Cross-Lingual Topical Relevance Models

Debasis Ganguly *Johannes Leveling* *Gareth J.F. Jones*
Centre for Next Generation Localisation (CNGL),
School of Computing, Dublin City University,
Dublin, Ireland
{dganguly, jleveling, gjones}@computing.dcu.ie

Abstract

Cross-lingual relevance modelling (CLRLM) is a state-of-the-art technique for cross-lingual information retrieval (CLIR) which integrates query term disambiguation and expansion in a unified framework, to directly estimate a model of relevant documents in the target language starting with a query in the source language. However, CLRLM involves integrating a translation model either on the document side if a parallel corpus is available, or on the query side if a bilingual dictionary is available. For low resourced language pairs, large parallel corpora do not exist and the vocabulary coverage of dictionaries is small, as a result of which RLM-based CLIR fails to obtain satisfactory results. Despite the lack of parallel resources for a majority of language pairs, the availability of *comparable* corpora for many languages has grown considerably in the recent years. Existing CLIR techniques such as cross-lingual relevance models, cannot effectively utilize these comparable corpora, since they do not use information from documents in the source language. We overcome this limitation by using information from retrieved documents in the source language to improve the retrieval quality of the target language documents. More precisely speaking, our model involves a two step approach of first retrieving documents both in the source language and the target language (using query translation), and then improving on the retrieval quality of target language documents by expanding the query with translations of words extracted from the top ranked documents retrieved in the source language which are thematically related (i.e. share the same concept) to the words in the top ranked target language documents. Our key hypothesis is that the query in the source language and its equivalent target language translation retrieve documents which share topics. The ovelapping topics of these top ranked documents in both languages are then used to improve the ranking of the target language documents. Since the model relies on the alignment of topics between language pairs, we call it the cross-lingual topical relevance model (CLTRLM). Experimental results show that the CLTRLM significantly outperforms the standard CLRLM by upto 37% on English-Bengali CLIR, achieving mean average precision (MAP) of up to 60.27% of the Bengali monolingual IR MAP.

Keywords: Cross-lingual Information Retrieval, Relevance Model, Topic Model, Pseudo-Relevance Feedback, Latent Dirichlet Allocation.

1 Introduction

Cross-language information retrieval (CLIR) involves retrieving documents in a language (target language), different from the language in which the users formulate their search (source language). A simple low resourced bi-lingual dictionary-based query translation followed by monolingual IR does not yield satisfactory results in CLIR mainly due to the poor vocabulary coverage of such a low resourced dictionary and the inherent ambiguities in query term senses (Hull and Grefenstette, 1996). More complex methods of query translation, e.g. statistical machine translation (SMT), perform better, but not entirely satisfactorily, due to the lack of availability of parallel resources such as sentence aligned corpora between resource-poor language pairs (Nie et al., 1999).

Regional languages of India are typical examples of languages with poor linguistic resources. The dominance of English, which has been used extensively as a medium of instruction and official work, can be exemplified by the fact that, while the English Wikipedia has almost 4M documents, the number of documents in the Hindi and Bengali Wikipedia are only around 100K and 23K respectively, although Hindi and Bengali ranks fourth and sixth respectively in terms of the number of native speakers[1]. The multi-linguality of Indian culture provides an ample motivation for the study of CLIR, e.g. a native Indian language speaker would often prefer to type his query in English due to his acquaintance with the English keyboard, although seeking to retrieve documents in his native language. Querying in English to a regional Indian language is thus a widespread real-life potential application for CLIR. The major hindrance to developing effective Indian language CLIR is the lack of *parallel corpora*, i.e. sentence aligned manually translated texts, to enable the development of effective standard translation tools. *Comparable corpora*, i.e. non sentence aligned texts which are not exact translations of each other but are roughly on the same topic, however are abundantly available owing to the growth of digital text in regional Indian languages. News articles published from the same source or from the same location in both English and a regional language over an identical time period are examples of such comparable corpora. Thus, despite the scarcity of parallel resources, significant comparable corpora are available for English and many Indian languages. This motivates us to research into new techniques effectively exploit these corpora for enhanced CLIR performance. This paper introduces and evaluates our proposed method to do this.

The main idea of our work can be outlined as follows. We assume that there exists a comparable corpus of documents in both the language of the query (source language) and the target language in which the documents need to be presented to the user. A query in the source language is translated into the target language by any available resource which may is typically a small bi-lingual dictionary for low resourced language pairs. Documents are then retrieved using both queries from the corresponding collections. It is a common practise in IR to improve upon the initial retrieval quality by utilizing information from the top-ranked documents (which are assumed to be relevant), the method being called pseudo-relevance feedback (PRF). PRF often does not work well when precision at top ranked documents is small. For our case, retrieval quality of the source language documents is expected to be much better than those of the target language ones, since the translated query in the target language is likely to be ambiguous and imprecise due to the lack of sufficient context in the short queries and the poor vocabulary coverage particularly for low resourced language pairs. Since the top ranked documents on the source side are more likely to be relevant to the information need, it can be hypothesized that utilizing this information for PRF can potentially improve retrieval results of the target language documents. Our proposed method thus relies on extracting translations of terms from the source language documents to expand the query

[1] http://en.wikipedia.org/wiki/Most_spoken_languages

in the target language so as to improve the retrieval quality on the target side.

The novelty of this paper is in the proposal of a PRF method which utilizes information from the source language documents to enhance retrieval effectiveness in the target language. The remainder of this paper is organized as follows. Section 2 describes the related work on CLIR and topic modelling. Section 3 describes the proposed method in details. Section 4 details the experimental setup, followed by Section 5, which evaluates the proposed method. Finally, Section 6 outlines the conclusions and directions for future work.

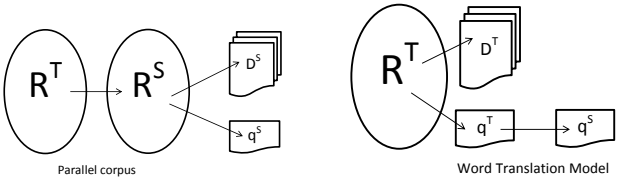
2 Related Work

This section starts with a brief review of the existing literature on general pseudo-relevance feedback (PRF). This is followed by a review on relevance models in IR, since our proposed method is a generalization of the cross-lingual relevance model (CLRLM). We then provide a brief survey on topic modelling applications in IR, as topic modelling is an integral part of our proposed method. Finally, we review existing work on combining pseudo-relevance feedback (PRF) evidences in multiple languages, since our method involves a combination of pseudo-relevance information from the source and the target languages.

Pseudo-Relevance Feedback (PRF). PRF is a standard technique in IR which seeks to improve retrieval effectiveness in the absence of explicit user feedback (Robertson and Sparck Jones, 1988; Salton and Buckley, 1990). The key idea of PRF is that the top ranked initially retrieved documents are relevant. These documents are then used to identify terms which can be added to the original query followed by an additional retrieval run with the expanded query often involving re-weighting of the query terms (Robertson and Sparck Jones, 1988; Hiemstra, 2000), and re-ranking initially retrieved documents by recomputing similarity scores (Lavrenko and Croft, 2001).

Relevance modelling in (CL)IR. Relevance modelling (RLM) is a state-of-the-art PRF technique involving estimation of a model of relevance generating terms both from pseudo-relevant documents and the query terms (Lavrenko and Croft, 2001). Terms which co-occur frequently with query terms are assigned a high likelihood of being generated from the RLM. In addition to monolingual IR, RLM has also been applied successfully to cross-lingual IR (CLIR) (Lavrenko et al., 2002), under the name of cross-lingual relevance model (CLRLM). A limitation of CLRLM is that it depends either on a parallel corpus or on a bilingual dictionary. However for low resourced languages, parallel corpus seldom exist and dictionaries have poor vocabulary coverage. We address this limitation by exploiting the topical overlap of top ranked documents retrieved in the source and target languages to improve the relevance model estimation. Our method thus only requires a comparable corpus instead of the more stringent requirement of a parallel corpus.

Topic modelling applications in IR. A widely used technique for topic modelling is the latent Dirichlet allocation (LDA) which treats every document as a mixture of multinomial distributions with Dirichlet priors (Blei et al., 2003). Various inference techniques have been proposed to estimate the probabilities in LDA, including variational Bayes, expectation propagation, and Gibbs sampling (Blei et al., 2003; Griffiths and Steyvers, 2004). We use the Gibbs sampling method for LDA inference because it is computationally faster and has been shown to outperform the other two (Griffiths and Steyvers, 2004). LDA was applied for monolingual IR by (Wei and Croft, 2006). Their work involves estimating the LDA model for the whole collection by Gibbs sampling and then linearly combining the LM term weighting with LDA-based term weighting. An early attempt to utilize the topical structure in language pairs for CLIR can be found in (Littman et al., 1998), which involved automatic construction of a multi-lingual semantic space using a topic modelling technique



(a) Parallel corpus (b) Bilingual dictionary

Figure 1: Schematic diagrams of a CLRLM.

called latent semantic indexing (LSI). The major limitation of their method is that it relies on the existence of a parallel corpus for the training phase. In contrast, we leverage upon the existence of a comparable corpus to improve search results in the target language.

A recent work (Vulić et al., 2011) overcomes the parallel corpus restriction of CLRLMs by training a CLRLM on a comparable corpus using topic models inferred by bilingual LDA (Bi-LDA), which is a special case of polylingual LDA (Mimno et al., 2009). A major difference of their work with ours is that their method requires a separate training phase on comparable corpora to estimate the latent topic models in Bi-LDA. In fact, the authors used external resources such as Europarl and Wikipedia for the training purpose. In contrast, our method does not require a separate training phase or external resources. Moreover, our method applies topic modelling only on the set of pseudo-relevant documents. Furthermore, the idea of topic modelling in (Vulić et al., 2011) is only loosely coupled within the CLRLM framework, whereas we tightly integrate the topic modelling step in a graphical model with added nodes for the latent topics.

Feedback model combination. (Chen et al., 2010) exploited comparable corpora for CLIR by training *learning-to-rank* methods on out-of-domain source data to improve the retrieval effectiveness of the target domain. The Multi-PRF method improves monolingual retrieval by applying PRF in an assisting language, and mapping the feedback terms back in the language of the query with the help of a dictionary (Chinnakotla et al., 2010). The similarity of our method with Multi-PRF is that both involve an intermediate retrieval step in a language different from the language of the query. However, there are several differences which are highlighted as follows. Firstly, Multi-PRF improves monolingual retrieval by information from another language (typically English), whereas our proposed method improves CLIR performance. Secondly, Muti-PRF does not take into consideration the latent topics in the pseudo-relevant documents of the two languages, whereas topic modelling plays a crucial role in our approach.

3 Cross-lingual Topical Relevance Models

This section describes our proposed model in detail. We start the section with a brief motivation, where we discuss the limitations of CLRLM and how these can possibly be addressed. We then describe the schematics of our model which is then followed by the estimation details. Finally we present the algorithmic details for implementing the model.

3.1 Motivation

A limitation of CLRLM is that it depends either on a parallel corpus or on a bilingual dictionary to estimate the target language document models essential for estimating the relevance model for the query (source) language. The schematic diagrams of Figure 1a and 1b illustrate this. In the parallel corpus based approach, for every top ranked document retrieved in the source language

D_j^S, the corresponding document D_j^T in the target language is used to compute R^T, the estimated relevance model for the target language. This is shown in Figure 2a where the edge from $\mathbf{D}^T$ to $\mathbf{D}^S$ represents the event of transforming each document of the target language to its equivalent in the source language. The estimated probability of relevance is thus

$$P(w^T|\mathbf{q}^S) = \sum_{j=1}^{R} \underbrace{P(w^T|D_j^T)\, P(D_j^S|\mathbf{q}^S)}_{D_j^S \text{ parallel to } D_j^T} \tag{1}$$

where $P(D_j^S|\mathbf{q}^S)$ is the standard LM similarity of the given query $\mathbf{q}^S$ with a document D_j^S (Hiemstra, 2000). The parallel corpus based approach to CLRLM thus involves document side translation. A complementary approach is query side translation, as done in the bilingual dictionary-based CLRLM method shown in Figure 2b. The edge from $\mathbf{q}^T$ to $\mathbf{q}^S$ indicates generation of the query vector in the source language from a query vector in the target language via translation. The estimated probability of relevance in this case is given by

$$P(w^T|\mathbf{q}^S) = \sum_{j=1}^{R} P(w^T|D_j^T)P(D_j^T|\mathbf{q}^S) = \sum_{j=1}^{R} P(w^T|D_j^T) \prod_{i'=1}^{n^T} P(D_j^T|q_{i'}^T) \underbrace{\sum_{i=1}^{n^S} P(q_{i'}^T|q_i^S)P(q_i^S)}_{\text{word-based query translation}} \tag{2}$$

While it is possible to apply CLRLM in the presence of a bilingual dictionary, these dictionaries for low resourced languages cover only a very small part of the vocabulary, as a result of which it becomes impossible to compute the probabilities $P(s|t)$ for most (s, t) pair (s and t refer to a word in the source language and the target language respectively). This in turn results in a poor performance of CLRLM for such low resourced languages.

The limitations of the current CLRLM method are that: a) it depends either on document translation with the help of a parallel corpus or on query translation with the help of a dictionary, without making provisions for a combination of the two; b) the document side translation depends on the availability of a parallel corpus which is rare for resource-poor languages; and c) it does not model the multiple aspects of relevance that are implicitly or explicitly expressed in a query. Assuming that there exists a comparable document collection in the two languages, the first two restrictions can be overcome with a two step retrieval process, one with the source language query and the other with a translated query in the target language to obtain separate working sets of documents in both the languages. The working set of the top ranked documents in the two languages, which

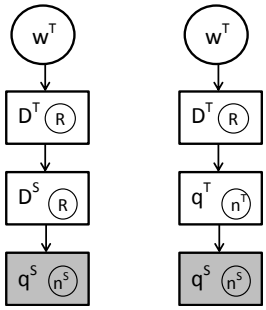


Figure 2: CLRLM dependence graphs for a) Parallel corpus (left) and b) Dictionary (right).

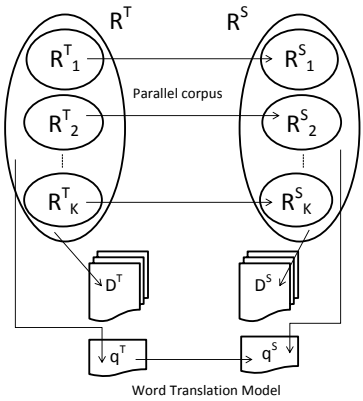


Figure 3: Schematic representation of a CLTRLM.

we refer to as pseudo-relevant document sets from now on, can thus potentially replace the parallel corpus requirement of the CLRLM. However, it is impractical to assume a one-one document level alignment between the pseudo-relevant documents in the two languages. Segmenting the pseudo-relevant documents into topics can result into a more accurate alignment at the level of topics rather than at the level of documents, with the underlying hypothesis that the documents retrieved in the two languages comprise of words related to overlapping concepts or topics. Topic modelling on the pseudo-relevant sets of documents also help in overcoming the third limitation where each topic addresses one particular sub information need associated with the query.

Our proposed methodology which we call cross-lingual topical relevance model (CLTRLM), involves estimation of two separate relevance models for both the source and target language pseudo-relevant documents. The target language relevance model is then updated by applying a word translation model to transform a word from each topic in the source language to a word in the target language. CLTRLM aims to achieve the benefits of a parallel corpus without actually having one. A topic level decomposition and mapping is helpful in adding contributions from the most likely topic, i.e. aspect of relevance, from the source language to a topic on target side. Note that the documents on which we apply topic modelling are the top ranked documents retrieved in response to the query in both the source language and its target language translation. Both document sets being focused on the query ensures that these documents share common topics. This is contrary to the approach of (Vulić et al., 2011), where the full comparable corpus is used for topic modelling. The working principle of a CLTRLM is illustrated schematically in Figure 3, which shows that the relevance models for both the source and the target languages have been split into topics. Each topic in a relevance model may refer to an individual aspect of the broad information need expressed in the query, and is thus associated with generating relevant documents related to that particular aspect. In contrast to the broad information need of a query, each particular aspect is more focused, and is thus easier to align from the target side to the source side. Although Figure 3 shows that the number of topics in the source and the target relevance models are identical, the number of topics may in fact be different on the source and the target sides. The next section presents the details of CLTRLM more formally.

3.2 Formal Description

Figure 4 shows the dependence network of CLTRLM in plate notation. Let w^T be a word in the target language. The top-most circle in Figure 4 represents a word in the target language for which

the objective is to calculate the probability $P(w^T|\mathbf{q}^S)$, i.e. to estimate the probability of generating this word from a hypothetical relevance model R^T. It is not possible to estimate this model directly because in ad-hoc retrieval, no prior information is provided about the relevant documents. The only observable entities are the query terms in the source language as shown in Figure 4. Let us denote this query by $\mathbf{q}^S$. The shaded rectangle at the bottom represents the vector q_i^S of observed variables having dimensionality n^S, each component representing the term frequency of a unique query term in the source language. The best way to estimate the probability $P(w^T|R^T)$ is thus to approximate it with $P(w^T|\mathbf{q}^S)$ i.e. $P(w^T|R^T) \approx P(w^T|\mathbf{q}^S)$. The rectangles $\mathbf{z}^T$ and $\mathbf{z}^S$ denote vectors of dimensionality K^T and K^S, the number of topics on the target and source sides respectively. While estimating the model, we assume that the topics for both the source and target langages have been obtained by latent Dirichlet allocation (LDA) (Blei et al., 2003), and thus we use the LDA estimated values viz. $\hat{\boldsymbol{\theta}}^T$, $\hat{\boldsymbol{\phi}}^T$, $\hat{\boldsymbol{\theta}}^S$ and $\hat{\boldsymbol{\phi}}^S$ in CLTRLM inference. The rectangles marked by $\mathbf{D}^T$ and $\mathbf{D}^S$ denote the set of top ranked R^T and R^S documents retrieved respectively for the target and source languages. $\mathbf{q}^T$ represents the translation of the observed query vector in the source language viz. $\mathbf{q}^S$, and is obtained by a bilngual dictionary or using a machine translation (MT) system. With these notations, we are now ready to work out the estimation details in the following subsection.

3.3 Estimation Details

With reference to Figure 4, the estimation of CLTRLM proceeds as follows.

$$\begin{aligned} P(w^T|\mathbf{q}^S) = & \underbrace{P(w^T|\mathbf{z}^T)P(\mathbf{z}^T|\mathbf{q}^S)}_{\text{target language generation event}} + \underbrace{P(w^T|\mathbf{w}^S)P(\mathbf{w}^S|\mathbf{q}^S)}_{\text{source language generation event}} \\ & = \sum_{k=1}^{K^T} P(w^T|z_k^T, \hat{\boldsymbol{\phi}^T}_{k,w^T})P(z_k^T|\mathbf{q}^S) + \sum_{i=1}^{t(w^T)} P(w^T|w_i^S)P(w_i^S|\mathbf{q}^S) \end{aligned} \tag{3}$$

Equation (3) represents two chains of events via the two components shown: one associated with the target language retrieved documents obtained by query translation, and the other associated with $t(w^T)$ possible translations of word w^T in the source language (denoted by the set w^S), which in turn correspond to the words in the source language retrieved documents. Note that the two generation events pertain to the topic level alignment introduced informally in the previous section, where a word in the source language query can either be generated from a topic in the source language or from an *equivalent* topic in the target language. Inferencing along the left chain proceeds as follows.

$$\begin{aligned} & P(z_k^T|\mathbf{q}^S) = \sum_{j=1}^{R^T} P(z_k^T|D_j^T)P(D_j^T|\mathbf{q}^S) = \sum_{j=1}^{R^T} P(z_k^T|D_j^T) \sum_{i'=1}^{n^T} P(D_j^T|q_{i'}^T)P(q_{i'}^T|\mathbf{q}^S) = \\ & \sum_{j=1}^{R^T} P(z_k^T|D_j^T) \sum_{i'=1}^{n^T} P(D_j^T|q_{i'}^T) \sum_{j'=1}^{n^S} P(q_{i'}^T|q_{j'}^S)P(q_{j'}^S) = \sum_{j=1}^{R^T} P(z_k^T|D_j^T, \hat{\theta}_{j,k}^T) \underbrace{\sum_{i'=1}^{n^T} \frac{P(q_{i'}^T|D_j^T)}{R^T}}_{\text{LM similarity}} \underbrace{\sum_{j'=1}^{n^S} P(q_{i'}^T|q_{j'}^S)}_{\text{query translation}} \end{aligned} \tag{4}$$

In the last line of Eq. (4) we have ignored the prior probability of $P(q_{j'}^S)$, and used the LDA estimated $\hat{\boldsymbol{\theta}}^T$ values for computing $P(z_k^T|D_j^T)$ and the standard LM similarity score $P(q_{i'}^T|D_j^T)$ to compute the probability of generating a target language query term from a target language document model.

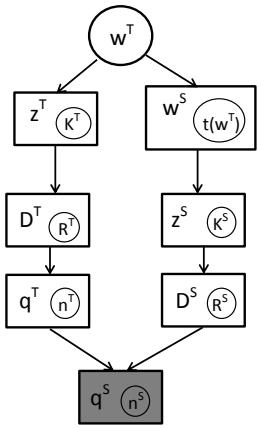


Figure 4: CLTRLM dependence graph in plate notation.

Similarily, the right side chain can be inferred as

$$P(w_i^S|\mathbf{q}^S) = \frac{1}{R^S}\underbrace{\sum_{k=1}^{K^S} P(w_i^S|z_k^S,\hat{\boldsymbol{\phi}}^S)\sum_{j=1}^{R^S} P(z_k^S|D_j^S,\hat{\boldsymbol{\theta}}^S)}_{\text{LDA document model of } D_j^S}\underbrace{P(D_j^S|\mathbf{q}^S)}_{\text{LM similarity}} = \sum_{j=1}^{R^S}\frac{P_{LDA}(w_i^S,D_j^S,\hat{\boldsymbol{\theta}}^S,\hat{\boldsymbol{\phi}}^S)P(D_j^S|\mathbf{q}^S)}{R^S} \tag{5}$$

where we have used the notation $P_{LDA}(.)$ to denote the LDA estimated probabilities marginalized over the latent topic variables. Substituting Eq. (4) and Eq. (5) into (3) gives the full expression of the probability of generating a target language word from the relevance model in case of CLTRLM.

$$P(w^T|\mathbf{q}^S) = \underbrace{\left(\sum_{k=1}^{K^T} P(w^T|z_k^T,\hat{\phi}^T_{k,w^T})\sum_{j=1}^{R^T} P(z_k^T|D_j^T,\hat{\theta}^T_{j,k})\sum_{i'=1}^{n^T}\frac{P(q_{i'}^T|D_j^T)}{R^T}\sum_{j'=1}^{n^S} P(q_{i'}^T|q_{j'}^S)\right)}_{\text{target language contribution estimated by query translation}} + \underbrace{\left(\sum_{i=1}^{t(w^T)} P(w^T|w_i^S)\times\sum_{j=1}^{R^S} P_{LDA}(w_i^S,D_j^S,\hat{\boldsymbol{\theta}}^S,\hat{\boldsymbol{\phi}}^S)\frac{P(\mathbf{q}^S|D_j^S)}{R^S}\right)}_{\text{source language contribution estimated by document word translation}} \tag{6}$$

In Equation 6, $\hat{\phi}^T_{k,w^T}$ denotes the probability of the word w^T belonging to the topic k, whereas $\hat{\theta}^T_{j,k}$ denotes the probability of the k^{th} topic in the j^{th} document. Both these quantities can be computed from the LDA estimation output matrices $\boldsymbol{\theta}^T$ and $\boldsymbol{\phi}^T$.

Also note that the CLTRLM as shown in Figure 4 involves two possible ways of generating the query in the source language, i.e. either directly using documents retrieved in the target language, or by using translations of words in documents retrieved in the source language. Thus, a natural question which arises is whether we need to introduce a new linear combination parameter to choose the two event paths with relative weights similar to (Chinnakotla et al., 2010). However, a closer look at Equation 3 reveals that the contribution from each path is inherently controlled by the two coefficients $P(w^T|\mathbf{z}^T)$ and $P(w^T|\mathbf{w}^S)$, thus eliminating the need for an extra parameter.

3.4 Estimation with Bi-LDA

In Section 3.3 we worked out the estimation details assuming that the source and the target language documents have different topic distributions denoted by the parameters $\hat{\boldsymbol{\theta}}^{\boldsymbol{S}}$ and $\hat{\boldsymbol{\theta}}^{\boldsymbol{T}}$ respectively. The CLTRLM estimation can also be performed with a tighter assumption that document pairs in the source and target languages share the same distribution of topics say $\hat{\boldsymbol{\theta}}$, with different topic-word distribution parameters say $\hat{\boldsymbol{\phi}}^{\boldsymbol{S}}$ and $\hat{\boldsymbol{\phi}}^{\boldsymbol{T}}$ respectively. This is a special case of polylingual LDA as proposed in (Mimno et al., 2009). For Bi-LDA estimation of CLTRLM, firstly we impose the restriction of $R^T = R^S$, i.e. to retrieve the same number of documents on the source and target sides, and secondly we set $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}^{\boldsymbol{T}} = \hat{\boldsymbol{\theta}}^{\boldsymbol{S}}$ in Equation 6. We refer to CLTRLM instances with Bi-LDA estimation as JCLTRLM (Joint CLTRLM). Note that a JCLTRLM has two parameters $R = R^T = R^S$ and $K = K^T = K^S$ as opposed to four of CLTRLM.

3.5 Algorithm

After presenting the estimation details, we now provide the implementation steps for CLTRLM.

1) Run initial retrieval on the source language query $\mathbf{q}^S$ using standard LM to get documents $\{D_j^S\}_{j=1}^R$ in the source language (Ponte, 1998) and let R^S be the number of top ranked documents assumed to be pseudo-relevant.
2) Use a source-target language dictionary to get the equivalent query $\mathbf{q}^T$ in the target language.
3) Retrieve documents $\{D_j^T\}_{j=1}^R$ using LM for the target language query $\mathbf{q}^T$, and assume that the top ranked among these are pseudo-relevant.
4) Perform LDA inference by N iterations of Gibbs sampling on the working sets $\{D_j^S\}_{j=1}^{R^S}$ and $\{D_j^T\}_{j=1}^{R^T}$ to estimate the parameters $\hat{\boldsymbol{\theta}}^S$, $\hat{\boldsymbol{\phi}}^S$, $\hat{\boldsymbol{\theta}}^T$ and $\hat{\boldsymbol{\phi}}^T$. For the case of JCLTRLM, use Bi-LDA to estimate parameters.
5) Let V^T be the vocabulary of $\{D_j^T\}_{j=1}^{R^T}$. For each word $w^T \in V^T$, use Eq. (6) to compute the probability of relevance $P(w|R^T) \approx P(w|\mathbf{q}^S)$.
6) Rerank every target language document $\{D_j^T\}_{j=1}^{R^T}$ by the KL divergence between its LM document model (as obtained by the initial retrieval) and the estimated $P(w^T|R^T)$ so as to get the final retrieval result.

The computational complexity of the above algorithm is $O((V^T + V^S)(R^T + R^S)(K^T + K^S)N)$ where V^T, V^S are the vocabulary sizes of $\{D_j^T\}_{j=1}^{R^T}$ and $\{D_j^S\}_{j=1}^{R^S}$ respectively, R^T and R^S is the number of pseudo-relevant documents, K is the number of topics, and N is the number of iterations used for Gibbs sampling. The computational complexity of CLRLM on the other hand is $O(V^T R^T)$. CLTRLM, as compared to CLRLM, has the added computational cost for the source language retrieved documents. However, it is expected that $V^S = O(V^T)$, $R^S = O(R^T)$, and that both K^T, K^S and N are small constant numbers independent of R^T and V^T. Thus, CLTRLM is only a constant times more computationally expensive than CLRLM.

4 Experimental Setup

In this section, we describe the details of our experimental setup for evaluating CLTRLM. Our experiments explore the following questions: a) Does integrating the event of generating a target language word from the source language lead to a better estimation of the generative model for relevance compared to CLRLM? b) Does the use of latent topics benefit the alignment between source and target language documents and thus leads to a better estimation of relevance? c) What is the effect of translation quality on the performance of CLTRLM estimation? and d) How does the performance of Bi-LDA estimation for JCLTRLM, and separate LDA estimation for CLTRLM compare against each other? To answer a), we compare CLTRLM against CLRLM on cross-lingual ad-hoc search. To explore question b), we instantiate CLTRLM with the number of topics on the

	Documents			Queries			
Language	Type	# Docs.	Vocab. size [words]	Field	# Queries	Avg. query len.	Avg. # rel. docs.
Bengali	News articles	123,048	641,308	title	50	3.6	13.6
English	News articles	125,586	318,922	title	50	4.9	10.2

Table 1: FIRE-2010 document and query characteristics.

source and target sides set to 1, i.e. we use $K^S = K^T = 1$ in Equation 6, and use this instantiation of CLTRLM as one of our baselines. To answer c), we obtain different translation qualities by applying a bilingual dictionary and Google translate[2], which is a free to use statistical machine translation service. The presence of OOV words across language pairs can impair the estimation quality of CLTRLM because for every target language word whose translation is not found in the dictionary, we fail to get the source language contribution in the generation probability (see Equation 3). To reduce the vocabulary gap we use transliteration of OOV words, since it has been reported that transliteration helps improve the retrieval quality of CLIR (Udupa et al., 2009). Finally to address question d), we evaluate the relative performance of CLTRLM and JCLTRLM.

Data set. We perform CLIR experiments from English to Bengali, i.e. the query is expressed in English (source language), and the objective is to retrieve documents in Bengali (target language). Experiments are conducted on the English and Bengali ad-hoc collections of FIRE-2010 dataset (Majumder et al., 2010), the documents of which comprise a comparable corpus of news articles published from the same news agency in Calcutta namely *Ananadabazar* and *The Telegraph* in Bengali and English respectively. Table 1 outlines the document collection and query characteristics. Note that we do not use any external parallel or comparable resources to train our model as was done in (Vulić et al., 2011; Littman et al., 1998; Chen et al., 2010).

Stopwords. The stopword list used for Bengali retrieval was the one provided by the track organizers[3] generated by following an approach of extracting the N most frequent words from the document collection (Fox, 1992; Savoy, 1999). This list comprises of 384 Bengali words. The stopword list used for English was the standard SMART stopword list which comprises of 573 words.

Stemming. We used a simple and moderately aggressive rule-based stemmer[4] for Bengali retrieval (Leveling et al., 2010). The stemmer used for our experiments is able to increase mean average precision (MAP) by 21.82% (0.2250 to 0.2741) on the monolingual title-only Bengali queries of the FIRE-2010 test collection. Although there are more complex corpus-based approaches reported for Bengali stemming (Majumder et al., 2007; Paik et al., 2011), the focus of this paper is not on improving stemming, but rather to improve on cross-lingual retrieval performance from English to Bengali. We thus applied a simple rule based approach to stemming which does not require computationally intensive pre-processing over the vocabulary of the corpus. The stemmer used on the English side is the default stemmer of SMART, a variant of Lovin's stemmer (Lovins, 1968).

Translation. One of the major components of CL(T)RLM is the bilingual dictionary to translate a given query in the source language viz. $\mathbf{q}^S$ to the corresponding representation in the target language

[2] http://translate.google.com/#en|bn|

[3] http://www.isical.ac.in/~fire/stopwords_list_ben.txt

[4] http://www.computing.dcu.ie/~dganguly/rbs.tar.gz

$\mathbf{q}^T$. In our experiments, we used the open source English-Bengali dictionary *Ankur*[5]. The dictionary at the time of writing this paper comprises of 9180 English words for each of which one or more Bengali word translations are provided. It can be seen that the vocabulary coverage of the *Ankur* English-Bengali dictionary is very small, in fact covering only 1.43% of the total vocabulary size of the English corpus (see Table 1), as a result of which a significant number of query words remain untranslated and hence play no role in estimating the CLTRLM. To increase the vocabulary coverage, and hence improve on the retrieval quality, we used Google translator, which is a statistical MT web-service, to translate the English queries to Bengali.

Transliteration. The out of vocabulary (OOV) words with respect to both the Google translator and the English-Bengali dictionary were left in the original English form. A manual inspection of these translated queries by one of the authors, who is a native Bengali speaker, revealed that most of these OOV words are English proper nouns. Proper nouns are important for retrieval (Xu and Croft, 2000), and thus need to be handled appropriately. An intuitive approach is to transliterate English names into Bengali which has proved to be beneficial for Indian language CLIR (Udupa et al., 2009). For transliteration, we applied *Google transliterate*[6] on the untranslated words of the Bengali queries as obtained by the dictionary-based and the Google translator approaches. Google transliterate returns five transliterations for each given word in decreasing order of likelihood. Out of these five Bengali transliterations for each English word, we use the top ranked one i.e. the most likely one. This simplistic approach of taking the most probable candidate from Google transliterate may not yield accurate transliterations. However, the focus of the paper is not to improve on the English-Bengali transliteration process itself, but rather to use transliteration as an intermediate tool to improve CLIR performance. Furthermore, an incorrect transliteration of a query term hardly has any effect on retrieval performance, since it is highly unlikely for an incorrectly transliterated word to match with the indexed vocabulary.

Table 2 shows the quality of the query translations obtained by the four methods. Translation quality of the English queries translated into Bengali is measured by the metric BLEU (Papineni et al., 2002), by using FIRE-2010 Bengali query titles as reference translations. We in fact report two versions, one which computes the BLEU score using the original word forms, and the other on the stemmed versions of the translated text and reference. The latter is a more appropriate measure for CLIR, because for the case of CLIR it is sufficient to match the stemmed forms rather than matching corresponding original word forms of a translated word with its reference. It can be seen that the BLEU scores are rather low compared to language pairs such as English-French (Dandapat et al., 2012), which is indicative of the fact that translation from English-Bengali is a difficult problem indeed. Application of transliteration however results in a significant improvement of BLEU score, indicating the importance of handling the OOV words and obtaining the equivalent form of proper nouns in the target language.

CL(T)RLM Implementation. CLTRLM has been implemented as an extension to SMART[7] along with the CLRLM approach which is used as one of the baselines. The other baseline using Google translator involves estimating a monolingual relevance model (Lavrenko and Croft, 2001) which is also implemented in SMART. GibbsLDA++[8] was employed for Gibbs sampling for LDA inference. A modified version of GibbsLDA++ was used for Bi-LDA inference for JCLTRLM estimation.

[5] http://www.bengalinux.org/english-to-bengali-dictionary/
[6] https://developers.google.com/transliterate/
[7] ftp://ftp.cs.cornell.edu/pub/smart/
[8] http://gibbslda.sourceforge.net/

Translation Method	BLEU	
	Unstemmed	Stemmed
Dictionary	6.13	6.71
Dictionary + Google transliterate	7.46	**8.79**
Google translate	6.50	7.61
Google translate + Google transliterate	7.47	8.64

Table 2: English-Bengali query translation qualities.

5 Results

This section reports the results of the experiments and analyzes the observations. We start with a detailed description of the retrieval runs and parameter settings.

Retrieval run description. The CLRLM baselines shown in Table 3 start with the prefix "CLRLM". Results are shown for each method of translation (Google translator or the dictionary-based translation), with or without transliteration on the translation output, thus yielding 4 cases. The CLRLM approach does not use any information from the documents retrieved in the source language. To show that source language information is beneficial for retrieval, we report runs which use only the target language path in the generative model shown in Figure 4, i.e. for these runs we set $P(w^T|\mathbf{w}^S)$ (see Equation 3) to zero. These runs are shown the prefix "TgtCLTRLM" in Table 3. To show that topic decomposition is essential, we set the number of topics to 1 on both the source and the target sides, i.e. we set K^T and K^S to 1. These runs benefit from the information on the source side, but do not use the topical structures of documents to help achieve a fine grained alignment of the topics. These run names are prefixed with "UCLTRLM" in Table 3. Finally, we report (J)CLTRLM results prefixed with (J)CLTRLM.

Parameter settings. Each run reported in Table 3 has been optimized with the best parameter settings. The parameters were varied as follows. The Jelinek-Mercer language modelling smoothing parameter (Hiemstra, 2000) of initial retrieval for all the runs was empirically chosen as 0.4. The hyper-parameters α and β which control the Dirichlet distributions for CLTRLM, were set to $\frac{50}{K}$ (K being the number of topics) and 0.1 respectively as described in (Griffiths and Steyvers, 2004). The number of iterations for Gibbs sampling i.e. N, was set to 1000 for all CLTRLM experiments. We tuned the common parameters i.e. R^S and R^T, i.e. the number of top ranked documents used for pseudo-relevance in the source and target languages respectively, within the range of $[10, 50]$ in steps of 10 so as to obtain the best settings. An important parameter to CLTRLM is the number of topics on the source and target sides viz. K^S and K^T. These parameters were empirically optimized within the range of $[5, 50]$. The justification of using a much smaller value range for the number of topics, in comparison to the global LDA based approach (Wei and Croft, 2006) which used much higher values of K in the range of 100 to 1500, comes from the fact that LDA estimation in the CLTRLM is done on only a small number of documents rather than on the full corpus.

Observations. With reference to Table 3, it can be seen that the initial retrieval run NOFDBK-DICT performs very poorly achieving only 21.58% of the MAP as compared to the monolingual retrieval run NOFDBK-MONOLINGUAL. This shows that a low resourced dictionary-based query translation does not yield satisfactory retrieval performance. A cross-lingual relevance model based feedback approach is able to significantly[9] improve on the initial retrieval MAP by 42.40% as can be

[9]Statistical *significance* or statistically *(in)distinguishable* henceforth refer to Wilcoxon test with 95% confidence measure.

Approach	Query Processing		Parameters					Results	
	Translation	Transliteration	PRF	R^T	R^S	K^T	K^S	MAP	P@5
NOFDBK-DICT	Dictionary	N	N	-	-	-	-	0.0592	0.0653
CLRLM-DICT	Dictionary	N	Y	30	-	1	-	0.0843	0.1102
UCLTRLM-DICT	Dictionary	N	Y	30	20	1	1	0.0935	0.1224
TgtCLTRLM-DICT	Dictionary	N	Y	30	0	15	0	0.1069	0.1388
CLTRLM-DICT	Dictionary	N	Y	30	20	15	5	**0.1130**	**0.1592**
JCLTRLM-DICT	Dictionary	N	Y	10	10	10	10	0.1086	0.1551
NOFDBK-DICT-TLIT	Dictionary	Y	N	-	-	-	-	0.0996	0.1120
CLRLM-DICT-TLIT	Dictionary	Y	Y	30	-	1	-	0.1156	0.1440
UCLTRLM-DICT-TLIT	Dictionary	Y	Y	30	20	1	1	0.1237	0.1560
TgtCLTRLM-DICT-TLIT	Dictionary	Y	Y	20	0	10	0	0.1369	0.1400
CLTRLM-DICT-TLIT	Dictionary	Y	Y	20	5	10	5	0.1446	0.1720
JCLTRLM-DICT-TLIT	Dictionary	Y	Y	10	10	20	20	**0.1588**	**0.1800**
NOFDBK-SMT	Google	N	N	-	-	-	-	0.0843	0.1080
CLRLM-SMT	Google	N	Y	30	-	1	-	0.1208	0.1520
UCLTRLM-SMT	Google	N	Y	30	20	1	1	0.1274	0.1640
TgtCLTRLM-SMT	Google	N	Y	30	0	15	0	0.1373	**0.1840**
CLTRLM-SMT	Google	N	Y	30	20	15	5	0.1425	0.1800
JCLTRLM-SMT	Google	N	Y	30	30	10	10	**0.1441**	0.1800
NOFDBK-SMT-TLIT	Google	Y	N	-	-	-	-	0.1024	0.1240
CLRLM-SMT-TLIT	Google	Y	Y	30	-	1	-	0.1393	0.1720
UCLTRLM-SMT-TLIT	Google	Y	Y	30	20	1	1	0.1483	0.1840
TgtCLTRLM-SMT-TLIT	Google	Y	Y	30	0	10	0	0.1523	0.1680
CLTRLM-SMT-TLIT	Google	Y	Y	30	20	10	5	0.1648	**0.2000**
JCLTRLM-SMT-TLIT	Google	Y	Y	20	20	5	5	**0.1652**	0.1920
NOFDBK-MONOLINGUAL	-	-	N	-	-	-	-	0.2741	0.3160

Table 3: Results for English-Bengali CLIR experiments.

seen by comparing run CLRLM-DICT with NOFDBK. The CLRLM run only retrieves documents in the target language i.e. Bengali in this case. MAP is further improved by 10.91% by using documents retrieved in the source language i.e. English as seen from the run UCLTRLM-DICT in comparison to CLRLM-DICT. This run is a corner-case of a CLTRLM without using topical decompositions on the source and target sides i.e. by setting $K^T = K^S = 1$. Topical decomposition on the target side alone (see "TgtCLTRLM" prefixed runs) produces better results than the CLRLM runs, but are outperformed by the runs which use information from the source side as well, as shown by the (J)CLTRLM runs. It can be seen that both topical decomposition and using information from the source-side play a crucial role in correctly estimating the relevance model.

It turns out that Bi-LDA inference of CLTRLM i.e. JCLTRLM performs better than separately inferencing the topic models on the source and target sides for all scenarios except the one which only uses the dictionary. The reason JCLTRLM performs poorly on dictionary-based approach is that the initial retrieval results on the target language is very poor (MAP: 0.0592). It is thus not a reasonable assumption that the document pairs on the source and target sides share the same topic distribution $\hat{\boldsymbol{\theta}}$. In such a scenario, it is helpful to use different number of documents on the source and target sides for the LDA estimation. However, when the initial retrieval quality improves on the

target side with the help of transliteration or SMT or both, it is observed that Bi-LDA estimation proves more beneficial because of a better level of agreement between the pseudo-relevant document sets in the source and target languages. The improvements obtained by JCLTRLM over CLTRLM are statistically indistinguishable with respect to MAP, except for the DICT-TLIT case.

To summarize, we observe that the CLTRLM runs (even with only one topic), using information from the source side, perform significantly better than CLRLM runs which in turn do not use any source information, thus indicating that *documents retrieved in the source language lead to a better relevance estimation*. This observation conforms to the Multi-PRF results where it was shown that using information from another language retrieval results helps improve the monolingual retrieval result in the language of the query (Chinnakotla et al., 2010). The CLTRLM approach achieves a similar benefit for CLIR. We also observe that the CLTRLM runs outperform the UCLTRLM prefixed runs, thus establishing that *the use of latent topics does benefit the alignment between source and target language documents*. The reason is that incorporating latent topic nodes in the estimation process helps, firstly in focusing co-occurrence computation on topics rather than on full documents, and secondly the alignment process of the fine grained topics is more reliable than aligning full documents in the source and target languages. The fact that CLTRLM improves retrieval effectiveness significantly over a range of query translation qualities, suggests that the method is robust to translation quality. Furthermore, JCLTRLM turns out to be slightly better than CLTRLM suggesting that Bi-LDA estimation is marginally better than separate LDA estimation.

Comparison to other reported results on Indian language CLIR. To the best of our knowledge, no results have been reported for fully automatic English-Bengali CLIR. (Leveling et al., 2010) report that manual English-Bengali query translation by native Bengali speakers achieves MAP up to 83.3% compared to monolingual Bengali retrieval, but on longer *TD* (title-description) queries. The fact that manual query translation by native Bengali speakers achieves 83.3% retrieval effectiveness in comparison to monolingual IR, demonstrates that English-Bengali CLIR is a considerably hard problem to solve. Our fully automatic approach achieves a satisfactory performance increasing MAP by 90.87% (see the rows CLTRLM-DICT and NOFDBK-DICT) compared to translation with a base dictionary, and achieves a MAP of 60.27% of the monolingual upper baseline (compare JCLTRLM-SMT-TLIT with NOFDBK-MONOLINGUAL).

6 Conclusions and Future work

This paper presented CLTRLM, a novel theoretical framework for exploiting the topical association of terms between pseudo-relevant documents of the source and target languages, to improve CLIR effectiveness. CLTRLM is a generalization of the standard cross-lingual relevance model, overcoming its limitations of: a) incompatibility with a comparable corpus; and b) co-occurrence computation at the level of whole documents, instead of likely relevant topics. CLTRLM uses a *comparable* corpus for IR on the retrieved set of top-ranked documents without the requirement of a separate training phase on the whole corpus. Empirical evidence of the effectivness of the method, specially on low resourced languages, is provided by achieving 41% (MAP: 0.1130) with a base dictionary of about 10K words, and 60.27% (MAP: 0.1652) with freely available SMT web-services of the monolingual MAP on English-Bengali CLIR (MAP: 0.2741).

The work presented in this paper treats an entire pseudo-relevant document as one document unit in the LDA estimation. A possible extension to this approach, which will be investigated as part of future work, would be to use smaller textual units, i.e. sentences or paragraphs as document units in the LDA estimation. This would naturally take into account the proximity evidence as well, in addition to the topical distribution of terms.

Acknowledgments This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (CNGL) project. The authors express their gratitude to the anonymous reviewers whose feedback helped improve the quality of this paper.

References

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Chen, D., Xiong, Y., Yan, J., Xue, G.-R., Wang, G., and Chen, Z. (2010). Knowledge transfer for cross domain learning to rank. *Inf. Retr.*, 13(3):236–253.

Chinnakotla, M. K., Raman, K., and Bhattacharyya, P. (2010). Multilingual PRF: English lends a helping hand. In *Proceedings of the SIGIR '10*, pages 659–666, New York, USA. ACM.

Dandapat, S., Morrissey, S., Way, A., and van Genabith, J. (2012). Combining EBMT, SMT, TM and IR technologies for quality and scale. In *Proceedings of ESIRMT and (HyTra)*, pages 48–58. ACL.

Fox, C. (1992). *Lexical analysis and stoplists*, pages 102–130. Prentice-Hall.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235.

Hiemstra, D. (2000). *Using Language Models for Information Retrieval*. PhD thesis, Center of Telematics and Information Technology, AE Enschede.

Hull, D. A. and Grefenstette, G. (1996). Querying across languages: A dictionary-based approach to multilingual information retrieval. In *Proceedings of SIGIR*, pages 49–57.

Lavrenko, V., Choquette, M., and Croft, W. B. (2002). Cross-lingual relevance models. In *Proceedings of the SIGIR '02*, pages 175–182, New York, USA. ACM.

Lavrenko, V. and Croft, B. W. (2001). Relevance based language models. In *Proceedings of the SIGIR '01*, pages 120–127. ACM.

Leveling, J., Ganguly, D., and Jones, G. J. F. (2010). DCU@FIRE2010: Term Conflation, Blind Relevance Feedback, and Cross-Language IR with Manual and Automatic Query Translation. In *Working Notes of the FIRE*.

Littman, M., Dumais, S. T., and Landauer, T. K. (1998). Automatic cross-language information retrieval using latent semantic indexing. In *Cross-Language Information Retrieval, chapter 5*, pages 51–62.

Lovins, J. B. (1968). Development of a stemming algorithm. *Mechanical translation and computation*, 11(1-2):22–31.

Majumder, P., Mitra, M., Pal, D., Bandyopadhyay, A., Maiti, S., Pal, S., Modak, D., and Sanyal, S. (2010). The FIRE 2008 evaluation exercise. *ACM Trans. Asian Lang. Inf. Process.*, 9(3).

Majumder, P., Mitra, M., Parui, S. K., Kole, G., Mitra, P., and Datta, K. (2007). YASS: Yet another suffix stripper. *ACM Trans. Inf. Syst.*, 25(4).

Mimno, D. M., Wallach, H. M., Naradowsky, J., Smith, D. A., and McCallum, A. (2009). Polylingual topic models. In *EMNLP*, pages 880–889.

Nie, J.-Y., Simard, M., Isabelle, P., and Durand, R. (1999). Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web. In *Proceedings of SIGIR '99*, pages 74–81.

Paik, J. H., Pal, D., and Parui, S. K. (2011). A novel corpus-based stemming algorithm using co-occurrence statistics. In *Proceedings of the SIGIR '11*, pages 863–872.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the ACL*, pages 311–318, Stroudsburg, PA, USA.

Ponte, J. M. (1998). *A language modeling approach to information retrieval*. PhD thesis, University of Massachusetts.

Robertson, S. E. and Sparck Jones, K. (1988). *Relevance weighting of search terms*, pages 143–160. Taylor Graham Publishing, London, UK.

Salton, G. and Buckley, C. (1990). Improving retrieval performance by relevance feedback. *JASIS*, 41(4):288–297.

Savoy, J. (1999). A stemming procedure and stopword list for general French corpora. *Journal of the American Society for Information Science*, 50(10):944–952.

Udupa, R., Saravanan, K., Bakalov, A., and Bhole, A. (2009). "They Are Out There, If You Know Where to Look": Mining Transliterations of OOV Query Terms for Cross-Language Information Retrieval. In *ECIR*, pages 437–448.

Vulić, I., De Smet, W., and Moens, M.-F. (2011). Cross-language information retrieval with latent topic models trained on a comparable corpus. In *Proceedings of the 7th Asia conference on Information Retrieval Technology*, AIRS'11, pages 37–48, Berlin, Heidelberg. Springer-Verlag.

Wei, X. and Croft, W. B. (2006). LDA-based document models for ad-hoc retrieval. In *SIGIR '06*, pages 178–185, New York, USA. ACM.

Xu, J. and Croft, W. B. (2000). Improving the effectiveness of informational retrieval with Local Context Analysis. *ACM Transactions on information systems*, 18:79–112.