Low-resource bilingual lexicon extraction using graph based word embeddings

Ximena Gutierrez-Vasques and Víctor Mijangos

UNAM, Mexico City.
xim@unam.mx, vmijangosc@ciencias.unam.mx,

1 Introduction

In this work we focus on the task of automatically extracting bilingual lexicon for the language pair Spanish-Nahuatl. This is a low-resource setting where only a small amount of parallel corpus is available. Most of the downstream methods do not work well under low-resources conditions. This is specially true for the approaches that use vectorial representations like Word2Vec. Our proposal is to construct bilingual word vectors from a graph. This graph is generated using translation pairs obtained from an unsupervised word alignment method.

We show that, in a low-resource setting, these type of vectors are successful in representing words in a bilingual semantic space. Moreover, when a linear transformation is applied to translate words from one language to another, our graph based representations considerably outperform the popular setting that uses Word2Vec.

2 Overall method

The overall procedure can be summarized into the following algorithmic steps. The next subsections contain the detailed explanation of each stage.

1. For each Spanish word, a scored list of translation candidates is extracted using a sampling-based method
2. Once the translation scores are obtained, a graph is computed. The nodes correspond to the vocabulary words of each language. The weighted connections between the nodes are obtained from the scored lists from previous step
3. Word vectors are computed from this graph, i.e., using the Node2Vec algorithm each node (a word) is transformed to a continuous vector in $\mathbb{R}^n$ [1].
4. A linear transformation is learned in order to map from Spanish vector space to Nahuatl vector space. A seed lexicon of correct translation pairs is used to learn this transformation

5. Once the transformation is calculated, it is applied to a set of evaluation Spanish words. For each projected vector, the nearest vectors correspond to the translation candidates in Nahuatl (L_2 metric was used).

3 Results

In order to evaluate the representations in a bilingual lexicon extraction task, we constructed an evaluation set by randomly selecting 130 Spanish words with frequency greater than 2 in the corpus. Human annotators wrote possible translations for this set of words. It is important to notice that the seed lexicon used for the linear mapping did not contain any of these words to avoid overfitting.

We used precision at one, precision at five and precision at ten, i.e., take into account up to 10 closest word vectors to a source word that we want to translate. First, we evaluated our graph based representations without using any linear mapping (*N2V-NOmap*). Then, we applied linear mapping to the Node2Vec representations (*N2V-map*) and the Word2Vec ones (*W2V-map*). Results are shown in Table 1.

Table 1. Evaluation of bilingual lexicon extraction using different vector representations

	P@1	**P@5**	**P@10**
N2V-NOmap	0.260	0.598	0.661
N2V-map	**0.614**	**0.835**	**0.866**
W2V-map	0.102	0.125	0.164

The bilingual graph based embeddings, clearly outperform the popular setting of mapping Word2Vec representations between languages [2]. This is consistent with recent works that have pointed out that the latter approach, tends to have a bad performance, specially if they lack of huge amounts of corpora to train the vector space representations [3]. In this sense, our proposal seems to be successful in dealing with low-resource languages. We only needed a small parallel corpus (sentence aligned) and we were able to generate word vector representations that are useful for bilingual lexicon extraction.

With this work, we would like to contribute to the development of automatic translation technologies for Spanish-Nahuatl, since there are practically no technologies developed for indigenous languages of Mexico.

References

1. Grover, A., Leskovec, J.: node2vec: Scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM (2016) 855–864
2. Mikolov, T., Le, Q.V., Sutskever, I.: Exploiting similarities among languages for machine translation. arXiv preprint arXiv:1309.4168 (2013)
3. Levy, O., Søgaard, A., Goldberg, Y., Ramat-Gan, I.: A strong baseline for learning cross-lingual word embeddings from sentence alignments. arXiv preprint arXiv:1608.05426 (2016)