Dimension Projection among Languages based on Pseudo-relevant Documents for Query Translation

Javid Dadashkarimi, Mahsa S. Shahshahani, Amirhossein Tebbifakhr, Heshaam Faili, and Azadeh Shakery
School of ECE, College of Engineering, University of Tehran, Iran
{dadashkarimi,ms.shahshahani,a.tebbifakhr,hfaili,shakery}@ut.ac.ir

ABSTRACT

Taking advantage of top-ranked documents in response to a query for improving quality of query translation has been shown to be an effective approach for cross-language information retrieval. In this paper, we propose a new method for query translation based on dimension projection of embedded vectors from the pseudo-relevant documents in the source language to their equivalents in the target language. To this end, first we learn low-dimensional representations of the words in the pseudo-relevant collections separately and then aim at finding a query-dependent transformation matrix between the vectors of translation pairs. At the next step, representation of each query term is projected to the target language and then, after using a softmax function, a query-dependent translation model is built. Finally, the model is used for query translation. Our experiments on four CLEF collections in French, Spanish, German, and Persian demonstrate that the proposed method outperforms all competitive baselines in language modelling, particularly when it is combined with a collection-dependent translation model.

Keywords
query translation, low-dimensional vectors, cross-language information retrieval, pseudo-relevant documents.

1. INTRODUCTION

Pseudo-relevance feedback (PRF) has been shown to be an effective approach for updating query language models in information retrieval (IR) [8]. In cross-language environments where there are a couple of top-ranked document sets in different languages building such models requires bridging the gap of the languages. To this end, cross-lingual topical relevance models (CLTRLM) aim at finding a way for transferring knowledge of the sets to the query using bilingual topic modeling and bilingual dictionaries [5, 2]. In this paper we propose a new method for building translation models on pseudo-relevant collections using neural network language modelling. The proposed cross-lingual word embedding translation model (CLWETM) takes advantage of a query-dependent transformation matrix between low-dimensional vectors of the languages. Indeed, we aim at finding a transformation matrix to bring the vector of each query term to dimensionality of the target language and then computing translation probabilities based on the vector similarities. To this aim, first we learn word representations of the pseudo-relevant collections separately and then focus on finding a transformation matrix minimizing a distance function between all the translation pairs. Finally, a softmax function is used for building a query-dependent translation model based on similarity of transformed vector of each query term with low-dimensional vectors of its translation candidates in the target language.

Unlike CLTRLM, CLWETM considers the relative positions of the co-occurred words and therefore we believe that the obtained model would capture the context of the query more accurately. Furthermore, the obtained model is incorporated within a language modelling framework, the state-of-the-art retrieval framework in the literature, and therefore the proposed method does not suffer from disadvantages of low-dimensional document representations in IR [13].

Experimental results on four CLEF collections in French, Spanish, German, and Persian demonstrate that the proposed method outperforms all competitive baselines in language modelling, particularly when it is combined with a collection-dependent or query-independent translation model.

2. RELATED WORKS

Pseudo-relevance feedback has long been employed as a powerful method for estimating query language models in a large number of studies [8, 9]. But, in multi-lingual environments there are only a limited number of methods to this end; cross-lingual relevance models (CLRLM) and CLTRLM are state-of-the-art methods in this area [2, 14, 5]. Unlike CLRLM that depends on parallel corpora and bilingual lexicons, CLTRLM aims at finding a number of bilingual topical variables from a comparable corpus in order to transfer relevance score of a term from one language to another. Ganguly et al. proposed to use these variables for query translation and demonstrated that CLTRLM is an effective method particularly for resource-lean languages [2].

In CLTRLM top-ranked documents \(T^s = \{d_1^s, d_2^s, \ldots, d_s^s\}\) retrieved in response to \(q^s\), the user-specified query, and top-ranked documents \(T^t = \{d_1^t, d_2^t, \ldots, d_t^t\}\) retrieved in response to \(q^t\), translation of the query, are assumed to be relevant documents. CLTRLM assumes that each word \(w^t\) in target language is generated either from a target language
follows:

\[ p(w^t|q^s) = |p(w^t|z^t)p(z^t|q^s) + p(w^t|w^s)p(w^s|q^s) \]  

(1)

in which \( z^t \) is a topical variable on \( F^t \) and \( w^s \) is a translation of \( w^t \) in the dictionary. For estimating \( p(z^t|q^s) \) and \( p(w^s|q^s) \) a topic model of \( F^t \) and a topic model of \( F^s \) are used, respectively and then topic projection is calculated based on translation probabilities provided in the dictionary \[3\]. It is noteworthy that topic modeling considers co-occurrences of the terms within documents which means that the positions of the terms do not play an important role in the model.

Language modeling based on neural networks is a popular technique for capturing both co-occurrences of the terms within a constant window \( c \) and also their positions in the context \[1]\. Learning vectors of terms relies on minimizing a loss function like a binary regression function \( L(\theta) \) as follows:

\[ L(\theta) = \sum_{(w,c) \in D \times D'} \log \left( \frac{1}{1 + \exp^{-v^T w}} \right)^z \left( \frac{1}{1 + \exp^{v^T w}} \right)^{1-z} \]  

(2)

in which \( v_w \in \mathbb{R}^{n \times 1} \) and \( v_c \in \mathbb{R}^{n \times 1} \) are the vectors of a term and its context respectively. \( z = 1 \) for valid pairs \((w, c) \in D \) and \( z = 0 \) for invalid ones \((w, c) \in D' \). But when it comes to cross-lingual environments a constraint is required for integrating dimensions in the languages. There are a couple of methods to this end; on-line methods and off-line methods. On-line methods aim at finding a unified space for the languages during the learning process as follows \[3\] :

\[ L(\theta) = L(\theta^t) + L(\theta^s) + \lambda f(\theta^t, \theta^s) \]  

(3)

Off-line methods learn the representation vectors separately and then aims at finding a transformation matrix by minimizing a constraint function \( f \) as follows \[10\] :

\[ f(W) = \sum_{x,z} ||W^T x - z||^2 \]  

(4)

in which \( x \in \mathbb{R}^{n \times 1} \) and \( z \in \mathbb{R}^{n \times 1} \) are the vectors of translation pairs from a dictionary or a parallel corpus.

### 3. Linear Projection Between Languages Based on Pseudo Comparable Documents

Skip-gram negative sampling has been shown to be an effective approach for learning low-dimensional vectors for each word of a language \[11\]. In this paper, we aim to employ an off-line approach for learning bilingual representations of the words related to the query by exploiting pseudo-relevant documents in both source and target languages as our input collection and a bilingual dictionary as a cross-lingual resource. In more details, first we learn word representations of the pseudo-relevant collections separately and then focus on finding a transformation matrix minimizing a distance function between all the translation pairs. As shown in Equation 4 our goal is to minimize \( f \) with respect to a transformation matrix \( W \in \mathbb{R}^{n \times n} \), a number of word pairs (i.e., \( w^s \) from the source collection and \( w^t \) from the bilingual dictionary appeared in the target collection), and their low-dimensional representations \( u_{w^s} \in \mathbb{R}^{n \times 1} \) and \( u_{w^t} \in \mathbb{R}^{n \times 1} \); \( f \) is defined as follow:

\[ f(W) = \sum_{w^s \in T(w^t)} \frac{1}{2} ||W^T u_{w^s} - v_{w^t}||^2 \]  

(5)

To solve this problem we opted the gradient descent algorithm (i.e., \( \frac{\partial f}{\partial W} = 0 \)):

\[ W^{t+1} = W^t - \eta (W^T u_{w^t} - v_{w^t}) u_{w^t}^T \]  

(6)

### 3.1 Bilinguial Representations and Translation Models

#### 3.1.1 Based on Dimension Projection

Vector space models based on low-dimensional vectors are suffering from a number of challenges in IR including but not limited to document representation, document length normalization, and similarity measures to some extent \[13\]. In this section we introduce a method based on bilingual word representations for building a translation model for query translation and then incorporating it within a state-of-the-art retrieval framework.

\[ T(w_s|w_t) = \frac{\exp \left( \frac{(W^T u_{w^s})}{||W^T u_{w^s}||} \right)}{\sum_{w^s \in T(w^t)} \exp \left( \frac{(W^T u_{w^s})}{||W^T u_{w^s}||} \right)} \]  

(7)

#### 3.1.2 Based on Mixed Word Embedding

Vulic et al., introduced a mixed word embedding method based on a shuffling approach over comparable corpora \[13\]. We employed this method as one of the baselines of the proposed method. We build translation models as follows:

\[ T(w_s|w_t) = \frac{\exp \left( \frac{(v_{w^s} + s_{w^s})}{||v_{w^s} + s_{w^s}||} \right)}{\sum_{w^s \in T(w^t)} \exp \left( \frac{(v_{w^s} + s_{w^s})}{||v_{w^s} + s_{w^s}||} \right)} \]  

(8)

#### 3.1.3 Combining Translation Models

It is noteworthy that since the obtained model is a probabilistic translation model we can interpolate it with other models using a constant controlling parameter as follow:

\[ T = (1 - \alpha)T_1 + \alpha T_2 \]  

(9)

### 4. Experiments

#### 4.1 Experimental Setup

Overview of the used collections are provided in Table 1. The source collection for learning \( u_{w^t} \) is a pool of Associated Press 1988-89, Los Angeles Times 1994, and Glasgow Herald 1995 collections. These collections are used in previous TREC and CLEF evaluation campaigns for ad-hoc retrieval. In all experiments, we use the language modeling framework with the KL-divergence retrieval model and Dirichlet smoothing method to estimate the document language models, where we set the dirichlet prior smoothing parameter \( \mu \) to the typical value of 1000. To improve the retrieval performance, we use the mixture model 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.

All European dictionaries, documents, and queries are normalized and stemmed using the Porter stemmer. Document and query normalization techniques are employed as the retrieval engine in our experiments.

We use the Google dictionaries in our experiments. As discussed in Section 3, we used stochastic gradient descent for learning $W$ which is initialized with random values in $[-1, 1]$; $\eta$ is set to a small value which also decreases after each iteration. $u_w$ and $v_w$ are computed based on negative sampling skip-gram introduced in [11]; the size of the window, the number of negative samples, and the size of the vectors are set to typical values of 10, 45, and 50 respectively.

As shown in [2] joint cross-lingual topical relevance model (JCLTRLM) outperforms CLTRLM and therefore we opted JCLTRLM as a baseline. The parameters of LDA are set to typical values of 10, 45, and 50 respectively.

4.2 Performance Comparison and Discussion

In this section we want to compare effectiveness of a number of competitive methods in CLIR with effectiveness of the proposed method. We consider the following dictionary-based CLIR methods to evaluate the proposed method: (1) the top-1 translation of each term in the bilingual dictionary.

### Table 1: Collections Characteristics

| ID | Lang. | Collection | Queries (title only) | #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 |

### Table 2: Comparison of different CLIR methods. Superscripts indicate that the MAP improvements are statistically significant.

| | Persian (FA) | French (FR) | Spanish (SP) | German (DE) |
|---|-------------|-------------|--------------|-------------|
| | MAP | %Mono | P@5 | P@10 | Mono | %M | P@5 | P@10 |
| Mono | 0.3659 | 0.588 | 0.562 | 0.407 | 0.5253 | 0.4697 |
| TOP-I | 0.2135 | 58.34 | 0.3480 | 0.3460 | 0.2524 | 0.3240 | 0.2760 |
| UNIFORM | 0.1977 | 54.03 | 0.3240 | 0.3120 | 0.2367 | 0.2912 | 0.2421 |
| PARALLEL | 0.1679 | 45.89 | 0.292 | 0.282 | 0.2266 | 0.2542 | 0.2212 |
| BiCTM | 0.2547 | 69.61 | 0.406 | 0.406 | 0.3042 | 0.3899 | 0.3515 |
| BiCTM | 0.2523 | 69.61 | 0.400 | 0.391 | JCLTRLM | 0.2266 | 0.3414 | 0.299 |
| JCLTRLM | 0.2347 | 64.14 | 0.3740 | 0.3600 | BiCTM | 0.3042 | 0.4334 | 0.3444 |
| CLWETM$^{123}$ | 0.2523 | 69.61 | 0.400 | 0.391 | JCLTRLM | 0.2266 | 0.3414 | 0.299 |
| MIXWETM$^{2,345}$ | 0.2695 | 73.65 | 0.4280 | 0.4190 | MIXWETM$^{2,345}$ | 0.2983 | 0.3919 | 0.3485 |
| MIXWETM$^{2,345}$ | 0.2774 | 75.81 | 0.4540 | 0.4250 | MIXWETM$^{2,345}$ | 0.3183 | 0.4202 | 0.3707 |

1 We use the stopword lists and the normalizing techniques available at http://www.unine.ch/info/clef/
2 http://www.lemurproject.org/
3 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 http://www.speech.sri.com/projects/srilm/
ies (TOP-1), (2) all the possible translations of each term with equal weights (UNIFORM), (3) model obtained from the parallel corpora (PARALLEL), (4) BiCTM) proposed in [12], (5) the JCLTRLM method proposed in [2], and (6) the mixed word embedding translation model (MIXWETM), introduced in Section 3.1.2. We provided the results of CLWETM for $\alpha = 0$ (i.e., without interpolation. See Equation [6]) and $\alpha^*$ which is determined by a 2-fold cross-validation over each query set. All the results are summarized in Table 2. As shown in the table, both MIXWETM and CLWETM outperform other methods in terms of MAP, P@5, and P@10 in all the collections. Except in Spanish, MIXWETM and CLWETM consistently achieved better results compared to others in terms of all the metrics; but, CLWETM is clearly more effective than MIXWETM, whose context is by far larger than of CLWETM. The results of JCLTRLM reveals that top-ranked documents in SP and DE are not as comparable as FA or FR. Therefore we hope more improvements in FA and FR than others for CLWETM and MIXWETM. It is noteworthy that in DE the performance of CLWETM is improved by increasing $n$ even for smaller values of $\alpha$ (see Figure 1):

In FA, FR, ES, and DE, CLWETM achieved 75.81%, 78.21%, 76.30%, and 66.67% of performance of the monolingual runs, respectively. CLWETM outperforms the original BiCTM in terms of MAP by 8.9%, 4.6%, 4.2%, and 10.93% on FA, FR, SP, and DE respectively. CLWETM$_{\alpha=0}$ outperforms TOP-1, UNIFORM, and PARALLEL, all the off-line translation models, on all the collections in almost all the metrics. It demonstrates the following issues: importance of tailoring top-ranked documents in both languages, incorporating a positional language model on the documents (compare the results of CLWETM and MIXWETM), and employing an effective transformation framework between the languages.

4.3 Parameter Sensitivity

We investigate the sensitivity of the proposed method to two parameters $\alpha$ and $n$ in Figure 1. We first fix one parameter to its optimal value and then try to get optimal value of the other one. It demonstrates that both parameters work stably across SP, FA and FR collections. The optimal $\alpha$ value is 0.6 and the optimal $n$ value is 0 in these three collections. In the DE collection, the optimal values for the both parameters are somehow different; the optimal $\alpha$ value is 0.2, suggesting that CLWETM works well without even being combined with BiCTM. The optimal $n$ value is 50 in DE, suggesting that top-retrieved documents are not good enough. Therefore increasing number of top-retrieved documents improves the retrieval performance.

5. CONCLUSION AND FUTURE WORKS

In this paper we presented a translation model for cross-lingual information retrieval, that uses feedback documents in source and target languages for creating word vectors in each one, and then learns a projection matrix to project word vectors in source language to their translations from dictionary in target language. Then we introduced a method for building a translation model that can be easily interpolated with other models using a constant controlling parameter. Our method showed more improvements in FA and FR since top-ranked documents in SP and DE are not as comparable as theirs; therefore applying the proposed method on comparable corpora is considered as an interesting future work.

6. REFERENCES

[1] A. Aker, M. Paramita, M. Pinnis, and R. Gaizauskas. Bilingual dictionaries for all eu languages. In LREC’14, may 2014.
[2] D. Ganguly, J. Leveling, and G. Jones. Cross-Lingual Topical Relevance Models. In COLING, pages 927–942, 2012.
[3] S. Gouws, Y. Bengio, and G. Corrado. Bilbowa: Fast bilingual distributed representations without word alignments. arXiv preprint:1410.2455, 2014.
[4] H. B. Hashemi and A. Shakery. Mining a Persian-English Comparable Corpus for Cross-language Information Retrieval. IPRI, 50(2):384–398, 2014.
[5] V. Lavrenko, M. Choquette, and W. B. Croft. Cross-lingual relevance models. In SIGIR ’03, pages 175–182, 2002.
[6] Q. V. Le and T. Mikolov. Distributed representations of sentences and documents. arXiv preprint:1405.4053, 2014.
[7] Y. Lv and C. Zhai. Positional relevance model for pseudo-relevance feedback. In SIGIR ’10, pages 579–586, 2010.
[8] Y. Lv and C. Zhai. Revisiting the divergence minimization feedback model. In CIKM ’14, pages 1863–1866, 2014.
[9] Y. Ma, J. Nie, H. Wu, and H. Wang. Opening machine translation black box for cross-language information retrieval. In CIRT ’12, pages 467–476, 2012.
[10] T. Mikolov, V. Q. Le, and I. Sutskever. Exploiting similarities among languages for machine translation. arXiv preprint:1309.4168, 2013.
[11] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In Adv. in Neur. Inf. Proc. Sys., pages 3111–3119, 2013.
[12] C. Moza and B. J. Dorr. Iterative Translation Disambiguation for Cross-language Information Retrieval. In SIGIR ’05, pages 520–527, 2005.
[13] I. Vulic and M. Moens. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In SIGIR ’15, pages 363–372, 2015.
[14] I. Vulic, W. D. Smet, J. Tang, and M. Moens. Probabilistic topic modeling in multilingual settings: An overview of its methodology and applications. IPRI, 51(1):111–147, 2015.