Word translation through shared embedding space

Mridul Garg
A53211083
m1garg@ucsd.edu

Abstract

I make an attempt to generalize the cross-lingual experiment setup given by Lazaridou et al. (2015). Their model is restricted to translate words only from one fixed language to another fixed language. I wish to extend it so that the model can translate words between multiple languages and, possibly, between language pairs which it did not see at the training time, thus making it a zero-shot learning technique.

1 Introduction

Our goal is to have a system which can translate words from any source language to any target language. Zero-shot learning model proposed by Lazaridou et al. (2015) works only for one specific source to target language pair. It is not practical to train a model for each language pair as languages evolve over time and to accommodate those changes, we would have to update the parameters for several models. To avoid these issue, I propose a model architecture which can be trained on a collection of word translations for multiple language pairs so that it can be used to translate words between any of those languages.

I hypothesize that this model would have the same benefits as noted by Johnson et al. (2016). These are (a) this model can easily be extended to train on new language pairs, (b) translation quality for low-resource languages increases and (c) the model learns to translate between language pairs it has never seen on training time, thus making it a *zero-shot learning technique*.

2 Background

I summarize the cross-lingual model used by Lazaridou et al. (2015) here. Their model is trained to learn a linear mapping function which is then applied to a vector representation of a word in language A to obtain a vector representation of its meaning in language B.

Thus, given a vector $\mathbf{x}$, it outputs a vector $\mathbf{y} = \mathbf{W}\mathbf{x}$. $\mathbf{W}$ is learned by the model using the max-margin loss. Thus, the loss for a given training pair $(\mathbf{x}_i, \mathbf{y}_i)$ and the prediction $\hat{\mathbf{y}}_i = \mathbf{W}\mathbf{x}_i$ is defined as

$$\sum_{j \neq i}^{k} \max\{0, \gamma + dist(\hat{\mathbf{y}}_i, \mathbf{y}_i) - dist(\hat{\mathbf{y}}_i, \mathbf{y}_j\}.$$

k and γ are hyperparameters and $dist$ is inverse cosine distance measure. We hereby refer to this model as FLPM (fixed language pair model).

3 Proposed model

I now extend FLPM to a new model, namely, MLPM (multiple language pair model). The input to MLPM is a collection of pairs in the form of $((\mathbf{x}, s), (\mathbf{y}, t))$ where $\mathbf{x}$ is a vector representation of a word in the source language, represented by the token s, and $\mathbf{y}$ is its translation or meaning in the target language t.

We can obtain $\mathbf{y}$ as

$$\mathbf{z} = \mathbf{A}_s\mathbf{x} \quad (1)$$

$$\mathbf{y} = \mathbf{B}_t\mathbf{z} \quad (2)$$

For each source language s, the model will learn a set of parameters $\mathbf{A}_s$ which linearly maps a word from semantic space of a source language s to a representation in a semantic space shared by all languages. This representation is then linearly mapped through $\mathbf{B}_t$ to a word in the semantic space of the target language t.

Thus, for N languages, we have reduced the number of parameter matrices from $O(N^2)$ to $O(N)$. This model should be able to translate from language A to B even if it would have only seen pairs from A to C and from C to B at the

training time. It, however, has to be validated through experiments.

4 Conclusion and Future work

I presented an outline for MLPM for multi-lingual word translation and claimed that it an instance of zero-shot learning. It is an initial idea to extend FLPM and it might change or improve for the second draft for this project.

I also need to identify a dataset and proper experiment setup which would establish or discard the validity of our model.

Acknowledgments

I thank Professor Ndapa Nakashole for valuable inputs during the initial stage of this project.

References

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's multilingual neural machine translation system: Enabling zero-shot translation. *CoRR* abs/1611.04558. http://arxiv.org/abs/1611.04558.

Angeliki Lazaridou, Georgiana Dinu, and Marco Baroni. 2015. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*. pages 270–280. http://aclweb.org/anthology/P/P15/P15-1027.pdf.