Learning Multilingual Embeddings for Cross-Lingual Information Retrieval in the Presence of Topically Aligned Corpora

Mitodru Niyogi  
Computer Science and Engineering,  
Indian Institute of Technology,  
Kanpur, India  
mitodru@cse.iitk.ac.in

Kripabandhu Ghosh  
Computer Science and Engineering,  
Indian Institute of Technology,  
Kanpur, India  
kripa@cse.iitk.ac.in

Arnab Bhattacharya  
Computer Science and Engineering,  
Indian Institute of Technology,  
Kanpur, India  
arabb@cse.iitk.ac.in

ABSTRACT
Cross-lingual information retrieval is a challenging task in the absence of aligned parallel corpora. In this paper, we address this problem by considering topically aligned corpora designed for evaluating an IR setup. To emphasize, we neither use any sentence-aligned corpora or document-aligned corpora, nor do we use any language specific resources such as dictionary, thesaurus, or grammar rules. Instead, we use an embedding into a common space and learn word correspondences directly from there. We test our proposed approach for bilingual IR on standard FIRE datasets for Bangla, Hindi and English. The proposed method is superior to the state-of-the-art method not only for IR evaluation measures but also in terms of time requirements. We extend our method successfully to the trilingual setting.

ACM Reference format:
Mitodru Niyogi, Kripabandhu Ghosh, and Arnab Bhattacharya. 2016. Learning Multilingual Embeddings for Cross-Lingual Information Retrieval in the Presence of Topically Aligned Corpora. In Proceedings of ArXiV, 2018, April, 5 pages.

DOI: XXX

1 INTRODUCTION
Cross-lingual information retrieval, where multiple languages are used simultaneously in an information retrieval (IR) task, is an important area of research. The increasing amount of non-English data available through the Internet and processed by several modern age IR/NLP (natural language processing) tasks has magnified the importance of cross-lingual IR manifold. In particular, we address the general ad-hoc information retrieval task where the query is in any of the n languages, and retrieval can be from any of the remaining languages. In countries such as India where multiple languages are used officially and regularly by a large amount of computer-educated citizens, the above task is particularly important, and can be a game changer for many of the digital initiatives that governments across the world are actively promoting.

Such queries can be quite common. For example, in nationwide events such as general elections, or an emergency situation, a sports event, etc., queries like “How many seats have party X won in state Y?” are quite common and will be issued in several languages. The proposed system should be able to retrieve the answer from documents written in any language.

Most of the previous work on cross-lingual IR [1, 5] require sentence-aligned parallel data and other language specific resources such as dictionaries. Vulic et al. [4] removed this extremely constraining requirement and learned bilingual word embedding using only document-aligned comparable corpora. However, such aligned corpora is not always readily available and need considerable effort to be built. Resource-poor languages such as the Indian languages specifically suffer from this setback.

To this end, we present a multi-lingual setup where we build a cross-lingual IR system that requires no such aligned corpora or language specific resources. It automatically learns cross-lingual embeddings using merely TREC-style test collections. We also propose to build a multi-lingual embedding on the same setup. This eliminates the requirement of building embeddings for collection pairs in a cross-lingual retrieval paradigm as well as the need to train bilingual embedding for all possible language pairs. Instead, this single multi-lingual embedding will leverage automatic cross-lingual retrieval between any two pairs of languages. The proposed setup is particularly useful in online situations in multi-lingual countries such as India.

To the best of our knowledge, this is the first cross-lingual IR work that works on more than 2 languages directly and simultaneously without targeting each pair separately.

Our proposed method yields considerable improvements over Vulic et al. [4] in the bilingual setting on standard Indian language test collections. We further demonstrate the efficacy of our method by using a trilingual embedding.

2 METHODOLOGY
A traditional ad-hoc information retrieval test collection (in the binary relevance setup) is defined as $C = \{D, Q, R\}$ where $D$ is a set of documents, $Q$ a set of queries, and relevance $R$ is a mapping defined as $R : Q \times D \rightarrow \{0, 1\}$. For each document $d \in D$ and query $q \in Q$, $R(d, q) = 1$ if $d$ is relevant for $q$, and $0$ otherwise.

In the multi-lingual retrieval setup, we consider a set $L$ of $n$ languages ($L_1, L_2, \ldots, L_n$). Corresponding to each language, there is a test collection $C_i = \{D_i, Q_i, R_i\}$ such that documents in $D_i$ and queries in $Q_i$ are in language $L_i$. Additionally, the queries in $Q_i$ are translations of each other. In other words, query $q_{ij} \in Q_k$ (the $j$th query in $Q_k$) is the translation in language $L_k$ of the query $q_{ij} \in Q_l$ ($i \neq k$) in language $L_l$. Each set $Q_i$ has exactly $m$ queries.

Note that since queries are generally very short phrases and/or just a set of words, finding translated queries in multiple languages is a much easier task.
Cross-lingual topical relevance hypothesis: Let $D_{kj}^R$ denote the set of documents in $D_k$ that are relevant to $q_{kj}$. We hypothesize that this set is topically similar to the set $D_{kj}^R(i \neq k)$ that is relevant to $q_{lj}$ where $q_{kj} \neq q_{lj}$ are translations of each other in languages $L_k$ and $L_l$ respectively. Note that the documents in $D_{kj}^R$ and $D_{lj}^R$ are not translations of each other but are supposed to be similar as they are relevant to the same “information need” expressed in the two languages $L_k$ and $L_l$ respectively. The sizes of the two sets $D_{kj}^R$ and $D_{lj}^R$ need not be equal as well.

This notion of topical relevance can be extended to the multilingual setting where for $q_{kj}$’s that are translations of each other, the corresponding relevant sets $\{D_{kj}^R\}$ are considered topically similar to each other.

We next describe our proposed method. We first create a multilingual vector space embedding for all the $k$ languages together, and then use that to generate cross-lingual queries that enable retrieval between any two languages in this multilingual setup.

### 2.1 Multilingual Embedding Construction

In this section, we describe our algorithm for creating a multilingual embedding from all the corpora $C_l$ designed on the cross-lingual topical relevance hypothesis for the set of queries $Q_l$. The algorithm is applied for both the training and testing set of queries.

**Training queries:** We describe the algorithm for creating multilingual embedding for one training query only, which we will generalize for all the training queries thereafter.

Let $q_{lj}$ be the query in language $L_l$ for which the number of relevant documents is the least among all the corresponding queries in the other languages $L_k$, i.e., $|D_{kj}^R| \leq |D_{lj}^R|$ for all $k \neq l$.

For each document $d_{kj} \in D_{kj}^R$, we choose $d_{kj}$ randomly without replacement from $D_{kj}^R$. Let $t_{min} = \min(t_k = |d_{kj}|)$ be the minimum document length measured as number of terms among all the $k$ documents. Let $d_{min} = \text{the corresponding document and } k_{min} = \arg\min \{t_k\}$ be the language index of $d_{min}$. Let $n_k^{\text{norm}} = \lfloor \frac{t_k}{t_{\text{min}}} \rfloor$. We create a multi-lingual document $D_{kj}^{\text{mult}}$ comprising of all the $n$ languages as follows. We start with an empty $D_{kj}^{\text{mult}}$. For each term in $d_{min}$, we append the term to $D_{kj}^{\text{mult}}$. Then, we select the next $n_k^{\text{norm}}$ terms from $d_{kj}$ (for all $k \neq k_{min}$) and append them to $D_{kj}^{\text{mult}}$. Thus, we create $D_{kj}^{\text{mult}}$ by placing each instance of a word of the document which has the least number of terms followed by the terms of the other documents (of the remaining languages) in the relative ratios of their document lengths.

This method of shuffling creates a better mix of the words from the multiple languages, thereby enabling a better learning of the embedded vector space.

For example, let $d_1 = (t_1, t_2)$ and $d_2 = (w_1, w_2, w_3, w_4, w_5)$ be two documents with terms $t_1$’s and $w_j$’s respectively. Suppose we are looking to create a bilingual embedding document from $d_1$ and $d_2$. Clearly, $t_{\text{min}} = 2$. Then, $n_j^{\text{norm}} = \lfloor \frac{t_j}{2} \rfloor = 3$. That is, for each term in $d_1$ there will follow 3 terms of $d_2$ until all the terms of $d_1$ (and $d_2$) are considered. Therefore, $D_{kj}^{\text{mult}} = (t_1, w_1, w_2, w_3, t_2, w_4, w_5)$.

This algorithm when run for all the training queries produce the set $D_{\text{mult}}^{\text{train}}$.

**Test queries:** To address the words that are missing in the training set of relevant documents (e.g., proper nouns present exclusively in the test queries), we run the same algorithm by replacing $D_{kj}^R$ with $PseudoRel$ created by running each test query $q_{kj}$ on the corresponding collection $D_k$. We select the top $k$ documents, and run this for all the test queries to produce $D_{\text{mult}}^{\text{test}}$.

**Final embedding:** Finally, we create $D_{\text{final}}^{\text{mult}}$ by taking the union of $D_{\text{train}}^{\text{mult}}$ and $D_{\text{test}}^{\text{mult}}$. We then running Word2Vec [2] on $D_{\text{final}}^{\text{mult}}$ to get the final multilingual word vector embeddings.

### 2.2 Cross-Lingual Query Generation

The main IR task is to perform cross-language information retrieval with a query $q_{kj}$ in language $L_k$ on any of the document collections $D_l$ in any other language $L_l$, $l \neq k$. (We exclude the monolingual setup $l = k$.) The language of the query, $L_k$, is referred to as the source language and the language of the document collection, $L_l$, is the target language.

The aim of cross-lingual query generation is to generate a target query version, $q'_{lj}$, in language $L_l$, of $q_{kj}$. Note that this is different from $q_{lj}$, which is the actual query in language $L_l$. In the results section, for reference, we will also state the results using $q_{lj}$ as the baseline monolingual setting, which is an expected upper-bound of performance.

**Query generation procedure:** Let $V_l$ be the vocabulary (set of unique terms) of $D_l$. We construct a vector $\tilde{q}_{kj}$ by aggregating the vectors corresponding to the constituent terms of $q_{kj}$ in the multilingual embedding space. For each vector in $\tilde{q}_{kj}$, we capture its top-$\tau$ semantically closest term vectors from $V_l$ in the multi-lingual embedding space. The semantic closeness is measured by cosine similarity. These closest term vectors are aggregated to form the target query vector $q'_{lj}$.

Thereafter, we perform cross-lingual retrieval with $q'_{lj}$ on $D_l$.

The overall scheme is shown in Figure 1. During training, for each of the queries, we consider the relevant documents $(R_{kj})$ from the corresponding corpus, and shuffle them to form a multi-lingual shuffled document. The multi-lingual document is further enriched by the pseudo-relevant documents of the test queries. A common word embedding is learned from this set of multi-lingual documents. During testing, cross-lingual retrieval is done by generating query $q'$ from the source query $q_k$ using the common word embedding space.
Cross-Lingual IR in Presence of Topically Aligned Corpora

3 EXPERIMENTS

3.1 Setup

Datasets: We use FIRE (http://fire.irs.i.res.in/fire/static/data) cross-lingual datasets in English, Hindi and Bangla (details given in Table 1). The documents were collected from the following newspapers: The Telegraph (http://www.telegraphindia.com) for English, Dainik Jagran (http://www.jagrancan) and Amar Ujala (http://www.amarujala.com) for Hindi, and Anandabazar Patrika (http://www.anandabazar.com) for Bangla. Query sets were created such that queries with the same identifier are translations of each other. For each language and collection, we choose randomly 10 queries for testing. The rest are used for training in a 5-fold cross validation manner.

For retrieval, only the title field of the queries were used. Stop-word removal was done. We use the default Dirichlet Language Model implemented in Terrier IR toolkit (http://terrier.org/) for all our retrieval experiments.

Baseline: We compare our method for cross-lingual IR with bilingual embeddings with Vulic et al. [4]. The shuffling code used is obtained from the authors.

Monolingual: In the monolingual setup, the results when the actual target language queries are used for retrieval on the target set sets the upper bound of performance that can be achieved with a multi-lingual setup.

3.2 Training

The Gensim implementation for Word2Vec (https://radimrehurek.com/gensim/models/word2vec.html) was used. The skip-gram model was used for the training using the following parameters: (i) vector dimensionality: 100, (ii) learning rate: 0.01, (iii) min word count: 1. The context window size was varied from 5 to 50 in intervals of 5. For bilingual embedding, window size 25 produced the best results on the training set and was subsequently used on the test queries, while for trilingual embedding, the best window size was 50. The parameters \(k\) and \(s\) were tuned on the training set over the values \{5, 10, 15, 20\} and \{5, 10, 15\} respectively.

3.3 Results

To assess quality, we report the Mean Average Precision (MAP), R-Precision (R-Prec) and Binary Preference (BPref).

We report our retrieval results in Table 3 and Table 4. We uniformly use the cross-lingual retrieval convention source language \(\rightarrow\) target language. For example, B\(\rightarrow\)E indicates that Bangla is the source language while English is the target language.

Bilingual Embeddings: Table 3 shows the results for bilingual retrieval, i.e., when the embedding space is built using only 2 languages. For all the language pairs, the proposed method outperforms Vulic et al. [4] significantly; the differences are statistically significant at 5% level of confidence (\(p < 0.05\)) by Wilcoxon signed-rank test [3]. We have reported the Monolingual results that does not require any cross-lingual IR as an upper bound of performance. Interestingly, our proposed method produces comparable MAP results for H\(\rightarrow\)B (FIRE 2010). It exhibits better BPref than Monolingual B\(\rightarrow\)B for H\(\rightarrow\)B (FIRE 2010) and the difference is statistically significant at 5% level of confidence by Wilcoxon signed-rank test. It is also comparable with Monolingual H\(\rightarrow\)H for B\(\rightarrow\)H (FIRE 2010), with Monolingual B\(\rightarrow\)B for E\(\rightarrow\)B, H\(\rightarrow\)B (FIRE 2011) and with Monolingual E\(\rightarrow\)E for H\(\rightarrow\)E (FIRE 2012). While evaluating with R-Prec, H\(\rightarrow\)B (FIRE 2010) is slightly better than Monolingual B\(\rightarrow\)B. This shows that the proposed method produces competitive performance even when compared with a strong baseline like Monolingual.

Time requirements: The time requirements comparison with Vulic et al. [4] is reported in Table 2. Our pre-retrieval time involves indexing time using Terrier (http://terrier.org/) and cross-lingual query generation time. Pre-retrieval time for Vulic is the time taken to create the document vectors for all the documents in a corpus. Retrieval time for us is the one taken by Terrier to produce the ranked list for only the test queries. Retrieval time for Vulic comprises of calculating the cosine score between the query vectors of the test queries and all the documents in the collection followed by sorting the documents of the whole collection in the decreasing order of this score for each query. Our proposed method clearly outperforms Vulic in terms of time requirements.

Trilingual Embeddings: We report the retrieval performance of the trilingual setting in Table 4. We chose not to compare with Vulic et al. [4] any further since we have already established our superiority over the latter in the bilingual setting. For FIRE 2010, our proposed method produces superior performance in both MAP and BPref over Monolingual B\(\rightarrow\)B for both E\(\rightarrow\)B and H\(\rightarrow\)B and the differences are statistically significant at 5% level of confidence by Wilcoxon signed-rank test. Using R-Prec, for FIRE 2010, E\(\rightarrow\)B is considerably better than Monolingual B\(\rightarrow\)B (\(p < 0.05\) by Wilcoxon signed-rank test). For FIRE 2011, our proposed method produces better results in BPref over Monolingual B\(\rightarrow\)B for E\(\rightarrow\)B and over Monolingual H\(\rightarrow\)H for E\(\rightarrow\)H. This shows that our proposed method is able to maintain its performance when compared with Monolingual even in a trilingual setting.

| Language | #Docs | #Queries | Mean rel docs per query |
|----------|-------|----------|-------------------------|
| English  | 1,25,586 | 50       | 13.06                   |
| Hindi    | 1,49,482 | 50       | 18.30                   |
| Bangla   | 1,23,047 | 50       | 10.02                   |

| Language | #Docs | #Queries | Mean rel docs per query |
|----------|-------|----------|-------------------------|
| English  | 89,286 | 50       | 55.22                   |
| Hindi    | 3,31,599| 50       | 57.70                   |
| Bangla   | 3,77,104| 50       | 55.56                   |

| Language | #Docs | #Queries | Mean rel docs per query |
|----------|-------|----------|-------------------------|
| English  | 89,286 | 50       | 70.78                   |
| Hindi    | 3,31,599| 50       | 46.18                   |
| Bangla   | 3,77,111| 50       | 51.62                   |

Table 1: Datasets.

| Method      | Language | Pre-retrieval time | Retrieval time |
|-------------|----------|--------------------|----------------|
| Proposed    | English  | 175.67s            | 5.51s          |
| Vulic       | English  | 105.59s            | 13.37s         |
| Proposed    | Hindi    | 2066.27s           | 0.92s          |
| Vulic       | Hindi    | 2620.04s           | 36.19s         |
| Proposed    | Bangla   | 1,027.92s          | 3.60s          |
| Vulic       | Bangla   | 1,5220.24s         | 118.86s        |

Table 2: Time requirements, averaged over three datasets.
### Table 3: Bilingual Retrieval. (Proposed method is always statistically significantly better than Vulic et al.\cite{Vulic}, $p < 0.05$.)

| Method | FIRE 2010 | FIRE 2011 | FIRE 2012 |
|--------|-----------|-----------|-----------|
|        | MAP | R-Prec | BPref | MAP | R-Prec | BPref | MAP | R-Prec | BPref |
|        |     |       |       |     |       |       |     |       |       |
| Monolingual | E→E | 0.4265 | 0.4044 | 0.3785 | 0.2836 | 0.3098 | 0.3528 | 0.4868 | 0.4785 | 0.4507 |
| Proposed | B→E | 0.1761 | 0.2041 | 0.2297 | 0.1148 | 0.1164 | 0.2204 | 0.2890 | 0.2899 | 0.3418 |
| Vulic | E→E | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| Monolingual | B→B | 0.3554 | 0.2951 | 0.2593 | 0.2127 | 0.2677 | 0.2164 | 0.3093 | 0.3188 | 0.3203 |
| Proposed | E→B | 0.1964 | 0.2429 | 0.2017 | 0.1302 | 0.1797 | 0.2105 | 0.2114 | 0.2409 | 0.2522 |
| Vulic | E→B | 0.0000 | 0.0000 | 0.0016 | 0.0000 | 0.0000 | 0.0032 | 0.0000 | 0.0000 | 0.0032 |
| Monolingual | H→H | 0.3169 | 0.2872 | 0.2691 | 0.2408 | 0.2756 | 0.2637 | 0.4221 | 0.4407 | 0.4226 |
| Proposed | E→H | 0.1497 | 0.1663 | 0.1681 | 0.1526 | 0.1806 | 0.2038 | 0.3094 | 0.3093 | 0.3325 |
| Vulic | B→H | 0.0000 | 0.0000 | 0.0030 | 0.0000 | 0.0000 | 0.0089 | 0.0000 | 0.0000 | 0.0089 |

### Table 4: Trilingual Retrieval.

| Type | Source language | Target language | Source query (with English translation) | Generated target query words (with English translation) |
|------|----------------|----------------|----------------------------------------|--------------------------------------------------------|
| Bilingual | Hindi | English | संजय दत्त का आश्रयसंगीत (surrender of sanjay dutt) | dutt sanjay sanjays munnabhai convicts namrata salem ak |
| Bilingual | Bangla | Hindi | স্তন্য রক্ত পরিনতি চিকিৎসা চিকিৎসা (cervical cancer awareness treatment vaccine) | cervical hpv human papillomavirus, infection, pregnant, silvia |
| Trilingual | English | Bangla | death of yasser arafat | arafat yasser ramallah palestine suha kurei plo |
| Trilingual | Hindi | Bangla | তাজ মহলের প্রতি নিষেধাজ্ঞা (taj mahal controversy) | taj mahal controversy, palestralia, world heritage site, taj mahal controversy, wakf archaeological survey link maintenance sunni |

### 3.4 Analysis

Figure 2 shows some example queries generated by our proposed method using bilingual and trilingual embeddings. For the query surrender of sanjay dutt (on the conviction of Bollywood actor Sanjay Dutt with relation to terrorist attack in Bombay in 1993), the generated query contains important words such as sanjay, dutt, salem (Abu Salem, a terrorist), ak (AK-47, a firearm), munnabhai (a popular screen name of Sanjay). The generated query for cervical cancer awareness treatment vaccine contains useful terms like cervical, hpv (Human papillomavirus), infection, pregnant, silvia (Silvia De Sanjose, a leading researcher in Cancer Epidemiology). The generated query for death of Yasser Arafat contains the terms arafat, yasser, ramallah (the headquarters of Yasser), palestine, suha (Suha Arafat, Yasser Arafat’s wife) and plo (Palestine Liberation Organization). The generated query for taj mahal controversy (regarding if Taj Mahal is a Waqf property as claimed by Uttar Pradesh Sunni Waf Board and subsequent statements by the Archaeological Survey of India) contains vital terms such as taj, mahal, wafk, archaeological and sunni. These examples clearly portray the effectiveness of our target query generation.

### 4 CONCLUSIONS

In this paper, we have proposed a cross-lingual IR setup in the absence of aligned comparable corpora. Our method used a common
embedding for all the languages and produced better performance than the closest state-of-the-art. In future, we would like to experiment with other embedding methods.

REFERENCES

[1] K. M. Hermann and P. Blunsom. 2014. Multilingual Models for Compositional Distributed Semantics. In ACL. 58–68.
[2] T. Mikolov, W. T. Yih, and G. Zweig. 2013. Linguistic Regularities in Continuous Space Word Representations. In NAACL HLT.
[3] S. Siegel. 1956. Nonparametric Statistics for the Behavioral Sciences. McGraw-Hill.
[4] I. Vulić and M.-F. Moens. 2015. Monolingual and Cross-Lingual Information Retrieval Models Based on (Bilingual) Word Embeddings. In SIGIR. 363–372.
[5] J. Wang and D. W. Oard. 2012. Matching Meaning for Cross-language Information Retrieval. Inf. Process. Manag. 48, 4 (2012), 631–653.