CPSC 290 Research Proposal: Multi-lingual Word Embeddings in Natural Language Processing

Jungo Kasai

January 15, 2017

Advisor: Professor Dragomir Radev at the Department of Computer Science
Meetings:
Term: Spring 2017
Topics: Word Embedding, Neural Networks, Distributional Semantics, Machine Translation

1 Background

One of the key lessons from the recent development in Natural Language Processing (NLP) is that vectorization of words, i.e. word embedding, plays an important role in various tasks. First, word embedding has benefited an approach to semantics called distributional semantics. Distributional semantics takes an inspiration from linguistics and philosophy that “you shall know a word by the company it keeps”, believing that the meaning of a word can be represented by the contexts it takes. This means that the primary task of distributional semantics comes down to compressing high dimensional information of raw contexts into lower dimension while keeping semantically sensible structures and discarding noises. Although there are many kinds of technique for such compression, including the singular value decomposition, neural networks have proven powerful. The neural network based word vectorization, Word2Vec, developed by Mikolov at Google appears to preserve a linguistically sensible linear structure [1]. For instance, it turns out that the closest vector to the vector obtained by a simple vector operation $\overrightarrow{king} - \overrightarrow{man} + \overrightarrow{woman}$ is $\overrightarrow{queen}$. Such results encourage us to see word embedding as a good representation of semantics. Moreover, pre-trained word embeddings are often used as an initialization point for many tasks such as Machine Translation and Part-of-Speech Tagging.

Though monolingual word embedding has been contributing to semantics and many other sub-fields in NLP as above, multilingual embedding has yet to be explored extensively. First of all, since monolingual embedding turns out to preserve semantic relations between words, it is encouraging to superpose multiple language word embeddings and extend our notion of semantics to multiple languages. Furthermore, multilingual embedding can potentially enrich our monolingual learning in languages with less data than English as we might be able to transfer a model in English to a model in another language. Motivated by these insights, this research project is intended to systematically explore cross-lingual aspects of word vectorization.

2 Outlines of the Project

2.1 Independent Training

A natural way to cross-lingually extend our notion of word embedding is training distinct models for different languages and superpose them through a linear transformation [2]. Concretely, suppose that we have trained models to obtain word embeddings for English and Spanish. Suppose also that we have pairs of vectors with the same meaning $\{e_i, s_i\}_{i=1}^{n}$ where $e_i \in \mathbf{e}$ and $s_i \in \mathbf{s}$ for $\forall i$ and $\mathbf{e}$ and $\mathbf{s}$ denote the sets of vectors for English and Spanish respectively. Mathematically, the approach boils down to a convex optimization problem:

$$\min_{\mathbf{X}} f_0\left(\mathbf{X}\right) = \sum_{i=1}^{n} \|\mathbf{X}e_i - s_i\|_2^2 + \lambda \|\mathbf{X}\|_F^2 \tag{1}$$

where $\mathbf{X}$ is an n by n translation matrix and λ is a coefficient for the regularization term. This optimization problem can be decomposed into n sub-problems which we can solve by at most n steps by the conjugate gradient method [3][1]. As intuitive and simple this approach is, the best results on paring of the test set are around 50%; we will treat it as a baseline over the course of the project and try to improve upon it.

2.2 Synchronized Training

One might be able to improve upon the aforementioned method by simultaneously training word embeddings in shared vector space. Such synchronized learning would have several advantages:

- We would not need a dictionary of word pairs between languages.
- We would not need to assume that two different vector spaces are related by a linear transformation.
- We would not need to limit ourselves to one-to-one correspondence from word to word, respecting diversity among languages.

The third advantageous point is particularly important. For example, in English, the word, “bank” has multiple meanings unlike some other languages such as Chinese and Japanese. In such cases, a linguistic problem arises if we represent bank in English and another language by the same vector.

BiSkip is a natural extension of Word2Vec to bilingual word vectorization [4]. Word2Vec [2] maximizes the likelihood of the contexts of all words, restricted by the neural network architecture and the conditional independence assumptions. The mathematical formulation of the vanilla Word2Vec is:

$$\max_{\theta} \sum_{i=1}^{n} \sum_{-m \leq k \leq m \ k \neq 0} logP(w_{i+k}|w_i;\theta) \tag{2}$$

where θ denotes the set of parameters in the neural network, and w_i represents the ith word in the corpus. m or $2m+1$ is called the window size in the literature, a formulation of the context of each word.

Now, Luong et al. propose the BiSkip formulation as follows [4]:

$$\max_{\theta} \sum_{i=1}^{n_1} \sum_{-m \leq k \leq m \ k \neq 0} logP(e_{i+k}|e_i;\theta_1) + logP(s_{a(e_i)+k}|e_i;\theta_2) \tag{3}$$

$$+ \sum_{j=1}^{n_2} \sum_{-m \leq k \leq m \ k \neq 0} logP(s_{j+k}|s_j;\theta_3) + logP(e_{a(s_j)+k}|s_j;\theta_4) \tag{4}$$

[1] Note that [2] uses the stochastic gradient descent algorithm instead. The conjugate gradient method terminates by a finite number of steps and usually takes much less computation.

[2] I mean the Skipgram algorithm, as opposed to the CBOW algorithm.

where e denotes an English word and s denotes a Spanish word for concreteness. This can be viewed as training four independent Word2Vec models. We intend to further improve their modification of the vanilla Word2Vec mechanism. I have some questions in mind as of now. First, more parameter sharing or relations would make much more sense. Having two independent Word2Vec modules for both directions of translation might not be ideal. And precisely because of such freedom, it is not obviously true that we are representing two languages in the same space; restricting the hypothesis space for the parameters could help the network to have the same space for the two languages. Second, architectural questions arise. In the BiSkip scheme, we would need to have an external aligner. It would be interesting to integrate a word prediction model across languages and an aligner.

2.3 Baselines and Metrics

Since multi-lingual embedding is an emerging topic, we also need to establish baselines and metrics. We generally follow the metrics established in [5].

- monolingual word similarity
- word translation
- extrinsic tasks (POS tagging, Dependency Parsing, etc)

In addition to these quantitative analyses, we will also look at sociological aspects of word embedding. It turns out that $\overrightarrow{man} - \overrightarrow{woman} \approx \overrightarrow{computer\ programmer} - \overrightarrow{homemaker}$ [6]. While this nature of our word embedding is sociologically intriguing[3], we want to just extract semantics for language processing. Bolukbasi et al. post-process word embedding to eliminate such biases [6]. Parallel word embedding across languages might give us a more systematic resolution for this issue.

3 Future Directions of Research

Should we well establish multilingual word vectorization, its applications would be of interest. As a potential direction of further research on top of this project, we might want to look at how to apply multilingual word vectorization to machine translation, which might improve upon the conventional way to generate translation based on the beam search algorithm [7].

References

[1] Mikolov et al. *Distributed Representations of Words and Phrases and their Compositionality* : `https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf`

[2] Mikolov et al. *Exploiting Similarities among Languages for Machine Translation* : `http://arxiv.org/pdf/1309.4168v1.pdf`

[3] I particularly find it interesting that despite the fact that “female computer scientist” occurs more often than “male computer scientist”, the network figures out that the co-occurrence implies a negative correlation. Recall that the Word2Vec algorithm is solely based off of predicting surrounding words.

[3] Shewchuk *An Introduction to the Conjugate Gradient Method Without the Agonizing Pain*
: `https://www.cs.cmu.edu/~quake-papers/painless-conjugate-gradient.pdf`

[4] Luong et al. *Bilingual Word Representations with Monolingual Quality in Mind*
: `http://www.aclweb.org/anthology/W15-1521`

[5] Ammar et al. *Massively Multilingual Word Embeddings*
: `https://arxiv.org/pdf/1602.01925v2.pdf`

[6] Bolukbasi et al. *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings*
: `https://arxiv.org/pdf/1607.06520v1.pdf`

[7] Johnson et al. *Googles Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation*
: `https://arxiv.org/pdf/1611.04558v1.pdf`