Zero-Shot Sentiment Analysis for Code-Mixed Data

Siddharth Yadav, Tanmoy Chakraborty
IIIT-Delhi, India
siddharth16268@iiitd.ac.in, tanmoy@iiitd.ac.in

Abstract

Code-mixing is the practice of alternating between two or more languages. A major part of sentiment analysis research has been monolingual and they perform poorly on the code-mixed text. We introduce methods that use multilingual and cross-lingual embeddings to transfer knowledge from monolingual text to code-mixed text for code-mixed sentiment analysis. Our methods handle code-mixed text through zero-shot learning and beat state-of-the-art English-Spanish code-mixed sentiment analysis by an absolute 3% F1-score. We are able to achieve 0.58 F1-score (without a parallel corpus) and 0.62 F1-score (with the parallel corpus) on the same benchmark in a zero-shot way as compared to 0.68 F1-score in supervised settings. Our code is publicly available on github.com/sedflix/unsacmt.

Introduction

Language used on social media platforms, from where we get most of the code-mixed text, differs from the standard language found in more formal texts such as Wikipedia, books. Monolingual NLP methods perform poorly on code-mixed data. Therefore, it is crucial to build technology for code-mixed text. Analysis of code-mixed text is hard due to lack of code-mixed text corpus and datasets, large amount of unseen constructions caused by combining lexicon and syntax of two or more languages, and large number of possible combinations of code-mixed languages. Constructing one dataset requires huge manual effort, and therefore, seems infeasible. It will be more beneficial to build technologies to transfer knowledge from related monolingual datasets to code-mixed domain efficiently.

(Vilares, Alonso Pardo, and Gómez-Rodríguez 2015) first introduced sentiment analysis on English-Spanish code-mixed data. (Pratapa, Choudhury, and Sitaram 2018) showed that using pretrained embeddings learned from code-mixed data performs better than bilingual embeddings.

In this work, we introduce methods that use different kinds of multilingual and cross-lingual embeddings to efficiently transfer knowledge from monolingual text to code-mixed text for the task of sentiment analysis on code-mixed text. We make our methods independent of any supervised signals between two languages by not using machine translation or language identification systems. Our goal in this work is to derive an embedding space and training techniques that assist efficient knowledge transfer from monolingual text to code-mixed text.

Embeddings

We choose the followings embeddings to work on our task.
• **Mapping based model:** First, we learn monolingual word representations[1] from monolingual corpus and then learn a transformation matrix to map representation from one language to other. We use MUSE[2], which returns a word-aligned embeddings for each language. We merge two different embeddings by taking an average of word vectors of common words and appending others. **MUSE-USUP** (MUSE Unsupervised) uses adversarial learning to learn the transformation matrix followed by refinement using Procrustes algorithm. **MUSE-SUP** (MUSE Supervised) uses bilingual dictionary to learn the mapping using Procrustes alignment.
• **Pseudo-multilingual model:** We obtain word/subword representation by training word embedding model on single corpus containing text from both the languages, formed by concatenation of monolingual corpus. We use subword based vocabulary, which helps us to get word vectors of even misspelled words without affecting the final vocabulary size. **MultiBPEmd** (Heinzerling and Strube 2018) is shared subword vocabulary and multilingual embeddings in 275 languages. It was formed by concatenating articles from all 256 languages on Wikipedia, learning a BPE subword segmentation model using SentencePiece, and using GloVe to train subword embeddings. We use a version with vocabulary size of 1,000,000 and dimensions of 300. **Custom-fastText**[3] is trained on a custom corpus, generated by concatenating 70 millions English, 70 millions Spanish and 20 millions code-mix tweets mined by us from twitter. Our embedding has vocabulary size of 1,343,436 and dimensions of 100.
• **Sentence level multilingual autoencoder model:** Due to abundant research in machine translation, a huge amount of sentence-aligned parallel data is available for popular lan-

[1] https://fasttext.cc/docs/en/crawl-vectors.html
[2] https://github.com/facebookresearch/MUSE
[3] https://github.com/facebookresearch/fastText

guage pairs. Using this data and machine translation models, we can learn cross-lingual sentence-level representation. **LASER** (Artetxe and Schwenk 2019) offers a shared multilingual sentence-level representation for 93 different languages. It has a common language-agnostic bidirectional LSTM encoder which provides the sentence embedding. The sentences embedding is then decoded by language-specific decoder. It is trained on 223 millions parallel sentences. It has a BPE vocabulary size of $50,000$ and builds 1024 dimensional sentence representation.

Experiments

Datasets: We use a combination of five different Twitter sentiment analysis datasets for training and testing: (i) **English:** 20,632 tweets from SemEval-2017 Task 4 (Rosenthal, Farra, and Nakov 2017) and 4,241 English tweets from Sentistrength[4]; (ii) **Spanish:** 7,217 tweets from Cerón-Guzmán and 3,202 tweets from (Villena-Román et al. 2015); (iii) **Code-mix:** 2,449 train and 613 test instances of code-mixed tweets from (Vilares, Alonso, and Gómez-Rodríguez 2016). All of the instances have three labels: neutral, positive, and negative. Tweets are prepossessed using TweetMotif[5] before embedding specific prepossessing.

Models: Our task is to classify code-mixed sentences into positive, negative, or neutral sentiment. Our choice of this tasks is motivated by the availability of annotated code-mixed and monolingual datasets. As our goal is to demonstrate effectiveness of embeddings, we use simple classifiers for our task. Models using sentence embedding have a single-layer dense neural network. Models using embedding have a single-layer bidirectional LSTM with 50 units for dropout and recurrent dropout of 0.3.

Training: We use ADAM optimiser with learning rate of 0.001 and momentum of 0.9. Each step in our training curriculum is trained for 30 epochs with early stopping at the patience of 10.

Pratapa, Choudhury, and Sitaram (2018) showed that training the models with monolingual instances and then on code-switched instances is a better curriculum. We use similar curriculum where the latter step is only used to get supervised results.

Results: Table 1 reports the result of our experiments. We beat the state-of-the-art on the task of English-Spanish code-mixed sentiment analysis by an absolute 3.11% using our custom-fastText embeddings. Custom-fastText gives us the best zero-shot performance(without parallel corpus) with 0.58 F1-score. This shows training on actual code-mixed corpus with subwords embedding helps in developing better embeddings because just concatenated monolingual or parallel corpus fail to capture the fine syntactic and semantic features of code-mixed text. LASER produces a remarkable zero-shot F1-score of 0.62, and performs at par with the previous state-of-the-art. This shows that training on translation tasks, with common encoder, and using parallel corpus indeed helps.

[4] http://sentistrength.wlv.ac.uk/

[5] https://github.com/brendano/tweetmotif

Model	Embeddings	Learning Type	F1-score
Previous Work			
LSTM	-	Supervised	54.40
LSTM	Bilingual Skip-gram[6]	Supervised	61.50
GirNet[7]	-	Supervised	63.40
LSTM	Synthetic CM based[8]	Supervised	**64.60**
Our Work			
LSTM	MUSE-USUP	Supervised	55.40
LSTM	MUSE-SUP	Supervised	56.61
LSTM	MultiBPEmb	Supervised	64.00
NN	LASER	Supervised	64.44
LSTM	FastText	Supervised	**67.71**
LSTM	MUSE-SUP	Zero-Shot	53.53
NN	LASER	Zero-Shot	**61.72**
LSTM	MUSE-USUP	Zero-Shot	51.99
LSTM	MultiBPEmb	Zero-Shot	54.59
LSTM	FastText	Zero-Shot	**58.40**

Table 1: F1-score of the competing models on code-mixed sentiment analysis task.

Conclusion

The best monolingual sentiment analysis technique performs poorly on code-mixed texts. We showed that usage of cross-lingual and multilingual embeddings improves the performance of sentiment analysis techniques on code-mixed text. We believe that the future of code-mix language analysis will lie in zero-shot approaches, where we efficiently transfer knowledge from monolingual datasets to code-mixed datasets using embeddings and models that can handle code-switching without much loss in performance.

References

Artetxe, M.; and Schwenk, H. 2019. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *TACL* .

Cerón-Guzmán, J. A. 2017. Classifier Ensembles That Push the State-of-the-Art in Sentiment Analysis of Spanish Tweets. In *TASS*.

Gupta, D.; Chakraborty, T.; and Chakrabarti, S. 2019. Girnet: Interleaved multi-task recurrent state sequence models. In *AAAI*.

Heinzerling, B.; and Strube, M. 2018. BPEmb: Tokenization-free Pre-trained Subword Embeddings in 275 Languages. In *LREC*.

Pratapa, A.; Choudhury, M.; and Sitaram, S. 2018. Word Embeddings for Code-Mixed Language Processing. In *EMNLP*.

Rosenthal, S.; Farra, N.; and Nakov, P. 2017. SemEval-2017 Task 4: Sentiment Analysis in Twitter. In *SemEval-2017*.

Vilares, D.; Alonso, M. A.; and Gómez-Rodríguez, C. 2016. EN-ES-CS: An English-Spanish Code-Switching Twitter Corpus for Multilingual Sentiment Analysis. In *LREC*.

Vilares, D.; Alonso Pardo, M.; and Gómez-Rodríguez, C. 2015. Sentiment Analysis on Monolingual, Multilingual and Code-Switching Twitter Corpora. In *WASSA*.

Villena-Román, J.; Martínez-Cámara, E.; García-Morera, J.; and Zafra, S. M. 2015. TASS 2014-The Challenge of Aspect-based Sentiment Analysis. *Procesamiento de Lenguaje Natural* .