} P(q_i|e_v)P(e_v|D_J) + \delta_3 P(q_i|Ref) \tag{17}$$

The summation goes over the top V target words from the ranked list of the lexicon entry for q_i. $P(q_i|e_v)$ is a translation probability for the words q_i and e_v from the lexicon entry calculated by Eq. (16) when only top V words are taken into account, while $P(e_v|D_J)$ can be computed as the first term of Eq. (6) (preceded by $(1 - \delta_2)$ in the equation). $P(q_i|Ref)$ is the background probability, needed in case there is no lexicon entry for the query word q_i. We call this model the **lex-only model**. It uses only evidences from the lexicon combined with the evidence of shared words.

6.2 LDA-lex model

The next model combines the knowledge from the lexicon (the *lex-only* model from the previous section) with the *LDA-only* model given by Eq. (4). The model follows Eq. (7), but instead of the *simple unigram* model utilized to model the probability $P_{lex}(q_i|D_J)$, it uses the *lex-only* model, where $P(q_i|D_J)$ is given by (6) for the words shared across vocabularies, and by (17) for all other words. This model has been named the **LDA-lex model**.

The model is underpinned by several a priori assumptions: (i) if a word occurs in both source and target vocabularies, it is reasonable to assume that the word speaks for itself more than its translations do (for instance, if someone searches for documents related to *Barack Obama*, no translation is needed[12]), (ii) if a word is not shared, one may use a list of

[10] In order to reduce complexity, we use only V top words from the ranked list. Although, it is clear that the amount of semantically related words and the strength of the relatedness varies for every single word, we believe that this fact might be partially captured by the probabilities in the lexicon, since they model "the level of relatedness" between two words in the context of other related words.

[11] Strictly speaking, by the target language vocabulary, we assume the vocabulary used during the BiLDA training, not the vocabulary extracted from the test collection. If we used the latter vocabulary instead, it would be necessary to adjust the vocabulary every time a new document is added to the test collection.

[12] However, that assumption clearly does not hold for the languages that do not share the same alphabet.

semantically related words in the target language, from the lexicon obtained from the per-topic word distributions learned during the training of the topic model. It is convenient, since it uses the same infrastructure as the topic model and does not require any additional resource nor dictionary, (iii) the "LDA-part"[13] of the retrieval model introduces additional topical knowledge, since it connects words in the source language with documents in the target language through the shared space of interlingual topics and groups together words appearing in similar contexts. The modeling of probability $P(q_i|D_J)$ follows these steps:

1. Calculate the probability $P_{lex}(q_i|D_J)$ for a source word q_i and a target document D_J:
   - If the source word q_i is found in the target vocabulary:

$$P_{lex}(q_i|D_J) = (1-\delta_2)\Big(\frac{N_d}{N_d+\mu}P_{mle}(q_i|D_J) + \Big(1-\frac{N_d}{N_d+\mu}\Big)P_{mle}(q_i|Coll)\Big) + \delta_2 P(q_i|Ref) \tag{18}$$

   - If the source word q_i is not found in the target vocabulary, take the top V items from the ranked list of its lexicon entry (if that entry exists) and calculate:

$$P_{lex}(q_i|D_J) = (1-\delta_3)\sum_{v=1}^{V} P(q_i|e_v)P(e_v|D_J) + \delta_3 P(q_i|Ref) \tag{19}$$

where $P(q_i|e_v)$ is calculated using Eq. (16), and $P(e_v|D_J)$ using Eq. (18).

2. Calculate the probability $P_{lda}(q_i|D_J)$:

$$P_{lda}(q_i|D_J) = (1-\delta_1)\sum_{k=1}^{K} \overbrace{P(q_i|z_k^S)}^{Source\ z_k}\underbrace{P(z_k^T|D_J)}_{Target\ z_k} + \delta_1 P(q_i|Ref) \tag{20}$$

3. Combine the calculated probabilities $P_{lex}(q_i|D_J)$ and $P_{lda}(q_i|D_J)$:

$$P(q_i|D_J) = \lambda P_{lex}(q_i|D_J) + (1-\lambda)P_{lda}(q_i|D_J) \tag{21}$$

If we deal with distant languages or languages that do not share the same alphabet, we can treat each word from the query in the source language as the word unobserved in the target vocabulary, and use only Eq. (17) to model the *lexical part* of the retrieval model.

7 Experimental setup

7.1 Training collections

The data used for training of the models is collected from various sources and varies strongly in theme, style and its "comparableness". The only constraint on the training data is the need for an initial document alignment (Sect. 3.1), and it is the only assumption our BiLDA model utilizes during training.

[13] By the "LDA-part" of the retrieval model, we assume the part of the model given by Eq. (4).

The first subset of our training data is the Europarl corpus (Koehn 2005), extracted from proceedings of the European Parliament and consisting of 6,206 documents in English and Dutch. We use only the evidence of document alignment during the training and do not benefit from the "parallelness" of the sentences in the corpus.

Another training subset is collected from Wikipedia *dumps*[14] and consists of paired documents in English and Dutch. Since the articles are written independently and by different authors, rather than being direct translations of each other, there is a considerable amount of divergence between aligned documents. The aligned articles often have a different focus on a subject, which results in different subtopics being addressed. Our Wikipedia training sub-corpus consists of 7,612 documents which vary in length, theme and style, discussing many different subjects including medicine, science, geographic locations, historical figures, industry, political issues, etc.

We removed stop words (429 in English and 110 in Dutch). Our final vocabularies consist of 76,555 words in English, and 71,168 words in Dutch.

7.2 Test collections and queries

We have carried out two conceptually different sets of experiments to evaluate our retrieval models. The first set of experiments tests the performance of the retrieval models on a less difficult task, where a subset of the training collections is used for testing. Another set of experiments has been conducted on test collections that were not used for training beforehand. Here, we deal with a more complex problem; we want to retrieve documents from a monolingual collection, which might be completely topically unrelated to our training collections (e.g. we train the BiLDA model on Wikipedia articles and Europarl documents, infer the BiLDA model on a newswire corpus, and use the BiLDA-based retrieval models on that newswire corpus). Despite the obvious topical disparity, we believe that by having enough training data to cover many different topics, we will be able to learn per-topic word document distributions and infer per-document topic distributions that will lead to quality CLIR models, even for topically unrelated monolingual corpora.

Parameters α and β for the BiLDA training are set to values $50/K$ and 0.01 respectively, where K denotes the number of topics (Steyvers and Griffiths 2007). The Dirichlet parameter μ is set to 2,000 in all models where it is used. Parameters δ_1, δ_2 and δ_3 are all set to negligible values.[15], while we set the interpolation parameter $\lambda = 0.3$ for all experiments, which assigns more weight to the topic model.

7.2.1 Wikipedia as a test collection for the known-item search

Being document-aligned, Wikipedia data might serve as a framework for the initial evaluation of our models in the less difficult task, where test articles have already been observed during the BiLDA training. The idea was to simulate the *cross-language known-item search*, since it provides a precise semantics and thus removes potential issues with defining an exact information need and assigning relevance judgments. The known-item search assumes that only one document is relevant for a specific query. For instance, a

[14] http://dumps.wikimedia.org/.

[15] These parameters contribute to the theoretical soundness of the retrieval models, but, due to the computational complexity, we did not use counts over a large monolingual reference collection. We used a fixed small-value constant in all our models instead, since we detected that it does not have any significant impact on the results.

known-item search in the cross-lingual setting might refer to finding a correct Wikipedia article in the target language with a query provided in the source language.

We did not have the ground truth nor existing queries for this task, so we decided to construct it by adapting the approach from Azzopardi et al. (2007) to the cross-language setting. Their approach has already proven useful for automatic generation of queries for a monolingual known-item search. As the first step, we randomly sample 101 pairs of Wikipedia articles from our training collection. The sampled articles will be regarded as known items we want to retrieve. After that, we generate a known-item query by selecting a document (in this case, the known-item) and constructing a query for that known item. For example, if we have a Wikipedia article pair (A_i^E, A_i^D), where A_i^E denotes the English article and A_i^D its Dutch counterpart, we are able to generate a known-item query from the article A_i^E, and then try to retrieve the article relevant to that query, which is implicitly A_i^D. For producing the automatic known-item queries (for instance, a query in English to retrieve a Dutch article), we have followed these steps:

1. Pick a Dutch article A_i^D for which an English query Q will be generated.
2. Initialize an empty English query $Q = \{\}$ for the current article A_i^D. Query words are extracted from the article A_i^E.
3. Choose the query length L with probability $P(L)$. The query length is drawn from a Poisson distribution, with the mean set to the integer closest to the average length of a query for that language from the CLEF collections in order to construct queries of similar length[16].
4. For each word w_e in the article A_i^E, calculate probability $P(w_e|M_{A_i^E})$, the probability that the word will be sampled from the document model of the article A_i^E. Formally, $P(w_e|M_{A_i^E})$ is a mixture between sampling from the article itself and from the entire collection as given by:

$$P(w_e|M_{A_i^E}) = (1 - \delta_4)P(w_e|A_i^E) + \delta_4 P(w_e|Coll^E) \tag{22}$$

Quality of the query is influenced by the δ_4 parameter which models noise in the sampling process. As δ_4 decreases to zero, the user is able to recall the content of the article in its entirety. Following the same line of thinking, as δ_4 increases to 1, the user knows that the article exists in the collection, but is not able to recollect any of the words relevant to the article. According to Azzopardi et al. (2007), setting $\delta_4 = 0.2$ reflects the average amount of noise within the queries for standard test collections. In order to define $P(w_e|A_i^E)$, the likelihood of selecting the word w_e from the document A_i^E, we have opted for the *Popular + Discrimination Selection* strategy which tries to compromise between *popular words* in a document (we assume that the user tends to use more frequent words as query words) and *discriminative words* for a document (the user considers information outside the scope of a document, and tries to construct a query from such query words that discriminate the particular document from the rest of the collection). The strategy is summarized in the following probability distribution:

$$P(w_e|A_i^E) = \frac{n(w_e, A_i^E) \cdot \log \frac{M}{df(w_e)}}{\sum_{w_j \in A_i^E} \left(n(w_j, A_i^E) \cdot \log \frac{M}{df(w_j)} \right)} \tag{23}$$

M is the number of documents in the entire collection, $n(w_e, A_i^E)$ denotes the number of occurrences of w_e in the article A_i^E, and $df(w_e)$ is the document frequency of w_e.

[16] Due to the fact that the length of the query is drawn from the Poisson distribution, English and Dutch queries for the same article pair are not necessarily of the same length and quality.

Table 2 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)	166,753
	Glasgow Her.95 (EN)	
NC+AD	NRC Hand. 94-95 (NL)	190,604
	Alg. Dagblad 94-95 (NL)	

CLEF Themes (Year: Topic Nr.)	# Quer.	Used for
(b) Statistics of used queries		
NL ’01: 41-90	47	LAT
NL ’02: 91-140	42	LAT
NL ’03: 141-200	53	LAT+GH
EN ’01: 41-90	50	NC+AD
EN ’02: 91-140	50	NC+AD
EN ’03: 141-200	56	NC+AD

5. Rank all words from the document A_i^E based on the scores obtained after employing (22) and (23).
6. Take the top L words from the ranked list as the query words of the known-item query for the article A_i^E.[17]

We perform this automatic query generation for all 101 article pairs in both directions, designing 101 Dutch queries to retrieve English documents and vice versa. For instance, for a Dutch article discussing *halfwaardebreedte (full width at half maximum)*, a query in English is $Q = \{width, hyperbolic, variable, deviation\}$.

7.2.2 CLEF test collections

Our experiments have been carried out on three data sets taken from the CLEF 2001-2003 CLIR campaigns: the LA Times 1994 (**LAT**), the LA Times 1994 and the Glasgow Herald 1995 (**LAT+GH**) in English, and the NRC Handelsblad 94-95 and the Algemeen Dagblad 94-95 (**NC+AD**) in Dutch.

We extracted queries from the *title* and *description* fields of all CLEF themes[18] for each year. Queries without relevant documents were removed from the query sets. Table 2a shows statistics of the CLEF collections, while Table 2b shows statistics of the queries used for testing.

8 Results and discussion

This section reports our experimental results for two main tasks: (i) retrieval models have been tested within a lenient experimental setup, where the goal is to perform a cross-

[17] More precisely, we have constructed the known-item query in the source language for the target language article A_i^D which is document-aligned to the article A_i^E.

[18] In order to avoid confusion, we use the term “topics” when speaking about latent variables of the BiLDA model, and “themes” when referring to the CLEF data.

language known-item search over Wikipedia articles (English queries, Dutch articles and vice versa) that have already been used to train the topic model, (ii) retrieval models have been tested with CLEF test collections for the tasks of English-Dutch and Dutch-English cross-language information retrieval. The cross-language topic model is trained just once on a large bilingual training corpus. It may then be used for both tasks[19].

First, we describe our training settings and evaluate lexicons obtained from per-topic word distributions of the BiLDA model trained with different parameters in order to choose the best lexicon for CLIR models. Second, we test our retrieval models from Sects. 4 and 6 in the known-item search of Wikipedia articles and report our findings. As the next step, we carry out different experiments for English-Dutch and Dutch-English cross-language information retrieval: (1) we compare our *LDA-only* to one baseline that has also tried to exploit latent topic spaces for CLIR (standard monolingual LDA trained on concatenated paired documents as described by Roth and Klakow 2010), and we also compare it to the *simple unigram* model from Sect. 4.2. We want to prove the soundness and the utility of the *LDA-only* model and, consequently, other models that later build upon the foundation established by the *LDA-only* model (*LDA-unigram* and *LDA-lex*), (2) we provide an extensive evaluation over all CLEF test collections with all BiLDA-based models (*LDA-only*, *LDA-unigram* and *LDA-lex*), (3) we compare our LDA-based models with similar models for monolingual retrieval (queries and documents in the same language) and a model that uses *Google Translate* tool to translate query words and then performs monolingual retrieval, and measure the decrease of performance for CLIR, (4) we also compare the best scoring combined *LDA-lex* model with the *lex-only* model that uses only evidences of the shared words and knowledge from the extracted lexicon, and, as the final step, (5) we compare results for all test collections when the BiLDA model is trained on different types of training data (parallel, comparable and combined) and show that comparable data boost retrieval performance.

We have trained our BiLDA model with different number of topics (400, 1,000 and 2,200) on the combined **EP+Wiki** corpus. Additionally, for the purpose of comparing retrieval performance when the BiLDA model is trained on different corpora, we have also trained the BiLDA model with $K = 1{,}000$ topics[20] on two different subsets of training corpora: (1) the parallel Europarl corpus (**EP**)[21], and (2) the comparable Wikipedia corpus (**Wiki**).

We have also empirically detected that the optimal value for V is 10,[22] so we have used the top 10 items from the ranked list for each lexicon entry in all experiments with the *lex-only* and the *LDA-lex* model.

[19] We must infer it on the appropriate test collection for the task (ii). Since we train our model on heterogeneous, out-of-domain corpora, topically unrelated to the test collections, results vary over different collections.

[20] Results with 400 and 2,200 topics are comparable and lead to the same conclusions as with $K = 1{,}000$.

[21] As mentioned, we never exploit the fact that Europarl is sentence-aligned. We use only knowledge of document alignments and nothing else beyond that.

[22] We have experimented with different values, $V = 1, 3, 5, 10, 20$, and have empirically detected that $V = 10$ displays the best results overall, although variations when using other values for V are in most cases minimal.

Table 3 Recall@1, MRR scores and the number of detected words for the test subset of 711 English query words and 724 Dutch query words

K	EN to NL Lexicon			NL to EN Lexicon		
	Recall@1	MRR	Detected	Recall@1	MRR	Detected
400	0.2124	0.2803	0.4121	0.1740	0.2467	0.3840
1,000	0.2869	0.3506	0.4444	0.2500	0.3141	0.4268
2,200	0.3263	0.3920	0.4867	0.2652	0.3338	0.4558

Extraction method is **TI+Cue**

8.1 Lexicon extraction and evaluation

8.1.1 Evaluation settings and results

Following the intuition that more data lead to better per-topic word distributions, we have decided to compare our lexicons extracted from the BiLDA model trained on the *EP+Wiki* corpus. Following the results from Vulić et al. (2011), we have used the **TI+Cue** lexicon extraction method.

As test sets, we use the set of words appearing in English and Dutch queries extracted from the CLEF themes. These sets form a true representation of a general vocabulary, since they contain high-frequency, medium-frequency and low-frequency words. Our test sets consist of 711 English words and 724 Dutch words. Our evaluation relies on Recall@1 scores[23] (the percentage of words where the first word from the list of translations is the correct one) and Mean Reciprocal Rank (MRR) scores as given by:

$$MRR = \frac{1}{V} \sum_{w \in E_V} \frac{1}{rank_w} \tag{24}$$

where E_V denotes the top V words from the ranked list of a lexicon entry. V is set to 10 for all lexicons, since we use the same V in further evaluations of our CLIR models.

We compare our candidates against translation candidates acquired by the *Google Translate* tool, which serves as the ground truth for evaluations, although one should be aware that *Google Translate* does not always necessarily return the best or even correct translation (Dolamic and Savoy 2010). We also provide the percentage of translations detected in the ranked lists (Recall@10)[23]. Results are presented in Table 3.

8.1.2 Discussion

As we can see in Table 3, lexicon entries are far from perfect, but we believe that lexicons will be useful for CLIR since, even when a correct translation is not found, other words from the ranked list carry enough semantics of the input word. By expanding a word from the query with the words from the list, we should be able to retrieve documents that cannot be entirely captured with the *simple unigram* or the *LDA-only* model. For instance, a correct translation for a Dutch word *zender (transmitter, sender)* is not found, but the first

[23] Since we assume that only one translation is correct, Recall@1 in this case is the same as Precision@1. On the other hand, Recall@10 is not the same as Precision@10. Assuming that only one translation is correct, Recall@10 counts how many times the correct translation appeared in the list of top 10 candidates for each word, and that is exactly the measure we need.

Table 4 Recall@1 and Recall@5 scores of the *simple unigram* model and the *lex-only* model for both search directions and all Wikipedia queries

Recall	EN queries, NL documents		NL queries, EN documents	
	Simple uni	Lex-only	Simple uni	Lex-only
@1	0.4059	0.5201	0.4851	0.6304
@5	0.5250	0.6104	0.5840	0.7230

five words in the list of related English words are *(radio, broadcast, broadcasting, television, broadcaster)*. This group of words can certainly help retrieving the correct documents associated with the query word *zender*, even when the correct translation is missing.

For the next experiments we have decided to use the best scoring English-Dutch and Dutch-English lexicons extracted from a model with 2,200 topics, based on the results from Table 3.[24]

8.2 Cross-language known-item search for Wikipedia articles

8.2.1 Experimental setup and results

The cross-language known-item search has been carried out for 101 pairs of Wikipedia articles randomly sampled from 7, 612 pairs of the English-Dutch Wikipedia training corpus. Experiments have been conducted for both possible retrieval directions (English to Dutch and Dutch to English). The BiLDA model was trained on the *EP+Wiki* corpus. To make the search a bit more difficult, we have also included the Europarl documents in the search space. Our search space then consisted of 13, 818 documents from all training document pairs. Scores for the *simple unigram* model and the *lex-only* model are given in Table 4. We report Recall@1 (the only relevant document is retrieved as the first in the list) and Recall@5 (the only relevant document is retrieved among the top 5 retrieved documents) scores of our BiLDA-based models for both search directions in Tables 5 and 6.

8.2.2 Discussion

We have drawn several conclusions based on the results presented in Sect. 8.2.1:

- Table 4 reveals that adding lexicon entries significantly helps in improving overall performance. However, these results are still much lower than results obtained by combining shared words and lexicon entries with the "LDA-based" part from the *LDA-only* model.
- The *LDA-only* model is outperformed by the *LDA-unigram* and the *LDA-lex* model which exploit more different evidences and try to use them in document modeling for retrieval. We conclude that the combination of translation evidences leads to better retrieval models, even when the evidences are not completely disjunct. That conclusion will be more firmly supported by later experiments on the CLEF collections.

[24] We have chosen the best scoring lexicon, although a real-life setting might exhibit the situation where only trained models are available, and training data is absent. All one has in possession are per-topic word distributions from the trained topic models that could be used for retrieval. We are then forced to extract the lexicon from these distributions.

Table 5 Recall@1 and Recall@5 scores of our LDA-based models for all 101 English queries and Dutch documents

K	Recall@1			Recall@5		
	LDA-only	LDA-uni	LDA-lex	LDA-only	LDA-uni	LDA-lex
400	0.1980	0.6728	0.6676	0.3960	0.7920	0.7920
1,000	0.4059	0.7237	0.7465	0.6735	0.8220	0.8515
2,200	0.4653	0.7573	**0.7865**	0.7725	0.9010	**0.9405**

Table 6 Recall@1 and Recall@5 scores of our LDA-based models for all 101 Dutch queries and English documents

K	Recall@1			Recall@5		
	LDA-only	LDA-uni	LDA-lex	LDA-only	LDA-uni	LDA-lex
400	0.3762	0.7821	0.7695	0.6040	0.8615	0.8810
1,000	0.5347	0.7803	0.7695	0.7920	0.9210	0.9305
2,200	0.5941	0.8405	**0.8408**	0.8615	0.9605	**0.9800**

- *LDA-unigram* and *LDA-lex* display comparable results, with a slight advantage for the lexicon-based method. The observation is explained if we investigate the structure of the query. Many Wikipedia articles describe people, toponyms or specific concepts where many words are shared between the Dutch and English vocabularies. In that setting, a lexicon helps to a lesser extent.
- We have successfully applied a method from Azzopardi et al. (2007) to automatically generate queries for known-item search and we have adapted it to a cross-language setting. Moreover, Azzopardi et al. (2007) assert that their method still suffers from the insufficient *replicative validity* and *predictive validity* (i.e., an automatically generated query should really behave as a query generated from the user, and retrieved articles should be similar in both cases). Using a thorough evaluation, they claim that automatically generated queries lead to lower retrieval scores, which leads to conclusion that the results with *real-life* manual queries might be even higher than presented in the tables.

8.3 Comparison of the LDA-only model with baseline systems

From now on, all experiments will be conducted on standard CLEF test collections. The main evaluation measure we use for all further experiments is the *mean average precision* (MAP).

The *LDA-only* model serves as the backbone of the two more advanced BiLDA-based document models (*LDA-unigram* and *LDA-lex*). Since we want to make sure that the *LDA-only* model constructs a firm and sound language-independent foundation for building more complex retrieval models, we compare it to another system which tries to build a CLIR system based around the idea of latent concept topics: the standard LDA model trained on the merged document pairs. We also compare the *LDA-only* model to the *simple unigram* model to make sure that our complex models do not draw its performance mainly from the shared words.

Table 7 MAP scores on all CLEF test collections for the LDA-unigram and the LDA-lex retrieval models, where BiLDA was trained with different number of topics (400, 1,000, 2,200)

Queries\K	LDA-only			LDA-uni			LDA-lex		
	400	1,000	2,200	400	1,000	2,200	400	1,000	2,200
NL 2001	0.1777	0.1969	0.2028	0.2330	0.2673	0.2813	**0.2995**	0.2943	0.2973
NL 2002	0.1117	0.1396	0.1371	0.2093	0.2253	0.2206	**0.2419**	0.2255	0.2241
NL 2003	0.0781	0.1227	0.0784	0.1608	0.1990	0.1658	0.2055	**0.2083**	0.1813
EN 2001	0.1270	0.1453	0.1624	0.2204	0.2275	0.2398	0.2294	0.2370	**0.2427**
EN 2002	0.0932	0.1374	0.1412	0.2455	0.2683	0.2665	0.2712	**0.2866**	0.2782
EN 2003	0.0984	0.1713	0.1529	0.2393	0.2783	0.2450	0.2388	**0.2784**	0.2499

Training corpus is EP+Wiki

Bold values denote the best MAP scores for the corresponding campaigns

We have trained the standard LDA model on the combined *EP+Wiki* corpus with 400 and 1,000 topics and compared the retrieval scores with our *LDA-only* model which uses the BiLDA model with the same number of topics. The *LDA-only* model outscores this model by a huge margin. The MAP scores for standard LDA are very low, and vary between the MAP of 0.01 and 0.03 for all experiments, which is significantly worse than the results of the *LDA-only* model as seen in Table 7. A problem with this baseline method might be in concatenation of document pairs, since one language might dominate the merged document. On the other hand, BiLDA keeps the structure of the original document space intact.

The MAP scores of the *simple unigram* model for NL 2001, NL 2002, and NL 2003 are 0.0274, 0.0343, and 0.0292, respectively, while the MAP scores for EN 2001, EN 2002, and EN 2003 are 0.0643, 0.1030, and 0.0827,[25] respectively. Comparison of precision-recall curves for the *LDA-only* model and the *simple unigram* model is presented in Figs. 5a, b.

Following these results, we are justified to use BiLDA in other retrieval models.

8.4 Comparison of LDA-only, LDA-unigram and LDA-lex

In this subsection, the idea was to compare three retrieval models that rely on per-document topic distributions of the BiLDA model, once it is inferred on test corpora. Besides that, we wanted to test whether the knowledge from shared words (as in the *LDA-unigram* model, and the knowledge of the shared words combined with the knowledge from BiLDA-induced lexicons (as in the *LDA-lex* model) positively affect retrieval.

8.4.1 Comparison of models with a fixed number of topics (K = 1,000)

The *LDA-only* model, the *LDA-unigram* model and the *LDA-lex* model have been evaluated on all test collections, with the number of topics initially fixed to 1,000. Figure 3a shows the precision-recall values obtained by applying all three models to English test

[25] As also presented in Sect. 8.4.3, the reason why the scores of the *simple unigram* model for English queries and Dutch documents are generally higher lies in the fact that more English words are present in Dutch documents than vice versa, so the *simple unigram* model finds more evidences of shared words in that retrieval direction.

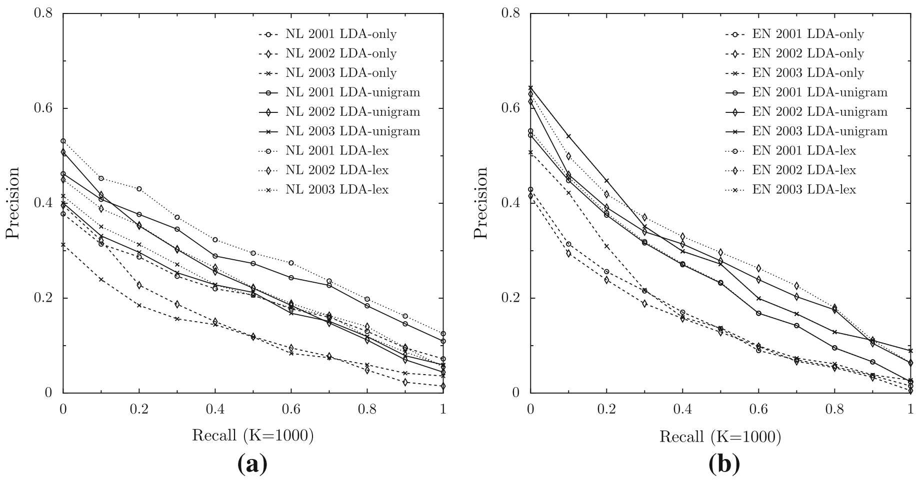

Fig. 3 Precision-recall for LDA-only, LDA-unigram and LDA-lex for both retrieval directions. K = 1,000. Training corpus is EP+Wiki. **a** NL queries, EN test collections. **b** EN queries, NL test collections

collections with Dutch queries, while Fig. 3b shows the precision-recall values for Dutch test collections and English queries.

As the corresponding figures show, the *LDA-only* model seems to be too coarse to be used as the only component of an IR model (e.g., due to its limited number of topics, words in queries unobserved during training). However, combining it with words shared across languages and lexicon entries from BiLDA-induced lexicons leads to a drastic increase in results. Results of the *LDA-lex* model which scores better than the *LDA-unigram* model seem especially promising. The *LDA-unigram* relies solely on shared words, which clearly makes it language-biased, since its performance relies heavily on the amount of shared words (or the degree of closeness between two languages). On the other hand, the *LDA-lex* has been envisioned for CLIR between distant language pairs.

8.4.2 *Varying the number of topics*

The main goal of the next set of experiments was to test the importance of the lexicon, and the behavior of our two best models if we vary the number of topics in BiLDA training. We have carried out experiments with the CLIR models relying on BiLDA trained with different numbers of topics (400, 1,000 and 2,200). The MAP scores of the *LDA-unigram* and the *LDA-lex* model for all campaigns are presented in Table 7, while Fig. 4 shows the associated precision-recall values.

8.4.3 *Discussion*

We observe several interesting phenomena from Table 7 and Fig. 4:

- The *LDA-lex* model obtains the best scores for all test collections which proves the intuition that additional evidences from lexicon entries will improve the retrieval scores and lead to a better model.
- The margins between scores of the *LDA-unigram* and the *LDA-lex* model are generally higher for campaigns with Dutch queries. It becomes even more interesting if we look

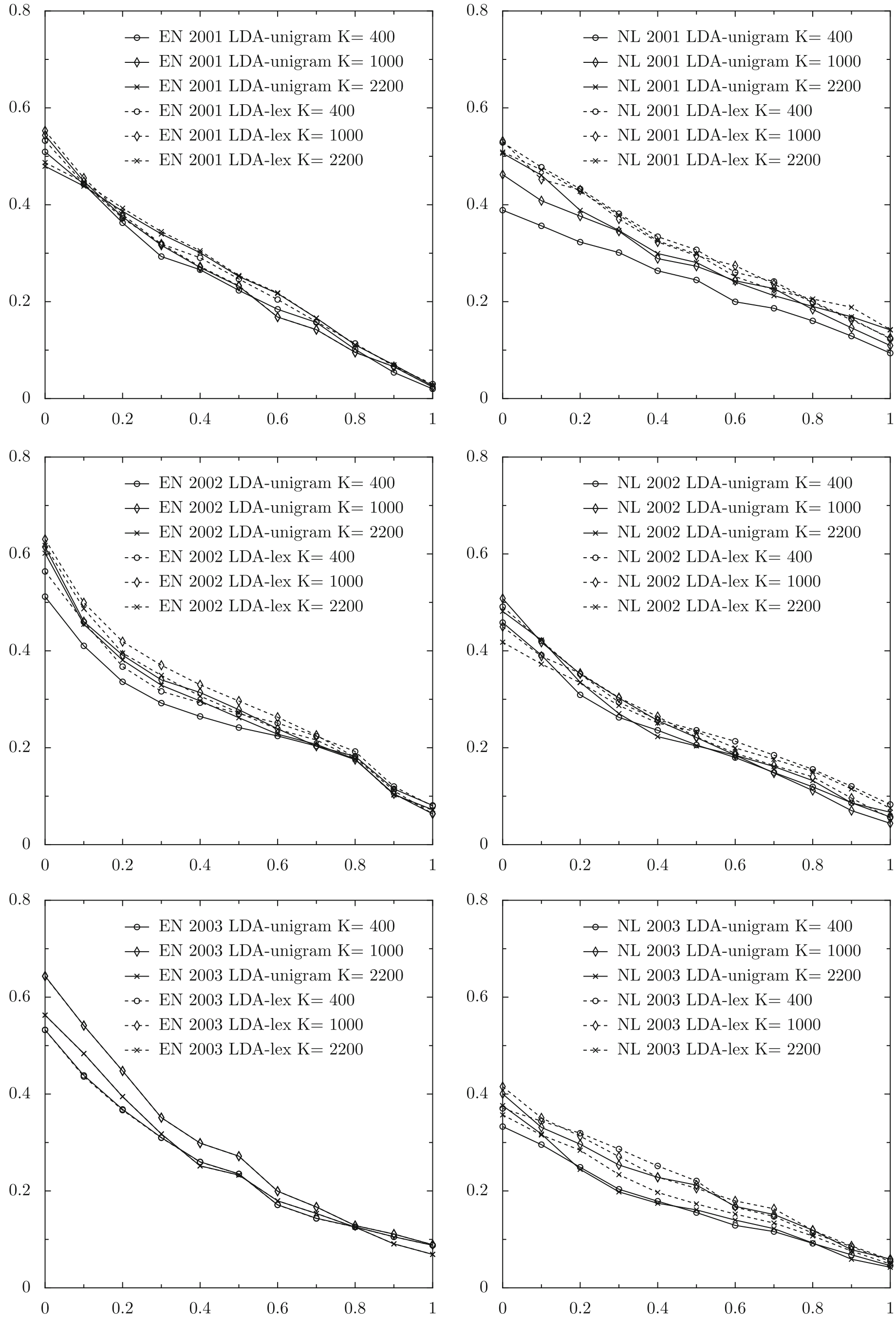

Fig. 4 Precision-recall for the LDA-unigram model and the LDA-lex model for all test collections. Training corpus is EP+Wiki

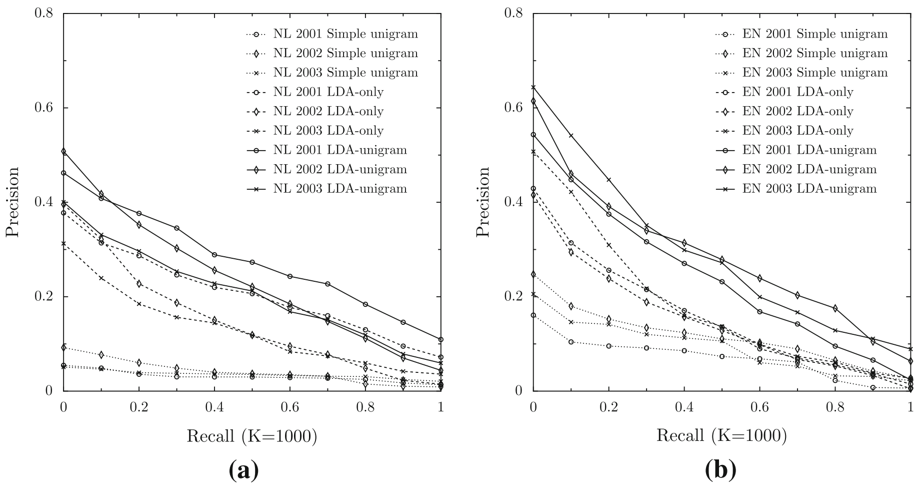

Fig. 5 Comparison of the simple unigram model, LDA-only model and LDA-unigram as their combination. K = 1,000. Training corpus is EP+Wiki. **a** NL queries, EN test collections. **b** EN queries, NL test collections

back at the results of the lexicons evaluation from Table 4, where the Dutch-English lexicons scored worse. The reason why the Dutch-English lexicon helps more might be in the fact that much more English words are observed in our Dutch vocabulary than vice versa. If that is the case, than the knowledge from the lexicon is used less frequently, and the *LDA-lex* model relies more on shared words, which brings it closer to the *LDA-unigram* model. On the other hand, less Dutch words are observed in the English vocabulary, and one needs to turn to the evidences from the lexicon entries more often. In order to support this intuition which explains the results from Fig. 4, we have computed the average percentage of shared words in both English and Dutch queries. The average percentage of shared words is 55.6 % per English query, and only 18.9 % per Dutch query.[26]

- Due to a high percentage of shared words, especially per English query (see the previous item), it may be possible that the *LDA-unigram* model draws its performance mainly from the part specified by the *simple unigram* model. However, as presented in Fig. 5a, b, that possibility has been denied, and the final *LDA-unigram* model clearly works as a positive synergy between the two simpler models, where *LDA-only* is more important for the overall performance of the combined model.
- The margins between scores of the *LDA-unigram* and the *LDA-lex* model are generally higher for the lower number of topics in campaigns with Dutch queries, where the lexicons are used more extensively. With less topics, per-topic word distributions and per-document topic distributions are too coarse, so more cross-language evidence comes from the lexicon itself. By increasing the number of topics, these distributions become more fine-grained, and more and more evidences that initially came from the lexicon, are now captured by the "LDA-part" of the *LDA-lex* model. We have

[26] This difference in percentage of shared words comes mostly from the English terms such as named entities that are often used in parallel with Dutch terms in Dutch news texts. For instance, when a Dutch news article discusses the London or New York Stock Exchange, it will use the exact English term, while an English article, of course, will not include the Dutch translation.

Table 8 MAP scores on all CLEF test collections for MLDA-only, MLDA-unigram, GT+MLDA-only and GT+MLDA-unigram

Model \queries	NL 2001	NL 2002	NL 2003	EN 2001	EN 2002	EN 2003
MLDA-only	0.2796	0.2163	0.2413	0.1315	0.1429	0.1300
MLDA-unigram	0.3993	0.3358	0.3786	0.2603	0.2891	0.3262
GT+MLDA-only	0.1856	0.1848	0.2261	0.1253	0.1148	0.1162
GT+MLDA-unigram	0.3066	0.2752	0.3481	0.2296	0.2401	0.2443

Standard monolingual LDA trained on monolingual English and Dutch data. EP + Wiki. K = 1,000

encountered overlaps of the evidences which lead to similar retrieval scores. The scores obtained by the *LDA-lex* model are still higher than the scores of the *LDA-unigram*, since some of the evidences can be found only in the lexicon, regardless of the number of topics.

- Although a larger number of topics should intuitively lead to a more fine-grained model with better per-topic word distributions and per-document topic distributions and, consequently, to a better retrieval model, this is clearly not the case. If we set the number of topics to a value too high, the topics will become less informative and descriptive as the evidences tend to disperse over all topics.[27]. One of the main disadvantages of the BiLDA model is the need to define and fix the number of topics before its training takes place. It does not have the ability to dynamically redefine the number of topics to adjust the training data in an optimal way.

8.5 Comparison with monolingual LDA-based models and LDA-based models that use an external translation resource

With this set of experiments, we investigated how efficient our LDA-based translation process actually is. Thus, we decided to compare our LDA-based models already evaluated in Sect. 8.4 with another four models: (1) a model that performs monolingual retrieval in the same fashion as our CLIR *LDA-only* model (**MLDA-only**), (2) a model that performs monolingual retrieval in the same fashion as our CLIR *LDA-unigram* model, as presented by Wei and Croft (2006) (**MLDA-unigram**), (3) a model that uses *Google Translate* to perform word-to-word translation of query words, and then performs monolingual retrieval using *MLDA-only* (**GT+MLDA-only**), (4) a model that uses *Google Translate* in the same way, and then employs the monolingual *MLDA-unigram* (**GT+MLDA-unigram**). In order to use these models, we have trained standard monolingual LDA with K = 1,000 topics for both English and Dutch side of our training corpora. MAP scores for these models are presented in Table 8,[28] while MAP scores for our CLIR models have already been presented in Table 7.

[27] In the most extreme case, each word could be a topic on its own, but how informative is that? And what can we learn from it?

[28] In order to remain consistent throughout the text, we did not change the naming conventions for the queries and document collections in this table. However, when dealing with the *MLDA-only* and *MLDA-unigram* models, e.g. NL 2001 actually means–English queries (instead of Dutch queries as for CLIR.) to retrieve English documents. We are then allowed to compare results of all the models (both monolingual and CLIR) for the NL 2001 campaign. The same goes for all other campaigns.

By examining the results in Tables 7 and 8, we derive several conclusions:

- As expected, the monolingual *MLDA-unigram* model outperforms our CLIR models, although the difference in scores is much more noticeable when performing monolingual retrieval in English. We attribute that observation to the quality of our training data. The English side of our Wikipedia data contains more information and articles of a higher quality, which altogether leads to latent topics of a better quality (so, although we deal with a shared topical space, our English topics are of a better quality than our Dutch topics), which then again leads to better statistical LDA-based retrieval models. While MAP scores for the *MLDA-only* model for Dutch are pretty similar to the scores of *LDA-only* when using English queries, *MLDA-only* for English scores much better than *LDA-only* with Dutch queries. The general problem is thus in the quality of the topics learned by BiLDA, and the intuition is that data of a higher quality should lead to the latent topics of a higher quality.
- Low results for *MLDA-only* for monolingual Dutch retrieval when we train standard LDA on monolingual data also refer to the fact that the Dutch side of our training corpus is of a lesser quality.
- A significant drop in performance for both retrieval directions is marked when we use *Google Translate* to translate words from queries and then perform monolingual retrieval. One-to-one word translation is clearly not always the best translational choice, and *Google Translate* might also introduce some errors in the translation process. That conclusion underpins the conclusions drawn by Dolamic and Savoy (2010).
- Our combined CLIR models outperform *GT+MLDA-unigram* for English queries and Dutch text collections. One of the reasons for that phenomenon might again be errors in the translation process performed by *Google Translate*. Moreover, many words from English queries are also found in Dutch documents, and our *LDA-unigram* and *LDA-lex* models are able to capture that tendency.
- For almost all CLEF campaigns, our *LDA-unigram* and *LDA-lex* models display performance that is comparable with or even better than performance of the *GT+MLDA-unigram* model, a model that uses knowledge from a dictionary to directly translate queries. Our models thus become extremely important for language pairs where such a dictionary or translation system is not available or of a low quality.

8.6 Comparison of Lex-only and LDA-lex

8.6.1 Motivation for comparison and results

We also want to compare our best scoring *LDA-lex* model that blends evidences from lexicons and shared words, and evidences from probability distributions of BiLDA, with the *lex-only* model which uses only the shared words and the lexicon knowledge as evidences. We have already proved that the combined, evidence-rich model yields better scores than the *LDA-only* that uses only evidences in the form of per-topic word and per-document topic distributions. We now want to prove that it also scores better than the more straightforward *lex-only* model that uses only *lexical* evidences. MAP scores for the *lex-only* model are 0.1998, 0.1810 and 0.1513 for NL 2001, NL 2002 and NL 2003 (Dutch queries, English documents), respectively, and 0.1412, 0.1378 and 0.1196 for EN 2001, EN 2002 and EN 2003 (English queries, Dutch documents). The best MAP scores for *LDA-lex* are given in Table 7. Figure 6a shows the comparison of the associated precision-recall

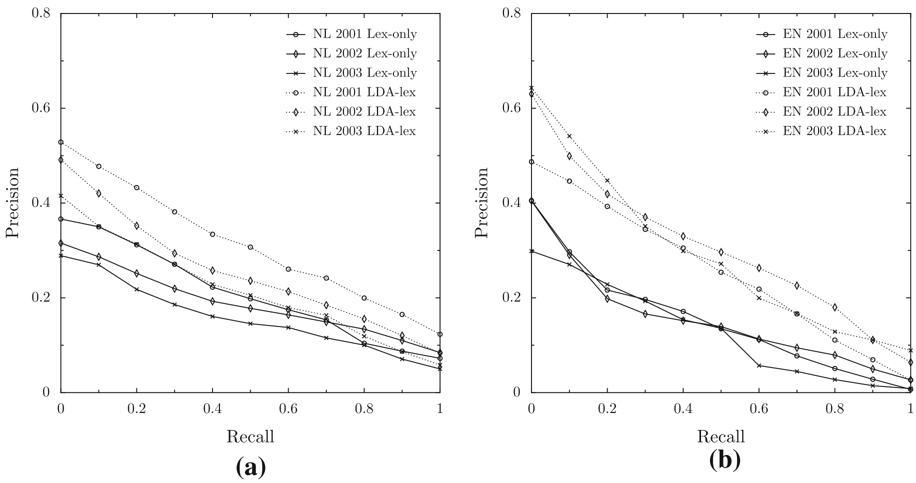

Fig. 6 Comparison of the precision-recall values for the lex-only model and the LDA-lex model. K = 400 or K = 1,000 for LDA-lex. Training corpus is EP + Wiki. **a** NL queries, EN test collections. **b** EN queries, NL test collections

diagrams for all English collections (with queries in Dutch), and Fig. 6b shows the comparison for all Dutch collections (with queries in English).

8.6.2 Discussion

Figure 3a and b have already shown the superiority of the *LDA-lex* model over the *LDA-only* model. The results in this subsection again show the superiority of the *LDA-lex* model over the *Lex-only* model. The fact that the *LDA-lex* model combines the evidences from the other two models makes it the strongest model. The other two models utilize only subsets of the available evidences which makes them more error-prone. For instance, if the semantics of a word from a query is not captured by the "LDA-part" (as in the *LDA-only* model), that model is unable to retrieve any documents strongly related to that word. On the other hand, if the same problem occurs for the *LDA-lex* model, it still has a possibility to look up for an aid in the lexicon. Additionally, if a document scores good for more than one evidence, it strengthens the belief that the document might be relevant for the query.

8.7 Training with different types of corpora

8.7.1 Motivation for comparison and results

In the final set of experiments with CLEF data, we measure the performance of our topic models trained on three different types of corpora (EP, Wiki, EP+Wiki) with K = 1,000 topics. We wanted to find out if and how Wikipedia training data help the retrieval. Moreover, we wanted to test our *"the more the merrier"* assumption that more training data lead to better probability distributions in the BiLDA model and, following that, better retrieval models. The *LDA-unigram* model and the *LDA-lex* model have been used for all the experiments. Table 9 shows the MAP scores over all CLEF test collections.

Table 9 MAP scores on CLEF test collections for the LDA-unigram and the LDA-lex retrieval models, where BiLDA was trained on different corpora (EP, Wiki, and EP+Wiki) K = 1,000

Queries	LDA-uni			LDA-lex		
	EP	Wiki	EP+Wiki	EP	Wiki	EP+Wiki
NL 2001	0.2590	0.1798	0.2673	0.2897	0.2800	**0.2943**
NL 2002	0.1788	0.1794	0.2253	0.2085	0.1993	**0.2255**
NL 2003	0.1813	0.1247	0.1990	0.2061	0.1896	**0.2083**
EN 2001	0.2285	0.1483	0.2275	0.2284	0.1512	**0.2370**
EN 2002	0.2373	0.2176	0.2683	0.2401	0.2322	**0.2866**
EN 2003	0.2398	0.1924	0.2783	0.2401	0.1957	**0.2784**

Bold values denote the best MAP scores for the corresponding campaigns

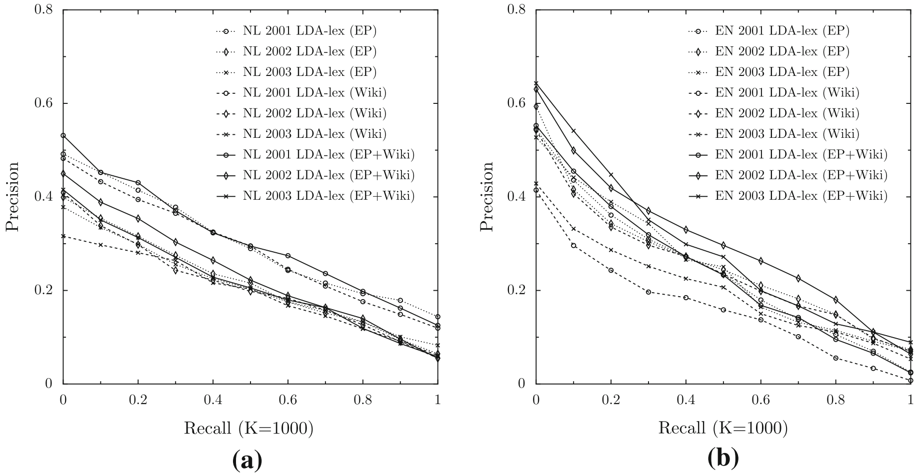

Fig. 7 Comparison of the precision-recall values for LDA-lex, where BiLDA was trained on different corpora (EP, Wiki and EP+Wiki). K = 1,000. **a** NL queries, EN test collections. **b** EN queries, NL test collections

Figure 7a shows precision-recall values of the *LDA-lex* for Dutch queries and English documents, and Fig. 7b for English queries.

8.7.2 Discussion

The results lead us to several conclusions:

- They show that the comparable out-of-domain Wikipedia data can be used to train the bilingual (parallel) LDA model (BiLDA) and reasonable CLIR results can still be expected. For some experiments (EN 2002, NL 2001 and NL 2002), the results with the model trained on the Wikipedia data are comparable to the results with the model trained on the parallel document-aligned Europarl corpus (especially for the *LDA-lex* model, when the lexicon knowledge is employed). For these campaigns we observe major improvement when the BiLDA model is trained on the combined corpus. On the other hand, some experiments where retrieval models rely on the BiLDA model trained solely on Wikipedia have led to much worse scores than the scores of models relying on BiLDA trained on Europarl (e.g. NL 2003, EN 2001). For these experiments, we don't observe the major improvement after we enrich our training data with Wikipedia data. However, we believe that extracting more Wikipedia articles for training data might solve this problem.

- Table 9 also reveals that accumulating more training data by adding comparable Wikipedia documents to a parallel corpus is definitely not harmful, and in most cases increases the quality of topic models which also leads to a better quality of the proposed CLIR systems.
- More training data lead to better per-topic word and per-document topic distributions and, consequently, to better CLIR models. The best results for all test collections are obtained by the BiLDA model trained on the combined *EP+Wiki* corpus.
- These initial experiments also reveal a clear advantage of using our automatically-extracted lexicons, since the *LDA-lex* model, which uses the lexicon, displays better results than *LDA-unigram* for all test collections. A more thorough analysis and comparison of these two models is provided in Sect. 8.4.

9 Conclusions and future work

We have proposed and constructed a novel language-independent framework for cross-language information retrieval, built upon the idea of cross-language topic models trained on document-aligned corpora, which does not use any type of an external translation resource such as a machine translation system or a dictionary that is hand-built or extracted from parallel data. The models employ translation dictionaries extracted directly from per-topic word distributions of the trained topic model instead. It makes the framework language-independent and applicable to any language pair. We have successfully integrated the automatically generated BiLDA-based cross-language lexicon into novel CLIR models. We have proved that models that exploit more different evidences (not necessarily disjunct) yield better retrieval results. Naturally, the cross-language lexicon proves to be of greater importance for source language queries where query words are not observed in target language test collections. We have also shown that adding out-of-domain comparable data to the training data boosts the quality of the topic model and, consequently, the proposed CLIR models that perform better even on topically unrelated test collections.

We have thoroughly evaluated all our models using our manually constructed Wikipedia test set and standard test collections from the CLEF 2001–2003 CLIR campaigns, presenting and explaining their key advantages and shortcomings. We have shown that our combined models, which fuse different evidences (probability distributions from the BiLDA model, unigrams shared across languages, knowledge from the BiLDA-induced lexicon) generally obtain the best scores.

Estimation of the cross-lingual BiLDA model is done *offline* (following the "*learn once, use many*" principle), so there are no restrictions in utilizing the proposed framework in real-world applications. We have shown that a large amount of Wikipedia articles paired through the interlingual links constitutes a quality dataset to train a cross-language topic model, which can later be used to cross-lingually retrieve documents in monolingual data collections. The BiLDA model can be easily expanded to cover more than two languages, while the CLIR framework and the methods for CLE underpinned by it remain completely unchanged and follow the same steps.

The BiLDA model was originally designated for parallel corpora, where an a priori assumption of a shared topical space is clearly valid. However, we have proved its applicability on comparable document-aligned data such as Wikipedia, where a greater divergence exists between topics and subtopics being addressed in different languages. Hence, in future work, we plan to expand the standard BiLDA or construct a novel cross-language model similar to the work presented by De Smet et al. (2011) which will be able to learn the number of topics dynamically during training. Those models should fit more divergent comparable training datasets. By using potentially novel models more suitable to comparable document-aligned corpora, we hope to learn better per-topic word and per-document topic distributions which will lead to retrieval models of higher quality. We also plan to combine our models with relevance models. Finally, we plan to apply the constructed framework to other cross-lingual tasks, such as document summarization and classification.

Acknowledgments We would like to thank the anonymous reviewers for their insightful and constructive comments. This research has been carried in the framework of the *TermWise* Knowledge Platform (IOF-KP/09/001) funded by the Industrial Research Fund, KU Leuven, Belgium, and the Flemish SBO-IWT project *AMASS++* (SBO-IWT 0060051).

Appendix: Gibbs sampling for BiLDA

The goal of training the bilingual LDA (BiLDA) model is double; given a document-aligned corpus, i.e., a collection of aligned document pairs:

1. Discover which words together form vocabulary topics (for both source and target sides), i.e., per-topic word distributions ϕ and ψ.
2. Discover which topics appear in each document pair, i.e., per-document topic distribution θ.

The most likely values of θ, ϕ, and ψ that explain the corpus are thus sought. The topics will then contain meaningful words that share a semantic meaning, relevant to the training corpus, as this configuration is more likely to generate the given corpus than a random collection of words. We will present how to learn those most likely values for BiLDA using Gibbs sampling (Geman and Geman 1984).

We will show the derivation and explain the notation for the source side of a bilingual corpus only (denoted by S, also in the superscripts of the variables involved in the derivation). The derivation for the target side of the corpus (T) follows completely analogously.

θ and ϕ will not be calculated directly, but rather inferred afterwards. As a result, they are integrated out of the calculations. The only hidden variable that is left then is z. Gibbs sampling then dictates that each z is cyclically updated, by being sampled from its posterior given all other variables (including all other z_{ji}^S-s). For the S part of each document pair d_j and each word position i, the probability is calculated that z_{ji}^S assumes, as its new value, one of the K possible topic indices. This new value is indicated with the variable k^*:

$$
\begin{aligned}
\text{sample}\quad z_{ji}^S &\sim P(z_{ji}^S = k^* | \mathbf{z}_{\neg ji}^S, \mathbf{z}^T, \mathbf{w}^S, \mathbf{w}^T, \alpha, \beta) \\
&\sim \int_\theta \int_\phi P(z_{ji}^S = k^* |, \mathbf{z}_{\neg ji}^S, \mathbf{z}^T, \mathbf{w}^S, \mathbf{w}^T, \alpha, \beta, \theta, \phi) d\phi d\theta \\
&\propto \int_\theta \int_\phi P(z_{ji}^S = k^* | \mathbf{z}_{\neg ji}^S, \mathbf{z}^T, \theta, \alpha) \cdot P(w_{ji}^S | z_{ji}^S = k^*, \mathbf{z}_{\neg ji}^S, \mathbf{w}_{\neg ji}^S, \phi, \beta) d\phi d\theta \\
&\propto \int_{\theta_j} P(z_{ji}^S = k^* | \theta_j) \cdot P(\theta_j | \mathbf{z}_{\neg ji}^S, \mathbf{z}^T, \alpha) d\theta_j \cdot \\
&\quad \int_{\phi_{k^*}} P(w_{ji}^S | z_{ji}^S = k^*, \phi_{k^*}) \cdot P(\phi_{k^*} | \mathbf{z}_{\neg ji}^S, \mathbf{w}_{\neg ji}^S, \beta) d\phi_{k^*} \\
&\propto \int_\theta \theta_j^k \cdot P(\theta_j | \mathbf{z}_{\neg ji}^S, \mathbf{z}^T, \alpha) d\theta \cdot \int_\phi \phi_{k^*}^{w_{ji}} \cdot P(\phi_{k^*} | \mathbf{z}_{\neg ji}^S, \mathbf{w}_{\neg ji}^S, \beta) d\phi.
\end{aligned}
$$

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). So, the expected values ($\int xf(x)dx$) for θ and ϕ become:

$$= E_{Dirichlet(n_{j,k^*,\neg i}^S + n_{j,k^*}^T + \alpha)}[\theta_{k^*}^j] \cdot E_{Dirichlet(v_{k^*,w_{ji}^S,\neg}^S + \beta)}[\phi_k^{ji}],$$

which, explicitly written in function of the formula for an expected value of a Dirichlet distribution gives:

$$= \frac{n_{j,k^*,\neg i}^S + n_{j,k^*}^T + \alpha}{n_{j,\cdot,\neg i}^S + n_{j,\cdot}^T + K \cdot \alpha} \cdot \frac{v_{k^*,w_{ji}^S,\neg}^S + \beta}{v_{k^*,\cdot,\neg}^S + |W^S| \cdot \beta}. \tag{25}$$

So the formulas for Gibbs sampling used in the BiLDA training are

$$P(z_{ji}^S = k^*) \propto \frac{n_{j,k^*,\neg i}^S + n_{j,k^*}^T + \alpha}{n_{j,\cdot,\neg i}^S + n_{j,\cdot}^T + K \cdot \alpha} \cdot \frac{v_{k^*,w_{ji}^S,\neg}^S + \beta}{v_{k^*,\cdot,\neg}^S + |W^S| \cdot \beta} \tag{26}$$

and

$$P(z_{ji}^T = k^*) \propto \frac{n_{j,k^*,\neg i}^T + n_{j,k^*}^S + \alpha}{n_{j,\cdot,\neg i}^T + n_{j,\cdot}^S + K \cdot \alpha} \cdot \frac{v_{k^*,w_{ji}^T,\neg}^T + \beta}{v_{k^*,\cdot,\neg}^T + |W^T| \cdot \beta}. \tag{27}$$

The last two formulas use important *counter* variables. The counter n_{j,k^*}^S denotes the number of times a source word w_{ji}^S occurs with a source topic k^* in the source document of a document pair d_j, while $n_{j,k^*,\neg i}^S$ has the same meaning, but not counting the current w_{ji}^S (i.e., $n_{j,k^*}^S - 1$). The same is true for the target side T.

When a "$\cdot$" appears 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_{j,k^*}^S counts the number of values of w_{ji}^S over one topic k^* in d_j, $n_{j,\cdot}^S$ does so over all topics in d_j.

The second counter variable, $v_{k^*,w_{ji},\neg}^S$ is the number of times w_{ji}^S occurs with source topic k^* on the source side of the corpus, but not counting the current w_{ji}^S (i.e., $v_{k^*,w_{ji}}^S - 1$).

We also have to define indicator variables w_{ji}^S and z_{ji}^S. w_{ji}^S denotes the word from the source vocabulary that can be found on position i in a source document from the document pair d_j, and z_{ji}^S is the source topic index associated with the source word w_{ji}^S.

Additionally, $\mathbf{z}^\mathbf{S}$ denotes all source topic indices for the document pair d_j, $\mathbf{z}_{\neg ji}^S$ denotes all source topic indices in d_j excluding w_{ji}^S. $\mathbf{w}^\mathbf{S}$ denotes all source words in a corpus, and $|W^S|$ is the number of source words in the corpus. K is the number of topics set a priori before training, and α and β are the parameters of the uniform conjugate Dirichlet priors (see Griffiths et al. 2007; Heinrich 2008).

$v_{k^*,\cdot,\neg}^S$ counts the total number of source words associated with source topic k^* in the whole corpus, as it is the sum over all possible source words (a "$\cdot$" appears instead of the w_{ji}^S). Again, because of the $\neg$ symbol in the superscript, the current w_{ji}^S is not counted (i.e., $v_{k^*,\cdot}^S - 1$).

With formulas (26) and (27), each z_{ji}^S (and z_{ji}^T) of each document pair is sampled and updated in turn. After a random initialization (usually a uniform distribution of probabilities), the sampled z values will converge to samples taken from the real joint distribution of θ, ϕ and ψ, after a time called the *burn-in period*. The estimations for θ^j, ϕ^k and ψ^k can then be calculated from these burned-in samples.

As can be seen from the first term of Eqs. (26) and (27), the document pairs are linked by the count variables n_j^S and n_j^T, as both z_{ji}^S and z_{ji}^T are drawn from the same θ. The vocabulary count variables operate only within the language of the term currently being considered.

Finally, one of the main advantages of the BiLDA model is an efficient procedure for inference. By taking the vocabulary distributions and the prior on the per-document topic distributions outside of the corpus, the per-document topic distribution of new documents can be inferred using the same Gibbs sampling formulas used for training.

References

Adriani, M. & Rijsbergen, C. J. V. (1999). Term similarity-based query expansion for cross-language information retrieval. In: *Proceedings of the third European conference on research and advanced technology for digital libraries* (pp. 311–322).

Azzopardi, L., de Rijke, M., & Balog, K. (2007). Building simulated queries for known-item topics: an analysis using six European languages. In: *Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 455–462).

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* (pp. 222–229).

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

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

Chen, D., Xiong, Y., Yan, J., Xue, G.-R., Wang, G., & Chen, Z. (2010). Knowledge transfer for cross domain learning to rank. *Information Retrieval*, (13), 236–253.

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* (pp. 143–152).

Chiao, Y.-C., & Zweigenbaum, P. (2002). Looking for candidate translational equivalents in specialized, comparable corpora. In: *Proceedings of the 19th international conference on computational linguistics* (pp. 1–5).

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* (pp. 1513–1518).

De Smet, 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* (pp. 57–64).

De Smet, W., Tang, J., & Moens, M.-F. (2011). Knowledge transfer across multilingual corpora via latent topics. In: *Proceedings of the PAKDD: the 15th Pacific-Asia conference on knowledge discovery and data mining* (pp. 549–560).

Déjean, H., Gaussier, E., & Sadat, F. (2002). An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In: *Proceedings of the 19th international conference on computational linguistics* (pp. 1–7).

Diab, M. T., & Finch, S. (2000). A statistical translation model using comparable corpora. In: *Proceedings of the 6th triennial conference on recherche d'Information Assistée par Ordinateur (RIAO)* (pp. 1500–1508).

Dolamic, L. & Savoy, J. (2010). Retrieval effectiveness of machine translated queries. *Journal of the American Society for Information Science and Technology, 61*(11), 2266–2273.

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).

Fung, P., & Cheung, P. (2004). Mining very-non-parallel corpora: Parallel sentence and lexicon extraction via bootstrapping and EM. In: *Proceedings of the conference on empirical methods in natural language processing* (pp. 57–63).

Fung, P., & Yee, L. Y. (1998). An IR approach for translating new words from nonparallel, comparable texts. In: *Proceedings of the 17th international conference on computational linguistics* (pp. 414–420).

Gaussier, E., 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* (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.

Grefenstette, G. (1998). *Cross-language information retrieval*. MA, USA: Norwell.

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* (pp. 771–779).

Harris, Z. S. (1954). Distributional structure. *Word, 10*(23), 146–162.

Heinrich, G. (2008). Parameter estimation for text analysis. *Technical report*.

Hofmann, T. (1999). Probabilistic Latent Semantic Indexing. In: *Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval* (pp. 50–57).

Jagarlamudi, J., & Daumé III, H. (2010). Extracting multilingual topics from unaligned comparable corpora. In: *Proceedings of the 32th annual European conference on advances in information retrieval* (pp. 444–456).

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In: *Proceedings of the MT summit 2005* (pp. 79–86).

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., & Herbst, E. (2007). Moses: open source toolkit for statistical machine translation. In: *Proceedings of the 45th annual meeting of the association for computational linguistics* (pp. 177–180).

Laroche, A. & Langlais, P. (2010). Revisiting context-based projection methods for term-translation spotting in comparable corpora. In: *Proceedings of the 23rd international conference on computational linguistics* (pp. 617–625).

Lavrenko, V., Choquette, M., & Croft, W. B. (2002). Cross-lingual relevance models. In: *Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 175–182).

Littman, M., Dumais, S. T., & Landauer, T. K. (1998). In: *Cross-language information retrieval, chapter 5* (pp. 51–62). Dordrecht: Kluwer.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA, USA: MIT Press.

Mathieu, B., Besançon, R., & Fluhr, C. (2004). Multilingual document clusters discovery. In: *Proceedings of the 7th triennial conference on recherche d'Information assistée par ordinateur (RIAO)* (pp. 116–125).

Mimno, D., Wallach, H. M., Naradowsky, J., Smith, D. A., & McCallum, A. (2009). Polylingual topic models. In: *Proceedings of the 2009 conference on empirical methods in natural language processing* (pp. 880–889).

Morin, E., Daille, B., Takeuchi, K., & Kageura, K. (2007). Bilingual terminology mining—using brain, not brawn comparable corpora. In: *Proceedings of the 45th annual meeting of the association for computational linguistics* (pp. 664–671).

Muramatsu, T., & Mori, T. (2004). Integration of pLSA into probabilistic CLIR model. In: *Proceedings of NTCIR-04*.

Navarro, G. (2001). A guided tour to approximate string matching. *ACM Computing Surveys, 33*(1), 31–88.

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* (pp. 1155–1156).

Nie, J.-Y. (2010). Cross-language information retrieval. *Synthesis Lectures on Human Language Technologies*.

Nie, J.-Y., Simard, M., Isabelle, P., & Durand, R. (1999). Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web. In: *Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval* (pp. 74–81).

Och, F. J. & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics, 29*(1), 19–51.

Platt, J. C., Toutanova, K., & Yih, W.-T. (2010). Translingual document representations from discriminative projections. In: *Proceedings of the conference on empirical methods in natural language processing* (pp. 251–261).

Rapp, R. (1995). Identifying word translations in non-parallel texts. In: *Proceedings of the 33rd annual meeting of the association for computational linguistics* (pp. 320–322).

Rapp, R. (1999). Automatic identification of word translations from unrelated English and German corpora. In: *Proceedings of the 37th annual meeting of the association for computational linguistics* (pp. 519–526).

Resnik, P., & Smith, N. A. (2003). The web as a parallel corpus. *Computational Linguistics, 29*(3), 349–380.

Roth, B., & Klakow, D. (2010). Combining Wikipedia-based concept models for cross-language retrieval. In: *Proceedings of the information retrieval facility conference* (pp. 47–59).

Savoy, J. (2004). Combining multiple strategies for effective monolingual and cross-language retrieval. *Information Retrieval, 7*(1-2), 121–148.

Sheridan, P., & Ballerini, J. P. (1996). Experiments in multilingual information retrieval using the spider system. In: *Proceedings of the 19th annual International ACM SIGIR conference on research and development in information retrieval* (pp. 58–65).

Shezaf, D., & Rappoport, A. (2010). Bilingual lexicon generation using non-aligned signatures. In: *Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 98–107).

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).

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* (pp. 72–79).

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* (pp. 843–851).

Vulić, I., De Smet, W., & Moens, M.-F. (2011). 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* (pp. 479–484).

Wang, A., Li, Y., & Wang, W. (2009). Cross-language information retrieval based on LDA. In: *Proceedings of the IEEE international conference on intelligent computing and intelligent systems* (pp. 485–490).

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* (pp. 178–185).

Xu, J., Weischedel, R., & Nguyen, C. (2001). Evaluating a probabilistic model for cross-lingual information retrieval. In: *Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 105–110).

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* (pp. 627–634).

Zhai, C., & Lafferty, J. (2004). A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems, 22*, 179–214.