Using Communities of Words Derived from Multilingual Word Vectors for Cross-Language Information Retrieval in Indian Languages

PAHELI BHATTACHARYA, PAWAN GOYAL, and SUDESHNA SARKAR, Indian Institute of Technology Kharagpur

We investigate the use of word embeddings for query translation to improve precision in cross-language information retrieval (CLIR). Word vectors represent words in a distributional space such that syntactically or semantically similar words are close to each other in this space. Multilingual word embeddings are constructed in such a way that similar words across languages have similar vector representations. We explore the effective use of bilingual and multilingual word embeddings learned from comparable corpora of Indic languages to the task of CLIR.

We propose a clustering method based on the multilingual word vectors to group similar words across languages. For this we construct a graph with words from multiple languages as nodes and with edges connecting words with similar vectors. We use the Louvain method for community detection to find communities in this graph. We show that choosing target language words as query translations from the clusters or communities containing the query terms helps in improving CLIR. We also find that better-quality query translations are obtained when words from more languages are used to do the clustering even when the additional languages are neither the source nor the target languages. This is probably because having more similar words across multiple languages helps define well-defined dense subclusters that help us obtain precise query translations.

In this article, we demonstrate the use of multilingual word embeddings and word clusters for CLIR involving Indic languages. We also make available a tool for obtaining related words and the visualizations of the multilingual word vectors for English, Hindi, Bengali, Marathi, Gujarati, and Tamil.

CCS Concepts: • **Information systems** → **Information retrieval**; **Information retrieval query processing**;

Additional Key Words and Phrases: Multilingual word embeddings, clusters, cross-language information retrieval, multilingual visualization

ACM Reference format:
Paheli Bhattacharya, Pawan Goyal, and Sudeshna Sarkar. 2018. Using Communities of Words Derived from Multilingual Word Vectors for Cross-Language Information Retrieval in Indian Languages. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 18, 1, Article 1 (December 2018), 27 pages.
https://doi.org/10.1145/3208358

Authors' addresses: P. Bhattacharya, P. Goyal, and S. Sarkar, Department of Computer Science and Engineering, Indian Institute of Technology Kharagpur, West Bengal, India; emails: paheli.bh@iitkp.ac.in, {pawang, sudeshna}@cse.iitkgp.ernet.in.

ACM 2375-4699/2018/12-ART1 $15.00
https://doi.org/10.1145/3208358

1 INTRODUCTION

English has been the dominant language of the web for a long time, but with the rising popularity of the web, there has been a substantial amount of content in multiple languages. With the availability of multilingual content, relevant and adequate information may not always be available in only one particular language but may be spread across other languages. This gives rise to the necessity of cross-language information retrieval (CLIR), where the query and the documents do not belong to a single language only. Specifically, in CLIR, the user query is in a language different than the collection. One of the main motivations behind CLIR is to gather a lot of knowledge from a variety of knowledge bases that are in the form of documents in various languages, helping a diverse set of users, who can provide the queries in the language of their choice.

Since the language of the query is different from the language of the documents in CLIR, a translation phase is necessary.

Translating documents is a tedious task. The general standard is to translate the query. We also follow the query translation approach for CLIR. Common or popular approaches for query translation for CLIR include leveraging bilingual or multilingual dictionaries (Hull and Grefenstette 1996; Pirkola 1998; Ballesteros and Croft 1996; Levow et al. 2005), statistical machine translation (SMT) systems (Schamoni et al. 2014; Türe et al. 2012a, 2012b; Sokolov et al. 2014), transliteration-based models (Udupa et al. 2009; Chinnakotla et al. 2007), graph-based models (Franco-Salvador et al. 2014), and online translators (Hosseinzadeh Vahid et al. 2015).

Each of these approaches has its own advantages and disadvantages. For instance, SMT systems require parallel corpora, and for Indian languages where such resources are scarce, SMTs are not very suitable. The dictionary-based approaches require substantial word-pair translations and suffer from coverage issues and data sparsity problems. We study the effectiveness of word embeddings in such a scenario where we want to have good-quality translations that can improve CLIR performance in spite of having a scarcity of data-aligned resources.

Word embeddings or word vectors are the representation of words as dense vectors of real numbers in a low-dimensional space. In the distributed space defined by the vector dimensions, syntactically and semantically similar words fall closer to each other. Good-quality word embeddings have become an important part of Natural Language Processing in recent years. Word embeddings have achieved tremendous success in natural language processing applications. It is now being used in a wide variety of tasks like sentiment analysis, named entity recognition, dependency parsing, part-of-speech tagging, and so forth.

For applications involving more than one language, either in the form of translation or cross-lingual lexical transfer, embeddings trained in a monolingual environment cannot be used directly. This gives rise to the necessity of constructing multilingual word embeddings.

Different words in different languages may share a similar sense. We may map words into a common vector space such that similar words across languages have similar vector representations. This will lead to having a unified representation of words from various languages. These representations are called multilingual word vector representations (Mikolov et al. 2013b). These embeddings have also become an integral part in many multilingual tasks, and it has been observed that approaches utilizing word embeddings prove beneficial in many applications. Multilingual word embeddings have been applied to cross-language document classification tasks (Minh-Thang Luong and Manning 2015; Hieu Pham and Manning 2015; Hermann and Blunsom 2013; Soyer et al. 2014; Chandar et al. 2014; Blunsom and Hermann 2014), cross-lingual dependency parsing (Huang et al. 2015), finding syntactic and semantic relations (Faruqui and Dyer 2014), part-of-speech (POS) tagging (Gouws and Søgaard 2015), and others.

Our motivation behind this work is to bring word representations for major Indian languages into a common space. Essentially, we want to have a unified representation of words across languages that preserves similarity across languages.

With a view to evaluate these multilingual word embeddings, we choose the task of cross-language information retrieval. We find that word embeddings serve as a potential tool for bridging the gap between the scarcity of aligned resources and good-quality translations. We find that word embeddings could easily find translations for out-of-vocabulary words like "*kaiMsara*" (meaning cancer), which are hard to obtain by dictionary-based or parallel-corpora-based methods. However, we observe that words that are not very relevant to the source language word also come up as potential translations. For instance, for the word "*desha*" (meaning country), although correct translations like "country" and "democracy" were obtained using word embeddings, irrelevant words like "aspiration" and "kind" also showed up as potential translations. Inclusion of such non-related words in a query greatly harms the IR performance.

To address this problem, we propose to use multilingual clustering. In multilingual clustering, words from the same language as well as across languages that are more likely to represent similar concepts fall into the same group. We use multilingual embeddings to build these clusters. When multilingual clusters were used, candidate English translations besides "country" and "democracy" for our running example "*desha*" were "nation" and "cities". We have also observed that if we use more languages for building the clusters, CLIR performance is improved. That is, if words from intermediate languages that are not the source or the target language are present in the cluster, the query translations are better and CLIR performance is improved. This idea is along the lines of Dabre et al. (2015), who show that assistive or pivot languages greatly help in machine translation.

Our proposed method has shown significant improvements over the dictionary-based method, a transliteration-based model, and Google Translate for the CLIR task.

We have constructed multilingual word embeddings for English and major Indic languages, namely, Hindi, Bengali, Marathi, Gujarati, and Tamil. That is, the vectors of words from all these languages are projected in the same space so that they can be used for various cross-lingual experiments. We have evaluated the quality of the multilingual word embeddings in CLIR.

We have also designed a multilingual visualization tool where these word representations can be visualized. We have provided a service through which one can translate a word in a language to multiple other languages at the same time and visualize their representations in a two-dimensional space. The word embeddings and the tool are available at https://github.com/paheli/indic-viz.

The work presented here is an extension of Bhattacharya et al. (2016). The added contributions of the article with respect to the workshop version are as follows: (1) We have explored in detail how multilingual clustering can benefit CLIR. We have included four languages—English, Hindi, Bengali, and Tamil—in the clusters. We analyze inclusion of words from more languages in the clusters and its effect in CLIR. (2) We have provided a service for translation and visualization of query terms using multilingual embeddings.

The rest of the article is organized as follows. In the next section, we give a brief overview of the related work in the fields of multilingual embeddings and cross-language information retrieval. In Section 3, we provide the motivation for our approach. Section 4 describes the settings and datasets we have used throughout our experiments. In Section 5, we show how the word embeddings perform in CLIR using a naive approach and bring out the drawbacks. In the subsequent section, we propose a clustering-based method to alleviate the problems of the naive approach. In Section 6, we construct multilingual word embeddings for major Indian languages and evaluate their quality in the task of CLIR. In Section 7, we introduce the multilingual visualization tool for Indic languages. We describe the detailed work in the Online Appendix. Finally, we conclude with some future extensions in Section 8.

2 RELATED WORK

2.1 Cross-Lingual and Multilingual Vector Representations in Distributed Space

We look at some related work in the field of vector representation methods in cross-lingual and multilingual spaces. A detailed and comprehensive list of methods has been presented in the survey paper by Ruder (2017).

Projection-based approaches like canonical correlation analysis (CCA) (Faruqui and Dyer 2014) and linear-regression-based approaches (Mikolov et al. 2013b) involve learning word embeddings for the languages a priori. CCA maps words from two different languages into a common, shared space where the embeddings across the two languages are maximally correlated. Mikolov et al. (2013b) show how a linear projection from a source language to a target language can be used to get bilingual word embeddings. Both these approaches require aligned data, for example, a bilingual dictionary besides monolingual corpora for each of the languages.

Adversarial auto-encoders have been used in Barone (2016), which do not require a parallel resource to learn the embeddings. The auto-encoder is used to reconstruct the source embeddings, while the discriminator is trained to differentiate the projection of the source language embedding from the actual target embedding.

Vulić and Korhonen (2016) argue that methods that use dictionaries do not pay much attention to their quality. To address this problem, they initially learn a shared bilingual space using an existing cross-lingual model and use this to form a dictionary. These dictionary translations are used to learn the embeddings following the idea of Mikolov et al. (2013b).

Matrix algebra decomposition approaches were leveraged by Huang et al. (2015), where they construct translation-invariant word embeddings using latent semantic analysis (LSA) on a multilingual co-occurrence matrix X. It finds a decomposition that can simultaneously explain the original matrix X and its various translations. The translations are in the form of a word dictionary matrix, with all the words from the two languages in the rows and columns and a context dictionary matrix with all context words from the two languages in the rows and columns.

Constructing multilingual word embeddings depending on a **particular task** has been proposed in Gouws and Søgaard (2015). They use a task-specific dictionary, that is, a list of word pairs that are equivalent in some respect, depending on the task. Using a nonparallel corpora, given a sentence in one language, for each word in the sentence, equivalent words are substituted in its place. Then word vectors are learned on this new corpora.

Joint-learning approaches using parallel corpora like bilingual bag-of-words (BilBOWA) (Gouws et al. 2015) use monolingual datasets coupled with sentence-aligned parallel data to learn word embeddings. They utilize the SkipGram model of *word2vec* to learn the monolingual features and a sample bag-of-words technique for each parallel sentence as the cross-lingual objective. Chandar et al. (2014) show that by learning to reconstruct the bag-of-words representations of aligned sentences, within and between languages, high-quality word representations can be learned. They use an auto-encoder for this purpose. Given an alignment link between a word w_1 in a language l_1 and a word w_2 in another language l_2, Luong et al. (2015) use the word w_1 to predict the neighbors of the word w_2 and vice versa. Klementiev et al. (2012) induce distributed representations for a pair of languages jointly. They treat it as a multitask learning problem where each task corresponds to a single word and task relatedness is derived from co-occurrence statistics in bilingual parallel data. Their model assumes word alignments to be available. Soyer et al. (2014) differ from others in the way that they consider phrases instead of only words and sentences. They exploit the observation that phrases are more closely related to their subphrases than to other randomly assigned phrases. In contrast to the Skip-Gram model (Mikolov et al. 2013a), where word vectors are similar if they appear as the center of similar word windows, this approach encourages word vectors to be similar if they tend to be embedded in similar phrases. Duong et al. (2016) use a high-coverage dictionary in an expectation-maximization (EM)-style training algorithm over monolingual corpora in two languages to build bilingual embeddings. Similar to Gouws and Søgaard (2015), they replace each word in the monolingual corpus with its translation. Solving the problem of selecting the most probable translation is achieved using EM. They train the vectors using a joint learning approach on the basis that words and their translations should appear in very similar contexts.

Levy et al. (2017) show that for learning cross-lingual embeddings using parallel data, the *SentenceID* feature plays a very important role. They compare the state-of-the-art methods for learning cross-lingual embeddings using a parallel corpus with traditional word alignment algorithms and propose that the Dice aligner can be a potential baseline for comparing the algorithms.

The **Compositional Vector Model (CVM)** introduced by Blunsom and Hermann (2014) learns semantic representations of larger syntactic units (like documents) given the semantic representations of their constituents (the sentences). Hermann and Blunsom (2013) utilize both the compositional semantic representations and shared word-level representations across languages to induce a method for learning word vectors in a multilingual setup. They train a model that learns to assign similar embeddings to aligned sentences and dissimilar ones to the unaligned sentences.

A **graph-based approach** has been leveraged in Soricut and Ding (2016). They present a framework of neural-network-based multilingual embedding models using multigraphs. It combines the existing SkipGram model (Mikolov et al. 2013a), the dependency embedding model of Levy and Goldberg (2014), and the BiSkip model (Luong et al. 2015). It uses sentence-parallel data with word alignments and parsing information available and builds a multigraph using these. It then builds SkipGram-based word embeddings on the multigraph with the contexts described by multiple attributes of alignment and dependency information.

Clustering approaches for multilingual word embeddings were introduced in Ammar et al. (2016). They construct massively multilingual word embeddings using the following approaches: multilingual cluster (multiCluster), where the problem of obtaining multilingual word embeddings is divided into two subproblems, where $E = E_{embed} \circ E_{cluster}$. $E_{cluster}$ maps words to clusters and E_{embed} assigns a vector to each cluster. Clusters are the connected components in a graph where nodes are (language-surface form) pairs and edges correspond to translation entries. Distributional similarities of the clusters in monolingual corpora from all languages are used to estimate an embedding for each cluster. They assign IDs to the clusters and replace each word in each monolingual corpus with the corresponding cluster ID. The resulting corpus consists of multilingual cluster ID sequences. Then they apply the skipgram model of *word2vec*. They also apply CCA (Faruqui and Dyer 2014), the BiSkip model of Luong et al. (2015), and translation-invariant word embeddings (Huang et al. 2015).

2.2 Evaluation

The performance of the word embeddings evaluated for various tasks are summarized as follows:

- Cross-Language Document Classification (CLDC) Task: Train the classifier to classify documents in one language and then apply to documents in a different language (Gouws et al. 2015; Minh-Thang Luong and Manning 2015; Hieu Pham and Manning 2015; Hermann and Blunsom 2013; Soyer et al. 2014; Chandar et al. 2014; Blunsom and Hermann 2014)
- Word Similarity Task: Monolingual word similarity (Faruqui and Dyer 2014; Huang et al. 2015; Qiu et al. 2014)
- Cross-Lingual Dependency Parsing: Train a dependency parser for one language using its word vectors and test it on another language using its data and word vectors (Huang et al. 2015)

- Identifying POS Tags and Translation Equivalent Classes (Gouws and Søgaard 2015)
- Analogical Reasoning Task (Qiu et al. 2014)
- Word Translation Task (Gouws et al. 2015; Mikolov et al. 2013b)

Upadhyay et al. (2016) compare the BiSkip model (Minh-Thang Luong and Manning 2015), BiCVM model (Blunsom and Hermann 2014), CCA-based approach (Faruqui and Dyer 2014), and BiVCD model (Vulić and Moens 2015) by applying them to the monolingual word similarity task for English, cross-lingual dictionary induction, cross-lingual document classification, and cross-lingual dependency parsing.

2.3 Cross-Lingual Information Retrieval

Information retrieval is the task of retrieving relevant documents with respect to a query. There are two variants of information retrieval. One is where both the query and the documents are in the same language. This is called monolingual information retrieval. The other is cross-language information retrieval, where the query is in one language and the documents are in another language.

Different researchers have tried viewing CLIR from various aspects. To start with, Pirkola (1998) use **dictionary-based translation** techniques for information retrieval. They use two dictionaries, one in which general translation of a query term is present, and the other in which domain-specific translation of the query term is present. The document collection is a subset of the TREC collection and queries were regarding TREC's health-related topics. Levow et al. (2005) discuss the key issues in dictionary-based CLIR. They have shown that query expansion effects are sensitive to the presence of orthographic cognates and develop a unified framework for term selection and term translation. Littmana et al. (1998) and Chew et al. (2007) perform CLIR by computing latent semantic indexing on the term-document matrix obtained from a parallel corpora. After reducing the rank, the queries and the documents are projected to a lower-dimensional space.

Statistical Machine Translation techniques have also been tried out (Schamoni et al. 2014; Türe et al. 2012a; Sokolov et al. 2014). Jagarlamudi and Kumaran (2007) use SMT for CLIR between Indian languages. They use a word alignment table that was learned using an SMT on parallel sentences to translate source language query to a query in the target language. In Sokolov et al. (2014), the SMT technique was trained to produce a weighted list of alternatives for query translation. In Ture and Boschee (2014), the combination of different query translation techniques is shown. The query translation techniques are one-best, n-best translations from a standard MT system and a probabilistic translation system obtained from a bilingual corpus with statistical word alignment techniques. Combination weights are learned uniquely for each query.

Transliteration-based models have also been looked into. Udupa et al. (2009) use transliteration of the out-of-vocabulary (OOV) terms. They treat a query and a document as comparable, and for each word in the query and each word in the document, they find out a transliteration similarity value. If this value is above a particular threshold, then the word is treated as a translation of the source query word. They iterate through this process, working on relevant documents retrieved in each iteration. A simple rule-based transliteration approach was proposed by Chinnakotla et al. (2007) for converting OOV Hindi terms to English. They then use a pageRank-based algorithm to resolve between multiple dictionary translations and transliterations.

Herbert et al. (2011) use **Wikipedia concepts** along with Google translate to translate queries. The Wikipedia concepts are mined using cross-language links and redirects and a translation table is built. Translations from Google are then expanded using these concept mappings. Explicit semantic analysis (ESA) is a method to represent a document in the Wikipedia article space as a vector whose components represent its association with the Wikipedia articles. ESA is used by Sorg and Cimiano (2008) in CLIR along with a mapping function that uses cross-lingual links to link documents in the two languages that talk about the same topic. Both the queries and the documents are mapped to this ESA space, where the retrieval is performed.

Graph-based approaches include Franco-Salvador et al. (2014). They employ BabelNet, a multilingual semantic network. They build a basic vector representation of each term in a document and a knowledge graph for every document using BabelNet and interpolate them in order to find the knowledge-based document similarity measure. Similarity Learning via Siamese Neural Network (S2Net) (Yih et al. 2011) trains two identical networks concurrently in which the input layer corresponds to the original term vector and the output layer is the projected concept vector. The model is trained by minimizing the loss of the similarity scores of the output vectors, given pairs of raw term vectors and their labels (similar or not).

Using the **online translation** services Google and Bing to translate queries from source language to target language has been studied in Hosseinzadeh Vahid et al. (2015). They conclude that no single perfect SMT or online translation service exists, but for each query one performs better than the others.

Bilingual word embeddings for CLIR have been introduced by Vulić and Moens (2015). They leverage document-aligned bilingual corpora for learning embeddings of words from both the languages. Given a document *d* in a source language and its aligned equivalent document *t* in the target language, they merge and randomly shuffle the documents *d* and *t*. They train this "pseudo-bilingual" document using *word2vec*. To get the document and query representations, they treat them as bag of words and combine the vectors of each word to obtain the representations of the query and document. Between a query vector and a document vector, they compute the cosine similarity score and rank the documents according to this metric.

Multilingual clustering has also been applied to CLIR. Wei et al. (2008) perform LSA. Their approach has three phases. The first is multilingual semantic space analysis, where terms and documents from either language from the parallel corpora are mapped. In the next phase, this multilingual semantic space is used to fold in the target multilingual documents to be clustered in this semantic space. Finally, a clustering algorithm is used to cluster the folded-in multilingual documents. A multiway generalization of SVD, PARAFAC2 (a variant of PARAFAC), has been used for clustering multilingual parallel documents and used for CLIR in Chew et al. (2007). They construct an irregular three-way array, each slice of which is a separate term-by-document matrix for a single language in the parallel corpus. The goal is to compute an SVD for each language such that V (the matrix of right singular vectors) is the same across all languages.

Sometimes adequate and efficient resources that can aid in translation between the source and target languages may not be present. **Intermediate languages**, also called **pivot languages or assistive languages,** are then used to fill the gap, if there exist sufficient resources between the source language-intermediate language and the intermediate language-target language. Ballesteros (2000) use transitive translation of a query using two bilingual dictionaries. Japanese to Dutch CLIR is done using English as the intermediate language. To reduce the error in the approach of Ballesteros (2000) and Gollins and Sanderson (2001) use lexical triangulation. They take only translations that are common from two ways of transitive translation using two pivot languages. Dabre et al. (2015) show how machine translation can be improved if more pivot language resources of small sizes are used between source and target languages. They use the phrase tables generated using multiple pivots for the source-target phrase table. They use the Multiple Decoding Paths (MDP) feature of Moses for this purpose.

3 MOTIVATION FOR THIS WORK

Most of the approaches discussed above use an auxiliary aligned resource. Dictionary-based and SMT-based approaches require dictionaries or parallel corpora, which are hard to find for

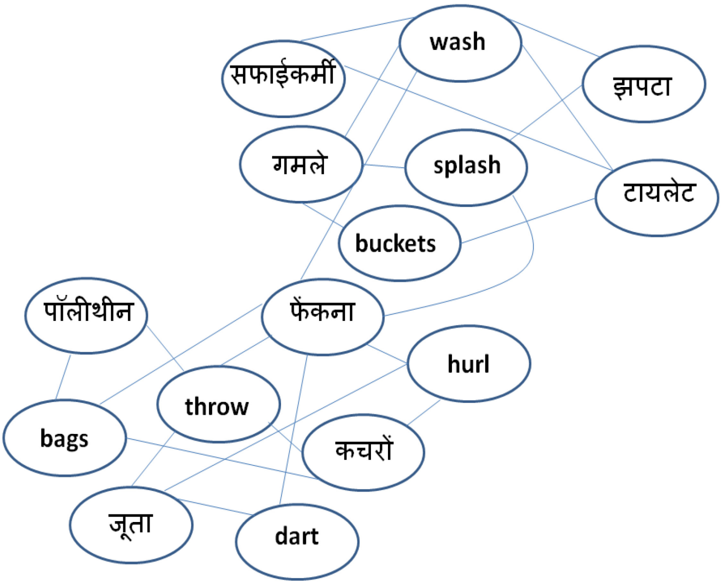


Fig. 1. An example graph for disambiguating the word *"pheMkanaa,"* meaning "throw," using two languages: English and Hindi.

resource-scarce languages such as Indic languages. Transliteration approaches require domain knowledge about the source and target language since they require an intricate design of transliteration rules but still have limited applicability. We propose to apply multilingual word embeddings in order to partially alleviate the data scarcity problem while maintaining a good translation performance. Multilingual word embeddings can be constructed efficiently using a small set of aligned or pseudo-aligned resources. Multilingual clustering groups together similar words across languages that share the same concept. We extend our approach from bilingual clustering to multilingual clustering for up to four languages. We hypothesize that the addition of more languages shall help improve cluster quality. The clusters shall be more well defined and coherent with the inclusion of more languages. Our idea is to construct these clusters using multilingual word embeddings so that the drawbacks of naively using word embeddings for query translation can be overcome. Our clustering approach is based on constructing a graph based on the cosine similarity metric and using this graph to form clusters. These clusters are now used for query translation CLIR.

Figure 1 is an example of graph formation and Figure 2 shows one of its corresponding clusters. We observe that the two senses of "*pheMkanaa*" are divided into two clusters—one meaning is "to throw away something" and the other meaning is "to throw something at someone." These differences in meaning are essentially captured in the clustering of the graph.

4 EXPERIMENTAL SETTINGS

In this section, we shall describe the dataset, experimental settings, and baselines we have used throughout our experiments.

4.1 Resources for Learning Word Embeddings

We first describe the datasets and resources used for learning the multilingual word embeddings.

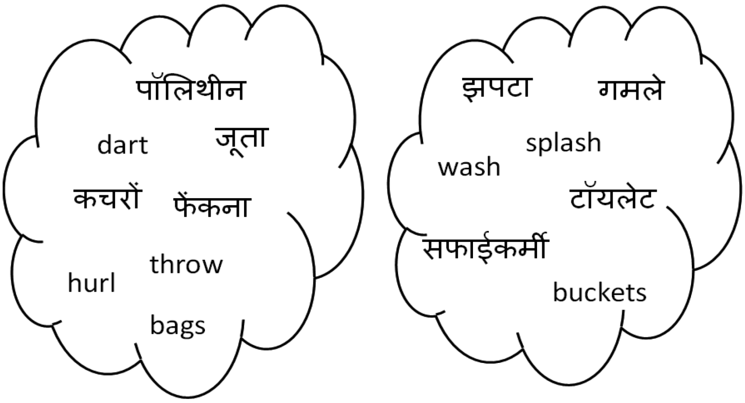


Fig. 2. Bilingual clustering of the above graph. Now the translations of the word *"pheMkanaa"* are in two different clusters.

Table 1. Statistics of the FIRE 2012 Corpora

Language	Source	# Documents
English	Telegraph BDNews24	303,291 89,286
Hindi	Amar Ujala Navbharat Times	54,266 331,599
Bengali	Anandabazar Patrika (IN) BDNews24 (BD)	374,203 83,167
Marathi	Maharashtra Times, Sakal	99,275
Gujarati	Gujarat Samachar	313,163
Tamil	Dinamalar	194,483

- **Comparable Corpora:** We have used Forum for Information Retrieval Evaluation (FIRE) datasets. FIRE is a South Asian counterpart of CLEF, TREC, and NTCIR. We have used the 2012 datasets developed for their shared tasks for CLIR.[1] The details of the data are presented in Table 1.
- **Document-Aligned Corpora:** We have used the Wikipedia dumps[2] available for download for the languages English (May 2016), Bengali (May 2016), Hindi (May 2016), and Tamil (October 2016). In order to get the cross-lingual articles, we made use of the inter-wiki links that exist in the corresponding Wikipedia pages. There were 55,949 English-Hindi pages, 34,234 English-Bengali pages, 12,324 English-Bengali-Hindi pages, and 8,024 English-Bengali-Hindi-Tamil pages.
- **Dictionary:** We used a Hindi-English dictionary[3] that had 26,485 translation pairs, Bengali-English (19,890 pairs), Marathi-English (12,017 pairs), Gujarati-English (6,951 pairs), and Tamil-English (11,848 pairs) dictionary.[4]

[1] http://fire.irsi.res.in/fire/data.
[2] https://dumps.wikimedia.org/backup-index.html.
[3] http://ltrc.iiit.ac.in/onlineServices/Dictionaries/Dict_Frame.html.
[4] http://www.cfilt.iitb.ac.in/Downloads.html.

- **Other NLP Resources:** The Hindi stopword list[5] was used and for the other languages the top 100 most frequently occurring words were considered as stopwords. We also used an English Named-Entity Recognizer.[6] Louvain (Blondel et al. 2008) was used for community detection.

4.2 Setup for the CLIR Task

- **Dataset for the CLIR Task:** We have experimented with FIRE 2008, 2010, 2011, and 2012. The topics are from topics 26 to 75, 76 to 125, 126 to 175, and 176 to 225, respectively. We use the title field for the experiments.
- **Information Retrieval Engine:** We used Apache Solr version 4.1 as the monolingual retrieval engine. The similarity score for the query and the documents was the default TF-IDF Similarity.[7]
- **Evaluation:** The human relevance judgments were available from FIRE. Each query had about 500 documents that were manually judged as relevant (1) or nonrelevant (0). We then used the *trec-eval* tool[8] to find the Precision at 5 and 10, Recall at 10 (P5, P10, R10), and the Mean Average Precision (MAP).

  P5 and P10 measure the precision at the top five and top 10 ranked retrieved documents, respectively. R10 is the recall at the top 10 ranked retrieved documents.

  For a single query, Average Precision (AP) is the average of the precision values obtained for the set of the top k documents existing after each relevant document is retrieved. When this value is averaged over queries, it is termed as MAP.

4.3 Transliteration of Named Entities

The source language query may contain named entities, whose embeddings are absent. Since no named-entity recognition (NER) tool was publicly available for Indian languages, we resorted to the transliteration-based process. For each character in an Indian language, we construct a table of its possible transliterations. For example, the first consonant in Hindi, *ka,* has three possible transliterations in English: *ka, qa, ca*. We apply several language specific rules: a consonant, for instance, *ka,* in Hindi can have two forms, one that is succeeded by a silent *a*, i.e., *ka,* and another that is not, i.e., *k*. The second case applies when it is succeeded by a vowel or another consonant in conjunction (also known as *yuktakshar*). For each transliteration of an OOV Hindi query word h and for each word e in the list of words returned as named entities in the English language, we apply the Minimum Edit Distance algorithm between h and e. We then take the word with the least edit distance. Our transliteration concept is based on the approach of Chinnakotla et al. (2007) and gives quite a satisfactory result, with an accuracy of 90%.

4.4 Baselines

In this section, we describe the baseline methods we have used to compare our proposed approaches.

- **English Monolingual:** This corresponds to the retrieval performance of the target language (English) queries supplied by FIRE.
- **Dictionary:** This is the dictionary-based method where the query translations have been obtained from the dictionary. For words that contain multiple translations, we include all

[5] http://www.ranks.nl/stopwords/hindi.
[6] http://nlp.stanford.edu/software/CRF-NER.shtml.
[7] https://lucene.apache.org/core/3_5_0/api/core/org/apache/lucene/search/Similarity.html.
[8] http://trec.nist.gov/trec_eval/.

of them. Translations with multiwords are not considered. Named entities are handled as described in Section 4.3.

- **Chinnakotla et al. (2007):** We also use the method proposed by Chinnakotla et al. (2007) as a baseline since they participated in the FIRE task.[9] They use a simple rule-based transliteration approach for converting OOV Hindi terms to English and then use a pageRank-based algorithm to resolve between multiple dictionary translations and transliterations. They represent the translations of the query terms as nodes in a graph, where there is an edge between the translations of the different query terms but no edge between translations of a particular query term. The edge weights are the Dice coefficients between the two connecting nodes. An iterative algorithm, based on the idea of pageRank, is then used to resolve between multiple translations and transliterations. Initially, each node is assumed to be equally likely and the weight of each node is a fraction of the number of translations or transliterations. At every iteration, the node weights are recomputed, using the link weights and the weight of the adjacent node of that particular link. The iteration stops on convergence and the two most probable translations/transliterations are chosen as candidates.
- **Vulić and Moens (2015):** We use the word-embedding-based approach proposed by Vulić and Moens (2015). We obtain the multilingual word embeddings using the merge-and-shuffle technique using Wikipedia articles. We add the individual word vectors of the query to get the query vector. Similarly, we add the individual word vectors of the document to get the document vector. IR is performed by computing the cosine similarity between the query vector and the document vector.
- **Google Translate:** This is also used as a baseline, where the Indian language query is translated using Google Translate to English. We consider the top one translation.
- **Google Translate + Dictionary (GT+DT):** We combine the translations from Google and the dictionaries and use it as a baseline.

5 CROSS-LANGUAGE INFORMATION RETRIEVAL USING MULTILINGUAL WORD EMBEDDINGS FOR INDIAN LANGUAGES

In this section, we discuss how multilingual word embeddings can be obtained using the different types of corpora available for Natural Language research, namely, comparable corpora, document-aligned corpora, and parallel corpora.

Comparable corpora are defined as a body of texts from two different languages that are not aligned but belong to the same domain, for example, news domain for a particular time duration.

Document-aligned corpora are text documents from different languages where the i^{th} document from each language talks about the same topic. For instance, Wikipedia articles in different languages are connected through inter-wiki links, as shown in Figure 3.

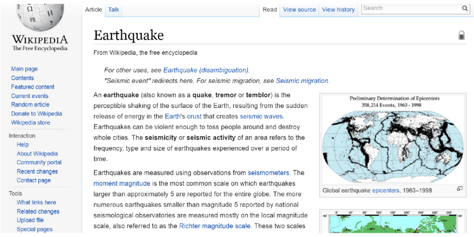
Parallel corpora are a body of texts where each sentence in a language has its translation in the other languages.

Since substantial parallel corpora are not available for Indian languages, techniques using such corpora did not perform well in our setting. We will limit our discussion to the methods involving the other two corpora.

In the setting of comparable corpora, we follow the approach proposed by Mikolov et al. (2013b). In the subsequent sections, we shall term this approach as the "dictionary projection" method.

For document-aligned corpora, we adopt the idea in Vulić and Moens (2015) for obtaining multilingual word vectors. In the subsequent sections, we shall term this approach as the "merge-and-shuffle" method.

[9] The work of Chinnakotla et al. (2007) is an improved version of Padariya et al. (2008).



Earthquake

From Wikipedia, the free encyclopedia

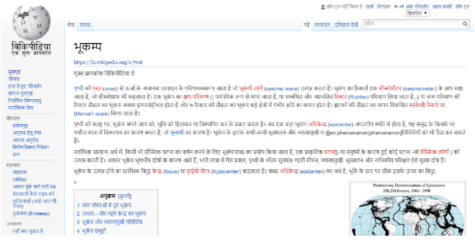
For other uses, see Earthquake (disambiguation).
"Seismic event" redirects here. For seismic migration, see Seismic migration.

An **earthquake** (also known as a **quake**, **tremor** or **temblor**) is the perceptible shaking of the surface of the Earth, resulting from the sudden release of energy in the Earth's crust that creates seismic waves. Earthquakes can be violent enough to toss people around and destroy whole cities. The **seismicity** or **seismic activity** of an area refers to the frequency, type and size of earthquakes experienced over a period of time.

Earthquakes are measured using observations from seismometers. The moment magnitude is the most common scale on which earthquakes larger than approximately 5 are reported for the entire globe. The more numerous earthquakes smaller than magnitude 5 reported by national seismological observatories are measured mostly on the local magnitude scale, also referred to as the Richter magnitude scale. These two scales

Global earthquake epicenters, 1963–1998

Document-Aligned Corpora for the topic
“Earthquake” in English from Wikipedia



भूकम्प

https://hi.wikipedia.org/s/7eut

मुक्त ज्ञानकोश विकिपीडिया से

पृथ्वी की परत (crust) से ऊर्जा के अचानक उत्पादन के परिणामस्वरूप आता है जो भूकंपी तरंगें (seismic wave) उत्पन्न करता है। भूकंप का रिकार्ड एक सीस्मोमीटर (seismometer) के साथ रखा जाता है, जो सीस्मोग्राफ भी कहलाता है। एक भूकंप का क्षण परिमाण () पारंपरिक रूप से मापा जाता है, या सम्बंधित और अप्रचलित रिक्टर (Richter) परिमाण लिया जाता है, ३ या कम परिमाण की रिक्टर तीव्रता का भूकंप अक्सर इम्परसेप्टीबल होता है और ७ रिक्टर की तीव्रता का भूकंप बड़े क्षेत्रों में गंभीर क्षति का कारन होता है। झटकों की तीव्रता का मापन विकसित मरकेली पैमाने पर (Mercalli scale) किया जाता है।

पृथ्वी की सतह पर, भूकंप अपने आप को, भूमि को हिलाकर या विस्थापित कर के प्रकट करता है। जब एक बड़ा भूकंप अधिकेन्द्र (epicenter) अपतटीय स्थति में होता है, यह समुद्र के किनारे पर पर्याप्त मात्रा में विस्थापन का कारण बनता है, जो सूनामी का कारण है। भूकंप के झटके कभी-कभी भूस्खलन और ज्वालामुखी ग ([[en:phenomenon|phenomenon]]तिविधियों को भी पैदा कर सकते हैं।

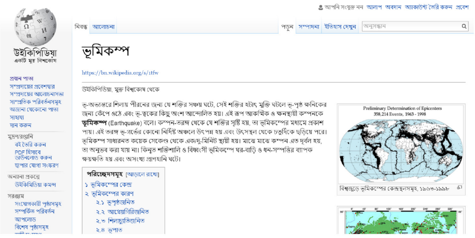
सर्वाधिक सामान्य अर्थ में, किसी भी सीस्मिक घटना का वर्णन करने के लिए *भूकंप* शब्द का प्रयोग किया जाता है, एक प्राकृतिक घटना]) या मनुष्यों के कारण हुई कोई घटना -जो सीस्मिक तरंगों) को उत्पन्न करती है। अक्सर भूकंप भूगर्भीय दोषों के कारण आते हैं, भारी मात्रा में गैस प्रवास, पृथ्वी के भीतर मुख्यतः गहरी मीथेन, ज्वालामुखी, भूस्खलन और नाभिकीय परिक्षण ऐसे मुख्य दोष हैं।

भूकंप के उत्पन्न होने का प्रारंभिक बिन्दु केन्द्र (focus) या हाईपो सेंटर (hypocenter) कहलाता है। शब्द अधिकेन्द्र (epicenter) का अर्थ है, भूमि के स्तर पर ठीक इसके ऊपर का बिन्दु.

=

अनुक्रम [छुपाएँ]
1 प्लेट सीमाओं से दूर भूकंप.
2 उथला - और गहरे केन्द्र का भूकंप
3 भूकंप और ज्वालामुखी गतिविधि
4 भूकंप समूहों

Document-Aligned Corpora for the topic
“Earthquake” in Hindi from Wikipedia



ভূমিকম্প

https://bn.wikipedia.org/s/1tfw

উইকিপিডিয়া, মুক্ত বিশ্বকোষ থেকে

ভূ-অভ্যন্তরে শিলায় পীরনের জন্য যে শক্তির সঞ্চয় ঘটে, সেই শক্তির হটাৎ মুক্তি ঘটলে ভূ-পৃষ্ঠ ক্ষনিকের জন্য কেঁপে ওঠে এবং ভূ-ত্বকের কিছু অংশ আন্দোলিত হয়। এই রূপ আকস্মিক ও ক্ষনস্থায়ী কম্পনকে **ভূমিকম্প** (Earthquake) বলে। কম্পন-তরঙ্গ থেকে যে শক্তির সৃষ্টি হয়, তা ভূমিকম্পের মধ্যমে প্রকাশ পায়। এই তরঙ্গ ভূ-গর্ভের কোনো নির্দিষ্ট অঞ্চলে উৎপন্ন হয় এবং উৎসস্থল থেকে চতুর্দিকে ছড়িয়ে পরে। ভূমিকম্প সাধারনত কয়েক সেকেণ্ড থেকে এক/দু-মিনিট স্থায়ী হয়। মাঝে মাঝে কম্পন এত দূর্বল হয়, তা অনুভব করা যায় না। কিন্ত শক্তিশালি ও বিধ্বংসী ভূমিকম্পে ঘর-বাড়ি ও ধন-সম্পত্তির ব্যাপক ক্ষয়ক্ষতি হয় এবং অসংখ্য প্রাণহানি ঘটে।

পরিচ্ছেদসমূহ [আড়ালে রাখো]
১ ভূমিকম্পের কেন্দ্র
২ ভূমিকম্পের কারণ
২.১ ভূপৃষ্ঠজনিত
২.২ আগ্নেয়গিরিজনিত
২.৩ শিলাচ্যুতিজনিত
২.৪ ভূপাত

বিশ্বজুড়ে ভূমিকম্পের কেন্দ্রস্থলসমূহ, ১৯৬৩-১৯৯৮

Document-Aligned Corpora for the topic
“Earthquake” in Bengali from Wikipedia

Fig. 3. Example of a document-aligned triplet for the topic “earthquake” from Wikipedia.

Table 2. Translations of Query Terms for “*2008 guvaahaaTii bama visphoTa se xati*” Using Word Embeddings

Query Term in Hindi	Meaning in English	Translations Using Embeddings
2008	2008, year	2008
guvaahaaTii	Guwahati, a place in India	Guwahati
bama	bomb	explosives, bomb, device
visphoTa	explosion	explosion, blast, accident
xati	loss	degradation, damage, distortion

5.1 Application of Multilingual Word Embeddings to Cross-Language Information Retrieval

In CLIR, the language of the query is different from the language of the collection from which the document is retrieved. There are three broad approaches to CLIR: translating the language of the query to the language of the documents (query translation), translating the language of the documents to the language of the query (document translation), and translating both the query and the documents to a third language. In our experiments, we use the query translation approach since it is the most standard approach for CLIR.

We take Hindi as the query language and English as the target language for CLIR. Given a query Q comprising terms $q_1 q_2 \ldots q_n$, we first remove the stop-words from the query. We then compute the cosine similarity between the vector of the query term q_i and each word vector from the English vocabulary and pick the top k most similar words and translations. The value of k is decided empirically. We report results for $k = 3$. An example query along with its translations is depicted in Table 2.

Table 3. Results of the Hindi-English and Bengali-English CLIR Using the Different Methods for Obtaining Multilingual Word Embeddings

Task	Method	Data	MAP	P5	P10
Hindi-English CLIR	Dictionary Projection (1)	Comparable Corpora + Dictionary	**0.25**	**0.384**	**0.372**
	Merge-and-Shuffle (2)	Doc-Align Pair (2a)	0.1832	0.272	0.27
		Common Doc-Align (2b)	0.1524	0.232	0.22
		All Doc-Align (2c)	0.1941	0.31	0.3
Bengali-English CLIR	Dictionary Projection (1)	Comparable Corpora + Dictionary	0.2368	0.334	0.318
	Merge-and-Shuffle (2)	Common Doc-Align (2b)	**0.3027**	**0.448**	**0.402**

We experiment with the following settings:

(1) Dictionary Projection: This method uses the FIRE comparable document collection for monolingual embeddings and a dictionary for projection.
(2) Merge-and-Shuffle: This method uses the document-aligned corpus from Wikipedia. We experiment with the following three variants of this approach:
  (a) Doc-Align Pair: Here we use Hindi-English-aligned document pairs.
  (b) Common Doc-Align: This corpus consists of common Wikipedia documents for English, Hindi, and Bengali.
  (c) All Doc-Align: In this setting, we have used all English-Hindi, English-Bengali, and Hindi-Bengali documents.

We also experiment with Bengali, Gujrati, Marathi, and Tamil as source languages and English as the target language for CLIR. For these languages, we only look into the performance of the cross-lingual word embeddings obtained from the dictionary projection approach.

5.2 Results

We present the results on Hindi-to-English and Bengali-to-English CLIR. We perform experiments with both the approaches presented in Section 5.1.

- Hindi-to-English CLIR: We perform experiments under various settings described earlier. The performance of the approaches are presented in Table 3. The dictionary projection method performs the best among the two approaches for obtaining cross-lingual word embeddings. Among the variants of the merge-and-shuffle approach for the document-aligned corpus, "All Doc-Align" performs the best, followed by "Doc-Align Pair." Some example translations are shown for the query word *Ahimsa* (meaning nonviolence) in Table 4.
- Bengali-to-English CLIR: We perform the same experiments under similar settings for Bengali-to-English CLIR. For Bengali-to-English CLIR, we report results for only the variant "Common Doc-Align" (2b) since it performed better than the other variants of method 2. The merge-and-shuffle (2) technique performed better than the dictionary projection (1) approach. This may be due to the fact that both the comparable corpora size and the dictionary size were less compared to the Hindi corpus and the Hindi-English dictionary.

5.3 Analysis

In Table 4, we show some example translations for the Hindi word *Ahimsa*, meaning nonviolence. We observe that the English and Bengali translations from all the approaches are contextually

Table 4. Translations of the Word *Ahimsa,* Meaning Nonviolence, by Different Approaches

Word	Method	Language	Most Similar Words
Ahimsa	Dictionary Projection (1)	English	idealism, compassion, mercy, civilization
	Doc-Align Pair (2a)	English	Ahimsa, himsa, nonviolence, hims, himsa
		Hindi	*ahiMsaa, praaNavadha, mahaavrata, vidyaatmaka, jiivahiMsaa*
	Common Doc-Align (2b)	English	Ahimsa, nonviolence, himsa, Dharmasastras
		Bengali	*ahiMsaa, satyaagraha*
		Hindi	*ahiMsaa, praaNavadha, shraavaka, asteya, satyaagraha*
	All Doc-Align (2c)	English	Ahimsa, himsa, nonviolence, noninjury, ahimsa
		Bengali	*ahiMsaa, hiMsaa, mahaabrata, anupaadisheSha*
		Hindi	*ahiMsaa, praaNavadha, aataMrika, hiMsaa*

relatable to the query word. Also, the words in Hindi are very similar to the query word, suggesting that even in the multilingual space, the quality of the monolingual word embeddings is good.

For OOV words that are actually in English and have been written in Hindi orthographic format (e.g., "housing," "speaker," and "cancer" in English have been written as "*haausiMga*," "*spiikara*," and "*kaiMsara*" in Hindi), word embeddings (WEs) can easily retrieve translations like "housing," "society," and "speaker" and "parliament," "cancer," and "disease," respectively, using contextual cues. It is thus evident that the word-embedding-based method is robust, the translations being very close in meaning to the source language words.

An issue that comes up while using the embedding-based methods is whether to include the embeddings of the named entities in the process. For a particular word in the source language *w*, similar words that showed up are relevant to *w* but are not translations. For example, the words that were most similar to the word *BJP* in Hindi (which is an Indian political party) also included the names of other political parties like *Congress* and also words like *Parliament* and *government* in the target language English. Inclusion of such terms can harm the retrieval process as named entities play a critical role in information retrieval. So we decided to exclude them from the embeddings and use a transliteration scheme as described in Section 4.3.

It is also observed that words that come up as potential translations with high cosine similarity value are always not correct translations. For instance, for the Hindi word "*pheMkanaa*" (meaning throw), besides giving the correct translation "throw," the method also came up with not-so-relevant translations like "wash" and "splashing." Another example is "*desha*" (meaning country) in Hindi. Although correct translations like "country" and "democracy" were provided, irrelevant words like "aspiration" and "kind" also showed up as potential translations.

6 USING MULTILINGUAL WORD EMBEDDING CLUSTERS FOR CLIR

We have seen that multilingual word embeddings can serve as a powerful tool in bridging the gap between good-quality translation and the scarcity of data-aligned resources. But the approach for directly translating a query term using a similarity match between the source and target language word vectors shows some serious problems. Sometimes words that are irrelevant to the source language word are output as translations. Inclusion of such nonrelated words in a query greatly harms the IR performance.

In order to alleviate the problem faced by directly using translations obtained by the cosine similarity score, we propose to use multilingual word clusters for the purpose of CLIR. In multilingual clustering, each group may contain words from multiple languages. Words from the same language as well as across languages that are more likely to represent similar concepts fall in the same group. We use multilingual word embeddings to build these clusters. On clustering, words in the languages that represent a certain concept will form dense clusters. Hence, the cluster containing "*pheMkanaa*" has similar and more related words like "hurl" and "dart" instead of "wash" and "splashing," which are now in a different clusters. Similarly, for the query word *"desha,"* the translations were "nation" and "cities."

We also hypothesize that for clustering, the performance of the system may be better. This is because more words from multiple languages make the clusters dense, coherent, and more well defined.

In the following sections, we discuss the approach of using multilingual word embedding clusters for Hindi-to-English and Bengali-to-English CLIR.

6.1 Methodology

We wish to create multilingual word clusters. For this, we first construct a multilingual graph. Next, we implement a community detection algorithm on the graph. We describe the details next.

(1) **Creating a Multilingual Graph:** After obtaining the multilingual embeddings separately by the methods of Mikolov et al. (2013b) (using comparable corpora and dictionary) and Vulić and Moens (2015) (using Wikipedia document-aligned articles), we compute the cosine similarities between the word vectors. Now a graph $G = (V, E)$ is constructed, where the vertex set V represents words from all the languages and E defines the set of edges—an edge exists between two vertices if the cosine similarity value of the word embeddings of the two vertices is greater than or equal to a threshold. We have selected this threshold as 0.5. The edge weights are the cosine similarity of the embeddings of the connecting vertices (words).

(2) **Clustering:** After the graphs have been obtained, we apply the Louvain algorithm for community detection (Blondel et al. 2008) separately for the graphs. The input to Louvain is the graph constructed previously. It hierarchically forms communities over the graph, optimizing the modularity at each level. Given a graph, Louvain looks for small clusters, optimizing the modularity in a local way. In the first pass, small communities are formed. In the subsequent passes, it combines communities from the lower level to create larger-sized clusters. The iteration stops once maximum modularity is achieved. Since it is a hierarchical clustering method, it does not assume a particular value k. It performs hard clustering; that is, a word belongs to only one cluster. It has a running time complexity of $O(nlogn)$ and hence executes fast.

6.2 Cluster Analysis

We perform experiments with two, three, and four languages. Tables 5, 6, and 7 show word count and cluster statistics of the multilingual word embedding clusters. The word count equals the number of vertices that have been used to create the graph. "Bilingual" indicates that the words (or vertices) are from two languages, while "Trilingual" and "Quadrilingual" indicate that the words (or vertices) are from three and four languages, respectively.

In lower levels, the number of clusters were more and words that should belong to the same cluster were scattered in other clusters. In the topmost level of clustering, although there were some clusters that had a large number of words and were unrelated, most of them were coherent. On observing the bilingual and multilingual clusters closely, we find that the bilingual clusters were mostly small and contained words that were translations and/or transliterations of each other. The multilingual clusters were large and the communities were well representative of the words. Words across languages that were similar in meaning now belonged to the same cluster. This suggests

Table 5. Table Illustrating the Number of Words, Levels, and Clusters from the Two Different Approaches Using Two Different Datasets for Bilingual Clustering

Method		English-Hindi		English-Bengali	
		En	Hi	En	Ben
Dictionary Projection	Words	129,688	84,773	129,688	93,057
	Levels	5		5	
	Clusters	403		384	
Merge-and-Shuffle	Words	106,746	35,361	77,302	24,794
	Levels	4		4	
	Clusters	19,611		20,627	

Table 6. Table Illustrating the Number of Words, Levels, and Clusters from the Two Different Approaches for Trilingual Clustering

Method		English-Hindi-Bengali		
		English	Hindi	Bengali
Dictionary Projection	Word	115,042	66,634	73,384
	Level	5		
	Cluster	294		
Merge-and-Shuffle	Word	50,620	16,534	13,490
	Level	5		
	Cluster	406		

Table 7. Table Illustrating the Number of Words, Levels, and Clusters from the Two Different Approaches for Quadrilingual Clustering

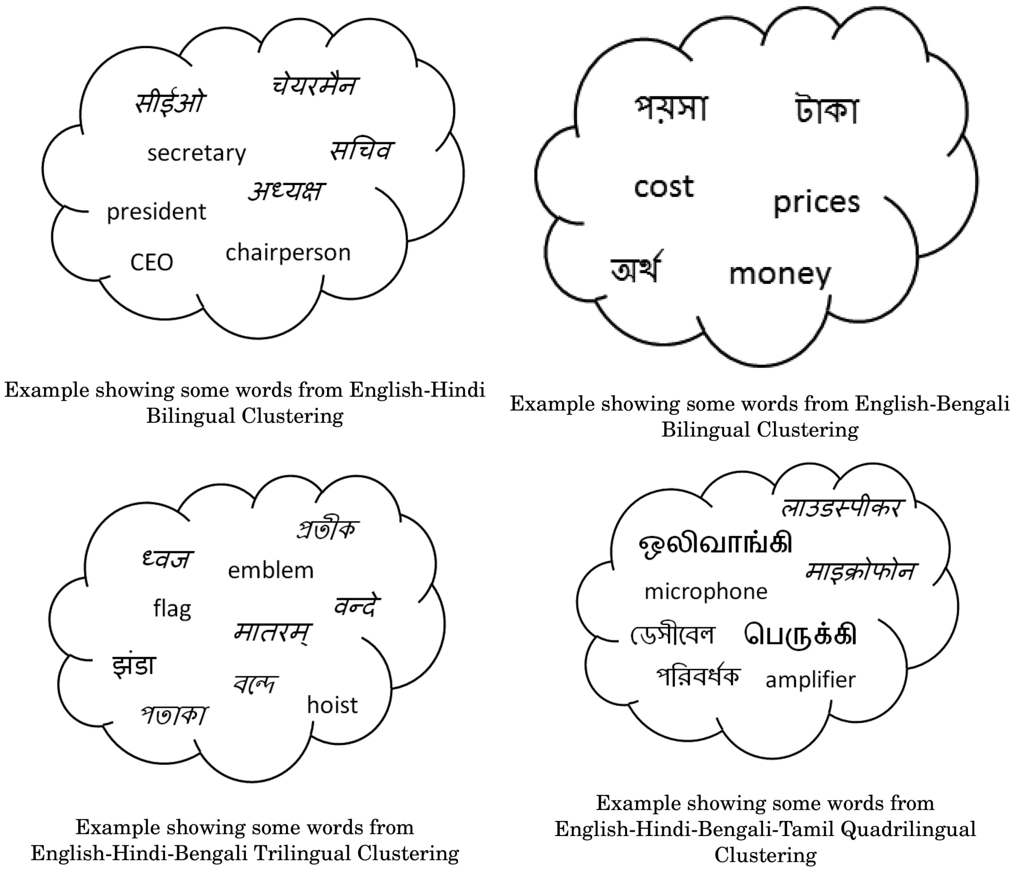
Method		English-Hindi-Bengali-Tamil			
		English	Hindi	Bengali	Tamil
Dictionary Projection	Word	102,726	48,831	55,794	65,476
	Level	5			
	Cluster	183			
Merge-and-Shuffle	Word	41,985	13,218	10,215	27,079
	Level	4			
	Cluster	234			

that if we include more languages, the graph becomes denser with more prominent substructures. The community detection algorithm is therefore able to create well-defined clusters out of the graph.

Some cluster examples are shown in Figure 4. The English-Hindi bilingual cluster shows words that mean some kind of a position of responsibility. The English-Bengali bilingual cluster shows words from the economy domain. The trilingual cluster is related to patriotism, while the quadrilingual cluster shows words from the concept of acoustics.

6.3 Query Translation from Word Clusters

After forming multilingual word clusters, we use them for the purpose of query translation for each query word q_i in CLIR. We perform Hindi-to-English and Bengali-to-English CLIR.



Example showing some words from English-Hindi Bilingual Clustering

Example showing some words from English-Bengali Bilingual Clustering

Example showing some words from English-Hindi-Bengali Trilingual Clustering

Example showing some words from English-Hindi-Bengali-Tamil Quadrilingual Clustering

Fig. 4. Examples of clusters from bilingual, trilingual, and quadrilingual clustering.

Given a query Q = $q_1 q_2 \cdots q_n$ in Hindi or Bengali, we first find the cluster c_k to which the query word q_i belongs. We then extract all the English words from c_k and pick the top t most similar (computed by cosine similarity) English words from the cluster c_k for the query word q_i. We repeat this step for all the query words and append them consecutively. Note that while the stopwords in the query are already filtered, the named entities do not have the embeddings because of filtering of words below the threshold frequency. These named entities are dealt separately, as described in Section 4.3.

We have experimented with various similarity thresholds and various levels of clustering and report the best results. We experimented with the following variants of our approach:

- **Cluster:** In this method, we simply pick the top three (experimentally chosen) most similar English words for each query term within the cluster and append them. We proportionally assign weights to each translation of a query term according to its similarity to the query word such that the weight of all the translations of a query term add up to 1. The named entities were assigned a weight of 1.
- **Cluster + DT:** We combine translations from the dictionary as well as from the clusters. We first take translations from the dictionary if a translation exists. If not, we take it only

Table 8. Performance of the Proposed Cluster-Based Approach for Hindi-to-English and Bengali-to-English CLIR for FIRE 2012 Datasets

Datasets		Methods	Hindi-to-English CLIR				Bengali-to-English CLIR			
			MAP	P5	P10	R10	MAP	P5	P10	R10
FIRE		English Monolingual	0.3218	0.56	0.522	0.158	0.3218	0.56	0.522	0.158
		Dictionary	0.1691	0.2048	0.2048	0.083	0.134	0.165	0.132	0.076
		Chinnakotla et al. (2007)	0.2236	0.3347	0.3388	0.089	0.11	0.15	0.147	0.084
		Vulić and Moens (2015)	0.287	0.386	0.372	0.096	0.254	0.33	0.326	0.0924
		Google Translate (August 2015)	0.3566	0.576	0.522	0.107	0.294	0.524	0.48	0.104
		GT+DT	0.348	0.562	0.51	0.098	0.17	0.232	0.228	0.0946
Pair En-Hi/ En-Ben	FIRE (Dictionary Projection)	Cluster	0.352	0.4503	0.427	0.091	0.3038	0.478	0.418	0.075
		Cluster+DT	0.362	0.537	0.52	0.098	0.326	0.495	0.464	0.081
		Cluster+DT+GT	0.452	0.627	0.578	0.112	0.342	0.534	0.49	0.104
	Wikipedia (Merge-and-Shuffle)	Cluster	0.2832	0.3760	0.35	0.084	0.3233	0.468	0.43	0.088
		Cluster+DT	0.324	0.408	0.386	0.091	0.361	0.482	0.458	0.096
		Cluster+DT+GT	0.42	0.526	0.501	0.103	0.389	0.517	0.487	0.105
Tri En-Ben-Hi	FIRE (Dictionary Projection)	Cluster	0.386	0.496	0.434	0.104	0.364	0.492	0.47	0.094
		Cluster+DT	0.414	0.567	0.531	0.113	0.407	0.543	0.529	0.112
		Cluster+DT+GT	0.482	0.640	0.585	0.125	0.444	0.568	0.542	0.124
	Wikipedia (Merge-and-Shuffle)	Cluster	0.3014	0.446	0.37	0.089	0.3557	0.476	0.418	0.097
		Cluster+DT	0.356	0.541	0.510	0.098	0.396	0.538	0.501	0.104
		Cluster+DT+GT	0.432	0.575	0.538	0.116	0.42	0.56	0.545	0.118
Quadri En-Hi-Ben-Ta	FIRE (Dictionary Projection)	Cluster	0.392	0.502	0.461	0.112	0.372	0.488	0.432	0.102
		Cluster+DT	0.427	0.606	0.548	0.121	0.404	0.525	0.476	0.108
		Cluster+DT+GT	**0.504**	**0.624**	**0.556**	**0.146**	0.437	0.578	0.56	0.116
	Wikipedia (Merge-and-Shuffle)	Cluster	0.3255	0.462	0.39	0.098	0.372	0.498	0.44	0.112
		Cluster+DT	0.36	0.552	0.514	0.105	0.405	0.55	0.50	0.119
		Cluster+DT+GT	0.451	0.582	0.542	0.122	**0.446**	**0.58**	**0.545**	**0.131**

The first six rows report the baselines.

from clusters. In case translations exist in both, we assign 80% weightage to the cluster translations and 20% weightage to the dictionary translations.[10]

- **Cluster + DT + GT:** In this scheme, we combine translations from Google Translate as well as with the dictionary. We assign equal weightage to cluster words and translations from Google, 40% each, and the rest to dictionary translations.
- **Pair En-Hi/En-Ben:** We experiment with bilingual clustering using English and Hindi languages for Hindi-to-English CLIR and bilingual clustering for English and Bengali languages for Bengali-to-English CLIR.
- **Tri En-Ben-Hi:** We form trilingual clusters using words from English, Bengali, and Hindi. We apply these clusters for Hindi-to-English CLIR and Bengali-to-English CLIR.
- **Quadri En-Hi-Ben-Ta:** We use quadrilingual clustering using English, Hindi, Bengali, and Tamil and apply them to Hindi-to-English and Bengali-to-English CLIR.

6.4 Results

The results of our proposed approached and the baselines are shown in Tables 8 through 11.

For Hindi-to-English CLIR, the dictionary projection method performs better than the merge-and-shuffle technique.

Multilingual and bilingual word clusters formed using Wikipedia document-aligned data perform better for Bengali-to-English CLIR compared to the dictionary-based approach using FIRE

[10] We experimented with other weightages like 70%-30% and 90%-10%, but the 80%-20% division gives the best results.

Table 9. Performance of the Proposed Cluster-Based Approach for Hindi-to-English and Bengali-to-English CLIR for FIRE 2011 Datasets

Datasets		Methods	Hindi-to-English CLIR				Bengali-to-English CLIR			
			MAP	P5	P10	R10	MAP	P5	P10	R10
FIRE		English Monolingual	0.2362	0.3760	0.3340	0.092	0.2362	0.376	0.334	0.092
		Dictionary	0.1063	0.1149	0.1447	0.052	0.0581	0.0682	0.06	0.0558
		Chinnakotla et al. (2007)	0.1105	0.194	0.189	0.0753	0.0936	0.107	0.11	0.071
		Vulić and Moens (2015)	0.136	0.198	0.109	0.1028	0.113	0.189	0.176	0.105
		Without Cluster	0.1262	0.1808	0.107	0.0987	0.102	0.124	0.118	0.096
		Google Translate (December 2017)	0.1827	0.304	0.27	0.124	0.1657	0.264	0.22	0.111
		GT+DT	0.1755	0.272	0.274	0.104	0.1696	0.2880	0.252	0.109
Pair En-Hi/ En-Ben	FIRE (Dictionary Projection)	Cluster	0.1704	0.251	0.242	0.093	0.140	0.23	0.217	0.082
		Cluster+DT	0.1796	0.269	0.258	0.098	0.147	0.252	0.226	0.087
		Cluster+DT+GT	0.234	0.318	0.30	0.104	0.161	0.28	0.274	0.103
	Wikipedia (Merge-and-Shuffle)	Cluster	0.13	0.214	0.208	0.082	0.152	0.254	0.232	0.104
		Cluster+DT	0.162	0.246	0.232	0.09	0.17	0.263	0.248	0.112
		Cluster+DT+GT	0.207	0.31	0.302	0.102	0.204	0.301	0.286	0.120
Tri En-Ben-Hi	FIRE (Dictionary Projection)	Cluster	0.192	0.29	0.267	0.106	0.205	0.284	0.271	0.094
		Cluster+DT	0.218	0.324	0.313	0.112	0.22	0.292	0.282	0.106
		Cluster+DT+GT	0.273	0.381	0.368	0.12	0.246	0.332	0.294	0.114
	Wikipedia (Merge-and-Shuffle)	Cluster	0.165	0.256	0.241	0.091	0.214	0.316	0.297	0.118
		Cluster+DT	0.182	0.275	0.262	0.104	0.23	0.328	0.315	0.124
		Cluster+DT+GT	0.22	0.324	0.318	0.118	0.261	0.354	0.343	0.132
Quadri En-Hi-Ben-Ta	FIRE (Dictionary Projection)	Cluster	0.224	0.316	0.294	0.114	0.234	0.337	0.324	0.109
		Cluster+DT	0.236	0.337	0.316	0.122	0.242	0.342	0.331	0.112
		Cluster+DT+GT	**0.297**	**0.404**	**0.39**	**0.138**	0.278	0.414	0.397	0.124
	Wikipedia (Merge-and-Shuffle)	Cluster	0.201	0.286	0.27	0.098	0.286	0.372	0.368	0.129
		Cluster+DT	0.234	0.314	0.292	0.106	0.294	0.396	0.384	0.134
		Cluster+DT+GT	0.25	0.357	0.301	0.124	0.168	**0.435**	**0.424**	**0.146**

The first six rows report the baselines. “Without Cluster” is added as a baseline, where the result reported uses word embeddings obtained by the dictionary projection method.

data. One possible reason may be that the dictionary for Bengali-English is not as rich as Hindi-English. For Hindi-English, the number of word pair translations used for training was 26,845, and for Bengali-English, the number was 19,890. Also, the number of documents in Bengali from the FIRE dataset was less, and this may be a probable cause for its poor performance. Multilingual word clusters alone perform well when compared in terms of MAP with Google Translate.

The clustering-based approach consistently outperforms dictionary and Chinnakotla et al. (2007). While the method of Vulić and Moens (2015) of combining query word and document word vectors performs better than naively using the word embeddings for query translations (i.e., without clustering), it shows poorer performance when compared with the clustering-based approach.

For the FIRE 2012, 2011, and 2008 datasets, “Cluster” alone is able to outperform Google Translate in the trilingual and quadrilingual setting. For the 2010 dataset, the clustering-based approach can beat Google Translate only when it is combined with translations from dictionary and Google Translate in the quadrilingual setting.

We perform Student’s t-test with a paired two sample for means to compare our clustering-based approach over multilingual languages with the baselines.

On comparing “Cluster” with the baseline approaches, we find that the improvements are statistically significant with $p < 0.05$. We find the improvement by Cluster+DT+GT over Cluster to be significant as well.

Table 10. Performance of the Proposed Cluster-Based Approach for Hindi-to-English and Bengali-to-English CLIR for FIRE 2010 Datasets

Datasets		Methods	Hindi-to-English CLIR				Bengali-to-English CLIR			
			MAP	P5	P10	R10	MAP	P5	P10	R10
FIRE		English Monolingual	0.1935	0.2240	0.1840	0.023	0.1935	0.2240	0.184	0.023
		Dictionary	0.1277	0.1306	0.1265	0.035	0.1202	0.1143	0.1082	0.026
		Chinnakotla et al. (2007)	0.1432	0.167	0.1636	0.038	0.124	0.132	0.127	0.034
		Vulić and Moens (2015)	0.163	0.168	0.16	0.064	0.141	0.229	0.227	0.052
		Without Clustering	0.152	0.2374	0.226	0.048	0.14	0.1462	0.142	0.043
		Google Translate (December 2017)	0.1941	0.1920	0.206	0.12	0.153	0.136	0.148	0.117
		GT+DT	0.1937	0.1840	0.18	0.104	0.161	0.124	0.125	0.100
Pair En-Hi/ En-Ben	FIRE (Dictionary Projection)	Cluster	0.178	0.274	0.261	0.082	0.158	0.257	0.204	0.068
		Cluster+DT	0.180	0.26	0.252	0.088	0.164	0.26	0.228	0.074
		Cluster+DT+GT	0.2442	0.326	0.304	0.114	0.182	0.291	0.091	0.082
	Wikipedia (Merge-and-Shuffle)	Cluster	0.122	0.214	0.191	0.076	0.166	0.282	0.276	0.072
		Cluster+DT	0.146	0.228	0.216	0.082	0.182	0.326	0.314	0.079
		Cluster+DT+GT	0.19	0.293	0.287	0.092	0.224	0.352	0.346	0.088
Tri En-Ben-Hi	FIRE (Dictionary Projection)	Cluster	0.184	0.253	0.217	0.097	0.22	0.28	0.264	0.092
		Cluster+DT	0.190	0.292	0.284	0.106	0.246	0.32	0.308	0.104
		Cluster+DT+GT	0.268	0.346	0.331	0.118	0.27	0.362	0.358	0.122
	Wikipedia (Merge-and-Shuffle)	Cluster	0.158	0.224	0.21	0.084	0.234	0.346	0.32	0.081
		Cluster+DT	0.16	0.24	0.236	0.088	0.26	0.38	0.365	0.084
		Cluster+DT+GT	0.232	0.313	0.30	0.096	0.2872	0.42	0.406	0.102
Quadri En-Hi-Ben-Ta	FIRE (Dictionary Projection)	Cluster	0.192	0.25	0.24	0.105	0.252	0.286	0.274	0.103
		Cluster+DT	0.208	0.31	0.308	0.113	0.276	0.34	0.317	0.112
		Cluster+DT+GT	**0.317**	**0.341**	**0.326**	**0.128**	0.294	0.386	0.374	0.124
	Wikipedia (Merge-and-Shuffle)	Cluster	0.176	0.238	0.232	0.095	0.264	0.301	0.298	0.092
		Cluster+DT	0.19	0.291	0.284	0.102	0.282	0.326	0.32	0.101
		Cluster+DT+GT	0.284	0.336	0.324	0.119	**0.306**	**0.343**	**0.34**	**0.123**

The first six rows report the baselines. "Without Cluster" is added as a baseline where the result reported uses word embeddings obtained by the dictionary projection method.

Another notable fact is that as we increase the number of languages in the clusters from two to four, the performance on the task is better.

The improvement of the quadrilingual clustering is statistically significant with $p < 0.05$. We also observe that increasing the number of languages from two to three and four causes statistically significant improvements with $p < 0.05$ over all the baselines. Improvement from Cluster to Cluster+DT+GT is also statistically significant.

To summarize, we have two major observations:

(1) The performance consistently improves when the translations from clusters are combined with dictionary translations and translations from Google. In each setting of Pair, Tri, and Quadri, Cluster+DT+GT shows statistically significant improvements over Cluster.
(2) As we increase the number of languages, that is, as we progress from Pair to Tri to Quadri, the performance improves and is statistically significant.

6.5 Analysis

Results suggest that incorporating cluster information rather than solely relying on translations from multilingual word vectors proves beneficial.

Cluster information improves and words in the clusters are more related with each other and aligned to the semantic information exhibited by the cluster. When more languages are

Table 11. Performance of the Proposed Cluster-Based Approach for Hindi-to-English and Bengali-to-English CLIR for FIRE 2008 Datasets

Datasets		Methods	Hindi-to-English CLIR				Bengali-to-English CLIR			
			MAP	**P5**	**P10**	**R10**	**MAP**	**P5**	**P10**	**R10**
FIRE		English Monolingual	0.1609	0.248	0.236	0.0587	0.161	0.248	0.236	0.0452
		Dictionary	0.084	0.1464	0.137	0.03	0.1042	0.135	0.096	0.024
		Chinnakotla et al. (2007)	0.11	0.15	0.147	0.065	0.062	0.112	0.107	0.0604
		Vulić and Moens (2015)	0.137	0.198	0.184	0.0842	0.081	0.151	0.145	0.079
		Without Clustering	0.121	0.165	0.154	0.0784	0.072	0.128	0.116	0.0642
		Google Translate (December 2017)	0.178	0.255	0.24	0.13	0.106	0.196	0.114	0.125
		GT+DT	0.173	0.241	0.226	0.092	0.164	0.107	0.086	0.086
Pair En-Hi/ En-Ben	FIRE (Dictionary Projection)	Cluster	0.174	0.26	0.249	0.092	0.114	0.185	0.162	0.083
		Cluster+DT	0.181	0.263	0.25	0.102	0.121	0.194	0.181	0.010
		Cluster+DT+GT	0.225	0.278	0.268	0.114	0.128	0.217	0.208	0.11
	Wikipedia (Merge-and-Shuffle)	Cluster	0.132	0.227	0.218	0.085	0.134	0.215	0.21	0.092
		Cluster+DT	0.144	0.241	0.246	0.094	0.142	0.238	0.226	0.108
		Cluster+DT+GT	0.164	0.268	0.26	0.123	0.158	0.242	0.238	0.126
Tri En-Ben-Hi	FIRE (Dictionary Projection)	Cluster	0.192	0.276	0.258	0.114	0.142	0.231	0.22	0.109
		Cluster+DT	0.216	0.294	0.286	0.118	0.151	0.254	0.241	0.122
		Cluster+DT+GT	0.242	0.33	0.324	0.129	0.168	0.262	0.253	0.146
	Wikipedia (Merge-and-Shuffle)	Cluster	0.151	0.245	0.236	0.104	0.167	0.28	0.274	0.114
		Cluster+DT	0.167	0.265	0.254	0.116	0.172	0.268	0.258	0.129
		Cluster+DT+GT	0.19	0.298	0.287	0.143	0.198	0.291	0.288	0.144
Quadri En-Hi-Ben-Ta	FIRE (Dictionary Projection)	Cluster	0.224	0.326	0.312	0.126	0.158	0.248	0.236	0.116
		Cluster+DT	0.242	0.338	0.326	0.132	0.165	0.261	0.252	0.124
		Cluster+DT+GT	**0.267**	**0.372**	**0.361**	**0.144**	0.184	0.292	0.285	0.139
	Wikipedia (Merge-and-Shuffle)	Cluster	0.174	0.268	0.254	0.113	0.188	0.276	0.265	0.128
		Cluster+DT	0.193	0.284	0.273	0.124	0.204	0.291	0.282	0.135
		Cluster+DT+GT	0.212	0.318	0.298	0.138	**0.226**	**0.344**	**0.342**	**0.147**

The first six rows report the baselines. "Without Cluster" is added as a baseline where the result reported uses word embeddings obtained by the dictionary projection method.

incorporated in the clusters, the clusters become more coherent and well defined. Each cluster now contains more focused words related to a concept. This improves the translation.

Table 12 shows two example queries. The first query is for Hindi-to-English CLIR and the second query is for Bengali-to-English CLIR. For the first two translation methods, no translation is available for "*unmuulana*" (meaning eradication), but multilingual clustering suggests the word "prevention." Also, for "*poliyo*," multilingual clustering comes up with the more related word "infection" rather than "vaccine" since "polio" is primarily a disease/infection and vaccination is a medication and is secondary. For the second query, the word "*Euro*" is related to sports and not economics. The No-Cluster method wrongly predicts the context and suggests words like "banknotes." On the other hand, pairwise clustering understands that "cup" is related to sports, "football" to be more specific. Multilingual clustering restricts to a shorter query and hence translates to only "trophy" and "cup."

We provide an example in Table 13 to illustrate the importance of clustering by including multiple languages. The word "express" can have two meanings, namely, "a superfast train" or in the sense of "expressing an emotion." In pairwise clustering with English and Hindi words, we have words from both these concepts. For instance, words like "react" and "*virodhutaa*" (meaning protest) suggest showing an emotion, and words like "railways," "car," "*rela*" (meaning rail), and "*basa*" (meaning bus) suggest names of vehicles. On including words from a third language, Bengali, in the graph, we obtain two different clusters, where one cluster of words belonging to

Table 12. Some Example Queries and Their Performances

Query	Gloss	Translation Method	Translation	MAP	P5	P10
poliyo *unmuulana* *abhiyaana*	polio eradication mission	No Cluster	vaccine polio campaign campaigns	0.4	0.55	0.48
		Wiki Pair	polio vaccine eradication mission	0.6	0.7	0.6
		Wiki Tri Cluster	polio infection prevention campaign	0.85	1	0.9
griisa iuro *kaapa 2004* *jaya*	2004 Greece Euro Cup victory	No Cluster	Greece 2004 euro banknotes tournament champions victory win defeat	0.5	0.7	0.6
		Wiki Pair	Greece 2004 Euro euro trophy football teams victory win winning	0.6	0.75	0.7
		Wiki Quadri Cluster	Greece 2004 Euro trophy cup champions winner	0.9	1	0.8

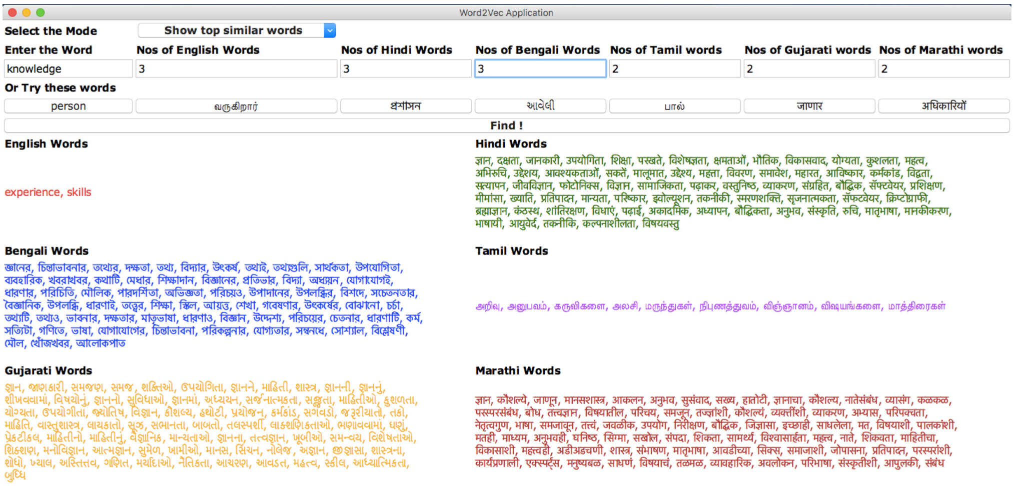
Table 13. A Clustering Example for the Query Word "Express" (Meaning Superfast Train)

Pair Cluster	Tri Cluster	Quad Cluster
express, represent, react, *rela* (H), demonstrate, protests, *relave (H)*, *Daranaa* (H), *plena* (H), *basa* (H), car, railways, *spaShTavaadii* (H), *virodhitaa* (H), rail	*rela* (H), *relave* (H), *paTarii* (H), EMU, platform, express, *Trena* (B), *kampaarTamenTa* (B), *gaaRi* (B), train, *Traka* (H), *kocha* (B)	*Rayil* (T), *Payirciyalar* (T), *Nataimetai* (T), EMU, *kampaarTamenTa (B)*, express platform, *kocha (B)*, *rela* (H), *relave* (H)

the sense of "expressing an emotion" are present and another cluster containing words relating to the concept of "vehicle" is present. The trilingual cluster that represents the concept of "vehicle" has words like "train," "platform," "*Traka*" (in Hindi, meaning truck), "*paTarii*" (in Hindi, meaning railway track), "*kocha*" (in Bengali, meaning coach), and "*gaaRi*" (in Bengali, meaning vehicle). When we incorporate the fourth language, Tamil, in our graph, we now get three distinct clusters—one in the sense of "expressing emotions," the second in the general sense of "vehicles," and the third cluster specifically in the sense of "railway" like "EMU," "platform," "kampaarTamenTa" (in Bengali, meaning compartment), "*Rayil*" (in Tamil, meaning rail), and "*relave*" (in Hindi, meaning railway).

Hence, by incorporating words from multiple languages, the cluster information improves. The subclusters, formed in each step of including one more language, make the clusters represent more specific concepts. The more the words from different languages are distributed, the more semantically coherent the clusters are. Having similar words from multiple languages makes the clusters well defined and semantically coherent. Such clusters help in translation since words are more relevant and specific to representing the concept.

While using bilingual and trilingual clustering, it was observed that for a query word that is not a named entity, translations of the word in other languages were named entities related to the word. But as we moved to quadrilingual clustering, it was observed that the named entities were mostly in one cluster, thus eliminating unwanted named entities as translations for a particular word.



Select the Mode: Show top similar words

Enter the Word: knowledge

Nos of English Words: 3

Nos of Hindi Words: 3

Nos of Bengali Words: 3

Nos of Tamil words: 2

Nos of Gujarati words: 2

Nos of Marathi words: 2

Or Try these words

English Words

experience, skills

Hindi Words

ज्ञान, दक्षता, जानकारी, उपयोगिता, शिक्षा, परखते, विशेषज्ञता, क्षमताओं, भौतिक, विकासवाद, योग्यता, कुशलता, महत्व, अभिरुचि, उद्देश्य, आवश्यकताओं, सकतें, मालूमात, उद्देश्य, महत्ता, विवरण, समावेश, महारत, आविष्कार, कर्मकांड, विद्वता, सत्यापन, जीवविज्ञान, फोटोनिक्स, विज्ञान, सामाजिकता, पढ़ाकर, वस्तुनिष्ठ, व्याकरण, संग्रहित, बौद्धिक, सॉफ्टवेयर, प्रशिक्षण, मीमांसा, ख्याति, प्रतिपादन, मान्यता, परिष्कार, इवोल्यूशन, तकनीकी, स्मरणशक्ति, सृजनात्मकता, सॉफ्टवेयर, क्रिप्टोग्राफी, ब्रह्माज्ञान, कंठस्थ, शांतिरक्षण, विधाएं, पढ़ाई, अकादमिक, अध्यापन, बौद्धिकता, अनुभव, संस्कृति, रुचि, मातृभाषा, मानकीकरण, भाषायी, आयुवेर्द, तकनीकि, कल्पनाशीलता, विषयवस्तु

Bengali Words

জ্ঞানের, চিন্তাভাবনার, তথ্যের, দক্ষতা, তথ্য, বিদ্যার, উৎকর্ষ, তথ্যই, তথ্যগুলি, সার্থকতা, উপযোগিতা, ব্যবহারিক, খবরাখবর, কথাটি, মেধার, শিক্ষাদান, বিজ্ঞানের, প্রতিভার, বিদ্যা, অধ্যয়ন, যোগাযোগই, ধারণার, পরিচিতি, মৌলিক, পারদর্শিতা, অভিজ্ঞতা, পরিচয়ও, উপাদানের, উপলব্ধির, বিশদে, সচেতনতার, বৈজ্ঞানিক, উপলব্ধি, ধারণাই, তত্ত্বের, শিক্ষা, স্কিল, আয়ত্ত, শেখা, গবেষণার, উৎকর্ষের, বোঝানো, চর্চা, তথ্যটি, তথ্যও, ভাবনার, দক্ষতার, মাতৃভাষা, ধারণাও, বিজ্ঞান, উদ্দেশ্য, পরিচয়ের, চেতনার, ধারণাটি, কর্ম, সত্যিটা, গণিতে, ভাষা, যোগাযোগের, চিন্তাভাবনা, পরিকল্পনার, যোগ্যতার, সম্বন্ধে, সোশ্যাল, বিশ্লেষণী, মৌল, খোঁজখবর, আলোকপাত

Tamil Words

அறிவு, அனுபவம், கருவிகளை, அலசி, மருந்துகள், நிபுணத்துவம், விஞ்ஞானம், விஷயங்களை, மாத்திரைகள்

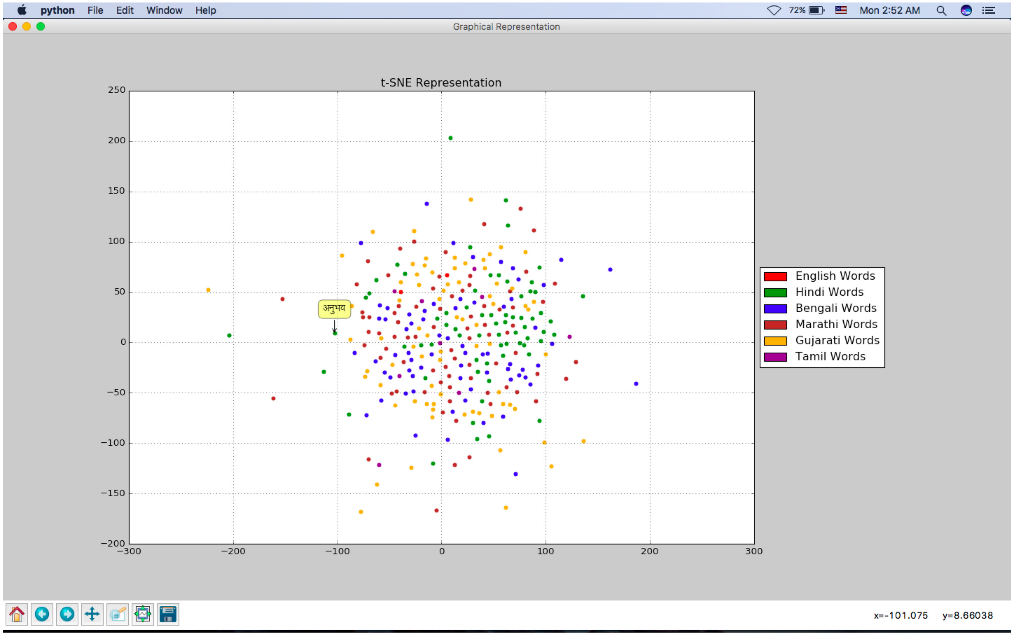
Gujarati Words

જ્ઞાન, જાણકારી, સમજણ, સમજ, શક્તિઓ, ઉપયોગિતા, જ્ઞાનને, માહિતી, શાસ્ત્ર, જ્ઞાનની, જ્ઞાનનું, શીખવવામાં, વિષયોનું, જ્ઞાનનો, સુવિધાઓ, જ્ઞાનમાં, અધ્યયન, સર્જનાત્મકતા, સજ્જતા, માહિતીઓ, કુશળતા, યોગ્યતા, ઉપયોગીતા, જ્યોતિષ, વિજ્ઞાન, કૌશલ્ય, હથોટી, પ્રયોજન, કર્મકાંડ, સગવડો, જરૂરીયાતો, તકો, માહિતિ, વાસ્તુશાસ્ત્ર, લાયકાતો, સૂઝ, સભાનતા, બાબતો, તલસ્પર્શી, લાક્ષણિકતાઓ, ભણાવવામાં, ઘણું, પ્રેક્ટીકલ, માહિતીનો, માહિતીનું, વૈજ્ઞાનિક, માન્યતાઓ, જ્ઞાનના, તત્વજ્ઞાન, ખૂબીઓ, સમન્વય, વિશેષતાઓ, શિક્ષણ, મનોવિજ્ઞાન, આત્મજ્ઞાન, સુમેળ, ખામીઓ, માનસ, સિંચન, નોલેજ, અજ્ઞાન, જીજ્ઞાસા, શાસ્ત્રના, શોધો, ખ્યાલ, અસ્તિત્તવ, ગણિત, મર્યાદાઓ, નૈતિકતા, આચરણ, આવડત, મહત્વ, સ્કીલ, આધ્યાત્મિકતા, બુધ્ધિ

Marathi Words

ज्ञान, कौशल्ये, जाणून, मानसशास्त्र, आकलन, अनुभव, सुसंवाद, सख्य, हातोटी, ज्ञानाचा, कौशल्य, नातेसंबंध, व्यासंग, कळकळ, परस्परसंबंध, बोध, तत्त्वज्ञान, विषयातील, परिचय, समजून, तज्ज्ञांशी, कौशल्यं, व्यक्तींशी, व्याकरण, अभ्यास, परिपक्वता, नेतृत्वगुण, भाषा, समजावून, तत्त्वं, जवळीक, उपयोग, निरीक्षण, बौद्धिक, जिज्ञासा, इच्छाही, साधलेला, मत, विषयाशी, पालकांशी, मतही, माध्यम, अनुभवही, घनिष्ठ, सिग्मा, सखोल, संपदा, शिकता, सामर्थ्य, विश्वासार्हता, महत्त्व, नाते, शिकवता, माहितीचा, विकासाशी, महत्त्वही, अडीअडचणी, शास्त्र, संभाषण, मातृभाषा, आवडीच्या, सिक्स, समाजाशी, जोपासना, प्रतिपादन, परस्परांशी, कार्यप्रणाली, एक्स्पर्ट्स, मनुष्यबळ, साधणं, विषयाचं, तळमळ, व्यावहारिक, अवलोकन, परिभाषा, संस्कृतीशी, आपुलकी, संबंध

Word Translation - showing all similar words

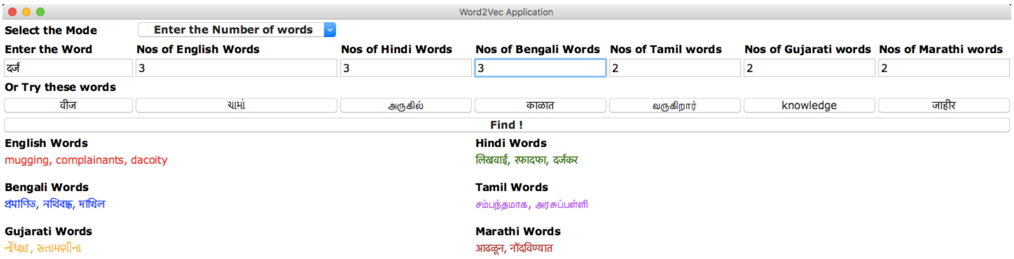


Word Visualization - all the words across different languages are now in the same space with words in a particular language being closer to one another

Fig. 5. Output of the GUI in the form of word translation and visualization for the input query word "knowledge" in English.

7 TOOL FOR PROVIDING RELATED QUERY WORDS ACROSS INDIC LANGUAGES

To get a complete picture, we have constructed multilingual word vectors for English and five Indian languages, namely, Hindi, Bengali, Marathi, Gujrati, and Tamil. We provide a service for visualizing the embeddings. Given a query word in a particular language, the tool retrieves similar words across the languages and also provides visualization of their positions in the two-dimensional coordinate system (using *t-SNE*).



Select the Mode	Enter the Number of words					
Enter the Word	Nos of English Words	Nos of Hindi Words	Nos of Bengali Words	Nos of Tamil words	Nos of Gujarati words	Nos of Marathi words
दर्ज	3	3	3	2	2	2
Or Try these words						
वीज	ચામાં	அருகில்	काळात	வருகிறார்	knowledge	जाहीर

Find !

English Words
mugging, complainants, dacoity

Hindi Words
लिखवाई, रफादफा, दर्जकर

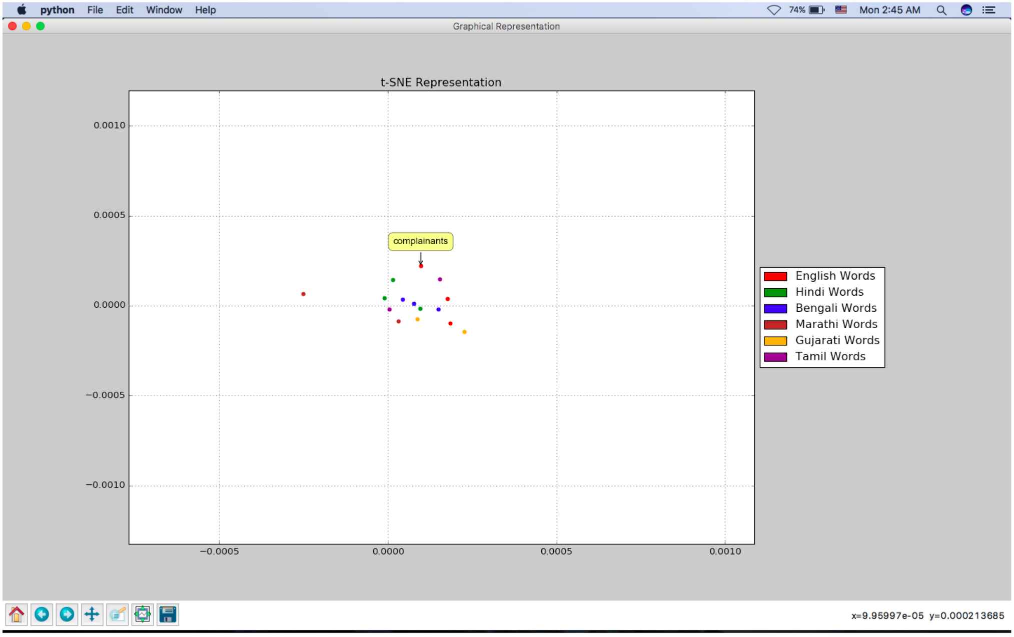
Bengali Words
প্রমাণিত, নথিবদ্ধ, দাখিল

Tamil Words
சம்பந்தமாக, அரசுப்பள்ளி

Gujarati Words
નોંધાઇ, સતામણીના

Marathi Words
आढळून, नोंदविण्यात

Word Translation - showing the user specified number of words per language



Word Visualization - all the words across different languages are now in the same space with words in a particular language being closer to one another

Fig. 6. Output of the GUI in the form of word translation and visualization for the input query word “*darja*” in Hindi, meaning “enter or register (a complaint).”

At the offline part, the word vectors and a list of similar words (cosine similarity between the vectors >= 0.5) have already been precomputed. This data is required as input to the system. In the online part of the system, the user needs to provide the query word and certain specifications, based on which the translations and the vectors shall be retrieved and shown. It takes about an average time of 5 to 6 seconds to retrieve and output the result (on a system with 8GB RAM and Intel Core i5 third-generation 2.8GHz processor).

The tool is available at https://github.com/paheli/indic-viz. A detailed README about how to install and execute the system is provided.

Figures 5 and 6 show the working of the tool.

8 CONCLUSION AND FUTURE WORK

This article utilizes multilingual word embeddings for cross-language information retrieval. It analyzes the performance in detail and proposes a novel cluster-based method to improve the performance. We show how multilingual word clusters help in the task. With the addition of more languages, the communities are well formed, representing coherent information. Additionally, it works toward constructing multilingual word embeddings for major Indic languages and provides a service that shall translate and visualize a query term in multiple languages. Our method mostly concentrates on how word embedding clusters can be useful for the task of cross-lingual information retrieval. We believe that if better word embeddings can be constructed, the performance shall be better. Methods including Xing et al. (2015), MarcoBaroni (2015), Guo et al. (2015), and Barone (2016) that work on the optimization objectives for better cross-lingual word embeddings may be applied for better performance. In this article, we have used the basic methods for obtaining multilingual word embeddings. We have made the tool and the multilingual word embeddings available.

As a future extension, one can incorporate other sources of knowledge like WordNet and see how it can be integrated with multilingual word embeddings for the task of CLIR. We shall also explore other tasks involving the usage of multilingual word embeddings.

Here we shall describe in detail the working of the tool. We have the word vectors and a list of similar words precomputed across different languages. We use these resources to build the interface. The tool has two parts: namely, the word translation part and the word visualization part. We describe the working of the tool subsequently in the Online Appendix.

REFERENCES

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith. 2016. Massively multilingual word embeddings. *CoRR* abs/1602.01925 (2016). Retrieved from http://arxiv.org/abs/1602.01925.

Lisa Ballesteros. 2000. Cross language retrieval via transitive translation. In *Advances in Information Retrieval: Recent Research from the Center for Intelligent Information Retrieval.* Kluwer, Boston, MA. 230–234.

Lisa Ballesteros and Bruce Croft. 1996. Dictionary methods for cross-lingual information retrieval. In *Proceedings of the 7th International Conference on Database and Expert Systems Applications.* Springer-Verlag, London, 791–801. http://dl.acm.org/citation.cfm?id=648309.754278.

Antonio Valerio Miceli Barone. 2016. Towards cross-lingual distributed representations without parallel text trained with adversarial autoencoders. *arXiv Preprint arXiv:1608.02996* (2016).

Paheli Bhattacharya, Pawan Goyal, and Sudeshna Sarkar. 2016. Query translation for cross-language information retrieval using multilingual word clusters. In *Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing*. 152–162.

Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008, 10 (2008), P10008.

Phil Blunsom and Karl Moritz Hermann. 2014. Multilingual Models for Compositional Distributional Semantics. (2014).

Sarath Chandar, Stanislas Lauly, Hugo Larochelle, Mitesh Khapra, Balaraman Ravindran, Vikas C. Raykar, and Amrita Saha. 2014. An autoencoder approach to learning bilingual word representations. In *Advances in Neural Information Processing Systems.* 1853–1861.

Peter A. Chew, Brett W. Bader, Tamara G. Kolda, and Ahmed Abdelali. 2007. Cross-language information retrieval using PARAFAC2. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM, 143–152.

Manoj Kumar Chinnakotla, Sagar Ranadive, Om P. Damani, and Pushpak Bhattacharyya. 2007. Hindi to English and Marathi to English cross language information retrieval evaluation. In *Workshop of the Cross-Language Evaluation Forum for European Languages*. Springer, 111–118.

Raj Dabre, Fabien Cromierès, Sadao Kurohashi, and Pushpak Bhattacharyya. 2015. Leveraging small multilingual corpora for SMT using many pivot languages. In *The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT'15).* 1192–1202. http://aclweb.org/anthology/N/N15/N15-1125.pdf.

Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. 2016. Learning crosslingual word embeddings without bilingual corpora. *CoRR* abs/1606.09403 (2016). http://arxiv.org/abs/1606.09403.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. 462–471.

Marc Franco-Salvador, Paolo Rosso, and Roberto Navigli. 2014. A knowledge-based representation for cross-language document retrieval and categorization. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL'14).* 414–423.

Tim Gollins and Mark Sanderson. 2001. Improving cross language retrieval with triangulated translation. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'01)*. ACM, New York, 90–95. DOI: http://dx.doi.org/10.1145/383952.383965

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. Bilbowa: Fast bilingual distributed representations without word alignments. In *International Conference on Machine Learning (ICML'15)*. 748–756.

Stephan Gouws and Anders Søgaard. 2015. Simple task-specific bilingual word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1386–1390.

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2015. Cross-lingual dependency parsing based on distributed representations. In *ACL (1)*. 1234–1244.

Benjamin Herbert, György Szarvas, and Iryna Gurevych. 2011. Combining query translation techniques to improve cross-language information retrieval. In *Proceedings of the 33rd European Conference on Advances in Information Retrieval (ECIR'11)*. 712–715.

Karl Moritz Hermann and Phil Blunsom. 2013. Multilingual distributed representations without word alignment. *arXiv preprint arXiv:1312.6173*.

Ali Hosseinzadeh Vahid, Piyush Arora, Qun Liu, and Gareth J. F. Jones. 2015. A comparative study of online translation services for cross language information retrieval. In *Proceedings of the 24th International Conference on World Wide Web*. ACM, 859–864.

Kejun Huang, Matt Gardner, Evangelos E. Papalexakis, Christos Faloutsos, Nikos D. Sidiropoulos, Tom M. Mitchell, Partha Pratim Talukdar, and Xiao Fu. 2015. Translation invariant word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP'15)*. Association for Computational Linguistics. 1084–1088.

David A. Hull and Gregory Grefenstette. 1996. Querying across languages: A dictionary-based approach to multilingual information retrieval. In *ACM SIGIR*.

Jagadeesh Jagarlamudi and A. Kumaran. 2007. *Cross-Lingual Information Retrieval System for Indian Languages*. In *Workshop of the Cross-Language Evaluation Forum for European Languages*. Springer, 80–87.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proceedings of COLING 2012*. 1459–1474.

Angeliki Lazaridou, Georgiana Dinu, and Marco Baroni. 2015. Hubness and Pollution: Delving into Cross-Space Mapping for Zero-Shot Learning. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Volume 1*, Long Papers. Association for Computational Linguistics, 270–280. DOI: http://dx.doi.org/10.3115/v1/P15-1027

Gina-Anne Levow, Douglas W. Oard, and Philip Resnik. 2005. Dictionary-based techniques for cross-language information retrieval. *Inf. Process. Manage.* 41, 3 (May 2005), 523–547.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Vol. 2.

Omer Levy, Anders Søgaard, and Yoav Goldberg. 2017. A strong baseline for learning cross-lingual word embeddings from sentence alignments. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, Vol. 1. 765–774.

Michael L. Littmana, Susan T. Dumais, and Thomas K. Landauer. 1998. Automatic cross-language information retrieval using latent semantic indexing. In *Cross-Language Information Retrieval, 1998*. Springer, 51–62.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*. 151–159.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv Preprint arXiv:1301.3781* (2013).

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation. *CoRR* abs/1309.4168 (2013).

Hieu Pham Minh-Thang Luong and Christopher D. Manning. 2015. Bilingual word representations with monolingual quality in mind. In *Proceedings of NAACL-HLT*. Association for Computational Linguistics. 151–159.

Nilesh Padariya, Manoj Chinnakotla, Ajay Nagesh, and Om P. Damani. 2008. Evaluation of Hindi to English, Marathi to English and English to Hindi CLIR at FIRE 2008. In *Working Notes of Forum for Information Retrieval and Evaluation (FIRE'08)*.

Hieu Pham, Thang Luong, and Christopher Manning. 2015. Learning distributed representations for multilingual text sequences. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*. 88–94.

Ari Pirkola. 1998. The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 55–63.

Siyu Qiu, Qing Cui, Jiang Bian, Bin Gao, and Tie-Yan Liu. 2014. Co-learning of word representations and morpheme representations. In *Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers (COLING'14)*. 141–150.

Sebastian Ruder. 2017. A survey of cross-lingual embedding models. *CoRR* abs/1706.04902 (2017). Retrieved from http://arxiv.org/abs/1706.04902.

Shigehiko Schamoni, Felix Hieber, Artem Sokolov, and Stefan Riezler. 2014. Learning translational and knowledge-based similarities from relevance rankings for cross-language retrieval. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Volume 2, Short Papers*, Vol. 2. 488–494.

Artem Sokolov, Felix Hieber, and Stefan Riezler. 2014. Learning to translate queries for CLIR. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR'14)*. 1179–1182.

Philipp Sorg and Philipp Cimiano. 2008. Cross-lingual information retrieval with explicit semantic analysis. In *Working Notes for the CLEF 2008 Workshop*.

Radu Soricut and Nan Ding. 2016. Multilingual word embeddings using multigraphs. *arXiv Preprint arXiv:1612.04732* (2016).

Hubert Soyer, Pontus Stenetorp, and Akiko Aizawa. 2014. Leveraging monolingual data for crosslingual compositional word representations. *arXiv preprint arXiv:1412.6334*.

Ferhan Türe and Elizabeth Boschee. 2014. Learning to translate: A query-specific combination approach for cross-lingual information retrieval. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*. 589–599.

Ferhan Türe, Jimmy Lin, and Douglas W. Oard. 2012a. Combining statistical translation techniques for crosslanguage information retrieval. In *Proceedings of COLING 2012*. 2685–2702.

Ferhan Türe, Jimmy Lin, and Douglas W. Oard. 2012b. Looking inside the box: Context-sensitive translation for cross-language information retrieval. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 1105–1106.

Raghavendra Udupa, K. Saravanan, Anton Bakalov, and Abhijit Bhole. 2009. They are out there, if you know where to look: Mining transliterations of OOV query terms for cross-language information retrieval. In *European Conference on Information Retrieval*. Springer, 437–448.

Shyam Upadhyay, Manaal Faruqui, Chris Dyer, and Dan Roth. 2016. Cross-lingual models of word embeddings: An empirical comparison. *arXiv preprint arXiv:1604.00425* (2016).

Ivan Vulić and Anna Korhonen. 2016. On the role of seed lexicons in learning bilingual word embeddings.

Ivan Vulić and Marie-Francine Moens. 2015. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM. 363–372.

Chih-Ping Wei, Christopher C. Yang, and Chia-Min Lin. 2008. A latent semantic indexing-based approach to multilingual document clustering. *Decision Support Systems* 45, 3 (June 2008), 606–620. DOI:http://dx.doi.org/10.1016/j.dss.2007.07.008

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *HLT-NAACL*. 1006–1011.

Wen-tau Yih, Kristina Toutanova, John C. Platt, and Christopher Meek. 2011. Learning discriminative projections for text similarity measures. In *Proceedings of the 15th Conference on Computational Natural Language Learning (CoNLL'11)*. Association for Computational Linguistics, 247–256.

Received September 2017; revised December 2017; accepted April 2018