MyanmarBERT:Myanmar Pre-trained Language Model using BERT

Saw Win
Natural Language Processing Lab
Universiy of Computer Studies
Yangon, Myanmar
sawwin@ucsy.edu.mm

Win Pa Pa
Natural Language Processing Lab
University of Computer Studies
Yangon, Myanmar
winpapa@ucsy.edu.mm

Abstract— **Myanmar language is a low-resource language as well as obtaining large-scale cleaned data for natural language processing(NLP) tasks, it is challenging and expensive with the progress in NLP. Deep learning has boosted the development of pre-trained language model has led to significant performance gains. Despite their popularity, the majority of available models have been either trained on English data or multi-language data concatenation. This makes very limited practical use of such models, in all languages except English. Currently, monolingual pre-trained language models based on Bidirectional Encoder Representations from Transformers (BERT) show that their performance outperforms multi-lingual models in many downstream NLP tasks, under same configurations. However, a large monolingual corpus and monolingual pre-trained language model for Myanmar language are not available publicly yet. In this paper, we introduce a large monolingual corpus called MyCorpus and also release Myanmar pre-trained language model(MyanmarBERT) based on BERT. Myanmar NLP tasks such as part-of-speech (POS) tagging and named-entity recognition (NER) have been used for evaluation on MyanmarBERT and Multilingual BERT(M-BERT). The comparative results over these two models are presented. MyanmarBERT will be useful for researchers working on the Myanmar NLP and pre-trained model is available at http://www.nlpresearch-ucsy.edu.mm/mybert.html.**

Keywords—BERT, Pre-trained Language Model, Named Entity Recognition, POS tagging, Myanmar Language

I. INTRODUCTION

Currently, Myanmar NLP is in the developing stage, and until now, well-prepared resources necessary for Myanmar NLP research has not been sufficiently available. The lack of resources is the main issue in resolving new NLP research in the Myanmar language. Myanmar language has few resources compared to other languages. Myanmar language has both rich morphology and ambiguity and complex as well. Besides, its writing structure is the free order, makes the data cleaning a complex process that Myanmar NLP researchers spend a lot of time on processing steps. To fill the gas, we released a pre-trained language model, MyanmarBERT and will release large cleaned text corpus called MyCorpus for future NLP researchers.

Nowadays transformers based pre-trained language models are enhancing NLP research areas as well as transfer learning is a popular in deep learning technique for reusing a model trained on a related predictive modeling problem. Currently, BERT[1] is the most commonly used pre-trained model and its derivatives are based on the transformer architecture[2] and also Multilingual BERT(M-BERT)[1] are released by Devlin in 2019[1], it is pre-trained on the connection of monolingual Wikipedia corpora as a single language model. While most work on BERT models has focused on high resource languages, particularly English, several recent efforts have introduced multilingual models that can be fine-tuned to tackle tasks in a wide range of languages. However, in particular for low-resourced languages, we still lack a detailed understanding of the capabilities of these models. While M-BERT has been shown to have a remarkable ability to generalize across languages[3], several studies have also demonstrated that monolingual BERT models, where available, can notably outperform M-BERT. And also Natural language processing (NLP) tasks have been enhanced by fine-tuned versions of BERT and BERT-derived models such as PhoBERT[4], BERTje[5], Finnish BERT[6], CamemBERT[7], etc. So BERT, M-BERT[1], and other monolingual language models motivate to release MyanmarBERT pre-trained language model.

Compared to a new MyanmarBERT, M-BERT[1] is thoroughly evaluative on a range of tasks such as POS tagging and NER. This article describes the following contributions:

a) We introduce a large monolingual corpus, MyCorpus and the first monolingual pre-trained language model, MyanmarBERT based on BERT for NLP researchers in Myanmar Language.

b) The performance of fine-tuning on our model versus M-BERT model has been investigated.

c) We evaluate MyanmarBERT on two downstream tasks: POS tagging and named entity recognition in Myanmar Language.

II. RELATED WORKS

In 2019, Devlin[1] released BERT is a contextualized word representation model that is based on a masked language model and pre-trained using bidirectional. Current techniques use unidirectional language models to learn general language representations and this restricts the use of architectures that can be applied during pre-training. BERT model addresses this unidirectional constraint by randomly masking some of the tokens from the input using a masked language model (MLM) for the objective of pre-training. Besides, this model also using next sentence prediction task that conjointly pre-trains text-pair representations. BERT is the first fine tuning based representation model and reduces the need for many specific architectures with a highly engineering task and up-to-date results on many natural language processing tasks according

[1] https://github.com/google-research/bert/blob/master/multilingual.md

to the result of GLUE. They use the English Wikipedia and book corpus for pre-training corpus.

Telmo Pires[3] showed that while large lexical overlap between languages enhances transfer, M-BERT is also able to transfer between languages written in diverse scripts which have zero lexical overlap that point out for occupy multilingual representations. They also illustrate that transfer tasks best for topologically same languages, indicating that while the multilingual representation of M-BERT is able to define learned structures to new vocabularies, in order to fit a target language of different word order, it does not tend to learn systematic transformations to such structures to accommodate a destination language with distinct word order. M-BERT is a 12-layer transformer, like the original English BERT model, but instead of only being trained with an English-derived vocabulary on monolingual English corpus, it is trained with a shared word piece vocabulary on the Wikipedia pages of 104 languages.

Antti Virtanen [6] presented that the new language-specific model has been shown to systematically and obviously succeed the multilingual. Although the multilingual model generally fails to achieve the performance of previously proposed methods, Finnish BERT model evaluate new results on all corpora for all reference tasks: POS tagging, NER, and dependency parsing. They also study the application of language-specific and M-BERT models to Finnish NLP.

III. Pre-trained language model

Learning word representations from a huge volume of unlabeled text is a long-established method. While previous models Word2Vec[8], GloVe[9] targeted on studying context independent word representations, recent works have targeted on studying context dependent word representations. For example, ELMo[10] uses a bidirectional language model, during CoVe[11] applies machine translation to embed context information into word representations.

BERT is contextualized word representation model that follows on the transformer architecture [2] and it is two-way deep neural network model. Main technical contributions of BERT are using the bidirectional training of transformer to language modeling which involves two distinct methods — an encoder that applies the text input and a decoder that produces a task prediction. Since the purpose of BERT is to demonstrate a language model, only the encoder mechanism is needed. BERT uses two phases: pre-training and fine-tuning.

a) Pre-training: the model is trained on unlabeled data over distinct pre-training tasks. The BERT pre-training phase includes of two unsupervised predictive tasks, one is the masked language model (MLM) and the other is next sentence prediction(NSP). For masked language model, BERT uses a masked language model that predicts randomly masked words in a sentence, and thus can be used for learning bidirectional representations. For next sentence prediction, in order to train a model that accepts sentence interrelationships along with semantic interrelationships between words, BERT also pre-trains for a bidirectional of next sentence prediction task that can be very simple generated from any text corpus.

b) Fine-tuning: the BERT model is first initialized with the pre-trained parameters, and all of the parameters are fine-tuned using labeled data from the up-to-date NLP tasks. Although they are applied with the similar pre-trained parameters, downstream task has different fine-tuned models.

In pre-training and fine-tuning procedures of BERT, the similar architectures are used in both these tasks except from output layers. This paper will omit a detail description of BERT and refer readers to Devlin [1].

IV. Building MyanmarBERT

In this section will describe the creation of training corpus, the text preprocessing, vocabulary generation, the model architecture and training configurations to build MyanmarBERT.

A. Data Collection and Preprocessing

Plain text corpora are valuable resources for language modeling in NLP and information extraction approaches that use machine learning techniques. Open Super-large Crawled ALMAnaCH corpus(OSCAR)[2] is a large multilingual corpus currently available in 166 different languages but it is a raw text form of Myanmar Language. Data collection is easy by online medias websites and social medias but data cleaning is heavy and time-consuming in Myanmar language. Therefore, cleaned corpus is very limited for low-resource language like Myanmar language. The following steps will discuss the process of corpus creation for pre-training data.

Firstly, we collected raw data from Wikipedia and official Myanmar news media websites. These raw data are most representative in technology, history, economics, opinions, tourism, politics, short stories, sports, articles, religion, crime, health world news, and entertainment.

Secondly, important preprocessing steps such as cleaning and spell check tools are not available in Myanmar language. That we will discuss detail preprocessing steps: encoding normalization, correcting typing errors, normalizing text and word segmentation for data cleaning.

a) Encoding Normalization: Myanmar media websites are now using Unicode, but some uses Zawgyi font before 2019, which needs to be uniformly encoded with different encodings. All the collected data are converted into Unicode encoding for encoding consistency that non-Unicode font is converted into Unicode by using a Rabbit converter[3].

b) Correcting Type Errors: Currently, typing errors are corrected using in-house spell corrector for all kinds of mistyped errors.

c) Normalizing Text: Myanmar language writing style is not limited now that we found unnecessary repetitive characters are cleaned using in-house tool.

d)Word Segmentation: Words composed of single or multiple syllables are typically not separated by white spaces in Myanmar texts. Spaces are used for easier reading and generally set among phrases, but there are no accurate rules for using spaces in the language of Myanmar. So Word segmentation is needed and segmentation errors will affect the language modeling performance. Therefore, texts are segmented into words by using Myanmar word-segmenter from the UCSY-NLP[4] lab.

Finally, we split sentences on MyCorpus for using pre-training dataset. This corpus consists of 3.5M sentences with 83M tokens. There is no other available cleaned text corpus that has as much data as MyCorpus. The statistics of MyCorpus is shown in table I.

TABLE I. SATATISTIC OF MYCORPUS

Domain	***Number of Sentences***	***Number of words***
Wikipedia	866,397	20,760,214
Online media websites	2,669,397	63,000,303
Total	3,535,794	83,760,517

B. *Vocabulary generation*

BERT uses WordPiece[12] for unsupervised tokenization of the input text. The vocabulary is created such that it includes the most frequently used words or subword units. SentencePiece library[5] was used to construct MyVocab, a new WordPiece vocabulary. MyVocab contains both Myanmar and English words because Myanmar people write letters mix of Myanmar words and English words. For instance, they usually use the technical term, medical term, political term, and foreign names are used by English words. Some Myanmar sentence is meaningless if the English words are removed that keep the original sentence structure.

C. *Model and Training Configuration*

a) Model Architecture: Devin[1] released three models: $\text{BERT}_{\text{BASE}}$, $\text{BERT}_{\text{LARGE}}$ and M-BERT[1] which contain a multi-layer bidirectional Transformer. The architecture of $\text{BERT}_{\text{BASE}}$ and M-BERT is the same. MyanmarBERT follows the same architecture of $\text{BERT}_{\text{BASE}}$, MyanmarBERT model size is (L=12, H=768 and A=12) with 187 parameters, L is the number of layers as Transformer blocks, H is the number of hidden size, and A is the number of self-attention heads.

b) Training Objective: In this work, MyanmarBERT was trained using MyCorpus is smaller than any other monolingual corpus. More interestingly, we show that a relatively small training data set leads to good results. For pre-training objectives, MyanmarBERT used the same masked language model(MLM) and next sentence prediction(NSP) tasks based on original BERT.

In MLM task, given an input text sequence composed of N tokens t1, ..., tN , training data generator select 15% of tokens for possible replacement. Among those selected tokens, 80% are replaced with the [MASK] token, 10% are left unchanged and 10% are replaced by a random token. The model is then trained to predict the initial masked tokens using cross-entropy loss. MyanmarBERT also used whole-word masking, where all pieces of a word are masked together rather than selecting masked word pieces. Example of whole-word masking shows in Fig. 1.

In NSP task, the model learns to whether sentence A is actual next of sentence B from two input sentences.

[Original sentence]:

သူမ သည် လှပသော မိန်းကလေး တစ်ယောက် ဖြစ် သည်။

[Word piece masking]:

သူမ သည် [MASK] ##သော မိန်း [MASK] တစ်ယောက် ဖြစ် သည်။

[Whole word masking]:

သူမ သည် [MASK] [MASK] တစ်ယောက် ဖြစ် သည်။

Fig. 1. Example of whole word masking

c) Training Setting: MyanmarBERT follows the same hyper-parameters setting of pre-training $\text{BERT}_{\text{BASE}}$, however, we use different sentence length and batch size. This model is trained until the training loss stops reducing and training time takes more than four weeks on *Tesla K80 GPU*. Table II describes training data, training objective and hyper-parameters of MyanmarBERT.

TABLE II. TRAINING SETTING OF *MYANMARBERT*

Size of Training Data	1.3G
Pre-training objective	MLM and NSP
Tokenizer	SentencePiece[5]
Hyper-parameters	learning rate: 2e-5, β1: 0.9, β2: 0.999, weight decay: 0.01, batch size: 64, sequence length:64 tokens

V. FINE-TUNING FOR POS TAGGING AND NER

In this section, we present MyanmarBERT and M-BERT fine-tuning results on two NLP tasks: POS tagging and NER and compare the result of the M-BERT and MyanmarBERT models on these two downstream tasks.The overall process of fine-tuning on NER and POS tagging using MyanmarBERT is illustrated in Figure 2. For fine-tuning, we use a batch size of 16, learning rate 2e-5, and tuned the number of training epoch from 3 to 4 and saved the model with the best performance. Fine-tuning on NER took 7 hours and fine-tuning on POS tagging took nearly 2 hours by training data size. For the evaluation on of NER and POS tagging, the accuracy of entity level F1 score and confusion matrix are used.

[2] https://oscar-corpus.com

[3] https://www.rabbit-converter.org/Rabbit/

[4] http://www.nlpresearch-ucsy.edu.mm/

[5] https://github.com/google/sentencepiece

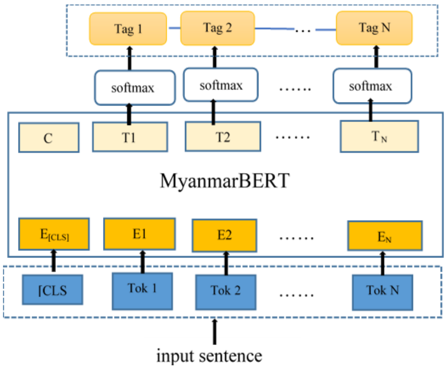

Fig. 2. Illustration of Fine-tuning MyanmarBERT on POS tagging and NER Tasks. [CLS] is a special symbol added in front of every input sentence and E is token embedding.

A. *Fine-tuning MyanmarBERT for POS Tagging*

POS tagging task is the attempt of correctly classify each token within a given set of grammatical categories (abb, adj, adv, conj, n, num, part, ppm, pron, etc). For fine-tuning POS tagging dataset on MyanmarBERT and M-BERT, we use mypos corpus[13] published from UCSY-NLP lab[4] and this corpus contains 11,000 sentence and 234,802 words with 16 POS taggings which show in table III. For fine-tuning POS tagging dataset, 9000 sentences for the training, and respectively 900 and 1100 sentences for the development and test sets.

TABLE III. POS TAG SET

N0.	*POS tag*	*Definition*
1	abb	Abbreviation
2	adj	Adjective
3	adv	Adverb
4	conj	Conjunction
5	fw	Foreign Word
6	num	Number
7	int	Interjection
8	n	Noun
9	part	Particle
10	part_neg	Negative Particle
11	ppm	Post Positional Marker
12	pron	Pronoun
13	punc	Punctuation
14	sb	Symbol
15	tn	Text Number
16	v	Verb

B. *Fine-tuning MyanmarBERT for Named Entity Recognition*

NER is the process of automatically tagging, identifying or labeling different named entities (NE) in text following the predefined sets of NE categories such as person, location and organization and so on. For fine-tuning NER dataset on MyanmarBERT and M-BERT, we use NER datasets [14] released from UCSY NLP lab, NER corpus contains 60,500 sentences and 174,133 NE samples such as PNAME, LOC, ORG, RACE, TIME and NUM. This version is provided with a standard split representing 58246 sentences for the training corpus, and respectively 1133 and 1121 sentences for the development and test sets.

C. *POS Tagging and NER Experiment Results*

In this work, we present the fine-tuning results of POS tagging and NER tasks. NER results of MyanmarBERT and M-BERT models are shown in Fig. 3 and 4. And, the POS tagging results of each model are shown in Fig. 5 and 6. The comparative results of each model are reported in table IV and the best F1 scores are in bold. Firstly, we observe that NER F1 score of MyanmarBERT was nearly equal to M-BERT.

Secondly, we notice that M-BERT, which was pre-trained on multilingual corpus is completely effective, but the POS tagging results of MyanmarBERT achieves F1 scores 0.048 higher.

Finally, this article explores the MyanmarBERT well learns of POS tagging task than M-BERT but less obvious in NER task.

TABLE IV. COMPARISON OF MYANMARBERT WITH M-BERT ON POS TAGGING AND NER TASKS IN TERMS OF F1-SCORE

Task	*MyanmarBERT model*	*M-BERT model*
NER	0.933	**0.934**
POS tagging	**0.914**	0.866

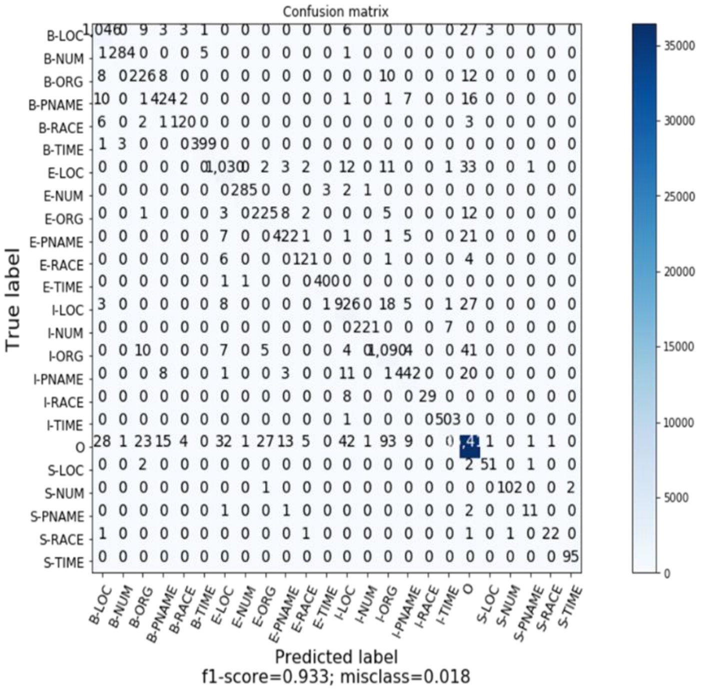

Fig. 3. Confusion matrix on NER task using MyanmarBERT model

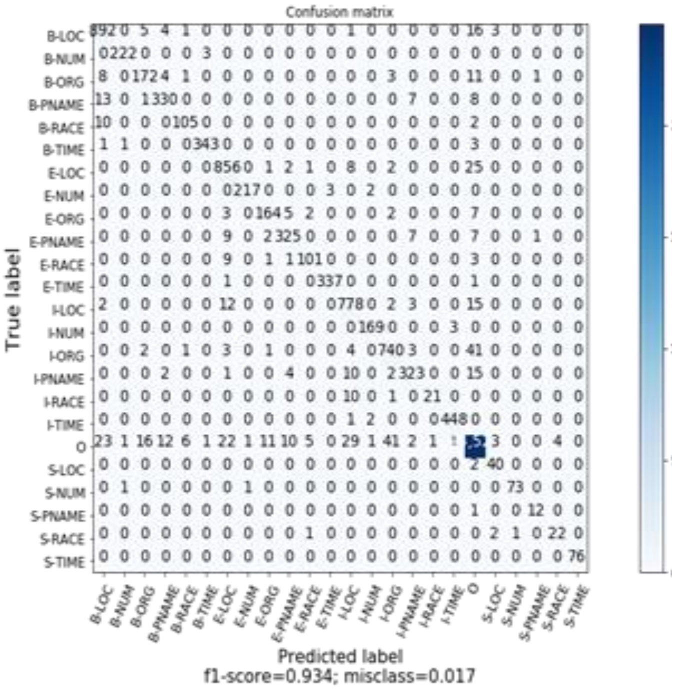

Fig. 4. Confusion matrix on NER task using M-BERT model

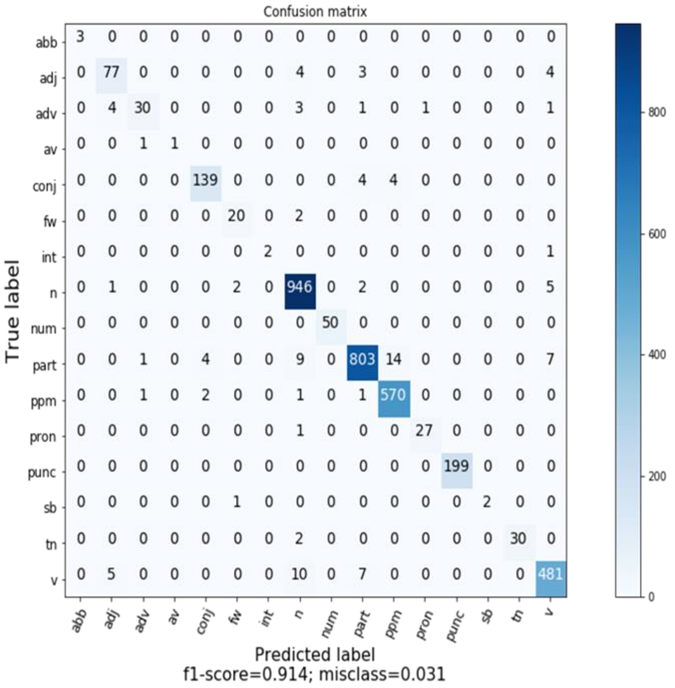

Fig. 5. Confusion matrix on POS tagging using MyanmarBERT

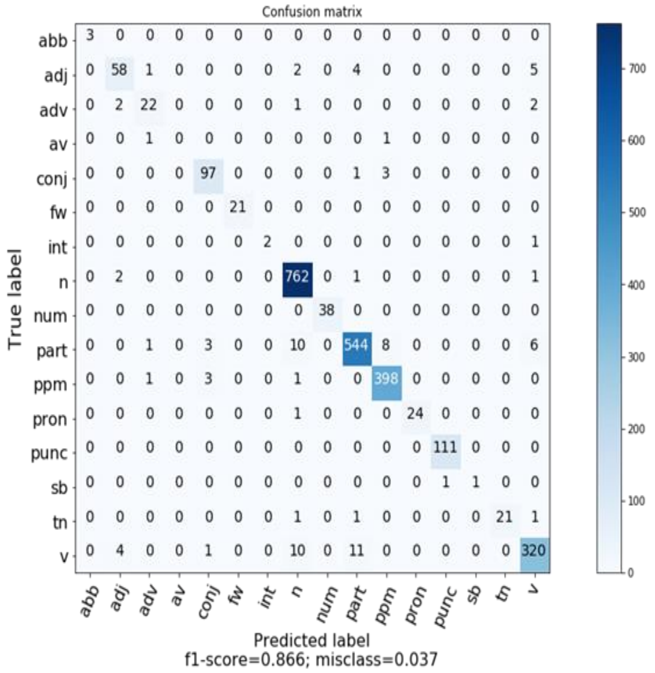

Fig. 6. Confusion matrix on POS tagging using M-BERT

VI. CONSLUSION

Language Modeling is an essential part of many language understanding systems. In this paper, the first monolingual large corpus and Myanmar pre-trained language model have been introduced. MyanmarBERT is adapted from BERT with evaluate two downstream tasks: POS tagging and NER. In particular, these comparative results showed that MyanmarBERT slightly outperformed on POS tagging task than M-BERT and it is less obvious in NER task.

REFERENCES

[1] Jacob Devlin,Ming-Wei Chang,Kenton Lee, Kristina Toutanova,” BERT: pre-training of deep bidirectional transformers for language understanding”, 24 May 2019.

[2] Ashish Vaswani, et al., ”Attention Is All You Need”, 6 DEC 2017.

[3] Telmo Pires, Eva Schlinger, Dan Garrette, ”How multilingual is multilingual BERT?”, arXiv:1906.01502v1 [cs.CL] 4 Jun 2019.

[4] Dat Quoc Nguyen and Anh Tuan Nguyen,” PhoBERT: pre-trained language models for Vietnamese”, arXiv:2003.00744v1 [cs.CL] 2 Mar 2020

[5] Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim,” BERTje: a dutch BERT model”, arXiv:1912.09582v1 [cs.CL] 19 Dec 2019

[6] Antti Virtanen, Jenna Kanerva Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo “Multilingual is not enough: BERT for Finnish”, arXiv:1912.07076v1 [cs.CL] 15 Dec 2019.

[7] Louis Martin,Benjamin Muller,Pedro Javier Ortiz Suárez,Yoann Dupont,Laurent Romary,Éric Villemonte de la Clergerie,Djamé Seddah,Benoît Sagot, ”CamemBERT: a Tasty French language model”, 21 May 2020.

[8] Tomas Mikolov, Ilya Sutskever , Kai Chen, Greg Corrado, Jeffrey Dean , “Distributed representations of words and phrases and their compositionality”. In: Burges,C.J.C. (eds.), Advances in Neural Information Processing Systems 26, Curran Associates, Inc.,2013, pp.3111–3119.

[9] Jeffrey Pennington, Richard Socher, Christopher D. Manning, Glove: “ Global vectors for word representation” In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar. pp. 1532–1543. Association for

Computational Linguistics. https://www.aclweb.org/anthology/D14-1162.

[10] Peters,M.E. et al. “Deep contextualized word representations” arXiv:1802.05365v2 [cs.CL] 22 Mar 2018

[11] Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher “Learned in Translation: Contextualized Word Vectors”, 20 Jun 2018

[12] Yonghui Wu, et al., “Google’s neural machine translationsystem: bridging the gap between human and machine translationar”, Xiv:1609.08144v2 [cs.CL] 8 Oct 2016

[13] Khin War War Htike,et al.,”Comparison of six POS tagging methods on 10K sentences Myanmar language (Burmese) POS Tagged Corpus”

[14] Hsu Myat Mo and Khin Mar Soe, “Myanmar named entity corpus and its use in syllable-based neural named entity recognition,", April 2020, pp. 1544~1551.