An Expectation-Maximization Algorithm for Query Translation Based on Pseudo-Relevant Documents

Javid Dadashkarimi[a], Azadeh Shakery[a,b,*], Heshaam Faili[a,b], Hamed Zamani[c]

[a] *School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran*
[b] *Institute for Research in Fundamental Sciences (IPM), Tehran, Iran*
[c] *Center for Intelligent Information Retrieval, University of Massachusetts Amherst, MA 01003*

Abstract

Query translation in cross-language information retrieval (CLIR) can be done by employing dictionaries, aligned corpora, or machine translators. Scarcity of aligned corpora for various domains in many language pairs intensifies the importance of dictionary-based CLIR which motivates us to use only a bilingual dictionary and two independent collections in source and target languages for query translation. We exploit pseudo-relevant documents for a given query in the source language and pseudo-relevant documents for a translation of the query in the target language with a proposed expectation-maximization algorithm for improving query translation. The proposed method (called *EM4QT*) assumes that each target term either is translated from the source pseudo-relevant documents or has come from a noisy collection. Since EM4QT does not directly consider term coherency, which is defined as fluency of the target translation, we investigate a crucial question: can EM4QT be improved using either coherency-based methods or token-to-token translation ones? To address this question, we combine different translation models via simple linear interpolation and a proposed divergence minimization method. Evaluations over four CLEF collections in Persian, French, Spanish, and German indicate that EM4QT significantly outperforms competitive baselines in all the collections. Our experiments also reveal that since EM4QT indirectly considers term coherency, combining the method with coherency-based models cannot significantly improve the retrieval performance. On the other hand, investigating the query-by-query results supports the view that EM4QT usually gives a relatively high weight to one translation and its combination with the proposed token-to-token translation model, which is obtained by running EM4QT for each query term separately, soothes the effect and reaches better results for many queries. Comparing the method with a competitive word-embedding baseline reveals superiority of the proposed model.

Keywords: dictionary-based cross-language information retrieval, query translation, expectation maximization, pseudo-relevant documents.
2010 MSC: 00-01, 99-00

[☆] A part of this work was done while Hamed Zamani was with the University of Tehran.

[*] Corresponding author
Email addresses: dadashkarimi@ut.ac.ir (Javid Dadashkarimi), shakery@ut.ac.ir (Azadeh Shakery), hfaili@ut.ac.ir (Heshaam Faili), zamani@cs.umass.edu (Hamed Zamani)

1. Introduction

Exponential growth of the Internet and availability of documents in different languages have turned the World Wide Web into a huge multilingual environment. Retrieval systems are obliged to retrieve documents in a language other than the users' native language since the users intend to find all relevant information available independent of the language. In these circumstances, it is easier for the users to formulate the queries in their native language [29]. Cross-language information retrieval (CLIR) is an approach for bridging the gap between the languages. To this end, several shared-tasks have been also focused on CLIR and related tasks, including the TREC and CLEF shared-tasks. The following techniques are proposed for CLIR: *(1)* translating queries to the target language, *(2)* translating documents to the source language, *(3)* translating queries and documents to a third language [29], *(4)* mapping queries and documents to a shared low-dimensional representation space, and *(5)* using cross-lingual semantic/concept networks [42, 14, 15, 9, 8, 46, 32]. Although it is shown that translating documents can outperform the query translation approach in a few number of languages, document translation is a time-consuming approach and demands re-indexing of the entire collection for each language [3]. That is why query translation is the most common technique for CLIR.

Queries can be translated using machine translation systems or various translation resources, such as dictionaries, comparable corpora, and parallel corpora. It is well known that building parallel corpora is highly expensive in terms of both time and cost. Moreover, current translation extraction methods are not able to purify noisy translations candidates from comparable corpora completely and this is why many language pairs are suffering from lack of these linguistic resources. In addition, these resources are usually domain-specific and employing them in domains other than the domain of the corpus can lead to low performance [29]. Furthermore, extracting reliable translation knowledge from comparable corpora heavily depends on the size of the collection in terms of the number of alignments [38]. On the other hand, bilingual machine readable dictionaries are known as available resources with high translation coverage in many language pairs for general domains [6, 13, 46]. All these facts intensify the importance of studying dictionary-based CLIR.

Dictionaries provide an unweighted list of target terms for each term of the source language. There is an important challenge in the dictionary-based CLIR: ambiguity in translation and swamping effect as a result[1] [13]. Indeed, in most cases each term in the source language has more than one translation candidate in the target language and thus detecting the correct translation for each term could be a big issue here. Several methods have so far been proposed to address these problems, such as structured query [30, 31], iterative translation disambiguation (ITD) [28], and maximum coherence model [23]. In structured queries all the translations of a word are dealt as members of a synonym set and the number of occurrences of the source word equals the sum of the number of the occurrences of the members. But the probabilistic approaches score documents based on a number of translation probabilities [28, 23, 6, 10]. Many of these methods aim at disambiguating the query using global mutual information of the translations in the target language. In this paper, our contribution is to use a couple of local in-the-context collections, one from the source language and the other from the target language, to compute a query-dependent translation model for each query. Pseudo-relevant documents in response to the query in both source and target languages comprise these collections.

Pseudo-relevant documents are a number of top-ranked documents in response to the query of the user and are expected to be relevant to the query and thus can potentially be suitable resources for extracting translation knowledge. The proposed method runs an on-line disambiguation process by

[1] retrieving irrelevant documents which is caused by translating the query to non-relevant candidates

incorporating two collections, one in the source language, and another in the target language. The method retrieves the source documents for the initial query and then, translates the query using a simple translation technique (e.g., uniform weighting of translations); at the next step target documents are retrieved for the translated query. We expect tha distribution of the context terms in the source collection to be similar to the distribution of their translations in the target collection, accepting a small amount of noise from the background collection. In more details, it is expected that each word in the target pseudo-relevant collection either is translated from the source pseudo-relevant collection or has come from a noisy background collection. Based on this expectation, we propose an expectation-maximization (EM) algorithm, an iterative hill-climbing algorithm, to extract a query-dependent translation knowledge for each query. This method is called Expectation-Maximization for query translation (EM4QT). We prove that the proposed method converges to a global optimum solution (see Appendix B).

Although the methods based on term coherency perform promising in dictionary-based CLIR, EM4QT does not directly consider the coherency between target terms. Therefore, in this paper we also investigate a crucial research question: can the performance of EM4QT be improved using the off-line coherency-based CLIR methods? To answer this question, we consider the simple linear interpolation method and also propose a statistical divergence minimization method to combine more than one translation model.

Since the extracted translation model from the proposed EM4QT method usually drifts to the translations which are more coherent in the pseudo-relevant documents and more discriminative through the collection, we investigate another research question: can the performance of EM4QT be improved using a token-to-token translation model? To this aim, we employ a token-to-token translation technique in which each term of the query is posed into the EM-based system individually and the obtained model is combined with that of EM4QT.

We evaluate our proposed method on four standard CLEF collections in four different languages: Spanish (CLEF-2002), German (CLEF 2002-03), French (CLEF 2002-03), and Persian (CLEF 2008-09). The proposed EM algorithm significantly outperforms competitive baselines in all the collections. It is also shown that combining the proposed method with iterative disambiguation method, a state-of-the-art coherency-based translation disambiguation method, cannot significantly improve the retrieval accuracy. This might be due to the fact that the proposed EM4QT method can indirectly take advantage of term coherency, since we use top-retrieved documents to improve the query translation. On the other hand, investigating the query-by-query results on the datasets reveals that EM4QT usually converges to one possible correct translation. So its combination with a proposed token-to-token translation model in which the proposed EM algorithm runs over terms individually soothes the effect and reaches better recall for many queries. Experimental results show that the proposed CLIR systems reach 80.34% monolingual MAP in the Persian collection, 77.67% monolingual MAP in the French collection, 79.06% monolingual MAP in the Spanish collection, and 75.54% monolingual MAP in the German collection.

In the rest of this paper, we review the previous works related to our research in Section 2. We propose our method in Section 3 and evaluate it in Section 4. Finally, we conclude our paper and provide some future works in Section 5.

2. Related work

Query/Document translation is a key step in the CLIR task. Translation knowledge can be extracted from various linguistic resources, such as parallel corpora [45, 1, 49] and comparable corpora

[34, 33]. Beside these methods, there are a number of studies [11, 28, 22, 6] whose aim is to exploit bilingual machine readable dictionaries as commonly available bilingual resources with high coverage, which are in focus of this paper. Ambiguity in translation is a pivotal challenge in dictionary-based CLIR. There have been multiple attempts to address the ambiguity problem in CLIR. Some methods resolve the ambiguity of a translation by selection-based approaches. These methods choose highly ranked translations of the query terms ([6, 1]) and leave the lower ranked ones out. Some methods define metrics for the ambiguity and select the least ambiguous translations [13]. Some other methods consider all the translations in their computations and aim at disambiguation by generating a number of weights for the translation candidates [23]. Pirkola et al. exploited structured queries which takes all translations of a query term into consideration as members of a synonym set [30]. Iterative translation disambiguation [28] is a method based on Page Rank algorithm which employs the converged translation weights within a vector space-based retrieval framework. Maximum coherence model [23] is also proposed to overcome a similar issue. To the best of our knowledge, there is no comparison between these methods but, since both approaches compute cohesion based on mutual information of translations, they are likely to reach a similar local optimum. However, query drift is a drawback of maximum coherence model for short queries [29].

2.1. Cross-language pseudo-relevance feedback

There have been some efforts to answer the users' queries either in single aggregated multi-lingual ranked lists [35, 37], in multiple in-the-target-language ones, or in single in-the-target ranked-list according to the users' languages in the CLIR task [43, 42, 41, 10]. Numerous methods are proposed in the subject to improve the quality of the translation process; some methods focus on a global approach in which the coherency of translations are considered through a collection, and some others do the translation locally based solely on the context of the query. Top-retrieved documents in response to the query are shown to be an informative local collection which can present the context of the query more accurately[4, 52]. Pseudo-relevance feedback (PRF) methods aim to update the user-specified query using these collections and improve the retrieval performance [4, 7, 20, 27, 44, 47, 48, 52]. There are a number of approaches for pseudo-relevance feedback in the CLIR task. Lee and Croft exploited feedback documents from both source and target languages using an aligned corpus of informal texts [21]. They translate source queries and target documents by a machine translator and then employ inter- and intra-language PRF to improve the retrieval performance. In their proposed method some bilingual features and some linguistic features (nouns, verbs, and named entities) have been exploited to train a support vector-based binary classifier. In addition to their dependency to aligned corpora, document translation and model training are very time-consuming tasks in their method.

Lavrenko presented a unified framework entitled cross-lingual relevance model (CLRM) for an English-Chinese CLIR task [19]. Ganguly et al. introduced a topical relevance model for this aim using top-ranked documents in source and target languages [10]. The proposed cross-lingual topical relevance language model (CLTRLM) exploits top-ranked documents in both source and target languages in a topic modeling framework for extracting word-topic and topic-document distributions based on latent Dirichlet allocation (LDA) [2]. The model assumes that each translation is generated either from a bilingual dictionary or from a number of relevant topics. The authors claim that this approach compensates limitations of the dictionaries in terms of coverage.

2.1.1. Cross-Language Topical Relevance Models

Topic modeling refers to study of text mining techniques for representing textual units in a low-

dimensional topical subspace [2]. When it comes to multilingual corpora, such dimensions can be interpreted as a number of multi-lingual topics in which each topic represents a fuzzy set of words frequently co-occurring with each other in a collection [43].

Ganguly et al. introduced a method entitled cross-lingual topical relevance language model (CLTRLM) for extracting multilingual topics from a pseudo-relevant corpus [10]. The authors exploit the model in a query translation task. CLTRLM considers the translation of a query as a result of a couple of generative models, one from a bilingual dictionary and the other from the topic model obtained from the target pseudo-comparable documents. Indeed, instead of document alignment, topics are aligned with each other. So a translation comes either form dictionary or from the equivalent topic of the source word in the target language.

Equation 2.1 shows the model as a probabilistic model between a translation word in the target language and the source query posed by the user [10]:

$$p(w^t|\mathbf{q}^s) = \sum_{k=1}^{K^t} p(w^t|z_k^t, \phi_{k,w^t}^t) \sum_{j=1}^{R^t} p(z_k^t|D_j^t, \theta_{j,k}^t) \sum_{i'=1}^{n^t} \frac{p(q_{i'}^t|D_j^t)}{R^t} \sum_{j'=1}^{n^s} p(q_{i'}^t|q_{j'}^s)$$
$$+ \sum_{i=1}^{\mathcal{T}\{w^t\}} p(w^t|w_i^s) \sum_{k=1}^{K^s} p(w_i^s|z_k^s, \phi_{k,w^s}^s) \sum_{j=1}^{R^s} p(z_k^s|D_j^s, \theta_{j,k}^s) \frac{p(\mathbf{q}^s|D_j^s)}{R^s} \quad (1)$$

In Equation 2.1, ϕ and θ are word-topic and document-topic distributions computed by LDA respectively and z denotes the topic labels. R is the number of relevant documents in the collection and K is the number of topics. s and t are source and target language indicators respectively. Both $p(q_{i'}^t|D_j^t)$ and $p(\mathbf{q}^s|D_j^s)$ can be computed based on monolingual document language models. $p(w^t|w_i^s)$ and $p(q_{i'}^t|q_{j'}^s)$ can also be computed using a simple translation model.

Ganguly et al. introduced an advanced version of CLTRLM entitled joint cross-language topical relevance model (JCLTRLM) tailoring bilingual LDA (BiLDA) on the pseudo-comparable documents. In JCLTRLM we have $\theta = \theta^t = \theta^s$, $R = R^t = R^s$, and $K = K^t = K^s$.

2.1.2. Cross-language dimension projection between languages

In this section, we briefly introduce a method recently proposed in [5] for finding equivalents of low-dimensional vectors learnt over a source collection in the target language. The obtained low-dimensional vectors can be used to re-weight translation candidates of each source term.

Cross-lingual word embedding translation model (CLWETM) tailors an off-line approach for learning bilingual term representations by exploiting pseudo-relevant documents in both source and target languages. To this end, first it learns word representations of the pseudo-relevant collections separately and then focuses on finding a transformation matrix minimizing a distance function between all translation pairs appeared in the collections. As shown in Equation 2, the goal is to minimize a cost function with respect to a transformation matrix $\mathbf{W} \in \mathbb{R}^{n \times n}$; the cost function f is defined as follow:

$$f(\mathbf{W}) = \sum_{(w^s, w^t)} \frac{1}{2} ||\mathbf{W}^T \mathbf{u}_{w^s} - \mathbf{v}_{w^t}||^2 \quad (2)$$

where $(w^s, w^t) \in (F^s, F^t)$ is a translation pair; $\mathbf{u}_{w^s} \in \mathbb{R}^{n \times 1}$ and $\mathbf{u}_{w^t} \in \mathbb{R}^{n \times 1}$ are their vectors respectively learnt on the source and the target pseudo-relevant collections . This problem can be solved using stochastic gradient descent as follows (i.e., $\frac{\partial f}{\partial \mathbf{W}} = 0$):

$$\mathbf{W}^{t+1} = \mathbf{W}^{t} - \eta(\mathbf{W}^{T}\mathbf{u}_{w^s} - \mathbf{v}_{w^t})\mathbf{u}_{w^s}^{T} \tag{3}$$

where η is a constant learning rate. $\mathbf{W}$ can be initialized randomly and then be updated incrementally.

Finally, the new vectors can be calculated as follows:

$$\hat{\mathbf{u}}_{w^s} = \mathbf{W}^{T}\mathbf{u}_{w^s} \tag{4}$$

where $\mathbf{W}^{T}\mathbf{u}_{w^s}$ transforms the source vector to the target low-dimensional space. The new translation model is built as follows:

$$p(w_t|w_s) = \frac{\exp\left(\frac{\hat{\mathbf{u}}_{w^s}.\mathbf{v}_{w^t}}{||\hat{\mathbf{u}}_{w^s}||\ ||\mathbf{v}_{w^t}||}\right)}{\sum\limits_{\bar{w}^t \in \mathcal{T}\{w^s\}} \exp\left(\frac{\hat{\mathbf{u}}_{w^s}.\mathbf{v}_{\bar{w}^t}}{||\hat{\mathbf{u}}_{w^s}||\ ||\mathbf{v}_{\bar{w}^t}||}\right)} \tag{5}$$

where $\mathcal{T}\{w^s\}$ is the list of translation candidates for w^s. The obtained translation model can be interpolated with another translation model for achieving a model which considers both local information (query dependent) and global information (collection dependent).

On one hand, CLTRLM uses a statistical approach with a mathematical framework to achieve the query-dependant translation model; on the other hand, CLWETM tailors a learning technique to capture both intra-language information and inter-language relations in the final model. In this work, we propose a novel theoretical technique for building a translation model of the query which also works well empirically. A fundamental difference of the proposed method with CLTRLM and CLWETM is in the configuration of hidden variables. The proposed method uses a pair-level configuration for the variables rather than a document-level one in CLTRLM and a language-level one in CLWETM. We believe that although the input collections might work well for some pairs of translations, this is not the case for some others. The former pairs should be rewarded and the latter ones should be punished. This reward/punish policy should be handled by a learning algorithm which is the focus of Section 3.

3. Methodology

Ambiguity is the main issue in dictionary-based CLIR since, in most cases each term in the source language has more than one translation candidate in the target language and thus weighting the correct translations for each term could be a big issue here. To address this issue, we propose a novel method which uses pseudo-relevant documents to extract translation knowledge. In fact, we consider two collections, one in the source language, and another one in the target language. Simultaneously, we retrieve documents of both source and target collections for a given query. To retrieve the documents in the target language, we use a simple translation method, such as giving equal weight to all translation candidates of each query term. Finally, by accepting a limited amount of noise from both collections, we can expect the term occurrence distribution in top-retrieved documents of the source language to be similar to their translations' occurrence distribution in top-retrieved documents of the target language collection. We propose an expectation-maximization (EM) algorithm to extract translation probabilities that minimize divergence of the distributions. Figure 1 shows an outline of the proposed framework. In the first step the source query q is imposed to retrieve a number of pseudo-relevant documents in the source language. In the next step, the query is translated to a target language by a simple translation model. The translated query is then incorporated to retrieve a number of pseudo-relevant documents in the target language. In the next step both the pseudo-relevant documents are incorporated in the proposed EM4QT algorithm to extract a query-dependent translation

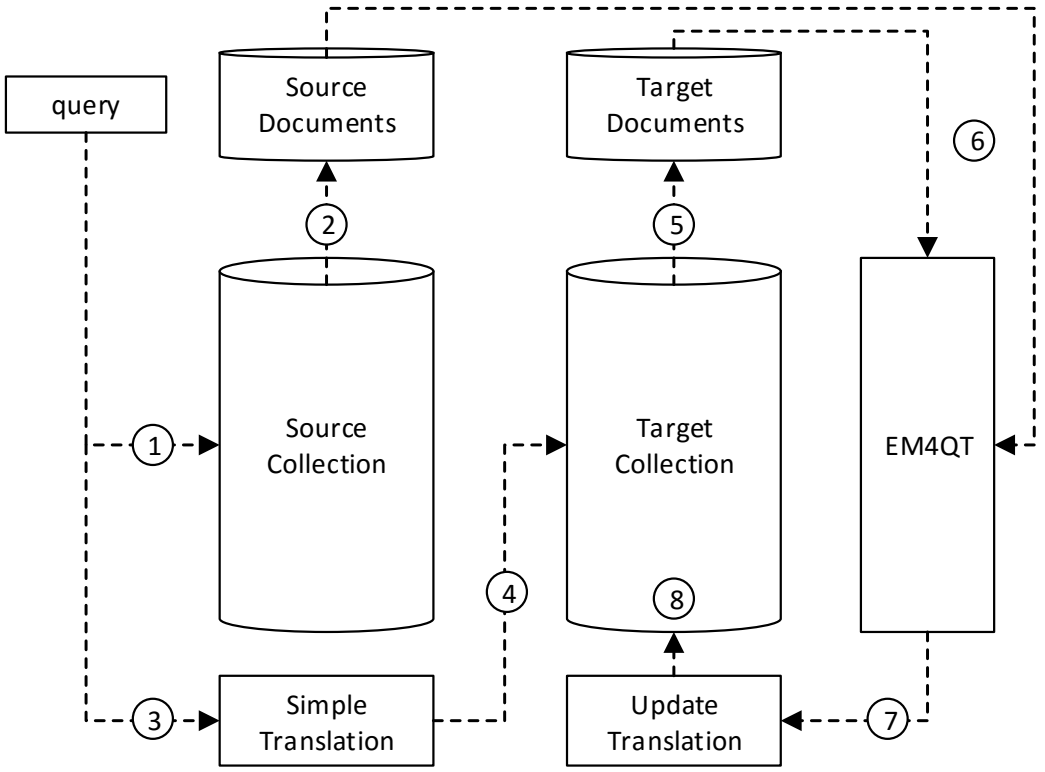


Figure 1: EM4QT framework.

model. Finally, q is updated by the model and then is used to retrieve the target documents again (Fig. 1).

Although term coherency between translations has been shown to be highly effective for dictionary-based CLIR, the proposed method does not consider term coherency directly. To also consider the coherence of terms in the target language, we propose to combine different translation resources by using linear interpolation or by minimizing the divergence between the given distributions (translation models). Furthermore, since the proposed EM4QT method usually drifts to the translations which are more coherent in the pseudo-relevant documents and more discriminative through the collection, we combine the model with an EM-based token-to-token translation model to soothe the effect and to achieve a higher recall. In the rest of this section, we first propose an EM algorithm to extract translation knowledge from top-retrieved documents in both languages. Then, we introduce two methods for combining different translation models.

3.1. Improving query translation quality using pseudo-relevant documents

Our purpose is to extract translation knowledge based on term occurrences in top-retrieved documents in both languages. To this aim, consider F and F' as the sets of top-retrieved documents for a given query q and its translation q', respectively. Note that q' is translated using a simple existing translation method. Our main idea is based on the fact that distributions of topical terms in F and F' are similar and the goal is to predict the value of $\mathcal{T}(w_t|w_s)$, the translation probability of the term w_s in the source language to the term w_t in the target language. To compute this probability, we first

define a hidden variable T where $p(T = 1|w_s, w_t)$ indicates the confidence that the term w_s in F is translated to term w_t in F'. Indeed given a source word in F and a target word in F', T controls the degree of comparability of F with respect to F'. Considering the Bayes rule we have:

$$p(T = 1|w_s, w_t) = \frac{p(w_t|T = 1, w_s)p(T = 1|w_s)}{\mathcal{T}(w_t|w_s)} \tag{6}$$

Based on the law of total probability we can calculate $\mathcal{T}(w_t|w_s)$ as:

$$\mathcal{T}(w_t|w_s) = p(w_t|T = 1, w_s)p(T = 1|w_s) + p(w_t|T = 0, w_s)p(T = 0|w_s) \tag{7}$$

Note that $p(T = 1|w_s)$ can also be re-written by:

$$p(T = 1|w_s) = \frac{p(w_s|T = 1)p(T = 1)}{p(w_s)} \tag{8}$$

Since $p(T = 1)$ is independent of the source and target terms, we can assume that it is equal to a constant parameter λ. $p(w_s)$ can be estimated by $p(w_s|\mathcal{C}_s)$ and $p(w_s|T = 1)$ can be estimated by $p(w_s|\theta_F)$. θ_F denotes the unigram language model of F and $\mathcal{C}_s$ is the collection language model of the source language. $p(w_t|T = 0, w_s)$ could be interpreted as the probability of w_t to be a translation of w_s if F and F' are not comparable to each other. Therefor, we can estimate $p(w_t|T = 0, w_s)$ as $p(w_t|\mathcal{C}_t)$ where $\mathcal{C}_t$ denotes the collection language model of the target language. Hence, based on Equation 6, we define the following iterative EM algorithm:

$$\begin{aligned} p(w_s) &= p(w_s|T = 1)\lambda + p(w_s|T = 0)(1 - \lambda) \\ &= p(w_s|\theta_F)\lambda + \frac{p(T = 0|w_s)p(w_s)(1 - \lambda)}{p(T = 0)} \\ &= p(w_s|\theta_F)\lambda + p(T = 0|w_s)p(w_s) \end{aligned} \tag{9}$$

So we have:

$$p(T = 0|w_s) = (1 - \frac{p(w_s|\theta_F)}{p(w_s|\mathcal{C})}\lambda) \tag{10}$$

Regarding to Equation 8 and Equation 10 we have the following inter-dependent equations:

$$p^{(i)}(T = 1|w_s, w_t) = \frac{\lambda p^{(i)}(w_t|T = 1, w_s)}{(1 - \frac{p(w_s|\theta_F)}{p(w_s|\mathcal{C}_s)}\lambda)p(w_t|\mathcal{C}_t) + \lambda p^{(i)}(w_t|T = 1, w_s)} \tag{11}$$

$$p^{(i+1)}(w_t|T = 1, w_s) = \frac{p(w_t|\theta_{F'})p^{(i)}(T = 1|w_s, w_t)}{\sum_{w'_t \in \mathcal{T}\{w_s\}} p(w'_t|\theta_{F'})p^{(i)}(T = 1|w_s, w'_t)} \tag{12}$$

where i and $\mathcal{T}\{w_s\}$ denote the iteration number and the set of translation terms of w_s in the dictionary, respectively. In the E-Step (Equation 11), we estimate the translation confidence probabilities and in the M-Step (Equation 12), we maximize the likelihood of the probabilities.

At the end, the translation distribution $\mathcal{T}$ can be calculated using Equation 7 in which $p(w_t|T = 1, w_s)$ is calculated in the mentioned EM algorithm. Note that convergence of the proposed EM algorithm to the global optimal solution is proven in Appendix B. A pseudo code of the EM4QT algorithm is shown in Algorithm 1.

Algorithm 1 Pseudo-code for the proposed EM4QT framework.

procedure EM4QT($q^s, \mathcal{T}, \mathcal{C}_s, \mathcal{C}_t$)
  $B \in \mathbb{R}^{1000\times|\{0,1\}|\times|\mathbb{V}_s|\times|\mathbb{V}_t|}$ ▷ $p^{(i)}(w_t|T \in \{1,0\}, w_s)$
  $q^t \leftarrow$ translate($q^s, \mathcal{T}$) ▷ uniform weighting of translations
  $F^s \leftarrow$ TopKRet($q^s, \mathcal{C}_s$) ▷ pseudo relevant source documents
  $F^t \leftarrow$ TopKRet($q^t, \mathcal{C}_t$) ▷ pseudo relevant target documents
  $\theta_F^s \leftarrow p(w_s|F^s)\ \forall\ w_s \in \mathbb{V}_s$ ▷ maximum likelihood estimation
  $\theta_F^t \leftarrow p(w_t|F^t)\ \forall\ w_t \in \mathbb{V}_t$
  $\epsilon \leftarrow 0.01$
  $\mathcal{E} \leftarrow 1.0$
  do
    $i \leftarrow 0$
    $\mathcal{E} \leftarrow 0.0$
    for w_s **in** q^s **do**
      $\alpha_{w_s} \leftarrow (1 - \frac{p(w_s|\theta_F)}{p(w_s|\mathcal{C})}\lambda)$ ▷ $\alpha_{w_s} \leftarrow p(T=0|w_s)$
      for w_t **in** $\mathcal{T}(w_s)$ **do**
        $B[0, T=1, w_s, w_t] \leftarrow \frac{1}{|\mathcal{T}\{w_s\}|}$ ▷ or initialize randomly
        $z_i \leftarrow B[i, T=1, w_s, w_t]$
        $\gamma_i(w_t) \leftarrow \frac{\lambda z_i}{\alpha_{w_s} p(w_t|\mathcal{C}_t) + \lambda z_i}$
        $B[i+1, T=1, w_s, w_t] \leftarrow \frac{p(w_t|\theta_{F'})\gamma_i(w_t)}{\sum_{w'_t \in \mathcal{T}\{w_s\}} p(w'_t|\theta_{F'})\gamma_i(w'_t)}$
        $\mathcal{E} \leftarrow \mathcal{E} + ||B[i+1, T=1, w_s, w_t] - B[i, T=1, w_s, w_t]||^2$
      end for
    end for
    $i \leftarrow i + 1$
  while $\mathcal{E} \geq \epsilon \mid i \geq 1000$
end procedure

3.1.1. Time complexity analysis

Time complexity of JCLTRLM is $\mathcal{O}((|V^t|+|V^s|)(R^t+R^s)(K^t+K^s)N)$ where V^t and V^s are the vocabulary sets of the source and the target documents respectively and N is the number of iterations of the algorithm. Whereas, time complexity of EM4QT is $\mathcal{O}((|\mathbf{q}^s| + |\mathbf{q}^t|)N)$.

3.2. Combining different translation models

In this section, we propose two translation combination methods to investigate whether these combinations can improve the performance or not. A simple method for combining two probabilistic distributions $\mathcal{T}_1$ and $\mathcal{T}_2$ is linear interpolation:

$$\mathcal{T} = (1-\alpha)\mathcal{T}_1 + \alpha\mathcal{T}_2 \tag{13}$$

where α controls the effect of each translation model. This simple method has been previously employed by [40] to combine different translation models.

As another combination method, we propose to estimate the translation model $\mathcal{T}$ by minimizing the divergence between the given translation models $\mathcal{T}_1$ and $\mathcal{T}_2$. Formally writing, the goal is to find $\mathcal{T}$ with the goal of minimizing the following objective function:

$$\arg\min_{\mathcal{T}} D(\mathcal{T}||\mathcal{T}_1) + \alpha D(\mathcal{T}||\mathcal{T}_2) \tag{14}$$

Table 1: Collections characteristics

ID	Lang.	Collection	Queries	#docs	#qrels
SP	Spanish	EFE 1994	CLEF 2002, topics 91-140	215,738	2,854
DE	German	Frankfurter Rundschau 94, SDA 94, Der Spiegel 94-95	CLEF 2002-03, topics 91-140	225,371	1,938
FR	French	Le Monde 94, SDA French 94-95	CLEF 2002-03, topics 251-350	129,806	3,524
FA	Persian	Hamshahri 1996-2002	CLEF 2008-09, topics 551-650	166,774	9,625

where parameter α controls the importance of each translation model: increasing the value of α means to pay more attention to the translation model $\mathcal{T}_2$. $D(\cdot||\cdot)$ denotes the divergence between two probabilistic distributions. We use the KL-divergence[2] formula [17] to compute this divergence. KL-divergence between two probabilistic distributions P and Q is calculated as:

$$D(P||Q) = \sum_i p(i|P) \log \frac{p(i|P)}{p(i|Q)} \tag{15}$$

Since $\mathcal{T}$ is a probability distribution, there is an obvious constraint:

$$\sum_{w_t \in \mathcal{T}\{w_s\}} \mathcal{T}(w_t|w_s) = 1. \tag{16}$$

Considering this constraint and the Lagrange multiplier method [39], we can find the optimum translation model $\mathcal{T}$ as follows:

$$\mathcal{T}(w_t|w_s) \propto \exp\left(\frac{1}{(1+\alpha)} \log \mathcal{T}_1(w_t|w_s) + \frac{\alpha}{(1+\alpha)} \log \mathcal{T}_2(w_t|w_s)\right) \tag{17}$$

Note that both linear interpolation and divergence minimization methods can be easily extended for combining more than two probabilistic distributions.

At the final step the query language model is constructed according to Equation 18 [36]. In this equation $q^s = q_1^s, q_2^s, .., q_n^s$ is the user's query in the source language. The probability of the target translation w_t w.r.t query can be computed as follows:

$$p(w^t|\theta_q) = \sum_{i=1}^{n} \frac{1}{n} \frac{\mathcal{T}(w^t|q_i^s)}{\sum_{j=1}^{k} \mathcal{T}(w_j^t|q_i^s)} \tag{18}$$

4. Experiments

4.1. Datasets

We use four standard CLEF CLIR collections in four different languages: Spanish, German, French, and Persian. The queries in all the collections are in English. The statistics of these collections and translation resources are reported in Table 1 and Table 2 respectively. As pointed out in Section 3, the proposed method needs a collection of documents in the source language for extracting the translation knowledge. Note that there is no need to have relevance judgments, since we use

[2]Kullback-Leibler divergence

Table 2: Dictionary Characteristics; $\#\mathcal{T}\{w\}$ is the expected number of translations for w.

Language	Resource	$\#\mathcal{T}\{w^s\}$	$\#\mathcal{T}\{w^t\}$
English-Persian	Google	2.77 ± 15.04	4.18 ± 5.50
English-French	Google	2.50 ± 13.34	3.08 ± 3.56
English-Spanish	Google	2.74 ± 16.60	2.7 ± 2.84
English-German	Google	3.75 ± 20.37	2.56 ± 2.80

pseudo-relevant documents in our EM algorithm. In our experiments, we use a pool of Associated Press 1988-89, Los Angeles Times 1994, and Glasgow Herald 1995 collections for English (source language) documents. These collections are used in previous TREC or CLEF evaluation campaigns for ad-hoc retrieval and are in domain of news same as the target retrieval collections.

4.2. Experimental setting

In all experiments, we use the language modeling framework with the KL-divergence retrieval model [18]. All the documents are smoothed using the Dirichlet prior smoothing method which has been shown to be highly effective in information retrieval [53]. The Dirichlet prior smoothing parameter μ is set to the typical value of 1000. To improve the retrieval performance, we use the mixture model [52] for pseudo-relevance feedback with the feedback coefficient of 0.5. The number of feedback documents and feedback terms are set to the typical values of 10 and 50, respectively. As mentioned in [5], We used stochastic gradient descent for learning $\mathbf{W}$ that is initialized randomly from $[-1, 1]$; η is set to a small value. $\mathbf{u}_{w^s}$ and $\mathbf{v}_{w^t}$ are obtained from negative sampling skip-gram introduced in [26]; the size of the window, the number of negative samples, and the lengths of the vectors are set to typical values of 10, 45, and 50, respectively [26, 25, 5].

All European dictionaries, documents, and queries are normalized and stemmed using the Porter stemmer. Documents and queries in Persian are not stemmed and remained intact due to the low performance of the Persian stemmers in IR [16, 6]. Stopwords are removed in all the experiments.[3].The Lemur toolkit[4] is employed as the retrieval engine in our experiments. All the source codes belonging to EM4QT and JCLTRLM are freely available at GitHub [5].

We use the Google dictionaries in our experiments.[6] In the European languages, we do not transliterate out of vocabulary (OOV) terms of the source languages. The target language OOVs are used as their original forms in in the source documents, since they are cognate languages; but, Persian has a different alphabet, we transliterate the OOVs. Note that we use uniform distribution as the initial translation model for retrieving top documents (see Section 3.1).

Similar to previous work [12, 27], the parameters λ, α, and n (number of top-retrieved documents that have been used in the proposed EM algorithm) are set via 2-fold cross-validation over the queries of each collection.

[3]We use the stopword lists and the normalizing techniques available at `http://www.unine.ch/info/clef/`.

[4]`http://www.lemurproject.org/`

[5]`https://github.com/javiddadashkarimi/translation`

[6]The Google dictionaries provide lists of translations and employ bilingual dictionaries for this purpose. So, they are different from the Google machine translation system.

4.3. Evaluation metrics

To evaluate retrieval effectiveness, we use mean average precision (MAP) of the 1000 top-ranked documents as the main evaluation metric. In addition, we report the precision of top 5 and top 10 retrieved documents (P@5 and P@10). Statistically significant differences of the performances are determined using two-tailed paired t-test computed at a 95% confidence level based on average precision per query.

4.4. Experimental results and discussion

In this section, we first compare the proposed method with competitive baselines and explore the effectiveness of combining the proposed model with a coherency-based translation disambiguation method. We further investigate sensitivity of the proposed method to the input free parameters.

4.4.1. Effectiveness analysis

In the first set of experiments, we consider the following dictionary-based CLIR methods to evaluate the proposed EM method:[7] *(1)* the top-1 translation of each term in the bilingual dictionaries (TOP-1) which is the most common and the most correct one based on the expert's view [6, 1], *(2)* all the possible translations of each term with equal weights (ALL), and *(3)* all the possible translations of each term with the collection frequency weighting[8] (COLL). In addition we use the mono-lingual retrieval results as a term of comparison (Mono). Note that we will consider a state-of-the-art coherency based method in the next set of experiments. As another term of comparison, we consider the proposed EM algorithm with term independence assumption. This method is called *EM4TT*[9]. This method assumes that query terms are independent and thus it retrieves each query term separately and run the proposed EM algorithm for the retrieved documents. The comparison of EM4QT and EM4TT will give us an insight into the properties of the proposed on-line query-based EM algorithm and its off-line token-to-token version.

The results obtained by the aforementioned methods are reported in Table 3. According to the results reported in Table 3, except for P@5 in the French collection, TOP-1 outperforms ALL in all collections in terms of all the reported evaluation metrics. The reason could be that we use the top-1 translation of the Google dictionaries in our experiments and this translation usually is the most common translation for each source term. Similar results were also achieved in the previous research studies [13, 6]. Furthermore, in most cases, COLL performs better than the other two baselines; since, it considers all the possible translations and weights the translations based on their commonness in the collection. Sometimes these kinds of considerations produce better performances for some types of queries.

According to Table 3, EM4TT outperforms all the baselines in terms of MAP. In a few experiments, these improvements are significant. Interestingly, EM4QT outperforms all the baselines and also EM4TT in terms of MAP and precision at top-retrieved documents. The MAP improvements are always statistically significant, which intensifies the effectiveness of the proposed method. Comparing the results achieved by EM4TT and EM4QT demonstrates the effectiveness of query-dependent

[7] To avoid apples-to-oranges comparisons, we do not consider the methods that use aligned corpora, those that are developed for a retrieval model other than the language modeling framework (such as structured queries), and the learning-based methods.

[8] This method is able to give higher weight to the more common translation.

[9] Expectation-Maximization for Term Translation

Table 3: Comparison of the proposed EM method with dictionary-based CLIR baselines that do not consider term coherency. Superscripts 1/2/3/4 indicate that the MAP improvement over the indicated baseline is statistically significant.

Language	ID	Method	MAP	%Mono	P@5	P@10
Persian (FA)		Mono	0.3659	-	0.5880	0.5620
	1	TOP-1	0.2135	58.34	0.3480	0.3460
	2	ALL	0.1977	54.03	0.3240	0.3120
	3	COLL	0.2372	64.82	0.3860	0.3850
	4	EM4TT	0.2643^{12}	72.23	0.4240	0.4040
	5	EM4QT	$\mathbf{0.2850}^{1234}$	**77.89**	**0.4520**	**0.4490**
French (FR)	1	TOP-1	0.2854	70.12	0.3515	0.3121
	2	ALL	0.2708	66.54	0.3596	0.3091
	3	COLL	0.2971	72.10	0.3980	0.3455
	4	EM4TT	0.2885^{2}	70.88	0.3717	0.3222
	5	EM4QT	$\mathbf{0.3123}^{1234}$	**76.73**	**0.4182**	**0.3677**
Spanish (SP)		Mono	0.5067	-	0.6680	0.5980
	1	TOP-1	0.3655	72.13	0.4440	0.4220
	2	ALL	0.3280	64.73	0.3720	0.3400
	3	COLL	0.3719	73.39	0.4520	0.4180
	4	EM4TT	0.3752^{2}	74.04	0.4360	0.4220
	5	EM4QT	$\mathbf{0.3980}^{1234}$	**78.54**	**0.4880**	**0.4480**
German (DE)		Mono	0.3912	-	0.5240	0.4840
	1	TOP-1	0.2661	68.02	0.3560	0.3040
	2	ALL	0.2589	66.18	0.3080	0.2720
	3	COLL	0.2613	66.79	0.3280	0.2760
	4	EM4TT	0.2704	69.12	0.3240	0.2900
	5	EM4QT	$\mathbf{0.2955}^{1234}$	**75.54**	**0.3800**	**0.3400**

translation. The reason is that queries usually contain phrases or collocations and thus their translations depend on the whole query. It should be noted that the translation knowledge extracted by EM4TT can be done off-line and thus EM4TT is more efficient than EM4QT.

Combining translation models: In the next set of experiments, we investigate the effectiveness of combining the results of the proposed query dependent EM method with (1) the iterative translation disambiguation (ITD) results [28], a state-of-the-art coherency-based translation disambiguation method, (2) JCLTRLM introduced in Section 2.1.1, (3), cross-lingual word-embedding translation model (CLWETM) proposed in Section 2.1.2, and (4) the proposed off-line EM4TT method. The results are reported in Table 4. With regards to this table, the EM method always outperforms ITD. It shows that using pseudo-relevant documents in both source and target languages to extract translation knowledge is an effective idea for CLIR. Although EM performs better than ITD, since they use completely different assumptions for translation disambiguation, their combination may improve the retrieval accuracy.

Interpolating with ITD: According to Table 4, the divergence minimized model between EM4QT and ITD (DIVMIN-I) outperforms the linear interpolation one (LINEAR-I) in three collections (FR, SP, and DE). The performance differences between LINEAR-I/DIVMIN-I and EM4QT are not sta-

Table 4: Combining the translation model obtained by the proposed query dependent EM method with ITD.

Language	ID	Method	MAP	%Mono	P@5	P@10
Persian (FA)		Mono	0.3659	-	0.588	0.562
	1	JCLTRLM	0.2288	62.53	0.3680	0.3690
	2	JCLTRLM†	0.2523	68.95	0.400	0.391
	3	CLWETM	0.2833	77.4	0.47	0.442
	4	ITD	0.2547	69.61	0.406	0.406
	5	EM4TT	0.2643	72.23	0.424	0.404
	6	EM4QT	0.285^{1245}	77.89	0.452	0.449
	7	LINEAR I	0.2867^{1245}	78.35	0.4400	0.454
	8	DIVMIN I	0.2801^{125}	76.55	0.456	0.448
	9	LINEAR II	$\mathbf{0.294}^{1245}$	**80.34**	0.4760	**0.4680**
	10	DIVMIN II	0.2882^{12345}	78.76	**0.4800**	0.4600
French (FR)		Mono	0.407	-	0.5253	0.4697
	1	JCLTRLM	0.1698	41.72	0.2242	0.2131
	2	JCLTRLM†	0.2266	55.67	0.3414	0.299
	3	CLWETM	**0.3186**	78.3	0.4162	0.3667
	4	ITD	0.2763^{12}	67.89	0.3657	0.3333
	5	EM4TT	0.2885^{12}	70.88	0.3717	0.3222
	6	EM4QT	0.3123^{1245}	76.73	**0.4182**	**0.3677**
	7	LINEAR I	0.3031^{124}	74.47	0.4000	0.3525
	8	DIVMIN I	0.3161^{1245}	**77.67**	0.4101	0.3616
	9	LINEAR II	0.3154^{1245}	77.49	0.4020	0.3556
	10	DIVMIN II	0.3096^{1235}	76.07	0.4020	0.3495
Spanish (SP)		Mono	0.5067	-	0.668	0.598
	1	JCLTRLM	0.2210	43.62	0.3200	0.3100
	2	JCLTRLM†	0.2734	53.96	0.4040	0.3500
	3	CLWETM	**0.4044**	**79.8**	**0.512**	**0.466**
	4	ITD	0.3709^{12}	73.20	0.4840	0.4440
	5	EM4TT	0.3752^{12}	74.04	0.436	0.422
	6	EM4QT	0.3980^{12}	78.55	0.4880	0.4480
	7	LINEAR I	0.3617	71.38	0.4400	0.396
	8	DIVMIN I	0.4006^{124}	79.06	0.4800	0.4360
	9	LINEAR II	0.3900	76.97	0.4800	0.4380
	10	DIVMIN II	0.3734^{12}	73.66	0.4360	0.4080
German (DE)		Mono	0.3912	-	0.524	0.4840
	1	JCLTRLM	0.1683	43.02	0.2000	0.1880
	2	JCLTRLM†	0.1520	38.85	0.2160	0.1880
	3	CLWETM	0.2636	67.4	0.368	0.322
	4	ITD	0.2351^{12}	60.10	0.3160	0.2760
	5	EM4TT	0.2704^{12}	69.12	0.324	0.29
	6	EM4QT	$\mathbf{0.2955}^{1234}$	**75.54**	**0.380**	**0.3400**
	7	LINEAR I	0.2628^{12}	67.18	0.3280	0.2860
	8	DIVMIN I	0.2711^{123}	69.30	0.3440	0.3040
	9	LINEAR II	0.2718^{123}	69.48	0.3640	0.3000
	10	DIVMIN II	0.2618^{12}	66.92	0.3200	0.2740

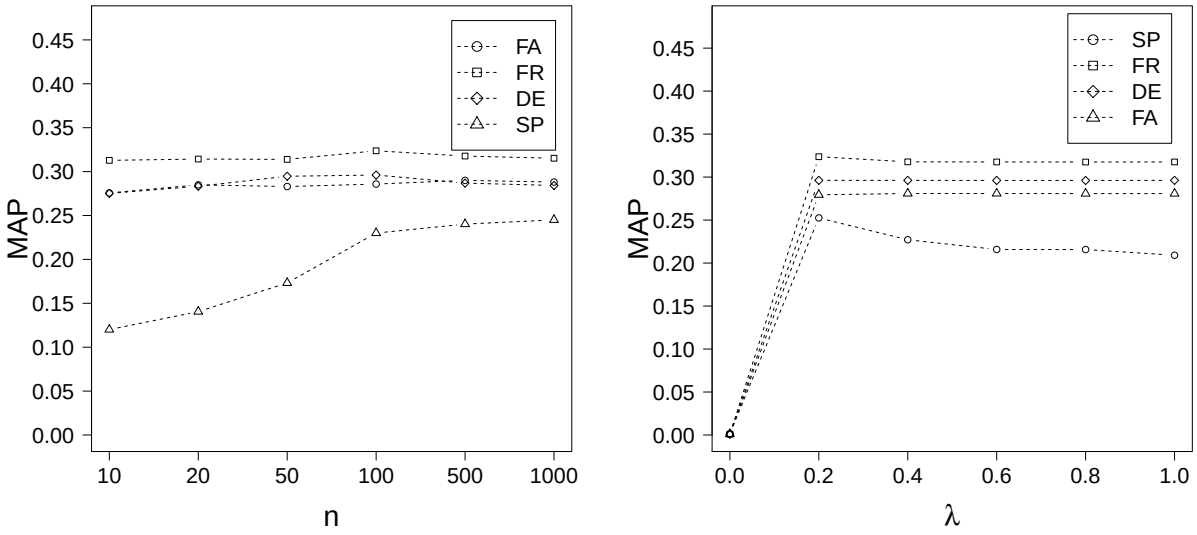


Figure 2: MAP achieved by EM4QT with different n and λ values.

tistically significant. It is worth noting that since the results obtained by ITD are sometimes by far lower than those achieved by EM, combining their results does not help to improve the results over the EM method. In addition, it should be noted that since EM4QT uses pseudo-relevant documents to the whole query, it considers term coherency in both source and target languages, indirectly. This is the reason for the high performance of the EM4QT method.

Combining EM4QT with EM4TT: Combining EM4QT with EM4TT is another option for studying features of EM4QT in more details. In Table 4 LINEAR-II and DIVMIN-II denote the results of this combination using linear interpolation and divergence minimization respectively. Interpolating EM4QT with its off-line version (EM4TT) brings variable results in each collection. As shown in Table 4, only in the FA collection there are overall improvements in all the evaluation metrics; the results in the FA collection show that interpolating EM4QT with EM4TT surpasses all the baselines particularly both EM4QT and EM4TT in terms of MAP, P@5, and P@10. Interestingly, the linear interpolation approach reaches 80.34% of the performance of monolingual retrieval in terms of MAP in this collection.

JCLTRLM: According to [10], JCLTRLM achieves better results compared to CLTRLM. Therefore the authors opted JCLTRLM in this set of experiments. As shown in Table 4, EM4QT performs better than JCLTRLM particularly in FR, SP, and DE. JCLTRLM assumes that although the coverage of the bilingual dictionaries are limited, word-topic distribution can compensate this deficiency (see $p(w^t|z_k^t, \phi_{k,w^t}^t)$ in Equation 2.1). But, this assumption can lead to query expansion with a number of non-relevant words which are not even translations of the user-specified query. Therefore, we added another set of experiments indicated by JCLTRLM† in which we estimate translation probabilities only for the terms provided in the dictionary. We believe that we have fair comparisons between the methods this way. As shown in Table 4, JCLTRLM† outperforms JCLTRLM in all the collections. In all the datasets even ITD, a collection-based translation model, outperforms JCLTRLM†. This might be due to the fact that JCLTRLM† heavily depends on the quality of the input collections in terms of their comparability [10].

CWETM: The results of the word embedding-based translation model are indicated by CLWETM. As discussed in Section 2.1.2, the obtained model from the low-dimensional word vectors can be interpolated by other models. To this aim, we interpolate the model with ITD; after finding a suitable interpolation parameter and proper number of pseudo-relevant documents computed by 2-fold cross-validation. According to [42], we opted the linear interpolation framework. As shown in Table 4, CLWETM achieved competitive results compared to all the baselines. However, interpolation of EM4QT with ITD or EM4TT outperforms CLWETM in almost all the collections. Although in FR, CLWETM outperforms EM4QT in terms of MAP, the improvement is not statistically significant. Furthermore, EM4QT (without interpolation) achieved higher P@5 and P@10 in FR. On the other hand, CLWETM outperforms all the baselines in terms of all the metrics in SP where the improvements compared to DIVMIN-I are marginal. In FA and DE, almost all the variants of EM4QT consistently have better results than CLWETM. In DE, EM4QT reachs 75.54% performance of Mono, which is by far better than all the baselines. This indicates that having an accurate general translation model for interpolation would lead to better results. Totally, it is hard to interpret the functionality of a neural network and then CLWET. But, it is more clear that having control on quite all the parameters of a model empowers the model for further improvements, investigations, and extensions. CLWETM neither has significant improvements compared to EM4QT nor is as interpretable as EM4QT over the hidden parameters.

Error analysis: Our query-by-query investigations reveal that a number of queries take advantage of the linear interpolation approach between EM4QT and EM4TT. Particularly, ambiguous queries benefit a lot from this approach since it prevents the system from converging to only one translation. For example, 48/100 queries in FA, 45/100 queries in FR, 24/50 queries in SP, and 23/50 queries in DE are improved by LINEAR-II. This denotes that finding an accurate α for each query can play a key role in the performance of the proposed CLIR system[10]. There are a couple of reasons for these outcomes; first, sometimes the correct translations are densely populated in a set of documents and have large term frequencies in those documents. Since EM4TT retrieves documents in response to each term of the query independently, it is expected to retrieve documents with frequent and discriminative query terms in higher ranks. The aim of EM4TT is to leave least common translations out from the set and to give more weights to topical translations. Our claim is that the proposed token-to-token method can recognize common tokens efficiently. As an evidence for our claim, we can mention the results of EM4TT and COLL in Table 3; although both EM4TT and COLL weight translations based on their frequencies in a collection, since EM4TT outperforms COLL (except in the FR collection), we can conclude that EM4TT finds the common translations more efficiently. Indeed, not only the collection frequency of a common token is important for the system, but also its distribution through the collection is a major factor. The second reason goes back to a property of the proposed EM4QT that mainly converges to only one translation. Indeed, in this method the most coherent translation absorbs the translation weightings considerably. This property can give an incorrect chance to an out-of-the-context translation and lead to retrieval of irrelevant documents. In these situations, the proposed combination framework soothes this effect and achieves a desirable performance.

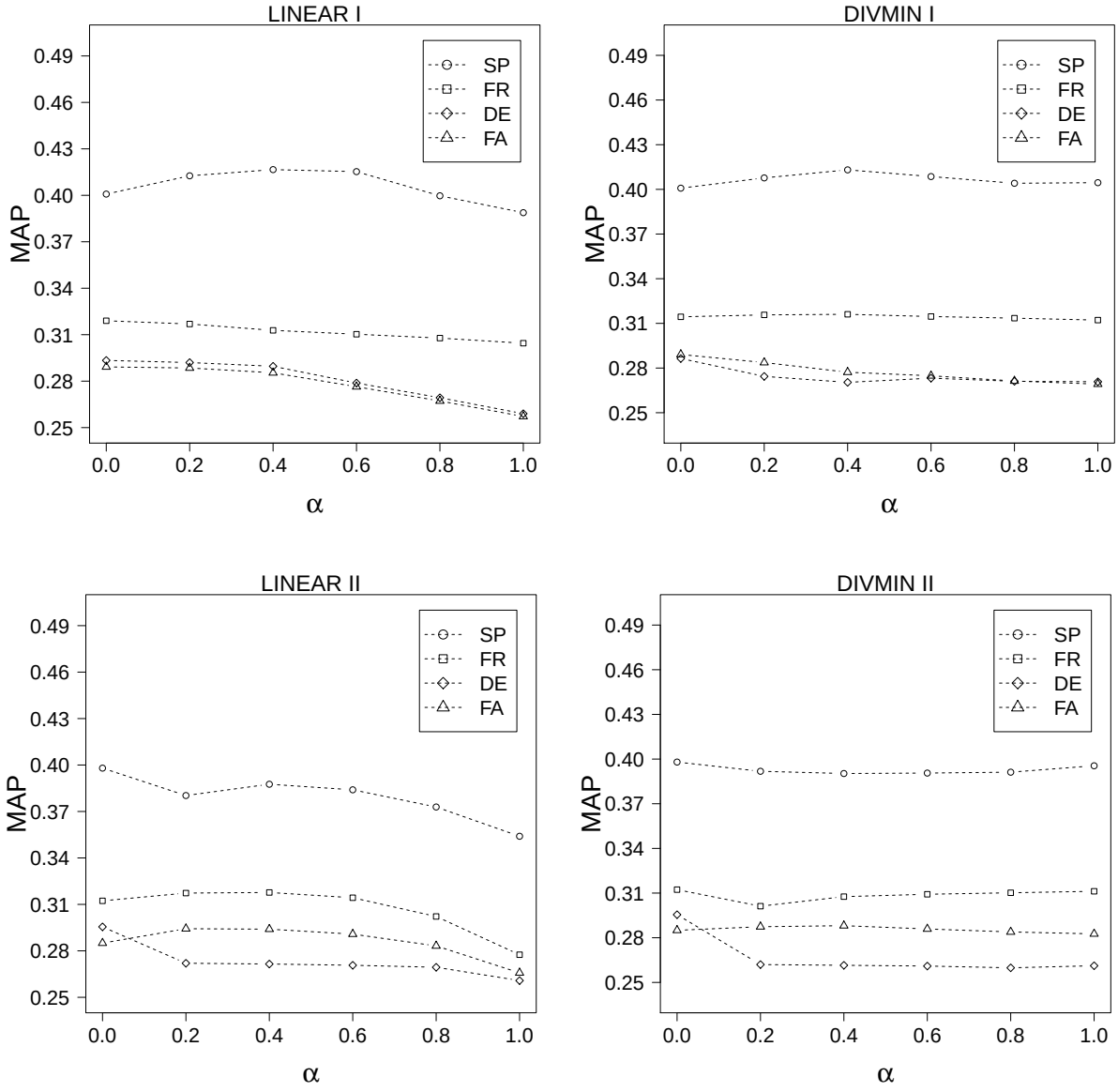


Figure 3: MAP achieved by LINEAR and DIVMIN methods with different α values.

4.4.2. Parameter sensitivity

We investigate the sensitivity of the proposed query dependent EM algorithm to two parameters λ and n in Figure 2. As mentioned in Section 3, λ controls the weight of the background collection in the EM algorithm and n denotes the number of pseudo-relevant documents that are used for translation knowledge extraction. According to Figure 2, by increasing the number of top-retrieved documents, the retrieval performance in the SP collection is also improved. The reason is that by increasing the number of top-retrieved documents, the diversities of the terms are also increased. In addition, the best translations in the SP collection seems to be the most common ones and thus by increasing the number of documents, it is more likely to select most translations. In the DE, FA, and FR collections, the retrieval performance is not very sensitive to the value of the parameter n. The reason is that

[10]We place this task as an interesting future work for the proposed system.

although by increasing the value of n, the coverage of terms is also enhanced, the amount of noise in the pseudo-relevant documents is also increased. When parameter n is greater than or equal to 100, the results in all the collections become stable and they do not have any significant differences.

Looking at the MAP obtained by different λ values in Figure 2, we notice that the EM method is not very sensitive to the parameter λ, except in the SP collection. As pointed out above, considering the SP queries, we realize that in most cases the best translation is the most common translation. That is why by increasing λ (decreasing the weight of the collection) the results are dropped. According to Equation (11), when λ is set to 0, the EM algorithm assigns zero weights to the pseudo-relevant documents and thus all the probabilities are also set to zero. This is why in this parameter setting the retrieval performance is equal to zero. It should be noted that in this experiment we use 2-fold cross-validation to find the values of the parameter n and when λ is set to a small value (giving high weight to the collection) the parameter n is also set to a small value (decreasing the amount of noise by decreasing the number of pseudo-relevant documents); otherwise, when λ value is close to 1, the selected value of the parameter n in cross-validation is also increased. This is why the results for different λ values are stable.

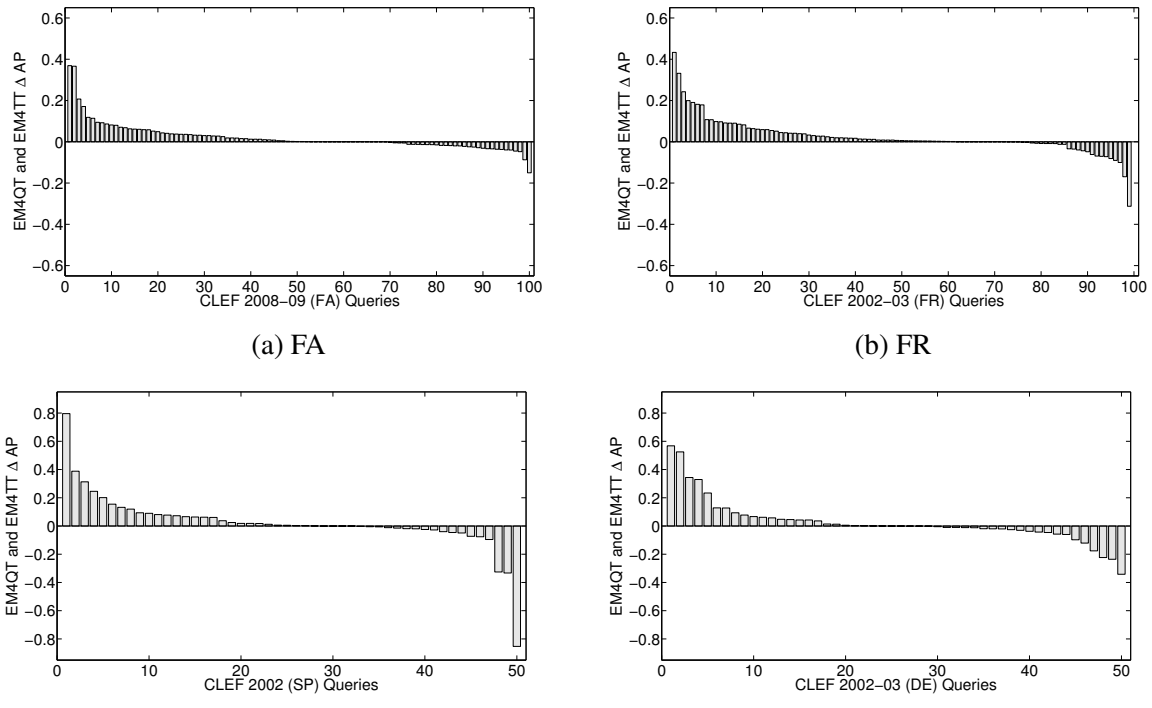
Figure 3 plots the sensitivity of MAP achieved by LINEAR and DIVMIN methods to the parameter α. According to this figure, interpolating the EM method with the ITD method (LINEAR-I and DIVMIN-I) does not lead to improving performance, in DE, FR, and FA collections. The reason is related to the low performance of the ITD method compared to the EM-based one in these collections (see Table 4). In contrast, in the SP collection increasing α to 0.4 and 0.8 in LINEAR-I and DIVMIN-I methods respectively, can help to improve the retrieval accuracy in terms of MAP. This observation demonstrates the effectiveness of the translation model combinations in the SP collection that the best translations are usually the common ones.

As shown in Figure 3 interpolating with EM4TT improves the performance of the proposed EM4QT method only in the FA and FR collections for some values of α. Amount of improvements heavily depends on the type of the queries and the statistics of the collection. Regarding to these results we can infer that EM4QT can take advantage of EM4TT in the proposed interpolation frameworks particularly in difficult and ambiguous queries.

4.4.3. Query-by-Query comparison of EM4QT and EM4TT

In this section we aim to elaborate on the effectiveness of the proposed on-line EM4QT method compared to EM4TT, its off-line version, in more details. Figure 4 presents average precision (AP) differences between EM4QT and EM4TT for a variety of topics in each dataset. 62%, 70%, 62%, and 52% of the queries in the Persian, French, Spanish, and German datasets are improved by EM4QT compared to the EM4TT method respectively. Amount of improvements are promising and AP differences of many degraded queries are bounded to 0.02.

There is a reason for obtaining improvements in a number of queries by EM4TT compared to EM4QT. In some queries the correct translations are densely populated in documents instead of appearing occasionally through the collection. Therefore, running the system for each query term separately retrieves in-the-context documents. Interestingly interpolation of these methods also achieves promising results in many queries (see Table 4). Nevertheless, considering the queries altogether, the improvements of EM4TT w.r.t EM4QT are not statistically significant.



(a) FA (b) FR

(c) SP (d) DE

Figure 4: AP differences between EM4QT and EM4TT in the CLEF datasets: Persian (FA), French (FR), Spanish (SP), and German (DE).

5. Conclusions and future work

In this paper, we investigate translation disambiguation for dictionary-based CLIR based on pseudo-relevant documents in both source and target languages. The main idea behind our approach is that each term in the target pseudo-relevant documents is either translated from the source pseudo-relevant documents or comes from a noisy background language model. We extract the translation knowledge based on the mentioned idea by proposing an expectation-maximization method, called EM4QT. This method was developed based on the statistical language modeling framework. We further investigate the possibility of improving the retrieval performance by combining the extracted translation model with a coherency-based CLIR method and a proposed token-to-token translation method. These combinations can be done using a simple linear interpolation or our proposed divergence minimization (DIVMIN) method.

Experimental results on four CLEF cross-language collections in four different languages: Spanish (CLEF-2002), German (CLEF 2002-03), French (CLEF 2002-03), and Persian (CLEF 2008-09) demonstrates that the proposed EM algorithm outperforms the baselines in terms of MAP, P@5, and P@10. The MAP improvements are always statistically significant. In addition, the results obtained by combining an extracted translation model by EM4QT and a coherence based model do not achieve a significantly better performance. This shows that EM4QT considers target term coherency, indirectly. On the other hand it has been shown that its combination with the proposed token-to-token translation model prevents the situations that the proposed EM4QT method drifts to irrelevant translation N-grams and thus achieves promising results.

Future research studies can be emerged on combining translation models obtained from comparable and parallel corpora. Moreover, since the amount of noise from the collection in the pseudo-

relevant documents varies from one query to another it seems to be an advantageous to define a number of documents in EM4QT for each query dynamically. It is also a demanding task to define a more accurate coefficient for each translation model in the proposed combination frameworks for each type of the query. It seems that difficult queries can take advantage of these combination approaches considerably.

Acknowledgments

This research was in part supported by two grants from Institute for Research in Fundamental Sciences (no. CS1395- 4-19 and no. CS1395-4-05) and in part by the Center for Intelligent Information Retrieval. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsors. The authors sincerely would like to thank anonymous reviewers for their constructive comments. We gratefully thank Ivan Vulić for providing the implementation of BiLDA, and Debasis Ganguly for his helpful comments on JCLTRLM.

References

[1] Azarbonyad, H., Shakery, A., Faili, H., 2012. Using learning to rank approach for parallel corpora based cross language information retrieval. In: ECAI. pp. 79–84.

[2] Blei, D. M., Ng, A. Y., Jordan, M. I., 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022.

[3] Chen, A., Gey, F. C., 2003. Combining query translation and document translation in cross-language retrieval. In: CLEF. Trondheim, Norway, pp. 108–121.

[4] Croft, W. B., Harper, D. J., 1979. Using probabilistic models of document retrieval without relevance information. Journal of Documentation 35 (4), 285–295.

[5] Dadashkarimi, J., Shahshahani, M. S., Tebbifakhr, A., Faili, H., Shakery, A., 2016. Dimension projection among languages based on pseudo-relevant documents for query translation. arXiv preprint arXiv:1605.07844.

[6] Dadashkarimi, J., Shakery, A., Faili, H., 2014. A probabilistic translation method for dictionary-based cross-lingual information retrieval in agglutinative languages. In: Conference of Computational Linguistic. Tehran, Iran.

[7] Dehghani, M., Azarbonyad, H., Kamps, J., Hiemstra, D., Marx, M., 2016. Luhn revisited: significant words language models. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. CIKM '16. pp. 1301–1310.

[8] Franco-Salvador, M., Rosso, P., Montes-y Gómez, M., 2016. A systematic study of knowledge graph analysis for cross-language plagiarism detection. Information Processing & Management.

[9] Franco-Salvador, M., Rosso, P., Navigli, R., 2014. A knowledge-based representation for cross-language document retrieval and categorization. In: EACL. pp. 414–423.

[10] Ganguly, D., Leveling, J., Jones, G., 2012. Cross-lingual topical relevance models. In: COLING. Mumbai, India, pp. 927–942.

[11] Gao, J., Nie, J.-Y., Xun, E., Zhang, J., Zhou, M., Huang, C., 2001. Improving query translation for cross-language information retrieval using statistical models. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '01. New Orleans, Louisiana, USA, pp. 96–104.

[12] Gao, J., Qi, H., Xia, X., Nie, J.-Y., 2005. Linear discriminant model for information retrieval. In: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, pp. 290–297.

[13] Gearailt, D. N., 2003. Dictionary characteristics in cross-language information retrieval. Ph.D. thesis, University of Cambridge.

[14] Gouws, S., Bengio, Y., Corrado, G., 2014. Bilbowa: fast bilingual distributed representations without word alignments. arXiv preprint arXiv:1410.2455.

[15] Gupta, P., Bali, K., Banchs, R. E., Choudhury, M., Rosso, P., 2014. Query expansion for mixed-script information retrieval. In: Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval. ACM, pp. 677–686.

[16] Hashemi, H. B., Shakery, A., 2014. Mining a persian-english comparable corpus for cross-language information retrieval. Inf. Process. Manage. 50 (2), 384–398.

[17] Kullback, S., Leibler, R. A., 1951. On information and sufficiency. Ann. Math. Statist. 22 (1), 79–86.

[18] Lafferty, J., Zhai, C., 2001. Document language models, query models, and risk minimization for information retrieval. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New Orleans, Louisiana, USA, pp. 111–119.

[19] 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. Tampere, Finland, pp. 175–182.

[20] Lavrenko, V., Croft, W. B., 2001. Relevance based language models. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '01. New Orleans, Louisiana, USA, pp. 120–127.

[21] Lee, C.-J., Croft, W. B., 2014. Cross-language pseudo-relevance feedback techniques for informal text. In: Proceedings of the 36th European Conference on Information Retrieval. ECIR '14. Amsterdam, The Netherlands, pp. 260–272.

[22] Levow, G.-A., Oard, D. W., Resnik, P., 2005. Dictionary-based techniques for cross-language information retrieval. Inf. Process. Manage. 41 (3), 523–547.

[23] Liu, Y., Jin, R., Chai, J. Y., 2005. A maximum coherence model for dictionary-based cross-language information retrieval. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Salvador, Brazil, pp. 536–543.

[24] McLachlan, G. J., Krishnan, T., 2008. The EM algorithm and extensions, 2nd Edition. Wiley series in probability and statistics. Wiley.

[25] Mikolov, T., Le, Q. V., Sutskever, I., 2013. Exploiting similarities among languages for machine translation. arXiv preprint arXiv:1309.4168.

[26] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., Dean, J., 2013. Distributed representations of words and phrases and their compositionality. In: Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K. Q. (Eds.), Advances in Neural Information Processing Systems 26. Curran Associates, Inc., pp. 3111–3119.

[27] Montazeralghaem, A., Zamani, H., Shakery, A., 2016. Axiomatic analysis for improving the log-logistic feedback model. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '16. ACM, New York, NY, USA, pp. 765–768.

[28] Monz, C., Dorr, B. J., 2005. Iterative translation disambiguation for cross-language information retrieval. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Salvador, Brazil, pp. 520–527.

[29] Nie, J.-Y., 2010. Cross-language information retrieval. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

[30] Pirkola, A., 1998. The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In: Proceedings of the 21th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Melbourne, Australia, pp. 55–63.

[31] Pirkola, A., Hedlund, T., Keskustalo, H., Järvelin, K., 2001. Dictionary-based cross-language information retrieval: problems, methods, and research findings. Information Retrieval 4 (3-4), 209–230.

[32] Platt, J. C., Toutanova, K., Yih, W.-t., 2010. Translingual document representations from discriminative projections. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, pp. 251–261.

[33] Rahimi, N., Shakery, A., Dadashkarimi, J., Aryannejad, M., Dehghani, M., Esfahani, H. N., 2016. Building a multi-domain comparable corpus using a learning to rank method. Natural Language Engineering 22 (2).

[34] Rahimi, R., Shakery, A., 2013. A language modeling approach for extracting translation knowledge from comparable corpora. In: Proceedings of the 35th European Conference on Information Retrieval. ECIR '13. Moscow, Russia, pp. 606–617.

[35] Rahimi, R., Shakery, A., King, I., 2015. Multilingual information retrieval in the language modeling framework. Inf. Retr. Journal 18 (3), 246–281.

[36] Shakery, A., Zhai, C., 2013. Leveraging comparable corpora for cross-lingual information retrieval in resource-lean language pairs. Inf. Retr. 16 (1), 1–29.

[37] Tabrizi, S. A., Dadashkarimi, J., Dehghani, M., Nasr Esfahani, H., Shakery, A., 2015. Revisiting optimal rank aggregation: a dynamic programming approach. In: Proceedings of the 2015 International Conference on the Theory of Information Retrieval. ICTIR '15. Northampton, Massachusetts, USA, pp. 353–356.

[38] Talvensaari, T., Laurikkala, J., Järvelin, K., Juhola, M., Keskustalo, H., 2007. Creating and exploiting a comparable corpus in cross-language information retrieval. ACM Transactions on Information Systems (TOIS) 25 (1), 4.

[39] Theodoridis, S., Koutroumbas, K., 1999. Pattern recognition. Academic Press.

[40] Ture, F., Lin, J., Oard, D., 2012. Combining statistical translation techniques for cross-language information retrieval. In: COLING. Mumbai, India, pp. 2685–2702.

[41] Vulic, I., Moens, M., 2014. Probabilistic models of cross-lingual semantic similarity in context based on latent cross-lingual concepts induced from comparable data. In: EMNLP. pp. 349–362.

[42] Vulic, I., Moens, M., 2015. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In: Proceedings of the 38th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 363–372.

[43] Vulic, I., Smet, W. D., Tang, J., Moens, M., 2015. Probabilistic topic modeling in multilingual settings: an overview of its methodology and applications. Inf. Process. Manage. 51 (1), 111–147.

[44] Xu, J., Croft, W. B., 1996. Query expansion using local and global document analysis. In: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '96. Zurich, Switzerland, pp. 4–11.

[45] Xu, J., Weischedel, R., 2005. Empirical studies on the impact of lexical resources on clir performance. Inf. Process. Manage. 41 (3), 475–487.

[46] Yih, W.-t., Toutanova, K., Platt, J. C., Meek, C., 2011. Learning discriminative projections for text similarity measures. In: Proceedings of the Fifteenth Conference on Computational Natural Language Learning. CoNLL '11. Portland, Oregon, pp. 247–256.

[47] Zamani, H., Croft, W. B., 2016. Embedding-based query language models. In: Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval. ICTIR '16. ACM, pp. 147–156.

[48] Zamani, H., Dadashkarimi, J., Shakery, A., Croft, W. B., 2016. Pseudo-relevance feedback based on matrix factorization. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. CIKM '16. ACM, pp. 1483–1492.

[49] Zamani, H., Faili, H., Shakery, A., 2016. Sentence alignment using local and global information. Computer Speech & Language 39, 88–107.

[50] Zhai, C., 2004. A note on the expectation-maximization (em) algorithm. Tech. rep., Carnegie Mellon University.

[51] Zhai, C., 2008. Statistical language models for information retrieval. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

[52] Zhai, C., Lafferty, J., 2001. Model-based feedback in the language modeling approach to information retrieval. In: Proceedings of the 10th International Conference on Information and Knowledge Management. Atlanta, Georgia, USA, pp. 403–410.

[53] Zhai, C., Lafferty, J., 2004. A study of smoothing methods for language models applied to information retrieval. ACM Trans. Inf. Syst. 22 (2), 179–214.

Appendix A. Computing the divergence minimization formula.

In this section we provide the mathematical details of the proposed divergence minimization framework. Our goal heare is to achieve a translation model obtained from combining two models in which one model is α times more important than the other.

$$\begin{aligned}\mathcal{D}(w_s, \mathcal{C}_s, \mathcal{C}_t, F, F') &= \mathcal{D}(\mathcal{T}||\mathcal{T}_1) + \alpha\mathcal{D}(\mathcal{T}||\mathcal{T}_2) \\ &= \textstyle\sum_{w_t \in \mathcal{T}(w_s)} p(w_t|\mathcal{T}) \log \frac{p(w_t|\mathcal{T})}{p(w_t|\mathcal{T}_1)} \\ &+\alpha \textstyle\sum_{w_t \in \mathcal{T}(w_s)} p(w_t|\mathcal{T}) \log \frac{p(w_t|\mathcal{T})}{p(w_t|\mathcal{T}_2)} \end{aligned} \tag{A.1}$$

In EquationA.1 we can compute $p(w_t|\mathcal{T})$ as follows:

$$p(w_t|\mathcal{T}) = \sum_{w_s \in \mathcal{T}(w_t)} \mathcal{T}(w_t|w_s) p(w_s|\mathcal{T}) \text{ and } \sum_{w_t \in \mathcal{T}(w_s)} \mathcal{T}(w_t|w_s) = 1, \tag{A.2}$$

We want to minimize $\mathcal{D}$ subject to the constraint in A.2:

$$\mathcal{D}_{min}(w_s, \mathcal{T}_1, \mathcal{T}_2) = min\Bigg(\mathcal{D}(\mathcal{T}||\mathcal{T}_1) + \alpha\mathcal{D}(\mathcal{T}||\mathcal{T}_2) + \lambda(\sum_{w_t \in \mathcal{T}(w_s)} \mathcal{T}(w_t|w_s) - 1)\Bigg) \tag{A.3}$$

To optimize the function $\mathcal{D}$ with respect to variables $\mathcal{T}(.|.)$ we can have the following equation:

$$\frac{\partial \mathcal{D}(w_s, \mathcal{T}_1, \mathcal{T}_2)}{\partial \mathcal{T}(w_t|w_s)} = p(w_s|\mathcal{T})\Bigg(\log \frac{p(w_s|\mathcal{T})\mathcal{T}(w_t|w_s)}{p(w_t|\mathcal{T}_1)} + 1\Bigg) \tag{A.4}$$

$$+\alpha p(w_s|\mathcal{T})\Bigg(\log \frac{p(w_s|\mathcal{T})\mathcal{T}(w_t|w_s)}{p(w_t|\mathcal{T}_2)} + 1\Bigg) \tag{A.5}$$

$$+\lambda = 0. \tag{A.6}$$

So we have:

$$\mathcal{T}(w_t|w_s) = \exp\Bigg(-1 - \tfrac{\lambda}{(1+\alpha)p(w_s|\mathcal{T})} \tag{A.7}$$

$$- \tfrac{1}{(1+\alpha)} \log \tfrac{p(w_s|\mathcal{T})}{p(w_t|\mathcal{T}_1)} \tag{A.8}$$

$$- \tfrac{\alpha}{(1+\alpha)} \log \tfrac{p(w_s|\mathcal{T})}{p(w_t|\mathcal{T}_2)} \Bigg), \tag{A.9}$$

Since $\exp(-(1+\frac{\lambda}{(1+\alpha)p(w_s|\mathcal{T})}+\frac{\log p(w_s|\mathcal{T})}{(1+\alpha)}))$ is constant for all translations of w_s we have following equation:

$$\mathcal{T}(w_t|w_s) \propto \exp\left(\frac{1}{(1+\alpha)} \log p(w_t|\mathcal{T}_1) + \frac{\alpha}{(1+\alpha)} \log p(w_t|\mathcal{T}_2)\right)$$

$$= \exp\left(\frac{1}{(1+\alpha)} \log \mathcal{T}_1(w_t|w_s) + \frac{\alpha}{(1+\alpha)} \log \mathcal{T}_2(w_t|w_s)\right). \tag{A.10}$$

Appendix B. Convergence of the proposed EM-based method

In this section, we investigate the convergence of the proposed EM algorithm. Equation (B.1) shows the likelihood function which is maximized subject to $\sum_{w_t \in \mathcal{T}\{w_s\}} \mathcal{T}(w_t|w_s) = 1$:

$$\log p(F|F') = \sum_{w_s \in V_s} p(w_s|F) \times \log\big(\sum_{w_t \in \mathcal{T}\{w_s\}} \mathcal{T}(w_t|w_s) p(w_t|F')\big) \tag{B.1}$$

where $\mathcal{T}(w_t|w_s)$ can be computed using Equation (7). For given values of λ, $\mathcal{C}$, F, and F' the maximum likelihood estimator aims to maximize $p(w_t|T=1,w_s)$ in the M-Step of the proposed EM algorithm (see Equation (12)) to reach a local optimum for the likelihood function. Since logarithm is a monotonic function and $p^{(n+1)}(w_t|T=1,w_s) > p^{(n)}(w_t|T=1,w_s)$ then we have:

$$\Delta\big(p^{(n+1)}(w_t|T=1,w_s)|p^{(n)}(w_t|T=1,w_s)\big)$$

$$\geq \Delta\big(p^{(n)}(w_t|T=1,w_s)|p^{(n)}(w_t|T=1,w_s)\big) = 0 \tag{B.2}$$

where $\Delta(\cdot|\cdot)$ denotes the difference between two estimations. As presented in Equation (B.2), we have increasing changes in each iteration. Hence, the likelihood function definitely will reach its stationary point and the algorithm will be eventually converged [24].

Furthermore, for the given values of λ, $\mathcal{C}$, F, and F' and due to the independence assumption for w_s, there is only one stationary point in the algorithm [50, 51]. As a result, the proposed algorithm would reach its global maximum. More details can be found in [50, 24].