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 obviously useful in Machine Translation, such embeddings also prove surprisingly adept for monolingual tasks in both source and target language. 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. In this paper we explore a specific strategy for learning crosslingual embeddings by replicating the work of a recent paper and extending its application to additional languages of interest. Our goal in doing so is not to generate an improvement on the author's results but to grow our own understanding of 1) how crosslingual word embeddings are trained 2) on what downstream tasks are they most effective and 3) how do they differ in concept, application and performance from other approaches.

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 own language.

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[1]. 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 [2]. Crosslingual word embeddings trained using a combination of unsupervised methods on monolingual corpora represent a promising alternative.

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 CBOW model to accurately learn crosslingual embeddings and handle polysemy (**?**; **?**). They train on two data sources: the PanLex bilingual dictionary (**?**) (bilingual signal) and monolingual texts from Polyglot (**?**) (cleaned and tokenized Wikipedia entries, basis of the word vectors). The authors measure the effectiveness of their embeddings in a variety of applications including both bilingual

[1] This is implied by Duong et.al but we need a real citation pending our ongoing literature review

[2] ditto re: citation, also I am not sure I fully understand the space of 'prior approaches' so I may be misrepresenting this issue.

and monolingual tasks. They achieve state of the art performance on a bilingual lexicon task modeled after Vulic and Moens (2015) and competitive performance on monolingual word similarity tasks and bilingual document classification tasks.

1.2 Our Proposed Work

For our project, we propose to replicate the methodology in this paper and apply it to a language not covered by the author's original work.

Our first step will be to convert the paper's pseudo-code into working Python code. The authors' C code, available on GitHub (*https://github.com/longdt219/XlingualEmb*) will provide a useful reference, as well as numerous other CBOW and Word2Vec tutorials. Because Duong et al. compare their model against several other field benchmarks, we will also be exploring the replication of those models in Python in order to adequately test the performance of the model's extension. As part of the replication task, we will apply our code to the original data from PanLex and Polyglot to see if we can produce the author's results on the bilingual lexicon induction, cross-lingual document classification, and mono-lingual word similarity tasks.

Finally, we will extend the analyses to include French and Japanese (also available in PanLex and Polyglot), and compare our results with the languages covered in the original paper (Spanish, Italian, Dutch and English).

1.3 Questions we hope to answer

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 seek to understand:

- How does Duong et al.'s method of training crosslingual word embeddings differ from earlier approaches to learning word representations in multiple languages?
- How to these differences in training method enable specific improvements on downstream tasks (eg. handling polysemy in translation tasks)?
- To what extent are these word embeddings competative on other monolingual and bilingual tasks?
- To what extent is there a linguistic theory for why these embeddings are effective at a variety of tasks?
- At the time of writing the authors (and some of their colleagues) cite the constraint of limited access to large volumes of parallel texts on which to train as a motivating factor for the development of training methods that can take their bilingual signal from a dictionary or small parallel text. Given advances in access to digital content and the comparative effectiveness of word embeddings trained on parallel corpora is it likely that this strategy will be the dominant one in the future or is this a niche strategy for use primarily in the case of resource-poor target languages?

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.

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 away to get around the lack of a task specific annotated corpus in the latter.
- This 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 for delexicalized direct transfer proposed by (McDonald et al., 2011) the authors use universal POS to model word relationships in one languge. Then using unsupervised learning in both languages to cluster similar words within and across langauges. Finally they apply the structure learned in the first language to the second using the clusterings to form the transition matrix[3]. Note that the second step

[3]Not sure I'm calling this the right thing here. Basically they develop a set of classes using clustering based on bilingual text. Then use these classes as features to transfer information about universal POS from the first language to the second

– training the clusterings, requires word aligned training data.

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

- This work extends the ideas of (Täckström et al., 2012) by using multi-lingual parallel data to induce representations.
- Each word is treated as a task in a multitask learning problem (MTL) and jointly induce representations of aligned words of different languages.
- The alignments are based on co-occurrence statistics from parallel data. Words that are likely to be translations of one another based on the bitext statistics are treated as related tasks.
- The shared representation is then fed into a neural network similar approach used by (Bengio et al. 2003).
- They test the representations by transferring a classifier trained on one language to another and found that classifiers based on the distributed representations outperformed all baselines.
- The induced representations were found particularly beneficial when the amount of training data was small. Thus, effectively taking advantage of plentiful unsupervised data used for inducing crosslingual word representations.

"Exploiting Similarities among Languages for Machine Translation"(Mikolov et al., 2013)

- The paper proposes a technique for automating the process of generating dictionaries and phrase tables for translation.
- Mikolov et al.'s method builds a monolingual model for a particular language based on a large body of text, and then creates a linear projection between languages using a ßmall"bilingual dictionary.
- Mikolov et al. use the CBOW model (combining the representations of surrounding words to predict the word in the middle), as CBOW can be trained on a large corpus in a short time due to low computational complexity.
- This method relies on linear transformation between languages, in which $x_i \in \mathbb{R}^{d_i}$ is the vector representation of word i, and we have a word pair $\{x_i, z_i\}_{\mathrm{i=1}}^{\mathrm{n}}$. Here, we rely on finding a transformation matrix W such that Wx_i approximates z_i.
- They conclude by comparing model performance to two baselines: (1) similarity of the morphological structure of words (e.g. edit distance), and (2) similarity of word co-occurrences.
- They find that although this model can translate words with high frequency ranks with good precision, but that incorporating edit distance into the translation guess improves precision even further.

3 Methods

As we continue our Literature Review process we expect to be able to explain Duong et al.'s method for learning the crosslingual word embeddings and contextualize them in relationship to earlier and subsequent methods.

4 Results and Discussion

Although our project is not oriented around a single measurement of word embedding effectiveness, we expect to be able to report results on the bilingual lexicon task and word similarity tasks.

5 Next Steps

Our immediate next step is continuing the literature review to build an understanding of the contextual questions we posed in our abstract.

We are currently part way through the translation from C and expect to have more information on the downstream tasks soon.

We have access to some of the baseline word embedding models in the original paper but as they are implmented in C we are concerned that it is not realistic for us to do that much translation. We will instead look for out of the box baselines for the downstream tasks with which we can compare.

References

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.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. *Inducing Crosslingual Distributed Representations of Words* 1459-1474. Proceedings of COLING 2012: Technical Papers.

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

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.