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 **[1]**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).

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. Cross-lingual 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*.

[1] **Jason**: Test