Learning a non-linear mapping for low-resource bilingual lexicon extraction

No Author Given

No Institute Given

Abstract. Finding translation pairs can be done by means of learning a mapping between word embeddings of different languages. Popular methods learn a linear transformation that maps words from a source language into a target one. This linear mapping minimizes the distance between word embeddings in two different languages. In this work we explore a non-linear mapping of word embeddings using a combination of denoising autoencoders and sparse autoencoders. We apply this neural architecture to extract translation pairs in a low-resource language pair setting: Spanish-Nahuatl

1 Introduction

In natural language processing (NLP), bilingual lexicon extraction is the task of obtaining a list of word pairs deemed to be word-level translations [1]. These translations pairs can be automatically extracted from parallel corpora, comparable corpora or even monolingual corpora of two languages.

There are plenty of methods for extracting, or inducing, bilingual lexicon. However, this task remains an active area of research since methods usually fail when dealing with small sized corpora or with languages that are very distant from each other [2].

Recently, there has been an increasing interest in methods that use word vector representations. These vectors usually encode the meaning of a word by using its contexts [3,4]. The general idea is that a word that occurs in a given context in a language should have a translation that occurs in a similar context in the other language. If we have vectors representing lexical units of two different languages, we can project words from a source language into a target one and compute distances in order to find translation candidates [5].

In this work, we propose a way to improve bilingual lexicon extraction for a low-resource language pair. On one hand, we work with bilingual vector representations that are suitable for languages with small parallel corpora available. On the other hand, instead of learning a linear map, we propose to learn a non-linear transformation that may find better translation pairs. This linear transformation is implemented by using a neural architecture, i.e., a denoising autoencoder.

We focus on the language pair Spanish-Nahuatl, these two languages are spoken in Mexico and previous NLP work has been conducted related to this language pair. Nahuatl is a low-resource language, it does not have a big amount of digital sources nor many language technologies.

Bilingual lexicons that are automatically extracted from corpora are useful resources, specially for languages that do not have many digital resources. Human crafted bilingual dictionaries are expensive resources that are not available for all language pairs in the world. Moreover, since bilingual lexicon extraction is closely related to machine translation, our extracted Spanish-Nahuatl bilingual lexicon could be useful for improving machine translation systems for this language pair.

For our experimental setting we focus in the low-resource language pair Spanish-Nahuatl. This type of scenarios are challenging since it is not easy to find bilingual correspondences.

We decide to choose this language pair because there is one available digital parallel corpus and previous work related to bilingual lexicon extraction has been done.

As a word vectors we use bilingual vector representations proposed by [6] for Spanish and Nahuatl, since they outperformed word2vec representations.

The main goal of this work is to propose a non-linear transformation between the vector spaces of two languages using a particular neural architecture called autoencoders [7,8,9,10], specifically denoising autoencoders and sparse autoencoders.

An autoencoder (AE) is a neural network that can be used to learn eficient data codings or to perform dimensionality reduction. This architectures are usually consider unsupervised.

For the bilingual lexicon extraction, we propose a neural architecture, specifically an autoencoder that replaces the linear transformation. We construct our bilingual architecture from two types of autoencoders: denoising and sparse autoencoders as describe above.

From the denoising autoencoder we integrate the corruption process $C(\hat{x}|x)$, assuming that word vectors in Spanish are a corrupted version of word vectors in Nahuatl. On the other hand, from a sparse autoencoder we integrate the notion of the hidden layer being of larger dimension than the input and output layers leaving aside the sparsity penalty.

Our proposed architecture has an input layer of 128 neurons, this 128 neurons corresponds to the dimensions of the Node2Vec word vectors, a hidden layer that consists of m neurons, where $m > 128$ for all experiments, and an output layer that consists also of 128 neurons that represent the reconstructed Node2Vec word vectors but in Nahuatl

This bilingual architecture must reconstruct the word vectors in Spanish by removing the existing noise, in such a way that the reconstructed vector at the output layer corresponds to its translation in Nahuatl (or it is close to it).

The major benefits of neural networks is that they are able to deal with non-linear problems. Our architecture combined with normalized Node2Vec vectors is able to outperform the traditional linear transformation for this task. Expanding the dimension of the hidden layer of an autoencoder, helps to improve the model performance since it is able to identify unique statistical features in the training set.We can also see that the normalization step helps our model to outperform the linear mapping even though these normalization does not helps the linear mapping to achieve a higher accuracy.

Learning non-linear transformations to find a mapping between languages is not very common in NLP, therefore, there is still many directions to explore. This is just a preliminary work. Moreover, with this work we would like to contribute to the devel-

opment of NLP technologies, specially for low-resource languages and particularly for Nahuatl since it is one of the most spoken languages in Mexico.

References

1. Haghighi, A., Liang, P., Berg-Kirkpatrick, T., Klein, D.: Learning bilingual lexicons from monolingual corpora. Proceedings of ACL-08: Hlt (2008) 771–779
2. Gutierrez-Vasques, X.: Bilingual lexicon extraction for a distant language pair using a small parallel corpus. In: NAACL-HLT 2015 Student Research Workshop (SRW). (2015) 154
3. Harris, Z.S.: Distributional structure. Word **10** (1954) 146–162
4. Firth, J.R.: A synopsis of linguistic theory, 1930-1955. Studies in linguistic analysis (1957)
5. Mikolov, T., Le, Q.V., Sutskever, I.: Exploiting similarities among languages for machine translation. arXiv preprint arXiv:1309.4168 (2013)
6. Gutierrez-Vasques, X., Mijangos, V.: Low-resource bilingual lexicon extraction using graph based word embeddings. arXiv preprint arXiv:1710.02569 (2017)
7. LeCun, Y.: Modèles connexionnistes de lapprentissage. PhD thesis, PhD thesis, These de Doctorat, Universite Paris 6 (1987)
8. Ballard, D.H.: Modular learning in neural networks. In: AAAI. (1987) 279–284
9. Bourlard, H., Kamp, Y.: Auto-association by multilayer perceptrons and singular value decomposition. Biological cybernetics **59** (1988) 291–294
10. Baldi, P., Hornik, K.: Neural networks and principal component analysis: Learning from examples without local minima. Neural Networks **2** (1989) 53–58