Final Project: Crosslingual Word Embeddings

Roseanna Hopper, Mona Iwamoto, Maya Miller-Vedam
w266: Natural Language Processing, Fall 2017
UC Berkeley School of Information, MIDS
{rhopper, miwamoto, mmillervedam}@berkeley.edu

Abstract

Add abstract here. Why are crosslingual word embeddings important? What we propose to do. (I am confused about whether to write this as if we are doing 'novel' work vs. we are trying to teach ourself something that others have done.

1 Introduction

Two exciting open problems in NLP are the challenge of modeling text that switches between multiple languages and the challenge of transferring learning between languages. For example, how might we take advantage of a wealth of training data in one language to perform modeling in another? Strategies based on dictionary translation commonly fail to represent the full nuance of word use but large bilingual corpora are rare. Crosslingual word embeddings offer a potential solution by representing lexical items from different languages in the same vector space.

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 use an Expectation-Maximization style algorithm and 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 and monolingual language modeling and 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 authors original work. Our first step will be to convert the papers pseudocode into working Python code. The authors C code, available on GitHub [4] will provide a useful reference as well as numerous other CBOW and Word2Vec tutorials. Next we will apply our code to the original data from PanLex and Polyglot to see if we can replicate the authors results on the bilingual lexicon induction, cross-lingual document classification and mono-lingual word similarity tasks. Finally we will repeat these analyses for 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).

2 Background

The course instructions say to use this space for a literature Review, Related Work, etc. I would say that means we need an explanation of big problems in Multilingual NLP and why word embeddings are an important part of this.

3 Methods

I am unclear what this would mean for us. See my notes on Slack.

4 Results and Discussion

I am unclear what this would mean for us. See my notes on Slack.

5 Next Steps

I am unclear what this would mean for us. See my notes on Slack.

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.