International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT217353

An Efficient Cross-Lingual BERT Model for Text Classification and Named Entity Extraction in Multilingual Dataset

Asoke Nath*, Rahul Gupta, Debapriya Kandar
Department of Computer Science, St. Xavier's College (Autonomous), Kolkata, West Bengal, India

ABSTRACT

Article Info
Volume 7, Issue 3
Page Number: 280-286

Publication Issue :
May-June-2021

Article History
Accepted : 15 May 2021
Published : 22 May 2021

In recent times, with the rise of the internet, everyone is being bombarded with tons of information and data from various sources like websites, blogs and articles, social media posts and comments, e-news portals etc. Now all these data are mostly unstructured. In this paper, the authors have tried to explore the efficiency of the cross-lingual BERT model i.e. M-BERT for text classification and named entity extraction on multilingual data. The authors have used datasets of three different languages namely: French, German and Portuguese to evaluate the model performance.

Keywords : Natural Language Processing, BERT, Transformers, Multilingual NER.

I. INTRODUCTION

According to a latest survey conducted by the Computer World Magazine, 70-80% of the entire web consists of unstructured data and text. Unstructured text has a very little utility in this digital world. We need to convert these unstructured data into structured form so that they can be used in various analytical, statistical and experimental settings. With help of entity extraction methods, we can bring some level of structure within a text document. Named entity recognition is the process in which we try to find predefined labels within a particular document.

For e.g. say there is a sentence *"Mr. Vajpayee left for Kashmir yesterday."*

If one feed the above sentence to an NER model which can identify entities: "Person" and "Location", then it will output "Mr. Vajpayee" as "Person" and "Kashmir" as "Location" and the rest of the words as "Unknown".

This is how named entity recognition works. It is mostly viewed as a classification task for sequences where each token is being classified to a particular label.

Bi-LSTM-CRF (Huang et al., 2015) proved very effective in different NER tasks due to their property of modelling temporal dependencies in a sequence well and learning the contextual representation of each token individually. The Transformer architecture (Vaswani et al., 2017) enabling parallel computation and self-attention was successful in modelling the long term contextual and positional dependencies among the words much better than the Bi-LSTMs. It gained much popularity in the NLP domain and became an active area of research. In the original paper, its efficiency was evaluated on machine translation but later it was applied on various downstream tasks by different researchers. It follows the encoder-decoder architecture and introduces multi-headed attention for significant parallelization. It has different components which governs the result that includes: word embeddings, positional encodings, attention mechanisms, feed-forward network. The positional encodings are responsible for understanding the relative positional dependency within a sentence and are generally computed using two sinusoidal equations.

Bidirectional Encoder Representation of Transformer, or BERT (Devlin et al., 2019) is a pre-trained Transformer model that has been successful in achieving state-of-the-art results in different downstream tasks like question answering, named entity recognition, sentiment analysis etc. The multilingual variant of the BERT (M-BERT) model has been trained on 104 different languages and is one of the largest NLP models. In this paper, the authors have fine-tuned the M-BERT model for evaluating on French, Portuguese and German dataset and verified the performance based on metrics such as precision, recall, f1-score and accuracy.

II. BACKGROUND

Earlier methods of named entity recognition and information extraction include rule-based approaches where there will be hand-crafted rules for identifying different labels in a sentence. They were not very effective as they lacked abilities of language and syntactical understanding, and modelling the underlying context of words.

Later, attention based Bi-LSTM-CRF (Huang et al., 2015) were incorporated for the CoNLL-2003 named-entity recognition dataset, which was quite successful and provided significant improvement in results. Although LSTM was quite successful in handling the exploding gradient problem in recurrent neural networks, they did not fully solve the problem. Thus, this model failed to learn the long term relative contextual dependency for comparatively larger sentences. Moreover, it allowed almost no scope for parallel computation and thereby making the training process very slow and wasting the GPU resource.

Inspired by the success of Bi-LSTM-CRF, TENER (Yan, et al., 2019), which is a transformer-based model, used a conditional random field (CRF) layer at the top and character embeddings to tackle out-of-vocabulary words (OOV) for named entity recognition. It performed significantly well over previously used Bi-LSTM-CRF.

III. FINE-TUNING M-BERT FOR ENTITY EXTRACTION

The BERT is a transformer based language model having only the encoder part of the transformer architecture stacked on top of one another. The BERT model has been pre-trained on two different training targets, the first one being Masked Language Modeling (MLM) and the second one is Next Sentence Prediction (NSP). In case of masked language modeling, some of the words in the input sentence are masked and at the output end the model tries to predict only the masked tokens. In BERT, 15% of the words are masked at random. The next sentence prediction is a binary classification task in which the model tries to predict whether the second sentence is a follow-up of the first sentence. For study, the authors have used the base variant of the multilingual BERT model which is pre-trained on a text corpus containing data in 104 different languages, consists of 12 encoder layers stacked on top of one another, embedding size of 768, 12 attention heads and a total of around 110 million parameters. They used the BERT tokenizer for tokenizing each sentence into a list of tokens and assigned the same label to each of the sub-word token as the word it has been broken down from, during both the training and evaluation phase.A feed forward neural network layer has been used on the top for the purpose of classifying the tokens into their corresponding labels. The authors masked out the loss generated from the padding tokens, thereby discouraging them to contribute to the learning process and only calculated the loss of the actual sentence otherwise it would have caused unnecessary training overhead.

ALGORITHMI

TRAINING OF THE NER BERT MODEL

```
Step 1: for each epoch do:
Step 2:     for each batch do:
              1. Forward pass of the BERT model
              2. Forward pass of the Feed-forward
                 layer
              3. Compute the loss
              4. Backward pass of the Feed-
                 forward layer
              5. Backward pass of the BERT model
              6. Update parameters
Step 3:  end for;
Step 4:  for each batch do:
              1. Forward pass of the BERT model
              2. Forward pass of the Feed-forward
                 layer
              3. Compute the loss
              4. Compute the F1-score
Step 5:  end for;
Step 6: end for;
```

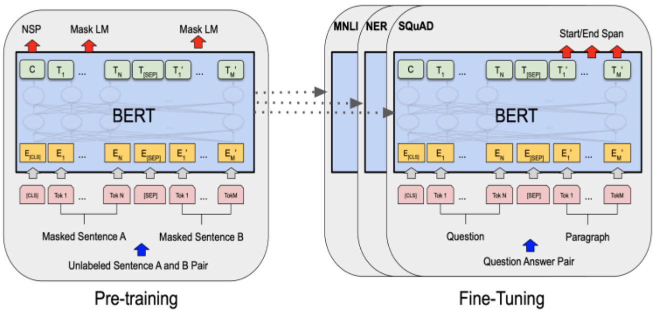

Figure 1. Fine-tuning BERT in different task (Devlin et al., 2019).

IV. RESULTS AND DISCUSSION

The authors have trained the model on 3 different multilingual dataset namely; WikiANN dataset for French NER, GermEval2014 dataset for German NER and leNER-Br legal document dataset for Portuguese NER.

TABLEII

TRAINING AND EVALUATION SIZE OF THE DATASETS

Dataset Name	Training Size	Evaluation Size
WikiANN	30000 sentences	10000 sentences
GermEval2014	24001 sentences	5099 sentences
leNER-BR	7827 sentences	1389 sentences

The leNER-Br dataset is quite imbalanced with very less samples of “LOC” and “TIM” labels.

The authors did a train-test split of 0.2, which means that 80% of our train data was used for training the model and the rest 20% was used for validation after each epoch.

The different label distribution of the three different datasets shown below with the help of pie charts:

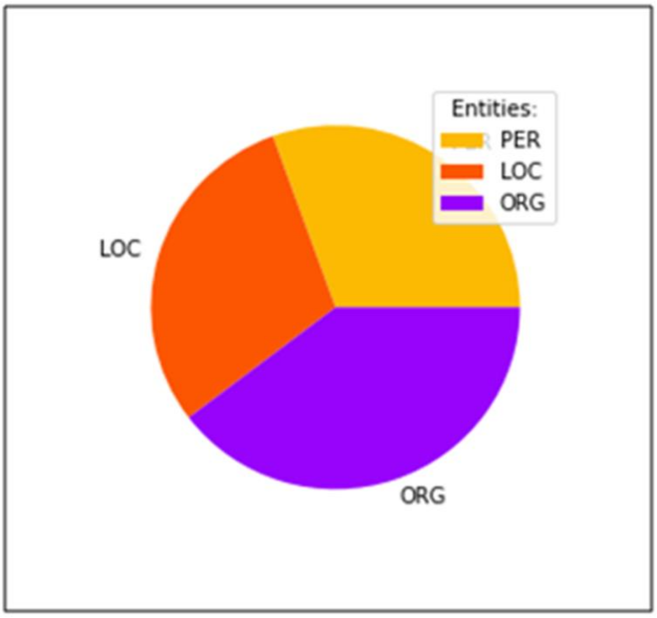

Figure 2. Merged label distribution of WikiANN training dataset.

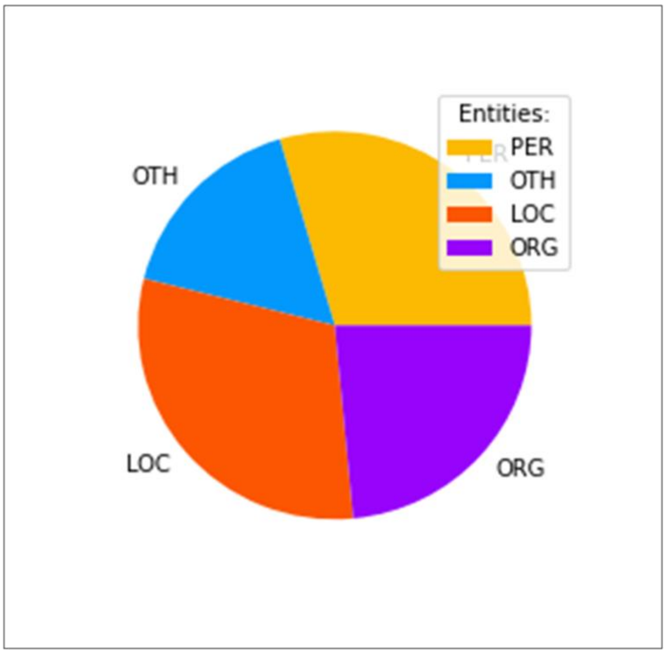

Figure 3. Merged label distribution of GermEval2014 training dataset.

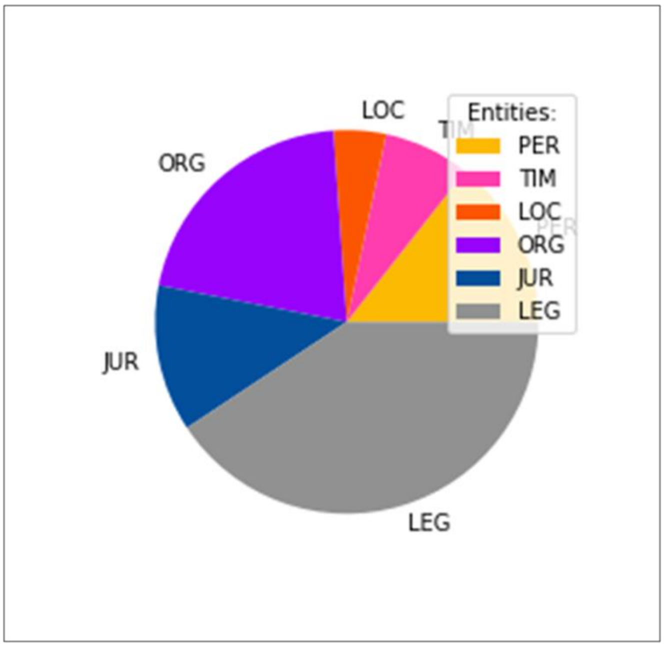

Figure 4. Merged label distribution of leNER-Br training dataset.

For each of the models they have used a maximum sequence length of 128 and have trained them in Google Colab (GPU) environment. The models have been trained on 3 epochs. The training batch size was 64 and the validation batch size was 32. A dropout of 0.1 has also been used.

The loss curve of the three different models are shown below:

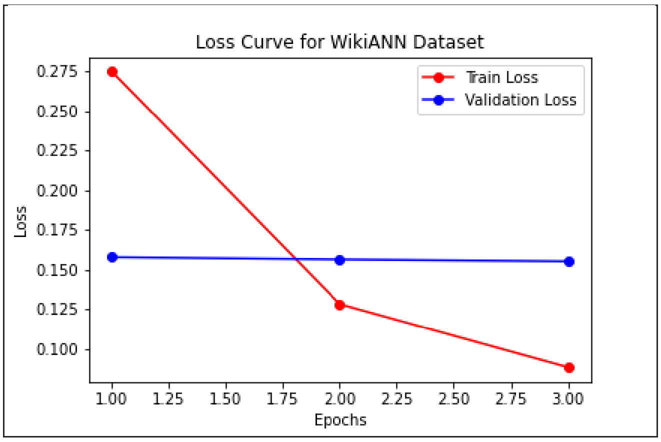

Figure 5. The loss curve for the French NER model.

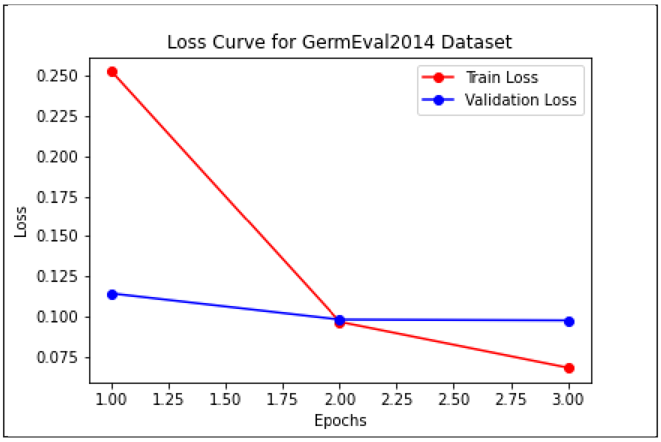

Figure 6. The loss curve for the German NER model.

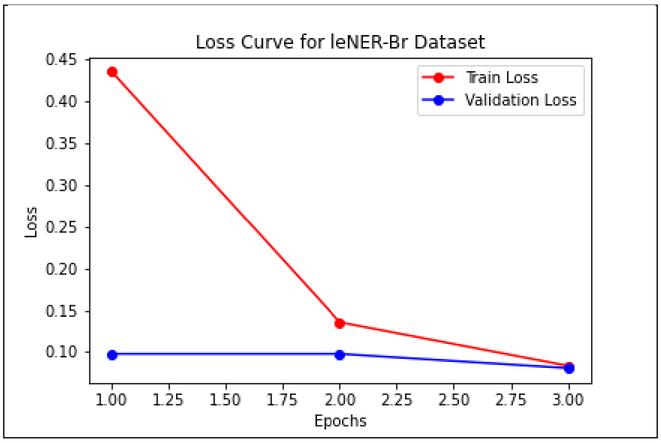

Figure 7. The loss curve for the Portuguese NER model.

Each of the three named entity recognition model was evaluated on the three different accuracy metrics:

1. **Precision**: It is the percentage of the positive identifications that were actually correct.

   Precision = TP / (TP + FP)

2. **Recall**: It is the percentage of the actual positives that were identified correctly.

   Recall = TP / (TP + FN)

3. **F1-score**: It is the harmonic mean of the precision score and recall score.

F1-score = 2 * Precision * Recall / (Precision + Recall)

The accuracy plot for the three different modelsareshown below:

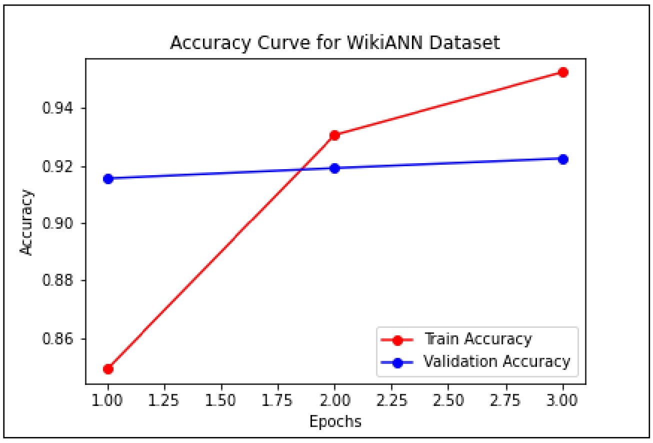

Figure 8. The accuracy curve for the French NER model

The BERT NER model performed significantly good for the WikiANN French dataset and yielded an average f1-score above 90%. The main reason for this is the dataset was well balanced with good amount of samples for each labels.

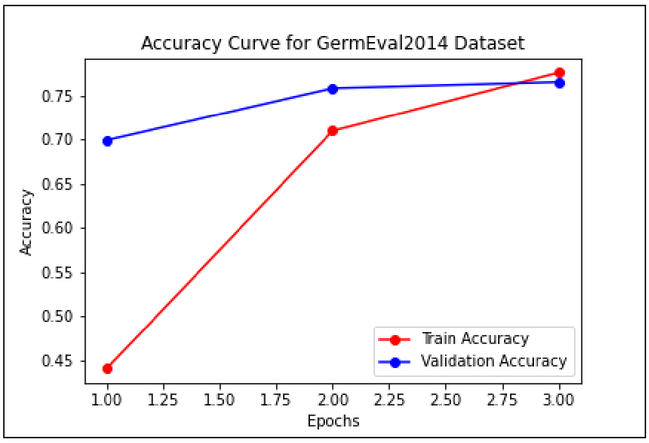

Figure 9. The accuracy curve for the German NER model

With the GermEval2014 dataset, the model yielded an f1-score of 81% and 83% for “PER” and “LOC” tags respectively but didn’t do well for the “OTH” tag.

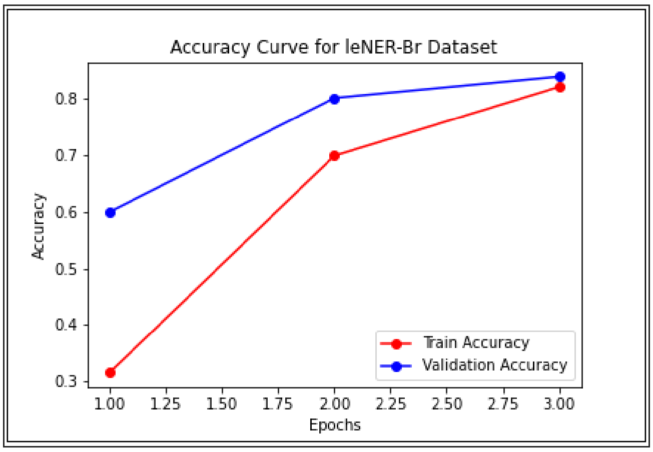

Figure 10. The accuracy curve for the Portuguese NER model

The Portuguese NER model did learned to classify significantly well for the “PER” entity and “LEG” entity but was not able to provide good results with the “LOC” whose f1-score was only 38%.

One important observation that one can derive here is that all the three different model were able to classify the “Person” tag very efficiently. This behaviour is because of the high density of the “PER” labels in each of the datasets.

TABLE II

EVALUATION RESULT OF THE FRENCH NER MODEL

Entity Name	**Precision**	**Recall**	**F1-score**
PER	0.95	0.95	0.95
LOC	0.92	0.92	0.92
ORG	0.88	0.91	0.90

TABLE III

EVALUATION RESULT OF THE GERMAN NER MODEL

Entity Name	**Precision**	**Recall**	**F1-score**
PER	0.85	0.81	0.81
LOC	0.83	0.79	0.83

ORG	0.77	0.77	0.78
OTH	0.59	0.54	0.55

TABLE IV
EVALUATION RESULT OF THE PORTUGUESE NER MODEL

Entity Name	Precision	Recall	F1-score
PER	0.87	0.92	0.90
LOC	0.38	0.43	0.38
ORG	0.77	0.80	0.79
TIM	0.84	0.76	0.80
JUR	0.76	0.74	0.74
LEG	0.91	0.86	0.88

One of the problem that has been noticed is that the model is failing to identify entities when they are intentionally clubbed together without whitespace and fed as input. Due to not having enough samples of some labels in the German and the Portuguese dataset the precision, recall and f1-score came down significantly which affected the overall accuracy score of the models but still it is above 70% despite the skewness of the datasets which is quite good. As there were significantly enough samples for "PER" tag in nearly all the datasets, the model learned to predict the tag very well with an overall accuracy of above 85%.

V. CONCLUSION

In this paper, the authors have tried to explore a BERT based transfer learning approach to effectively label words in a sentence and thereby extract relevant information from a given text, and also understand the limits of the cross lingual sequence understanding capabilities of the BERT model. This method has worked gracefully for the used multilingual datasets and has provided a satisfactory outcome. One of the major drawback is that there is a lack of appropriate dataset for testing the approach. Most of the dataset that are available are either veryskewed i.e. concentration of some of the labels are very high compared to the others or there are not enough labels to give a good prediction. But still this model was quite successful in giving good results mostly because of the size of this pre-trained model. In the future,the authors would like to extend this approach and incorporate Bidirectional Long Short Term memory or Bi-LSTM with CRF to understand the effectiveness of the approach even further.

VI. REFERENCES

[1]. Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", Google AI Language, 24 May, 2019, arXiv:1810.04805.

[2]. Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, Andrew McCallum, "Linguistically-Informed Self-Attention for Semantic Role Labeling", Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 5027–5038, , October 31 - November 4, 2018.

[3]. Hang Yan, Bocao Deng, Xiaonan Li, XipengQiu, "TENER: Adapting Transformer Encoder for Named Entity Recognition", 2019, arXiv:1911.04474.

[4]. Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li, "A Survey on Deep Learning for Named Entity Recognition", IEEE Transactions on Knowledge and Data Engineering", 2020

[5]. Andrea Galassi, Marco Lippi, Paolo Torroni, "Attention in Natural Language Processing", IEEE Transactions on Neural Networks and Learning Systems, 20 Aug, 2020, arXiv:1902.02181.

[6]. Ashish Vaswani, Noam Shazeer, Niki Parmar, JakobUszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, IlliaPolosukhin, "Attention Is All

You Need", 31st Conference on Neural Information Processing Systems, 2017, arXiv:1706.03762.

[7]. Zhiheng Huang, Wei Xu, Kai Yu, "Bidirectional LSTM-CRF Models for Sequence Tagging", 2015, arXiv:1508.01991.

[8]. Guillaume Lample, Alexis Conneau, "Cross-lingual Language Model Pretraining", 2019, arXiv:1901.07291

VII. AUTHOR'S PROFILE

1. **Dr. AsokeNath** is working as Associate Professor in the Department of Computer Science, St. Xavier's College (Autonomous), Kolkata. He is engaged in research work in the field of Cryptography and Network Security, Steganography, Green Computing, Big data analytics, Li-Fi Technology,Mathematical modelling of Social Area Networks, MOOCs etc. He has published 253 research articles in different Journals and conference proceedings.

2. **Mr. Rahul Gupta** is a final year student of M.Sc. Computer Science at St. Xavier's College (Autonomous), Kolkata. He has keen interest in the fields of Artificial Intelligence, Computational Intelligence and Data Science like Machine Learning, Deep Learning, Natural Language Processing etc. and is very passionate about working in real world projects surrounding these topics. Apart from these, he also holds interest in Full-stack Web Development and Web Designing.

3. **Ms. DebapriyaKandar** is currently a final year post-graduate student of Computer Science at St. Xavier's College (Autonomous), Kolkata. Her interest lies in the fields of Front-end Web Development, Web Designing, Data Science and Computer Networking and thereby bringing new ideas to life.

Cite this article as :

Asoke Nath, Debapriya Kandar, Rahul Gupta, "An Efficient Cross-Lingual BERT Model for Text Classification and Named Entity Extraction in Multilingual Dataset", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 7 Issue 3, pp. 280-286, May-June 2021. Available at doi : https://doi.org/10.32628/CSEIT217353

Journal URL : https://ijsrcseit.com/CSEIT217353