# Final Project: Crosslingual Word Embeddings Replication Study

**Roseanna Hopper, Mona Iwamoto, Maya Miller-Vedam**
w266: Natural Language Processing, Fall 2017
UC Berkeley School of Information, MIDS
{rhopper, miwamoto, mmillervedam}@berkeley.edu
https://github.com/r-hopper/W266-Fall-2017-Final-Project

## Abstract

Crosslingual word embeddings create a representation of both a source and target language in the same vector space. While useful in machine translation, such embeddings have surprising utility for monolingual tasks in both the source and target languages. This raises interesting questions about modeling semantic relationships between languages and transfer learning for improving multilingual systems' performance on tasks involving resource-poor target languages. We explore a specific strategy for learning crosslingual embeddings by replicating the work of a recent paper by Duong et al., through a study of its pseudocode. We describe our replication of key components of Duong et al.'s embedding strategy, and compare our results to baselines.

## 1 Introduction

If you ask a natively bilingual human what is hardest about switching between two languages there's a good chance they'll mention a certain pause: the pause that happens when we know what we want to say but can't find a word to fit. Usually, this pause is spent considering options in the target language that almost but don't quite fit the desired word in the source language. Sometimes the speaker may land on the perfect translation but often they'll settle on a word that is "close but not quite."

This is the bilingual lexicon induction task. The human has learned each language not by studying large lists of parallel phrases but by engaging in large amounts of monolingual conversation in each language separately, while also inferring similarities in meaning based on how and when words from each language are used in relationship to each other. The notion of "closeness" that informs the speaker's chosen translation isn't a one-to-one or one-to-few mapping of words and translations, but a multidimensional relationship that captures things like syntax and connotation as well as literal meaning. Assessing this distance requires not only a mental image of relationships between words in different languages, but also relationships between words in their respective languages.

While humans do this effortlessly, until recently it has been hard for machines to learn representations of words which are flexible enough to be used in different multilingual NLP tasks. Instead, high performance was achieved by training task specific word representations by backpropagating errors from that task, or by adding language specific features to multilingual systems to account for the peculiarities of specific languages.

With increasing research in the field of neural machine translation, crosslingual word embeddings are not state-of-the-art as a basis for a salable machine translation system (Wu et al., 2016). However, crosslingual word embeddings trained using a combination of unsupervised methods on monolingual corpora represent an interesting linguistic and theoretical task in understanding crosslingual induction.

### 1.1 Original Author's Approach

In a 2016 paper titled "Learning Cross-lingual Word Embeddings Without Bilingual Corpora", Duong et al. propose a method to extend the contextual bag-of-words (CBOW) model to accurately learn crosslingual embeddings and handle polysemy (Duong et al., 2016). Training of word embeddings is based most frequently on large parallel text corpora, where words and sentences are often paired by auto-alignment. Duong et al. depart from this, using cleaned and tokenized Wikipedia entries from Polyglot (Al-Rfou et al., 2014) as the monolingual corpora, and the PanLex bilingual dictionaries (Kamholz et al., 2014) as the source of the bilingual signal. They achieve state of the art performance on a bilingual lexicon task modeled after Vulić and Moens (2015), and competitive performance on monolingual word similarity tasks and bilingual document classification tasks.

## 2 Background

In this section, we will review key papers related to the domain of word embeddings for machine translation as a way to build context for Duong et al.'s approach. Though there is an extremely deep body of related work, we have selected the following papers as Duong et al. use them as key formations or make key digressions from them.

### 2.1 Literature Review

"Cross-lingual Word Clusters for Direct Transfer of Linguistic Structure" (Täckström et al., 2012)

In this paper, the authors apply a model built on a resource-rich language to a resource-poor language, as a stopgap for the lack of task-specific annotated corpora in the latter. Their model enables the transfer of information about structure learned in the source language to perform dependency parsing and named entity recognition in the target language.

Starting from the methodology of delexicalized direct transfer proposed by McDonald et al. (2011), the authors use universal parts of speech to model word relationships in one language. They then utilize unsupervised learning in both languages to cluster similar words within and across languages, finally applying the learned features of the first language to the second. The main issue occurs in this step, as training the clusters requires word-aligned training data. This issue describes the basis of one the key difficulties with low-resource language translation that Duong et al. are keen to resolve.

"Inducing Crosslingual Distributed Representation of Words" (Klementiev et al., 2012)

Klementiev et al. extend the ideas of Täckström et al. (2012) by using multi-lingual parallel data to induce word representations.

Each word is treated as a task in a multi-task learning problem (MTL), with joint induction of representations of aligned words of different languages. In this case, word alignments are based on co-occurrence statistics from parallel data, and words that are likely to be translations of one another based on the bitextual statistics are treated as related tasks. The representations are tested by transferring a classifier trained on one language to another, with the finding that classifiers based on the distributed representations outperform all baseline comparisons.

One significant result was that the induced representations were found particularly beneficial when the amount of training data was small, particularly since resource-poor languages are a significant motivator for Duong et al.'s work (though they do not test on an extremely resource-poor language).

"Bilingual Word Embeddings from Non-Parallel Document-Aligned Data Applied to Bilingual Lexicon Induction" (Vulić and Moens, 2015)

Vulić and Moens's paper is the direct precursor to Duong et al.'s work. In this paper, they propose a model which induces bilingual word embeddings from non-parallel data, without any other translation resources. This model focuses on learning lexicons from document-aligned comparable corpora (subject-aligned Wikipedia articles), relying on TreeTagger for part of speech tagging and lemmatization.

For a pair of aligned documents with a source vocabulary ($V^S$) and target vocabulary ($V^T$), the goal is to learn word embeddings for all words in both $V^S$ and $V^T$ such that the embeddings are "semantically coherent and closely aligned" over languages in a single embedding space. To achieve this, they first merge two such aligned documents to form a single pseudo-bilingual document, and then randomly shuffle the pseudo-bilingual document, to ensure that each word obtains surrounding context words from both languages. A version of a monolingual Word2vec skip-gram model is then trained on the shuffled pseudo-document.

Cosine similarity as a proxy for semantic similarity can be computed both monolingually and bilingually once the model is trained. They evaluate the model on one thousand "ground truth"translation pairs for the relevant languages. The authors compute what they refer to as an ($Acc_1$) score (which Duong et al. and we will refer to below as $rec_n$), which they define as the proportion of source language words from ground-truth translation pairs for which the top-ranked word cross-lingually is the correct translation.

Vulić and Moens's evaluation task is adapted in Duong et al.'s paper, which also uses Vulić and Moens's results as a baseline. Duong et al. additionally rely on the training parameters specified by Vulić and Moens's (e.g. window size, learning rate, etc.). Though not part of the originally proposed methology, Duong et al. test a lemmatized version of their model as well, specifically for comparison to Vulić and Moens.

### 2.2 Key Explorations

In replicating Duong et al.'s work our goal is to better understand the role of word embeddings in multilingual NLP. As part of this process we particularly seek to understand:

1. How does Duong et al.'s method of training crosslingual word embeddings differ from earlier approaches to learning word representations in multiple languages?

   Duong et al.'s method differs from earlier approaches in both the mathemetical construction of the embedding strategy and the source of the bilingual signal. Duong et al. claim that for low-resource languages, bilingual dictionaries are necessarily more available than large parallel bilingual corpora, and thus adopt the strategy of relying on a bilingual dictionary as the sole source of bilingual signal. Difficulties in automatic alignment between parallel bilingual corpora also illustrate problems with relying on parallel bilingual texts, but Duong et al.'s dictionary-only methodology sidesteps this issue.

   Another peculiarity of Duong et al.'s method is the utilization of both learned embedding spaces $V$ and $U$, which is related to the joint training of the source and target languages; they choose to combine or interpolate these spaces. Although the aforementioned joint training - as opposed to cascade-style training - is not unique to Duong et al., this choice also represents a split in methodologies within this research space.

2. To what extent is there a linguistic theory for why these embeddings are effective at a variety of tasks?

   Duong et al. posit that embedding space $V$ best represents monolingual similarity while space $U$ represents bilingual similarity, so utilization of both embeddings (by linear combination or interpolation) produces a better crosslingual representation than simply using $V$, as was most common at the time of publication.

   Though Levy and Goldberg address skip-gram as opposed to CBOW in their 2014 paper "Neural Word Embedding as Implicit Matrix Factorization", they assert that for skip-gram with noise-contrastive estimation (NCE), the objective is factorizing a word-context matrix $V \times U^T = M$ in which each row corresponds to a word, each column corresponds to a context, and each cell contains a quantity $f(word, context)$ reflecting the strength of association between that particular word-context pair, finally detailing a proof that $f(w, c) = PMI(w, c)$.

   This result seems almost self-evident upon reflection, as it follows linguistically that word-context PMI would be an effective measure of association in both monolingual and bilingual contexts, as joint probabilities (and therefore information) are symmetric ($P(x, y) = P(y, x)$, $I(x, y) = I(y, x)$), (Church and Hanks, 1990). Though Church and Hanks make this point to propose that PMI does not capture the importance of word order, this point is salient with respect to making a symmetric translation between a source and target word pair.

## 3 Methods

Alongside our study of the theory behind their work, we have reconstructed several key elements of the methodology in Duong et al. (2016). While we originally proposed to translate the authors' model from C, after instructor feedback and running into some roadblocks with the C code we pivoted and instead recreate their model in Python using their pseudocode as our guide. While we fall short of replicating their reported accuracy (due to train time constraints) we did develop three models implementing their joint training strategy, including their main innovation: the use of context embeddings to explicitly handle polysemy. We successfully train and compare embeddings from these three models on all three of the language pairs studied in the original paper (English-Italian, English-Dutch, English-Spanish) plus one additional language pair: English-Japanese. Each of the subsection headings below link to Jupyter Notebooks in our GitHub repository which further explain and demonstrate our work.

### 3.1 Parsing and Preprocessing the Corpora

The corpora provided by Polyglot (Al-Rfou et al., 2014) required some preprocessing. Each token in the corpus received a prefix identifying its source language, and we then made a random draw of sentences in each monolingual corpus (Duong et al. select five million sentences, as did we). To create a bilingual corpus, two monolingual corpora are shuffled together with sentences arranged in random order to avoid learning the two language's weights in sequence and to ensure that topic-specific words and contexts were not overrepresented in the training data.

### 3.2 Modification of the Word2vec Model

We began with the basic Word2vec code provided by TensorFlow, modifying it to suit our data. This included tokenizing and writing a batch iterator, and developing methods for working with a bilingual vocabulary. From here, we implemented a continuous bag-of-words (CBOW) model and a corresponding batch generation method. This model was tested against skip-gram. Our TensorFlow graph was then modified to use CBOW and was augmented with a comparison of full softmax, noise-contrastive estimation (NCE) and sampled softmax. Duong et al. employed NCE training loss, but we found that sampled softmax produced the best results in our tests. To create the crosslingual embeddings, this monolingual CBOW model was further modified to replace the center word in the context with a translation from a bilingual dictionary and to perform joint optimization on both the source and target word. We did not implement the regularization term that Duong et al. further add to the objective function.

### 3.3 Dictionary Replacement

For the bilingual signal, we used preprocessed dictionaries from Duong, et al., which were based on the PanLex data (Kamholz et al., 2014). The PanLex dictionaries are known to be very comprehensive, yet very noisy. Therefore, a single word can have many translations; for example, we discovered that the English word “break” has 179 possible translations in Spanish. We implemented three different strategies for making this center-word substitution with the goal of comparing each method's effectiveness on the bilingual induction task. These three strategies are:

- Model 1: A random translation was selected from the candidate translations.
- Model 2: The highest ranked translation word was selected (where rank is based on vocabulary trained from a monolingual corpus).
- Model 3: The final strategy selected the word based on its cosine similarity with the context embedding. This model represents our closest replication of Duong et al's strategy.

In implementing these three strategies we hypothesized that Model 3 would out perform the others and that the performance difference between Model 1 and 2 would depend on the training corpora and the noisiness of the bilingual signal.

## 4 Results and Discussion

To the extent possible, in all our models we used the training parameters specified by Duong et al. (which were in turn taken from Vulić and Moens (2015)), with a few exceptions[1]. Our biggest departure, and the likely reason we failed to replicate their results is in the training time. Duong et al. state that they run for "15 epochs"however their released example corpus (English-Italian) is much smaller than their reported 5million sentences so we are a little unclear which length ’epochs’ they mean. They also do not specify the size of their vocabulary which has a direct and dramatic effect on train time (especially with the cosine similarity translation selection). We used a 20K word vocabulary. Because we were not training on a GPU configuration we were unfortunately limited in the number of batches we could train each model. (Our final embeddings were learned on 5K

[1] Learning rate=0.15, sampled softmax n=64, embedding size d=200, window size cs=8. Our window size represents a significant departure from theirs (48) which was a conscious choice due to the nature of our "pseudo-bilingual"corpus. The window size used by Duong et al. (48) appears to use an entire sentence as context in most cases. Which this paradigm makes sense, the result in our batch generator was the inclusion of a large number of start and end tokens which the window was too large, which in turn resulted in poor performance due to sublimation of the true context signal. We have outstanding questions about whether Duong et al. really used their reported window size or if their data were preprocessed differently than ours.

batches for cosine selection and to 100K batches for random and probabilistic selection). We realize this is far too short (much less than 1 epoch) to see a model's full capacity. We also found that the learning rate specified by Duong et al. (0.025) did not produce good performance, (again related to our too-short training) so we used a higher alpha, another point of divergence.

Evaluating the caliber of cross-lingual word embeddings presents a challenge: on the one hand we now know that this is not a state of the art component in machine translation; on the other hand these embeddings retain interest for their flexible ability to capture word meaning for a variety of tasks. To evaluate our work we chose to focus on just one of the three downstream tasks discussed by Duong et al.

### 4.1 Bilingual Lexicon Induction Task

As proposed by Vulić and Moens (2015), the bilingual lexicon induction (BLI) task is to predict the translation of a single source-language word in a target language. Though Vulić and Moens and Duong et al. evaluate their methods on recall at one, "where each term has only a single gold translation" (Duong et al., 2016), we utilize a less strict evaluation, in which any of the top ten nearest neighbors that match the ground-truth "gold" translation in the target language may be counted as a successful translation. To make this comparison, we utilize the ground-truth translation dictionaries released by Conneau et al. (2017). Each dictionary provides translations for a single source-target language pair, with each unique word in the dictionary having one or more "accurate" translations, depending on appropriateness. (For example, for English to Spanish, multiple correct translations are provided for "the".) These dictionaries were also processed by prepending the source language prefix, for precise comparison with the training data.

In our work, we tested English/Italian, English/Spanish, and English/Dutch to replicate the work of Duong et al and English/Japanese by way of extension. Unfortunately as discussed above our results, visible in Appendix Table 1, are fundamentally not comparable to the results of the baseline models and Duong et al.

### 4.2 Qualitative Word Similarity

In Appendix Table 2, we compare the top ten closest crosslingual words in both source and target languages for Spanish/English and Italian/English. Duong et al. make this comparison for the Spanish word "gravedad" and Italian word "tassazione". As these words do not appear in our vocabulary, we created the comparison for the more common words "primero" (Spanish) and "suo" (Italian). While the top English predictions for "suo" (bold) are correct, the other top words, and all top words for "primero" have no qualitative appearance of relation. Our interpretation is that this is an additional indication that although our methodological replication of the embeddings was successful, a significant extension in training time and vocabulary size is needed to improve the embeddings.

## 5 Conclusions

Duong et al. propose a crosslingual word embedding method that relaxes the resource requirements, relying on monolingual corpora and a high-coverage but noisy dictionary. We replicate much of their methodology through a study of their pseudocode, and compare our results to the baseline results presented in their paper. Although for common words, our qualitative interpretation of the predictions is promising, significantly more training is needed to properly replicate their results.

Our methodological representation also does not include every component of their approach - particularly the interpolation between embedding spaces $V$ and $U$, which would be a concrete extension of this project. Additionally, we performed training on a Japanese corpus in order to test Duong et al.'s results on a non-Western European language, but found that due to string comparison issues with the Japanese characters, we were unable to complete the BLI task with the available ground-truth translation dictionaries.

Despite extensive discussion of the benefits of this method for low-resource languages, and although Duong et al. restrict the tested vocabulary sizes, the model is not tested on extremely resource-poor languages (though there is some discussion with respect to the term (Duong, 2017)). Therefore, a straightforward extension in testing would be to train embeddings for truly low-resource languages, particularly for evaluation on the BLI task. Revising our methodology to be friendly to non-Western characters would also be helpful for incorporating our trained embeddings in Japanese, as well as other languages.

As the speed is a key consideration for real-time machine translation tools, we would also want to make an extension to test how these embeddings affect the total speed of translation. In particular we would be interested in using our embeddings in the context of neural machine translation to see whether they might support improvement in train time or performance.

## References

[Al-Rfou et al. 2013] Rami Al-Rfou, Bryan Perozzi, Steven Skiena. 2013. *Polyglot: Distributed Word Representations for Multilingual NLP* 183-192. Proceedings of the Seventeenth Conference on Computational Natural Language Learning 2013.

[Church and Hanks 1990] Kenneth Ward Church and Patrick Hanks. 1990. *Word association norms, mutual information, and lexicography* 22-29. Journal of Computational Linguistics archive Volume 16 Issue 1, March 1990.

[Conneau et al. 2017] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. *Word Translation Without Parallel Data*. arXiv preprint arXiv:1710.04087.

[Duong 2017] Long Duong. 2017. *Natural language processing for resource-poor languages*. The Department of Computing and Information Systems, University of Melbourne.

[Duong et al. 2016] Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. 2016. *Learning Crosslingual Word Embeddings without Bilingual Corpora* 1285-1295. 10.18653/v1/D16-1136.

[Kamholtz et al. 2014] David Kamholz, Jonathan Pool, and Susan M. Colowick. 2014. *PanLex: Building a Resource for Panlingual Lexical Translation* 3145-3150. Proceedings of The International Conference on Language Resources and Evaluation 2014.

[Klementiev et al. 2012] Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. *Inducing Cross-lingual Distributed Representations of Words* 1459-1474. Proceedings of COLING 2012: Technical Papers.

[Levy and Goldberg 2014] Omar Levy and Noav Goldberg. 2014. *Neural Word Embedding as Implicit Matrix Factorization* 2177-2185. NIPS'14 Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2.

[McDonald et al. 2011] Ryan McDonald, Slav Petrov, and Keith Hall. 2011. *Multi-source transfer of delexicalized dependency parsers*. Proceedings of EMNLP.

[Mikolov et al. 2013] Tomas Mikolov, Quoc V. Le, Ilya Sutskever. 2013. *Exploiting Similarities among Languages for Machine Translation.* CoRR abs/1309.4168.

[Täckström et al. 2012] Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. 2013. *Cross-lingual Word Clusters for Direct Transfer of Linguistic Structure* 477-487. 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

[Vulić and Moens 2015] Ivan Vulić and Marie-Francine Moens. 2015. *Bilingual Word Embeddings from Non-Parallel Document-Aligned Data Applied to Bilingual Lexicon Induction* 719-725. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL 2015), Beijing, China, 2015.

[Wu et al. 2016] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, and Mohammad Norouzi. 2016. *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation.* Technical Report, 2016.

## 6 Appendix

Table 1: Bilingual Lexicon Induction

| Model | es-en | es-en | it-en | it-en | nl-en | nl-en | Average | Average |
|---|---|---|---|---|---|---|---|---|
| | $rec_1$ | $rec_5$ | $rec_1$ | $rec_5$ | $rec_1$ | $rec_5$ | $rec_1$ | $rec_5$ |
| Gouws and Søgaard (2015) + Panlex | 37.6 | 63.6 | 26.6 | 56.3 | 49.8 | 76.0 | 38.0 | 65.3 |
| Gouws and Søgaard (2015) + Wikt | 61.6 | 78.9 | 62.6 | 81.1 | 65.6 | 79.7 | 63.3 | 79.9 |
| BilBOWA: Gouws et al. (2015) | 51.6 | - | 55.7 | - | 57.5 | - | 54.9 | - |
| Vulić and Moens (2015) | 68.9 | - | 68.3 | - | 39.2 | - | 58.8 | - |
| Duong et al. (2016) (EM selection) | 67.3 | 79.5 | 66.8 | 82.3 | 64.7 | 82.4 | 66.3 | 81.4 |
| Duong et al. (2016) (EM + lemmatization) | 71.8 | 85.0 | 79.6 | 90.4 | 77.1 | 90.6 | 76.2 | 88.7 |
| | es-en | it-en | nl-en | Average | | | | |
| **Our Models** | $rec_{10}$ | $rec_{10}$ | $rec_{10}$ | $rec_{10}$ | | | | |
| **Random center selection** | 0.0025 | 0.011 | 0.0018 | 0.0051 | | | | |
| **Probabilistic center selection** | 0.003 | 0.0011 | 0.0024 | 0.0065 | | | | |
| **Cosine center selection** | 0.003 | 0.0011 | 0.0024 | 0.0065 | | | | |

Table 2: Crosslingual Word Similarity

| es (primero) | es (primero) | it (suo) | it (suo) |
|---|---|---|---|
| *es* | *en* | *it* | *en* |
| primero | afl | suo | **his** |
| plantel | dating | un | her |
| atl | macedonian | proprio | marketing |
| japoneses | highest | quale | entertaining |
| autoridades | dense | questo | medalist |
| fall | freud | loro | landscape |
| ampliaci | blockade | rito | violation |
| directora | pits | momento | continue |
| mateo | article | lo | ruined |
| andina | hydrogen | rara | inform |