Learning to Respond to Mixed-code Queries using Bilingual Word Embeddings

Chia-Fang Ho[1], **Jhih-Jie Chen**[2], **Ching-Yu Yang**[2], **Jason S. Chang**[2]
[1]Institute of Information Systems and Applications
National Tsing Hua University
[2]Department of Computer Science
National Tsing Hua University
{fun, jjc, chingyu, jason}@nlplab.cc

Abstract

We present a method for learning bilingual word embeddings in order to support second language (L2) learners in finding recurring phrases and example sentences that match mixed-code queries (e.g., *"接受 sentence"*) composed of words in both target language and native language (L1). In our approach, mixed-code queries are transformed into target language queries aimed at maximizing the probability of retrieving relevant target language phrases and sentences. The method involves converting a given parallel corpus into mixed-code data, generating word embeddings from mixed-code data, and expanding queries in target languages based on bilingual word embeddings. We present a prototype search engine, *x.Linggle*, that applies the method to a linguistic search engine for a parallel corpus. Preliminary evaluation on a list of common word-translation shows that the method performs reasonably well.

1 Introduction

Many queries are submitted to search engines on the Web every day to retrieve linguistic information for learning a second language (L2), and an increasing number of search engines specifically target queries for finding translations of phrases and sentences. For example, *Linguee* (www.linguee.com) accepts L1 queries and retrieves bilingual sentences (L1+L2), while *Google Translate* (translate.google.com) is used to translate (mixed-code) texts, and return L2 results.

Due to limited L2 vocabulary knowledge, users often submit mix-coded queries, but search engines such as *Linguee* only retrieve sentences similar to queries without converting them into target language queries.

By transforming L1 keywords in the original query into relevant L2 keywords, we can bias the search engine toward retrieving relevant L2 phrases and sentences for language learning.

We present a system, *x.Linggle*, that automatically processes mixed-code queries into monolingual queries and retrieves relevant phrases and examples to users. See Figure 1 for an example of *x.Linggle* search results of the query "接受 *education*". As shown in Figure 1, *x.Linggle* is accessible at https://x.linggle.com. *x.Linggle* has determined several L2 keywords for the L1 keyword "接受" by calculating cosine similarities between word vectors in the bilingual embedding space and convert the query into L2 queries (e.g., "receive education", "obtain education", "accept education"). Then, *x.Linggle* retrieves and ranks the results of these L2 queries according to occurrence counts, and finally returns relevant phrases with example sentences.

The rest of the article is organized as follows. First, we present our method for deriving bilingual word embeddings to support mixed-code queries. Next, we introduce the search engine in which we integrate our mixed-code query system. Then, we conduct a preliminary evaluation on the most common 7000 vocabulary for ESL learners. Finally, we conclude in Section 6.

2 Related Work

Word representation or word embedding has been an area of active research. It has been shown that predicting instead of counting context words leads to better representation of lexical semantics and relation between words (Mikolov et al., 2013; Pennington et al., 2014). We consider the specific case of learning word representation of two languages simultaneously, instead of a single language.

Previously proposed methods use a rotation matrix to learn the relation between word embeddings of the two languages. Conneau et al. (2017);

Proceedings of NAACL-HLT 2019: Demonstrations, pages 24–28
Minneapolis, Minnesota, June 2 - June 7, 2019.

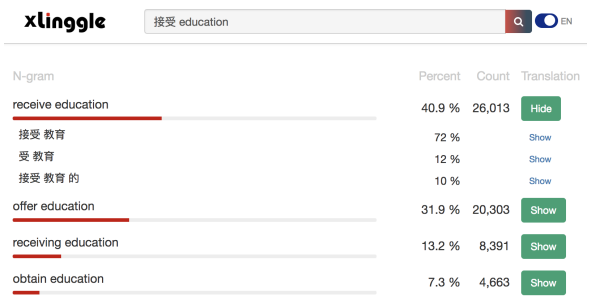

Figure 1: The prototype system, *x.Linggle*

Figure 2: The extended example mechanism of *x.Linggle*

Duong et al. (2017) relate cross-lingual information based on a small set of word-translation pairs. Our approach is different in that we use mixed-code data converted from a parallel corpus, to derive directly an embedding space with word tokens in two languages, instead of learning a matrix transforming between two independent language embedding spaces.

In a study more closely related to our work, Gouws and Søgaard (2015); Vulić and Moens (2015) process a document-aligned comparable corpus as training data while Luong et al. (2015) processes mixed-code sentences for Cross-lingual Document Classification (CLDC) task. We use a similar training methodology for the different purpose of responding to mixed-code queries by converting mixed-code query into L2 queries based on the bilingual embedding.

In area of evaluating embedddings, researchers have typically used Spearman's rank correlation coefficients and Word Similarity to measure the quality of word embeddings (e.g., Mikolov et al. (2013); Pennington et al. (2014)). In contrast, we evaluate bilingual embedding by measuring the coverage of appropriate and relevant translation of the mixed-code queries.

In contrast to the previous research in word representation and bilingual word embeddings, we present a system that automatically converts mixed-code linguistic queries (which may contain L1 keywords or part of speech wildcards) so as to retrieve relevant phrase and sentences to assist language learning.

3 Bilingual Word Embeddings

Combining two separate word embeddings using a limited set of word-translation pairs to form a bilingual word embedding might not work very well. Word embedding vectors typically represent many word senses, while translations may cover only the dominant sense. To develop bilingual word embeddings, a promising approach is to artificially generate a mixed-code dataset based on a parallel corpus.

*Problem Statement:*We focus on the preprocessing step of mixed-code answering process: training bilingual word embedding model. We are given a mixed-code query Q_{mc}, a parallel corpus, and an L2 linguistic search engine. Our goal is to respond to the query, and retrieve relevant recurring L2 phrases and sentences. For this, we derive a bilingual word embedding V, such that $V(W)$ for an L1 keyword W (e,g,, "接受") in Q_{mc} is close to $V(T)$ for most L2 word T (e.g., "receive") relevant to W. Therefore the system can use V to retrieve "receive education" for the query of "接受 education".

The method involves (i) training a bilingual word embedding beforehand, (ii) searching for similar L2 words for L1 keyword in the embedding space, (iii) convert and expanding the mixed-code query for retrieving relevant phrases and sentences in the target language. To train word embeddings, we adopt the approach proposed by Mikolov et al. (2013), to derive a continuous, *semantic* representation of words based on context. Consider the flexibility, our method provides a framework of methodology and elements can be change according to different target. For example, bilingual word embedding model in different languages can be trained simply replace training data with other language corpus. Moreover, any word embedding training method (e.g., Mikolov et al. (2013); Pennington et al. (2014)) can be applied to train bilingual word embeddings.

However, if we only train with monolingual sentences, we can not find cross-lingual relation

for our purpose. Therefore, we transplant the translation of a word into the sentence to generate artificially mixed-code sentences, and then train a word embedding model to encode cross-lingual information.

3.1 Transplanting Translations

In order to train word embeddings that with cross-lingual information, we generate mixed-code sentences from parallel sentences by transplanting word translation into the source sentences. For this, we used *Hong Kong Parallel Text* (HKPT), which consists of pairs of Chinese and English sentence with word-level alignments. The HKPT corpus consists of nearly 3M parallel sentences with 59M English and 98M tokens.

However, the alignment of Chinese and English does not correspond exactly word by word, and some even involve non one-to-one (1-1) alignment, leading to difficulties in transplanting. To cope with this problem, we perform the following training data preprocessing procedure.

Preprocessing Parallel Sentence

First, we merge possible multi-words as inseparable units whenever a word aligns to consecutive multiple words. Due to the differences between Chinese and English segmentation, for example, the alignment of English token "power plant", "發電廠"(which could be segmented into "發電" and "廠"). If that is the case, then the model can learn fine-grained information (e.g., "power" → "發電", "plant" → "廠") during training. For this reason, we change the word segmentation and realign, in order to derive more 1-1 correspondances. A transformation table is built to convert alignment of two English words and one Chinese word (e.g., "power plant" → "發電廠") into two pairs of 1-1 word alignment (e.g., "power" → "發電", "plant" → "廠") based on lexical translation probability derived from the dataset itself. With the transformation table, parallel corpus sentences are re-aligned and our model can perform better because of more information is available for individual words, which was previously not possible due to non 1-1 alignments.

Transplanting Translations

After preprocessing, we generate mixed-code sentence by replacing words with their alignment counterpart. It is important to note that we only replace one token at a time for simplicity. As it turns out, this approach worked just fine.

First, for each of the two languages, we generate mixed-code sentences by replacing one token in the source sentence with its corresponding foreign token. This process repeats for each content word in the L2 sentence to generate mixed-code sentences (e.g., 'I 有 a dream .', 'I have 一個 dream .' ...)

3.2 Word Embedding Training

We apply Skip-Gram models with negative sampling technique which reduces the noise distribution by logistic regression while using parallel corpus data as our training data. With parallel dataset, we generate training sentences by replacing source language tokens with target ones to obtain the neighbors of a token not only in the source language but also in the target language. Skip-gram model tries to predict current word's neighbors (its context) by giving a set of sentences (also called corpus), and the model loops on the words of each sentence and learn relation between words in a vector space. We train word embeddings model with the mixed-code sentences by putting them into pairs of a target word and its context words. (e.g., target word: **have**, context word: [我, 有, 一個, 夢]) Finaly, words in both languages can be represented in the same embeddings space.

When user submits an mixed-code query, L1 tokens in query are converted into candidate tokens in L2 by calculating distances of token vectors in a bilingual word embeddings model.

We convert and expand L1 keywords into L2 queries and re-rank the results to these queries by frequency.

4 x.Linggle: a mixed-code Linguistic Search Engine

We build our system based on an underlying linguistic search engine, *Linggle*, by (Boisson et al., 2013), supporting a set of wildcard query symbols. Figure 1 shows an example search performed by the system. Figure 3 describes the query symbols with examples. In addition to mixed-code query, we also offer on-demand display of example sentences to assist learners in writing or translation. We introduce the query symbol in the next section.

Operators	Description	Example
_	match any single word	drive _ car
*	match zero or more words	ready * change
?	search for TERM optionally	discuss ?about the issue
~	search for similar words	play an/a ~important role
/	either TERM1 or TERM2	in/at/on the afternoon
{}	order of TERM1, TERM2, TERM3, ...	{know where is she}
PoS.	search for words with specific PoS. (v, n, adj, adv, prep, det, conj, pron)	v. death penalty

Figure 3: Query operator instruction

4.1 Query Symbols

An underline (_) match any single word (e.g., "drive _ car"), while wildcards (*) match zero or more words (less than 4) (e.g., "ready * change"). Additionally, the question mark (?) before a word or part of speech symbol match nothing or the word/pos that follows.(e.g., "discuss ?about the issue") Use tilde (~) before a word to search for synonyms(e.g., "play an/a ~important role"). To match any of a list of words, use the symbols (/) (e.g., "in/at/on the afternoon"). Use curly brackets ({}) to match a list of words in any order (e.g., "{know where is she}"). Finally, a set of part of speech symbols can be used to match any single word with the designated POS (e.g., "v. death penalty")

4.2 Example and Translation

The original *Linggle* provides example sentences containing retrieved phrases to help learners learn the usage. We take a step further and extend the example mechanism. In our system, possible translations are shown first, and then parallel examples are provided. In so doing, learners not only learn the actual usage but also understand the nuance between phrases through the examples in their native language. The extended version of example is shown in Figure 2.

5 Preliminary Evaluation

The goal of this work is to enable a cross-language search engine to answer mixed-code queries, the model should be evaluated according to how well it covers relevant translations. We conduct a preliminary evaluation on a list of the most common 7000 words for intermediate high school ESL learners[1], with translations from an official Website of Ministry of Education in Taiwan. With the dataset, we compare our model with *Cambridge Dictionary* in terms of covering the words and translation. The evaluation results show the proposed model model perform on par with the *Cambridge English-Chinese Dictionary* covering around 51% of the word-translation list.

[1]https://zh.wikibooks.org/zh-tw/英語/高中7000辭彙

6 Conclusion

Many avenues exist for future research and improvement of our system. For example, existing methods for ranking relevant phrases from queries could be implemented. Ranking phrases according to TF-IDF score instead of frequency could be used to improve relevance between queries and phrases. Additionally, an interesting direction to explore is disambiguating word sense by constructing a graph linking context words to sense translations based on bilingual word embeddings. Yet another direction of research would be to derive word embedding for units large than a single word, including collocations and compounds in more one language.

In summary, we have introduced a method for learning bilingual word embeddings that supports second language (L2) learners in finding recurring phrases and example sentences. The method involves converting a given parallel corpus into mixed-code data, generating word embeddings from mixed-code data, and expanding queries in the target language based on bilingual word embeddings. We have implemented the method on an underlying linguistic search engine, *Linggle*. Through the evaluation, we have shown that the method performs reasonably well and is useful for L2 learners.

References

Joanne Boisson, Ting-Hui Kao, Jian-Cheng Wu, Tzu-Hsi Yen, and Jason S Chang. 2013. Linggle: a web-

scale linguistic search engine for words in context. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 139–144.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. 2017. Multilingual training of crosslingual word embeddings. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, volume 1, pages 894–904.

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

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

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

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Ivan Vulić and Marie-Francine Moens. 2015. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 363–372. ACM.