Fall 2021 CS 410 Technology Review

Date: 11/7/2021

Title: Word Embedding approaches and cross-lingual embeddings

By: Sree Alaparthi

Introduction

Humans can learn and understand the languages with ease when compared to computers because it is easy for humans to understand the relationship between the words. For example, we humans understand the words such as man and woman, good and bad, king and queen have a certain type of relationship. However, computers do not have the prior knowledge of the relationship between the words.

Word embeddings are basically a form of word representation that helps the computers understand the language. The words are represented in an n-dimensional space, where the words with similar meaning have a similar representation. This means two similar words are represented by the similar vectors that are very close to each other in vector space. Word embeddings have become an integral part of Natural Language Processing (NLP) tasks such as Machine Translation, Question Answering, Information Retrieval and several other machine learning tasks.

In this paper, we will explore the different word embedding approaches and how they can be used in the context of low resource languages (for example, Asian Languages).

Overview of Word Embedding Techniques

Word Embeddings can be categorized into two categories:

1) Monolingual - Embeddings for a single language.
2) Cross-lingual - Embeddings for a pair of languages generated using methods such as MUSE.

Monolingual embeddings can further be classified into ***contextual*** and ***non-contextual*** embeddings. Word Embeddings trained using methods such as ***Word2Vec***, ***GloVe*** and ***FastText*** are often called as ***non-contextual embeddings*** because the output of such techniques is a vocabulary of words. The elements of this vocabulary are words and its corresponding word embeddings. Therefore, for a given word, it's embeddings are always the same in any sentence it occurs. The pre-trained word embeddings are ***static***. For example, the word "left" has the same representation in these two sentences: "Steve left his computer in the meeting room." and "Steve's house is on left side of Brian's house." However, the word "left" has two different meanings and needs to have tow different representations in the embedding space.

C***ontextual embeddings*** are generated by modern techniques such as ***BERT***, ***CoVe*** and ***Elmo***. The embeddings are generated by passing the entire sentence to the pre-trained model. In this case, the embedding for each word depends on the other words in each sentence. The other words in a sentence are referred as ***context***, Therefore, the given word will not have a static embedding, but the embeddings are dynamically generated from pre-trained model.

Cross-lingual embeddings are used to represent meaning and transfer knowledge across different languages. ***mBERT***, ***XLM*** and ***MUSE*** are some of the multilingual pre-trained models available.

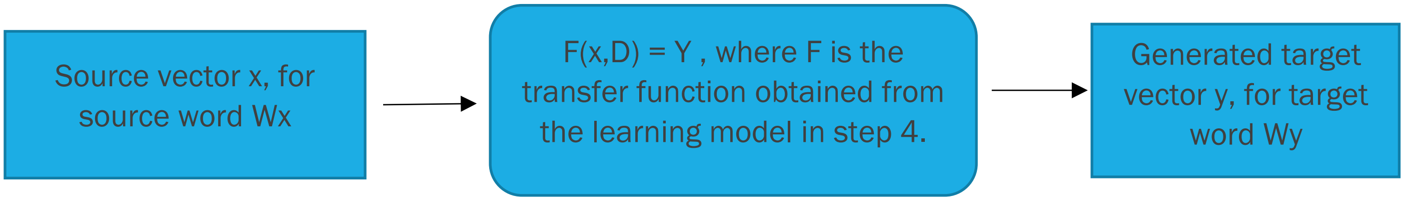
Why do we need Cross-lingual embeddings?

Word Embeddings provide rich set of features for downstream NLP tasks. One of the shortcomings of the pre-trained language models is the corpora size used for their training. Almost all the models use Wikipedia corpus to train models which is insufficient for resource constraint languages (such as Hindi, Arabic etc.) as Wikipedia does not have significant number of articles or text in these languages. The goal of Cross-lingual word embeddings is to bridge the gap between high-resource language such as English and low-resource languages such as Hindi, Arabic, Telugu etc. by enabling the machine to learn multi-lingual word representations even without any parallel data.

Cross-lingual embeddings

Cross-lingual embedding is accomplished by mapping the vectors from one language's embedding space into that of the other language through a transfer function. Here is the flowchart of how the learning model can be used to generate cross-lingual embeddings. The general steps followed to generate the bi-lingual embeddings are as follows:

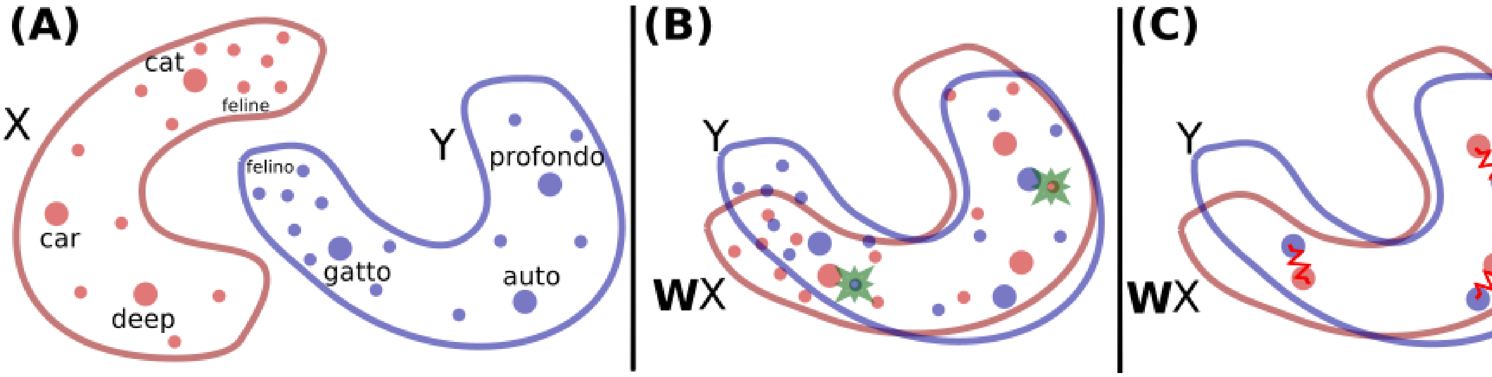
1) Obtain the pre-trained source embedding, X. The source language is a High Resource language such as English.
2) Use the same algorithm (Word2Vec/GloVe/FastText) as the source embedding(X) to generate the Target Embedding - Y.
3) Source-Target Dictionary - D: Minimal dictionary for learning bilingual information
4) The pre-trained source embedding X, target embedding Y and source-target dictionary D are fed to the machine learning model M for learning linear transfer function F. F = M (X, Y, D).


Source vector x, for source word Wx
F(x,D) = Y , where F is the transfer function obtained from the learning model in step 4.
Generated target vector y, for target word Wy

MUSE: Multilingual Unsupervised and Supervised Embeddings

MUSE is a python library for multilingual word embeddings. MUSE provides state-of-the-art multilingual word embeddings. The word embeddings for each language are generated using the FastText algorithm. MUSE also provides large-scale high-quality bilingual dictionaries for training and evaluation. MUSE supports both supervised and unsupervised learning. MUSE is trained simultaneously on multiple languages using a deep learning technique called as "Multi-task Dual

Encoder Model". This means that sentences from different languages are mapped in the vector space using the same model.



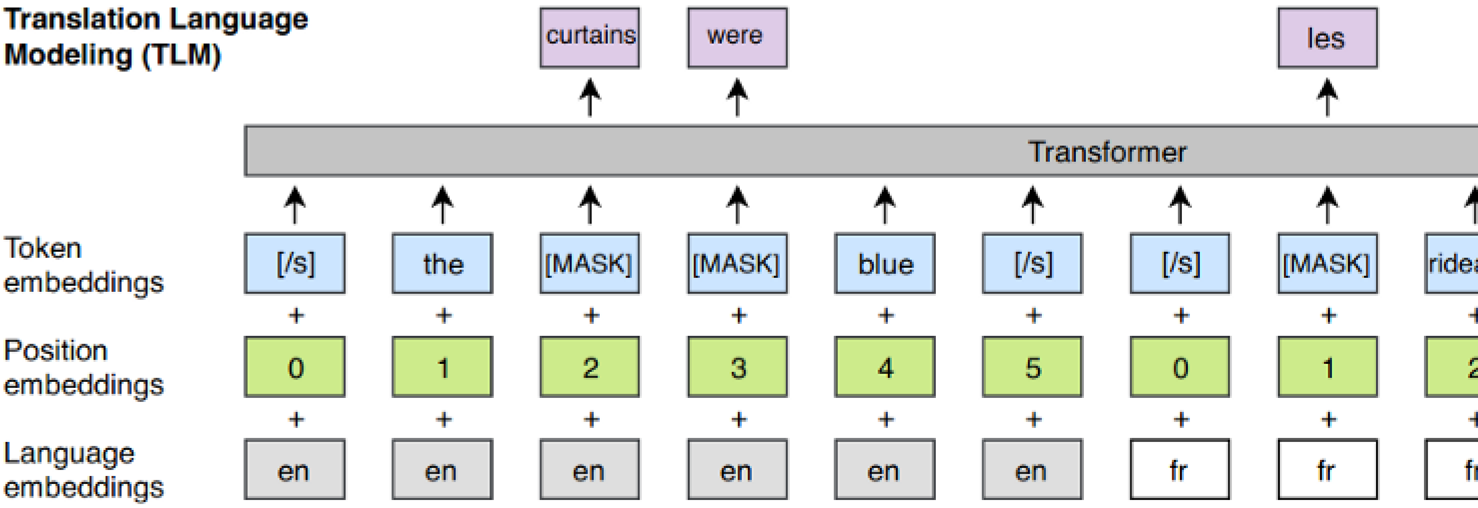
In MUSE, we start with a fixed set of word embeddings in each language. We learn a mapping Y = XW between two spaces, where X and Y are source and target language word embedding vectors respectively. The mapping distribution of vectors in each of the two languages should be ultimately same. MUSE does this by using Generative Adversarial Network. The discriminator is trained to determine of the two vectors are from the same language and the generator is trained to map the vectors from one language to other.

XLM

XLM is a Transformer based architecture that is pre-trained using one of the following language models.

1) Casual Language Modeling (CLM) – models the probability of a word based on the previous words in a sentence.
2) Masked Language Modeling (MLM) – using the masked language modeling of BERT.
3) Translation language modeling – a new translation language modeling to improve the cross-lingual pre-training.

The CLM and MLM tasks work well on monolingual corpora, and they do not take advantage of available parallel translation data. However, TLM takes a sequence of parallel sentences from the translation data and randomly masks tokens from the source and the target sentence.



For example, in the figure above, the words are masked both from English and French sentences. All the words in the sequence contribute to the prediction of a given masked word. This provides the cross-lingual mapping among the tokens. The authors in the paper (https://arxiv.org/abs/1901.07291) recommend cross-lingual model pre-training using either CLM, MLM or MLM used in combination with TLM.

XLM pre-training can be leveraged for tasks such as Zero-shot cross lingual classification, supervised and unsupervised neural machine translation, language models for low-resource languages and Unsupervised cross-lingual word embeddings.

Conclusion

There are about 7000 languages around the world and therefore the multilingual NLP has been a long-standing goal. Current NLP systems mostly support the English language due to limited set of available parallel corpora even for a resource-rich languages. The cross-lingual techniques help remove the constraint on the development of multilingual NLP systems. Cross-lingual Embeddings allow comparing a word's meaning in multiple languages and enable cross-lingual transfer learning. These properties of Cross-lingual embeddings are beneficial for both resource-rich and low resource languages in downstream NLP tasks.

References

- *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
- *GloVe* - https://nlp.stanford.edu/projects/glove/
- *Cross-lingual and Multilingual Word Embeddings* - http://www.iitp.ac.in/~shad.pcs15/data/CL-ML-WE.pdf
- *Refinement of Unsupervised Cross-Lingual Word Embeddings* - https://arxiv.org/pdf/2002.09213.pdf
- *A Survey of Cross-lingual Word Embedding Models* - https://www.jair.org/index.php/jair/article/view/11640/26511
- *Cross-lingual Language Model Pretraining* - https://arxiv.org/abs/1901.07291

- *MUSE* - https://github.com/facebookresearch/MUSE
- *MUSE: Modularizing Unsupervised Sense Embeddings* - https://aclanthology.org/D17-1034.pdf