International Journal of Information Science and Management
Vol. 21, No. 1, 2023, 161-180
DOI: 10.22034/ijism.2022.1977779.0 / https://dorl.net/dor/20.1001.1.20088302.2023.21.1.9.9

Original Research

Training vs. Post-training Cross-lingual Word Embedding Approaches: A Comparative Study

Masood Ghayoomi
Assistant Prof. Faculty of Linguistics
Institute for Humanities and Cultural Studies, Tehran, Iran.
M.Ghayoomi@ihcs.ac.ir
ORCID iD: https://orcid.org/0000-0001-6685-1332

Received: 02 April 2022
Accepted: 26 June 2022

Abstract

This paper provides a comparative analysis of cross-lingual word embedding by studying the impact of different variables on the quality of the embedding models within the distributional semantics framework. Distributional semantics is a method for the semantic representation of words, phrases, sentences, and documents. This method aims at capturing as much information as possible from the contextual information in a vector space. The early study in this domain focused on monolingual word embedding. Further progress used cross-lingual data to capture the contextual semantic information across different languages. The main contribution of this research is to make a comparative study to find out the superior impact of the learning methods, supervised and unsupervised in training and post-training approaches in different embedding algorithms, to capture semantic properties of the words in cross-lingual embedding models to be applicable in tasks that deal with multi-languages, such as question retrieval. To this end, we study the cross-lingual embedding models created by BilBOWA, VecMap, and MUSE embedding algorithms along with the variables that impact the embedding models' quality, namely the size of the training data and the window size of the local context. In our study, we use the unsupervised monolingual Word2Vec embedding model as the baseline and evaluate the quality of embeddings on three data sets: Google analogy, mono- and cross-lingual words similar lists. We further investigated the impact of the embedding models in the question retrieval task.

Keywords: Semantic Word Representation, Cross-lingual Context, Vector Space Model, Distributional Semantics.

Introduction

Various approaches are proposed to represent the semantic information and to show how concepts are conveyed from one person to another or from one language to another language, including using predicate logic (first-order logic), semantic network, conceptual dependency, frame-based representation (frame semantics), and vector representation, known as vector space model (Jurafsky & Martin, 2000: 502, 538, 647). The vector representation has shown great flexibility in capturing the semantic information of the words and representing it numerically, known as word embedding. This type of representation has two advantages

(Mikolov, Sutskever, Chen, Corrado, & Dean, 2013; Chaudhary, Zhou, Levin, Neubig, Mortensen & Carbonell, 2018): a) due to the numerical representation of semantic information, the representation is independent of any word form in any language; b) the vector representation adds flexibility to the data to use a wide range of algorithms. In Natural Language Processing (NLP) tasks, word embedding can be utilized to model a language. One issue that should be considered is the quality of the embedded information in the vector space to be applicable in tasks like machine translation or question retrieval.

In this paper, we look closely at the semantic information of the words that can be captured in the vectors from a monolingual or cross-lingual Corpus. There is no doubt that cross-lingual embedding contains enriched vectors, and it reduces ambiguities that, is important in tasks like machine translation (Ha, Niehues, Sperber, Pham & Waibel, 2018; Artetxe, Labaka & Agirre, 2018a), entity linking, and parallel sentence mining (Pan, Gowda, Ji, May & Miller, 2019), and dialog systems (Schuster, Gupta, Shah & Lewis, 2019). To this end, a parallel bi- or multilingual Corpus is required. Ruder, Vulic and Søgaard (2019) stated two reasons for cross-lingual word embedding: a) comparing words' meaning across languages; b) enabling a transfer model from rich-resource to low-resource languages. Furthermore, cross-lingual word embedding can be used in cross-lingual information retrieval (HajiAminShirazi & Momtazi, 2020).

This paper contributes to a comparative study on the impact of the training and post-training approaches in cross-lingual word representation. The outcome of this research is beneficial when one aims to use an embedding method and faces a different algorithm. Semantic word representation models are divided into two main categories: a) static word embedding, such as Word2Vec (Mikolov et al., 2013), and b) contextualized transformer-based embedding, such as BERT[1] (Devlin, Chang, Lee & Toutanova, 2019). The advantage of contextualized transformer-based embedding over static word embedding is that based on each local context of a target word, a vector is created to capture the semantic properties better. In contrast, in static word embedding, only one vector is created for each word in all contexts. In addition, although contextualized transformer-based embeddings obtained state-of-the-art results in the field, training such models is very time-consuming and requires advanced hardware resources, such as Tensor Processing Unit (TPU), due to the complex structure of their neural network models. This problem intensifies when low-resource languages are studied, or a domain-specific model is required. Hence, considering the promising results of static models, they are still prevalent in research studies for low-resource languages, and it is impossible to ignore their rapid model development. In this paper, we use static embedding and not contextualized embedding algorithms. The evaluation data sets used to compare the models, such as Google analogy and Wordsim353, are decontextualized and developed for evaluating static models, which provide one vector for each word.

Since developing data annotation requires human intervention, any method that uses this data is known as 'supervised', such as BilBOWA (Gouws, Bengio & Corrado, 2015), which uses a sentence-aligned bilingual Corpus to create words' vectors. It is also possible to post-train two independent monolingual embeddings created from two separate monolingual corpora to develop cross-lingual embeddings of the words, known as 'unsupervised', such as VecMap (Artetxe, Labaka, & Agirre, 2017) or MUSE (Conneau, Lample, Ranzato, Denoyer & Jégou, 2018). It is also possible to use a bilingual dictionary in addition to the monolingual embeddings to create cross-lingual embeddings. This process is also considered as 'supervised' due to using the bilingual dictionary. We further study the impact of the quantity of the training data as a parameter on the quality of the vector representation and the parameter related to the vectors, such as the window size of the neighboring words used in the vectorization process. Furthermore, we investigate the achievement of this research in a cross-lingual question retrieval task for external evaluation of the embedding quality.

Research Background

Several algorithms are proposed to encode the semantic properties in a vector space model. This resulted in two research directions: one is the embedding algorithm by using either supervised or unsupervised machine learning approaches, which is correlated with hardware equipment and available data sources, and the other is the parameters of the algorithms. The current paper contributes to quality checking to compare the impact of embedding algorithms and their parameters. In this section, we review the papers that have studied the embedding properties; in the next section, the distributional representation and the relevant algorithms are introduced and reviewed.

Hadifar and Momtazi (2018) studied the impact of the corpus domain on word embeddings. To this end, they used four monolingual Persian corpora in different domains, including Persian Twitter, Persian Wikipedia, Hamshahri Corpus (AleAhan et al., 2009), and irBlog (AleAhan et al., 2016), to create the embedding models. Furthermore, two English test sets originally developed by Mikolov et al. (2013), namely Google analogy and Wordsim353, were adapted for Persian. According to the experimental results, the domain of the training corpus has a direct impact on the quality of the embedding model.

Zahedi et al. (2018) studied the impact of the static embedding algorithm, including variations of Word2Vec and FastText[2], such as Continuous Skip gram (Skip-gram) and Continuous Bag Of Words (CBOW) models (Mikolov et al., 2013), in addition to the Global Vector representation (GloVe)(Pennington, Socher & Manning, 2014). They used the developed Google analogy test for internal evaluation of the embedding models. The two other metrics they studied were the window size and the number of dimensions. This study used the Wikipedia dump of Persian to develop the embedding models.

In addition to the above studies that focused on analyzing monolingual embedding models for Persian, the following research studies have focused on cross-lingual models. Camacho-Collados, Pilehvar, Collier & Navigli (2017) introduced a dataset to measure the semantic similarity of word pairs cross-lingually. They targeted five languages: English, German, Italian, Spanish, and Persian. They manually developed a set of 500-word pairs with their relevance score as the seed dataset. English Wikipedia was used as the core language, and all data in this language was translated into other languages automatically to be used for embedding. For Persian, movie subtitles were used. The Pearson and Spearman correlation coefficient was used as evaluation metrics.

Doval, Camacho-Collados, Espinosa Anke and Schockaert (2020) studied the robustness of unsupervised and semi-supervised cross-lingual word embedding models. In their study, three corpora, including Wikipedia, Web corpora, and social media, were used for creating the embedding models of the languages, including English, Spanish, Italian, German, Finish, Russian, and Persian. In their study, the embeddings by VecMap (Artetxe et al., 2017; Artetxe, Labaka, Agirre, & Cho, 2018b), MUSE (Conneau et al., 2018), Meemi (Doval, Camacho-Collados, Espinosa-Anke & Schockaert, 2018) were studied, where English was used as the source language. They evaluated the embedding models internally by using Spearman correlation coefficient for various cross-lingual word embedding models in the cross-lingual word similarity task developed by Camacho-Collados et al. (2017), and externally in a bilingual dictionary induction task. According to the results, the unsupervised mode of VecMap was the most robust embedding model.

Espinosa-Anke, Palmer, Corcoran, Filimonov, Spasić and Knight (2021) studied the development of the English-Welsh cross-lingual embedding model. They used Word2Vec (Mikolov et al., 2013) and FastText to develop monolingual static embeddings from Wikipedia for the two languages. Then, they used three cross-language alignment strategies: cosine similarity, inverted softmax, and cross-domain similarity local scaling. After the internal evaluation, they evaluated the embedding models externally in the sentiment analysis task.

The differences between our current study with cross-lingual analysis are in two issues: (1) Camacho-Collados et al. (2017) and Dova et al. (2020) focused on post-training cross-lingual embedding. Still, in our research, we make a more comprehensive study of training and post-training cross-lingual embedding models. It should be mentioned that in training approaches, a bilingual corpus is required, which has not been taken into consideration in post-training approaches; (2) Camacho-Collados et al. (2017) and Dova et al. (2020) did not study the impact of different variables; while we aim to study the variables that are effective on creating embedding models for English and Persian; e.g., the window size.

Distributional Representation

Distributional semantics is based upon the "distributional hypothesis." This hypothesis is formed based on the idea proposed by Harris (1954), who explores that different words utilized in the same context tend to have similar meanings. Harris pays attention to the local context and believes that the meaning of a word is reflected in the context. This idea has been previously proposed by Wittgenstein (1953), who says that "the meaning of words lies in their use." Firth (1957) emphasizes Harris' idea and adds that "[y]ou shall know a word by the company it keeps." These ideas indicate that contextual information plays a vital role in determining the meaning of a word. As a result, Miller and Charles (1991) proposed a robust contextual hypothesis that expresses that "two words are semantically similar to the extent that their contextual representations are similar." In this statement, a great emphasis is made on the local context of the words.

Two general approaches are used to represent the contextual information of the distributional semantics (Song, Wang, Mi & Gildea, 2016): Bayesian methods using topic modeling approaches; and feature-based methods using the vector representation of the contextual information. While topic modeling approaches, such as latent Dirichlet allocation (Blei, Ng & Jordan, 2003) and hierarchical Dirichlet process (The et al., 2006), represented successful results in various NLP tasks, the flexibility of vector space models has received researchers' attention to capturing words' meanings in the vector space.

The vector space model exploited in information retrieval (Salton, Wong & Yang, 1975) contributes to distributional semantics to represent information about a word and its context. In other words, compressing the words' information and their contexts in vectors explores the semantic distribution of the words. This way of representing word information in the literature is known as 'word embedding' (Mikolov et al., 2013). Computing the geometric distance between the vectors results in the similarity between the words.

Monolingual word embedding was the first attempt toward a text's distributional representation, which has been widely studied in recent years. The model was proposed by Mikolov et al. (2013) and then exploited in various NLP applications. The representation of words in the vector space was further extended in various dimensions. One of the main extensions of word embedding is the cross-lingual embedding of words (Ruder et al., 2019). This approach represented words' semantic information from two (or more) languages in the same vector space. It was successfully used in various cross-lingual NLP applications, including cross-lingual part-of-speech tagging (Gouws & Sogaard, 2015), dependency parsing (Guo, Che, Yarowsky, Wang, & Liu, 2015), document classification (Shi, Liu, Liu & Sun, 2015; Xu, Ouyang, Ren, Wang & Jiang, 2018), and machine translation (Gu, Hassan, Devlin, & Li, 2018a; Gu, Wang, Chen, Cho & Li, 2018b).

Word Embedding

Mono-lingual Word Embedding:

Recent studies have considered the word embedding approach to building the words' vectors. This approach's promising results caused researchers to propose different techniques to achieve high-quality vectors.

The available studies on monolingual word embedding lay in two different dimensions: a) the approach that is used for building embeddings, and b) the type of context that is used for discovering word relations. Recently two approaches have been widely studied to model the contextual information: a) using the matrix decomposition techniques: GloVe (Pennington et al., 2014) is an unsupervised learning method that follows this approach to provide the distributional representation of words. And b) using neural network-based techniques. Skip-gram and CBOW models (Mikolov et al., 2013) use this approach to represent the contextual information of words in vectors.

Precise encoding of the word's contextual information directly impacts finding the most similar words. Since the context plays a significant role, Peirsman and Geeraerts (2009) introduced three types of linguistic contexts: a) document-based model: the words which are used in the same paragraph or the same documents are similar (Landauer & Dumais, 1997; Sahlgren, 2006); b) syntax-based model: words are compared according to their syntactic relations, more precisely using the dependency relations (Harris, 1991; Lin, 1998; Pado and Lapata, 2007; Levy & Goldberg, 2014), or the combinatory categorial grammar (Hermann & Blunsom, 2013); and c) window-based model: it extracts word-word co-occurrence statistics from a large corpus. These word co-occurrences resemble the Bag-Of-Words (BOW) model (Sahlgren, 2006).

Cross-lingual Word Embedding:

Static word embedding models have discovered semantic relations of words in various languages with promising results in NLP applications. However, word embeddings of two different languages are not comparable; and cross-lingual NLP applications, such as cross-lingual information retrieval, cannot benefit from these vector representations. This shortcoming motivated researchers to focus on this topic to provide distributional word representation across languages. Such embeddings help researchers in the following aspects:

a) To compare words' meanings in different languages;

b) To transfer trained models from one language to another language, mainly from dominant languages to low-resource languages;

c) To enhance the performance of cross-lingual NLP tasks, such as machine translation and cross-lingual information retrieval.

To reach the goal, parallel or comparable corpora are required. The main classification of the developed model is based on the type of resource that is available for mapping words in two languages. The available researches mainly focus on the following types of textual resources:

a) Word-aligned resources;

b) Sentence-aligned parallel corpora;

c) Document-aligned comparable corpora.

Like monolingual word embedding, matrix factorization and Skip-gram models are the main approaches for cross-lingual word embedding. Moreover, other approaches, such as auto-encoders, can be used for the task.

The idea of bilingual auto-encoders is reconstructing a sentence in the target language from the corresponding sentence in the source language instead of minimizing the distance between two sentence representations. To this end, a sentence from the source language is encoded as the sum of its word embeddings. The auto-encoder is trained using a binary BOW (Chandar et al., 2014) or hierarchical softmax (Lauly, Boulanger & Larochelle, 2014).

The idea of bilingual matrix factorization is based on the sparse monolingual representation. Vyas and Carpuat (2016) proposed a model based on sparse coding. Similar to other approaches, they benefited from both monolingual and cross-lingual contexts. They argued that their proposed model has more flexibility in representing data than the dimensionality reduction techniques, such as principle component analysis.

The Skip-gram model for cross-lingual word embedding follows the same idea as the monolingual word embedding, such that in addition to predicting neighboring words by the current word in the source language, all words in the target language should also be predicted by that word. Like monolingual embedding, negative sampling is also used in cross-lingual word embeddings. BilBOWA (Gouws et al., 2015), Trans-gram (Coulmance, Marty, Wenzek & Benhalloum, 2015), and BiSkip (Luong, Pham, & Manning, 2015) are the leading researchers that use this approach.

In the proposed model by Gouws et al. (2015), called Bilingual Bag-of-Words without Alignments (BilBOWA), the Skip-gram model and negative sampling were used to create words' vectors. The window size was set to 8 words in the local context. For the cross-lingual objective, they used a naïve assumption such that any word in a sentence of the source language has the potential to be considered in the context of any word in the aligned sentence of the target language.

VecMap is a Python library used for cross-lingual word embeddings mappings by utilizing either a supervised learning method (Artetxe et al., 2017) or an unsupervised one (Artetxe et al., 2018b). A self-learning approach for embedding mapping and dictionary induction technique is implemented in the core algorithm of this embedding method. The supervised mode of this embedding method uses a bilingual dictionary developed through a word translation process to better map monolingual embedded words in the vector space. In contrast, this bi-lingual dictionary is not used in the unsupervised mode.

Multilingual Unsupervised and Supervised Embeddings (MUSE) is a Python library for multilingual word embedding proposed by Conneau et al. (2018). Like VecMap, this embedding method does not require a parallel corpus; therefore, the embeddings are created from two independent monolingual corpora and are enhanced to represent two languages in the same vector space using post-trainings. MUSE tries to rotate the vector space in the source language on the vector space of the target language. A supervised or unsupervised machine learning approach is used to reach the goal. In the supervised approach, the predefined bilingual dictionaries for a set of 30 languages with a size of 500 words are used to find out the coefficient scores to rotate the vector spaces. Figure 1 shows the rotation of vector X over vector Y graphically. In the unsupervised mode, this multilingual dictionary is not used.

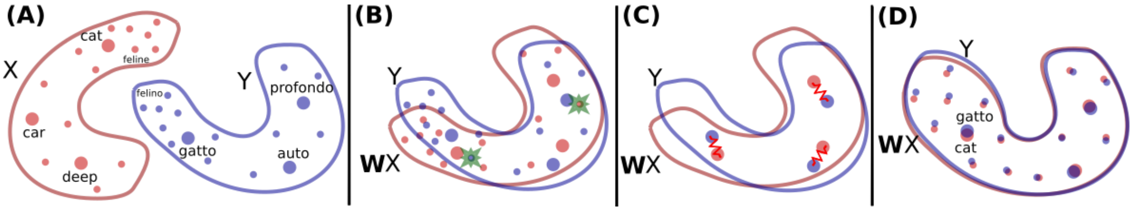

Figure 1: Rotation of vector spaces in MUSE (Conneau et al., 2018)

Duong, Kanayama, Ma, Bird & Cohn (2016) proposed an extended CBOW model for cross-lingual word embeddings without using bilingual corpora; in Duong, Kanayama, Ma, Bird & Cohn (2017), they proposed another model to use multilingual data for cross-lingual word embeddings.

Proposed Framework

In this research, we aim to provide a comparative study about the quality of the vector representations of the semantic information captured cross-lingually in the training scenario, such as BiLBOWA, as well as post-training scenarios, such as Vecmap, and MUSE. Then, we use the created embeddings in the cross-lingual question retrieval task for external evaluation. We select the Persian and English languages in our study; as a result, to create the embedding models, we require both monolingual and cross-lingual Persian and English datasets.

Method

Figure 2 represents the overall framework of our research. The main components of the framework are cross-lingual embedding methods. The other essential parts are datasets, consisting of a monolingual corpus for each of the two languages and bilingual data, either a bilingual corpus or a bilingual dictionary.

Embedding methods:

To create the words' vectors in the supervised scenario, we use the BiLBOWA model (Gouws et al., 2015) to create embedding in the training phase. For the supervised and unsupervised scenarios of cross-lingual embedding, we used VecMap (Artetxe et al., 2017; 2018b) and MUSE (Conneau et al., 2018), which provide bilingual vectors based on a post-training process. We require a sentence-aligned parallel corpus and the monolingual Corpus to perform the experiments.

Embedding parameters:

Some variables have an impact on the quality of the vector representation. Some variables are corpus-based, such as the quantity of the data, and there are model-based parameters, such as the window size of the neighboring words used in the vectorization process. The experiments can be run using training data in different sizes such that instead of combining the resources as one Corpus, separated corpora can be used. The window size that impacts capturing the semantic information about the words is set to 2, 4, 8, 10, 20, 25, 30, 35, and 40 neighboring words. We set the vector size for word embedding as 300 dimensions to embed the local context information.

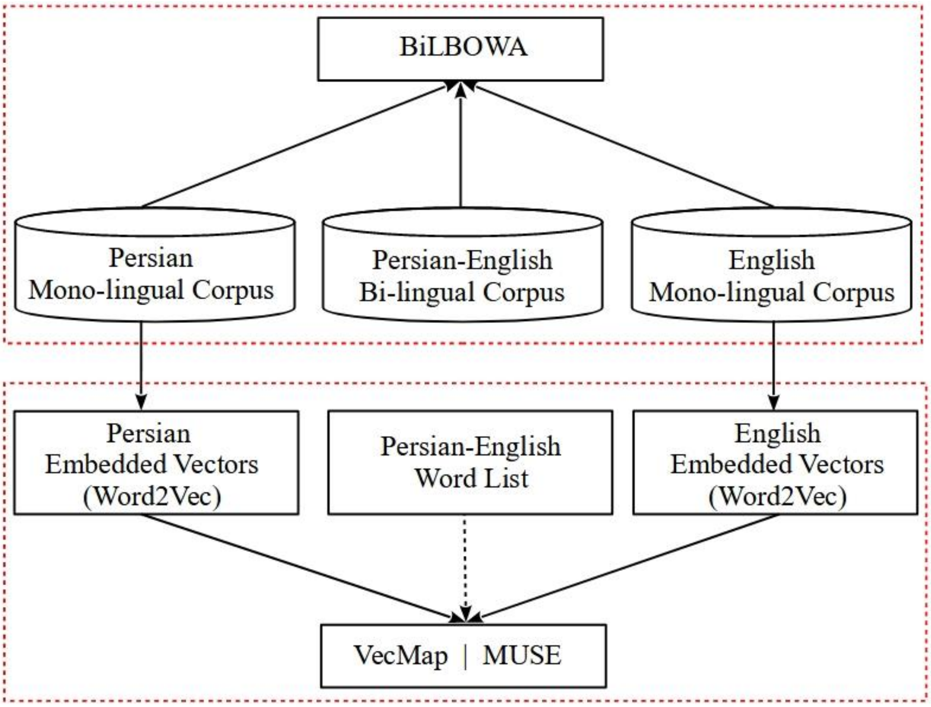
BiLBOWA
Persian Mono-lingual Corpus
Persian-English Bi-lingual Corpus
English Mono-lingual Corpus
Persian Embedded Vectors (Word2Vec)
Persian-English Word List
English Embedded Vectors (Word2Vec)
VecMap | MUSE

Figure 2: The Proposed Framework

Data Sets:

In the supervised scenario, either a bilingual dictionary or a parallel corpus minimally aligned at the sentence level is required; therefore, the data is not raw anymore. In the development of a parallel corpus, human intervention is required. It should be added that this type of data differs from the comparable Corpus collected automatically. Although several Persian-English comparable corpora exist, a limited number of parallel corpora are briefly introduced in the following.

The Amirkabir bilingual Persian-English Corpus (Jabbari, Bakshaei, Ziabary & Khadivi, 2012), named Aut, is a corpus constructed from a large amount of data collected through crawling websites. Then, parallel and comparative texts are extracted from this data. In addition to the web data, other data sources, including books, software manuals, laws in several countries, and news from CNN and BBC news agencies, are manually translated and added to this data. Persian-English movie subtitles are also downloaded and added.[3] Additionally, Persian-English phrases and sentences in the VerbMobil Project (Ney, Och & Vogel, 2000) required by tourists are added to this data. In the next step, the alignment of sentences belonging to the comparative texts, i.e., news and movie subtitles, is done automatically.

Targoman, after that called Tar, is another Persian-English parallel corpus developed at the

Information and Communications Technology Research Institute (ICTRI)[4]. This data is developed within a project on machine translation. The domain of the data is news such that 269،329 sentences from CNN are translated manually.

Ansari, Sadreddini, Tabebordbar & Wallace (2014) developed the Persian-English Parallel Corpus, called Pepc, from Wikipedia; then they developed a comparative corpus by crawling 1200 Persian-English articles, and they aligned the data at the sentence level.

The statistical information of the Persian and English sections of the parallel corpora used in this research is summarized in Tables 1 and 2. The shortage of available data indicates the difficulty of developing a language resource; therefore, the research using an unsupervised approach seems preferable.

In the post-training scenario, a monolingual Corpus is required. The complexity of data development in this scenario is not high because independent monolingual corpora are required. There is a free, large, open-domain corpus for English collected from a dump of English Wikipedia articles in April 2010. This Corpus is composed of more than 990 million tokens, called the Westbury Lab Wikipedia Corpus (Shaoul & Westbury, 2010). It should be mentioned that documents with less than 2000 characters long are excluded from this Corpus.

Table 1
Statistical Information of Persian-English Corpus (Persian Section)

Corpus Name	Provider	Sentences	Persian Words		
			Token	Type	Freq. 1
Amirkabir University Fa-En Bilingual Corpus	Jabbari et al. (2012)	2164408	37043534	426347	232730
Targoman Fa-En Bilingual Corpus	ICTRI	269329	6755234	125577	69233
Fa-En Parallel Corpus	Ansari et al. (2014)	199936	4194187	159426	80484
Total		2633673	47992955	536179	262396

Table 2
Statistical Information of Persian-English Corpus (English Section)

Corpus Name	Provider	Sentences	Persian Words		
			Token	Type	Freq. 1
Amirkabir University Fa-En Bilingual Corpus	Jabbari et al. (2012)	2164408	36762192	310484	148261
Targoman Fa-En Bilingual Corpus	ICTRI	269329	6428089	122447	55770
Fa-En Parallel Corpus	Ansari et al. (2014)	199936	4456994	123795	53599
Total		2633673	47992955	395980	179813

The Persian monolingual Corpus is composed of several other corpora: a) Persian Linguistic DataBase (PLDB) (Assi, 1997) is a balanced Persian corpus containing both historical and contemporary Persian. In this research, we only use the current section of the Corpus; b) The Newspaper Corpus, which is a collection of news crawled from the online archive of several Persian newspapers (Ghayoomi, 2019); c) The Hamshahri Corpus (AleAhmad, Amiri, Darrudi, Rahgozar & Oroumchian 2009) which is also another corpus in the news domain collected from the online archive of the Hamshahri Newspaper. d) The

Bijankhan Corpus (Bijankhan, 2004), which is a fraction of Peykare (Bijankhan, Sheykhzadegan, Bahrani & Ghayoomi, 2011), the Persian Text Corpus; e) The Persian Wikipedia corpus, which contains 361,479 articles downloaded from the dump of Persian Wikipedia articles in July 2016.[5] This collected data contains over 538 million tokens. The statistical information of this dataset, after that called bigMono, is reported in Table 3.

Table 3
Persian Resources of the Mono-lingual Corpus (bigMono Corpus)

Corpus Name	Provider	Word Token	Word Type
PLDB	Assi (1997)	23848655	307726
Newspaper Corpus	Ghayoomi (2019)	301186277	1324317
Hamshahri Corpus	AleAhmad et al. (2009)	157841123	608503
Bijankhan Corpus	Bijankhan (2004)	2602536	77143
Persian Wikipedia	Wikipedia	80995743	956655
Total		**538586487**	**1957541**

In addition to the monolingual raw data, a bilingual Persian-English dictionary is required to map the words from one language to another. A small dictionary contains 5,000 words, and it is available online.[6] Furthermore, we use the Word2Vec toolkit to create monolingual embedding by using 8 words of the local context in 300 dimensions of the Skip-gram model. It needs to be added that in MUSE, it is possible to rotate the source language on the target language or vice-versa. In our experiments, we also investigate the impact of the source and target languages on creating the embedding models by MUSE.

Evaluation

Based on the described settings in the previous section, we compare the quality of the vectors created within the supervised or unsupervised scenario of the training or post-training approaches. To this end, we compare the vectors of the configurations against two well-known monolingual data sets: Google analogy and Wordsim353. Mikolov et al. (2013) introduced an analogy data set developed at Google with four-arity in the form of A:B::C:D. In this data set, the analogy of the instance A to B is similar to instance C to D. To use the data in the evaluation, the fourth instance, D, is missed, and the system should guess the most similar instance from the words' vectors that have the minimum distance-based cosine similarity score. In this data set, 14 types of relations are defined, and in total, 19,544 English analogies are defined. Hadifar and Momtazi (2018) translated the English dataset into Persian using the Google translate tool and manual correction. This process resulted in defining 15,763 analogies for Persian.

Finkelstein et al. (2002) developed a gold dataset that contains 353-word pairs with their semantic relations scored from 1 to 10. The cosine similarity metric between two-word embeddings was computed and listed according to the similarities. The ranked list based on their gold similarity and the estimated similarity based on the embedding were compared using the Pearson and Spearman correlations. The data is organized in three columns. Hadifar and Momtazi (2018) manually translated the English word pairs into Persian and selected the best translation based on the agreement between the translations. This process resulted in a set of 285-word pairs in Persian to evaluate the monolingual embedding quality, i.e., Persian.

Camacho-Collados et al. (2017) developed ten bilingual datasets for English, German, Spanish, Italian, and Persian such that the bilingual word pairs are organized in three columns, similar to Finkelstein et al. (2002) and Hadifar and Momtazi (2018) with this difference that the headword is in a language, for instance, English. Its similar word is in another language, for instance, Persian, and their semantic similarity scored 0 to 5 with a step size of 0.25. We use the English-Persian dataset, which contains 703 pairs, to evaluate the quality of both supervised and unsupervised cross-lingual embedding models.

Table 4 summarizes and compares the obtained results of the embedding models using the bigMono Corpus, either Persian or English, with the window size 8 in 300 dimensions. We considered the Word2Vec model the baseline using either Persian or English raw plain corpus. Although it is possible to evaluate the monolingual embedding model with the monolingual Persian Wordsim data, it is not possible to evaluate it with the cross-lingual Wordsim data; This determines the shortcoming of this embedding method. Utilizing a cross-lingual embedding method can be a solution.

As seen in the table, the BilBOWA model that uses parallel data cannot beat the baseline based on the Google analogy score; This determines that accessing a high amount of cross-lingual data to create the model from scratch cannot end in an acceptable performing model. Among the two post-training embedding models, namely VecMap and MUSE that take monolingual embedding of languages separately, MUSE performed the best according to the Google analogy score and had a slightly better performance than the baseline, but the difference was not significant. Although selecting the learning method (supervised vs. unsupervised) and changing the source and the target languages have no impact on the quality of embeddings in MUSE, the unsupervised mode has a slightly better performance than the supervised mode in VecMap. When cross-lingual Wordsim data is used for evaluating VevMap, the unsupervised model performs better with a significant difference according to the two-tailed *t*-test ($p < 0.05$). Comparing the results of the models created based on post-training, we saw no impact by changing the source and target languages when monolingual data is used for evaluation.

Meanwhile, the MUSE model that used an unsupervised method with the rotation of the English vector space over the Persian vector space performed the worst when cross-lingual Wordsim data was used in the evaluation. This result determines the loss of the cross-lingual concept. According to the reported results in Table 4, the overall comparison between the embedding methods indicates that the MUSE model performed the best.

Table 4
Comparing embedding methods using bigMono Corpus, window size 8 in 300 dimensions

Embedding Method	Mode	Google Analogy	Mono-ling. Wordsim		Cross-ling. Wordsim	
			Pearson	Spearman	Pearson	Spearman
Word2Vec	unsup	40.18	54.31	57.37	-	-
BilBowa	sup	21.52	20.95	22.98	8.62	21.29
VecMap	sup	36.18	52.71	55.52	18.03	25.25
	unsup	37.64	54.24	56.51	55.62	54.63
MUSE	Sup (Fa-En)	40.52	54.12	57.12	54.46	53.38
	unsup (Fa-En)	40.52	54.12	57.12	53.55	52.27
	Sup (En-Fa)	40.52	54.12	57.12	54.00	52.93
	unsup (En-Fa)	40.52	54.12	57.12	8.81	6.03

In the next step of our experiments, we investigated the impact of two variables on embeddings, namely the data size and the window size. Since Persian is the target language, the embedding model, which used either the supervised or unsupervised approaches, rotates the Persian vector space over the English vector space. The window size and the vector size were set to 8 in 300 dimensions, respectively, To study the impact of the corpus size. The results of the supervised model (sup (Fa-En)) and the unsupervised model (unsup (Fa-En)) are reported in Tables 5 and 6.

Table 5
Comparing the performance of MUSE embedding method (sup (Fa-En)) using different data sizes for window 8 in 300 dimensions

Data Size	OOV	Google Analogy	Mono-ling. Wordsim		Cross-ling. Wordsim	
			Pearson	Spearman	Pearson	Spearman
bigMono	0.00	40.52	54.12	57.12	54.46	53.38
AutTarPepcMono	0.00	30.67	55.50	58.19	50.73	49.81
AutMono	2.81	28.57	53.54	54.80	47.59	46.27
TarMono	7.02	25.47	48.67	49.75	31.03	28.47
PepcMono	1.40	14.37	51.65	55.11	38.95	37.59
AutTarMono	2.46	28.68	54.41	56.21	47.03	46.35

Table 6
Comparing the performance of MUSE embedding method (unsup (Fa-En)) using different data sizes for window 8 in 300 dimensions

Data Size	OOV	Google Analogy	Mono-ling. Wordsim		Cross-ling. Wordsim	
			Pearson	Spearman	Pearson	Spearman
bigMono	0.00	40.52	54.12	57.12	53.55	52.27
AutTarPepcMono	0.00	30.67	55.50	58.19	51.08	50.40
AutMono	-	-	-	-	-	-
TarMono	-	-	-	-	-	-
PepcMono	-	-	-	-	-	-
AutTarMono	2.46	28.68	54.41	56.21	47.26	46.72

As can be seen in Tables 5 and 6, the bigMono model that used the big Persian and English monolingual corpora outperformed the other models based on the Google analogy score without facing the Out Of Vocabulary (OOV) problem. Merging the Persian and English parts of the Aut, Tar, and Pepc corpora together created a corpus called AutTarPepcMono for embedding. This Corpus is used for embedding. Although there is no OOV problem, the Google analogy score reduced significantly by almost 10%. The data size reduction has almost no impact on the evaluation using monolingual Wordsim data. Still, approximately 4% reduction was achieved for the evaluation using cross-lingual Wordsim data compared to the upper-bound, the bigMono model. This result determines the impact of size, but due to the time consumption to create the vector space models, a relatively good result is achieved in the model.

We also studied the impact of each bilingual Corpus to build the models. To this end, each bilingual Corpus was used separately. The Pepc, built from Wikipedia and has less OOV problem compared to other models, obtained a severe score reduction for Google analogy. The scores for monolingual and cross-lingual Wordsim data were also reduced. Although the model that used the Tar corpus, the TarMono model, had a high rate of OOV problem, it had a much better performance than the PepcMono model; This showed that the content of the Google analogy data was closer to the Tar corpus than the Pepc corpus. The Aut corpus, larger than the two other corpora, had the OOV problem by 2.81%. This dataset performed better than the other bilingual corpora based on the three evaluation metrics.

Among the individual models created from the bilingual Corpus, the Aut and Tar performed better than the Pepc corpus. We merged the Aut and Tar corpora and created a corpus called AutTarMono. Reducing the corpus size compared to the AutTarPepcMono model caused the OOV problem by 2.46%; increasing the corpus size in comparison to the AutMono reduced the OOV problem by 0.35%. This reduction had about a 2% reduction of the Google analogy score compared to the AutTarPepcMono model that determines the slight impact of the Pepc corpus on the result. Adding the Tar corpus to the Aut corpus had almost no impact on the analogy score and mono- and cross-lingual Wordsim data.

Similar results to the supervised model were obtained in the unsupervised mode of the embedding model. The only difference was that the training data should be in a specific size to create the model. Since the individual bilingual corpora, namely Aut, Tar, or Pepc, did not rich the size, the models were not created. Based on the results, it can be concluded that the corpus size, either in supervised or unsupervised modes, matters in creating a model concerning the cross-lingual properties; therefore, the more significant the amount of training data, the better the results.

According to the results in Table 5, the MUSE embedding method that used bigMono data performed the best. This model used the supervised approach to rotate the vector space of Persian on the English vector space.

In Table 7, the impact of the window size on the quality of embeddings from Windows 2 to 40 words was investigated. Although enlarging the window size increased the local context and encoded many word information in the vectors, too broad a context hurt the embedded information. As can be seen in the table, window size 25 performed the best according to the Google analogy score and the monolingual Wordsim data. But the window size 10 obtained a slightly better result for cross-lingual Wordsim data; This determines that languages such as Persian require a significant context to capture the semantic properties of the words in the vector space, which is not the case for other languages like English. We also compared the best performance of MUSE, i.e., window size 25, and the Word2Vec model. The Google analogy score of Word2Vec for this window size was 42.63% which determined that MUSE still performed the best but without a significant difference.

Table 7
Comparing performance of MUSE embedding method (sup (Fa-En)) using bigMono data in different window sizes in 300 dimensions

Window Size	Google Analogy	Mono-ling. Wordsim		Cross-ling. Wordsim	
		Pearson	Spearman	Pearson	Spearman
2	32.80	50.62	53.06	52.63	52.03
4	36.99	52.78	55.72	54.39	53.40
8	40.52	54.12	57.12	54.46	53.38
10	42.04	54.38	57.55	**54.74**	**53.63**
20	42.51	55.66	58.98	53.66	52.51
25	**43.02**	**55.98**	**59.34**	52.75	51.74
30	41.93	55.09	58.27	52.28	51.36

35	42.07	55.49	58.86	52.35	51.61
40	42.30	55.78	59.34	51.40	50.46

Discussion

To show the superiority of the models and the studied variables, we used the models with window size 8 as the local context of the words and utilized the embedding models created by BiLBOWA, supervised and unsupervised modes of VecMap and MUSE in a question retrieval application. Cross-lingual question retrieval aims to retrieve similar questions in the queries expressed by the user in another language.

To this end, we used the question retrieval corpus developed by HajiAminShirazi and Momtazi (2020) as the test data in our experiments. In this evaluation dataset, firstly, 103908 English questions from 6 different Stack Exchange community question-answering platform categories were crawled. Then, 100 questions in each category were translated into Persian by Google translation tool and manually corrected.

In this evaluation, we represent the cross-lingual question retrieval corpus as the test data based on the developed embedding models. The models should retrieve as many similarities as possible in questions in another language simply by using the cosine similarity metric. It should bear in mind that this evaluation aims to study the impact of the embedding models on a task, such as question retrieval, and improving the model for question retrieval is out of the scope of this research.

Two metrics are used in the evaluation. Mean Reciprocal Ranks (MRR) is a metric that finds the 10 most similar questions in the target language and assigns a ranking weight at the retrieved question. S@*k* is a metric that takes the success of the *k* most similar retrieved questions without considering the weight. Table 8 represents the results of the various embedding models when the whole Corpus retrieves the questions in the target language. Additionally, Table 9 reports the results of the various embedding models when the search space is limited to the referent category in the test data to retrieve the questions in the target language. As expected, the overall results of the models with the defined categories and limited search space are higher than the model with a larger search space.

Table 8

Comparing the performance of embedding models in cross-lingual question retrieval tasks in the whole data

Embedding Model	MRR	S@1	S@2	S@3	S@4	S@5
BiLBOWA	0.43	0.36	0.43	0.44	0.48	0.51
VecMap (Supervised)	0.45	0.37	0.46	0.61	0.63	0.61
VecMap (Unsupervised)	0.52	0.40	0.53	0.60	0.65	0.67
MUSE (Supervised)	0.54	0.42	0.54	0.64	0.65	0.66
MUSE (Unsupervised)	0.56	0.47	0.56	0.59	0.62	0.67

Table 9

Performance of embedding models in cross-lingual question retrieval task in the relevant category

Embedding Model	MRR	S@1	S@2	S@3	S@4	S@5
BiLBOWA	0.48	0.40	0.47	0.49	0.52	0.53

VecMap (Supervised)	0.50	0.41	0.51	0.58	0.62	0.68
VecMap (Unsupervised)	0.56	0.43	0.57	0.65	0.68	0.70
MUSE (Supervised)	0.60	0.47	0.63	0.65	0.68	0.72
MUSE (Unsupervised)	0.61	0.52	0.64	0.67	0.69	0.72

According to the experimental results in Table 8, BiLBOWA obtained the lowest results and the MUSE embedding in the unsupervised mode obtained the best performance in MRR, S@1, S@2, and S@5. Moreover, the MUSE embedding in the supervised mode and the VecMap embedding in the unsupervised mode performed the same for S@3 and S@4, and they performed better than the MUSE embedding in the unsupervised mode. This result indicates that post-training models obtain better results than the training models, and rotation of the vectors in the target language over the source language is highly effective. The conclusion of comparing results in Table 9 with Table 8 was similar such that BiLBOWA obtained the lowest results and the MUSE embedding in the unsupervised model obtained the best performance both in MRR and all S@*k* metrics.

Conclusion

Nowadays, various embedding methods exist to capture as much information as possible from the local context of the words mono-lingually or cross-lingually. One side of this research direction is introducing the algorithm from artificial intelligence and computer engineering point of view; the other direction is selecting appropriate parameters for the method. Deciding on selecting the appropriate embedding method for a task is controversial. It can be done by either the internal evaluation of the embedding method or an external application. The research outcome eases making this decision by comparing various methods and parameters based on internal and external evaluation techniques. In this paper, we studied the impact of the semantic distribution of the words on cross-lingual embedding models that are usable in tasks, say cross-lingual information retrieval. To this end, we compared three different embedding algorithms, BilBOWA, VecMap, and MUSE, to capture the semantic properties of the words.

Moreover, we used Word2Vec monolingual word embedding as the baseline because this model is developed based on a language to capture the semantic information for acquiring reasonable accuracy. This baseline was the upper bound, and we did not expect a better performance than this baseline. However, we aimed to build a cross-lingual embedding model that works with monolingual embeddings and has cross-lingual abilities. According to the compared experimental results, we list our findings as follows:

a) The cross-lingual, post-training methods for embedding, namely VecMap and MUSE, outperformed the other cross-lingual embedding algorithm that requires training a model from scratch, such as BilBOWA. Because in these models, the original embedding models were created based on the existing monolingual embeddings built with richer corpora. The embedding models were further processed to make their vector spaces closer together. This shows that although the input data to BilBOWA is vibrant due to using bilingual Corpus, building the model from scratch does not create a high-quality embedding model.

b) We further compared the quality of this embedding with Word2Vec, which did not benefit from knowledge from another language. The accuracy of cross-lingual embedding was comparable with monolingual embedding, with a slight improvement. Although the different result was not statistically significant, it is worth using in cross-lingual tasks, including question retrieval.

c) Although the cross-lingual embedding model performed better than other models, the difference was insignificant.

d) Among the two post-training cross-lingual embedding models, MUSE outperformed VecMap.

e) The training data of the supervised approach in MUSE could be at any size because it uses a bilingual dictionary; the unsupervised approach ought to be in a specific minimum size.

f) Increasing the window size of the local context to 25 words improved the accuracy of embedding results against the Google analogy score.

g) Rotation of the vector space model of the source and the target languages had no impact on the accuracy of the embedding model in either supervised or unsupervised approach when the monolingual similar word list is used for the evaluation. But the rotation of the target language on the source language decreased the accuracy of the embedding model dramatically when the cross-lingual word similar list was used for evaluation.

h) The post-training embedding models, such as MUSE and VecMap, performed better than the training embedding model, such as BiLBOWA.

i) The MUSE unsupervised embedding model performed the best among the compared models.

j) Limiting the search space by defining the category of the question had a positive impact on the model to retrieve the questions.

Endnotes

1. Bidirectional Encoder Representations from Transformers (BERT)
2. https://fasttext.cc/
3. http://www.ted.com/talks
4. http://parsigan.ir/datasources/targoman/2
5. https://archive.org/details/fawiki-20160720
6. https://github.com/artetxem/vecmap

References

AleAhmad, A., Amiri, H., Darrudi, E., Rahgozar, M. & Oroumchian, F. (2009). Hamshahri: A standard Persian text collection. *Knowledge-Based Systems*, 22(5), 382–387.

Ansari, E., Sadreddini, M., Tabebordbar, A. & Wallace, R. (2014). Extracting Persian–English parallel sentences from document level aligned comparable Corpus using bi-directional translation. *Advances in Computer Science: An International Journal*, 3(5), 59–65.

Artetxe, M., Labaka, G. & Agirre, E. (2018a). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*,1, (pp: 789–798), Melbourne, Australia. Association for Computational Linguistics.

Artetxe, M., Labaka, G. &Agirre, E. (2017). Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*,1, (pp: 451–462), Vancouver, Canada. Association for Computational Linguistics.

Artetxe, M., Labaka, G., Agirre, E. & Cho, K. (2018b). Unsupervised neural machine translation.

Assi, S. (1997). Farsi linguistic database (FLDB). *International Journal of Lexicography*, 10(3), 5.

Bijankhan, M. (2004). naqše peykarehāye zabāni dar neveštane dasture zabān: mo'arrefiye yek narmafzāre rāyāneyi ["The role of corpora in writing a grammar: Introducing a software"]. Journal of Linguistics, 19(2), 48–67.

Bijankhan, M., Sheykhzadegan, J., Bahrani, M. & Ghayoomi, M. (2011). Lessons from building a Persian written corpus: Peykare. *Language Resources and Evaluation*, 45(2), 143–164.

Blei, D. M., Ng, A. & Jordan, M. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022.

Camacho-Collados, J., Pilehvar, M., Collier, N. & Navigli, R. (2017). SemEval-2017 task 2: Multilingual and cross-lingual semantic word similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation* (SemEval-2017), (pp. 15–26), Vancouver, Canada. Association for Computational Linguistics.

Chandar, S., Lauly, S., Larochelle, H., Khapra, M. M., Ravindran, B., Raykar, V. & Saha, A. (2014). An auto-encoder approach to learning bilingual word representations. In *Proceedings of the 27th Annual Conference on Neural Information Processing Systems*, , (pp. 1853–1861).

Chaudhary, A., Zhou, C., Levin, L., Neubig, G., Mortensen, D. R. & Carbonell, J. G. (2018) Adapting word embeddings to new languages with morphological and phonological subword representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, (pp. 3285–3295).

Conneau, A., Lample, G., Ranzato, M., Denoyer, L. & Jégou, H. (2018). Word translation without parallel data. In *sixth International Conference on Learning Representations*.

Coulmance, J., Marty, J. M., Wenzek, G. & Benhalloum, A. (2015). Transgram, fast cross-lingual word-embeddings. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, (pp. 1109–1113).

Devlin, J., Chang, M. W., Lee, K. & Toutanova, K. (2019). BERT: Pretraining of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (pp. 4171–4186), Minneapolis, Minnesota. Association for Computational Linguistics.

Doval, Y., Camacho-Collados, J., Espinosa Anke, L. & Schockaert, S. (2020) On the robustness of unsupervised and semi-supervised cross-lingual word embedding learning, In *Proceedings of the 12th Language Resources and Evaluation Conference*, (pp. 4013–4023), Marseille, France, European Language Resources Association.

Doval, Y., Camacho-Collados, J., Espinosa-Anke, L. & Schockaert, S. (2018). Improving cross-lingual word embeddings by meeting in the middle, In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, (pp. 294–304), Brussels, Belgium. Association for Computational Linguistics.

Duong, L., Kanayama, H., Ma, T., Bird, S. & Cohn, T. (2017). Multilingual training of cross-lingual word embeddings, In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, (pp. 894–904), Valencia, Spain. Association for Computational Linguistics.

Duong, L., Kanayama, H., Ma, T., Bird, S., & Cohn, T. (2016). Learning cross-lingual word embeddings without bilingual corpora, In *Proceedings of the 2016 Conference on*

Empirical Methods in Natural Language Processing, Austin, Texas, USA, Novem

Espinosa-Anke L., Palmer G., Corcoran P., Filimonov M., Spasić I. & Knight D. (2021). English–Welsh cross-lingual embeddings, *Applied Sciences*, 11(14), 6541.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G. & Ruppin, E. (2002). Placing search in context: The concept revisited. A*CM Transactions on Information Systems*, 20, 116–131.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*. (Special Volume of the Philological Society), 1–32.

Ghayoomi, M. (2019). Finding the meaning of Persian words automatically using word embedding. *Iranian Journal of Information Processing & Management*, 35(1), 25–50.

Gouws, S. & Søgaard, A. (2015). Simple task-specific bilingual word embeddings. In *HLT-NAACL* (pp. 1386-1390).

Gouws, S., Bengio, Y. & Corrado, G. (2015). Bilbowa: Fast bilingual distributed representations without word alignments. In *International Conference on Machine Learning* (pp. 748-756). PMLR.

Gu, J., Hassan, H., Devlin, J. & Li, V. O. K. (2018a). Universal neural machine translation for extremely low resource languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (pp. 344–354).

Gu, J., Wang, Y., Chen, Y., Cho, K. & Li, V. O. K. (2018b). Meta-learning for low-resource neural machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, (pp. 3622–3631).

Guo, J., Che, W., Yarowsky, D., Wang, H. & Liu, T. (2015). Cross-lingual dependency parsing based on distributed representations. 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)* (pp. 1234-1244).

Ha, T. L., Niehues, J., Sperber, M., Pham, N. Q. & Waibel, A. (2018). KIT-Multi: A translation-oriented multilingual embedding corpus. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*, (pp. 3904–3907), Miyazaki, Japan. European Language Resources Association (ELRA).

Hadifar, A. & Momtazi, S. (2018). The impact of corpus domain on word representation: a study on persian word embeddings. *Language Resources and Evaluation*, 52(4), 997–1019.

HajiAminShirazi, S. & Momtazi, S. (2020). Cross-lingual embedding for cross-lingual question retrieval in low-resource community question answering. *Machine Translation*, *34*(4).

Harris, Z. S. (1954). Distributional structure. *Word*, 23(10), 146–162.

Harris, Z. S. (1991). *A Theory of Language and Information: A Mathematical Approach*. Oxford University Press, Oxford, England.

Hermann, K. M. & Blunsom, P. (2013). The role of syntax in vector space models of compositional semantics. In Pr*oceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, 1, (pp. 894–904), Sofia, Bulgaria.

Jabbari, F., Bakshaei, S., Ziabary, S. M. M. & Khadivi, S. (2012). Developing an open-domain English-Farsi translation system using AFEC: Amirkabir bilingual Farsi-English corpus. In *Fourth Workshop on Computational Approaches to Arabic-Script-based Languages* (pp. 17-23).

Jurafsky, D. & Martin, J. H. (2000). *Speech and Language Processing: An Introduction to*

Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall, Upper Saddle River, New Jersey.

Landauer, T. K. & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2), 211–240.

Lauly, S., Boulanger, A. & Larochelle, H. (2014). Learning multilingual word representations using a bag-of-words autoencoder. *arXiv preprint arXiv:1401.1803*.

Levy, O. & Goldberg, Y. (2014). Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 302-308).

Lin, D. (1998). Automatic retrieval and clustering of similar words. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 2* (pp. 768-774).

Luong, M. T., Pham, H. & Manning, C. D. (2015). Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st workshop on vector space modeling for natural language processing* (pp. 151-159).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Miller, G. A. & Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1), 1–28.

Ney, H., Och, F. J. & Vogel, S. (2000). Statistical translation of spoken dialogues in the Verbmobil system. In *Workshop on Multilingual Speech Communication*, (pp. 69–74), Kyoto, Japan.

Padó, S. & Lapata, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2), 161–199.

Pan, X., Gowda, T., Ji, H., May, J. & Miller, S. (2019). Cross-lingual joint entity and word embedding to improve entity linking and parallel sentence mining. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)* (pp. 56-66).

Peirsman, Y. & Geeraerts, D. (2009) Predicting strong associations on the basis of corpus data. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)* (pp. 648-656).

Pennington, J., Socher, R. & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).

Ruder, S., Vulic, I. & Søgaard, A. (2019). A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65(1), 569–631.

Sahlgren, M. (2006). *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces* (Doctoral dissertation, Institutionen för lingvistik).

Salton, G. M., Wong, A. & Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11), 613–620.

Schuster, S., Gupta, S., Shah, R. & Lewis, M. (2019). Cross-lingual transfer learning for multilingual task oriented dialog. In *Proceedings of the 2019 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, (pp. 3795–3805), Minneapolis, MN, USA.

Shaoul, C. & Westbury, C. (2010). *The Westbury Lab Wikipedia Corpus*. Retrieved from http://www.psych.ualberta.ca/~westburylab/downloads/westburylab.wikicorp.download.html.

Shi, T., Liu, Z., Liu, Y. & Sun, M. (2015). Learning cross-lingual word embeddings via matrix co-factorization. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, 2, (pp.567–572).

Song, L., Wang, Z., Mi, H. & Gildea, D. (2016). Sense embedding learning for word sense induction. arXiv preprint arXiv:1606.05409.

Vyas, Y. & Carpuat, M. (2016). Sparse bilingual word representations for cross-lingual lexical entailment. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies* (pp. 1187-1197).

Wittgenstein, L. (1953). *Philosophical Investigations*. Blackwell Publishing Ltd, Oxford, UK.

Xu, L., Ouyang, W., Ren, X., Wang, Y. & Jiang, L. (2018). Enhancing Semantic Representations of Bilingual Word Embeddings with Syntactic Dependencies. In *IJCAI* (pp. 4517-4524).

Zahedi, M. S., Bokaei, M. H., Shoeleh, F., Yadollahi, M. M., Doostmohammadi, E. & Farhoodi, M. (2018). "Persian word embedding evaluation benchmarks," In *Proceedings of IEEE Iranian Conference of Electrical Engineering*, (pp:1583-1588).