HIT_SUN@Dravidian-CodeMix-FIRE2020:Sentiment Analysis on Multilingual Code-Mixing Text Base on BERT

Huilin Sun*[a], Jiaming Gao[b] and Fang Sun[a]

[a]*Heilongjiang Institute of Technology, Harbin, Heilongjiang, 150000, China*
[b]*Harbin Engineering University, Harbin, Heilongjiang, 150000, China*

Abstract
This paper mainly introduces the method used in the FIRE2020@Sentiment Analysis for Davidian Languages in the Code-Mixed Text evaluation task [1]. Sentiment analysis mainly identifies the sentiment tendencies of a given text, such as positive, negative, unknown, mixed emotions. This evaluation task is to sentiment analysis on Code-Mixed text, including Tamil-English[2] and Malayalam-English[3] mixed text analysis. This paper uses a bidirectional pre-training language model (BERT) to solve the problem of sentiment classification of cross-language text. The BERT model is divided into two parts: pre-training and fine-tuning. The pre-training part uses two novel unsupervised prediction tasks: the masked language model and the next sentence prediction. The fine-tuning part is to connect a fully connected neural network as the classification layer after the pre-training model, then use the classification dataset to fine-tune the parameters of the whole network. In the fine-tuning process, only need to use a small amount of sentiment classification data to get a good classification result. The BERT model achieved ranked 2nd in the Malayalam-English evaluation and ranked 4th in the Tamil-English evaluation.

Keywords
BRET, fine-turning, Code-Mixed, Sentiment Analysis

1. Introduction

Sentiment analysis is an important research problem in natural language processing. With the development of social media, mixed-language texts have gradually become a common phenomenon in media communication. The task[4] of code-mixed text sentiment recognition mainly focuses on two mixed languages in Dravidian languages (Malayalam-English and Tamil-English). sentiment categories mainly include five categories: Positive, Negative, Mixed feelings, unknown state, not-Tamil/not-Malayalam[5]. To solve the problem of cross-language sentiment analysis, this paper directly uses the multi-language pre-training language model(multi_cased_L-12_H-768_A-12 https://github.com/google-research/bert) to solve the code-mixed sentiment classification problem. Bidirectional Encoder Representations from Transformers (BERT) is a bidirectional pre-training language model based on the transformer proposed by Google in 2018. This model has achieved new state-of-the-art results on eleven natural language processing

RIRE2020: Forum for Information Retrieval Evaluation, December 16–20, 2020, Hyderabad, India
EMAIL: sunhuilin24@163.com (H. Sun*); gaojiaming24@163.coml (J. Gao); 15846044997@163.coml (F. Sun)
ORCID:

tasks and has become a new benchmark model. Therefore, the two-stage training method has become the mainstream technology in the NLP field.

2. Model

BERT is designed to pre-train deep bidirectional representations by jointly conditioning on both left and right contexts in all layers[6]. The training process of the BETR model can be divided into two parts: the pre-training part and the fine-tuning part. The pre-training task of BERT adopts the method of masked language model and the next sentence prediction task. pre-training the model using large-scale unsupervised text, and storing the semantic information of the text in the pre-training model. After the model pre-training stage is completed, the model is fine-tuned according to different natural language processing tasks. In the fine-tuning stage, only a small amount of training data is needed to make the deep learning model achieve well generalization performance. The pre-training model can be effectively transferred to other tasks, such as text classification and natural language inference.

BERT's model architecture is a multi-layer bidirectional Transformer encoder based on the original implementation described in "Attention is all you need"[7]. The BERT-base model contains 12 layers of Transformer blocks, each of which the hidden size is 768, each layer contains 12 self-attention heads, and each attention head has a size of 64. The input of BERT contains a sequence of 512 tokens, including token embeddings, position embeddings, segment embeddings. The first token input to the model must be the [CLS] token, which means the beginning of the sentence, and the last token must be [SEP] token, which means the end of the sentence. The input of the model can contain one or two sentences, and the different sentences are separated by [SEP] token.

For classification tasks, we only use the information in the [CLS] token output by the model as the feature vector of the input sentence. Then use a single-layer fully connected neural network as the classifier, and the SoftMax layer is added to the top of BERT to predict the probability of label Y:

$$P(Y \mid X) = softmax(WX),$$

Where W represents the parameter matrix of the classification layer, X represents the [CLS] token contains the semantic information of the sentence. We fine-tuning all the parameters of the BERT model, including the parameter W of the classification layer, to maximize the log probability of the correct label.

3. Experiments

3.1. Datasets

This evaluation dataset comes from YouTube video comments, including Tamil-English and Malayalam-English. The datasets contain three Code-Mixed sentences: Inter-Sentential switch, Intra-Sentential switch, and Tag switching. Most of the comment content is composed of the Roman script and contains the grammatical structure of Tamil and Malayalam. At the same time, part of the comment content contains English content.

Table 1
The number of tags in the Tamil-English data

Category	Train + Dev	Test
Positive	8484	2075
Negative	1613	424
Mixed feelings	1424	377
unknown state	677	173
not-Tamil	397	100
All	12595	3149

Table 2
The number of tags in the Malayalam-English data

Category	Train + Dev	Test
Positive	2246	565
Negative	600	398
Mixed feelings	333	177
unknown state	1505	138
not-Malayalam	707	70
All	5391	1348

This sentiment classification task mainly involves two Code-Mixed texts, including Tamil-English and Malayalam-English. Tamil-English train dataset contains 11335 data, the dev dataset contains 1260 data, the test dataset contains 3149 data, total data volume is 15744. Malayalam-English train dataset contains 4851 data, the dev dataset contains 540 data, the test dataset contains 1348 data, total data volume is 6739.

Because we merge the original train dataset and dev dataset like the final train dataset, we put them together for quantity statistics. The number of classification tags in each dataset is shown in the following table:

3.2. Experiments Results

The pre-training model version we chose is BERT-base, Multilingual Cased version. The pre-training corpus contains 104 languages (including English, Malayalam, Tamil). The model contains 12 layers of transformers, each layer contains 768-hidden, 12-attention-heads, a total of 110M parameters. In the model fine-tuning stage, first use the original 11335 (4851) pieces of train datasets to fine-tune the model, and adjust the parameters of the model fine-tuning according to the sentiment classification result of the fine-tuned model on the dev dataset. The parameters include learning rate, batch-size, maximum sentence length, epochs. The specific parameter settings are shown in the following table:

In the final classification prediction stage, we merge the data from the original train dataset and the original dataset into new 12595 (5391) pieces of training data, use the fine-tuning parameters in the above table to re-tune the model, and then make predictions on the test dataset. The classification result is as follows:

Table 3
BERT Fine-tuning parameter settings

Parameter	Value
learning rate	2e-5
batch size	128
max-seq-length	32
train-epochs	3

Table 4
BERT Experimental Result Tamil-English

classification metrics	Value
Precision	0.61
Recall	0.64
F-Score	0.62
Rank	4

Table 5
BERT Experimental Result Malayalam-English

classification metrics	Value
Precision	0.73
Recall	0.73
F-Score	0.73
Rank	2

4. Conclusions

In this evaluation, we use the pre-trained language model BERT (Bidirectional Encoder Representations from Transformers) to solve the sentiment classification problem of Code-Mixed text. Because the BERT model uses unsupervised corpus covering 104 languages (including English, Tamil, and Malayalam) in the pre-training stage, the cross-language problems in sentiment classification can be alleviated from the root cause. The BERT model uses multiple layers of transformers to extract the deep semantic features of the text. In the fine-tuning stage, only need a small amount of sentiment classification data to fine-tune the BERT to achieve good classification performance. Finally, the BERT model achieved ranked 2nd in the Malayalam-English evaluation and ranked 4th in the Tamil-English evaluation.

References

[1] B. R. Chakravarthi, Leveraging orthographic information to improve machine translation of under-resourced languages, Ph.D. thesis, NUI Galway, 2020.

[2] B. R. Chakravarthi, R. Priyadharshini, V. Muralidaran, S. Suryawanshi, N. Jose, E. Sherly,

J. P. McCrae, Overview of the track on Sentiment Analysis for Dravidian Languages in Code-Mixed Text, in: Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2020). CEUR Workshop Proceedings. In: CEUR-WS. org, Hyderabad, India, 2020.

[3] B. R. Chakravarthi, R. Priyadharshini, V. Muralidaran, S. Suryawanshi, N. Jose, E. Sherly, J. P. McCrae, Overview of the track on Sentiment Analysis for Dravidian Languages in Code-Mixed Text, in: Proceedings of the 12th Forum for Information Retrieval Evaluation, FIRE '20, 2020.

[4] B. R. Chakravarthi, V. Muralidaran, R. Priyadharshini, J. P. McCrae, Corpus creation for sentiment analysis in code-mixed Tamil-English text, in: Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL), European Language Resources association, Marseille, France, 2020, pp. 202–210. URL: https://www.aclweb.org/anthology/2020.sltu-1.28.

[5] B. R. Chakravarthi, N. Jose, S. Suryawanshi, E. Sherly, J. P. McCrae, A sentiment analysis dataset for code-mixed Malayalam-English, in: Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL), European Language Resources association, Marseille, France, 2020, pp. 177–184. URL: https://www.aclweb.org/anthology/2020.sltu-1.25.

[6] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, CoRR abs/1706.03762 (2017). URL: http://arxiv.org/abs/1706.03762. arXiv:1706.03762.