Inigo Alonso
April 17th, 2019

CAPSTONE PROPOSAL

DOMAIN BACKGROUND

The domain of this problem is Natural Language Processing. The more specifically, with the purpose of getting some practice with sequential models, it will be focused on the use of word embeddings for sentiment classification.

As I wanted to work on something recent, I have decided to have as a goal to evaluate the use of cross-lingual word embeddings for sentiment classification, without requiring parallel corpora.

In particular the MUSE embeddings provided by the developers on facebook will be used.

PROBLEM STATEMENT

Since the representation of words by a vector representation of a number different semantic spaces came out first in 2013 (Mikolov et al), it has become the most ubiquitous technique in the field of Natural Language Processing (NLP). There is a variety of problems for which its use has been researched, text classification (Peng Jin et al , Qian Liu et al), document clustering (Inzamam Rahaman et al), part of speech tagging(Yu Dong et al), named entity recognition (Cıcero Nogueira dos Santos et al) sentiment analysis (Ruggero Petrolito et al). In particular, the project will focus on the latest of these tasks.

As it happened in the field of image recognition, where the weights from the most successful models in the ImageNet challenge where used from transfer learning, the embeddings that have performed the best until this day have been made available for the public. Being the most popular of this resources the GloVe embedding dataset(Jeffrey Pennington et al).

The current project utilizes a similar resource, the MUSE embeddings (Alexis Conneau et al). This MUSE embeddings are a cross-lingual set of pre-trained embeddings, which means that, they were created with the goal of fixing the traditional approach to cross-lingual document classification. This old approach involved translating from a language to English and using an English classifier to classify the text originally written in another language. With this procedure, the errors in the translation get propagated and increase the error rate in the classification.

The resulting MUSE embeddings are language independent, that is, each language gets its own set of embeddings. With the constraint that all these embeddings are represented in the same vector space.

Hence, these state-of-the-art embeddings allow to take advantage of the data in high resource languages to perform NLP tasks over other low resource languages.

Inigo Alonso
April 17th, 2019

In this paper, we examine the performance of cross-lingual embeddings. This evaluation is performed for Sentiment Classification of movie reviews, a task more complex than the traditional topic classification problem. The multilingual embeddings that are compared to prove their validity are the ones obtained for English (high resource language).and Spanish (lower resource language).

DATASET AND INPUTS

High Resource Language Data

The training data consists of 50,000 reviews in English from IMDB, allowing no more than 30 reviews per movie. The constructed dataset contains an even number of positive and negative reviews, so a naive classifier would yield 50% accuracy. Following previous work on polarity classification, we highly polarized reviews have only been considered. In particular, the review is labelled as a negative review if it has a score ≤ 4 out of 10, and positive, if has a score ≥ 7out of 10. Neutral reviews are not included in the dataset.

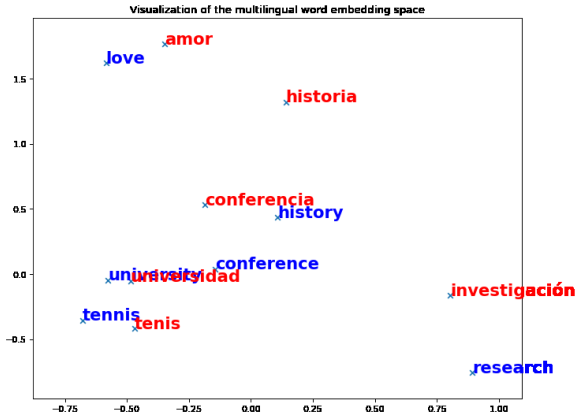
This dataset was made available to the public after (Andrew L. et al.) provided a benchmark for future work in this area. The training and test sets are evenly divided. The training set is the same 25,000 labeled reviews used to induce word vectors in their work. The access to the dataset has become the most popular for sentiment classification, thanks to its availability through keras.

Low Resource Language Data

The test data is provided by Dr. Fermín L. Cruz Mata from the University of Seville (http://www.lsi.us.es/~fermin/index.php/Main_Page) containing 4380 movie reviews in Spanish, extracted from the site www.muchocine.net. This movie review are ranked from 1 to 5. Which, so as to follow the polarized approach in the training set, are considered as positive if the rate is 4 or 5 and negative if it is 1 or 2. Hence, the review of rating equal to 3 are discarded.

Input: MUSE Embeddings

The particularity of the cross lingual word embeddings is the following: if you represent the embedding of a word in English and the one of the same word in a different language, using t-SNE or a similar technique, you will be able to see both words closely together in the vector space. This is just a way that shows the similarity across a chosen dimensionality space between the vectors representing 2 words that have the same meaning, but belong to different languages.



This magic, does not occur automatically when training each embedding on its own language, it requires to project the language independent embedding on a common space.

To make the projection happen a projection matrix W, is learnt. This matrix W, is the result of the optimization over a cost function with the goal of minimizing the loss between the word x_i in English and its translation in the objective language, y_i.

$$M = argmin_W \sum_i ||x_i - W y_i||^2$$

When one analyzes the process to obtain the MUSE embeddings presented in the work by (Alexis Conneau et al), 3 main steps are identified:

1. Get monolingual embeddings: trained separate embeddings for each language using fastText and a combination of data from Facebook and Wikipedia.

2. Alignment of monolingual embeddings: The projection of the embeddings into a common space can be done thorugh 2 approaches.

a) The supervised way: This method uses parallel data, using a bilingual dictionary, it learns a mapping from the source to the target space by solving the orthogonal Procrustes problem.

b) The unsupervised way: Does not require parallel data. In this case it uses, adversarial training to do the mapping, and instead of, by minimizing the loss function the mapping is performed by maximizing the similarity measure CSLS.

3. Evaluation of Cross-lingual embeddings for several tasks.

https://github.com/facebookresearch/MUSE

SOLUTION STATEMENT

The Cross-lingual Embeddings Problem

While the accuracy of the embeddings claimed by the Facebook Developers Group is remarkable, there are some concerns about their application to sentiment analysis.

First, the projection of the words to a common space is fails to explain how words that suffer from polysemy and synonymy in their original language are translated into this common space.

Other issue that will have to be solved in the preprocessing steps, is the absence of some of the words from the movie dataset in the embeddings and vice versa. This happens because some of the movie characters and titles are meaningless words that are not found in the resources used to train the multilingual embeddings. In the same way, the corpus on which the MUSE emebddings are train is way larger than the 50,000 movie reviews in English.

The nature of this corpus is not composed of reviews for the most part, which makes it less sensible to capture sentiments. A case of this could be the words sucks, used for reviews in a very different way to its contextual meaning.

The Sentiment Analysis Problem

The distinction between positive and negative reviews is a fairly easy task to perform for humans, especially in comparison to that of the standard text categorization problem, where topics can be closely related. Nonetheless, machine learning algorithms have thrived to a major extent with the latter type of problems.

As pointed Shivakumar Vaithyanathan et al, the complexity of the problem is not appreciated by human intuition, as we fail to determine that a sufficiently long list of positive and negative sentiment adjectives such as excellent and terrible would suffice to achieve considerable performance on the task. Nonetheless, this approach fails to acknowledge the importance of comparative structures like still or as a matter of fact that having proved be as paramount of indicators as the previous adjectives. As a result, the strategy mentioned beforehand only yielded 69% accuracy.

Additionally, some preprocessing techniques that identify negation should be considered to speed up calculations. An example of this could be the sentence *I didn't like the scene where…,* the vector representation of like should not be the same as the one of not_like, for this reason the sentence could be rewritten as *I not_like the scene where…* for better performance. Although the standard LSTM architecture should be able to get feedback from previous words, the previous technique could be helpful to speed up calculations.

EVALUATION METRICS

The network will be trained and tested on English data, the accuracy that is expected to be reached in this scenario will be somewhere around 80%.

After that Spanish that will be used for testing purposes solely. Here the expected results wil highly depend on the limitations of the network architecture, the quality of the Spanish data and the effectiveness of the MUSE cross lingual embeddings. An

ambitious result for this test could be somewhere between 10-15% below the English accuracy.

PROJECT DESIGN

1. Preprocessing of English data
2. Design network architecutre
3. Train the network and run it on test data
4. Preprocessing of Spanish Data
5. Testing on Spanish Data
6. Evaluation of initial thoughts and execution

REFERENCES

Tomas Mikolov, Ilya Sutshever, Kai Chen, Greg Corrado and Jeffrey Dean. *Distributed Representations of Words and Phrases and their Compositionality,* 2013.

Peng Jin, Yue Zhang, Xingyuan Chen, Yunqing Xia. *Bag-of-Embeddings for Text Classification*, 2016.

Qian Liu, Heyan Huang, Yang Gao, Xiaochi Wei, Yuxin Tian, Luyang Liu. *Task-oriented Word Embedding for Text Classification*, 2018

Inzamam Rahaman, Patrick Hosein*. Exploiting Gaussian Word Embeddings for Document Clustering*, 2017

Yu, Dong Jin Wang, Wei. *Part-Of-Speech Tag Embedding for Modeling Sentences and Documents* 2016

Cicero Nogueira dos Santos, Victor Guimaraes. *Boosting Named Entity Recognition with Neural Character Embeddings, 2015.*

Ruggero Petrolito, Felice Dell'Orletta*. Word Embeddings in Sentiment Analysis*, 2017

Jeffrey Pennington, Richard Socher, Christopher D. Manning. *GloVe: Global Vectors for Word Representation*, 2014.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, Herve Jegou. *Word Translation Without Parallel Data*, 2017.

Ying Lin, Shengqi Yang, Veselin Stoyanov, Heng Ji. *A Multi-lingual Multi-task Architecture for Low-resource Sequence Labeling,* 2018.

Shyam Upadhyay, Manaal Faruqui, Chris Dyer, Dan Roth. *Cross-lingual Models of Word Embeddings: An Empirical Comparison,* 2016.

Jiang Guo, Wanxiang Che1, David Yarowsky, Haifeng Wang, Ting Liu. *Cross-lingual Dependency Parsing Based on Distributed Representations,* 2016.

Tianze Shi Zhiyuan Liu Yang Liu Maosong Sun. *Learning Cross-lingual Word Embeddings via Matrix Co-factorization,* 2015.

Sebastian Ruder, Parsa Ghaffari, John G. Breslin. *Deep Learning for Multilingual, Aspect-based Sentiment Analysis*, 2016.

Jan Deriu, Aurelien Lucchi, Valeria De Luca, Aliaksei Severyn, Simon Müller, Mark Cieliebak, Thomas Hofmann, Martin Jaggi. *Leveraging Large Amounts of Weakly Supervised Data for Multi-Language Sentiment Classification,* 2017.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzatoy , Ludovic Denoyerx , Herve Jegouy. *Word Translation Without Parallel Data,* 2017.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng and Christopher Potts. *Learning Word Vectors for Sentiment Analysis,* 2011.

Shivakumar Vaithyanathan, Bo Pang and Lillian Lee. *Thumbs up? Sentiment Classification using Machine Learning Techniques,* 2002.