Inf Retrieval J (2015) 18:246–281
DOI 10.1007/s10791-015-9255-1

Multilingual information retrieval in the language modeling framework

Razieh Rahimi[1] · Azadeh Shakery[1,2] · Irwin King[3]

Received: 10 September 2014 / Accepted: 16 April 2015 / Published online: 6 May 2015

Abstract Multilingual information retrieval (MLIR) provides results that are more comprehensive than those of mono- and cross-lingual retrieval. Methods for MLIR are categorized as: (1) *Fusion-based* methods that merge results from multiple retrieval runs, and (2) *Direct* methods that build a unique index for the entire collection. Merging results of individual runs reduces the overall effectiveness, while more effective direct methods suffer from either time complexity and memory overhead, or over-weighting of index terms. In this paper, we propose a direct MLIR approach by using the language modeling framework that includes a novel multilingual language model estimation for documents, and a new way to globally estimate word statistics. These contributions enable ranking documents in multiple languages in one retrieval phase without having the problems of the previous direct methods. Moreover, our approach has the advantage of accommodating multilingual feedback information which helps to prevent query drift, and consequently to improve the performance. Finally, we effectively address the common case of incomplete coverage of translation resources in our proposed estimation methods. Experimental results show that the proposed approach outperforms the previous MLIR approaches.

✉ Azadeh Shakery
shakery@ut.ac.ir

Razieh Rahimi
r.rahimi@ece.ut.ac.ir

Irwin King
king@cse.cuhk.edu.hk

[1] School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran

[2] School of Computer Science, Institute for Research in Fundamental Sciences (IPM), P.O. Box 19395-5746, Tehran, Iran

[3] Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, NT, Hong Kong

Keywords Multilingual information retrieval · Multilingual language models · KL-divergence framework · Language modeling framework · Multilingual feedback

1 Introduction

With the rapid proliferation of Web contents in different languages, *multilingual information retrieval* (MLIR) has been inevitably linked to Web search. MLIR enables retrieving documents in multiple languages in response to a user's query. MLIR is more challenging than mono- and cross-lingual information retrieval in that in the former, document collection contains documents in multiple languages, while in the latter, all documents of the collection are in one language. A major challenge in MLIR is the appropriate use of translation knowledge to score documents which might or might not be in the query language.

There are two main architectures for MLIR which are referred to as *distributed* and *centralized* architectures in (Lin and Hsi 2003), and *query translation* and *document translation* approaches in (Peters et al. 2012). Herein, to avoid confusion with federated MLIR (Si et al. 2008) and cross-language IR terminologies (Nie 2010), the two architectures are called fusion-based and direct approaches, respectively. In *direct* approaches, the entire multilingual collection has a unique index (Lin and Hsi 2003) and thus, the multilingual result list can be obtained in one retrieval phase. On the other hand, in *fusion-based* approaches, the multilingual retrieval problem is transformed to a number of retrieval tasks, each of which corresponds to a language in the collection. These retrieval tasks are then followed by a merging phase to create the final ranked list.

Using direct approaches is motivated by substantial performance degradation in the merging phase of fusion-based approaches (Peters et al. 2012). However, existing direct approaches are not as practical either due to time and memory overhead, or over-weighting of index terms. For example, in one direct approach, each document is translated into all languages (typically by using machine translation systems), which not only is time-consuming, but also needs updating with improvements in translation resources (Peters et al. 2012). To ameliorate these problems, the interlingua approach can be adopted (Kraaij and de Jong 2004; Sorg and Cimiano 2012). This approach although reduces the required translation resources by a factor of the number of languages, may decrease the accuracy of retrieval, in that a query and a document in two languages different from the pivot language are indirectly matched through translation to the pivot language. This is because employing direct translation resources when they are available, usually outperforms the use of transitive translations (Nie 2010). In another direct approach (Nie and Jin 2002), instead of translating documents, the query is expanded by translations of query terms in all languages. The drawback of this method is that terms of a language having fewer documents may be over-weighted (Kishida 2005). Therefore, taking advantage of direct approaches requires overcoming the mentioned problems.

In this paper, we propose a direct approach for MLIR with only one retrieval phase which does not suffer from memory or time overhead of translating all documents, and also preserves relative term statistics. Our approach is based on the language modeling framework (Lafferty and Zhai 2001). In this framework, documents are represented by a language model. The language model of a document is a representation of the queries that would be submitted by users interested in that document. The idea behind our approach is

that a document in a multilingual environment is relevant to queries in different languages. Thus, the language models of these documents should reflect this fact. To achieve this, we represent each document by a *multilingual unigram language model* (MULM).

This paper improves the performance of MLIR through the following contributions:

1. We propose a novel way of estimating document language models in a multilingual environment. Using these document models in the KL-divergence framework, documents in multiple languages are ranked in only one retrieval phase, without translating all documents into all languages.
2. We provide more accurate global estimates of two retrieval heuristics, namely term and document frequencies, by simultaneously considering all subcollections in different languages. This way, we avoid over-weighting of index terms.
3. In items 1 and 2 above, we adjust each estimation approach in such a way that it also performs well when translation resources do not provide full coverage of collection words.
4. We show that feedback information from documents in multiple languages can be naturally incorporated to improve the ranking of documents in each subcollection. This feature prevents query drift when documents in one language do not cover the topic of a query.

Results of our experiments reveal that our approach outperforms the previous MLIR approaches; namely fusion-based methods (Powell et al. 2000; Savoy 2003; Martinez-Santiago et al. 2006) and a direct method (Nie and Jin 2003). Furthermore, we achieved MLIR effectiveness between 68 and 93 % of the theoretical optimal effectiveness achievable by any fusion-based method. This is also higher than the percentage reported in the previous work (Martinez-Santiago et al. 2006), which is the state-of-the-art among unsupervised fusion-based methods.

The rest of the paper is organized as follows. In Sect. 2, we review previous work on MLIR. Section 3 discusses basics of the KL-divergence retrieval model. The main approach is presented in Sect. 4 and its features are discussed in Sect. 5. Section 6 describes the experimental setup and evaluation of our approach. Finally, we conclude in Sect. 7.

2 Related work

We divide the MLIR approaches into two groups, fusion-based and direct approaches; fusion-based approaches merge retrieval results from separate indexes, while direct approaches avoid the merging phase by retrieving documents from a single index.

Direct approaches In the direct method proposed by Braschler et al. (2002), all documents of a multilingual collection are translated into the query language. Thus, multilingual retrieval task is reduced to monolingual retrieval on the translated document collection. Subsequent approaches overcome the problem of translating all documents into all languages. Nie and Jin (2002, 2003) build a unique index for a multilingual collection in which each word has a language tag. To rank the documents with respect to a query, the query is expanded by adding the translations of each query term in all languages with their associated probabilities. Documents and multilingual queries are then matched using a TF-IDF weighting scheme. The *inverse document frequency* (IDF) of a term is estimated on the unified multilingual index and is higher than that calculated on the term's respective subcollection. However, the amount of increase in IDF is not equal for terms of different languages due to the difference in the size of respective subcollections. This can cause

improper higher weights for terms of a language with fewer number of documents. In another study, Sorg and Cimiano (2012) make a unique index from inter-lingual representations of documents using Wikipedia.

Fusion-based approaches are further divided into three sub-groups based on their merging strategy.

First group of fusion-based methods use only the ranks and/or scores of documents to merge the result lists from mono- and cross-lingual runs. Traditional approaches such as *round-robin* merging (Savoy 2002; Chen and Gey 2004), *raw-score* merging (Savoy 2003), and *normalized-score* merging (Savoy 2004b; Savoy and Berger 2005; Savoy 2004a; Jones et al. 2005) belong to this group. These approaches are based on some assumptions on the distribution of relevant documents across subcollections, or comparability of retrieval scores from different subcollections.

Second group of fusion-based methods includes methods that try to partially relax the aforementioned limiting assumptions by exploiting more information from underlying subcollections or retrieved lists. For this purpose, Braschler and Schäuble (2000) align documents in different languages in the collection and use this information to make the scores of retrieved documents comparable. Braschler (2004) investigates performance improvement on simple merging strategies through contributing intermediate lists to the final result based on their respective subcollection sizes. Lin and Hsi (2004) consider subcollection characteristics and translation qualities in the merging phase, and propose a weighted combination of intermediate results. Multiple parameters are used for determining the combination weights which makes tuning difficult. Martinez-Santiago et al. (2006) perform an additional retrieval step to merge the intermediate results. The second retrieval step is to rank the top-k documents, retrieved in the first step, with respect to an expanded query through translations of its terms in all languages.

Third group of fusion-based approaches apply machine learning techniques to the merging problem. Le Calvé and Savoy (2000) explore the use of logistic regression to learn the probability of relevance for each document based on its score and the logarithm of its rank. Si and Callan (2006) also propose a query- and language-specific result merging algorithm by using the logistic model. Gao et al. (2009) define features of the learning method based on a document's similarities with a query, with other retrieved documents in its language, and with retrieved documents in other languages. Tsai et al. (2008) define several features, such as technical terms and person and organization names, for learning the document ranks. Extracting some of these defined features highly depends on the availability of language-specific tools. Therefore, these features might not be available for languages with limited resources.

Language modeling framework In addition to monolingual information retrieval, the language modeling framework has been used for cross-lingual (bilingual) information retrieval (Xu et al. 2001; Lavrenko et al. 2002; Kraaij et al. 2003). These approaches try to adopt the idea of translation models in monolingual information retrieval, proposed in (Berger and Lafferty 1999), to CLIR. Thus, these approaches are applicable when documents of a collection are written in one language and the goal is to rank these documents w.r.t. a query in another language. Adopting these approaches for MLIR needs a merging phase to combine the results which are separately generated by the language modeling framework for each language (monolingual and bilingual runs). Although the language modeling framework has a good performance in ranking documents of each subcollection, the merging phase can degrade the performance of MLIR. Therefore, ranking documents of a multilingual collection in one retrieval phase is preferred to avoid performance degradation through merging. To the best of our knowledge, this work is the first direct

approach for MLIR using the language modeling framework. We use different merging algorithms on the individual results produced by the language modeling framework, as baselines for evaluating our approach.

Our approach is similar to the approaches of (Kraaij et al. 2003; Xu et al. 2001) in terms of building new language models for documents. But, in (Kraaij et al. 2003; Xu et al. 2001), the new language model of a document has the same number of parameters as the number of words in the source (query) language, while in our approach, the number of parameters is the same as the number of words in all languages in order to enable direct retrieval.

In addition to removing the merging phase, our approach naturally allows adopting feedback information from one subcollection to improve the retrieval performance on other subcollections. Getting assistance of subcollections in other languages is not directly applicable in fusion-based methods for MLIR, independent of which retrieval model is used for generating the individual lists. Therefore, using the language modeling approaches of (Kraaij et al. 2003; Xu et al. 2001) for MLIR cannot provide this advantage.

3 Language modeling approach

In this section, we briefly review the basics of the language modeling framework which are required to describe our approach.

Monolingual information retrieval The Kullback–Leibler (KL) divergence retrieval model is considered as the state-of-the-art for retrieval using the language modeling approach. Using the KL-divergence model for retrieval, the score of a document D with respect to a query Q is calculated as (Lafferty and Zhai 2001):

$$\begin{aligned} \text{Score}(Q,D) &= -D_{KL}(\theta_Q \| \theta_D) \\ &\stackrel{\text{rank}}{=} \sum_{w \in v} p(w|\theta_Q) \log p(w|\theta_D), \end{aligned} \tag{1}$$

where θ_Q and θ_D are the estimated query and document language models, respectively, and v is the vocabulary set. Assuming a multinomial model, the basic approach to estimate document language models is the *Maximum likelihood* (ML) estimator. According to this estimator, word probabilities are estimated as follows:

$$p_{ml}(w|\theta_D) = \frac{c(w,D)}{|D|}, \tag{2}$$

where $c(w,D)$ is the count of word w in document D and $|D|$ is the length of D. Maximum likelihood estimator assigns zero probabilities to unseen words in a document, causing problems in scoring the document using Eq. (1) (Zhai and Lafferty 2001b). Smoothing methods address this problem by discounting the probabilities of observed words in the document and assigning non-zero probabilities to unseen words. One commonly used smoothing technique is *Dirichlet Prior* (Zhai and Lafferty 2001b) in which the language model for a document D is estimated as:

$$p(w|\theta_D) = \frac{|D|}{|D|+\mu} p_{ml}(w|D) + \frac{\mu}{|D|+\mu} p(w|C), \tag{3}$$

where μ is the smoothing parameter and $p(.|C)$ is the collection language model.

Feedback The KL-divergence framework provides a principled way to leverage feedback information in order to improve the estimation of the query language model. In *model-based feedback* (Zhai and Lafferty 2001a), the query language model is updated using the feedback model estimated based on the feedback documents:

$$p\left(w|\theta'_Q\right) = \lambda p(w|\theta_Q) + (1-\lambda)p(w|\theta_F), \tag{4}$$

where θ'_Q is the new language model for the query, θ_F is the estimated feedback model, and $0 \leq \lambda \leq 1$ is the interpolation parameter.

Cross-Lingual IR The KL-divergence retrieval model can also be used for cross-language information retrieval (CLIR). In CLIR, the language of the query is different from that of the documents. Therefore, to score documents with respect to a given query, we need to integrate the translation model into either the query or the document language model (Nie 2010). The translation model, in the basic form, includes a translation probability for each pair of source- and target-language words. In *query translation* approach, a new language model is built for the query and documents are ranked using:

$$\text{Score}(Q,D) = \sum_{w_t \in V_t} p(w_t|\tilde{\theta}_Q) \log p(w_t|\theta_D), \tag{5}$$

$$\begin{aligned} p(w_t|\tilde{\theta}_Q) &= \sum_{w_s \in V_s} p(w_t|w_s,\theta_Q)p(w_s|\theta_Q) \\ &\approx \sum_{w_s \in V_s} p(w_t|w_s)p(w_s|\theta_Q), \end{aligned} \tag{6}$$

where w_s (w_t) are source (target) words belonging to V_s (V_t) which are the source (target) language vocabulary, respectively. $p(w_t|w_s)$ indicates the probability of translating the source word w_s to the target word w_t. In *document translation* approach, the translation model is integrated into the document language models and the score of a document D is given by:

$$\text{Score}(Q,D) = \sum_{w_s \in V_s} p(w_s|\theta_Q) \log p(w_s|\tilde{\theta}_D), \tag{7}$$

$$\begin{aligned} p(w_s|\tilde{\theta}_D) &= \sum_{w_t \in V_t} p(w_s|w_t,\theta_D)p(w_t|\theta_D) \\ &\approx \sum_{w_t \in V_t} p(w_s|w_t)p(w_t|\theta_D), \end{aligned} \tag{8}$$

where $p(w_s|w_t)$ indicates the probability of translating the target word w_t to the source word w_s. We call this approach *LM-based document translation* to avoid confusion with *traditional document translation* approach that literally translates the whole document and then indexes the translated document.

4 Extension of the language modeling framework to MLIR

In this section, we describe our approach for MLIR.

Problem definition Suppose that we have a multilingual collection C where its documents are written in N different languages $\{l_i\}_{i=1}^{N}$. For each language pair, we are given a

translation model of the form $p(w|u)$ which indicates the probability of translating word u in one language to word w in another language. The goal is to optimize the effectiveness of multilingual information retrieval given the available translation models. In particular, we aim to estimate the score of a document D ($D \in C$) with respect to a query Q in language l_j ($1 \leq j \leq N$) in order to provide a ranking of documents in multiple languages $\{l_i\}_{i=1}^{N}$.

Before presenting the solution, let us first define some notations. By dividing the collection C based on document languages, we get N subcollections $\{C_i\}_{i=1}^{N}$ where $C = \cup_{i=1}^{N} C_i$ and all documents in each subcollection C_i are in the same language l_i. We define v_i as the vocabulary set of language l_i and $V = \cup_{i=1}^{N} v_i$ as the vocabulary set of the entire collection. Words of each language are labeled with a language tag similar to (Nie and Jin 2002). Thus, common words between languages are considered as separate words and $|V| = \sum_{i=1}^{N} |v_i|$. The reason behind this decision is described in Sect. 5.1. The probabilities in the translation model for each language pair (l_i, l_j), $1 \leq i, j \leq N$ and $i \neq j$, are denoted by $p_{ij}(w|u)$ which indicates the probability of translating word u in language l_j to word w in language l_i. The given translation probabilities are normalized such that $\sum_{w \in v_i} p_{ij}(w|u) = 1$ for each pair (l_i, l_j), $1 \leq i, j \leq N$, of languages. In addition, self translation probabilities are set to one, i.e., $p_{ii}(w|w) = 1$. In the following, we describe how to perform multilingual information retrieval using the language modeling framework.

4.1 Multilingual unigram language model

Multilingual language model for document representation As mentioned before, a document in a multilingual environment can be retrieved with respect to queries in different languages. Hence, we should extend the basic estimation of document language models to further support these queries. In all basic language modeling approaches, the parameters of the document language model are estimated by considering the document as the observed data, which has the problem that the document might be a small sample of its language model. Extending language models of the documents to support queries in different languages makes the estimation of document language models more challenging, because the document has no term in many query languages. To tackle this issue, we define the *probabilistic count* of term $w \in v_i$ (belonging to language l_i) in document D written in language l_j as:

$$c_p(w, D) = \sum_{u \in D} p_{ij}(w|u) c(u, D), \tag{9}$$

where $c(u, D)$ is the real count of term u in document D. Defining probabilistic counts of terms is intuitively analogous to expanding each document with terms of other languages than the document language. The expanded documents are then considered as bags of multilingual words, i.e., we assume independence between terms of expanded documents. This simplifying assumption is also made in estimating document language models in monolingual information retrieval (Zhai 2008).

We build a new multilingual language model for document D, denoted by $\hat{\theta}_D$, considering the probabilistic counts of words instead of only the real counts. To estimate the parameters of $\hat{\theta}_D$, we follow the well-established *unigram multinomial* language model, where the probability of generating a sequence of words is obtained by multiplying the probabilities of generating each of its words, assuming that words are generated

independently. Therefore, the parameters of this multilingual model are $\{p(w_i|\hat{\theta}_D)\}_{i=1}^{|V|}$, i.e. all terms of all languages. The maximum likelihood estimator gives us:

$$p_{ml}\left(w|\hat{\theta}_D\right) = \frac{c_p(w,D)}{\sum_{u\in V} c_p(u,D)} = \frac{c_p(w,D)}{N|D|}, \tag{10}$$

where N is the number of different languages in the collection, $|D|$ is the length of document D accounting real counts of its terms, and $N|D|$ represents the new size of document D considering the probabilistic counts. For illustration, consider the example in Fig. 1a. In the basic ML-estimated language model for document D_1 in the figure, term a has probability 1, while term a as well as its translation, term α, has probability 0.5 within the multilingual language model built by our approach (Eq. (10)) using the probabilistic counts shown in Fig. 1b.

The estimates provided in Eq. (10) suffer from the same problem of underestimating probabilities for words that have zero probabilistic counts in a document, computed according to Eq. (9). To address this issue, we can generalize existing smoothing techniques to be applicable on our multilingual language model. We consider here a smoothing method that uses a reference language model. First, we proceed with the estimation of the reference language model in our retrieval model.

New reference language model To estimate the reference language model for smoothing techniques, probabilistic counts of words in all documents of the entire collection are used. That is,

$$p'(w|C) = \frac{\sum_{D\in C} c_p(w,D)}{\sum_{D\in C}\sum_{u\in V} c_p(u,D)} = \frac{\sum_{D\in C} c_p(w,D)}{N\sum_{D\in C}|D|}, \tag{11}$$

where $p'(.|C)$ denotes the new reference language model. This estimate of the reference language model considers the probabilistic counts of each word in all subcollections, rather than only the subcollection that actually includes that word. Therefore, this collection language model can be considered as an *expanded estimate* of the reference language model, compared to the *ML-estimate*,

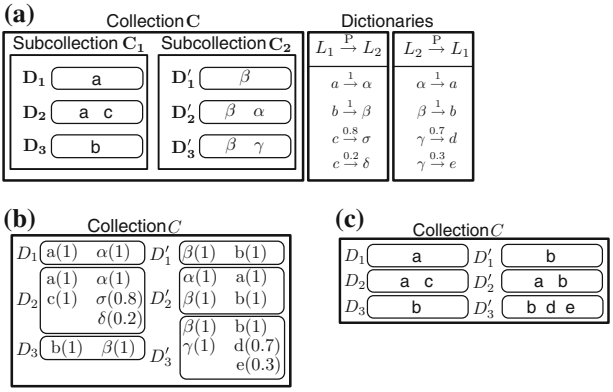

Fig. 1 A sample collection. **a** A sample collection and dictionary. *Numbers* above the *arrows* indicate translation probabilities. **b** Conceptual representations of documents considering probabilistic counts of terms (*numbers in parentheses*). **c** Literally translated documents into one language l_1

$$p(w|C) = \frac{\sum_{D \in C} c(w, D)}{\sum_{D \in C} |D|}, \tag{12}$$

which is equal to $\frac{\sum_{D \in C_i} c(w,D)}{\sum_{D \in C} |D|}$ if word w belongs to language l_i. The effect of counting word occurrences in all subcollections is investigated in more detail in Sect. 5.3.

Smoothing The new reference language model (Eq. (11)) should be used in smoothing techniques that need a fallback model. For example, the smoothed multilingual language model for a document D using Dirichlet Prior smoothing technique is estimated as:

$$p\left(w|\hat{\theta}_D\right) = \frac{N|D|}{N|D| + \mu} p_{\text{ml}}\left(w|\hat{\theta}_D\right) + \frac{\mu}{N|D| + \mu} p'(w|C). \tag{13}$$

Ranking documents By substituting the smoothed multilingual language models for documents in Eq. (1), the score of each document in the collection can be calculated with respect to any given query, independent of the original language of the document. Ranking based on these scores gives us a multilingual result list without the need to merge different ranked lists.

4.2 Dictionary coverage

The estimates of the document and the reference language models, described in the previous subsection, are valid when the translation resource provides full coverage of the words in the vocabulary set V, which may not be satisfied in practice. Despite having high quality translation resources, there may be several words in the collection with no entry in the translation resources, due to out of vocabulary, misspelled, or informal words in the collection. The first solution can be extending each translation model by translation relations implied by the transitivity through a pivot language. But, this solution only reduces the severity of the problem and does not yield a translation resource with full coverage. In case of incomplete coverage of dictionary, $\sum_{u \in V} c_p(u, D)$ in Eq. (10) is not equal to $N|D|$. Therefore, we have:

$$p_{ml}\left(w|\hat{\theta}_D\right) = \frac{c_p(w, D)}{\sum_{u \in V} c_p(u, D)}. \tag{14}$$

Length ratio However, estimation of document language models using Eq. (14) implies disregarding words with no entry in the translation resource, which does not preserve the length ratio of documents. In particular, the length ratio of documents may differ when we count real occurrences of words in the documents compared to when we sum the probabilistic counts of words. This contradicts a retrieval axiom which is explored in detail in Sect. 5.2. To resolve this problem, we consider dummy words as the translations of words with no entry in the translation resource. These dummy words do not match any query term, but help to preserve the length ratio of documents. Therefore, the language model of a document D using the maximum likelihood estimator is estimated by the following equation considering dummy words:

$$p_{ml}\left(w|\hat{\theta}_D\right) = \frac{c_p(w, D)}{N|D|}. \tag{15}$$

Particularly, it suffices to consider $N|D|$ as the length of the document D. The new reference language model is also estimated as:

$$p''(w|C) = \frac{\sum_{D \in C} c_p(w, D)}{N \sum_{D \in C} |D|}. \tag{16}$$

Term discrimination value (IDF heuristic) Another side effect of words without translations emerges in discriminating query words based on the reference language model. A frequent term in the collection would have a high probability in the reference language model $p''(w|C)$. Smoothing document language models with the reference language model makes the weights of matched terms between a query and a document in Eq. (1) to have a factor of $1/p''(w|C)$ which consequently causes frequent terms to get penalized[1] (Zhai 2008).

A word with no entry in the translation resource, may get an artificially high discrimination value because of no increase in the frequency of the word with documents of other subcollections, although its translations may be available in other subcollections. Therefore, we cannot solely rely on the reference language model estimated based on the expanded documents to determine the frequent terms. To address this issue, we propose two solutions. The solutions are based on employing the reference language model estimated without considering translations.

The first solution is to combine the expanded and the maximum likelihood estimates of the reference language model. Toward this, we adopt linear interpolation:

$$\hat{p}(w|C) = \beta p''(w|C) + (1 - \beta) p(w|C), \tag{17}$$

where $p''(.|C)$ is the global estimate of word statistics (the expanded estimate of the reference language model), while $p(.|C)$ is the ML-estimate of the reference language model and depends on occurrences of words only in their respective subcollections (Eq. (12)), and $0 \leq \beta \leq 1$ is a weighting parameter that can be determined based on the dictionary coverage. $\hat{p}(.|C)$ can subsequently be employed as the reference language model for smoothing.

The second solution to avoid rewarding words with no entry in the translation resource is to use 2-stage smoothing to estimate the document language models as:

$$p\left(w|\hat{\theta}_D\right) = (1 - \lambda) \frac{c_p(w, D) + \mu p''(w|C)}{N|D| + \mu} + \lambda p(w|C). \tag{18}$$

As mentioned in (Zhai and Lafferty 2002), the purpose of the first stage of smoothing is to explain unseen words in a document. Therefore, for the first stage, we use the reference language model estimated globally using probabilistic counts. The second stage of smoothing is then supposed to reduce the effect of noise words in ranking the documents, for which we use the reference language model estimated based only on the real counts of the words in the collection. In all that follows, we use only this solution and leave the first solution for future work.

4.3 Incorporation of feedback information

In this section, we study the feedback concept in a multilingual environment and its incorporation in our MLIR approach. The purpose of using feedback in the retrieval task is to update a query with feedback information to achieve better performance. In the *relevance feedback* technique, feedback information is obtained from sample relevant

[1] Exact term weights are shown in Eq. (19).

documents, which are substituted in *pseudo relevance feedback* by documents that seem to match the query in an initial retrieval run.

Multilingual feedback model In multilingual retrieval, feedback information should be extracted from documents in different languages, since documents relevant to the query as well as the top ranked documents in an initial run, are generally in different languages. Feedback information, extracted from either set of documents, would subsequently be in multiple languages. Incorporating this conceptual perception of feedback into the *model-based feedback* technique, the topic model (θ_F in Eq. (4)) extracted from feedback documents should be multilingual. In particular, parameters of the topic language model include terms of different languages available in a multilingual collection. Building such a feedback model is not possible with existing approaches mentioned in Sect. 2. In our approach, we can naturally build a multilingual topic model, since language models of documents are multilingual.

Multilingual query After estimating the feedback topic model, the next step is to update the query language model using Eq. (4). Interpolating the query language model with the multilingual feedback model results in a query language model different from the initial query model. The new query model may have terms in different languages, i.e., incorporating feedback information results in a multilingual query. The next step is to score documents of the multilingual collection with respect to this new query model, which imposes additional complexity due to query terms in multiple languages.

To score documents of a multilingual collection with respect to a multilingual query using the basic retrieval models, the query should be translated into one language. Otherwise, only query terms in the language of a document have impact on the score of that document and thus documents of different subcollections are scored with respect to different parts of the query which are not equivalent. In contrast, our proposed approach allows directly retrieving relevant documents to a multilingual query, without any additional query translation, since the new document models have parameters equivalent to the terms of all languages.

Therefore, the great advantage of our approach is that all components of a retrieval framework including query expansion, relevance feedback, and pseudo relevance feedback are directly applicable to multilingual information retrieval.

Knowledge transfer One problem of query expansion through pseudo relevance feedback is query drift that occurs when the collection has few relevant documents with respect to a query. In retrieval on a multilingual collection, one subcollection may have fewer relevant documents to a query compared to the others. Merging the ranked list of documents in this subcollection, generated by applying the pseudo relevance feedback technique, may harm the overall multilingual performance.

Leveraging multilingual feedback information has the remarkable benefit of transferring knowledge between subcollections which prevents query drift. Considering C_i as the subcollection with few relevant documents w.r.t a query, most of the top-ranked documents as well as most of the feedback terms would be in languages other than l_i. In our approach, these feedback terms can also increase the retrieval performance on subcollection C_i although they are in other languages. Since the new language models of documents assign probabilities to the terms of all languages, feedback terms can directly match with documents in subcollection C_i without translation. Therefore, feedback terms in other languages can help to increase the recall measure in subcollection C_i. We deal with this issue in a similar way to (Chinnakotla et al. 2010), but with a lower overhead. In (Chinnakotla et al. 2010), feedback information is obtained from top-ranked documents from an assisting

collection in a language different from the query language. The original query is then updated by translating the obtained feedback terms.

5 Discussions of the proposed multilingual retrieval model

In this section, we analytically study some aspects of the proposed retrieval model using axiomatic analysis (Fang et al. 2011) and also discuss the computational complexity of our approach. Axiomatic analysis is based on formal constraints that any reasonable retrieval model should satisfy.

5.1 Common terms between languages

We first discuss the reason of labeling terms with language tags in the presence of common terms between languages. By common terms, we mean words of two languages that after performing preprocessing steps such as normalization and stemming, have the same spelling. Therefore, common terms include a part of cognates whose spellings are identical, and may include some proper nouns. The reason of labeling terms is described through the following constraint.

MLIR Constraint 1 Consider a collection in two languages l_1 and l_2 which share a common term x. Let $q = \{xz\}$ be a two-term query in language l_1. We are interested in the relative ranking of two documents D_1 and D_2 in language l_1 w.r.t query q, where D_1 contains the common term but D_2 does not. Suppose the following assumptions hold for the documents and given dictionaries:

- $|D_1| = |D_2|$, $c(x, D_1) = c(z, D_2)$, $z \notin D_1$, and $x \notin D_2$.
- Terms x and z have the same discrimination value considering the entire collection.
- Term x in l_1 translates to term x in l_2 with probability greater than 0, but translations of z into l_2 do not belong to v_1. In particular, we have: $p_{21}(x|x) > 0$ and if $p_{21}(\zeta|z) > 0$, then $\zeta \notin v_1$.

Given these assumptions, D_1 and D_2 should get the same score.

To analyze the mentioned constraint on our MLIR approach, we first calculate the probabilistic counts of words in each document given the probabilistic dictionary. If we do not distinguish term x in two languages, then for calculating the probabilistic count of term x in document D_1, we count x in both languages, i.e., $c_p(x, D_1) > c(x, D_1)$. On the other hand, for counting z in D_2, we only have z in l_1, i.e. $c_p(z, D_1) = c(z, D_1)$. Therefore, $c_p(x, D_1) > c_p(z, D_2)$. This causes document D_1 to artificially have more occurrences of query terms and hence get a higher rank than document D_2, which is not desirable. But, this problem does not arise when terms are labeled with language tags.

5.2 Incomplete dictionary coverage

Another point to discuss is the proposed estimation approaches in the case that available translation resources do not provide full coverage of words (Eqs. (15), (16)). Estimating document language models using Eq. (14) does not preserve the length ratio of documents. As a consequence, terms of a document containing term(s) with no entry in the translation dictionary are artificially enhanced compared to those of a document that all its terms are

available in the dictionary. This leads to an improper ranking of documents which we show using the second MLIR constraint.

MLIR Constraint 2 Let D_1 and D_2 be two documents in the same language in a multilingual collection and the two documents differ only in one term. Therefore, $|D_1| = |D_2|$. Let terms x and y belonging to D_1 and D_2, respectively, represent the only difference between these two documents. Also, assume that the dictionary contains translations for x, but not for y. Let q be a query that contains neither x nor y. Under these assumptions, D_1 and D_2 should get the same score w.r.t. q.

Analyzing the second constraint on our approach The matched terms between a query and a document contribute to the document's score in our approach. Let z denote a matched term between the document D_1 (D_2) and query q. We are thus interested in $p(z|D_1)$ and $p(z|D_2)$ to figure out the relative ranking of the two documents in the result list. If we estimate term probabilities using Eq. (14), then the value of the denominator for D_2 is one less than that for D_1, which means that the length ratio of these documents is changed considering the initial equal lengths of these documents. Since the probabilistic counts of z in both documents are equal, we have $p(z|D_2) > p(z|D_1)$ and consequently $\text{Score}(q, D_2) > \text{Score}(q, D_1)$. This scoring is contrary to the reasonable scores of documents, $\text{Score}(q, D_2) = \text{Score}(q, D_1)$. But, considering $N|D|$ as the denominator in Eq. (15), we achieve the expected ranking.

5.3 Term discrimination value

The *term discrimination constraint* (TDC), introduced in (Fang et al. 2004), regulates the impact of discrimination values of query terms on a document's score. TDC states that between two equal-length documents with the same total occurrences of query terms, the document containing more occurrences of the more specific query term should get a higher score.

TDC axiom in a multilingual environment The main objective is to depict that discrimination values of terms in a multilingual collection should be determined considering term occurrences in all documents of the collection, independent of their languages. To clarify, consider the example collection in Fig. 1a. Let $q = \{ab\}$ be a two term query in language l_1. The goal is to investigate the reasonable relative ranking of documents D_1 and D_3 in language l_1. Note that their relative ranking depends only on the discrimination values of query terms. To consider the entire collection in determining term discrimination values, all documents are translated into one language (l_1), using the dictionary of Fig. 1a. Given the translated collection, depicted in Fig. 1c, term a occurs in three documents, while term b has been used in four documents. Therefore, term a is more specific than term b and according to TDC, D_1 should get a higher score compared to D_3 in the final result.

Our MLIR approach Our approach satisfies the mentioned reasonable ranking of documents, because the reference language model is estimated using probabilistic counts of words (Eq. (11)). As shown in Fig. 1b, terms a and b have also non-zero probabilistic counts in documents in language l_2. Therefore, in our reference language model, the probability of term a is less than that of term b. Smoothing using this reference language model leads to the desired ranking of documents.

Fusion-based methods Almost all fusion-based methods fail to satisfy TDC, because of the limitation that the relative ranking of documents in the individual result lists should be preserved in the final ranked list. These methods combine the results of two retrieval runs to produce results for collection C: monolingual retrieval on subcollection C_1, and cross-lingual retrieval on subcollection C_2. An ideal monolingual retrieval model should prefer

D_3 over D_1 in response to q, because in subcollection C_1, query term b is more specific than term a. Under the constraint of preserving relative ranks of documents in the individual results, the rank of D_3 will be lower than that of D_1 in the final result of these fusion-based methods in response to query q on collection C, which is not desirable considering the statistics on the entire collection.

The mentioned sample case is common in practice. Documents in one language might cover the query topic, while documents in another language do not. Hence, decisions on term and document features to derive a ranked list of documents should be based on the entire collection, which is not possible in fusion-based methods, but is a strong point of our approach.

5.4 Computational complexity

The next point to consider is the run time analysis of the proposed MLIR approach. We first discuss the efficient implementation of monolingual retrieval based on the KL-divergence framework, investigated in (Zhai and Lafferty 2001b). If we use a smoothing technique in which the probability of an unseen word w in a document D is equal to $\alpha_D p(w|C)$ (α_D is a document-dependent constant), then Eq. (1) can be calculated very efficiently. The reason is that the summation in Eq. (1) is computed only for matched terms between the query and the document. In this case, the computational complexity of scoring documents with respect to query Q is estimated as $O(K|Q|)$, where K is the average number of documents containing a query term.

Using multilingual language models, the summation in the KL-divergence scoring function of Eq. (1) can be computed only for words that have non-zero probabilities in the query language model, and non-zero probabilistic counts in a document D as follows:

$$\text{Score}(Q,D) = \sum_{\substack{w\,:\,p(w|\theta_Q)>0,\\ c_p(w,D)>0}} p(w|\theta_Q)\log\frac{p_{\text{s}}(w|\theta_D)}{p_{\text{u}}(w|\theta_D)} + \sum_{w:p(w|\theta_Q)>0} p(w|\theta_Q)\log p_{\text{u}}(w|\theta_D), \tag{19}$$

where $p_s(w|D)$ is the smoothed probability of word w seen in document D, and $p_u(w|D)$ is the probability assigned to unseen word w in the document. The only efficiency issue for computing this equation is the estimation of $c_p(w,D)$ in $p_s(w|D)$.

We employ probabilistic word-to-word translation models to estimate probabilistic counts of words. There are two strategies to filter probabilistic translation models, obtained from parallel corpora, for use in CLIR; selecting n best translations for each word, or selecting translations whose probabilities are higher than a threshold (Nie et al. 2012). After filtering and renormalizing the translation models, an inverted index is built on translation models such that for each word u a list of words in all languages that translate into u along with their translation probabilities is kept. Model parameters are estimated using this inverted index on translation models and the inverted index of the collection. This estimation can be done either at index time or at retrieval time. Following, we discuss the complexity of the two options in more details.

1. Estimating the parameters of document language models at index time: Probabilistic counts are precomputed for all documents at index time (Xu et al. 2001). In this case, each document will be added to the document list of all translations of its words. The size of new index, containing probabilistic counts of words, depends on the number of

selected translations per word. On average, the index size will be about N times larger than that on the original document collection where N is the number of languages in the collection. Additional offline processing for building this index is just the calculation of probabilistic counts for documents. Using the new probabilistic index, multilingual retrieval can be performed as efficiently as monolingual retrieval. This way, the scoring complexity of our MLIR approach for query Q is $O(K'|Q|)$, where K' is the average number of documents that have a query term with a non-zero probabilistic count.[2]

2. Estimating the parameters of document language models at retrieval time: In this case, there is no increase in the size of collection index, or offline processing time. The index is multilingual, however it is built on the original documents, not the expanded ones. Thus, this index is identical to concatenation of individual indexes of subcollections since we use language tags and there cannot be a common term between documents of different languages. Probabilities of words given a document are then estimated at retrieval time. However, this estimation does not significantly increase the runtime complexity, because retrieval score is computed only for documents that have a non-zero probabilistic count of a query term. For each query term in monolingual retrieval, the score of documents containing that term, obtained using the inverted index of the collection, will be increased. In this implementation of multilingual retrieval, in addition to documents containing each query term, the score of documents that contain a term that translates to a query term is also increased. These documents are obtained using the built inverted index on translation models. Therefore, the scoring complexity of our MLIR approach for query Q using this strategy is $O(K(|Q| + |T|))$, where K is the average number of documents that have a query term or a term that translates to a query term[3], and T is the set of terms that translate to query terms.

In our experiments, we follow the second strategy.

6 Experiments

Datasets We use two CLEF datasets for evaluation: (1) CLEF 2001–2002 multilingual test collection, and (2) CLEF 2003-Multilingual 4 test collection. Table 1 lists statistics of these collections. We index the TEXT and TITLE fields of documents in both collections for retrieval and use the three query sets, CLEF2001, CLEF2002, and CLEF2003. Each query set includes equivalent topics in multiple languages.

Translation models We build a word-to-word translation model for each language pair using the *Europarl Corpus* (European Parliament Proceedings Parallel Corpus) (Tiedemann 2012). Statistical translation models (IBM model 1) are obtained using the GIZA++ toolkit (Och and Ney 2003). Before word alignment, the below-mentioned preprocessing steps are done on both sides of each parallel corpus. Obtained translation probabilities are

[2] Parameter K' in the setting of our experiments (based on CLEF multilingual datasets and using the top 3 translations for each word from translation models trained on Europarl corpus) is 29,753 and K is 7,524 where the total number of documents in a multilingual index is on average 5.57 times of that in a monolingual index.

[3] A query term or a term that translates to a query term, occurs only in a part of the multilingual index equivalent to the index of its respective language. This is the reason that parameter K is used in the scoring complexity of the second implementation option.

Table 1 Dataset statistics

Collection	Languages	Number of documents	Query sets	Number of queries
CLEF2001–2002	English French German Italian Spanish	749,883	CLEF2001 CLEF2002	50 (topics 41–90) 50 (topics 91–140)
CLEF2003-Multilingual 4	English French German Spanish	1,048,137	CLEF2003	60 (topics 141–200)

then linearly normalized by selecting the top x translations for each word where x is 3 in our experiments, unless otherwise specified.

Preprocessing Diacritic characters are mapped to the corresponding unmarked characters. Stopwords are removed using stopword lists provided in *IR Multilingual Resources at UniNE*.[4] Next, we use Snowball stemmers[5] for all languages. Same preprocessing steps are done on the documents of all collections, either used for retrieval or for training a translation model (Kraaij et al. 2003).

Experimental setup All experiments are done using the Lemur toolkit.[6] We use only the TITLE of topics for evaluation and retrieve 1000 results per query. We evaluate the performance of multilingual information retrieval with respect to query sets in different languages. For each experiment, we report Mean Average Precision (MAP) and Precision at top 10 documents (Prec@10). Two-tailed paired t-test is used to test whether the differences between performance of approaches are statistically significant. MAP values reported in the tables are marked with ▲ (▼) and △ (▽) to indicate statistical significance at the levels of 0.01 and 0.05 respectively.

6.1 Effectiveness of multilingual unigram language models

MULM performance In the first set of experiments, we examine the effectiveness of the proposed approach for MLIR. Since translation models do not provide full coverage of the words in test collections, we use Eq. (18) to estimate document language models. Parameters of the 2-stage smoothing method are not tuned and default values are used as: $\mu = 2000$ and $\lambda = 0.5$ (Zhai and Lafferty 2001b). In these experiments, the top 3 translations for each word are used. The effectiveness of our model (**MULM**) on each test collection is reported in Table 2. The retrieval performance is measured using English queries of each query set.

In the results of MULM, we further explore the correlation between the number of retrieved documents in a language and the number of relevant documents in that language. The larger the number of relevant documents in subcollection C_i for a query, the more retrieved documents in language l_i is expected in the search results for that query. To check

[4] http://members.unine.ch/jacques.savoy/clef/.

[5] See http://snowball.tartarus.org/.

[6] See www.lemurproject.org.

Table 2 Performance of the proposed MLIR approach with different estimation methods

Query set	Query language	Method					
		MULM		MULM without language tags		MULM without dummy words	
		MAP	Prec@10	MAP	Prec@10	MAP	Prec@10
CLEF2001	English	0.3491	0.6500	**0.3701**▲	0.6820	0.3475	0.6500
CLEF2002	English	0.2764	0.5920	**0.3046**	0.6160	0.2734	0.5900
CLEF2003	English	0.3049	0.4967	**0.3308**	0.5017	0.3023	0.5033

The best performance per query set is marked in bold. Statistically significant differences between MULM and the run without language tags are marked

this correlation in the results of our approach, we estimate two ratios for each query in a query set:

1. The ratio between the number of relevant documents in language l_1 and the number of relevant documents in language l_2.
2. The ratio between the number of retrieved documents in l_1 and the number of retrieved documents in l_2, extracted from MULM results.

Any pair of languages can be selected for this study. Figure 2 shows the correlation between these two ratios, computed for French and English languages, in the MULM results for English queries of each query set. Queries that make the ratios or denominators of the ratios equal to zero are removed. As indicated in the diagrams, our approach provides high correlation between the two mentioned ratios which is desirable.

In addition, we report the performance of our approach on queries formed from TITLE and DESCRIPTION fields of English topics in Table 3 in order to evaluate the performance of our approach on longer queries and to also simplify the comparison with other approaches that use this type of queries.

In the next step, we investigate the effect of each design decision we have made, mentioned in Sect. 5, on the performance of our approach.

Language tag effect First, we evaluate the performance of our multilingual unigram language model when terms are not labeled with language tags. Hence, the common terms between languages are considered as the same term. The results are shown in the "MULM without language tags" column of Table 2. Contrary to what we expected according to *MLIR Constraint 1*, we observe substantial improvement in performance when we do not distinguish the common terms between languages (improvements are not statistically significant except for CLEF2001 English queries). The main reason for this observation is that not using language tags impacts retrieval in two ways:

1. Asymmetric changes of term frequencies in a document: frequencies of terms that are common between languages would increase while the frequencies of other terms would not. This can result in an undesirable ranking of documents as shown in the constraint 1.
2. Direct match between a document and a query in different languages when they both contain a common term: This direct match will increase the recall measure when translation models are noisy and common terms with similar meanings do not translate to each other. In particular, suppose a query Q in language l_i includes query term q which is common to two or more languages, denoted by the set L_c. If q does not

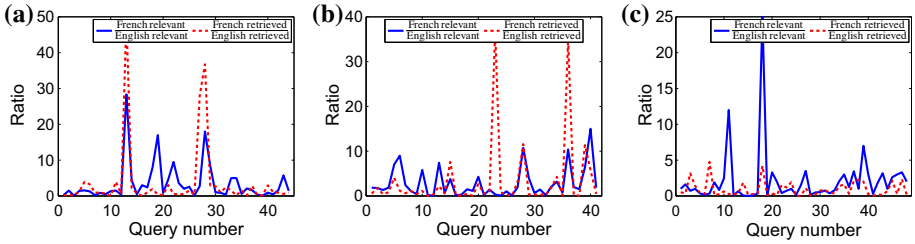

Fig. 2 Correlation between relevant and retrieved ratios of French to English. **a** CLEF2001 English query set. **b** CLEF2002 English query set. **c** CLEF2003 English query set

Table 3 Performance of the proposed MLIR approach on queries formed from TITLE + DESCRIPTION fields of topics

Query set	Query language	MULM	
		MAP	Prec@10
CLEF2001	English	0.3869	0.7000
CLEF2002	English	0.3469	0.6660
CLEF2003	English	0.3760	0.5633

translate to itself in the respective translation models, then using language tags avoids matching of documents in languages $L_c \setminus \{l_i\}$ containing term q with query Q. But, not using language tags allows these documents to directly match the query term q. Therefore, not using language tags in this case will increase the recall and consequently the MAP measure.

The results of our experiment show that given translation models trained on the Europarl corpus, the second item has more impact on the performance of retrieval. To explore this impact in more detail, let us describe the reason that query 44 ("Indurain Wins Tour") has the highest increase in *average precision* when comparing the run with non-labeled terms to the run with labeled terms. Stems of all terms of this query, indurain, win, and tour, are available in the indexes of all languages. To show the impact of direct matching between query terms and documents, we perform an experiment in which query 44 in English is searched over indexes of all other languages with and without employing translation models. In the former runs, bilingual runs using translation models, translations of document terms are matched against query terms, while in the latter runs, bilingual runs without using a translation model, direct match between query and document terms helps to rank documents. Table 4 shows the average precision of query 44 in both strategies for

Table 4 Comparison of AP performance of query 44 (CLEF2001 English query set) between the runs using terms with and without language tags

Documents' language	Number of relevant documents	Bilingual retrieval	
		Direct matching	Using a translation model
French	40	**0.2241**	0.0536
German	0	–	–
Italian	2	0.1385	**0.5105**
Spanish	6	**0.3135**	0.1829

The best average precision of query 44 per monolingual index is marked in bold

bilingual retrieval. As the results show, for French and Spanish indexes where the numbers of relevant documents are greater than 5, direct matching significantly outperforms using translation models. Using language tags in multilingual information retrieval is equivalent to relying only on translation models for matching documents against queries, while not using language tags also allows direct matching between documents and queries. This is the reason that query 44 has higher average precision in the multilingual run with non-labeled terms compared to the multilingual run with labeled terms. In addition, both runs, using labeled and non-labeled terms, have the same performance for queries that do not contain a common term such as query 58 in English, "Euthanasia". Our exploration shows that on average, each English query term of CLEF2001 query set is available in 3.97 indexes among five monolingual indexes built on CLEF2001–2002 collection, which in turn shows high overlap between the languages of the multilingual collection. We believe the possibility of direct matching when translation models are not perfect, is the main reason for the performance improvement of the multilingual run without language tags. Therefore, an interesting future research direction is to predict that given available translation models which strategy, using or not using language tags, achieves higher performance.

Dictionary coverage effect The next experiment is done to study the behavior of our approach when dictionaries do not have full coverage of the words used in the collection. We demonstrated in Sect. 5.2 that ignoring dummy words can lead to an incorrect ranking of documents which subsequently results in a lower MAP value. To examine this, we measure the performance of multilingual retrieval when dummy words are not included in the estimation of document language models. For this purpose, we use the maximum likelihood estimate in Eq. (14) and a similar estimate of the reference language model given by:

$$\overline{p}(w|C) = \frac{\sum_{D \in C} c_p(w, D)}{\sum_{D \in C} \sum_{u \in V} c_p(u, D)} , \tag{20}$$

which is used for smoothing as follows:

$$p(w|\overline{\theta}_D) = (1 - \lambda) \frac{c_p(w, D) + \mu \overline{p}(w|C)}{\sum_{w \in V} c_p(w, D) + \mu} + \lambda p(w|C) . \tag{21}$$

Table 2, column "MULM without dummy words", summarizes the results of using these language models for documents which show a slight reduction in MAP values compared to the "MULM" runs. To explore the reason for slight reduction, we first provide the completeness of trained translation models. Given the translation models trained for each pair of languages, Fig. 3a illustrates the number of unique words in relation to the number of translation models they have entries in. As the figure clearly shows, only a minority of unique words can be fully translated into all other languages and about 81 % of unique words cannot be translated into any of the other languages. However, the frequencies of almost all untranslatable words in the collections are very low, mostly just one occurrence. Therefore, the diagram of the total frequencies of words according to their coverage in the translation models in Fig. 3b is much different from the previous one. According to this diagram, 92 % of total words in the collections are fully translated into all other languages. From this viewpoint, the translation models provide adequate coverage of words of the collections. This implies that the number of added dummy words is low. To validate this, we consider all documents that have an English query term with a non-zero probabilistic count, and compare their lengths with and without dummy words. As depicted in Fig. 4,

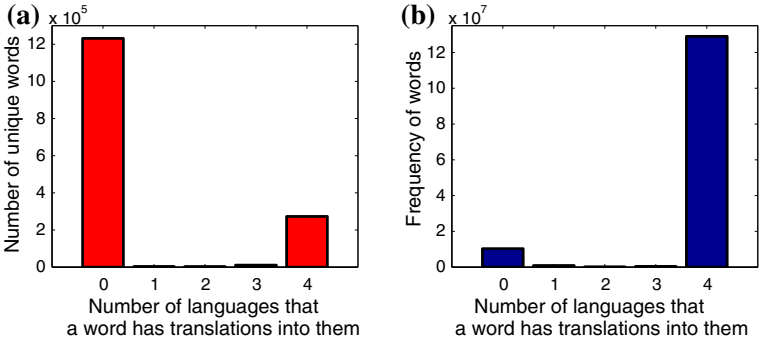

Fig. 3 Statistics on the coverage of the words of CLEF2001–2002 collection by the trained translation models. *Note* the figures are not to the same scale. We only show the coverage statistics for CLEF2001–2002 collection since they are representative of the statistics for CLEF2003 collection. **a** Unique words. **b** Total words

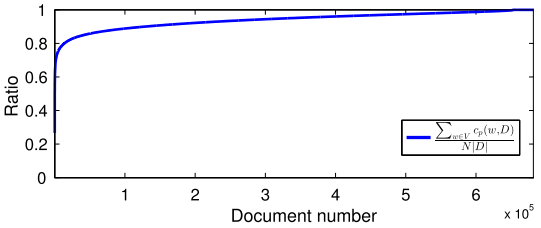

Fig. 4 Ratio of document lengths with and without considering dummy words

the ratio of two possible document lengths is very close to one for a large number of documents. Therefore, a slight alteration by replacing $N|D|$ in Eq. (18) with $\sum_{w \in V} c_p(w, D)$ in Eq. (21) does not give significantly different values. Because of this, we see a slight reduction in MAP values when dummy words are ignored.

To show the importance of the second MLIR constraint, we modify the translation models. For this purpose, we reduce the coverage of dictionaries by removing the top 15 % of most frequent words. The performance of our approach given the decreased translation models with and without considering dummy words are reported in Table 5. Statistically significant improvements of performance of runs with dummy words over those without dummy words across all three query sets show the importance of satisfying the second MLIR constraint.

Global estimation Finally, we examine the effect of global estimations of retrieval heuristics on the retrieval performance. To explain the effect of this global estimation, consider two documents in the same language. Their relative ranking may differ in the multilingual results using our approach compared to the results obtained from monolingual retrieval targeted on their underlying subcollection. Although the global estimation is necessary for MLIR, it can be arguable in the case that a user is interested in finding documents only in the query language from a multilingual collection. Hence, we study the effect of global estimation on the performance of retrieving documents in the query language using our approach. For English queries of each query set, we restrict the results of our MLIR approach to the documents in English, and compare the MAP value of these results against the results obtained from the original monolingual language modeling

Table 5 Performance of the proposed MLIR approach with and without considering dummy words when the translation models are slightly modified

Query set	Query language	Method			
		MULM		MULM without dummy	
		MAP	Prec@10	MAP	Prec@10
CLEF2001	English	**0.2761**▲	0.5820	0.2696	0.5660
CLEF2002	English	**0.2226**△	0.5300	0.2166	0.5220
CLEF2003	English	**0.2228**▲	0.4417	0.2160	0.4233

Statistically significant differences between two runs are marked. The best performance per query set is marked in bold

approach only on the English subcollection. We repeat this experiment for other languages of each query set. The results are shown in Table 6 which show improvements (mainly non-statistically significant) over the performance of monolingual retrieval in most cases. The improvements confirm our hypothesis that the global estimates of retrieval heuristics help to obtain more accurate term features.

The above experiment shows that using the expanded reference language model for smoothing allows to better distinguish specific query terms in most cases. The reason for better discrimination of query terms is that a subcollection may not cover the topic of the query or may have much fewer documents relevant to the query compared to the other subcollections. In this case, estimating word discrimination values based on only that subcollection may not be reliable. Adopting the expanded reference language model can

Table 6 Comparing the performance of retrieving documents in the query language using the proposed MLIR approach with the performance of monolingual retrieval

Query set	Query language	Performance of monolingual results obtained from MULM		Monolingual performance	
		MAP	Prec@10	MAP	Prec@10
CLEF2001	English	**0.5004**	0.4106	0.4626	0.4000
	French	**0.4458**	0.3918	0.4327	0.4265
	German	0.3546	0.4750	0.3546	0.4583
	Italian	**0.3736**	0.4128	0.3644	0.4170
	Spanish	**0.4924**▲	0.5878	0.4826	0.5837
CLEF2002	English	**0.4106**	0.3690	0.4096	0.3833
	French	**0.3690**	0.3580	0.3655	0.3640
	German	**0.3328**	0.4500	0.3287	0.4417
	Italian	**0.3177**▲	0.3694	0.3047	0.3571
	Spanish	**0.4226**	0.5440	0.4152	0.5240
CLEF2003	English	**0.4401**▲	0.3167	0.4190	0.3111
	French	**0.4070**▲	0.3212	0.3803	0.3019
	German	0.2921	0.3732	**0.2962**	0.3732
	Spanish	**0.3947**	0.4684	0.3909	0.4561

Statistically significant differences between two runs are marked. The best MAP performance per query set is marked in bold

lead to a different relative order of two documents in results produced by our multilingual retrieval model than those generated by a monolingual retrieval model. Table 7 shows a sample of this change in the relative order of two documents observed for query 45 of CLEF2001 query set ("Israel/Jordan Peace Treaty"). As indicated in the table, our approach assigns a higher rank to the relevant document compared to the monolingual retrieval model. This is because the most specific query term according to the two estimates of the reference language model is different. The reference language model estimated on the entire multilingual collection better discriminates the query terms than the one estimated only on the English subcollection. The reason is that English subcollection has the fewest documents relevant to query 45 among other subcollections. The numbers of relevant documents are 34, 42, 47, 105, and 191 in English, Italian, French, German, and Spanish subcollections, respectively.

Sensitivity to the retrieval parameters To gain a better understanding of the behavior of our MLIR approach, we study the sensitivity of performance of our approach, focusing on three parameters. First, we vary the parameter of Dirichlet Prior smoothing method; μ in Eq. (18) and keep $\lambda = 0.5$. Fig. 5a shows the effect of this parameter on the MLIR performance. All test collections show a decreasing trend for large μ values. This behavior is compatible with the impact of smoothing parameters on the performance of the monolingual language modeling approach. Second, we sweep the λ parameter in Eq. (18) and set μ to 2000. The obtained MAP values in Fig. 5b show that there is a general trend to prefer the lower values of λ, because dictionaries have adequate coverage of words in the target test collections. The sharp decrease in MAP value for $\lambda = 1$ is expected, because in this case, we completely ignore the content of the documents and use only the reference language model. Third, to study the impact of the number of translations selected for each word, we vary this number from 1 to 10. Fig. 5c shows how the performance varies according to the number of selected translations. In general, using only the top translation for each word results in a substantially lower MAP value since we lose some synonymous

Table 7 Statistics of terms of query 45 (CLEF2001 English query set) in two English documents, English subcollection, and the entire multilingual collection

		Document ID	
		LA120694-0031	LA071694-0006
Relevance status		0	1
Document rank in MULM		23	**13**
Document rank in monolingual		**9**	10
Document length		983	580
$c(q, D); (p_{\text{ml}}(q\|\theta_D))$	Israel	36 (**0.037**)	19 (0.033)
	Jordan	3 (0.003)	10 (**0.017**)
	Peace	23 (0.023)	14 (**0.024**)
	Treaty	6 (**0.006**)	3 (0.005)
$p(w\|C_{\text{En}})$	Israel	0.000159055	
	Jordan	0.000113171	
	Peace	0.000294397	
	Treaty	**4.26199e−05**	
$p''(w\|C)$	Israel	7.18164e−05	
	Jordan	**2.22658e−05**	
	Peace	0.000142423	
	Treaty	0.000112275	

The higher probability of each query term in document language models is marked in bold. In addition, the most specific query term in each estimate of the reference language model is indicated in bold

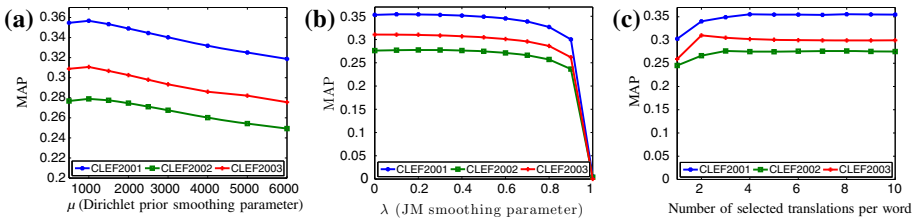

Fig. 5 Sensitivity to the model parameters

Table 8 Performance of the proposed MLIR approach when the parameters (the smoothing parameters and the number of selected translations for each word) are tuned

Query set	Query language	MULM	
		MAP	Prec@10
CLEF2001	English	0.3582	0.6560
CLEF2002	English	0.2797	0.5940
CLEF2003	English	0.3137	0.5283

or related translations. Therefore, all curves have an increase when the number of selected translations is low. Then, our approach has stable performance when the number of selected translations increases. Finally, we optimize the three parameters with respect to MAP. Table 8 shows the results. We find that using default values for parameters does not yield very different results from the optimized parameters.

6.2 Comparison with previous approaches

In this set of experiments, we compare our approach with the existing methods for MLIR. First, we provide the results of traditional merging algorithms for MLIR, i.e. *Raw scoring*, *Round-Robin*, *Max normalized scoring*, and *Min-Max normalized scoring*. Intermediate result lists for merging are produced using the language modeling framework to rule out the effect of retrieval models on the performance. For cross-language results using the language modeling framework, we use both strategies mentioned in Sect. 3:

1. Integrating translation knowledge into the query language model (Eq. (56)): We consider the fusion of lists produced using this strategy as a baseline for comparison with our approach. This is because most fusion-based methods use the query translation strategy for producing intermediate lists due to its efficiency.
2. Integrating translation knowledge into the document language models (Eq. (78)): This approach is exploited in (Xu et al. 2001; Kraaij et al. 2003) to perform cross-language (bilingual) information retrieval by representing documents with new language models in the query language. Similarly, our approach for MLIR is based on building new multilingual language models for documents. Therefore, to provide a fair comparison and preclude the impact of translation direction on the performance, we also regard the merging of lists produced using this strategy as a baseline.

Tables 9 and 11 show the performance of merging the lists generated using the first and second strategies, respectively, in terms of MAP. We also report precision at top 10

Table 9 MAP performance of MULM and different merging strategies

Query set	Query language	Method					
		MULM (% optimal)	Raw scoring	Max Norm.	Min–max Norm.	R.R.	Optimal
CLEF2001	English	**0.3491**▲ (82 %)	0.2658	0.2518	0.2819	0.2723	0.4224
	French	**0.3608**▲ (88 %)	0.2470	0.2431	0.2758	0.2613	0.4055
	German	**0.2653** (70 %)	0.2291	0.2033	0.2400	0.2324	0.3746
	Italian	**0.3525**▲ (87 %)	0.2504	0.2320	0.2727	0.2574	0.4006
	Spanish	**0.3603**▲ (88 %)	0.2424	0.2238	0.2664	0.2597	0.4071
CLEF2002	English	**0.2764**▲ (85 %)	0.2022	0.1613	0.1990	0.1966	0.3220
	French	**0.2715**▲ (88 %)	0.1699	0.1474	0.1964	0.1849	0.3064
	German	**0.2257** (68 %)	0.1765	0.1640	0.2028	0.1973	0.3283
	Italian	**0.2503**▲ (81 %)	0.1568	0.1573	0.1964	0.1853	0.3062
	Spanish	**0.2963**▲ (83 %)	0.1898	0.1581	0.2164	0.1973	0.3555
CLEF2003	English	**0.3049**▲ (93 %)	0.2161	0.1860	0.2150	0.2034	0.3262
	French	**0.2823**▲ (85 %)	0.2257	0.1974	0.2224	0.2062	0.3313
	German	**0.2260** (70 %)	0.1985	0.1696	0.2032	0.1878	0.3227
	Spanish	**0.2817**▲ (81 %)	0.2036	0.1894	0.2175	0.2048	0.3470

The reported MAP values for merging strategies are for fusing the intermediate lists obtained using **query translation** approach. The best performance per query set, except the theoretical optimal performance, is marked in bold. Statistically significant differences between MULM and the best merging strategy run are marked

documents for strategies 1 and 2 in Tables 10 and 12, respectively. We include P@10 performance only for English queries of each query set, since all approaches behave similarly for all metrics. The first remark about the results is the performance difference between queries in different languages belonging to a query set which is related to different translation qualities between corresponding language pairs. The second observation is that merging the lists generated using document translation strategy outperforms that using query translation strategy in all cases. Third, our approach outperforms traditional merging approaches on the results of both strategies across all three collections and all query languages, and the improvements are statistically significant in most cases.

To gain more insight into the results, we display the difference in average precision (AP) between our multilingual approach and raw score merging method for each English query in Figs. 6 and 7. The figures clearly show that our approach behaves remarkably different from the raw score merging and outperforms that for many queries. For instance, our approach improves the average precision of query 45 ("Israel/Jordan Peace Treaty") by 50 % compared to the raw score merging of results obtained by the LM-based document translation approach. The expanded estimate of the reference language model helps to effectively retrieve documents of interest w.r.t. this query. To further explain the reason for improvement of the performance, we measure the popularity of query terms in both the entire multilingual collection and the individual subcollections. We find that probabilities of the query terms given one or another of the subcollections vary significantly. Query term "Treaty" has the lowest probability given the English subcollection among all query terms given any subcollection. This leads to higher ranks for English documents that have more

Table 10 P@10 performance of MULM and different merging strategies

Query set	Query language	Method					
		MULM (% optimal)	Raw scoring	Max Norm.	Min–max Norm.	R.R.	Optimal
CLEF2001	English	**0.6500** (81 %)	0.5940	0.5320	0.5340	0.5460	0.7980
CLEF2002	English	**0.5920** (78 %)	0.5180	0.4480	0.4580	0.4440	0.7600
CLEF2003	English	**0.4967** (89 %)	0.4167	0.3767	0.3833	0.3667	0.5583

The reported performance for merging strategies are for fusing the intermediate lists obtained using **query translation** approach. The best performance per query set, except the theoretical optimal performance, is marked in bold

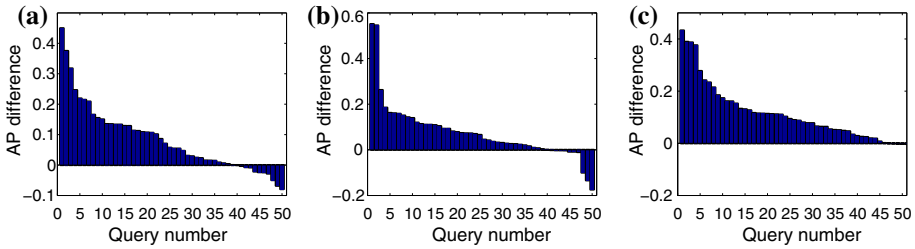

Fig. 6 Per query AP difference between MULM and raw score merging. The reported MAP values for raw score merging are for fusing the intermediate lists obtained using **query translation** approach. **a** CLEF2001 query set. **b** CLEF2002 query set. **c** CLEF2003 query set

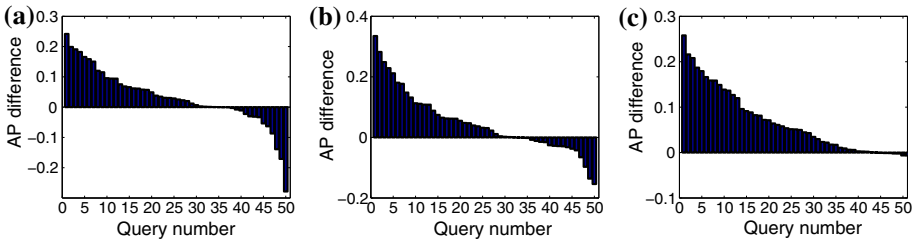

Fig. 7 Per query AP difference between MULM and raw score merging. The reported MAP values for raw score merging are for fusing the intermediate lists obtained using **document translation** approach. **a** CLEF2001 query set. **b** CLEF2002 query set. **c** CLEF2003 query set

occurrences of "Treaty". However, query terms "Jordan" and "Israel" are more specific than "Treaty" in the expanded reference language model as shown in Table 7. The more accurate estimation of term discrimination values using the entire collection helps to better rank the documents w.r.t this query.

On the other hand, performance of some queries is hurt using our approach. The highest performance loss compared to the raw score merging of the results of LM-based document translation is observed for query 59 of CLEF2001 query set ("Computer Viruses"). Both the raw score merging and our approach retrieve 19 of 20 documents relevant to this query. However, better ranking of these relevant documents by the raw score merging is due to the higher discrimination value of query term "Computer" in Spanish subcollection compared to

other subcollections. The probability of query term "Computer" given Spanish subcollection is between 4 to 12 times smaller than that given other subcollections. The smaller value of this probability leads to higher scores for Spanish documents compared to the documents in other languages. This observation, together with the fact that 7 of 20 documents relevant to this query are in Spanish, explains the higher average precision of results retrieved by the raw score merging method. However, considering two documents listed in Table 13 shows that our approach behaves reasonably w.r.t. this query. The non-relevant document "LASTAMPA94-022927" which is ranked higher using our approach, in contrast to the raw score merging, has more occurrences of the two query terms. Therefore, independent of discrimination values of query terms, the provided ranking is reasonable. The provided ranking is compatible with the *Term Frequency Constraint 1* (TFC1) defined in (Fang et al. 2011). The TFC1 axiom states that increasing the occurrences of query terms causes increase in the retrieval score. The reason for lower performance might then be inexact estimates of term frequencies in other languages due to noise in the translation models.

Our analysis also shows that when MULM or the raw score merging outperforms the other for a query in a language, the same trend does not necessarily hold for the corresponding query in other languages. The reason is that the raw score merging highly depends on the difference between the popularity of a query term in different subcollections, and this difference can vary according to the language of the query term. Our approach, on the other hand, exhibits relatively stable performance over different languages as demonstrated in robustness evaluation of MLIR approaches. To summarize, the raw score merging can succeed when query terms are more specific with respect to a subcollection that has more relevant documents than other subcollections. In contrast, our approach employs the expanded estimate of the reference language model which can help in better distinguishing query terms.

Moreover, we report the optimal performance achievable by fusion-based approaches given the intermediate results. In (Chen 2002), a method for deriving theoretically optimal performance that could possibly be achieved from merging multiple ranked lists is presented. This method builds a merged list with the maximum MAP value such that the relative order of documents in each intermediate list is preserved in the final list. However, since the approach estimates an upper bound of MLIR performance using relevance status of documents, it applies only to test collections. The optimal performance that could be achieved from merging the intermediate results produced using the LM-based query and document translation approaches are reported in the last columns of Tables 9 and 11, respectively. Our direct approach achieves 68–93 % of this optimal performance.

To provide a comprehensive evaluation, we also provide the results of our approach in comparison with different merging strategies when translation models are filtered such that only translations with probabilities higher than 0.1 are kept. These results are reported in Table 14 which show that our approach outperforms traditional merging strategies when translation models are filtered based on translation probabilities.

We also compare our approach with 2-step RSV (Martinez-Santiago et al. 2006) and the direct method of (Nie and Jin 2003) which is reported as *direct-NJ*. Result lists of these methods are generated using the OKAPI retrieval model. In both methods, queries are expanded by translations of query terms in all languages, while document words are translated using our approach. This makes the comparison difficult because effectiveness of an approach also depends on the translation direction (Nie 2010). We make comparisons between our approach, 2-step RSV, and direct-NJ methods in Tables 15 and 16 in terms of MAP and P@10, respectively. According to the results reported in Table 15, MULM outperforms both methods. The trend of the reported results for 2-step RSV method differs

Table 11 MAP performance of MULM and different merging strategies

Query set	Query language	Method					
		MULM (% optimal)	Raw scoring	Max Norm.	Min–max Norm.	R.R.	Optimal
CLEF2001	English	**0.3491**$^{\triangle}$ (79 %)	0.3164	0.3035	0.3030	0.2898	0.4400
	French	**0.3608**$^{\blacktriangle}$ (82 %)	0.3131	0.3064	0.3050	0.2850	0.4364
	German	**0.2653**$^{\blacktriangle}$ (81 %)	0.2391	0.2090	0.2114	0.2048	0.3268
	Italian	**0.3525**$^{\blacktriangle}$ (81 %)	0.3182	0.3016	0.3010	0.2859	0.4326
	Spanish	**0.3603**$^{\blacktriangle}$ (81%)	0.3263	0.3103	0.3094	0.2936	0.4434
CLEF2002	English	**0.2764**$^{\blacktriangle}$ (78 %)	0.2283	0.2193	0.2207	0.2178	0.3532
	French	**0.2715**$^{\blacktriangle}$ (79 %)	0.2353	0.2314	0.2349	0.2185	0.3418
	German	**0.2257**$^{\blacktriangle}$ (78 %)	0.2017	0.1845	0.1842	0.1723	0.2875
	Italian	**0.2503**$^{\triangle}$ (78 %)	0.2268	0.2102	0.2133	0.2020	0.3176
	Spanish	**0.2963**$^{\triangle}$ (78 %)	0.2701	0.2286	0.2356	0.2229	0.3752
CLEF2003	English	**0.3049**$^{\blacktriangle}$ (83 %)	0.2527	0.2545	0.2547	0.2318	0.3645
	French	**0.2823**$^{\blacktriangle}$ (82 %)	0.2352	0.2334	0.2335	0.2118	0.3433
	German	**0.2260**$^{\blacktriangle}$ (82 %)	0.1969	0.1680	0.1687	0.1495	0.2743
	Spanish	**0.2817**$^{\blacktriangle}$ (79 %)	0.2505	0.2305	0.2269	0.2135	0.3541

The reported MAP values for merging strategies are for fusing the intermediate lists obtained using **document translation** approach. The best performance per query set, except the theoretical optimal performance, is marked in bold. Statistically significant differences between MULM and the best merging strategy run are marked

Table 12 P@10 performance of MULM and different merging strategies

Query set	Query language	Method					
		MULM (% optimal)	Raw scoring	Max Norm.	Min-Max Norm.	R.R.	Optimal
CLEF2001	English	**0.6500** (79 %)	0.6060	0.5680	0.5620	0.5520	0.8220
CLEF2002	English	**0.5920** (76 %)	0.5160	0.4700	0.4740	0.4800	0.7820
CLEF2003	English	**0.4967** (79 %)	0.4317	0.4483	0.4483	0.3900	0.6250

The reported performance for merging strategies are for fusing the intermediate lists obtained using **document translation** approach. The best performance per query set, except the theoretical optimal performance, is marked in bold

from that of the results in (Martinez-Santiago et al. 2006); in contrast to the results of our experiments in Table 15, 2-step RSV method outperforms traditional merging algorithms in (Martinez-Santiago et al. 2006). This difference arises almost entirely from three sources: (1) different translation resources: machine-readable bilingual dictionaries are used in (Martinez-Santiago et al. 2006), while we extract translations from parallel corpora in our experiments. (2) different queries: queries in (Martinez-Santiago et al. 2006) are consisted of TITLE and DESCRIPTION fields, while queries in our experiments are formed only from the TITLE field of topics. (3) different numbers of selected translations for each word. We hypothesize that the difference in results mainly derives from the third factor as the authors in (Martinez-Santiago et al. 2006) also mentioned that they achieved better performance using only the top translation for each word.

Table 13 Probabilistic counts of terms of query 59 (CLEF2001 English query set) in two documents

Document ID (D)	Document language	$c_p(q_i\|D)$		Document size
		Computer	Viruses	
LASTAMPA94-022927	Italian	12.6541	2.28905	316
EFE19940211-06409	Spanish	7.06986	1.76224	310

Table 14 MAP performance of MULM and different merging strategies when translation models are filtered by selecting translations whose probabilities are greater than 0.1

Query set	Query language	Method					
		MULM (% optimal)	Raw scoring	Max Norm.	Min–max Norm.	R.R.	Optimal
CLEF2001	English	**0.3381**$^{\triangle}$ (80 %)	0.3075	0.2914	0.2910	0.2760	0.4250
CLEF2002	English	**0.2667**$^{\blacktriangle}$ (79 %)	0.2209	0.2099	0.2112	0.2071	0.3376
CLEF2003	English	**0.2971**$^{\blacktriangle}$ (82 %)	0.2507	0.2521	0.2536	0.2299	0.3632

The reported MAP values for merging strategies are for fusing the intermediate lists obtained using document translation approach. The best performance per query set, except the theoretical optimal performance, is marked in bold

Table 15 MAP performance comparison between MULM, 2-step, and Direct-NJ

Query set	Query language	Method		
		MULM	2-Step	Direct-NJ
CLEF2001	English	**0.3491**	0.2155$^{\blacktriangledown}$	0.1620$^{\blacktriangledown}$
	French	**0.3608**	0.2291$^{\blacktriangledown}$	0.1643$^{\blacktriangledown}$
	German	**0.2653**	0.1758$^{\blacktriangledown}$	0.1649$^{\blacktriangledown}$
	Italian	**0.3525**	0.1850$^{\blacktriangledown}$	0.1641$^{\blacktriangledown}$
	Spanish	**0.3603**	0.1858$^{\blacktriangledown}$	0.1677$^{\blacktriangledown}$
CLEF2002	English	**0.2764**	0.1436$^{\blacktriangledown}$	0.1173$^{\blacktriangledown}$
	French	**0.2715**	0.1253$^{\blacktriangledown}$	0.1250$^{\blacktriangledown}$
	German	**0.2257**	0.1208$^{\blacktriangledown}$	0.1385$^{\blacktriangledown}$
	Italian	**0.2503**	0.1033$^{\blacktriangledown}$	0.1225$^{\blacktriangledown}$
	Spanish	**0.2963**	0.1636$^{\blacktriangledown}$	0.1489$^{\blacktriangledown}$
CLEF2003	English	**0.3049**	0.1567$^{\blacktriangledown}$	0.1039$^{\blacktriangledown}$
	French	**0.2823**	0.1703$^{\blacktriangledown}$	0.1126$^{\blacktriangledown}$
	German	**0.2260**	0.1533$^{\blacktriangledown}$	0.1144$^{\blacktriangledown}$
	Spanish	**0.2817**	0.1864$^{\blacktriangledown}$	0.1260$^{\blacktriangledown}$

The best performance per query set is marked in bold. Statistically significant differences between MULM and 2-step, and between MULM and Direct-NJ are marked

To mitigate this mismatch, we also report the performance of MULM, 2-step RSV, and direct-NJ methods when only the top translation for each word is used. The results of these runs are listed in Tables 17 and 18 for queries that are generated from only the TITLE field and both the TITLE and DESCRIPTION fields of the topics, respectively. For comparison purpose, Table 19 also summarizes the performance of traditional merging algorithms on queries generated from the TITLE field of topics when only the top translation of each

Table 16 Comparison of P@10 performance between MULM, 2-step, and Direct-NJ

Query set	Query language	Method		
		MULM	2-Step	Direct-NJ
CLEF2001	English	**0.6500**	0.5020	0.2980
CLEF2002	English	**0.5920**	0.4260	0.2500
CLEF2003	English	**0.4967**	0.3367	0.1700

The best performance per query set is marked in bold

Table 17 MAP performance comparison between MULM, 2-step and Direct-NJ approaches when only the top translation for each word is used

Query set	Query language	Method			
		MULM	2-step	Direct-NJ	Optimal
CLEF2001	English	**0.3025 (78 %)**	0.2957 (76 %)	0.1451▼	0.3846
CLEF2002	English	**0.2452 (83 %)**	0.2243 (75 %)	0.1047▼	0.2952
CLEF2003	English	**0.2591 (83 %)**	0.2423 (78 %)	0.1004▼	0.3096

The best performance per query set, except the theoretical optimal performance, is marked in bold

Table 18 Comparison of MAP performance between MULM, 2-step and Direct-NJ approaches when queries are generated from TITLE and DESCRIPTION fields of the topics and only the top translation for each word is used

Query set	Query language	Method		
		MULM	2-step	Direct-NJ
CLEF2001	English	**0.3419**	0.2988*	0.1594▼
CLEF2002	English	**0.3077**	0.2570▽	0.1095▼
CLEF2003	English	**0.3232**	0.2818▽	0.1173▼

The best performance per query set is marked in bold

* Indicates significance at the level of 0.1

word is used. The results in Tables 17, 18, and 19 follow the same trend as the results in (Martinez-Santiago et al. 2006); 2-step RSV method outperforms traditional merging algorithms. In this setting, our approach also outperforms other methods in all cases. In addition, MULM achieves consistently higher percentages of the theoretical optimal performance compared to the 2-step RSV method.

Finally, in Fig. 8, we compare the precision-recall curves of different MLIR approaches which shows that MULM performs better than all other methods.

Robustness evaluation Robustness of a model for multilingual information retrieval is interpreted as having "stable performance over all topics instead of high average performance" (Mandl et al. 2008). Robustness of a model has been assessed using *Geometric Mean Average Precision* (GMAP) in robust tasks organized at CLEF 2006 and 2007 (Di Nunzio et al. 2007; Nunzio et al. 2008). The experiments conducted in these CLEF tasks demonstrate that MLIR models behave more differently w.r.t. performance measures compared to monolingual retrieval models since more topics become difficult in MLIR.

Table 19 MAP performance of MULM and different merging strategies when only the top translation for each word is used

Query set	Query language	Method					
		MULM	Raw scoring	Max Norm.	Min–max Norm.	R.R.	Optimal
CLEF2001	English	**0.3025**△	0.2561	0.2160	0.2578	0.2421	0.3846
CLEF2002	English	**0.2452**▲	0.1854	0.1420	0.1782	0.1775	0.2952
CLEF2003	English	**0.2591**△	0.2142	0.1758	0.2032	0.1878	0.3096

The reported performance for merging strategies are for fusing the intermediate lists obtained using query translation approach. The best performance per query set, except the theoretical optimal performance, is marked in bold

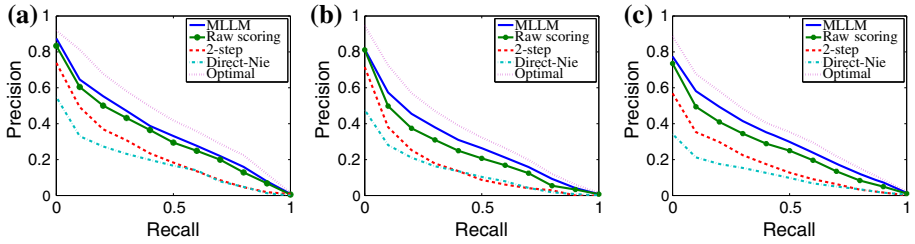

Fig. 8 Precision-recall curves comparing MULM, raw score merging, 2-step RSV, direct-NJ, and optimal merging methods. The curves for merging strategies are for fusing the intermediate lists obtained using document translation approach. **a** CLEF2001 query set. **b** CLEF2002 query set. **c** CLEF2003 query set

Therefore, from a practical point of view, evaluation of MLIR models based on measures that better reveal the robustness is important. In this regard, we also compare our approach with existing approaches for MLIR based on GMAP. The results are listed in Tables 20 and 21, which indicate that our approach has more robust performance across different query languages compared to different merging strategies, 2-step, and Direct-NJ methods.

6.3 Impact of feedback on retrieval performance

In these experiments, we study the effect of pseudo relevance feedback on the performance of multilingual retrieval. Topic model of feedback documents is built adopting mixture model-based feedback. Mixture model involves two parameters: (1) feedback mixture noise and (2) feedback coefficient. We do not tune theses parameters, and both parameters are set to default values of 0.5. In addition, feedback information is extracted from top 10 documents of previously retrieved documents and is integrated into the query language model, similar to Eq. (4), as:

$$p\left(w|\hat{\theta}'_Q\right) = \lambda p(w|\theta_Q) + (1-\lambda)p(w|\hat{\theta}_F)\,, \tag{22}$$

where $\hat{\theta}_F$ is the multilingual topic model, estimated based on the multilingual language models of the top retrieved documents. In addition, the new language model of the query,

Table 20 GMAP performance of MULM and different merging strategies

Query set	Query language	Method					
		MULM (% optimal)	Raw scoring	Max Norm.	Min–max Norm.	R.R.	Optimal
CLEF2001	English	**0.2056** (67 %)	0.1669	0.1605	0.1681	0.1759	0.3086
CLEF2002	English	**0.1582** (68 %)	0.1289	0.1198	0.1240	0.1198	0.2336
CLEF2003	English	**0.1752** (71 %)	0.1276	0.1389	0.1395	0.1323	0.2474

The reported performance for merging strategies are for fusing the intermediate lists obtained using document translation approach. The best performance per query set, except the theoretical optimal performance, is marked in bold

Table 21 Comparison of GMAP performance between MULM, 2-step, and Direct-NJ

Query set	Query language	Method		
		MULM	2-step	Direct-NJ
CLEF2001	English	**0.2056**	0.1062	0.0759
CLEF2002	English	**0.1582**	0.0727	0.0515
CLEF2003	English	**0.1752**	0.0487	0.0374

The best performance per query set is marked in bold

Table 22 Performance of accommodating pseudo relevance feedback in MULM

Query set	Query language	Method					
		MULM			MULM+FB		
		MAP	P@10	R@1000	MAP	P@10	R@1000
CLEF2001	English	0.3491	0.6500	0.7020	**0.3635**▲	**0.6660**	**0.7163**
CLEF2002	English	0.2764	0.5920	0.6244	**0.2978**▲	**0.6260**	**0.6524**
CLEF2003	English	0.3049	0.4967	0.7407	**0.3200**▲	**0.5233**	**0.7567**

Bold face indicates best score per metric. Statistical significance is tested against MULM runs

$\hat{\theta}'_Q$, is a multilingual unigram language model. To investigate the effect of feedback, we also report Recall at 1000 documents (R@1000), since one purpose of feedback techniques is to increase the recall measure. We can see the positive effect of feedback in Table 22, which shows improvements in all measures for all query sets.

We also consider the effect of feedback on fusion-based methods. To incorporate feedback in these methods, we update each individually retrieved list with feedback information before the merging phase. Individually retrieved lists, which need to be updated, are obtained using either monolingual or cross-lingual information retrieval using the language modeling framework. The language models of queries in monolingual runs are expanded with feedback information according to Eq. (4). Employing feedback information in cross-lingual retrieval depends on the selected strategy for the initial retrieval phase.

In case of using query translation approach (Eq. (56)), the new language model of queries can be directly updated with feedback information, similar to monolingual runs, by using:

$$p\left(w|\tilde{\theta}'_Q\right) = \lambda p(w|\tilde{\theta}_Q) + (1-\lambda)p(w|\theta_F)\,, \tag{23}$$

where language models of the query and feedback documents are monolingual and their parameters are words of the target language. In document translation approach, the feedback topic model can be estimated based on the new language models of documents, $\tilde{\theta}_D$, in the source language. To estimate such a topic model, a reference language model in the source language is required. Therefore, we build a new reference language model as follows:

$$\tilde{p}(w_s|C) = \sum_{w_t \in V_t} p(w_s|w_t)p(w_t|C)\,. \tag{24}$$

The following equation is then used to update the query language model based on the estimated feedback topic model in the source language, denoted by $\tilde{\theta}_F$, as:

$$p\left(w|\theta'_Q\right) = \lambda p(w|\theta_Q) + (1-\lambda)p(w|\tilde{\theta}_F), \tag{25}$$

where parameters of all language models are words of the source language.

MAP performance of different merging strategies on updated intermediate lists with feedback information are listed in Table 23. We also report P@10 and R@1000 performance measures for these methods in Tables 24 and 25, respectively. Results show that MULM outperforms the merging strategies in terms of MAP and P@10 across all three

Table 23 MAP performance of MULM and different merging strategies

Query set	Query language	Method					
		MULM	Raw scoring	Max Norm.	Min–max Norm.	R.R.	Optimal
CLEF2001	English	**0.3635**$^{\triangle}$	0.3316	0.3118	0.3117	0.2969	0.4515
	French	**0.3736**$^{\blacktriangle}$	0.3268	0.3122	0.3111	0.2949	0.4496
	German	**0.2882**$^{\blacktriangle}$	0.2395	0.2238	0.2254	0.2161	0.3462
	Italian	**0.3669**$^{\blacktriangle}$	0.3225	0.3070	0.3067	0.2913	0.4459
	Spanish	**0.3728**$^{\blacktriangle}$	0.3318	0.3185	0.3172	0.3034	0.4591
CLEF2002	English	**0.2978**$^{\blacktriangle}$	0.2494	0.2380	0.2366	0.2292	0.3754
	French	**0.2905**$^{\blacktriangle}$	0.2526	0.2466	0.2478	0.2298	0.3694
	German	**0.2561**$^{\blacktriangle}$	0.2050	0.1961	0.1958	0.1822	0.3084
	Italian	**0.2691**$^{\blacktriangle}$	0.2365	0.2295	0.2316	0.2105	0.3348
	Spanish	**0.3170**$^{\blacktriangle}$	0.2698	0.2437	0.2501	0.2262	0.3981
CLEF2003	English	**0.3200**$^{\blacktriangle}$	0.2756	0.2704	0.2728	0.2442	0.3866
	French	**0.2920**$^{\triangle}$	0.2567	0.2472	0.2477	0.2248	0.3713
	German	**0.2418**$^{\blacktriangle}$	0.1925	0.1733	0.1746	0.1563	0.2867
	Spanish	**0.2990**$^{\blacktriangle}$	0.2621	0.2434	0.2398	0.2258	0.3761

The reported MAP values for merging strategies are for fusing the intermediate lists obtained using document translation approach. The best performance per query set, except the theoretical optimal performance, is marked in bold. Statistically significant differences between MULM and the best merging strategy run are marked

Table 24 P@10 performance of MULM and different merging strategies

Query set	Query language	Method					
		MULM	Raw scoring	Max Norm.	Min–max Norm.	R.R.	Optimal
CLEF2001	English	**0.6660**	0.6160	0.5620	0.5620	0.5600	0.8200
CLEF2002	English	**0.6260**	0.5440	0.4860	0.4920	0.4920	0.7940
CLEF2003	English	**0.5233**	0.4383	0.4267	0.4367	0.3883	0.6367

The reported performance for merging strategies are for fusing the intermediate lists obtained using document translation approach. The best performance per query set, except the theoretical optimal performance, is marked in bold

Table 25 R@1000 performance of MULM and different merging strategies

Query set	Query language	Method					
		MULM	Raw scoring	Max Norm.	Min–max Norm.	R.R.	Optimal
CLEF2001	English	0.7163	0.7149	0.7218	**0.7226**	0.7090	0.7925
CLEF2002	English	**0.6524**	0.6292	0.6446	0.6452	0.6380	0.7016
CLEF2003	English	**0.7567**	0.7203	0.7338	0.7312	0.7295	0.7889

The reported performance for merging strategies are for fusing the intermediate lists obtained using document translation approach. The best performance per query set, except the theoretical optimal performance, is marked in bold

datasets and all query languages, and the improvements are statistically significant in all cases. Except for one case (English queries of CLEF2001), MULM also achieves the best R@1000. Therefore, our approach that accommodates multilingual feedback information performs better than other approaches that use only local feedback information. This certifies that feedback information from one subcollection can help to also improve the retrieval performance on other subcollections, which is possible through multilingual feedback information.

7 Conclusion and future work

In this paper, we have investigated the estimation of multilingual unigram language models for documents, as well as global estimations of retrieval statistics in MLIR. These estimations enable retrieving a ranked list of documents in multiple languages in one retrieval phase and incorporation of multilingual feedback information. We have further adapted the proposed estimation approaches to the common case of incomplete coverage of dictionaries. Experimental results demonstrate that MLIR performance using the proposed approach is higher than the performance of the existing approaches in almost all cases. Meanwhile the proposed approach maintains two following advantages. First, it is independent of any assumption about the distribution of relevant documents in the subcollections. Second, tuning the performance of the proposed MLIR approach is straightforward similar to monolingual IR using the KL-divergence model.

Obtained results stimulate further research activities. A promising line is to study the MLIR performance with respect to long verbose queries using our approach. Another direction is to extend the proposed document language model in a way to efficiently consider document contexts in estimating the probabilistic counts of words. In addition, based on available translation models, determining whether using or not using language tags gives the higher performance is an important research question to be investigated in the future. How to compensate the incomplete coverage of translation models is also a crucial research direction for resource-limited languages in a multilingual collection. Finally, evaluating MLIR approaches on collections with documents in resource-limited languages or on collections containing mixed-language documents is another interesting and valuable future direction.

Acknowledgments This research was in part supported by a Grant from Institute for Research in Fundamental Sciences (No. CS1393-4-43), the National Grand Fundamental Research 973 Program of China (No. 2014CB340405), the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CUHK 413213), and Microsoft Research Asia Regional Seed Fund in Big Data Research (Grant No. FY13-RES-SPONSOR-036).

References

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). New York, NY, USA, SIGIR '99: ACM. doi:10.1145/312624.312681.

Braschler, M. (2004). Combination approaches for multilingual text retrieval. *Information Retrieval*, *7*(1–2), 183–204. doi:10.1023/B:INRT.0000009445.19495.46.

Braschler, M., & Schäuble, P. (2000). Using corpus-based approaches in a system for multilingual information retrieval. *Information Retrieval*, *3*(3), 273–284. doi:10.1023/A:1026525127581.

Braschler, M., Ripplinger, B., & Schäuble, P. (2002). Experiments with the eurospider retrieval system for CLEF 2001. In *CLEF* (pp. 102–110). http://dl.acm.org/citation.cfm?id=648264.753545

Chen, A. (2002). Cross-language retrieval experiments at CLEF 2002. In *Proceedings of advances in cross-language information retrieval, third workshop of the CLEF, 2002* (pp. 28–48).

Chen, A., & Gey, F. (2004). Combining query translation and document translation in cross-language retrieval. In *Comparative evaluation of multilingual information access systems* (vol. 3237, pp. 108–121). Springer.

Chinnakotla, M. K., Raman, K., & Bhattacharyya, P. (2010). Multilingual PRF: English lends a helping hand. In *SIGIR* (pp. 659–666). ACM. doi:10.1145/1835449.1835559.

Di Nunzio, G. M., Ferro, N., Mandl, T., & Peters, C. (2007). Clef 2006: Ad hoc track overview. In *Proceedings of the 7th international conference on cross-language evaluation forum: Evaluation of multilingual and multi-modal information retrieval* (pp. 21–34). Berlin, Heidelberg, CLEF'06: Springer-Verlag, http://dl.acm.org/citation.cfm?id=2393955.2393960

Fang, H., Tao, T., & Zhai, C. (2004). A formal study of information retrieval heuristics. In *SIGIR* (pp. 49–56). ACM, doi:10.1145/1008992.1009004.

Fang, H., Tao, T., & Zhai, C. (2011). Diagnostic evaluation of information retrieval models. *ACM Transactions on Information Systems*. doi:10.1145/1961209.1961210.

Gao, W., Niu, C., Zhou, M., & Wong, K. F. (2009). Joint ranking for multilingual web search. In *ECIR* (pp. 114–125). Springer.

Jones, G. J., Burke, M., Judge, J., Khasin, A., Lam-Adesina, A., & Wagner, J. (2005). Dublin city university at CLEF 2004: Experiments in monolingual, bilingual and multilingual retrieval. In *CLEF* (pp. 207–220). Springer.

Kishida, K. (2005). Technical issues of cross-language information retrieval: A review. *Information Processing and management*, *41*(3), 433–455, doi:10.1016/j.ipm.2004.06.007, http://www.sciencedirect.com/science/article/pii/S0306457304000767.

Kraaij, W., & de Jong, F. (2004). Transitive probabilistic CLIR models. In *Proceedings of RIAO 2004*.

Kraaij, W., Nie, J. Y., & Simard, M. (2003). Embedding web-based statistical translation models in cross-language information retrieval. *Computational Linguistics*, *29*(3), 381–419. doi:10.1162/089120103322711587.

Lafferty, J., & Zhai, C. (2001). Document language models, query models, and risk minimization for information retrieval. In *SIGIR* (pp. 111–119). ACM, doi:10.1145/383952.383970.

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). New York, NY, USA, SIGIR '02: ACM, doi:10.1145/564376.564408.

Le Calvé, A., & Savoy, J. (2000). Database merging strategy based on logistic regression. *Information Processing and Management*, *36*(3), 341–359. doi:10.1016/S0306-4573(99)00036-9.

Lin, W. C., & Hsi, C. H. (2003). Description of NTU approach to NTCIR3 multilingual information retrieval. In *NTCIR workshop*.

Lin, W. C., & Hsi, C. H. (2004). Merging multilingual information retrieval result based on prediction of retrieval effectiveness. In *NTCIR workshop*.

Mandl, T., Womser-Hacker, C., Di Nunzio, G., & Ferro, N. (2008). How robust are multilingual information retrieval systems?. In *Proceedings of the 2008 ACM symposium on applied computing* (pp. 1132–1136). New York, NY, USA, SAC '08: ACM.

Martinez-Santiago, F., Urena Lopez, L., & Martin-Valdivia, M. (2006). A merging strategy proposal: The 2-step retrieval status value method. *Information Retrieval*, *9*, 71–93. doi:10.1007/s10791-005-5722-4.

Nie, J. Y. (2010). *Cross-language information retrieval. Synthesis lectures on human language technologies*. San Rafael: Morgan & Claypool Publishers.

Nie, J. Y., & Jin, F. (2002). Merging different languages in a single document collection. In *CLEF* (pp. 59–62). Springer.

Nie, J. Y., & Jin, F. (2003). A multilingual approach to multilingual information retrieval. In *CLEF, vol 2785* (pp. 101–110). Springer.

Nie, J. Y., Gao, J., & Cao, G. (2012). Translingual mining from text data. In C. C. Aggarwal & C. Zhai (Eds.), *Mining text data* (pp. 323–359). New York: Springer.

Nunzio, G. M., Ferro, N., Mandl, T., & Peters, C. (2008). *Advances in multilingual and multimodal information retrieval*. Berlin, Heidelberg: Springer-Verlag, chap CLEF 2007: Ad Hoc Track Overview, pp. 13–32.

Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational linguistics*, *29*(1), 19–51. doi:10.1162/089120103321337421.

Peters, C., Braschler, M., & Clough, P. (2012). *Multilingual information retrieval: From research to practice*. Berlin: Springer.

Powell, A. L., French, J. C., Callan, J., Connell, M., & Viles C. L. (2000). The impact of database selection on distributed searching. In *SIGIR* (pp. 232–239). doi:10.1145/345508.345584.

Savoy, J. (2002). Report on clef-2001 experiments: Effective combined query-translation approach. In *CLEF* (pp. 27–43). Springer, http://dl.acm.org/citation.cfm?id=648264.761432.

Savoy, J. (2003). Report on CLEF 2002 experiments: Combining multiple sources of evidence. In *CLEF* (vol 2785, pp. 66–90). Springer.

Savoy, J. (2004a). Combining multiple strategies for effective monolingual and cross-language retrieval. *Information Retrieval*, *7*(1–2), 121–148. doi:10.1023/B:INRT.0000009443.51912.e7.

Savoy, J. (2004b). Report on clef-2003 multilingual tracks. In *Comparative evaluation of multilingual information access systems* (vol. 3237, pp. 64–73). Springer.

Savoy, J., & Berger, P. Y. (2005). Selection and merging strategies for multilingual information retrieval. In *CLEF* (pp. 27–37). Springer.

Si, L., & Callan, J. (2006). Clef 2005: Multilingual retrieval by combining multiple multilingual ranked lists. In *CLEF* (pp. 121–130). Springer.

Si, L., Callan, J., Cetintas, S., & Yuan, H. (2008). An effective and efficient results merging strategy for multilingual information retrieval in federated search environments. *Information Retrieval*, *11*(1), 1–24. doi:10.1007/s10791-007-9036-6.

Sorg, P., & Cimiano, P. (2012). Exploiting wikipedia for cross-lingual and multilingual information retrieval. *Data and Knowledge Engineering*, *74*, 26–45. doi:10.1016/j.datak.2012.02.003.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In *Proceedings of the eight international conference on language resources and evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA).

Tsai, M. F., Wang, Y. T., & Chen, H. H. (2008). A study of learning a merge model for multilingual information retrieval. In *SIGIR* (pp. 195–202). ACM, doi:10.1145/1390334.1390370.

Xu, J., Weischedel, R., & Nguyen, C. (2001). Evaluating a probabilistic model for cross-lingual information retrieval. In *SIGIR* (pp. 105–110). doi:10.1145/383952.383968.

Zhai, C. (2008). Statistical language models for information retrieval: A critical review. *Foundations and Trends in Information Retrieval*, *2*(3), 137–213.
Zhai, C., & Lafferty, J. (2001a). Model-based feedback in the language modeling approach to information retrieval. In *CIKM* (pp. 403–410). ACM, doi:10.1145/502585.502654.
Zhai, C., & Lafferty, J. (2001b). A study of smoothing methods for language models applied to ad hoc information retrieval. In *SIGIR* (pp. 334–342). ACM, doi:10.1145/383952.384019.
Zhai, C., & Lafferty, J. (2002). Two-stage language models for information retrieval. In *SIGIR* (pp. 49–56). ACM.