English-Malay Cross-Lingual Embedding Alignment using Bilingual Lexicon Augmentation

Lim Ying Hao, Jasy Liew Suet Yan
School of Computer Sciences, Universiti Sains Malaysia
11800 Penang, Malaysia
yinghaoly@student.usm.my, jasyliew@usm.my

Abstract

As high-quality Malay language resources are still a scarcity, cross lingual word embeddings make it possible for richer English resources to be leveraged for downstream Malay text classification tasks. This paper focuses on creating an English-Malay cross-lingual word embeddings using embedding alignment by exploiting existing language resources. We augmented the training bilingual lexicons using machine translation with the goal to improve the alignment precision of our cross-lingual word embeddings. We investigated the quality of the current state-of-the-art English-Malay bilingual lexicon and worked on improving its quality using Google Translate. We also examined the effect of Malay word coverage on the quality of cross-lingual word embeddings. Experimental results with a precision up till 28.17% show that the alignment precision of the cross-lingual word embeddings would inevitably degrade after 1-NN but a better seed lexicon and cleaner nearest neighbours can reduce the number of word pairs required to achieve satisfactory performance. As the English and Malay monolingual embeddings are pre-trained on informal language corpora, our proposed English-Malay embeddings alignment approach is also able to map non-standard Malay translations in the English nearest neighbours.

1 Introduction

Distributional semantic models produce word embeddings that allow us to compare the relationship between words. In (static) word embeddings, each word is associated to a continuous real-valued vector such that words that are semantically similar to each other will be in close proximity when we visualize them. Word embeddings that have been adopted extensively include but are not limited to CBOW (Mikolov, Chen, et al., 2013), Skip-gram (Mikolov, Chen, et al., 2013) and GloVe (Pennington et al., 2014).

Word embeddings that are pre-trained monolingually are limited to tasks solely in its own language. For this reason, we are unable to compare the meaning of words between languages or transfer models trained on one language to another language (Ruder et al., 2019). Cross-lingual word embeddings could overcome the language constraint and make it possible for the more abundant English resources to be leveraged for emotion or other text classification in Malay (i.e., the resource poor language of interest). In cross-lingual word embeddings, words that are semantically similar regardless of the languages, will appear to be close to each other in the vector space.

The current state-of-the-art multilingual language models like mBERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020) for cross-lingual transfer are computationally expensive. Cross-lingual word embeddings offers an alternative that is cost-effective for cross-lingual transfer requiring a model to be trained only using the source language, which can then subsequently be applied to perform zero-shot or few-shot learning (Ghasemi et al., 2020) on the target language.

In this study, we attempt to align English and Malay monolingual static word embeddings pre-trained on informal text (i.e., tweets or Instagram posts) using the transformation method proposed by Smith et al. (2017). Corpora used to pre-train the embeddings are neither parallel nor aligned, and the only bilingual signal comes from the training bilingual lexicon. Instead of constructing the

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Student Research Workshop, pages 229 - 238

training bilingual lexicon from parallel corpora (Dinu et al., 2015; Smith et al., 2017), we exploit and extend an existing English-Malay bilingual lexicon. The drawback of using the set of word pairs in this bilingual lexicon is that there are numerous invalid word pairs that need to be filtered out. However, using the bilingual lexicon, we are able to generalize the mapping to word embeddings trained on corpora of other domains. We also created a new bilingual lexicon that was shown to be better than the baseline seed lexicon in alignment precision.

To the best of our knowledge, there is no gold-standard bilingual lexicon available for our evaluation task. We select a portion of the bilingual lexicon available on Malaya Documentation (Husein, 2018) and manually extracted the Malay translations for the English-side seed words from *Dewan Bahasa dan Pustaka Malaysia*[1] (DBP). Malaya Documentation offers a current state-of-the-art Malay-English lexicon, which is currently not validated.

The contributions of this study are three-fold.:

a) We created a better English-Malay bilingual lexicon in terms of alignment precision than that the one from Malaya Documentation that has been widely used.
b) We created a gold-standard bilingual lexicon containing approximately 1,200 word pairs to be used as the seed dictionary to induce a better embedding mapping or as a test set to evaluate the quality of cross-lingual word embeddings for future research.
c) We aligned monolingual word embeddings trained independently on informal corpora to create the first English-Malay cross-lingual word embeddings in the social media domain and evaluated its quality using bilingual lexicon induction.

2 Related Work

Word embeddings alignment is one of the approaches used to bridge the language gap between the source (resource rich) and target (resource poor) languages. Prior studies can be categorized into those that require and do not require a set of bilingual seed lexicons.

Mikolov et al. (2013) were one of the earliest and influential studies using bilingual seed lexicons for word embeddings alignment. A transformation matrix was learnt from the seed word pairs to linearly map the source word embeddings to the target embeddings space. Dinu et al. (2015) enhanced this approach by introducing an L2-regularization least-squares error in the objective function.

Xing et al. (2015) improved the method proposed by Mikolov et al. (2013) by restricting the word vectors to a unit length and constraining the transformation matrix to be orthogonal. They also redefined the objective function by using cosine similarity between the transformed source and target embeddings. These additional steps solved the inconsistency uncovered in Mikolov et al. (2013) and achieved better performance. On top of these constraints, Artetxe et al. (2016) enforced dimension-wise mean centering on the word embeddings so that randomly chosen words would not be semantically similar. Their study also discovered that the improvement attained by Xing et al. (2015) was solely from the orthogonal constraint instead of solving the inconsistency problem. Smith et al. (2017) proved that an orthogonal transformation matrix must also be self-consistent.

Faruqui and Dyer (2014) used Canonical Correlation Analysis (CCA) to learn the transformation matrices from the seed lexicon. Unlike Mikolov et al. (2013), a transformation matric was learnt for both source language and target language, respectively. The source and target word embeddings were then mapped to a new shared vector space where the seed word pairs from a lexicon would be maximally correlated.

Lu et al. (2015) adopted a non-linear extension of CCA to train the transformation matrices. Two neural networks were trained to obtain the transformation matrices by maximizing the correlation of the transformed source and target word embeddings in the new vector space.

Barone (2016) aligned word embeddings by eliminating the need for seed lexicons. They adopted an adversarial autoencoder in mapping the source embeddings to target embeddings. The source embeddings were transformed using an encoder, and the discriminator then tried to match

[1] A government body that coordinates the use of the Malay language in Malaysia.

the latent representations to the distribution of the target embeddings.

Zhang et al. (2017) matched the distribution of the transformed source embeddings to target embeddings using adversarial training. They learnt an orthogonal transformation matrix using the generator, and the discriminator would then try to distinguish the transformed source embeddings from target embeddings. Additionally, they also attempted to relax the orthogonality constraint by using an adversarial autoencoder.

Artetxe et al. (2018) induced the initial seed word pairs by exploiting the similarity distribution between words using an unsupervised method. These initial seed word pairs were then refined through iterative self-learning. They also enforced the transformation matrices used to map the source and target embeddings to the new vector space to be orthogonal.

Feng and Wan (2019) proposed an approach to nonlinearly map the source word embeddings and target word embeddings to a new vector space. They induced the seed word pairs using nearest neighbour retrieval, and the seed word pairs were then refined iteratively using self-learning. Instead of orthogonality, they adopted the Euclidean Norm to guide the learning of the transformation matrices.

Recent studies proposed to align contextual embeddings. Schuster et al. (2019) generalized the approach by Mikolov et al. (2013) and Conneau et. al (2018) to align embeddings from mBERT. Since one word can have different embeddings based on the context, Schuster et al. represented each word using the embedding anchor that was obtained by averaging a subset of its contextual embeddings.

Aldarmaki and Diab (2019) adopted the approach by Mikolov et al. (2013) to map contextual embeddings from the ELMo (Peters et al., 2018). Instead of using the embedding anchor, they constructed a dynamic bilingual lexicon from a parallel corpus with word alignment. Additionally, they also proposed sentence-level mapping in which the transformation matrix was learnt on aligned sentences.

Wang et al. (2020) also extended the method by Mikolov et al. (2013) to contextual embeddings. Similar to Aldarmaki and Diab (2019), they formed the alignment matrix based on aligned word pairs extracted from parallel corpora. The representations extracted from mBERT were then aligned by multiplying with the alignment matrix.

Existing studies have not explored static or contextual word embeddings alignment between English and Malay languages. As word embeddings alignment has shown promising performance in cross-lingual transfer tasks in other language pairs, it provides strong motivation for us to explore how word embeddings alignment can also benefit the English-Malay language pair, and subsequently any future study that may require English-Malay cross-lingual word embeddings.

3 Data Sources

The method requires a monolingual English embedding, a monolingual Malay embedding and a bilingual English-Malay lexicon to map the two monolingual embeddings into a bilingual Malay-English embedding. The quality of the bilingual English-Malay lexicon plays an important role because it serves as the only bridge to map two separate English and Malay monolingual lexicons into a single shared space. Malay is written in the Latin alphabet and shares lexical similarities with Indonesian as they are from the same language family.

3.1 Word Embeddings

Our study used the word2vec **English monolingual word embedding (EWE)** pre-trained on tweets by Godin (2019) using the Skip-gram architecture and contained approximately 3 million words. The words were represented by 400-dimensional vectors.

Word2vec **Malay monolingual word embedding (MYWE)** were pre-trained on tweets and Instagram posts by Husein (2018) using the Skip-gram architecture and contained approximately 1.3 million words. Normalization and spell-check were performed to standardize non-standard Malay words in the embeddings.

We trimmed the vocabulary of the embeddings to the top 200,000 most frequent words from the subset of the training corpora used to train the word embeddings (**top200k-MYWE**). This naïve filter was an attempt to remove non-Malay words from the vocabulary. Additionally, we trimmed the original embeddings separately to the top 800,000 most frequent words from the same corpora and compared them against the words extracted from selected corpora by DBP written in standard Malay to remove non-standard Malay words from the vocabulary (**top800k-MYWE**).

3.2 Bilingual Lexicon

An **English-Malay bilingual lexicon** was obtained from Malaya Documentation (Husein, 2018). Invalid words, non-English words and non-Malay words were filtered out. We used English spell-check in Microsoft Excel to filter English words and *Dewan Eja Pro*[2] to filter Malay words. We randomly selected 90% of these lexicon word pairs for mapping in the training phase (**T-BL**) and regarded it as our baseline, while the remaining 10% were used to create a set of gold standard test English-Malay word pairs. For every word pair, we retained its English side, for which we manually extracted its corresponding Malay translations from the English-Malay dictionary by DBP[1] to create a gold standard bilingual lexicon (**G-BL**). G-BL contains 1273 entries of which one English word can have one or many Malay translations from G-BL. This bilingual lexicon consists of 3675 unique Malay words.

4 Methodology

4.1 Cross-lingual Word Embeddings

To create cross-lingual word embeddings, we mapped the English embeddings, $\boldsymbol{E}$ to the Malay embeddings space using the orthogonal transformations approach proposed by Smith et al. (2017). Malay embeddings were first made to have the same dimensions as English embeddings by post-padding with arrays of zeros. We also normalized both embeddings to a unit length.

From the bilingual lexicons (T-BL) containing n word pairs, two ordered matrices $S_D \in \mathbb{R}^{n\times 400}$ and $T_D \in \mathbb{R}^{n\times 400}$ were formed where i^{th} row of the matrices corresponded to the English and Malay word vectors of the i^{th}word pairs. We then performed Singular Value Decomposition (SVD) operation on the matrix product $P = {S_D}^T T_D \in \mathbb{R}^{400\times 400}$ and subsequently, P was represented by $U\sum V^T$. The English embeddings, $\boldsymbol{E}$, were then aligned to the Malay embeddings space by multiplying it with the transformation matrix $\boldsymbol{O} = UV^T$ that was subject to the orthogonal constraint:

$$\max_{O} \sum_{i=1}^{n} {t_i}^T \boldsymbol{O} s_i \text{, subject to } \boldsymbol{O}^T\boldsymbol{O} = \boldsymbol{I} \quad (1)$$

4.2 Experiment Extensions

We explored three different directions to extend the initial embeddings mapping using T-BL. The first direction examined how the coverage of the Malay words in the training lexicon could affect the translation accuracy. The second direction investigated if the quality of T-BL is satisfactory, and the third direction aimed to improve the quality of the training bilingual lexicon used for mapping.

Direction 1: We hypothesized that a higher coverage of the Malay words in the training lexicon would improve the translation accuracy of English words. To investigate this hypothesis, we augmented T-BL by using the English-side words from the lexicon as the seed words. A different number (1, 5, 10) of nearest neighbours (NN) of the seed English words was selected using the dot product from their respective embeddings space. This is equivalent to using cosine similarity after normalizing the embeddings to a unit length as shown below:

$$cos(\theta) = \frac{\boldsymbol{V}_i \cdot \boldsymbol{V}_j}{\| \boldsymbol{V}_i \| \| \boldsymbol{V}_j \|} \; for \; i \neq j \quad (2)$$

where $\boldsymbol{V}_i$ is the vector representation of the i^{th} seed word and $\boldsymbol{V}_j$ is the vector representation of other English words in the embeddings space.

Selected nearest neighbours were then translated into Malay language using either Google Cloud Translation API or Google Translate function in Google Sheets. This comparison is to help us determine which tool returns better translation as we notice they could return different translations for the same English word. Translated Malay words that are longer than one token were omitted as words in the vocabulary were restricted to be one-token long. Furthermore, the English nearest neighbours that happened to be in G-BL were also discarded to prevent possible data leakage.

Direction 2: We extracted the English-side words of T-BL and translated them into Malay language using Google Cloud Translation API to form a completely new set of seed lexicon (**N-BL**). This resulted in a completely new set of training word pairs having the same size as T-BL to allow direct comparison of the quality of the word pairs.

Direction 3: We observed that the English nearest neighbours could contain noise as the

[2] A Malay proofing tool produced by Dewan Bahasa dan Pustaka.

vocabulary of the English embeddings was not trimmed. Therefore, to remove this noise, we sent selected English nearest neighbours through a word filter to ensure that the nearest neighbours only comprised English words. Two different and independent filters were applied, resulting in two different sets of augmented training lexicons.

The **first filter** was built using words from WordNet. WordNet resembles a thesaurus in which words were grouped into synonymous sets (synsets) based on their concept and these words are known as lemmas. This filter gathered lemmas extracted from every synset into a list and omitted nearest neighbour words that did not match the words in the list.

The **second filter** was built using words from the Words Corpus. Words Corpus is a list of dictionary words attainable from /usr/share/dict/words file in Unix that some spell checkers use. It is a built-in corpus in the Natural Language Toolkit (NLTK) (Bird et al., 2009). Similarly, this filter gathered all words in this corpus into a list and omitted nearest neighbour words that did not have a match in the list.

4.3 Evaluation Metric

We used bilingual lexicon induction to evaluate the quality of our embeddings mapping by finding the top-*N* most semantically similar Malay words to the English words in the test set (G-BL) using cosine similarity from the shared vector space (P@N), where *N* is 1, 5 or 10. To avoid confusion from the translations obtained from Google Translate or bilingual English-Malay dictionary, we use "induced translation" to specifically indicate Malay words from the Malay embeddings that are mapped to the corresponding English words from the English embeddings. P@N measures the proportion of English words in G-BL which have at least one true Malay translation among the *N* Malay induced translations. Formally, P@N can be computed using the following equation:

$$P@N = \frac{\sum_{i=1}^{M} I_i}{M} \quad (3)$$

where M is the number of English words in G-BL. I_i is an indicator function that will take 1 if and only if i^{th} English word in G-BL has at least one correct Malay translation appearing in its corresponding N Malay induced translations, and take 0 otherwise. Therefore, the numerator indicates the number of English words that have at least one correct Malay induced translation. An induced translation will not be counted as correct if it does not appear in G-BL.

As our word embeddings were not trained on English-Malay parallel or aligned corpora, the investigation of a different number of Malay nearest neighbours is necessary to determine the extent of correct translations from the English words,. Furthermore, the word embeddings were pre-trained on tweets or Instagram posts known to mostly contain informal words.

5 Results and Discussion

5.1 Comparing Malay Embeddings Coverage

While we fixed the number of word pairs in G-BL, top200k-MYWE and top800k-MYWE have a smaller vocabulary size, and hence different number of effective word pairs for evaluation as reflected in the denominator in the P@10 column of Table 1. We only adopted P@10 in this experiment to justify the choice of the embeddings for subsequent experiments and not to compare the quality of the bilingual lexicon.

Embeddings	P@10
MYWE	22.2041 (274/1234)
top200k-MYWE	25.4036 (299/1177)
top800k-MYWE	24.9167 (299/1200)

Table 1: Performance comparison between embeddings using T-BL

The improvement when using top200k-MYWE and top800k-MYWE was attributed to the reduced noise in the cross-lingual space since the filters removed numerous non-Malay words from the Malay embeddings space. In other words, the English words were not obscured by irrelevant 'Malay' neighbours and could induce the correct Malay translations more easily.

The seemingly higher P@10 from top200k-MYWE was actually due to the lower number of effective word pairs (1177 in the denominator) for evaluation than top800k-MYWE when both embeddings, in fact, returned an exact number of correct translations (299). In this regard, we conclude that there is no difference in the mapping quality using these embeddings. However, given that our downstream task is cross-lingual emotion classification, we are inclined towards

		Google Translate function			Googletranslate API		
	0-NN	1-NN	5-NN	10-NN	1-NN	5-NN	10-NN
P@1	9.2500	9.4167	9.5000	8.9167	9.7500	9.3333	9.1667
P@5	19.0833	19.5833	18.7500	19.5000	19.5000	18.6667	18.0000
P@10	24.9167	25.2500	24.1667	23.6667	25.5000	25.0000	23.1667

Table 2: Performance comparison using T-BL as the seed lexicon and augmenting using different translation tools (0NN: T-BL without augmentation, 1-NN: T-BL augmented with one nearest neighbour, 5NN: T-BL augmented with 5 nearest neighbours, 10-NN: T-BL augmented with 10 nearest neighbours).

top800k-MYWE, which has a broader coverage of Malay words. This would ensure that fewer Malay words in our downstream task get encoded with zero vectors. Thus, for any subsequent extensions of the experiment, we would be using top800k-MYWE.

5.2 Augmentation of T-BL for Bilingual Lexicon Extension

As shown in Table 2, regardless of the translation tools, we managed to obtain a maximum P@1 of 9.8%, P@5 of 19.6% and P@10 of 25.5% on the test bilingual lexicon (G-BL) after augmentation using T-BL. However, the mapping quality seems to be generally better when we augmented T-BL with its 1-NN and 5-NN using Google Cloud Translation API. In other words, Google Cloud Translation API returned more accurate Malay translations independent of the context of English words. For this reason, we used it for subsequent translations.

The coverage of Malay words will always increase with the number of nearest neighbours. In other words, the more number of nearest neighbours included, the higher the coverage. Based on Table 2, our hypothesis positing that higher coverage of Malay words in T-BL does not hold beyond 1-NN using either translation tool. The precisions that initially increased with Malay words coverage using 1-NN augmentation are still aligned with our hypothesis.

However, the precision started to degrade afterwards even though our augmentation broadened the coverage significantly using either translation tool. We speculate that the drop in precision after 1-NN is due to the additional noise (English nearest neighbours that are not legitimate English words) introduced to our training bilingual lexicon, thus lowering the quality of T-BL. This noise will affect the transformation matrix induction adversely when the corresponding Malay translation returned by Google happened to be in the Malay word embeddings vocabulary, as we observed that the translation tool would attempt to correct the spelling of the noise before translation. For example, 'zeroo' was translated to *sifar* (zero), 'weeka' to *mingguan* (weekly), 'talkn' to *bercakap* (talking).

5.3 Quality of T-BL

As shown in Table 3, we managed to obtain better mapping quality in terms of alignment precision using just the naïve approach.

Lexicons	P@1	P@5	P@10
T-BL	9.2500	19.0833	24.9167
N-BL	10.9167	23.3333	27.3333

Table 3: Performance comparison between T-BL and N-BL

The new set of word pairs in the bilingual lexicon (N-BL) improves P@1 by 1.7%, P@5 by 4.2%, and P@10 by 2.4%, suggesting that there is still room for improvement in T-BL quality. It is possible that the words were paired up imprecisely or paired with less frequently used words. In fact, we removed a large number of word pairs from the original non-validated English-Malay bilingual lexicon to form T-BL. This removal gave us the signal that T-BL was below par. Hence, for the experiments in Section 5.4, N-BL is used as the seed lexicon.

5.4 Augmentation of N-BL for Bilingual Lexicon Extension

As shown in Table 4, we managed to achieve a maximum P@1 of 10.9%, P@5 of 23.3% and P@10 of 28.2% on the test bilingual lexicon (G-BL) after filtering the nearest neighbours using the WordNet filter. We also observed marginal improvement over N-BL in the P@10 when augmenting with 1-NN and 5-NN using the NLTK filter. However, the best P@10 using the NLTK filter (27.8% at 1-NN) is still slightly lower than

		NLTK filter			WordNet filter		
	0-NN	1-NN	5-NN	10-NN	1-NN	5-NN	10-NN
P@1	10.9167	10.5000	8.9167	9.5000	10.9167	9.9167	9.7500
P@5	22.333	22.5000	21.2500	19.0833	23.2500	21.5000	19.5000
P@10	27.3333	27.8333	27.4167	25.4167	28.1667	26.2500	24.5000

Table 4: Performance comparison using N-BL as the seed lexicon and filtering the nearest neighbours using different filters.

the best precision using the WordNet filter (28.2% at 1-NN). While the precisions generally degraded after 1-NN, they are still higher than most of the combinations in Section 5.2 without filters.

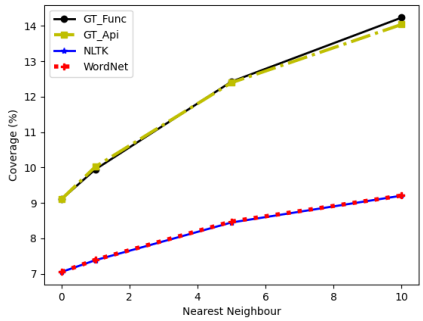

Figure 1: Comparison of the changes of coverage for experiments in Table 2 and Table 4.

Moreover, it is also worth noting from Figure 1 that the general coverage of the Malay words using N-BL is significantly lower than when using T-BL, but we managed to achieve better precisions most of the time with lower amount of computational time. This implies again that our N-BL along with the filter, is better in terms of word pairing quality than T-BL. We hit a P@10 of 28.2% at a lower coverage of about 7.4% when augmenting N-BL with 1-NN along with the WordNet filter. However, at a higher coverage of about 10%, augmenting T-BL with 1-NN using Google Translate API only resulted in a P@10 of 25.5%.

We acknowledged that having an enormous training set of English-Malay word pairs is not desirable. Using N-BL + 10NN still took us significantly longer than N-BL + 5NN and N-BL + 1NN to perform the embeddings mapping, yet the general performance degraded. From the results in Table 4, augmentation using 1-nearest neighbour is deemed ideal as it required the least translation time and training time but yielded the best mapping quality based on our experiments.

Additionally, we also compare our bilingual lexicon (N-BL + 1NN with WordNet filter) with the bilingual lexicons used for MUSE (Conneau et al., 2018) and by Anastasopoulos and Neubig (2020), which are also currently publicly available. We refer to the bilingual lexicons by Anastasopoulos and Neubig (2020) as AN-BL. Both MUSE and AN-BL have two sets of bilingual lexicons: full and train.

Lexicons	**P@1**	**P@5**	**P@10**
N-BL+1NN	10.9167	23.25	28.1667
MUSE-full	8.4167	18.3333	23.0000
MUSE-train	6.6667	16.0833	20.7500
AN-BL-full	8.0000	19.000	23.5833
AN-BL-train	7.4167	16.0833	21.0000

Table 5: Performance comparison between N-BL + 1NN, MUSE-full, MUSE-train, AN-BL-full and AN-BL-train.

As seen in Table 5, our bilingual lexicon also outperformed these bilingual lexicons by a minimum of 4.6%. We observe that there are bilingual lexicons contain word pairs of identical strings in English like *your-your*, *state-state* and *old-old* (i.e., the second word in the pair is identical to the first English word instead of being paired with its corresponding Malay word). While it is possible for English words to appear in Malay embeddings, these word pairs may disrupt the mapping to some extent if the English word also appears in the vocabulary of the Malay embeddings. In addition, our bilingual lexicon is smaller in size and requires 5 times less computational time than MUSE-full and AN-BL-full, but we show that the quality of our bilingual lexicon is better in terms of alignment precision.

5.5 Nearest Neighbours Analysis

Table 6 shows some interesting Malay translations of the English words from our G-BL test set. The term "rrc" in Table 6 is not a legitimate

Malay word as it could be the abbreviation of any words depending on the context, or possibly a result of typing errors that slipped through the cracks in spell-check. Thus, we consider this word as noise. Although none of the induced Malay translations returned matched the gold-standard translation for the word "criminal", some translations are related to this word in a way. For example, *korup* (corrupt), *pelacur* (prostitute), *siber* (cyber), *liwat* (sodomi) and *bersenjata* (armed). The word *kriminal* is not a legitimate Malay word but it is 'homophonically translated' to Malay and has been used to mean "criminal" instead of the correct Malay words *penjenayah* (the person) or *jenayah* (the noun). Such observation proves that Malay words that share similar semantic meaning to their English counterparts are mapped correctly in the bilingual word embedding space and is further strengthened by the mapping to informal Malay words even though our training bilingual lexicon contains only formal words. Since *kriminal* is non-standard and thus not included in our gold-standard bilingual lexicon (G-BL), we did not count this as a correct translation. Similarly, for the word "research", the gold standard only contains *penyelidikan* (the noun) or *menyelidik* (the verb) as its translations, completely leaving out other correct translations such as *kajian*. Moreover, the word *studi* is also an informal Malay word commonly used to represent "study" or "research".

criminal	**research**	**vegetable**	**sad**
korup	pemantauan	perasa	cemburu
kriminal	penyelidikan	makaroni	jijik
rrc	analisis	pete	kecewa
pelacur	penulisan	petai	menyampah
siber	pembelajaran	pindang	cuak
liwat	kajian	bayam	rimas
zalim	riset	tomat	berdosa
rasis	statistik	salmon	terharu
lgbt	studi	sayuran	sebak
bersenjata	penelitian	sardin	sedih

Table 6: Nearest neighbours of selected English words

For the Malay word "vegetable", we observed several semantically similar words to "vegetable" are returned such as *petai* (bitter bean), *bayam* (spinach) and *sayuran* (a variety of vegetables). Regardless of the plurality, *sayuran* is the closest Malay word to vegetable among the induced translations. *Tomat* is possibly a result of typing error for the word *tomato*. On the other hand, we observed Malay words of negative emotions are also returned in addition to the correct translation *sedih* for the word "sad" such as *cemburu* (jealous), *kecewa* (disappointed), *cuak* (scared) and *rimas* (uneasy or anxious). Although there are some Indonesian words in our Malay embeddings space which were not filtered out during the pre-training, such as *riset* and *pete*, these words will be eliminated as Indonesian tweets are removed in our downstream task.

6 Conclusion

In this study, we attempted to create English-Malay cross-lingual word embeddings using an English-Malay bilingual lexicon to map the English and Malay monolingual word embeddings into a single representation that was empirically and intrinsically evaluated based on word pair coverage and alignment precision. Despite the fact that the bilingual lexicon from Malaya Documentation being the current state-of-the-art, we demonstrated that its quality has room for improvement. Our bilingual lexicons (N-BL) obtained using a naïve approach easily outperformed Malaya Documentation in terms of the English-Malay alignment precision. We also investigated the effect of Malay word coverage on bilingual lexicon induction and discovered that a higher coverage would not necessarily improve the alignment precision. Also, we did not select our training or test lexicons based on word frequency in any corpora, thus our evaluation is more unbiased and generalizable.

We are aware that there are semi-supervised and unsupervised approaches in creating cross-lingual word embeddings that require limited or almost no bilingual signals. However, we did not adopt such an approach because both our embeddings were pre-trained on informal corpora, especially our Malay monolingual embeddings still containing significant noise even after applying the filter. Hence, legitimate words would easily be paired up with noise, or vice versa, without bilingual supervision. We adopted word2vec in this study as we were not aware of any existing Malay fastText embeddings pre-trained on the social media domain, and pre-training it ourselves is not within the scope of this study. In the future, we wish to pre-train Malay fastText embeddings that may work better on informal corpora and subsequently explore the feasibility of creating embeddings using semi-supervised and unsupervised methods.

We also plan to evaluate the performance of our English-Malay cross-lingual word embeddings on downstream tasks such as emotion classification.

Acknowledgement

This study was supported by the Ministry of Higher Education Malaysia for Fundamental Research Grant Scheme with Project Code: FRGS/1/2020/ICT02/USM/02/3.

References

Aldarmaki, Hanan, & Diab, Mona. (2019). Context-Aware Cross-Lingual Mapping. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 3906–3911.

Anastasopoulos, Antonios, & Neubig, Graham. (2020). Should All Cross-Lingual Embeddings Speak English? Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 8658–8679.

Artetxe Mikel, Labaka Gorka and Agirre Eneko. (2016). Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2289–2294.

Artetxe Mikel, Labaka Gorka and Agirre Eneko. (2018). A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 789–798

Barone Antonio Valerio Miceli. (2016). Towards cross-lingual distributed representations without parallel text trained with adversarial autoencoders. Proceedings of the 1st Workshop on Representation Learning for NLP, 121–126.

Bird, Steven, Klein, Ewan, & Loper, Edward. (2009). Natural Language Processing With Python: Analyzing Text With The Natural Language Toolkit. O'Reilly Media, Inc.

Conneau, Alexis, Khandelwal, Kartikay, Goyal, Naman, Chaudhary, Vishrav, Wenzek, Guillaume, Guzmán, Franciso, Grave, Edouard, Ott, Myle, Zettlemoyer, Luke, & Stoyanov, Veselin. (2020). Unsupervised Cross-Lingual Representation Learning At Scale. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 8440–8451.

Conneau, Alexis, Lample, Guillaume, Ranzato, Marc' Aurelio, Denoyer, Ludovic, & Jégou, Hervé. (2018). Word Translation Without Parallel Data. ArXiv:1710.04087 [Cs].

Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, & Toutanova, Kristina. (2019). BERT: Pre-Training Of Deep Bidirectional Transformers For Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 4171–4186.

Dinu, Georgiana, Lazaridou, Angeliki, & Baroni, Marco. (2015). Improving zero-shot learning by mitigating the hubness problem. ArXiv:1412.6568 [Cs].

Faruqui Manaal and Dyer Chris. (2014). Improving Vector Space Word Representations Using Multilingual Correlation. Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, 462–471.

Feng Yanlin and Wan Xiaojun. (2019). Learning Bilingual Sentiment-Specific Word Embeddings without Cross-lingual Supervision. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 420–429.

Ghasemi, Rouzbeh, Ashrafi Asli, Syed Arad, & Momtazi, Saeedeh. (2020). Deep Persian Sentiment Analysis: Cross-Lingual Training For Low-Resource Languages. Journal of Information Science, 016555152096278.

Godin Fréderic. 2019. "Improving and Interpreting Neural Networks for Word-Level Prediction Tasks in Natural Language Processing." Ghent University, Belgium.

Husein Zolkepli. (2018). Malaya, Natural-Language-Toolkit library for bahasa Malaysia, powered by Deep Learning Tensorflow [Github]. Malaya.

Lu Ang, Wang Weiran, Bansal Mohit, Gimpel Kevin, and Livescu Karen. (2015). Deep Multilingual Correlation for Improved Word Embeddings. Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 250–256.

Mikolov Tomas, Chen Kai, Corrado Greg and Dean Jeffrey. (2013). Efficient Estimation of Word Representations in Vector Space. ArXiv:1301.3781 [Cs].

Mikolov Tomas, Le Quoc V., and Sutskever Ilya. (2013). Exploiting Similarities among Languages for Machine Translation. ArXiv:1309.4168 [Cs].

Pennington Jeffrey, Socher Richard and Manning Christoper. (2014). Glove: Global Vectors for Word Representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1532–1543.

Peters, Matthew. E., Neumann, Mark, Iyyer, Mohit, Gardner, Matt, Clark, Christopher, Lee, Kenton, & Zettlemoyer, Luke. (2018). Deep Contextualized Word Representations. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2227–2237.

Ruder Sebastian, Vulić Ivan and Søgaard Anders. (2019). A Survey of Cross-lingual Word Embedding Models. Journal of Artificial Intelligence Research, 65, 569–631.

Schuster, Tal, Ram, Ori, Barzilay, Regina, & Globerson, Amir. (2019). Cross-Lingual Alignment of Contextual Word Embeddings, with Applications to Zero-shot Dependency Parsing. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 1599–1613.

Smith Samuel L., Turban David H. P., Hamblin Steven and Hammerla Nils Y. (2017). Offline bilingual word vectors, orthogonal transformations and the inverted softmax. ArXiv:1702.03859 [Cs].

Wang, Zirui, Xie, Jiateng, Xu, Ruochen, Yang, Yiming, Neubig, Graham, & Carbonell, Jaime. (2020, May). Cross-lingual Alignment vs Joint Training: A Comparative Study And A Simple Unified Framework. Proceedings of the 8th International Conference on Learning Representations, ICLR 2020.

Xing Chao, Wang Dong, Liu Chao and Lin Yiye (2015). Normalized Word Embedding and Orthogonal Transform for Bilingual Word Translation. Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1006–1011.

Zhang Meng, Liu Yang, Luan Huanbo and Sun Maosong. (2017). Adversarial Training for Unsupervised Bilingual Lexicon Induction. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 1959–1970.