Iran University of Science and Technology
Deep Learning (Spring 2020)

Final report template

Kiamehr Rezaee
Department of Computer Engineering
Iran University of Science
and Technology
k_rezaee@cs.iust.ac.ir

Abstract

In the field of machine-learning, transferring knowledge from one domain to another (known as transfer-learning), especially in problems with less available resources, is a main area of research. Transferring knowledge from English language to another, can be an example. In this work, we observe how we can transfer learned contextualized embeddings between languages, only using a multi-lingual dictionary as a source of supervision.

1 Introduction

In this work, we investigates the problem of transferring knowledge learned in one language to another, by learning a linear transformation from contextual sentence alignments to align the contextualized embeddings, independently trained in different languages. We propose a method for using BabelNet multilingual dictionary, in order ot extract the contextual aligned pairs, rather than using parallel corpora. This method is especially effective in the case of languages with less parallel resources. We further demonstrate the effectiveness of this approach on a zero-shot cross-lingual sentiment analysis task.

2 Related work/Background

Retrofitting monolingual word embeddings of different languages has been used as an off-line approach for static cross-lingual embedding learning in many previous works, among which there has been some successful results. In these off-line methods, a transformation matrix is computed to map aligned word pairs in two monolingual embedding spaces. The the matrix is used to map unseen word embeddings from one language to the other. In contrast to the offline-approach, there are some so called on-line approaches in which monolingual and cross-lingual objectives are integrated to learn cross-lingual word embeddings in a joint manner.

In the case of contextualized word embeddings, which we are interested in, there have been some attempts to use such off-line and on-line methods recently. The work by Pires et al. (1) shows that Multilingual BERT(M-BERT), released by Devlin et al. (2) as a single language model pre-trained from monolingual corpora in 104 languages, is surprisingly good at zero-shot cross-lingual model transfer, in which task-specific annotations in one language are used to fine-tune the model for evaluation in another language. This paper also argues that using a simple linear transformation matrix computed with aligned pairs, one can obtain the translation of unseen sentences with an acceptable accuracy, suggesting that cross-lingual embedding spaces can be mapped by a linear transformation.

(LateX template borrowed from NeurIPS 2019)

Furthermore, a work by Wang et al. (3) proposes Cross-Lingual BERT Transformation (CLBT), a simple and efficient off-line approach that learns a linear transformation from contextual word alignments. This paper also reports the effectiveness of the proposed method on a cross-lingual dependency parsing task setting.

The downside of these methods, however, is that they heavily rely on cross-lingual aligned corpora, which can be a critical problem, especially in the case of languages with less available resources. In this work we introduce a novel method, to use BabelNet (4), a multilingual encyclopedic dictionary, as a resource for extracting required contextual alignment pairs.

3 Proposed method

In this section, we provide a comprehensive explanation of our method.

3.1 Contextual Alignment

Traditional methods of learning static cross lingual word embeddings have been relying on various sources of supervision such as bilingual dictionaries, parallel corpus or online Google Translate. To learn contextualized cross-lingual word embeddings, however, we require supervision at context-level rather than token-level. One solution would be using parallel corpus as the source of supervision. In this approach, unsupervised bidirectional word alignment is first applied to the parallel corpus to obtain a set of aligned word pairs with their contexts. Then using these contextual aligned pairs, a transformation matrix is computed to map embeddings from one language to the other. Figure 1 is a toy illustration of the method.

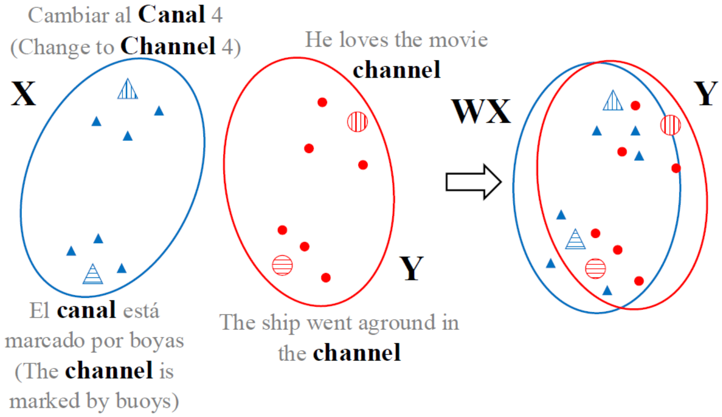
Cambiar al **Canal** 4
(Change to **Channel** 4)
He loves the movie **channel**
X
WX
Y
Y
El **canal** está marcado por boyas (The **channel** is marked by buoys)
The ship went aground in the **channel**

Figure 1: A simple illustration of how the method works. The contextualized embeddings of the word canal from Spanish is transformed to the semantic space of English

3.2 Source of Supervision

The downside of the suggested method, as we discussed earlier, is that it requires a parallel corpus as the source of supervision, which may be not available in many languages. In order to overcome this issue, we suggest using BabelNet as the source of supervision. There are some core concepts related to BabelNet worth mentioning:

- **Babel-Synset:** A Babel Synset is a set of multilingual lexicalizations that are synonyms expressing a given concept or named entity.
- **Babel-Gloss:** A Babel Gloss is the definition of a concept (Synset) in a given language. These definitions come from various sources, among which we use the WikiPedia source.

Suppose we want to find the alignment between English and Spanish languages. In order to retrieve the required alignment pairs, we take the following steps:

1. We first gather a list of frequent words in English. The number of words should be at least equal to the dimension of the embedding space. For this experiment, we gathered 3000 frequent English words.
2. Then we gather the corresponding Babel-Sysnet to these words. For the ambiguous words, we choose one of it's corresponding synsets which is marked as a main concept in BabelNet.
3. For each gathered synset, we extract it's gloss in both target and source language (in this case, Spanish and English).
4. Finally, using our multilingual pre-trained model (in our case, M-BERT), we extract the embedding for these gloss pairs. We use these embeddings as our contextual aligned pair.

3.3 Off-Line Transformation

Given a set of contextual pairs $\{x_i, y_i\}_{i=1}^{n}$ where $x_i \in \mathbb{R}^{d_1}$ is the contextualized embedding of Babel-Gloss g_i in the target language, and $y_i \in \mathbb{R}^{d_2}$ is the representation of its alignment in the source language, we can project the embeddings of target language to the space of the source language. We aim at finding an appropriate linear transformation $\mathbf{W} \in \mathbb{R}^{d_1 \times d_2}$, such that $\mathbf{W}x_i$ approximates y_i. This can be achieved by solving the following optimization problem:

$$\min_{\mathbf{W}} \sum_{i=0}^{n} ||\mathbf{W}x_i - y_i||^2 \tag{1}$$

In this case, an analytical solution can be found through singular value decomposition (SVD) of $\mathbf{Y^T X}$:

$$\mathbf{W} = \mathbf{VU^T} \tag{2}$$

where $\mathbf{U}\Sigma$ $\mathbf{V^T}$ = **SVD**($\mathbf{Y^T}$ $\mathbf{X}$). Here **X** and **Y** are embedding matrices in source and target languages respectively.

3.4 Experimental Setting

To show the effectiveness of our approach, we test it in a document classification task setting. We chose Jigsaw Multilingual Toxic-Comment Classification task since it is provided with a rich training-set in English and a test-set in many different languages including Spanish. In this task, given a document, the model has to label it either as toxic or non-toxic, where toxicity is defined as anything rude, disrespectful or otherwise likely to make someone leave a discussion.

3.5 Model Implementation

After computing English to Spanish alignment matrix, we use it in our model as it is shown in Figure 2. We take as input the document to be identified as toxic or non-toxic, and the language the sentence is written in. We first feed the document to M-BERT to obtain it's embedding (the CLS token embedding represents the embedding of the whole sequence), then using the language provided as input, we choose the appropriate transformation matrix. In this case, EN-EN transformation matrix is the identity matrix, since there is no transformation needed to map English to itself, and EN-ES transformation matrix is the matrix we previously computed using SVD. Then we simply transform the document embedding from the source language to the target language by multiplying it by the transformation matrix. finally we feed the transformed embedding to the classifier we previously trained in the source language.

4 Results

In this work, we used a sub-sample of 250K English training samples and tested the performance on 2500 Spanish test samples. We used accuracy as the evaluation metric. To see the effectiveness of the proposed approach, we ran two different experiments. In both experiments, the model is trained on English data and tested the performance on Spanish data. However, in the first experiment we

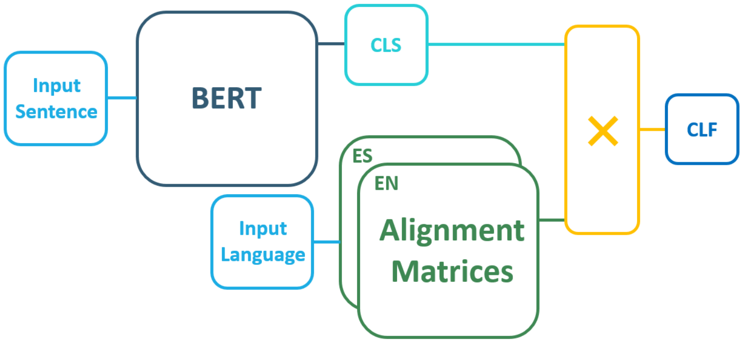

Figure 2: A simple illustration of how the method works. The contextualized embeddings of the word canal from Spanish is transformed to the semantic space of English

provided the model with English as input language (which is essentially equal to use the default model with no alignment), while in the second experiment we fed to the model Spanish as input language. In the former setting model performed with 82.13% accuracy while in the later it performed with 83.16% accuracy. Since no other factor is changed in these two different experiments, we can conclude that the alignment directly improves the performance of the model in terms of accuracy.

5 Discussion

In this work, we proposed a method for using BabelNet as a supervision resource for off-line zero-shot knowledge transfer between two languages. we also tested the proposed method on a sentiment analysis task setting and observed a direct improvement in the performance. This approach has the following advantages over on-line and other off-line methods:

- The first common advantage of off-line methods over on-line methods, is that they are generally less resource demanding.
- The other and more important advantage of our approach over other off-line methods, is that we use BabelNet as a resource instead of parallel corpora, which can be helpful especially in the case of languages with less available resources.

Future work can focus on analyzing the effectiveness of this method on a language pair with less topological similarity (i.e. with different order of verb/subject/object etc.). Furthermore, embeddings provided by different layers of M-BERT can be tested to see if there is any improvement in the performance. Finally, this method can be tested on other model such as RoBerta, which have explicit multilingual pre-training objectives.

References

[1] Telmo Pires and Eva Schlinger and Dan Garrette. *How multilingual is Multilingual BERT?*. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019.

[2] J. Devlin and Ming-Wei Chang and Kenton Lee and Kristina Toutanova. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. NAACL-HLT, 2019.

[3] Yuxuan Wang and W. Che and Jiang Guo and Yijia Liu and Ting Liu. *Cross-Lingual BERT Transformation for Zero-Shot Dependency Parsing*. EMNLP/IJCNLP, 2019.

[4] Roberto Navigli and Simone Paolo Ponzetto. *BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network*. Artificial Intelligence, 2012.