Cross-Language Information Retrieval with Latent Topic Models Trained on a Comparable Corpus

Ivan Vulić, Wim De Smet, and Marie-Francine Moens

Department of Computer Science, K.U. Leuven, Belgium
{ivan.vulic,wim.desmet,marie-francine.moens}@cs.kuleuven.be

Abstract. In this paper we study cross-language information retrieval using a bilingual topic model trained on comparable corpora such as Wikipedia articles. The bilingual Latent Dirichlet Allocation model (BiLDA) creates an interlingual representation, which can be used as a translation resource in many different multilingual settings as comparable corpora are available for many language pairs. The probabilistic interlingual representation is incorporated in a statistical language model for information retrieval. Experiments performed on the English and Dutch test datasets of the CLEF 2001-2003 CLIR campaigns show the competitive performance of our approach compared to cross-language retrieval methods that rely on pre-existing translation dictionaries that are hand-built or constructed based on parallel corpora.

Keywords: Cross-language retrieval, topic models, comparable corpora, document models, multilingual retrieval, Wikipedia

1 Introduction

With the ongoing growth of the World Wide Web and the expanding use of different languages, the need for cross-language models that retrieve relevant documents becomes more pressing than ever. Cross-language information retrieval deals with the retrieval of 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 machine translation system. Translation dictionaries do not exist for every language pair, and they are usually trained on large parallel corpora, where each document has an exact translation in the other language, or are hand-built. Parallel corpora are not available for each language pair. In contrast, comparable corpora in which documents in the source and the target language contain similar content, are usually available in abundance. In this paper we address the question whether suitable cross-language retrieval models can be built based on the interlingual topic representations learned from comparable corpora. We accomplish this goal by means of a cross-language generative model, i.e., bilingual Latent Dirichlet Allocation (BiLDA), trained on a comparable corpus such as one composed of Wikipedia articles. The resulting probabilistic translation model is incorporated in a statistical language model for information retrieval. The language models for retrieval have a sound statistical foundation and can easily incorporate probabilistic evidence in order to optimize the cross-language retrieval process.

The contributions of the paper are as follows. Firstly, we show the validity and the potential of training a bilingual LDA model on bilingual comparable corpora. Secondly, we successfully integrate the topic distributions resulting from training the bilingual LDA model in several variant retrieval models and perform a full-fledged evaluation of the retrieval models on the standard CLEF test collections. We show that the results obtained by our retrieval models, which do not exploit any linguistic knowledge from a translation dictionary, are competitive with dictionary-based models. Our work makes cross-language information retrieval portable to many different language pairs.

2 Related Work

Probabilistic topic models such as probabilistic Latent Semantic Indexing [9] and Latent Dirichlet Allocation [1] are both 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. Cimiano et al. [6] use standard LDA trained on concatenated parallel and comparable documents in a document comparison task. Roth and Klakow [23] try to use the standard LDA model trained on concatenated Wikipedia articles for cross-language information retrieval, but they do not obtain decent results without the additional usage of a machine translation system.

Recently, the bilingual or multilingual LDA model was independently proposed by different authors ([17, 14, 7, 2]) who identify interlingual topics of different languages. These authors train the bilingual LDA model on a parallel corpus. Jagarlamudi and Daumé III [10] extract interlingual topics from comparable corpora, but use additional translation dictionary information. None of these works apply the bilingual LDA model in a cross-lingual information retrieval setting.

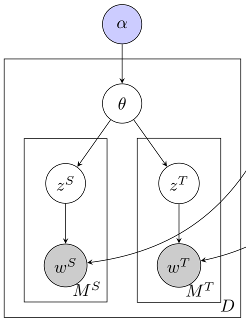
Cross-language information retrieval is a well-studied research topic (e.g., [8, 19, 24, 18]). As mentioned, existing methods rely on a translation dictionary to bridge documents of different languages. In some cases interlingual information is learned based on parallel corpora and correlations found in the paired documents [13], 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, document and query are projected in a lower dimensional space ([3, 15, 5, 29]). Our work follows this line of thinking, but uses generative probabilistic approaches. In addition, the models are trained on the individual documents in the different languages, but paired by their joint interlingual topics. Cross-language relevance models [12] have also been applied for the task, but they still require either a parallel corpus or a translation dictionary. LDA-based monolingual retrieval has been described by Wei and Croft [28].

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 pLSA was proposed by Xue et al. [29] 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. [4]. Takasu [26] proposes cross-language keyword recommendation using latent topics. Except for Wang et al. [27], where the evaluation is vague and unsatisfactory (the same dataset is used for training and testing), and relies solely on 30 documents and 7 queries, none of the above works use LDA-based interlingual topics in cross-language retrieval.

3 Bilingual LDA

The topic model we use is a bilingual extension of a standard LDA model, called *bilingual LDA* (BiLDA) ([17, 14, 7, 2]).

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. 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. Algorithm 3.1 summarizes the generative story, while Figure 1 shows the plate model.



Algorithm 3.1: GENERATIVE STORY FOR BILDA()

```
for each document pair $d_j$
do
    for each word position $i \in d_{jS}$
    do
        sample $z_{ji}^S \sim Mult(\theta)$
        sample $w_{ji}^S \sim Mult(\phi, z_{ji}^S)$
    for each word position $i \in d_{jT}$
    do
        sample $z_{ji}^T \sim Mult(\theta)$
        sample $w_{ji}^T \sim Mult(\psi, z_{ji}^T)$
```

Fig. 1. Generative description and plate model of the bilingual BiLDA model

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.

The described BiLDA model serves as a framework for modeling our retrieval models. After the training using Gibbs sampling ([25]), two sets of probability distributions are obtained for each of the languages. One set consists of per-topic word probability distributions, calculated as $P(w_i|z_k) = \phi_{k,i}^S = \frac{n_k^{(w_i)}+\beta}{\sum_{j=1}^{|W^S|} n_k^{(w_j)}+W^S\beta}$, where $n_k^{(w_i)}$ denotes the total number of times that the topic z_k is assigned to the word w_i from the vocabulary W^S. 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 consists of per-document topic probability distributions, calculated as $P(z_k|D_J) = \theta_{J,k} = \frac{n_J^{(k)}+\alpha}{\sum_{j=1}^{K} n_J^{(j)}+K\alpha}$, 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.

4 LDA-Based CLIR

This section provides a theoretical insight to cross-language retrieval models relying on per-topic word distributions and per-document word distributions.

4.1 LDA-Only CLIR Model

Given the set $\{D_1, D_2, \dots, D_L\}$ 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 [28]. 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:

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

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 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. We propose to use sets of per-topic word distributions and per-document topic distributions, assuming the shared space of latent topics. We calculate the right-hand side of equation (1) as

$$\begin{aligned} P(q_i|D_J) &= \delta_1 \sum_{k=1}^{K} \overbrace{P(q_i|z_k^S)}^{Source\, z_k} \underbrace{P(z_k^T|D_J)}_{Target\, z_k} + (1-\delta_1)P(q_i|Ref) \\ &= \delta_1 \sum_{k=1}^{K} \phi_{k,i}^S \theta_{J,k}^T + (1-\delta_1)P(q_i|Ref), \end{aligned} \tag{2}$$

by using the two BiLDA-related probability distributions $\phi_{k,i}^S$ and $\theta_{J,k}^T$. The parameter δ_1 is an 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.

The per-topic word distributions for the source language are used to predict the probability that the word q_i from the query Q will be sampled from the topic z_k^S, and 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[1]) is assigned to a token in the target document D_J. As LDA is a generative model, we may infer the source or target language part of a pre-trained bilingual model on any monolingual collection in the source or the target language, using the same formulas for $\phi_{k,i}^S$ or $\psi_{k,i}^T$ and θ_J, k as in Section 3.

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**:

1. Infer the trained model on a test corpus in the target language to learn $P(z_k^T|D_J)$
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, $k = 1, \ldots, K$
   (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:
$$P(Q|D_J) = \prod_{i=1}^{m} \left(\delta_1 \sum_{k=1}^{K} \phi_{k,i}^S \, \theta_{J,k}^T + (1 - \delta_1) P(q_i|Ref) \right) \tag{3}$$

This gives the score for one target language document D_J. Finally, documents are ranked based on their 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 Subsection 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 (1) may be specified by a document model with the Dirichlet smoothing. We adopt smoothing techniques according to evaluations and findings from [30]. The Dirichlet smoothing acts as a length normalization parameter and penalizes long documents. The model is then:

$$\begin{aligned} P_{lex}(q_i|D_J) =& \delta_2 \Big(\frac{N_d}{N_d + \mu} P_{mle}(q_i|D_J) + (1 - \frac{N_d}{N_d + \mu}) P_{mle}(q_i|Coll) \Big) \\ &+ (1 - \delta_2) P(q_i|Ref), \end{aligned} \tag{4}$$

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

---

[1] z_k^S and z_k^T basically refer to the same cross-language topic z_k, but z_k^S is interpreted as a cross-language topic used by source language words, and z_k^T by the target language words.

Coll, 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 the large corpus *Ref*. 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:

$$
\begin{aligned}
P(q_i|D_J) &= \lambda P_{lex}(q_i|D_J) + (1-\lambda)P_{lda}(q_i|D_J) && (5) \\
&= \lambda\Big(\delta_2\big(\frac{N_d}{N_d+\mu}P_{mle}(q_i|D_J) + (1-\delta_2)P(q_i|Ref)\big)\Big) \\
&\quad + (1-\lambda)P_{lda}(q_i|D_J) && (6)
\end{aligned}
$$

where P_{lda} is the *LDA-only* model given by (2), P_{lex} the simple unigram model given by (4), 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* ([16]) 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. We believe that a more advanced ”non-LDA” part[2] of the document model may result in even higher scores, since knowledge from other translation resources may be used to model the probability $P_{lex}(q_i|D_J)$.

5 Experimental Setup

5.1 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 document alignment, and it is the only assumption our BiLDA model utilizes during training.

The first subset of our training data is the Europarl corpus [11], extracted from proceedings of the European Parliament and consisting of $6,206$ parallel 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*[3] 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 con-

---

[2] By the “LDA-part” of the retrieval model, we assume the part of the model in equation (2).

[3] `http://dumps.wikimedia.org/`

siderable amount of divergence between aligned documents. Our Wikipedia training sub-corpus consists of $7,612$ documents which vary in length, theme and style[4].

As a preprocessing step we remove stop words, and our final vocabularies consist of $76,555$ words in English, and $71,168$ words in Dutch.

5.2 Test Collections

Our experiments have been conducted on three data sets taken from the CLEF 2001-2003 CLIR campaigns: the LA Times 1994 (**LAT**), the LA Times 1994 and Glasgow Herald 1995 (**LAT+GH**) in English, and the NRC Handelsblad 94-95 and the Algemeen Dagblad 94-95 (**NC+AD**) in Dutch. Statistics of the collections are given in Table 1.

Table 1. Statistics of the experimental setup

(a) Statistics of test collections

Collection	Contents	# of Docs
LAT	LA Times 94 (EN)	110,861
LAT+GH	LA Times 94 (EN) Glasgow Her.95 (EN)	166,753
NC+AD	NRC Hand. 94-95 (NL) Alg. Dagblad 94-95 (NL)	190,604

(b) Statistics of used queries

CLEF Topics (Year: Topic Nr.)	# Queries	Used for
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

Queries are extracted from the *title* and *description* fields of CLEF topics for each year. Stop words have been removed from queries and documents. Table 1(b) shows the queries used for the test collections.

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 following [25]. The Dirichlet parameter μ in the LDA-unigram retrieval model is set to 1000. The parameters δ_1 and δ_2 are set to negligible values[5], while we set $\lambda = 0.3$, which gives more weight to the topic model.

6 Results and Discussion

This section reports our experimental results for both English-Dutch CLIR and Dutch-English CLIR. The cross-language topic model is trained just once on a large bilingual training corpus. After training, it can be used for both retrieval directions, after we

---

[4] We will make the corpus publicly available at `http://www.cs.kuleuven.be/groups/liir/software.php`.

[5] 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.

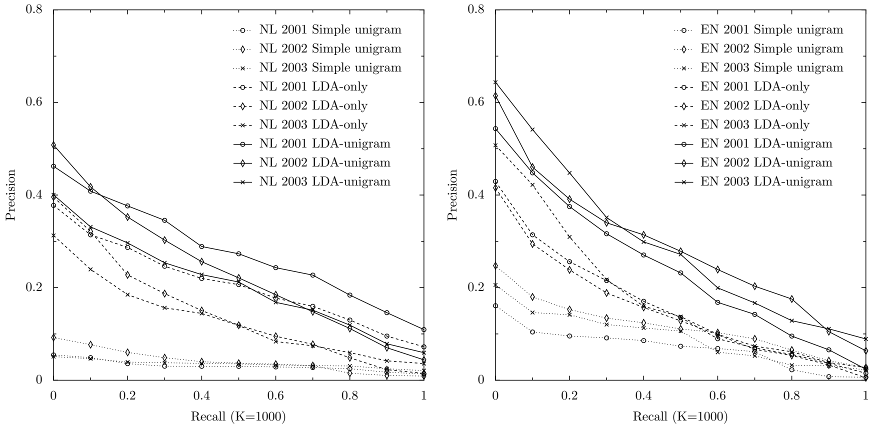
infer it on the appropriate test collection. We have carried out the following experiments: (1) we compare our LDA-only model to several baselines that have also tried to exploit latent concept spaces for cross-language information retrieval, such as cross-language Latent Semantic Indexing (cLSI) and standard LDA trained on concatenated paired documents. We want to prove the soundness and the usefulness of the basic LDA-only model and, consequently, other models that might later build upon the foundation established by the LDA-only model. (2) We provide an extensive evaluation over all CLEF test collections with all our retrieval models, and provide a comparison of the best scoring LDA-unigram model with some of the best CLIR systems from the CLEF 2001-2003 campaigns. We have trained our BiLDA model with a different number of topics (400, 1000 and 2200) on the combined **EP+Wiki** corpus. The main evaluation measure we use for all experiments is the *mean average precision* (MAP). For several experiments, we additionally provide precision-recall curves.

6.1 Comparison with Baseline Systems

The LDA-only model serves as the backbone of other, more advanced BiLDA-based document models. 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 state-of-the-art systems which try to build a CLIR system based around the idea of latent concept spaces: (i) the cross-language Latent Semantic Indexing (cLSI) as described by [3], which constructs a reduced (latent) vector space trained on concatenated paired documents in two languages, and (ii) the standard LDA model trained on the merged document pairs [23].

We have trained the cLSI model and the standard LDA model on the combined *EP+Wiki* corpus with 400 and 1000 dimensions (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 the other two models by a huge margin. The MAP scores for cLSI and standard LDA are similar and 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. The MAP scores of the LDA-only model for NL 2001, NL 2002, and NL 2003 for K=1000 are 0.1969, 0.1396, and 0.1227, respectively, while the MAP scores for EN 2001, EN 2002, and EN 2003 for K=1000 are 0.1453, 0.1374, and 0.1713, respectively.

One reason for such a huge difference in scores might be the ability to infer the BiLDA model on a new test collection (due to its fully generative semantics) more accurately. Cross-language LSI for CLIR reported in the literature always uses the same corpus (or subsets of the same corpus) for training and testing, while this setting asks for inferring on a test corpus which is not by any means content-related to a training corpus. BiLDA has a better statistical foundation by defining the common per-document topic distribution θ, which allows inference on new documents based on the previously trained model and also avoids the problem of overfitting inherent to the pLSI model and, consequently, the cLSI model. Another problem with the baseline methods might be the 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.



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

Fig. 2. Precision-recall for all models. K=1000, training corpus is EP+Wiki.

6.2 Comparison of Our CLIR Models

Using a Fixed Number of Topics (K=1000) In this subsection, the LDA-only model, the simple unigram model and the combined LDA-unigram model have been evaluated on all test collections, with the number of topics initially fixed to 1000. Table 2 contains MAP scores for the LDA-unigram model, Figure 2(a) shows the precision-recall values obtained by applying all three models to the English test collections and the Dutch queries, while Figure 2(b) shows the precision-recall values for the Dutch test collections and the English queries.

Table 2. MAP scores of the LDA-unigram model for all test collections and different number of topics K. Training corpus is EP+Wiki.

Queries	**K=400**	**K=1000**	**K=2200**
NL 2001	0.2330	0.2673	0.2813
NL 2002	0.2093	0.2253	0.2206
NL 2003	0.1608	0.1990	0.1658
EN 2001	0.2204	0.2275	0.2398
EN 2002	0.2455	0.2683	0.2665
EN 2003	0.2393	0.2783	0.2450

Varying the Number of Topics The main goal of the next set of experiments was to test the performance of our models if we vary the number of topics set for BiLDA training. We have carried out experiments with the CLIR models relying on BiLDA trained with different numbers of topics (400, 1000 and 2200). Figure 3 shows the precision-recall values of the LDA-only and the LDA-unigram model, while the associated MAP scores of the best scoring LDA- unigram model are presented in Table 2.

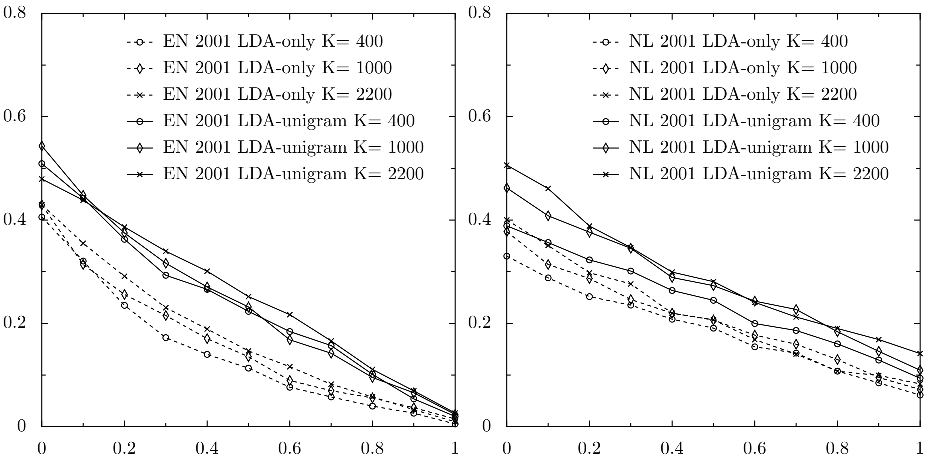


Fig. 3. Precision-recall for the LDA-only and the LDA-unigram model for the 2001 test collections. Training corpus is EP+Wiki.

Discussion 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, the combination of the LDA-only and the simple unigram model, which allows retrieving relevant documents based on shared words across the languages (e.g. personal names), leads to much better scores which are competitive even with models which utilize cross-lingual dictionaries or machine translation systems. For instance, our LDA-unigram model would have been placed among the top 5 retrieval systems for the CLEF 2002 Bilingual to Dutch task, would have been placed among the top 3 retrieval systems for the CLEF 2001 Bilingual to Dutch task, and outperforms the only participating system in the CLEF 2002 Dutch to English task (MAP: 0.1495) [20, 21]. All these state-of-the-art CLEF systems operated in a similar settings as ours and constructed queries from *title* and *description* or *title*, *description* and *narrative* fields from the CLEF topics. They, however, rely on translation resources which were hand-built or trained on parallel corpora. We obtain competitive results by using the BiLDA model trained on comparable corpora. We believe that our results could still improve by training the BiLDA model on a corpus which is topically related with the corpus on which we perform the retrieval.

7 Conclusions and Future Work

We have proposed a novel language-independent and dictionary-free framework for cross-language information retrieval that does not use any type of a cross-lingual dictionary or translation system. The framework is built upon the idea of cross-language topic models obtained by applying a bilingual Latent Dirichlet Allocation model (BiLDA), where the only prerequisite is the availability of abundant training data consisting of comparable document-aligned documents.

We have thoroughly evaluated this cross-language retrieval model using standard test collections from the CLEF 2001-2003 CLIR campaigns and have shown that our combined model, which fuses evidence from the BiLDA model and the unigram model, is competitive with the current top CLIR systems that use translation resources that are hand-built or are trained on parallel corpora.

In future work, we will accumulate more comparable document-aligned data, exploiting Wikipedia and other sources. We also plan to construct other models that will combine topical knowledge with other evidences (for instance, using cognates instead of exactly the same words shared across languages). Additionally, we plan to expand the standard BiLDA to fit more divergent comparable training datasets. In addition, the cross-language knowledge transfer based on the proposed generative topic models that are trained on comparable corpora might be useful in many other multilingual information management tasks including categorization and summarization.

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. Journal of Machine Learning Research (3), 993–1022 (2003)
2. Boyd-Graber, J., Blei, D.M.: Multilingual topic models for unaligned text. In: Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence. pp. 75–82 (2009)
3. Carbonell, J.G., Yang, J.G., Frederking, R.E., Brown, R.D., Geng, Y., Lee, D., Frederking, Y., E, R., Geng, R.D., Yang, Y.: Translingual information retrieval: A comparative evaluation. In: Proceedings of the 15th International Joint Conference on Artificial Intelligence. pp. 708–714 (1997)
4. Chen, D., Xiong, Y., Yan, J., Xue, G.R., Wang, G., Chen, Z.: Knowledge transfer for cross domain learning to rank. Information Retrieval (13), 236–253 (2010)
5. Chew, P.A., Bader, B.W., Kolda, T.G., Abdelali, A.: Cross-language information retrieval using PARAFAC2. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 143–152 (2007)
6. Cimiano, P., Schultz, A., Sizov, S., Sorg, P., Staab, S.: Explicit versus latent concept models for cross-language information retrieval. In: Proceedings of the 21st International Joint Conference on Artifical Intelligence. pp. 1513–1518 (2009)
7. De Smet, W., Moens, M.F.: 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 (2009)
8. Grefenstette, G.: Cross-Language Information Retrieval. Norwell, MA, USA (1998)
9. Hofmann, T.: Probabilistic Latent Semantic Indexing. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 50–57 (1999)
10. Jagarlamudi, J., Daumé III, H.: Extracting multilingual topics from unaligned comparable corpora. In: Proceedings of the 32th Annual European Conference on Advances in Information Retrieval. pp. 444–456 (2010)
11. Koehn, P.: Europarl: A parallel corpus for statistical machine translation. In: Proceedings of the MT Summit 2005. pp. 79–86 (2005)

12. Lavrenko, V., Choquette, M., Croft, W.B.: Cross-lingual relevance models. In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 175–182 (2002)
13. Mathieu, B., Besançon, R., Fluhr, C.: Multilingual document clusters discovery. In: Proceedings of the 7th Triennial Conference on Recherche d'Information Assistée Par Ordinateur (RIAO). pp. 116–125 (2004)
14. Mimno, D., Wallach, H.M., Naradowsky, J., Smith, D.A., McCallum, A.: Polylingual topic models. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. pp. 880–889 (2009)
15. Muramatsu, T., Mori, T.: Integration of pLSA into probabilistic CLIR model. In: Proceedings of NTCIR-04 (2004)
16. Navarro, G.: A guided tour to approximate string matching. ACM Computing Surveys 33(1), 31–88 (2001)
17. Ni, X., Sun, J.T., Hu, J., Chen, Z.: Mining multilingual topics from Wikipedia. In: 18th International World Wide Web Conference. pp. 1155–1156 (2009)
18. Nie, J.Y.: Cross-Language Information Retrieval. Synthesis Lectures on Human Language Technologies (2010)
19. Nie, J.Y., Simard, M., Isabelle, P., Durand, R.: Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 74–81 (1999)
20. Peters, C., Braschler, M., Gonzalo, J., Kluck, M. (eds.): Evaluation of Cross-Language Information Retrieval Systems, CLEF 2001, Revised Papers, Lecture Notes in Computer Science, vol. 2406 (2002)
21. Peters, C., Braschler, M., Gonzalo, J., Kluck, M. (eds.): Advances in Cross-Language Information Retrieval, CLEF 2002, Revised Papers, Lecture Notes in Computer Science, vol. 2785 (2003)
22. Platt, J.C., Toutanova, K., Yih, W.T.: Translingual document representations from discriminative projections. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 251–261 (2010)
23. Roth, B., Klakow, D.: Combining Wikipedia-based concept models for cross-language retrieval. In: IRFC. pp. 47–59 (2010)
24. Savoy, J.: Combining multiple strategies for effective monolingual and cross-language retrieval. Information Retrieval (7)(1-2), 121–148 (2004)
25. Steyvers, M., Griffiths, T.: Probabilistic topic models. Handbook of Latent Semantic Analysis 427(7), 424–440 (2007)
26. Takasu, A.: 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 (2010)
27. Wang, A., Li, Y., Wang, W.: Cross-language information retrieval based on LDA. pp. 485–490 (Nov 2009)
28. Wei, X., Croft, W.B.: 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 (2006)
29. Xue, G.R., Dai, W., Yang, Q., Yu, Y.: 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 (2008)
30. Zhai, C., Lafferty, J.: A study of smoothing methods for language models applied to information retrieval. ACM Transactions on Information Systems 22, 179–214 (2004)