A Unified Framework for Monolingual and Cross-Lingual Relevance Modeling Based on Probabilistic Topic Models

Ivan Vulić and Marie-Francine Moens

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

Abstract. We explore the potential of probabilistic topic modeling within the relevance modeling framework for both monolingual and cross-lingual ad-hoc retrieval. Multilingual topic models provide a way to represent documents in a structured and coherent way, regardless of their actual language, by means of language-independent concepts, that is, cross-lingual topics. We show how to integrate the topical knowledge into a unified relevance modeling framework in order to build quality retrieval models in monolingual and cross-lingual contexts. The proposed modeling framework processes all documents uniformly and does not make any conceptual distinction between monolingual and cross-lingual modeling. Our results obtained from the experiments conducted on the standard CLEF test collections reveal that fusing the topical knowledge and relevance modeling leads to building monolingual and cross-lingual retrieval models that outperform several strong baselines. We show that that the topical knowledge coming from a general Web-generated corpus boosts retrieval scores. Additionally, we show that within this framework the estimation of cross-lingual relevance models may be performed by exploiting only a general non-parallel corpus.

Keywords: Cross-lingual information retrieval, relevance models, multilingual topic models, probabilistic retrieval models, comparable multilingual corpora.

1 Introduction

Following the ongoing expansion of the World Wide Web and its omnipresence in today's increasingly connected world, more and more content on the Web is available in languages other than English. Additionally, the advent of the Web 2.0 was characterized by the possibility for end users to generate data directly and easily. With user blogs and social websites such as Wikipedia or Twitter users have created huge amounts of data in numerous different languages. Consequently, the Web has truly become a multilingual data-driven environment. A need to successfully navigate through that sea or rather ocean of multilingual information becomes more pressing than ever. Two key questions have emerged from that need: (Q1) How to represent documents written in different languages in a structured and coherent way, regardless of their actual language?, and (Q2) How to perform the effective retrieval of information (monolingually and across languages) that relies on such language-independent representations?

P. Serdyukov et al. (Eds.): ECIR 2013, LNCS 7814, pp. 98–109, 2013.

In this paper, we try to combine the answers to these two questions into a powerful language-independent unified framework for the task of ad-hoc information retrieval, with a special focus on Cross-Lingual Information Retrieval (CLIR) which deals with the retrieval of documents written in a language that differs from the language of the user's query. To answer question Q1, we utilize recent advances in probabilistic *multilingual topic modeling* (MuTM). MuTM provides a way to build structured representations of documents regardless of their language. Probabilistic topic models can then be used in the probabilistic *language modeling* (LM) framework for IR [17,2], as already proven for both monolingual [22,24] and cross-lingual retrieval [20]. However, the prior work dealt with only simpler query likelihood models [22,20], or did not formally define the relation between MuTM and CLIR [22,24,20]. In this work, in order to satisfy the requirements from question Q2, we opt for the more complex and robust *relevance-based LM retrieval framework* [12,11], and exploit the knowledge from multilingual topic models within that framework. We make several important contributions:

(1) We show that it is possible to estimate a quality relevance model in both monolingual and cross-lingual settings by means of a topic model trained on a general easily obtainable user-generated corpus such as Wikipedia.

(2) We present a novel way of estimating *relevance models* by means of a multilingual topic model in the cross-lingual setting. The estimation is performed without any additional translation resource, while previous estimation techniques for cross-lingual relevance models critically relied on either a machine-readable bilingual dictionary or an in-domain parallel corpus [11], not available for many languages and domains.

(3) We additionally show that by our estimation procedure we create a unified formal framework that does not make any conceptual distinction between monolingual retrieval and CLIR. The proposed framework combines the strength and robustness of relevance modeling (e.g., its implicit query expansion and disambiguation) with the strength of MuTM (e.g., shallow semantic analysis of documents, representation by means of language-independent cross-lingual topics).

The reported results from the experiments on the standard CLEF datasets show the validity of our unified approach as (1) Relevance modeling clearly benefits from the additional knowledge coming from a topic model, and it is visible in both monolingual and cross-lingual retrieval settings, (2) Cross-lingual relevance models estimated by means of a multilingual topic model produce results which are better than or comparable to several strong monolingual baselines, (3) Cross-lingual relevance models may be estimated by using only comparable user-generated data, which is especially important for language pairs and domains that lack readily available machine-readable bilingual dictionaries or parallel corpora.

The remainder of the paper is structured as follows. We formally define multilingual topic modeling in Sect. 2. In Sect. 3, we provide a short overview of relevance modeling, and present our novel estimation technique. In Sect. 4, we evaluate our novel retrieval models and show their validity in the monolingual and cross-lingual retrieval tasks of the CLEF campaigns. Our conclusions and future work are summarized in Sect. 5.

2 Multilingual Topic Modeling

Current state-of-the-art multilingual topic models [14,6,4,9,26,15] are multilingual extensions of probabilistic topic models (PTM) initially tailored for the monolingual setting, such as probabilistic Latent Semantic Analysis (pLSA) [8] and Latent Dirichlet Allocation (LDA) [3]. They provide a robust and unsupervised framework for performing shallow latent semantic analysis of themes (or topics) discussed in text. These models are all based upon the idea that there exist latent variables, i.e., topics, which determine how words in documents have been generated. Fitting such a generative model denotes finding the best set of those latent variables in order to explain the observed data. With respect to that generative process, documents are seen as mixtures of latent topics, while topics are seen as probability distributions over vocabulary words.

A multilingual topic model learns a set of language-independent concepts or *cross-lingual topics*. Each document in a document collection can then be represented as a mixture of these topics which is modeled by *per-document topic distributions*. They provide a probability that a certain topic is found in a certain document. Moreover, each topic is represented as a probability distribution over vocabulary words as modeled by *per topic-word distributions*. Each language possesses its own language-specific per-topic word distributions which serve as an interface towards the language-independent concepts, that is, cross-lingual topics. Per-document topic distributions allow uniform representations of all the documents in the language-independent space spanned by cross-lingual topics, while per-topic word distributions provide a way to represent these documents in actual languages. Monolingual topic models could be interpreted as a degenerate special case of multilingual topic models where only one language is involved, and all the definitions and assumptions remain the same.

We will not analyze specific multilingual topic models along with their specific assumptions and generative stories, but only sketch a broad outline and define the concepts that all these models share.

Def. 1. Theme-aligned multilingual corpus. Assume that we are given a theme-aligned multilingual corpus $\mathcal{C}$ of $l = |\mathcal{L}|$ languages, where $\mathcal{L} = \{L_1, L_2, \ldots, L_l\}$ is the set of languages. $\mathcal{C}$ is a set of text collections $\{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_l\}$ where each $\mathcal{C}_i = \{d_1^i, \ldots, d_{nd_i}^i\}$ is a collection of documents in language L_i with vocabulary $V^i = \{w_1^i, w_2^i, \ldots, w_{nw_i}^i\}$. Collections $\{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_l\}$ are *theme-aligned* if they discuss at least a portion of similar themes (e.g., Wikipedia articles in different languages discussing the same subject, news stories about the same event). Here, nd_i denotes the total number of documents in the corpus $\mathcal{C}_i$, while nw_i is the total number of words in V^i, and d_j^i denotes the j-th document in collection $\mathcal{C}_i$.

Def. 2. Multilingual topic modeling. A multilingual topic model of a multilingual corpus $\mathcal{C}$ is a set of semantically coherent multinomial distributions of words with values $P_i(w^i|z_k)$, $i = 1, \ldots, l$, for each vocabulary $V^1, \ldots, V^i, \ldots, V^l$ associated with text collections $\mathcal{C}_1, \ldots, \mathcal{C}_i, \ldots, \mathcal{C}_l \in \mathcal{C}$ given in languages $L_1, \ldots, L_i, \ldots, L_l$. w^i denotes a word from vocabulary V^i, and $P_i(w^i|z_k)$ is calculated for each $w^i \in V^i$. The probability scores $P_i(w^i|z_k)$ build *per-topic word distributions*. They constitute a language-specific representation (e.g., a probability value is assigned only for words from V^i) of a language-independent cross-lingual concept, that is, cross-lingual topic $z_k \in \mathcal{Z}$.

$\mathcal{Z} = \{z_1, \ldots, z_K\}$ represents the set of all K cross-lingual topics present in the multilingual corpus. Each document in the multilingual corpus is thus considered a mixture of K cross-lingual topics from the set $\mathcal{Z}$. That mixture for a document $d^i \in \mathcal{C}_i$ is modeled by the probability scores $P_i(z_k|d^i)$ that build *per-document topic distributions*.

For instance, having a multilingual collection in English, Italian and Dutch and discovering a cross-lingual topic on *Tourism*, that topic would be represented by words (actually probabilities over words) {*tourist, hotel, travel, ...* } in English, {*albergo (hotel), viaggio (journey), viaggiatore (traveller), ...* } in Italian, and {*reis (travel), toerisme (tourism), hotel (hotel), ...* } in Dutch. We have $\sum_{w^i \in V^i} P_i(w^i|z_k) = 1$, for each vocabulary V^i representing language L_i, and for each topic $z_k \in \mathcal{Z}$. Documents in Italian, English or Dutch discussing themes related to tourism will assign a high importance (by their per-document topic distributions) to this specific cross-lingual topic.

We say that a topic is *semantically coherent* if it assigns high probability scores to words that are semantically related. A desirable property of the cross-lingual topics learned from a theme-aligned corpus is to display both a strong *intra semantic coherence*, that is, words from the same vocabulary grouped together in the same topic are closely semantically related, as well as a strong *inter semantic coherence*, i.e., words across languages that represent the same cross-lingual topic are also closely semantically related. These properties are satisfied when a multilingual topic model is trained on a theme-aligned corpus.

Def. 3. Cross-lingual topic extraction. Given a theme-aligned multilingual corpus $\mathcal{C}$, the goal is to learn and extract a set $\mathcal{Z}$ of semantically coherent K cross-lingual topics $\{z_1, \ldots, z_K\}$ that optimally describe the observed data, that is, the multilingual corpus $\mathcal{C}$. Extracting cross-lingual topics actually implies learning *per-document topic distributions* for each document in the corpus, and discovering language-specific representations of these topics given by *per-topic word distributions* in each language. In the monolingual setting, the set $\mathcal{Z}$ contains monolingual topics only.

Multilingual topic models could be learned on one multilingual corpus, and then inferred on previously unseen documents, where the *inference* in this context denotes inferring per-document topic distributions for the new documents based on the training output of the model. We will exploit this property in our estimation technique.

3 Estimating Cross-Lingual Relevance Models

In this section, we operate in the cross-lingual setting, and present the modeling steps of a CLIR model that combines relevance modeling and multilingual topic modeling. The modeling in the monolingual setting may be observed as an easier special case.

In recent years, numerous language modeling techniques were proposed to deal with the task of cross-lingual information retrieval. The common approach is to perform a word-by-word translation of a query in the source language to the target language by means of word translation probabilities [1,7,2,21]. The translation probabilities are obtained from a bilingual dictionary or are induced from parallel corpora using alignment models for statistical machine translation [5,16], or association measures based on hypothesis testing. However, cross-lingual relevance models [11] proved superior compared to these models in the CLIR tasks, but they still critically rely on a translation resource such as a bilingual dictionary or an in-domain parallel corpus.

3.1 Estimating Cross-Lingual Relevance Models by a Multilingual Topic Model

An Insight into Relevance Modeling. In general, the term *relevance model* addresses a probability distribution that specifies the expectancy that any given word is observed in a set of documents relevant to the issued query. Assume that we are given the query $Q^S = q_1^S, \ldots, q_m^S$ in the source language S with vocabulary V^S, and let R_Q^T be the set of documents that are relevant to the source query Q^S. Let us assume that we operate in the cross-lingual context, with the set R_Q^T and the document collection $\mathcal{D}^T = \{D_1^T, \ldots, D_J^T\}$ given in the target language T with vocabulary V^T. The ranking of documents in $\mathcal{D}^T$ could be achieved if one had a way to estimate the relevance model of the source query Q^S, but in the target language, that is, the set of probabilities $P(w^T|R_Q^T)$ for each word $w^T \in V^T$, where $P(w^T|R_Q^T)$ denotes the probability that we will randomly sample exactly the target word w^T from a relevant document in the target language. Relevance models serve as a powerful and robust retrieval framework, due to its implicit massive query expansion (since the value $P(w^T|R_Q^T)$ is calculated for each $w^T \in V^T$) and its implicit disambiguation [12,11].

Cross-Lingual Estimation. Here, we face two major problems in the cross-lingual setting: (1) We typically do not possess any knowledge of which documents comprise the set R_Q^T, (2) We have to bridge the gap between different languages, and model the concept of sampling a source query word from a target language document.

In order to estimate the relevance model in the absence of any prior knowledge about the set R_Q^T, we follow the usual heuristic presented by Lavrenko et al. [12,11]:

$$P(w^T|R_Q^T) \approx P(w^T|Q^S) = \frac{P(w^T, q_1^S, \ldots, q_m^S)}{P(q_1^S, \ldots, q_m^S)} \tag{1}$$

The probability $P(w^T|Q^S)$ denotes the chance to observe a target word w^T, with respect to a set of underlying distributions $\mathcal{U}$ from which the words are sampled, conditioned on observing m source words $q_1^S, \ldots, q_m^S$ that constitute the source query Q^S. The set $\mathcal{U}$ is typically the target document collection $\mathcal{D}^T$ [11].

Further, Lavrenko and Croft [12] propose a method for estimating the joint probability $P(w^S, q_1^S, \ldots, q_m^S)$ in the monolingual setting when $w^S, q_1^S, \ldots, q_m^S \in V^S$. We adopt their method and adjust it to the cross-lingual setting. The estimate is then:

$$P(w^T, q_1^S, \ldots, q_m^S) = \sum_{D_i^T \in \mathcal{D}^T} P(D_i^T) \left(P(w^T|D_i^T) \prod_{r=1}^{m} P(q_r^S|D_i^T) \right) \tag{2}$$

This estimation model assumes that Eq. (2) is calculated over every document in $\mathcal{D}^T$, and it is repeated for each word $w^T \in V^T$. In case of a large vocabulary and a huge document collection, the estimation is almost computationally infeasible. Therefore, we need to an approximate, computationally tractable estimation of the probability $P(w^T|R_Q^T)$. We adapt the solution proposed by Lavrenko et al. [11]. The probability $P(w^T|R_Q^T)$ may be decomposed as:

$$P(w^T|R_Q^T) = \sum_{D_i^T \in \mathcal{D}^T} P(w^T|D_i^T) P(D_i^T|q_1^S, \ldots, q_m^S) \tag{3}$$

The posterior probability $P(D_i^T|q_1^S,\ldots,q_m^S)$ then may be expressed as:

$$P(D_i^T|q_1^S,\ldots,q_m^S) = \frac{P(D_i^T)\prod_{r=1}^{m} P(q_r^S|D_i^T)}{\sum_{D_j^T\in\mathcal{D}^T} P(D_j^T)\prod_{r=1}^{m} P(q_r^S|D_j^T)} \tag{4}$$

This probability has negligible near-zero values for all but a few documents D_i^T from the collection. These target documents are exactly the documents that obtain the highest scores for the source query Q^S. In order to speed up the retrieval process, we have decided to calculate Eq. (3) over only the top M target documents for the query Q^S (e.g., initially ranking them by a query likelihood model as described by Eq. (6)), instead of calculating Eq. (2) over the entire collection [11,13].

Now, we have to model the probabilities that constitute Eq. (3) and Eq. (4). $P(D_i^T)$ denotes some prior distribution over the dataset which is usually assumed as uniform. For estimation of the probabilities $P(w^T|D_i^T)$ and $P(q_r^S|D_i^T)$, we will utilize the knowledge from a multilingual topic model.

Assume that we have a multilingual topic model trained on a general theme-aligned corpus $\mathcal{C}$ comprising languages S and T (see Def. 1). The model is then inferred on the document collection $\mathcal{D}^T$, that is, each $D_i^T \in \mathcal{D}^T$ may be represented by per-document topic distributions with scores $P(z_k|D_i^T)$, where $z_k \in \mathcal{Z}$ is a cross-lingual topic (see Def. 2). Additionally, since each topic is actually a probability distribution over vocabulary words, each word w, regardless of its language, is assigned a probability $P(w|z_k)$. Thus, if words $q_r^S \in V^S$ and $w^T \in V^T$ were observed during the training of the topic model, they will get the corresponding scores $P(q_r^S|z_k)$ and $P(w^T|z_k)$. We can now easily calculate the probabilities $P(w^T|D_i^T)$ and $P(q_r^S|D_i^T)$ using the shared cross-lingual topic space:

$$P(w^T|D_i^T) = \sum_{k=1}^{K} P(w^T|z_k)P(z_k|D_i^T) \qquad P(q_r^S|D_i^T) = \sum_{k=1}^{K} P(q_r^S|z_k)P(z_k|D_i^T) \tag{5}$$

Note that there is conceptually no difference between the monolingual calculation and the calculation across languages. However, Wei and Croft [22] detected that a document representation that relies only on a topic model is too coarse to be used as the only representation. To obtain the final estimation model, the MuTM representation from Eq. (5) may be linearly combined with the original document model (DM) [22,20]:

$$P(q_r^S|D_i^T) = \lambda\Big((1-\delta)\big(\frac{N_d}{N_d+\mu}P'(q_i^S|D_i^T) + (1-\frac{N_d}{N_d+\mu})P'(q_i^S|\mathcal{D}^T)\big) + \delta P(q_i^S|Ref^S)\Big) + (1-\lambda)\sum_{k=1}^{K} P(q_r^S|z_k)P(z_k|D_i^T) \tag{6}$$

Due to a lack space, we omit the similar equation for estimating $P(w^T|D_i^T)$. Here, $P'(q_r^S|D_i^T)$ denotes the maximum likelihood estimate of the word q_r^S in the target document D_i^T, $P'(q_r^S|\mathcal{D}^T)$ denotes the maximum likelihood estimate of the word q_r^S in the entire document collection $\mathcal{D}^T$. $P(q_i^S|Ref^S)$ is the background probability of observing the word q_r^S in a large source reference corpus. Finally, δ is a tunable parameter which gives a non-zero probability for words that have zero occurrences in the test

collection, λ is an interpolation parameter which assigns weights to the MuTM representation and the DM representation, N_d denotes the length of the document in the number of words, and μ is the parameter of the Dirichlet prior [25]. The final combined estimation is called the *MuTM+DM* model. This estimation model assumes that a proportion of words, such as named entities, remains intact across languages (e.g., when a user searches for Ban Ki-moon, his name remains unchanged in Italian, English or Dutch), which is mostly true for related languages. For more distant languages, other methods were proposed [23,20], but it is out of the scope of this work.

Final Retrieval Model. We may now summarize the entire retrieval process that combines the knowledge from multilingual topic models with the framework of cross-lingual relevance modeling:

1. Train a multilingual topic model on a large theme-aligned corpus and obtain a set $\mathcal{Z}$ of language-independent cross-lingual topics.
2. Infer the topic model on a target document collection $\mathcal{D}^T$.
3. Perform the *first retrieval round* with a query-likelihood PTM-based cross-lingual retrieval model (we use Eq. (6), but other models are also possible).
4. Keep only M top scoring documents from the previous step as pseudo-relevant documents. Estimate the probability scores $P(q_r^S|D_i^T)$ and $P(w^T|D_i^T)$ using the *MuTM+DM* estimation procedure (again, Eq. (6)), but only for the M documents.
5. Estimate the relevance model $P(w^T|R_Q^T)$ for each $w^T \in V^T$ by calculating Eq. (3) and Eq. (4) over these M documents.
6. Perform the *second retrieval round* over the entire collection $\mathcal{D}^T$.[1] Each document D_i^T is assigned a score that is the relative entropy (the Kullback-Leibler divergence) between a relevance model R_Q^T and a target document D_i^T:

$$KL(R_Q^T||D_i^T) = \sum_{w^T \in V^T} P(w^T|R_Q^T) \log \frac{P(w^T|R_Q^T)}{P(w^T|D_i^T)} \tag{7}$$

7. Rank documents in terms of their increasing relative entropy score.

Note that the proposed framework is able to process source and target words in an uniform way (see Eq. (5) and Eq. (6)), and therefore the same model may be used for monolingual and cross-lingual information retrieval. Moreover, since documents have the same language-independent representation given by the distributions over cross-lingual topics, it allows for retrieving documents from a target collection given in multiple languages. In other words, documents relevant to the query may be in different languages, and the proposed framework is able to process it in an uniform way.

4 Experiments and Results

4.1 Experimental Setup

Topic Model. The multilingual topic model we use in our experiments is a straightforward bilingual extension of the standard monolingual LDA model [3] called bilingual

[1] In a real-life retrieval setting, it is more common and less time-consuming to perform only the re-ranking of the top best scoring documents retrieved in the first retrieval round.

LDA (BiLDA) [14,6,15]. BiLDA is trained on a document-aligned bilingual corpus such as Wikipedia articles or news stories discussing the same events. For the details regarding the modeling assumptions, generative story, training and inference procedure of the BiLDA model, we refer the interested reader to the aforementioned relevant literature. It has already been used in a myriad of cross-lingual tasks such as cross-lingual document classification [15], cross-lingual information retrieval [20] or machine translation [14,19]. We use Gibbs sampling for training and set the number of topics K to 1000. Other parameters of the model are set according to [18]. The output after training is composed of the sets of per-topic word distributions in two languages, and the sets of per-document topic distributions (see Def. 2). In the monolingual setting, we use only one half of our training corpus containing that language, and train the standard monolingual LDA model [3] with the same parameters as for BiLDA.

Training Collections. We use a set of 7,612 document-aligned English-Dutch Wikipedia article pairs to train the BiLDA model. To reduce data sparsity, as in [20], we augment the dataset with 6,206 Europarl English-Dutch document pairs [10]. We do not exploit its alignment at the sentence level. Our final vocabularies consist of 76,555 words in English and 71,168 words in Dutch.

Test Collections. All our experiments were performed on the standard dataset used in the cross-lingual evaluation of the CLEF campaigns. The target collection comprises 190,604 Dutch news articles from the NRC Handelsblad 94-95 and the Algemeen Dagblad 94-95 (NC+AD) newspapers. English queries were extracted from the title and description fields of the CLEF themes for the years 2001-2003. Stop words were removed from queries and documents. We also extracted Dutch queries in order to test the monolingual performance of our systems. The overview is provided in Table 1.

Model Parameters. The parameter of the Dirichlet prior from Eq. (6) is set to the standard value of 1000 [22,24]. The parameter δ contributes to the theoretical soundness of our models, but, due to simplicity, we fix it to a negligible near-zero value. The interpolation parameter λ is set to the value of 0.3 which assigns more weight to the MuTM representation. To estimate the relevance model of a query in all models, we use $M = 50$ top scoring documents from the first retrieval round, according to Lavrenko and Allan [13]. They present the full analysis of the impact of reducing the number of documents to only top M documents considered for expansion on the speed and effectiveness of relevance-based retrieval models.

Retrieval Models. We carry out an evaluation of the following models:

Table 1. Statistics of the CLEF Dutch corpus and the CLEF themes. Net queries denote the number of queries that have at least one relevant document.

Collection	Contents	# of Docs	CLEF Themes	Net queries	Campaign label
NC+AD	NRC Handelsblad 94-95 & Algemeen Dagblad 94-95	190,604	41-90	50	CLEF-2001
			91-140	50	CLEF-2002
			141-200	56	CLEF-2003

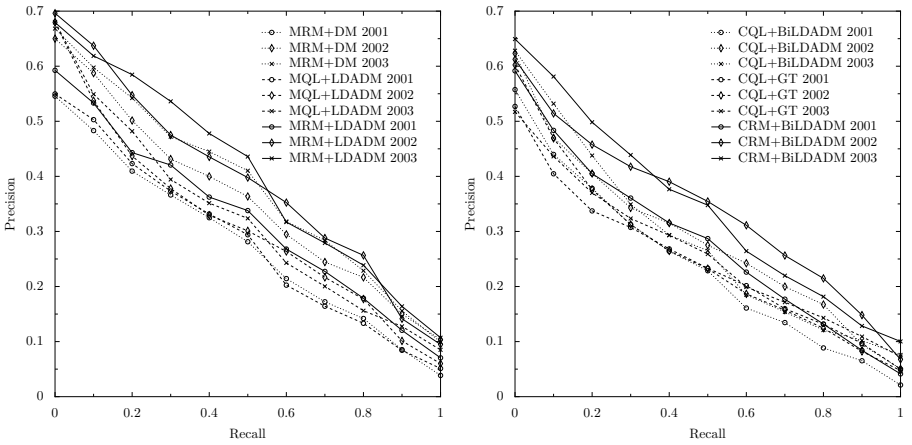
1. Monolingual relevance model estimated using only the document model representation (the first row of Eq. (6)).The model is estimated according to [12]. It was used before as a strong monolingual baseline [11,24] (the **MRM+DM** model).
2. Monolingual query likelihood LDA-based retrieval model that linearly combines the DM and the topic model (LDA) representation as in Eq.(6) [22] (**MQL+LDADM**).
3. Monolingual relevance model estimated using both the DM and the topic model representation (according to Eq. (6)). Our goal is to test whether combining relevance modeling and topic modeling in the monolingual setting also leads to a better model and, consequently to a stronger monolingual baseline (**MRM+LDADM**).
4. Cross-lingual query likelihood BiLDA-based retrieval model that linearly combines the DM and the topic model (BiLDA) representation as given by Eq. (6) [20] (**CQL+BiLDADM**).
5. Cross-lingual translation model which uses *Google Translate* to perform a word-by-word translation of the original query as formulated by [23] and then effectively performs monolingual retrieval using both the DM and the topic model representation as in the previous MQL+LDADM model (**CQL+GT**).
6. Cross-lingual relevance model estimated by Eq. (3), (4) and (6) (see Sect. 3.1), which combines both document representation and MuTM (BiLDA) representation within the relevance modeling framework (**CRM+BiLDADM**).

4.2 Results and Analysis

Our main evaluation criterion is the standard measure of the *mean average precision* (MAP). The MAP scores over all retrieval tasks are displayed in Table 2. Additionally, 11-pt recall-precision curves are presented in Fig. 1(a) and Fig. 1(b), that respectively compare our monolingual and cross-lingual models. Based on these results, we can derive several interesting conclusions. The general important conclusion is that combining the advantages of topic modeling and relevance modeling leads to a better performance of language models for retrieval in both monolingual and cross-lingual

Table 2. MAP scores on the CLEF monolingual and cross-lingual retrieval task with English (and Dutch) queries and Dutch document collection. All relative performances are given with respect to the baseline MRM+DM model performance. Each model is also assigned a unique symbol. The symbols indicate statistically significant differences between the MAP scores in each campaign of every two models to which these symbols are assigned. We use the one-tailed t-test ($p < 0.05$).

Model	CLEF-2001		CLEF-2002		CLEF-2003	
MRM+DM ($\circ$)	$0.2637^{\bullet\diamond\clubsuit}$		$0.3340^{\star\bullet\diamond\clubsuit}$		$0.3539^{\bullet\diamond\clubsuit}$	
MQL+LDADM ($\star$)	$0.2603^{\bullet\clubsuit}$	-1%	$0.2891^{\circ\bullet\diamond\clubsuit\triangle}$	-13%	$0.3262^{\bullet\clubsuit}$	-8%
MRM+LDADM ($\bullet$)	$0.3042^{\circ\star\diamond\clubsuit}$	+15%	$0.3709^{\circ\star\diamond\clubsuit\triangle}$	+11%	$0.3836^{\circ\star\diamond\clubsuit\triangle}$	+8%
CQL+BiLDADM ($\diamond$)	$0.2275^{\circ\bullet\triangle}$	-14%	$0.2683^{\circ\star\bullet\triangle}$	-20%	$0.2783^{\circ\bullet\clubsuit\triangle}$	-21%
CQL+GT ($\clubsuit$)	$0.2296^{\circ\star\bullet\triangle}$	-13%	$0.2401^{\circ\star\bullet\triangle}$	-28%	$0.2443^{\circ\star\bullet\diamond\triangle}$	-31%
CRM+BiLDADM ($\triangle$)	$0.2689^{\diamond\clubsuit}$	+2%	$0.3372^{\star\bullet\diamond\clubsuit}$	+1%	$0.3351^{\bullet\diamond\triangle}$	-5%



(a) Monolingual retrieval (b) Cross-lingual retrieval

Fig. 1. 11-pt recall-precision curves for all models over all campaigns. The positive synergy between probabilistic topic modeling and relevance modeling is clearly visible in both the monolingual setting and the cross-lingual setting. The similar relative performance is observed in the reverse retrieval direction (Dutch queries, English documents) and in the English monolingual retrieval task, but we do not report it due to space constraints.

contexts. The MRM+LDADM model that uses both the original document representation and the topic model representation outperforms a strong monolingual baseline (the MRM+DM model) that also relies on relevance modeling, but utilizes only the original document representation to estimate the relevance model. Thus, the MRM+LDADM should be used as a stronger monolingual baseline. Additionally, comparisons between MRM+LDADM and MQL+LDADM, and MRM+LDADM and MRM+DM reveal that both relevance modeling and topic modeling are significant factors in constructing high quality retrieval models, and the most powerful and robust retrieval models are built by combining the two. Another important remark is that all previous work on topic models in ad-hoc monolingual retrieval relied on in-domain corpora to train the models and learn the topical structure [22,24] (i.e., they train on newswire corpora and perform retrieval on another newswire corpus). Here, we show that such models may also benefit from the topical knowledge coming from a general corpus such as Wikipedia.

In the cross-lingual setting, it is again visible that the CRM+BiLDADM model, which combines relevance modeling and two different representations of a document, outperforms the two other CLIR models by a significant margin. A simple probabilistic word-to-word translation model (CQL+GT) is not sufficient to fully capture the semantics of the query and disambiguate the query terms. On the other hand, cross-lingual topics have an ability to capture the semantics of the query, as the query words are likely to be generated by particular cross-lingual topics and, consequently, a higher preference is assigned to documents dominated by these most likely topics in their topic representation. Cross-lingual topics serve as a bridge between two languages and as implicit query disambiguation tool, but a simple query likelihood model such as CQL+BiLDADM

[20] is still not sufficient to obtain results comparable to the monolingual retrieval models. However, by integrating the topical knowledge in the proposed cross-lingual relevance modeling framework, we are able to build a CLIR model (CRM+BiLDADM) that outscores that simple query likelihood CLIR model. The CRM+BiLDADM model is more complex and has a higher computational complexity, but it is more robust and effective. A comparison of the CRM+BiLDADM model with the monolingual baselines reveals that its performance is on a par with the MRM+DM model which does not rely on any topical knowledge, and it reaches up to 90% of the average performance of the MRM+LDADM model, which is conceptually the same model, but operating in the monolingual setting. We believe that CRM+BiLDADM displays an excellent overall performance, especially taking into account that it does not utilize any translation resource and relies only on a general non-parallel corpus for training.

5 Conclusions

We have proposed a unified framework for ad-hoc monolingual and cross-lingual information retrieval that combines the modeling advantages of multilingual topic modeling and relevance modeling. Multilingual topic models have a capability to represent each document in a collection as a mixture of language-independent concepts, that is, cross-lingual topics, regardless of the actual language of the documents. Relevance models additionally provide a robust framework for a massive query expansion and disambiguation. We have presented an estimation procedure for the relevance models by means of a multilingual topic model that relies only on general data easily obtainable from the Web (e.g., Wikipedia articles). The proposed framework is generic, language-independent and model-independent, as it allows for inputting any multilingual topic model that outputs the sets of per-topic word and per-document topic distributions in the relevance modeling framework. Additionally, the framework is able to process documents in the target collection in an uniform way regardless of their actual language.

We have conducted a thorough analysis of our models within a real-life setting of the CLEF retrieval tasks, with the CLEF test collection of news stories comprising nearly 200,000 documents. Our results show that the topical knowledge learned on a general corpus is useful when combined with the framework of relevance modeling in both monolingual and cross-lingual contexts. Additionally, current state-of-the-art CLIR models that exploit the topical knowledge [22,20] are outperformed by the model built within this novel framework. In this work, we have used the standard multilingual extension of the LDA model, but one path of future research might lead to designing other topic models that better fit specific retrieval tasks.

References

1. Ballesteros, L., Croft, W.B.: Phrasal translation and query expansion techniques for cross-language information retrieval. In: Proceedings of ACM SIGIR, pp. 84–91 (1997)
2. Berger, A., Lafferty, J.: Information retrieval as statistical translation. In: Proceedings of ACM SIGIR, pp. 222–229 (1999)
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. Journal of Machine Learning Research (3), 993–1022 (2003)

4. Boyd-Graber, J., Blei, D.M.: Multilingual topic models for unaligned text. In: Proceedings of UAI, pp. 75–82 (2009)
5. Brown, P.F., Pietra, V.J.D., Pietra, S.A.D., Mercer, R.L.: The mathematics of statistical machine translation: parameter estimation. Computational Linguistics 19(2), 263–311 (1993)
6. De Smet, W., Moens, M.-F.: Cross-language linking of news stories on the Web using interlingual topic modeling. In: Proceedings of the CIKM Workshop on Social Web Search and Mining (SWSM), pp. 57–64 (2009)
7. Hiemstra, D., de Jong, F.: Disambiguation Strategies for Cross-Language Information Retrieval. In: Abiteboul, S., Vercoustre, A.-M. (eds.) ECDL 1999. LNCS, vol. 1696, pp. 274–293. Springer, Heidelberg (1999)
8. Hofmann, T.: Probabilistic Latent Semantic Indexing. In: Proceedings of ACM SIGIR, pp. 50–57 (1999)
9. Jagarlamudi, J., Daumé III, H.: Extracting Multilingual Topics from Unaligned Comparable Corpora. In: Gurrin, C., He, Y., Kazai, G., Kruschwitz, U., Little, S., Roelleke, T., Rüger, S., van Rijsbergen, K. (eds.) ECIR 2010. LNCS, vol. 5993, pp. 444–456. Springer, Heidelberg (2010)
10. Koehn, P.: Europarl: A parallel corpus for statistical machine translation. In: Proceedings of the MT Summit, pp. 79–86 (2005)
11. Lavrenko, V., Choquette, M., Croft, W.B.: Cross-lingual relevance models. In: Proceedings of ACM SIGIR, pp. 175–182 (2002)
12. Lavrenko, V., Croft, W.B.: Relevance-based language models. In: Proceedings of ACM SIGIR, pp. 120–127 (2001)
13. Lavrenko, V., Allan, J.: Real-time query expansion in relevance models. CIIR Technical Report IR-473 (2006)
14. Mimno, D., Wallach, H.M., Naradowsky, J., Smith, D.A., McCallum, A.: Polylingual topic models. In: Proceedings of EMNLP, pp. 880–889 (2009)
15. Ni, X., Sun, J.T., Hu, J., Chen, Z.: Cross lingual text classification by mining multilingual topics from Wikipedia. In: Proceedings of WSDM, pp. 375–384 (2011)
16. Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. Computational Linguistics 29(1), 19–51 (2003)
17. Ponte, J.M., Croft, W.B.: A language modeling approach to information retrieval. In: Proceedings of ACM SIGIR, pp. 275–281 (1998)
18. Steyvers, M., Griffiths, T.: Probabilistic topic models. Handbook of Latent Semantic Analysis 427(7), 424–440 (2007)
19. Vulić, I., De Smet, W., Moens, M.-F.: Identifying word translations from comparable corpora using latent topic models. In: Proceedings of ACL, pp. 479–484 (2011)
20. Vulić, I., De Smet, W., Moens, M.-F.: Cross-language information retrieval models based on latent topic models trained with document-aligned comparable corpora. Information Retrieval (2012)
21. Wang, J., Oard, D.W.: Combining bidirectional translation and synonymy for cross-language information retrieval. In: Proceedings of ACM SIGIR, pp. 202–209 (2006)
22. Wei, X., Croft, W.B.: LDA-based document models for ad-hoc retrieval. In: Proceedings of ACM SIGIR, pp. 178–185 (2006)
23. Xu, J., Weischedel, R., Nguyen, C.: Evaluating a probabilistic model for cross-lingual information retrieval. In: Proceedings of ACM SIGIR, pp. 105–110 (2001)
24. Yi, X., Allan, J.: A Comparative Study of Utilizing Topic Models for Information Retrieval. In: Boughanem, M., Berrut, C., Mothe, J., Soule-Dupuy, C. (eds.) ECIR 2009. LNCS, vol. 5478, pp. 29–41. Springer, Heidelberg (2009)
25. 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)
26. Zhang, D., Mei, Q., Zhai, C.: Cross-lingual latent topic extraction. In: Proceedings of ACL, pp. 1128–1137 (2010)