# **Deep Neural Models for Bilingual Lexicon Extraction**

#### Jon Gauthier

Symbolic Systems Program Stanford University Stanford, CA 94305

jgauthie@stanford.edu

#### **Arthur Tsang**

Department of Computer Science Stanford University Stanford, CA 94305

atsang2@stanford.edu

#### **Abstract**

We present a new method for producing bilingual lexicons from very non-parallel corpora. In contrast with previous approaches, our method separates the steps of (1) constructing distributed word representations from monolingual corpora, and (2) building a mapping between these representations.

We demonstrate that a deep neural model can effectively learn a translational mapping between two monolingual vector spaces for the purposes of bilingual lexicon extraction. We evaluate the performance of this model relative to X and Y.

### 1 Introduction

In the task of bilingual lexicon extraction (BLE), we accept as input some pair of corpora in different languages—perhaps parallel or perhaps not—and learn associations of translational equivalence between one language (a "source language") and the other (a "target language").

Bilingual lexicon extraction is obviously of most utility when applied to under-resourced language pairs which do not already benefit from a surplus of lexicon materials. These same language pairs, lacking resources as simple as translation dictionaries, also lack sufficient aligned text for training BLE models. For this reason, most recent work on this task focuses on minimally supervised approaches which do not require parallel corpora (Rapp, 1995; Peirsman and Padó, 2010).

Rapp (1995) suggested that there is "a correlation between the patterns of word co-occurrences in different languages." In line with this hypothesis, recent work has determined the likelihood of two words as translations of one another by comparing the translations of the contexts in which

they appear. The standard approach is to manually\*\*UNCLEAR\*\*construct a "bilingual vector space" which contains distributed word representations of words in both the source and target language (Fung and Yee, 1998; Peirsman and Padó, 2010; Vulić and Moens, 2013). The axes of this space are built from bilingual seed pairs. A given word representation has a large value along an axis if it co-occurs often with the word corresponding to the axis. Because axes are defined with bilingual pairs, word representations for both source and target language words can lie in the same space. With a bilingual vector space constructed, translations for a given source-language word can be retrieved by simply finding the nearest target-language word representation neighboring the word's own vector representation in the bilingual vector space.

The present paper is motivated by the hypothesis that the standard bilingual vector space approach to BLE is restrained by the nature of the space itself. We suggest that the construction of the space along axes composed of a hand-picked set of bilingual seeds limits the expressiveness of the word vectors contained therein.

We present a novel method for bilingual lexicon extraction centered around two monolingual vector spaces, constructed via a standard neural language model which minimizes absolute costs of word representation\*\*UNCLEAR\*\*(?). Our "translation function" maps between the two monolingual vector spaces using a deep neural network, trained on a small set of seed translations.

Section 2 describes the corpora used to test this hypothesis. Section 3 details the actual process of translation. In Sections 4 and 5 we analyze the performance of our model and examine how it supports our motivating hypothesis.

#### 2 Data

We train our models on English and Spanish Wikipedia.

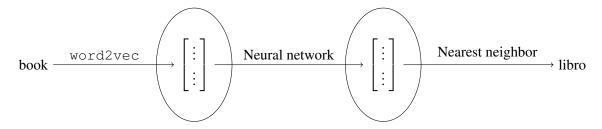


Figure 1: The process of word translation, using a neural network to map between two monolingual VSMs

## 3 Model

# 3.1 Word representations

#### 3.2 Translation with a neural network

The computation central to the process of translation is the transformation of the source-language word representation into a target-language word representation via a neural network. This neural network is trained on a minimal set of broaddomain translation "seeds," each a pair relating a single source language word to a single translationally equivalent target language word.

In line with previous approaches we fill this seed set with cognates shared between the two languages (?). We choose to source seeds in this way simply because it is less labor-intensive than the alternative manual collection of seeds. The model could train equally well on non-cognate pairs, however — an approach which would perhaps be necessary in the case where the source and target languages do not share many words.

For each seed pair of words (e, f), we provide as a training example to the neural network the input and output  $(w_e, w_f)$ , where  $w_e$  and  $w_f$  are word representations generated as described in Section 3.1.

## 4 Evaluation

- 4.1 Baseline
- 4.2 Results
- 4.3 Error analysis
- 5 Conclusion

# Acknowledgments

The acknowledgments should go immediately before the references. Do not number the acknowledgments section. Do not include this section when submitting your paper for review.

#### References

Pascale Fung and Lo Yuen Yee. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 17th International Conference on Computational linguistics-Volume 1*, pages 414–420. Association for Computational Linguistics.

Yves Peirsman and Sebastian Padó. 2010. Crosslingual induction of selectional preferences with bilingual vector spaces. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 921–929. Association for Computational Linguistics.

Reinhard Rapp. 1995. Identifying word translations in non-parallel texts. In *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics*, pages 320–322. Association for Computational Linguistics.

Ivan Vulić and Marie-Francine Moens. 2013. A study on bootstrapping bilingual vector spaces from non-parallel data (and nothing else). In *Proceedings of EMNLP 2013: Conference on Empirical Methods in Natural Language Processing*.