Received: June 22, 2020. Revised: July 18, 2020.

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Exploiting Comparable Corpora to Enhance Bilingual Lexicon Induction from Monolingual Corpora

Rizka Wakhidatus Sholikah[1]* **Yasuhiko Morimoto[2]** **Agus Zainal Arifin[1]**
Chastine Fatichah[1] **Ayu Purwarianti[3]**

[1]*Department of Informatics, Faculty of Intelligent Electrical and Information Technology,*
Institut Teknologi Sepuluh Nopember, Indonesia
[2]*Graduate School of Advanced Science and Engineering,*
Hiroshima University, Japan
[3]*Informatics Engineering, School of Electrical Engineering and Informatics,*
Institut Teknologi Bandung, Indonesia
* Corresponding author's Email: rizkaws@gmail.com

Abstract: Bilingual lexicons are essential resources in natural language processing (NLP) and information retrieval (IR). Automatic bilingual lexicon acquisition relies on a large number of parallel corpora that can be scarce or even unavailable for several languages. On the other hand, there are other resources that can be used to build bilingual lexicon such as comparable corpora (aligned documents) and monolingual corpora that are easily to get and available in any language, including resource-limited languages. Hence, this paper proposes a two stages framework that can learn bilingual lexicons from monolingual corpora enhanced using comparable corpora without any additional resources. The framework consists of two stages: comparable dictionary building and monolingual mapping. Comparable dictionary building is a process to create coarse dictionary from comparable corpora by utilizing topic modeling approach. The second stage is monolingual mapping by using the result from the previous stage as seed initialization for the bi-directional projection learning. The utilization of comparable corpora can replace the need of bilingual dictionary. The experiment was conducted using three kinds of language pairs: English-®Indonesia, English-®Arabic and Arabic-®Indonesia. The result of the experiment showed that the proposed method can enhance the accuracy from monolingual corpora and outperform other previous methods.

Keywords: Bilingual lexicon, Enhanced-mono, Comparable corpora, Monolingual corpora, Linear mapping, Hubness problem.

1. Introduction

Bilingual lexicon induction (BLI) is the task of extracting word pairs with the same meaning from two different languages. In the last few decades, BLI has gained a great attention in the field of natural language processing (NLP) due to its usefulness in a variety of tasks, for example in cross-lingual information retrieval [1-3], cross-lingual part-of-speech (POS) tagging [4], cross-lingual document classification [5], cross-lingual semantic text similarity [6], cross-lingual name entity recognition (NER) [7], and statistical machine translation (SMT) [8, 9]. However, building a bilingual lexicon manually is time-consuming and requires much effort. The automatic acquisition of BLI can overcome this problem. Automatic BLI is based on the assumption that words in different languages possibly translate to one another if they have the same distribution [10].

In general, methods of acquiring bilingual lexicons can be divided into two categories: those based on online learning [11, 12] and those based on offline learning [13, 14]. In online learning, the source and target languages are trained together in a shared space to produce a bilingual lexicon. This approach usually uses bilingual signals such as parallel words or parallel sentences. The bilingual signals are used to penalize words in both languages that are similar to one another to be closed in the vector space. Meanwhile, offline learning obtains translations by independently training the source language and the target language. During post-processing, the vector of the source language is mapped onto the target language with the help of bilingual signals. The use of parallel corpora such as a parallel translation or a bilingual dictionary has been very successful in obtaining bilingual lexicons. However, parallel corpora can be scarce, especially in poor language resources. Therefore, the current research focused on extracting a bilingual lexicon from non-parallel resources (e.g. comparable corpora, monolingual corpora) that can be obtained easily in the event of poor language resources.

Comparable corpora consist of document pairs for two different languages that have similar topic. The development of comparable corpora takes less time and costs less than the parallel corpora. Nevertheless, comparable corpora are able to provide resources in a large domain and can be used on a wide range of cross-lingual applications. Several methods use comparable corpora to extract bilingual lexicons [15, 16, 17]. Most of these methods are based on the assumption that words and their translations tend to occur in the same context across languages [18]. Nevertheless, they need additional resources such as a bilingual dictionary to create alignment between the vocabularies of the source language and the target language. Recent research has proposed a strategy to eliminate the need for additional resources by using hybrid learning [19]. This strategy uses linear mapping from source to target to get translation pairs. However, it does not exploit information from the other direction (target to source), which could improve the performance of bilingual lexicon extraction.

Comparable corpora are a profitable resource because it has been paired between documents based on the similarity of topics. This makes the vocabulary collection from the comparable corpora between source languages with the target language likely to have almost the same distribution. Even though in comparable corpora there is no guarantee of parallel sentence which is a direct translation from the source language to the target language. However, comparable corpora availability is less than monolingual corpora. In addition, when compared to monolingual corpora, the number of vocabularies from comparable corpora is more limited. On the other hand, the acquisition of bilingual lexicon using only the monolingual corpora will produce lower results than when using comparable corpora.

Therefore, in this paper, we propose a two stages framework to enhanced bilingual lexicon induction from monolingual corpora by utilizing the widely available of comparable corpora. This framework generates automatic seed initialization from comparable corpora based on topic modelling and linear projection to omit the need of bilingual dictionary and using the result to map between source and target languages in monolingual corpora. This proposed framework consists of the following contribution:

1. Proposes a strategy to enhanced bilingual lexicon from monolingual corpora
2. Proposes comparable dictionary from comparable corpora based on a language topic model and linear projection that can substitute the need for a bilingual dictionary during the mapping process; and
3. Shows the need for hubness mitigation to minimize the side effect of mapping and that combining several hubness mitigation algorithm can help to improve the accuracy of the proposed framework.

The remainder of this paper is organized as follows: in Section 2, we review the related work on BLI. In Section 3, we describe our proposed framework in detail. We discuss our experimental results in Section 4. In Section 5, we conclude our paper with a summary.

2. Related work

2.1 Bilingual lexicon extraction from comparable corpora

Extracting BLI from comparable corpora has gained attention in recent decades. Their huge number and availability even for poor language resources, makes comparable corpora an important resource for bilingual lexicon extraction. Most methods use a statistical approach to get bilingual lexicons from raw data. Stajner and Mladenic proposed a method to extract bilingual lexicon from comparable corpora by using the non-linear projection between source and target languages [15]. The non-linear projection is achieved by utilizing kernel mapping. Instead of performing non-linear regression, which is less effective in a large dimension of data, they sample the source language and map them using kernel mapping. Artetxe et al. [20] proposed an approach to build bilingual lexicon by generating synthetic parallel data. The synthetic parallel data is used as resource in unsupervised machine translation to generate bilingual dictionary.

Rather than directly inducing bilingual dictionary, this method uses the automatic generate dictionary as seed in BLI process. Tang et al. proposed to build the bilingual lexicon from Chinese·®Thai comparable corpora by using bilingual word correlation [16]. The correlation between bilingual words is computed by the Pearson correlation method. Then the translation probabilities are estimated by computing the natural correlation score. A proposed method by Chebel et al. utilizes a combination of context vector and concept vector to obtain a bilingual lexicon from comparable corpora [21]. The context vector can be obtained by using a word embedding method (e.g. Skip-gram, CBoW), while the concept vector can be extracted using formal concept analysis (FCA). However, the model still requires additional parallel resources. Vulic et al. proposed a method to eliminate the need for additional resources, such as a bilingual dictionary or word alignment, to get translation pairs [17]. They used a topic modelling method based on Bilingual Latent Dirichlet Allocation (BiLDA) to create groups of words in a bilingual setting. They applied various term weighting and similarity measures using Term Frequency-Inverse Topic Frequency (TF-ITF), the Kullback–Leibler method, the Cue method and combinations of them. Another approach by Vulic et al. [19] proposed hybrid method that combine the result of pseudo documents to create initial seed lexicon for mapping training.

2.2 Offline method

In general, BLI can be categorized into two variants: online methods and offline methods. Online methods have direct interaction between the source and the target languages to create a shared semantic representation. Online methods also involve joint training. The idea behind these methods is that similar words in different languages have the same semantic structure. While, offline approaches perform bilingual induction in a post-hoc setting. The bilingual lexicon is obtained from both the source language and the target language by learning their own embedding representations independently. The idea is based on the work of Mikolov et al. [22]. The bilingual lexicon can be extracted from monolingual embedding by performing a linear projection from the source language's embedding space to the target language's embedding space. The projection has the objective to minimize the distance between the source and the target dictionary. The result can capture unseen translations from the vocabularies. However, this method requires a large number of dictionary entries, usually around a thousand bilingual pairs. Another offline method has been proposed by Faruqui and Dyer [13]. The idea behind this is almost the same as for the previous method, but instead of simple linear mapping, canonical correlation analysis is used to project the source and the target embedding space into a shared space. This method also needs a large bilingual dictionary to train the projection. Orthogonal projection is another idea to map the source onto the target space [23]. The objective function used is similar to ordinary linear mapping, but instead the constraint of orthogonality of the projection is used. Orthogonality means that the projection matrix is $AA^{-1} = I$. This constraint assumes that the translation must be symmetric, and we can use the inverse of the projection from the source to the target to project the target to the source. Experimental testing showed that orthogonal mapping produces better results than ordinary linear mapping.

The mapping process can also be done by using ridge regression, which gives the L2 regularization as an objective function [19, 22, 24]. Artetxe et al. [23] proposed self-learning by repeating the learning process until certain criteria are satisfied. The dictionary is used as the initial seed for the first iteration; in the next iteration, the seed is replaced by the output from the previous iteration. This method can increase the performance of BLI compared with the previous method. In offline methods, the seed dictionary plays a crucial role in the training process. Some research attempted to generate the seed dictionary automatically to minimize the resources required for building a bilingual lexicon. One such method has been proposed by Vulic et al. [19]. They used pseudo-documents from comparable corpora to obtain the translation pairs. The translation pairs are then used as the seed lexicon to train the projection matrix. With this strategy, the need of thousands of translation pairs can be overcome automatically. Another approach came frome Karan et al. [25] that inducing classification based process into self learning that allows the integration of features in each iteration.

3. Proposed framework

The proposed framework consists of two stages, as shown in Fig. 1. First, comparable dictionary building is performed to get bilingual lexicon from comparable corpus. The second step is monolingual mapping to get bilingual lexicon from monolingual corpora using the result from previous step as input for training process.

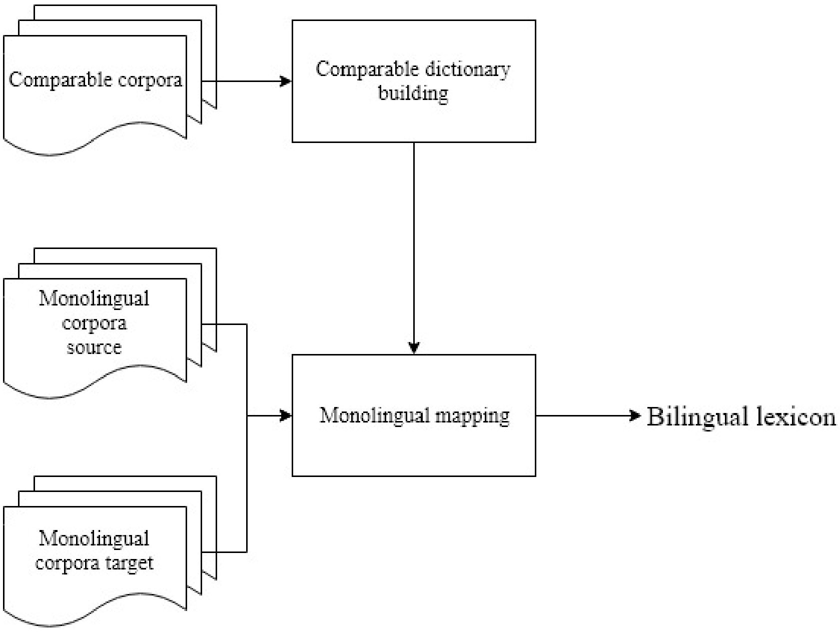
Comparable corpora
Comparable dictionary building
Monolingual corpora source
Monolingual corpora target
Monolingual mapping
Bilingual lexicon

Figure. 1 Two stages framework

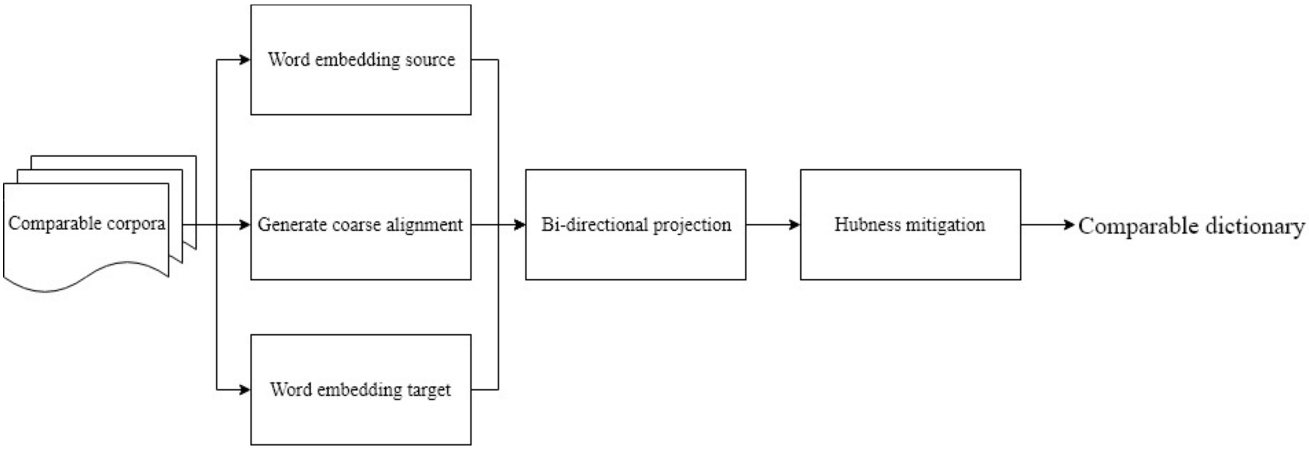

Figure. 2 Comparable dictionary building

3.1 Comparable dictionary building

The first stage is comparable dictionary building that consists of four main processes as shown in Fig. 2. First, the topic modelling method is performed using BiLDA [24, 26] to generate coarse alignment. The second step is independently converting the source documents and the target documents from comparable corpus into a dense vector representation using Skip-Gram Negative Sampling (SGNS) [22, 27]. Then, the mapping process is performed using bi-directional projection in an iterative manner. The coarse alignment is then used as the initial seed. Finally, we minimize the hubness problem by using a combination method, called re-ranking global correction. Re-ranking is done by applying CSLS in similarity matrix before using GC to retrieve the target language. Every step is explained in more detail in the next subsection.

3.1.1 Generate coarse alignment

Generate coarse alignment is done by using a bilingual topic modelling strategy. The bilingual topic model can represent the contents of bilingual documents from comparable corpora. Comparable corpora mean that a document from one language is aligned with a document in another language based on their similarity in topic or theme. Let comparable corpus C consist of a pair documents C = d_1 , d_2, ... ,d_n, where d_j = ($d_1^{L_S}$, $d_1^{L_T}$)) represents the document in language source L_S that has a link to target L_T.

BiLDA is a generalization of Latent Dirichlet Allocation (LDA) that enables us to compute a topic model of more than one language [24, 26].

Received: June 22, 2020. Revised: July 18, 2020.

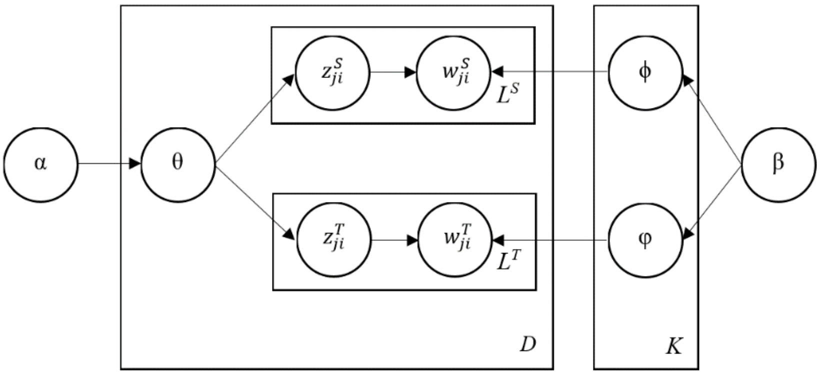

Figure. 3 BiLDA model

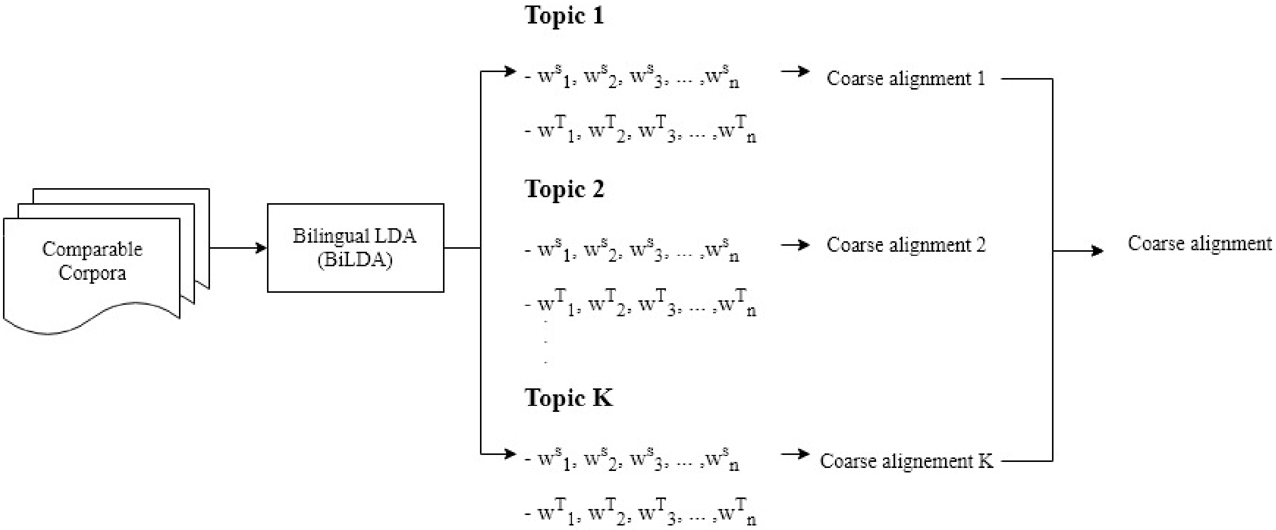
Comparable Corpora
Bilingual LDA (BiLDA)
Topic 1
- w^s_1, w^s_2, w^s_3, ... ,w^s_n
- w^T_1, w^T_2, w^T_3, ... ,w^T_n
Coarse alignment 1
Topic 2
- w^s_1, w^s_2, w^s_3, ... ,w^s_n
- w^T_1, w^T_2, w^T_3, ... ,w^T_n
Coarse alignment 2
Topic K
- w^s_1, w^s_2, w^s_3, ... ,w^s_n
- w^T_1, w^T_2, w^T_3, ... ,w^T_n
Coarse alignement K
Coarse alignment

Figure. 4 Coarse alignment generation

The structure of BiLDA can be seen in Fig. 3. This model has the common variable Θ as the topic distribution that is shared by both languages. A topic for each document is sampled from Θ with K hyper-parameter $\alpha_i = \alpha_1, \alpha_2, ..., \alpha_K$, where K denotes the number of topics. Then, the cross-lingual latent topic for each token $z_(ji)$ is sampled with respect to θ. For each language, the word in document j at position i, $w_(ji)$ is sampled with respect to word distributional φ (for the source) or ϕ (for the target) with single hyper-parameter β.

The training process of BiLDA tries to learn the optimal values of Θ, φ and ϕ so that we can detect which word is important for a particular topic and which topic is important for particular documents. In this research, we used Gibbs sampling as the monolingual LDA, with $\alpha = 50/K$ and $\beta = 0.01$. At each iteration, the assignment of the topic of the source document is updated by Eq. (1):

$$P\left(z^S_{ji} = z_k \middle| z^S_{\neg ji}, z^T_j, w^S, w^T, \alpha, \beta\right) \propto \frac{n^S_{j,k,\neg i} + n^T_{j,k} + \alpha}{n^S_{j,\neg i} + n^T_j + K\alpha} \cdot \frac{V^S_{k,w^S_{ji,\neg}} + \beta}{V^S_{k,.,\neg} + |V^S|\beta}, \tag{1}$$

where $n^S_{j,k}$ is the number of source words in the source document pair d_j assigned to topic z_k, and $n^S_{j,k,\neg i}$ is the same as $n^S_{j,k}$, except for the current word. $V^S_{k,w^S_i,\neg}$ denotes the frequency of word w^S_{ji} assigned to topic z_k, except for the current word, w^S_{ji}. $V^S_{k,.,\neg}$ represents the number of words in vocabulary source V^S that are associated with topic z_k. We can see that the first part of Eq. (1) shows the per document topic distribution, while the second part shows the per topic word distribution. For the target document, the update is computed in an analogical manner.

From this step, we obtain the group of words that best match a certain topic in both the source and target languages. The process to generate coarse alignment can be seen in Fig. 4. The source and target words for the same topic are then paired with each

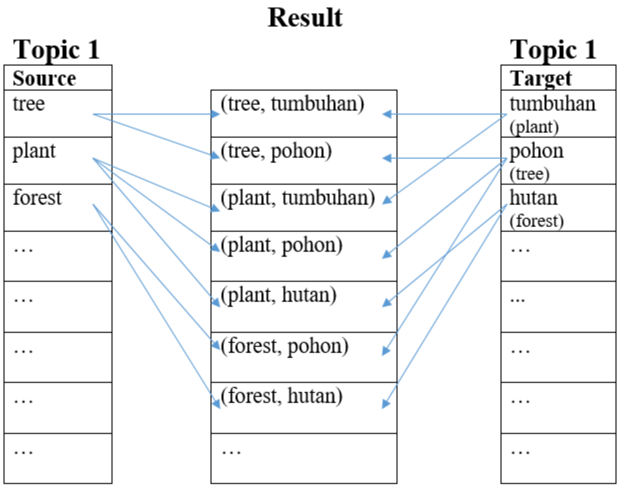

Figure. 5 Example of coarse alignment in topic 1

other to produce a coarse alignment. Because the top words for one topic are more likely to represent the semantic of the corpus, we only take the top-n for each topic to be paired. The list of top-n words contained in source languages often has a relationship with top-n words from the target language, even direct translations, although in different index sequences. Therefore, the process of pairing is based on the index of the words for a topic. Each source word is paired with the target word with the same index, the target word in the previous index and the target word in the next index. The number of words in previous and next index is controlled by the number of window size z. This paper introduces Eq. (2) to get the target words w^T from given source word w_i^s. Fig. 5. shows the example of coarse alignment that obtained from topic 1 using window size equals 1.

$$target(w_i^s) = \{w_{i-z}^T, w_{i-(z-1)}^T, \dots, w_i^T, ..., w_{i+(z-1)}^T, w_{i+z}^T | i - z \geq 0, i + z <= n\} \quad (2)$$

3.1.2 Word embedding

Word embedding can well capture the semantic relationship between words. In our method, each language is trained using Skip-Gram Negative Sampling (SGNS) [22] [27] independently. Currently, SGNS has become the state-of-the-art method to represent the distribution of words in a dense vector space. For this step, the training process can be done by any word embedding method.

In this process, word embedding is formed by using comparable corpus as input. However, documents in the source language and target language are trained independently to form vector embedding. The comparable corpora used in the generation of vector embedding at this stage is based on the assumption that, similar words from source documents and target documents will have a similar word distribution even though the number of collection documents in a comparable corpus is limited.

Let T be the number of words in a sequence of words $w = w_1, w_2, \dots, w_T$. The objective function of skip-gram is to maximize the log probability of all words within the fixed window around the centre word. Given w_x as the centre word and w_{x+i} as the words surrounding w_x with fixed window size c, the objective function can be calculated using Eq. (3):

$$L_{SG} = \frac{1}{T}\sum_{t=1}^{T} \sum_{-c \leq i \leq c, c+i \neq 0} log(p(w_{x+i}|w_x))(3)$$

The condition probability $p(w_{x+i}|w_x)$ can be calculated using the Softmax function, as shown in Eq. (4):

$$p(w_{x+i}|w_x) = \frac{\exp(v_{w_x} v_{w_{x+i}})}{\sum_{w=1}^{V} \exp(v_{w_x} v_w)}, \quad (4)$$

where $v_{w_{x+i}}$ and v_{w_x} are the context and target of vector representation, and V is the number of words in the vocabulary.

Since the above objective function is to expensive to compute, Mikolov et al. [27] speed up the training process by using negative sampling based on skip-gram model. The objective function of SGNS can be seen in Eq. (4). Goldberg and Levy [28] give clear explanation about Eq. (5):

$$L_{SGNS} = \frac{1}{T}\sum_{t=1}^{T} \sum_{-c \leq i \leq c, c+i \neq 0} log\sigma(v_{w_x}, v_{w_{x+i}}) + k\mathbb{E}_{n \sim q(n)}[log\sigma(-v_{w_x}, v_n)], \quad (5)$$

Received: June 22, 2020. Revised: July 18, 2020.

where $\sigma(x)$ is a sigmoid function.

In this research, word2vec from Gensim was used to implement SGNS on the training data. We used an embedding dimension of 100 and ignored words with a frequency smaller than 5. The rest of the setting parameters were the same as the default settings in word2vec.

3.1.3 Bi-directional projection

The standard approach to map source language L_S and target language L_T is by adopting the solution of linear equation $y = W_x$ [8]. Let $X \in R^{d_T}$ and $Y \in R^{ds}$ be the embedding vectors of the source language and the target language, respectively, where d_S and d_T denote the dimensions of source and target in the monolingual word embedding space. The seed lexicon (i.e. aligned pairs) $D_{tr} = (x_i, y_i)$ is used as the bilingual signal in the training process, where $x_i \in R^{ds}$, $y_i \in R^{d_T}$ and $x_i \in V^S$, $y_i \in V^T$. In this research, for the initialization process, we used the coarse mapping from the result of BiLDA, so that we did not need any prior knowledge, such as a bilingual dictionary or word alignment, as a bilingual signal during training.

The learning process assumes that the mapping function $W \in R^{ds \times d_T}$ is linear. The objective function of the model uses the L2 regularization of least-square error (ridge regression) as shown in Eq. (6):

$$W = arg\, min_W \Big||XW - Y||_F^2 + \lambda\Big| |W||_F, \quad (6)$$

where X and Y are training words for the source and target languages, respectively. The scalar $\lambda \leq 0$ is referred to as the regularization parameter. As the value of W is obtained, any unseen translation from the source language will be automatically mapped onto target language space R^{d_T} as W_x.

In our framework, mapping is accomplished in both directions. The source is mapped onto the target ($x \to y$) and the target onto source ($y \to x$) independently. Then, we introduce Eq. (7) to measure the combine similarities in two directions.

$$sim(x, y) = \gamma cos(Wx, y) + (1 - \gamma) cos(x, Ay), \quad (7)$$

where γ is a weight that defines the importance of each direction. For example γ = 0.5 means that both directions have the same degree of importance. By combining the similarities from both directions, we aimed to give a reward to words that appear to be close in both directions and decrease the value of words that only seem to be a translation in one direction. This idea is based on the assumption that if two words in a word pair tend to be nearest neighbours of one another, then the chance of those words being the translation is higher than in other cases. The similarity between the source projection and the target is calculated by using cosine measures.

The linear mapping is performed iteratively by following the self-learning method by Artetxe et al. [23]. In the first iteration, the training process is achieved by using the coarse mapping as the initialization. Then, in the next iteration, the result from the previous iteration is used as the seed dictionary to train the projection. The process is repeated until convergence or a certain epoch.

3.1.4 Hubness mitigation

The research conducted by Dinu et al. [29] showed that mapping elements from the source onto the target space can increase the hubness problem. The hubness problem is a side effect of mapping that occurs because there are words that appear as a nearest neighbour (NN) of several elements. Hubness can decrease the accuracy of the model because these words tend to appear in the top-1 nearest neighbours and lead to the wrong translation.

In this research, we combine the method CSLS from from Conneau et al. [30] with GC from Dinu et al. [29]. This combination is done to combine hubness mitigation using local features and global features. From our perspective, combining two different features can increase the ability to minimize hubness. Conneau et al. minimize the hubness problem by performing re-ranking into similarity matrix based on their local neighbour. The re-ranking will give a penalty to words that have high similarity with several elements. The re-ranking process can be done by following Eq. (8):

$$\begin{aligned} sim(x, y) &= 2 \times sim(x, y) \\ &- \frac{\sum_{i=1}^{K} sim(x, y_i)}{K} - \frac{\sum_{j=1}^{K} sim(x_j, y)}{K}, \end{aligned} \quad (8)$$

where K is the number of the nearest neighbours that were considered in the training process.

The second step, following Dinu et al. [29], rather than doing a query from source x to get the nearest neighbour of target y, we do it in the opposite way. We choose y that has x as their highest ranking. If there are more than one y, then we pick one y that has highest similarity among them. This method, as expressed in Eq. (9), proved to be effective in minimizing the effect of hubness.

$$GC(x, V^T) = \arg min_{y \in V^T} \left(Rank_y, p(x) - sim(x, y) \right), \quad (9)$$

where y denotes as target word and V^T is the vocabulary of the target language.

3.2 Monolingual mapping

The second stage is monolingual mapping which consists of two processes, monolingual vector embedding and bi-directional projection, as shown in Fig. 6. The input of this stage is monolingual corpora and comparable dictionary (output from previous stage), while the output are pairs of bilingual lexicon from source to target language. Monolingual vector embedding is the same process as generating embedding vector from comparable corpora in the previous stage. The difference lies in the source used to form word embedding. In this process, we use monolingual corpora from source language and target language, so there is no alignment between documents as found in comparable corpora. We used the same algorithm as the previous one (section 3.1.2), SGNS, to generate vector embedding using the same setting parameter. The reason for using the monolingual corpora at this last stage is because the monolingual corpora has more vocabulary collections and more availability compared to comparable corpora.

The next process is bi-directional mapping. This process uses a comparable dictionary that has been formed from the previous stage as a seed dictionary in the training process. While, the representation vector used by seed dictionary is a vector embedding obtained from monolingual corpora. The usage of comparable dictionary is able to produce better mapping compared to random pair as initial seed. In this process, a bi-directional projection strategy is used as in subsection 3.1.3, where mapping is done from both direction source to target and target to source. At this process, iterative training is not carried out as in stage 1. This is based on the results of experiments that show an insignificant increase and even decrease for each additional iteration. The output of this stage is bilingual lexicon which is a list of pairs of words from the source language with the target language.

4. Result and discussion

4.1 Data

In this research, we used three language pairs, English ·® Indonesian (EN ·® ID), English ·® Arabic(EN·®AR) and Arabic-Indonesian (AR·®ID). For all language pairs, we used comparable corpora and monolingual corpora, the list of data that we used is explained below:

- Comparable corpora are different from parallel corpora, which provide an exact translation from the source to the target language. Comparable corpora only have topic alignment and do not necessarily provide an exact translation. We obtained comparable corpora from Wikipedia and used a tool called WikiDocsAligner [31] to get documents alignment. In this experiment we only took 15,000 document pairs from each, language pairs.
- Monolingual corpora are used to build vector embedding as a representation of each term. In each language, we used Wikipedia dump database from July 2018 to create embedding vector. The statistics of monolingual corpora used in the experiment are shown in Table 1.
- Data test that we used in this experiment came from MUSE [30] data set that provide 1,500 records. However, MUSE data set only provide data from English to other languages. So, in this experiment, we build our own data test for Arabic·®Indonesia with 1,500 records.
- In our experiments we used bilingual dictionaries to compare the results of using bilingual dictionary and automatic dictionary. The bilingual dictionary used consists of 5,000 records. For English·®Indonesia and English·®Arabic bilingual dictionary obtained from MUSE data set [30]. Whereas for Arabic·®Indonesia, the bilingual dictionary comes from Al-Munawir's dictionary with 5,000 records taken randomly.

We performed pre-processing on the corpus, i.e. tokenization, stopword removal and part-of-speech (POS) tagging, and only took the words that have POS nouns (N), verbs (V) and adjectives (J). Then, the frequency of each, word was calculated. We only took the words that had a frequency more than 5. This process was done to remove rare words that do not give significant information. The generation of a coarse mapping was done from the comparable corpora using the BiLDA method. We used number of topics T = 300 with α = 50/K and β = 0.01. From the result of BiLDA, we took the top-5 from each group. The total number of pairs that were obtained from the coarse mapping was 3,900. Before using the pairs in the linear mapping process, we did filtering on the pairs that had no vocabulary in the source and

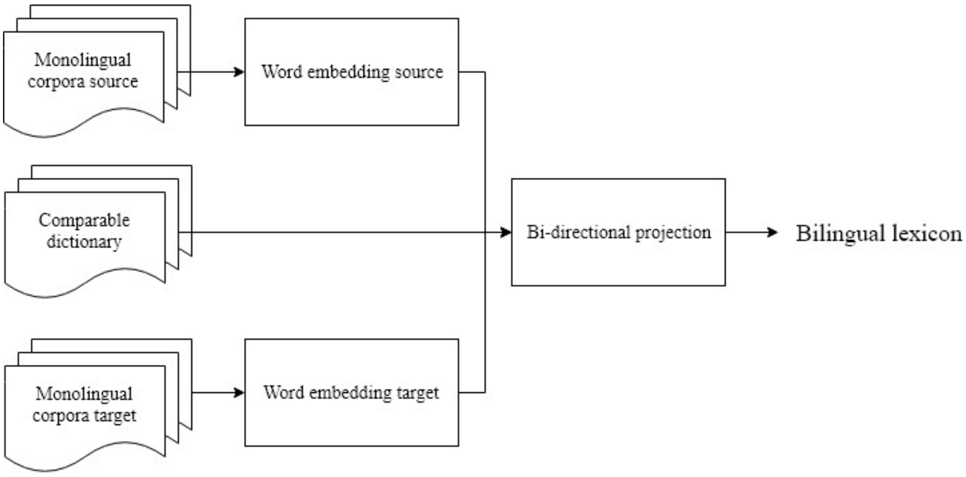

Figure. 6 Monolingual mapping

Table 1. Statistic of monolingual resources

Language	Token	Vocabulary
English (EN)	2,492M	9M
Indonesia (ID)	90M	1M
Arabic (AR)	129M	2M

the target embedding and remove words that consist less than 2 characters. The final result was 2,396 entries for EN· ®ID, 2,378 entries for EN· ®AR, and 2,369 entries for AR· ®ID.

In our method, the training process of linear mapping is done between the source and target languages with the result from the coarse mapping as the seed dictionary. Training is repeated until convergence or a certain epoch. We assumed convergence of the training process if there were no more changes in accuracy.

4.2 Matrix evaluation

We measured the performance of our bilingual lexicon framework by using accuracy on the top-t ranked translations. the experiment was done using accuracy with t equals to 1, 5 and 10. Let L be the number of test sets and $trans_i^t$ be the list of top-t ranked translations for test set i. The accuracy can then be computed with the following equation:

$$acc_t = \frac{\sum_{i=1}^{L} f(trans_i^t)}{L} \quad (10)$$

with x defined as correct translation. Then $f(trans)$ can be calculated by Eq. (11).

$$f(trans) = \begin{cases} 1, if x \in trans \\ 0, otherwise \end{cases} \quad (11)$$

We set $f(trans)$ equal to 1 when there is at least 1 translation in top-t ranked translation. Then, the results of all test set are summed. From Eq. (10), we can see that the accuracy increases as the number of t increases.

4.3 Bilingual lexicon induction

4.3.1 Parameter testing

In the first stage, to get coarse alignment, we only use top-5 with widow size 1 for each topic produced by BiLDA. The use of top-n aims to reduce noise, besides that the top-n of each topic can better represent the contents of the document collections.

In mapping process, a comparable dictionary is established on only 5,000 vocabularies. This is because if the number of vocabularies increases, the noise from the data also increases so that it can reduce the precision of the dictionary produced. However, the cut-off method like this has a disadvantage if the translation pair is not on 5,000 taken vocabularies.

In bi-directional procedure, we use γ = 0.5 based on the result of the experiment as shown in Table 2. In Table 2, for all data set, we can see that γ = 0.5 have higher accuracy only on EN· ®AR, however, the average accuracy is higher compared to others γ. Gamma equals 0.5 means that we use similarity from both directions equally. This proves that bi-directional use in the mapping process is able to

Table 2. Gamma parameter testing for bi-directional projection

Gamma	EN· ® ID	EN· ® AR	AR· ® ID	Avg
0.1	40.59	13.10	18.94	24.21
0.3	40.99	13.41	19.71	24.70
0.5	41.22	13.88	19.86	**24.99**
0.7	41.27	13.49	20.01	24.92

increase the value of accuracy by utilizing the degree of importance from both directions.

4.3.2 Hubness mitigation methods

Hubness is a side effect of mapping. This problem can decrease the performance of BLI because words that tend to be hubs appear in the top-n candidate translations. In this research, we conducted an experiment to compare a number of different strategies to minimize the effect of hubness, i.e. ordinary KNN, GC, CSLS with neighbours 10, CSLS with neighbours 15, and combining CSLS with GC. Ordinary KNN is the same as the mapping result without hubness mitigation. Meanwhile, GC reverses the process by choosing a word in target language that has query word as their top-n neighbor. While CSLS method uses Eq. (8) to modify the similarity matrix. The modification has the purpose to decrease the similarity score if a word is detected as a hub. CSLS-GC is a method that combines between CSLS and GC with a certain predefined number of neighbours.

In this experiment, we compared several hubness methods by using 2 different strategies, mono and enhanced-mono. Mono means that we apply the hubness method in mapping process by using only monolingual data. Meanwhile, enhanced-mono using our strategy that used the mapping result of comparable corpus as an initial seed of monolingual mapping.

Table 3 shows the results for accuracy at 1 from each method in 2 different strategy. We use three data set EN· ®ID, EN· ®AR, and AR· ®ID. Each method was applied to both strategy with the same seed initialization (in our strategy, this seed initialization is used at the first stage), mapping method and parameter setting. The result shows that almost in every data the hubness methods improved the performance compared to ordinary KNN with the highest average accuracy 23.47%. Among several hubness methods, CSLS-GC, a combination of methods that use global features and local features produces the highest average accuracy value. However, the increase in accuracy of this combined method is not significant. Table 3 also shows that the proposed method is able to get better accuracy than just using monolingual data. This is indicated by the accuracy of all data set that produce better results in the enhanced-mono strategy in each scenario. For EN· ®ID, EN· ®AR, and AR· ®ID the average accuracy is increased to 1.24%, 0.56%, and 0.95% respectively.

4.3.3 Comparison with other methods

The experiment was conducted by retrieving 100,000 vocabularies from the target language with queries from the source language. The list of queries comes from data test that has been mentioned in previous sub-section. There are two kinds of scenarios in this experiment, comparing different seed initialization and comparing with previous methods. The first scenario, we compare bilingual dictionary as seed initialization with coarse alignment. We use three strategies, monolingual, comparable and enhanced-mono (two stages strategy). Table 4 shows that the average accuracy of the bilingual dictionary outperforms coarse alignment for each data set and strategy. Although the average accuracy of coarse alignment is lower, but the difference with the bilingual dictionary is still acceptable. Coarse alignment can be used as an alternative of seed initialization. This is considering that coarse alignment does not use word alignment or dictionary, which is a resource that not all languages have.

In Table 4, we can see the performance comparison of the three types of strategies. From all data set, the average accuracy of comparable corpora is higher than monolingual. However, comparable corpora have less availability than monolingual corpora, besides the vocabulary covered by comparable corpora is also limited. The high accuracy of comparable corpora combined with the wide coverage of monolingual vocabulary can enhance the result of lexicon induction compared to monolingual in coarse alignment. Nevertheless, enhanced-mono using a bilingual dictionary has lower average accuracy among other strategies. This might be caused by the seed dictionary that already has good quality, so the use of the result in first stage as seed initialization in second stage can be decreased the performance. As we know that hand-crafted bilingual dictionary has higher precision combine with automatic bilingual lexicon.

The second scenario is combining our strategy with previous methods, i.e. linear mapping by Mikolov et al. [22], orthogonal projection by Artetxe et al. [23], and ridge regression with GC [29]. For all methods, we divide the experiment based on the seed initialization, the first one using bilingual dictionary and the second one using coarse alignment. All the methods using iterative learning during mapping process, except Mikolov methods, because the performance of Mikolov decreased as the number of iterations increased. The experiment was done for EN· ®ID, EN· ®AR, and AR· ®ID. Table 5 shows the result of the experiment. The average accuracy in

Table 3. Accuracy for different hubness mitigation methods

Hubness method	EN· ®ID		EN· ®AR		AR· ®ID	
	Mono	Enhance-mono	Mono	Enhance-mono	Mono	Enhanced-mono
KNN	31.96	34.53	10.52	10.89	20.30	**21.41**
GC	34.35	**35.48**	**14.83**	15.25	19.15	21.22
CSLS (10)	34.22	35.28	14.47	15.25	**20.54**	20.88
CSLS (15)	34.17	35.24	14.54	15.13	20.54	20.97
CSLS (10) - GC	**34.70**	35.28	14.80	**15.32**	19.64	21.07

Table 4. Accuracy for different hubness mitigation methods

Seed initialization	EN· ®ID	EN· ®AR	AR· ®ID	Avg
Bilingual dictionary 5,000 (monolingual)	36.39	**17.47**	19.78	24.55
Bilingual dictionary 5,000 (comparable)	40.81	13.92	20.34	**25.02**
Bilingual dictionary 5,000 (enhanced-mono)	35.55	15.76	20.69	24.00
Coarse alignment (monolingual)	34.70	14.80	19.63	23.04
Coarse alignment (comparable)	**41.26**	13.34	19.86	24.28
Coarse alignment (enhanced-mono)	35.28	15.32	**21.17**	23.92

Table 5. Accuracy with different methods and seed initialization

Method	EN· ®ID	EN· ®AR	AR· ®ID	Avg
Mikolov [22] (5,000 bilingual dictionary)	**35.72**	**17.40**	10.36	21.26
Artetxe [23] (5,000 bilingual dictionary)	33.86	15.09	19.20	22.72
Dinu [29] (5,000 bilingual dictionary)	34.88	15.80	18.38	23.02
Ours (5,000 bilingual dictionary)	35.55	15.76	**20.88**	**24.06**
Mikolov [22] (coarse alignment)	24.07	4.02	3.07	10.39
Artetxe [23] (coarse alignment)	33.68	14.57	20.59	22.95
Dinu [29] (coarse alignment)	34.79	14.76	19.68	23.08
Ours (coarse alignment)	35.28	15.32	21.21	**23.94**

all data set using a bilingual dictionary, and coarse alignment shows that our proposed strategy outperformed the other methods. The average accuracy was 24.06% for bilingual dictionary and 23.94% for coarse alignment. We can see that our two stages strategy by utilizing wide availability resource, comparable corpora, can increase the performance, especially while using coarse alignment. From the result, we also can conclude that hubness method increases the performance in iterative learning if the seed initialization does not have good quality. It can be seen in Mikolov that does not apply hubness method has poor performance while using coarse alignment.

In Table 6 we give an example of our result in retrieving translation from the top-5 candidate translation. We used EN·®ID with three different queries: hotels, ambassadors and certificate. These translations were retrieved from 100.000 vocabularies in target language. We get the translation based on their similarity with the query words. In the pre-processing process, for English language we do not apply lematization so that there are plural words in vocabulary collection. Meanwhile, in Indonesian, plural words are written using double singular, for example, plurals from the word ’*hotel*’ (hotel) are ’*hotel-hotel*’ (hotels), ’*mobil*’ (car) are ’*mobil-mobil*’ (cars). So the query ’hotels’ should be returned the translation ’*hotel-hotel*’ (hotels). However, because in this study we only use single words as a vocabulary collection, for ’hotels’ query, which are translated as ’hotel’ are considered correct.

In this experiment, we categorized errors into two groups: ‘not completely wrong’ and ‘completely wrong’. ‘Not completely wrong’ is an error that returns the wrong translation but still has a relation with the query word. The relation can be antonym, association, hypernym or hyponym, etc. Another error is when for several words in the test data only have one translation, whereas those words actually have more than one translation. So even though the translation is actually correct, it can still be counted

Table 6. Example of result from proposed method

EN-ID		
hotels	**ambassadors**	**certificate**
hotel (hotel)	*duta* (ambassador)	*sertifikat* (certificate)
perbelanjaan (shopping)	*delegasi* (delegation)	*ijazah* (certificate)
pertokoan (shopping complex)	*perwakilan* (representative)	*ijasah* (certificate)
hipermarket (hypermarket)	*diundang* (invited)	*diploma* (diploma)
ritel (retail)	*konferensi* (conference)	*ijin* (permission)

as an error. ‘Completely wrong’ means an error where there is no relation at all with the query word.

5. Conclusions

In this work, we proposed a two stage strategy to learn projection from monolingual corpora enhanced with comparable corpora. It exploits the result from the topic model in the comparable corpora to create a coarse mapping. Then, the coarse mapping is used as initial seed projection to replace the need of the bilingual dictionary. In stage one, we used embedding vector and vocabulary from comparable corpora. The result of the first stage is used as initial seed in stage two. In this stage, the embedding vector comes from monolingual data that has a wide range of coverage. The experiment showed that our proposed strategy is competitive with previous methods. This can be seen from the results of our proposed strategy, that can reach accuracy with 35.28%, 15.32%, and 21.21% for EN·®ID, EN·®AR, and AR·®ID, respectively. Although the result are slightly lower compared with the use of dictionaries, but the difference can still be considered. This is because the proposed strategy does not use dictionaries or word alignments in the development of bilingual lexicon.

Conflicts of Interest

The authors have no conflict of interest to declare.

Author Contributions

R. W. Sholikah conceived of the presented idea. R. W. Sholikah and Y. Morimoto developed the method. C. Fatichah and A. Purwarianti verified the experiment and analysis. A. Z. Arifin and Y. Morimoto supervised the finding of this work. R. W. Sholikah wrote the manuscript with support from Y. Morimoto, A. Z. Arifin, C. Fatichah and A. Purwarianti.

Acknowledgments

This work was supported by the Ministry of Research, Technology and Higher Education of Republic Indonesia (No. 135 /SP2H/LT/DRPM/IV/ 2017).

References

[1] I. Vulic and S. Moens, “Monolingual and cross lingual information retrieval models based on (bilingual) word embeddings”, In: *Proc. of SIGIR '15 the 38th International ACM SIGIR Conf. on Research and Development in Information Retrieval*, 2015.

[2] R. Rahimi, A. Shakery, and I. King, “Extracting translations from comparable corpora for cross-language information retrieval using the language modeling framework”, *Information Processing & Management*, Vol. 52, No. 2, pp. 199-318, 2016.

[3] J. Dadashkarimia, A. Shakeryab, H. Failiab, and H. Zamani, “An expectation-maximization algorithm for query translation based on pseudo-relevant documents”, *Information Processing & Management*, Vol. 53, No. 2, pp. 371-387, 2017.

[4] Y. Zang, D. Gaddy, R. Barzilay, and T. Jaakkola, “Ten pairs to tag - multilingual pos tagging via coarse mapping between embeddings”, In: *Proc. of 15th Annual Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016.

[5] S. Gouws, Y. Bengio, and G. Corrado, “Bilbowa: Fast bilingual distributed representations without word alignments”, In: *Proc. of International Conf. on Machine learning*, 2015.

[6] G. Glavaš, M. F. Salvador, S. P. Ponzetto, and P. Rosso, “A resource-light method for cross-lingual semantic textual similarity”, *Knowledge-based systems*, Vol. 143, pp. 1-9, 2018.

[7] D. Wang, N. Peng and K. Duh, “A multitask learning approach to adapting bilingual word embeddings for cross-lingual named entity recognition”, In: *Proc. of the 8th International Joint Conf. on Natural Language Processing*, 2017.

[8] T. Mikolov, Q. V. Le, and I. Sutskever, “Exploiting similarities among languages for machine translation”, In: *Proc. of the International Conf. on Learning Representations (ICLR 2013)*, 2013.

[9] J. Gao, X. He, W. T. Yih, and L. Deng, “Learning continuous phrase representations for translation modeling”, In: *Proc. of the 52nd Annual Meeting of the association for computational linguistics*, 2014.

[10] R. Rapp, “Identifying word translations in non-parallel texts”, In: *Proc. of the 33th Annual Meeting on Association for Computational Linguistics*, 1995.

[11] A. P. S. Chandar, S. Lauly, H. Larochelle, M. M. Khapra, B. Ravindran, V. Raykar, and A. Saha, “An autoencoder approach to learning bilingual word representations”, In: *Proc. of the 27th International Conf. on Neural Information Processing Systems*, 2014.

[12] K. M. Hermann and P. Blunsom, “Multilingual distributed representations without word alignment”, In: *Proc. of the 2014 International Conf. on Learning Representations,* 2014.

[13] M. Faruqui and C. Dyer, “Improving vector space word representations using multilingual correlation”, In: *Proc. of 14th Conf. of the European Chapter of the Association for Computational Linguistics*, 2014.

[14] S. L. Smith, D. H. Turban, S. Hamblin, and N. Y. Hammerla, “Offline bilingual word vector, orthogonal transformation and inverted softmax”, In: *Proc. of ICLR,* 2017.

[15] T. Stajner and D. Mledenic, “Cross-lingual documents similarity estimation and dictionary generation with comparable corpora”, *Knowledge and Information Systems,* pp. 1-15, 2018.

[16] P. Tang, J. Zhao, Z. Yu, and Y. Xian, “A method of chinese and thai cross-lingual query expansion based on comparable corpus”, *Journal od Information Processing Systems,* Vol. 13, pp. 805-817, 2017.

[17] Vulic, W. D. Smet, and M. Moens, “Identifying word translations from comparable corpora using latent topic models”, 2011.

[18] P. Fung, “A statistical view on bilingual lexicon extraction: from parallel corpora to non-parallel corpora”, In: *Proc. of the 3th Conf. of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup,* 1998.

[19] Vulic and A. Korhonen, “On the role of seed lexicons in learning bilingual word embeddings”, in *ACL 2016*, 2016.

[20] M. Artetxe, G. Labaka, and E. Agirre, “Bilingual lexicon induction through unsupervised machine translation”, In: *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics 2019,* 2019.

[21] M. Chebel, C. Latiri, and E. Gaussier, “Bilingual lexicon extraction from comparable corpora based on close concept mining”, In: *Proc. of Pacific-Asia Conf. on Knowledge Discovery and Data Mining*, 2017.

[22] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space”, In: *Proc. of the International Conf. on Learning Representations (ICLR 2013),* 2013.

[23] M. Artetxe, G. Labaka, and E. Agirre, “Learning bilingual word embeddings with (almost) no bilingual data”, In: *Proc. of the 55th Annual Meeting of the Association for Computational Linguistics,* 2017.

[24] Vulic, D. Smet, J. Tang and M. F. Moens, “Probabilistic topic modeling in multilingual settings: an overview of its methodology and applications”, *Information Processing and Management,* Vol. 51, No. 1, pp. 111-147, 2015.

[25] M. Karan, I. Vulic, A. Korhonen, and G. Glavas, “Classification-Based Self-Learning for Weakly Supervised Bilingual Lexicon Induction”, In: *Proc. of the 58th Annual Meeting of the Association for Computational Linguistics 2020,* 2020.

[26] D. Mimno, H. M. Wallach, J. Naradowsky, D. A. Smith, and A. McCallum, “Polylingual topic model”, In: *Proc. of the 2009 Conf. on Empirical Methods in Natural Language Processing,* 2009.

[27] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality”, in *Advanced in Neural Information Processing Systems*, 2013.

[28] Y. Goldberg and O. Levy, “Word2vec explained: deriving Mikolov et al.'s negative sampling word embedding method”, *ArXiv:1402.3722 [cs, stat]*, 2014.

[29] G. Dinu, A. Lazaridou and M. Baroni, “Improving zero-shot learning by mitigating the hubness problem”, in *ICLR Workshop Papers*, 2015.

[30] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jegou, “Word translation without parallel data”, in *ICLR*, 2018.

[31] M. Saad and B. Alijla, “WikiDocsAligner: an off-the-shelf wikipedia documents alignment tool”, In: *Proc. of Palestinian International Conf. on Information and Communication Technology,* 2017.

[32] E. Morin and A. Hazem, “Exploiting unbalanced specialized comparable corpora for bilingual lexicon extraction”, *Natural Language Engineering*, Vol. 22, pp. 575-601, 2016.