Learning Indonesian-Chinese Lexicon with Bilingual Word Embedding Models and Monolingual Signals

Xinying Qiu
CISCO School of Informatics
Guangdong University of Foreign Studies
Guangzhou, China
qiuxinying@gdufs.edu.cn

Gangqin Zhu
The Faculty of Asian Languages and Cultures
Guangdong University of Foreign Studies
Guangzhou, China
199210621@oamail.gdufs.edu.cn

Abstract

We present a research on learning Indonesian-Chinese bilingual lexicon using monolingual word embedding and bilingual seed lexicons to build shared bilingual word embedding space. We take the first attempt to examine the impact of different monolingual signals for the choice of seed lexicons on the model performance. We found that although monolingual signals alone do not seem to outperform signals coverings all words, the significant improvement for learning word translation of the same signal types may suggest that linguistic features possess value for further study in distinguishing the semantic margins of the shared word embedding space.

1 Introduction

We explore the latest development of bilingual lexicon learning (BLL) research and investigate their application on inducing Indonesian-Chinese lexicon. In particular, due to the limitation of parallel and comparable Indonesian-Chinese bilingual corpora, we study the state-of-the-art bilingual word embedding (BWE) models built with seed lexicons and monolingual corpora to project Indonesian and Chinese word pairs onto the same transformed space. We further explore the impact of Indonesian linguistic signals on these models to provide insights on the implications of monolingual signals and challenges for bilingual lexicon learning

Bilingual word embedding models have proven to be effective in many cross-lingual tasks such as document classification, POS tagging, and phrase generation. As illustrated in Figure 1, two sets of words (numbers and animals) in two languages (English and Spanish) have similar geometric arrangements. This is achieved by constructing wore embedding vectors for both languages and projecting the vectors down into two dimensions, rotated to show similarity. The Figure demonstrates that the relations between words are similar across languages. This finding inspired a series of research on generating a bilingual dictionary with cross-lingual word embedding space. The general steps involve 1) building a word space for each individual language; 2) projecting the two spaces into one shared space or from one to the other; and 3) learning or retrieving the target language word most similar to the source language word in the projection.

Our paper attempts to contribute to this line of research by examining the monolingual signals from Indonesian in building the bilingual word embedding model. The rest of the paper is organized as follows. In Section 2, we review the latest development in BLL with BWE. We summarize the related research and propose our research questions. In Section 3, we discuss details of our methodologies. We present our data preparation, experiment design, results and analysis in Section 4, and conclude with Section 5.

2 Research Framework and Related Work

Bilingual lexicon learning aims at enriching existing bilingual dictionaries and building new dictionaries to cross the language barriers between under-resourced languages and resourced languages. Many research endeavors such as the dictionary extraction from Wiktionary (Sérasset and Tchechmedjiev

Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing,
pages 188–193, Osaka, Japan, December 11-17 2016.

(2014)), and the SisTec-embt Project (Al-Adhaileh et.al. (2002)) have explored the automatic dictionary construction methods and linguistics features for machine learning systems. To our limited knowledge, the electronic version of Indonesian-Chinese dictionary is currently only available as hardware devices for language learners, the content of which is not extractable as stand-alone softcopy. The hard-copy of Indonesian-Chinese dictionary is a little out-dated which makes the digitization work not much desirable. Both Google and Bing provide translation between Indonesian and Chinese (and vice versa), but with great deal of errors. Therefore, we find it a challenging research project to learn automatically Indonesian-Chinese lexicon, with many application opportunities for example, as building blocks for machine translation systems, document classification, and sentiment analysis. Our work is of explorative nature. We currently focus on learning simple vocables excluding collocations, and not restricted to specific domain or distinguishing senses. We aim at examining the performance of bilingual word embedding model complemented with monolingual signals in learning Indonesian-Chinese lexicon. We hope that the development and improvement of such models and algorithms would support the more efficient generation of large-volume and high quality bilingual dictionary.

We define our Indonesian-Chinese bilingual learning problem along the following dimensions: usage of monolingual word embedding and signals, bilingual signals, bilingual word embedding model, and learning algorithm, as inspired by the frameworks proposed by Upadhyay et al. (2016) and Vulic and Korhonen (2016).

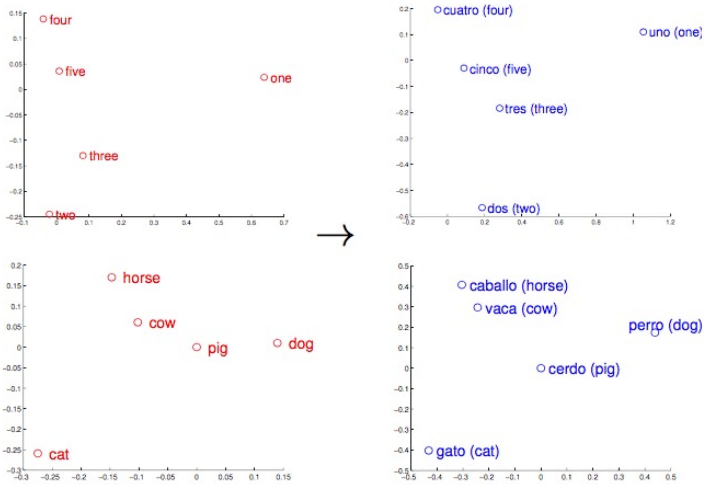

Figure 1: The idea behind transformation matrix. (from Mikilov et al. 2013)

Upadhyay et al. (2016) compared empirically some of the most recent development on cross-lingual models of word embeddings. They come up with a general schema as shown in Figure 2. Their empirical comparisons focus on the “bilingual corpus” component covering parallel corpus (Luong et. al. 2015), comparable corpus (Vulic and Moen, 2015), sentence-aligned corpus (Hermann and Blunsom, 2014), and bilingual lexicon (Faruqui and Dyer, 2014; Mikolov et al. 2013; Dinu et al. 2015). Their findings suggest that the most expensive supervision of training data such as word alignment may be more suitable for bilingual lexicon learning.

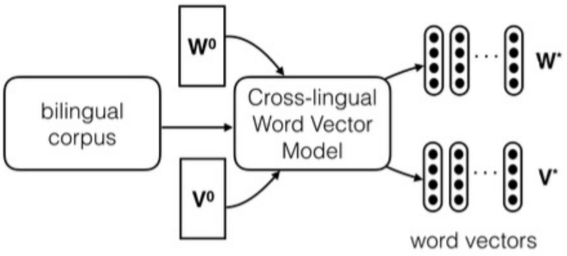

Figure 2: Cross-Lingual Word Embedding Schema as from Upadhyay et al. (2016)

Similarly, Vulic and Korhonen (2016) defined Bilingual Word Embedding (BWE) model as "induction of *a shared bilingual word embedding space (SBWES)*". They further proposed two desirable properties for BWE model as 1) usage of monolingual training sets tied with bilingual signals; and 2) inexpensive bilingual signal. In their setting, the "bilingual signals" are equivalent with the "bilingual corpus" in the schema by Upadhyay et al. (2016). The "monolingual training set" property is in consistent with Upadhyay et al. (2016) in their generalization of the loss function with monolingual corpora. The "inexpensive" requirement is in line with the "supervision cost" discussed by Upadhyay et al. (2016). They also suggest that for Bilingual Lexicon Learning, careful selection of seed lexicon (thus more expensive human supervision) may produce better results (Vulic and Korhonen (2016)).

By integrating these two frameworks, we demonstrate our research framework to induce Indonesian-Chinese lexicon as shown in Figure 3. Due to the lack of parallel and comparable corpora, and also because BLL is proven to be better supported with more expensive knowledge, we opt for using seed bilingual lexicon as our bilingual signal.

For learning algorithms, previous researchers have examined supervised or distantly supervised models (Irvine and Callison-Burch, 2015; Gouws and Sogaard, 2015), and unsupervised models (Mikolov et al. 2013; Luong et. al. 2015). Dinu et al. (2015) modified Mikolov's nearest neighbour method with zero-shot paradigm to correct the bilingual translations by considering the hubness of the candidate target language words. In this paper, we experiment with both Dinu's and Mikolov's unsupervised learning algorithms. We will explore the supervised learning approaches in our future work.

Many researchers have suggested that monolingual signals or features may impact on the learning the cross-lingual word embedding models, such as Irvine and Callison-Burch (2015), Vulic and Korhonen (2016) and Dinu et al. (2015). Inspired by the research discussed above, we propose to examine different monolingual signals to analyse their impact on bilingual word embedding models for Indonesian-Chinese lexicon learning.

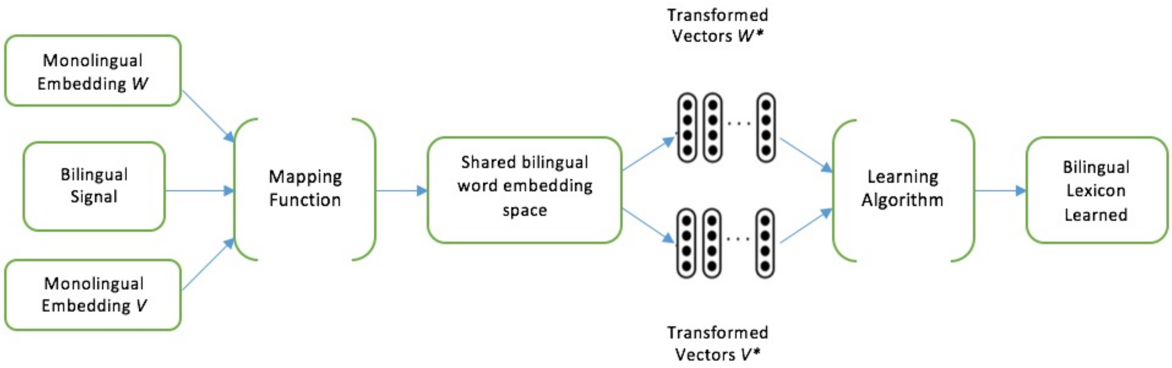
Monolingual Embedding *W*
Bilingual Signal
Monolingual Embedding *V*
Mapping Function
Shared bilingual word embedding space
Transformed Vectors *W**
Transformed Vectors *V**
Learning Algorithm
Bilingual Lexicon Learned

Figure 3: Our Research Framework

3 Methodologies

As discussed in Section 2, we choose the following methodologies within our proposed research framework:

1) We build monolingual embedding models for Indonesian and Chinese respectively. We use a seed Indonesian-Chinese lexicon as bilingual signals to tie up the monolingual word embeddings.

2) We experiment with Mikolov's mapping function to generate transformation matrix from which we could generate vectors of the two languages projected onto the same space. We also experiment with Dinu's method to mitigate the impact of hub vectors in the vector space.

3) Learning of translated Chinese words for the test data set is based on nearest neighbour retrieval. Evaluation method is the standard Precision@k for bilingual lexicon learning, with which we report results for k as 1, 5, and 10.

4) We examine the impact of the following monolingual signals on the performance of the word embedding models: nouns, root words, high-frequency words, and unambiguous words.

3.1 Mikolov's Mapping

Mikolov et al. (2013) proposed a method to use distributed representation of words and learns a linear mapping between vector space of different languages. More specifically, the model is as illustrated in the following equation:

$$\min_{W} \sum_{i=1}^{n} \|Wx_i - z_i\|^2$$

where $W$is the transformation matrix; x_i the vector of the source language word; z_i the vector of the target language word. When such a transformation matrix is learned, to retrieve the translation of a new word with its vector x, we may compute a new vector $z = Wx$, and find the nearest neighbour vector in the target language space.

3.2 Dinu's Hub-Correction

Dinu et al. (2015) proposed to improve over Mikolov's retrieval method by solving the *hubness* problem when retrieving target words with the following *globally-corrected* approach:

$$GC(x, T) = arg \min_{y \in T}(Rank_{y,P}\,(x) - \cos(x, y))$$

where x is the vector in the source language space; $Rank_{y,P}(x)$ measures the rank of x in the set of pivot vectors P with respect to its similarity to y in the target space; *cosine* score is used to break ties for the candidate target words.

3.3 Monolingual Signals

Inspired by previous research that discuss monolingual signals, the importance of seed lexicon choice, and the problem caused by "hub" words in the vector space, we propose to examine the following monolingual signals' impact on bilingual word embedding models.

Nouns: As we study "Kamus Besar Bahasa Indonesia" (the Grand Indonesian Dictionary), we found that out of the 7 POS tags available for Indonesian, the NOUN words take up the largest proportion of of 56.6%, with the second popular POS tag being VERB, taking up only 29.7% of Indonesian words. Considering this phenomenon, we experiment to see if words of NOUN type alone could serve as better monolingual signal in learning the bilingual word embedding models.

Root words: Indonesian is an agglutinative language, with many words derived as inflectional forms of root words with prefixes and/or suffixes attached. For example:

Root words:		*Derived words:*

kerja (work, *n.*)	---------->	bekerja (work, *v.i.*); mengerjakan (work, *v.t.*);
		pekerja (worker, *n.*); mempekerjakan (employment, *n.*);

We propose that by selecting root words and their Chinese translation as seed lexicon, we might be able to generate a more coherent transformation matrix that reduces the semantic similarity between word of inflectional variations within the bilingual word embedding space.

High-frequency words: Vulic and Korhonen (2016) suggested that words with higher frequency are more reliably translated to guarantee the quality of the seed lexicon. In addition to that, we also hypothesize that by selecting the more frequently used words to construct bilingual seed lexicons, we might be able to cover the more popularly discussed semantics in the transformation matrix.

Unambiguous words: The polysemy phenomenon in Indonesian languages may give us multiple translation entries for a single word in the seed lexicon. It is also quite common for a single Indonesian word to be matched with multiple similar Chinese translations. For example:

Polysemy:

Peringatan ----> 纪念 (commemorate); 警告 (warning)

Multiple translations with similar meanings:

Berkah ----- > 恩赐 (bestow), 祝福(blessing)

We hypothesize that by selecting highly unambiguous and monosemous translation pairs, we may be able generate vector space with more semantic margins between word vectors, and therefore improving the target word selection performance.

4 Experiments and Results

4.1 Data and Evaluation

For building monolingual word embedding models, we use Chinese and Indonesian Wikipedia articles as training set. We collected and processed the Chinese Wikipedia dump of Aug. 1 2016 and the Indonesian Wikipedia dump of July 20, 2016 and generate 727k Chinese word vectors and 190k Indonesian vectors.

For bilingual lexicons, we take the complete vocabulary from "Kamus Besar Bahasa Indonesia" (the Grand Indonesian Dictionary) and run the Google translation and Bing translation. Since both translation systems generate a great deal of errors, we take the same translation from both systems hoping for better accuracy. One of our authors (an Indonesian language teaching professor) manually filtered out the correct word-pairs from this translation set. We also take the vocabulary from the Indonesian language textbooks for Chinese learners to include with the word-pairs from the Grand Dictionary. Therefore, we have a collection of 10436 Indonesian-Chinese word-pair lexicon. Out of this base seed lexicon, we select nouns, root words, high-frequency words (as from the basic-level and medium-level Indonesian textbook for Chinese language learners), and highly-unambiguous words. The statistics are as follows:

All words	High-frequency	Nouns	Root-words	Unambiguous
10436	5037	4078	5493	2502

For each of the above 5 monolingual signals, we experiment with Mikolov's and Dinu's methodologies respectively. We take 10% of the data as test set, 90% as training set. We perform two types of experiment designs: Design 1: We build test data by randomly selecting 10% of all-words data. We build 5 training models with the five signal data without overlapping with the test set; Design 2: For each of 5 signal data, we randomly select 10% for testing, and the rest for training. In other words, each experiment is performed within the data of the same signal themselves.

These experiments evaluate the impact of signal for learning a general lexicon, and for learning lexicon of their own signal types. We evaluate performance with the standard Precisions @ 1, 5, and 10.

4.2 Results and Analysis

We first present some examples of learning results, with the correct translations retrieved in bold. There are many cases where the retrieved translation words rank as far as the hundredth.

Indonesian word:	Retrieved translation: rank: Chinese (English, cosine score)
murid (student) ----->	**#1: 学生 (student, 0.705);** #2: 老师 (teacher, 0.636); #3: 女生(female student, 0.624); #4: 班级(class, 0.621); #5: 毕业生(graduates, 0.588)
gembira (happy) ----->	#1: 难过(sad, 0.685); #2: 想念(yearn for, 0.669); #3: 伤心(grief, 0.663); #4: 吃惊(surprised, 0.643); **#5: 高兴(happy, 0.640)**

Table 1 presents results for testing on the same data set randomly selected from all-words lexicon, i.e. 1043 word-pairs. We find that the 4 special monolingual signals alone do not seem to improve the learning performance over the model built with all-words. The best performance is highlighted for "all-words" with Mikolov's method at 0.514 for precision at 10.

Table 2 presents results for testing within the same signal type. For the All-words data, this experiment design is the same as Design 1. We repeat it in the table for comparison purposes.

We have the following findings from Design 2:

1) NOUNs, root words, and highly unambiguous words all perform better in retrieving the correct translations for words of their own signal types.

2) Model with unambiguous words performs the best with a 0.632 precision at 10, much higher than even the all-words signal. We may infer that the better performance may come from the fact the transformation space is composed of highly distinguished vectors representing the drastic difference in words' semantics.

3) Dinu's method with hubness correction performs well with root words signal and test on its own data. This may be because root words data set support the elimination of similar target words that may push down the ranking of the correct translation.

		All words	High-Frequency	Nouns	Root words	Unambiguous
Mikolov's	Pre@1	0.244	0.213	0.204	0.201	0.217
	Pre@5	0.434	0.419	0.403	0.399	0.394
	Pre@10	**0.514**	0.509	0.474	0.478	0.472
Dinu's	Pre@1	0.248	0.223	0.201	0.218	0.227
	Pre@5	0.422	0.420	0.392	0.398	0.385
	Pre@10	0.481	0.470	0.443	0.449	0.435

Table 1: Design 1 -- Test on All-words Lexicon

		All words	High-Frequency	Nouns	Root words	Unambiguous
Mikolov's	Pre@1	0.244	0.110	0.266	0.265	0.318
	Pre@5	0.434	0.344	0.457	0.458	0.538
	Pre@10	**0.514**	0.407	**0.531**	0.529	**0.632**
Dinu's	Pre@1	0.248	0.114	0.269	0.272	0.323
	Pre@5	0.422	0.319	0.454	0.473	0.498
	Pre@10	0.481	0.392	0.529	**0.544**	0.516

Table 2: Design 2 – Test with data from the same signal type

5 Conclusions

We present a research on learning Indonesian-Chinese bilingual lexicon using monolingual word embedding and bilingual seed lexicons to build shared bilingual word embedding space. The aim of the work is to develop and improve bilingual lexicon learning models, as building block for research on machine translation and cross-language NLP. We apply the latest development on BWE framework and also take the first attempt to examine the possible impact of different monolingual signals for the choice of seed lexicons on the model performance. We found that although monolingual signals alone do not seem to outperform signals coverings all words, the significant improvement for learning word translation of the same signal types may suggest that linguistic features possess value for further study in distinguishing the semantic margins of the shared word embedding space. For our future work, we plan on studying the impact of word senses, collocation, and other lexical features on the BWE model.

References:

Sérasset G, Tchechmedjiev A. Dbnary: Wiktionary as Linked Data for 12 Language Editions with Enhanced Translation Relations. 3rd Workshop on Linked Data in Linguistics. 2014.

Al-Adhaileh M H, Kong T E, Yusoff Z. A synchronization structure of SSTC and its applications in machine translation. Proceedings of the 2002 COLING workshop on Machine translation in Asia. Volume 16. 2002

Mikolov T., Quoc V. Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. CoRR, abs/1309.4168. 2013.

Upadhyay S, Faruqui M, Dyer C, et al. Cross-lingual Models of Word Embeddings: An Empirical Comparison[J]. arXiv preprint arXiv:1604.00425, 2016.

Vulic I, Korhonen A. On the Role of Seed Lexicons in Learning Bilingual Word Embeddings. ACL, 2016.

Luong T, Pham H, Manning C D. Bilingual word representations with monolingual quality in mind. Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing. 2015: 151-159.

Vulic ́ I and Marie-Francine Moens. Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In Proc. of ACL, 2015

Hermann K. and Phil Blunsom. Multi-lingual models for compositional distributed semantics. In ACL, 2014.

Faruqui M, Dyer C. Improving vector space word representations using multilingual correlation. In ACL, 2014.

Dinu G, Lazaridou A, Baroni M. Improving zero-shot learning by mitigating the hubness problem. ICLR, 2015.

Irvine A, Callison-Burch C. Discriminative Bilingual Lexicon Induction. Computational Linguistics, 2015, 1(1).

Gouws S, Søgaard A. Simple task-specific bilingual word embeddings. Proceedings of NAACL-HLT. 2015: 1386-1390.