Incorporating Transformer Models for Sentiment Analysis and News Classification in Khmer

Md Rifatul Islam Rifat[1(✉)] and Abdullah Al Imran[2]

[1] Rajshahi University of Engineering and Technology, Rajshahi, Bangladesh
[2] American International University-Bangladesh, Dhaka, Bangladesh

Abstract. In recent years, natural language modeling has achieved a major breakthrough with its sophisticated theoretical and technical advancements. Leveraging the power of deep learning, transformer models have created a disrupting impact in the domain of natural language processing. However, the benefits of such advancements are still inscribed between few highly resourced languages such as English, German, and French. Low-resourced language such as Khmer is still deprived of utilizing these advancements due to lack of technical support for this language. In this study, our objective is to apply the state-of-the-art language models within two empirical use cases such as Sentiment Analysis and News Classification in the Khmer language. To perform the classification tasks, we have employed FastText and BERT for extracting word embeddings and carried out three different type of experiments such as FastText, BERT feature-based, and BERT fine-tuning-based. A large text corpus including over 100,000 news articles has been used for pre-training the transformer model, BERT. The outcome of our experiment shows that in both of the use cases, a pre-trained and fine-tuned BERT model produces the outperforming results.

Keywords: Khmer · Deep learning · Sentiment analysis · News classification · Transformer models

1 Introduction

Last two decades have seen a growing trend towards a field named Natural Language Processing (NLP) that is concerned with the interactions between computers and human (natural) languages. Numerous applications of NLP have already been implemented in our real-life such as Sentiment Analysis, Question Answering, Chatbots, Machine Translation, Speech Recognition and so on.

In this study, we have aimed to work with the Khmer language, the official language of Cambodia. According to Wikipedia, Khmer is the second most widely spoken Austroasiatic language in the world with approximately 16 million speakers. Unlike English, French, and German, over the past two decades, few studies have been conducted for the Khmer language related to NLP. That

D. Mohaisen and R. Jin (Eds.): CSoNet 2021, LNCS 13116, pp. 106–117, 2021.
https://doi.org/10.1007/978-3-030-91434-9_10

is why, there are few NLP resources available for the Khmer language that have been built based on the traditional techniques.

Considering the inadequacy of Khmer NLP resources, in this work, we have provided a pre-trained BERT model for the Khmer language along with two downstream benchmarks for Sentiment Analysis and News Category Classification. BERT [1], which stands for Bidirectional Encoder Representations from Transformers, developed by Google, is the first fine-tuning based representation model. Through this architecture, the models can learn the context of a sentence from both of the direction (left and right). This model outperforms many task-specific architectures as well as achieves state-of-the-art performance on a large suite of sentence-level and token-level tasks.

Apart from English, Google also provided a multilingual BERT model that supports for 104 languages of different countries but unfortunately Khmer has not been included in the list of supported languages. This might be due to the unique grammatical architecture of the Khmer. Unlike in Latin languages such as English or French, Khmer has more complex structure to word form and permits different orders of the character components that lead to the same visual representation. So far, there is no available BERT pre-trained model, which has been trained exclusively using the Khmer language, based on which the Cambodian people can develop NLP models for different applications. That is why, we aim to provide a pre-trained BERT model which will definitely add a new dimension to Khmer language and its computational advancement.

We have pre-trained our BERT model with a large corpus composed of Khmer online newspapers and also performed some downstream tasks such as Sentiment Analysis and News Classification by fine-tuning our pre-trained BERT model. Both of these two applications, Sentiment Analysis and News Classification, have a significant impact on business. More specifically, for a consumer-centric industry, which focuses on making a positive experience by maximizing the quality of services or products, the Sentiment Analysis model can help to figure out the customers' demand based on the product or service reviews. In contrast, every day a huge number of employees of the news industry spend their time arranging the news based on the category of the news. By the utilization of our News Classification model, the news industry of Cambodia will be able to save their time as well as the manpower.

In order to validate the performance of our BERT model, we have applied 4 different evaluation metrics namely accuracy, precision, recall, and f1-score. In addition to this, we have included a comparative analysis with a context-free word embeddings named FastText. For both of the applications, we have conducted the experiment through three distinct approaches such as FastText, BERT Feature-based, and BERT Fine-tuning. Significantly, the BERT Fine-tuning approach outperforms the other two approaches by accuracy and f1-score.

2 Background Study

In recent years, there has been an increasing amount of literature on language modeling. But, there is a relatively small body of literature that is concerned

with the Khmer language. Among these studies, the authors proposed some basic resources of Khmer language such as character recognition, word segmentation, and POS tagging. This section briefly discusses some of the latest and relevant background studies.

The authors in [2] proposed techniques for recognizing the character and text from Khmer ancient palm leaf documents. For isolated character recognition, they applied different types of neural network architectures such as CNN, LSTM-RNN and found the outperforming result with the combination of both convolutional and recurrent architectures. On the other hand, they used both one-dimensional and two-dimensional RNN to recognize the word/text image patches of variable length and found that two-dimensional RNN performs better than one-dimensional.

For building lemmatizers as well as extracting relation between words, POS tagging is one of the essential tasks. Considering the sentence structure and word classes' ambiguties of Khmer, in 2007, the eminent authors [3] modified the applying rule algorithms and proposed a supervised transformation-based POS tagger. Moreover, to handle the unknown words, they proposed a hybrid approach which combines rule-based and tigram approach. On the other hand, based on the Conditional Radom Fields (CRFs), in 2017, the authors [4] proposed a new approach to Khmer POS tagging by incorporating 5 groups of features such as contextual, morphological, word-shape, named-entity, and lexical features.

Like Khmer, for each of the languages, word segmentation is an integral part of all of the language modeling related tasks. Unlike English, however, there are no spaces in the writing system of the Khmer language to separate the words that make it more complex. The authors in [5–7] have worked on the word segmentation of the Khmer language. In paper [5], the authors applied Maximum Matching Algorithm and a Khmer manual corpus to make word boundaries in each sentence. Then intended to solving the unknown words, they created 21 grammar rules based on the principle of Khmer grammar books. On the other hand, to reduce the frequency of dictionary lookup and Khmer text manipulation tweaks, the paper [6] presented a study on Bi-directional Maximal Matching (BiMM). They implemented their study for Khmer word segmentation by focusing on both Plaintext and Microsoft Word document. In 2015, Chea and co-workers [7] presented a word segmenter for the Khmer language based on a supervised CRF segmentation method. Their proposed segmenter outperformed the baseline in terms of precision (=0.986), recall (0.983), and f-score (=0.985) by a wide margin. Surprisingly, they obtained substantial increases in the BLEU score of up to 7.7 points, relative to a maximum matching baseline, in their evaluation in a statistical machine translation system.

For many aspects of machine translation, the parallel corpus is essential. But, the existing parallel corpus contains some problems such as difficulty in obtaining, narrow fields, small quantity, and poor timeliness. To overcome this insufficiency of bilingual parallel corpus of low resource languages, in a recent paper [8], published in 2020, the prominent authors proposed a parallel fragment extraction method based on the Dirichlet process. On the basis of the empirical

results, the authors concluded that the method based on the Dirichlet process is better than that based on the LDA model.

From the above discussion, it is obvious that there are few studies related to the Khmer language where the researchers applied traditional techniques on different NLP tasks. However, there does not exists enough NLP resources for the Khmer language that has been built based on state-of-the-art techniques. This is where this study will play a vital role in mitigating the gap by providing a state of the art resource for the Khmer language.

3 Data Description

For the three different experiments, such as creating the pre-trained BERT model for Khmer language, sentiment analysis, and news classification, we have collected three different corpora. The entire data collection process has been illustrated in the following Fig. 1.

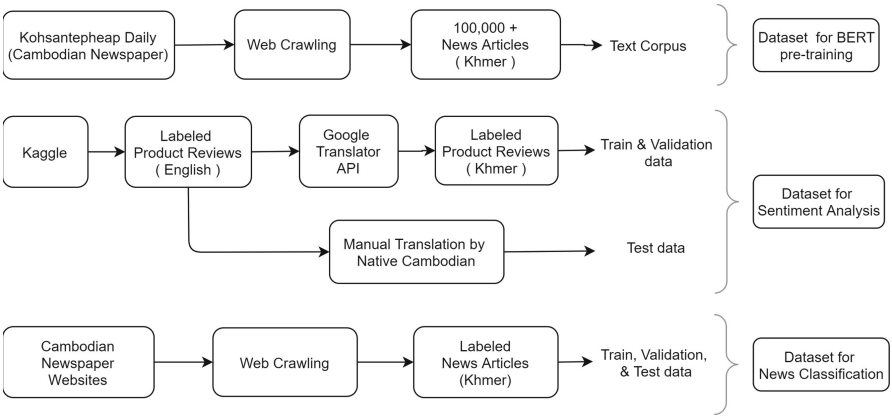

Fig. 1. Data collection pipeline.

3.1 Dataset for BERT Pre-training

In NLP, the concept behind pre-training refers to train a model with a large corpus for learning the underlying knowledge from the data. Then, the pre-trained model is used to initialize the model parameters of various downstream tasks. For creating the pre-trained BERT model of Khmer language, we have collected a large corpus from a popular Cambodian newspaper named Kohsantepheap Daily [9]. The daily newspaper Koh Santepheap was founded in 1967 by Chou Thany. The corpus contains more than 100,000 news articles that have been scraped from the official website.

3.2 Dataset for BERT Fine-Tuning

For two different downstream tasks, we have prepared 2 different datasets for fine-tuning each of the tasks.

1. **Dataset for Sentiment Analysis:** The Sentiment Analysis dataset contains 10710 instances including 5367 for positive, and 5343 for negative sentiment. For model testing purpose, we have collected additional 400 samples validated by a native Cambodian.
2. **Dataset for News Classification:** We have collected the News Classification data through scraping several Cambodian newspaper websites. The dataset includes 7418 instances with 8 different news categories as class labels. For training purpose we have used 70%(=5192), for validation 21%(=1558), and for testing 9%(=668) instances from the whole dataset.

4 Data Preprocessing

One of the crucial phases of solving a real-world problem using NLP is to preprocess the raw dataset properly. Usually, the NLP pipeline includes a bunch of tasks such as removing the punctuation, tokenization, removing the stop words, and so on. In this study, we have arranged this phase according to the form of our datasets that includes some basic preprocessing followed by word segmentation.

4.1 Basic Preprocessing

To keep only the letters and digits of the Khmer and English languages, we have applied some regular expressions on raw data using the Unicode. For basic Khmer characters, the Unicode block is U+1780 – U+17FF. After that, we have removed the punctuations of the Khmer language. The most commonly used punctuations used in the Khmer language are ។, ៕, ៖, ។, ។ល។, ๏, ๛ and their Unicode range is U+17D4 – U+17DA.

4.2 Word Segmentation

Word segmentation is an essential task in every application of Natural Language Processing. In the case of Khmer language, word segmentation is not a trivial task because, unlike English, spaces are not used here to separate words. Moreover, multiple Khmer words can be joined together to build up a new Khmer word (compound word) that conveys different meaning. Some examples of Khmer compound word are given in Table 1.

Table 1. Khmer Compound Words

Examples	Explanation
ក៏ប៉ុន្តែ	ក៏ (or) + ប៉ុន្តែ (but) = ក៏ប៉ុន្តែ (but)
ផងដែរ	ផង (also) + ដែរ (also) = ផងដែរ (as well)
ទទួលខុសត្រូវ	ទទួល (receive) + ខុស (wrong) + ត្រូវ (right) = ទទួលខុសត្រូវ (be responsible for)

Another complexity in the writing system of the Khmer language is, a single sentence can be tokenized in several ways based on its meaning in the context [7]. Table 2 shows the distinct segmentation of an identical sentence.

Table 2. Two distinct segmentation of an identical sentence.

Khmer	English
ខុំ ចង់ឱ្យ < អ្នកស្តាប់ > យល់ ពី បញ្ហា នេះ	I want listener to understand this problem
ខុំ ចង់ឱ្យ < អ្នក > < ស្តាប់ > យល់ ពី បញ្ហា នេះ	I want you to listen in order to understand this problem

All these complications address many challenges in the tokenization of Khmer text. In this study, we have applied a conditional random fields (CRFs) based approach from [7] to separate the words in Khmer sentences. The authors developed a large manually-segmented corpus and also provided a set of word segmentation strategies usually used by humans.

5 Methodology

In this section, we have described the workflow of our experiment in a proper sequence. Our entire experiment can be divided into two steps: one is using the FasText and the other one using BERT. Again, BERT is composed of three distinct parts such as pre-training, fine-tuning, and feature-based. Each of these steps has been described extensively in the subsections below.

5.1 FastText

In order to make the baseline, firstly, we have conducted our experiment using the FastText word embeddings. The FastText library of the Khmer language contains 242,732 vocabs and against each of the words, it provides a vector of length 300. On the basis of these embeddings, in this phase, we have designed two DNN models for sentiment analysis and news classification.

In the architecture of a DNN model, the number of neurons in its input layer depends on the length of the feature embeddings. As the length of the feature vectors is 300, we have specified 300 neurons in the input layer for both of the models. Determining the number of hidden layers and the number of neurons in the hidden layers is a crucial task as it defines the complexity and efficiency of the network and has an enormous impact on the learning process of the model. In both the sentiment analysis and news classification model architectures, we have specified the identical number of hidden layers as well as the number of neurons in hidden layers that are shown in Table 3.

Table 3. Parameters of DNN Architectures.

Sentiment analysis			News classification		
Layers	Number of neurons	Activation function	Layers	Number of neurons	Activation function
Input Layer	300		Input Layer	300	
Dense 1	1024	tanh	Dense 1	1024	tanh
Dense 2	512	tanh	Dense 2	512	tanh
Dense 3	256	tanh	Dense 3	256	tanh
Dense 4	128	tanh	Dense 4	128	tanh
Output Layer	2	softmax	Output Layer	8	softmax

On the other hand, in the case of the classification model, the number of neurons in the output layer depends on the unique class labels. According to this terminology, we have specified the number of neurons in the output layer 2 and 8 for the sentiment analysis and news classification, respectively.

As our extracted embeddings contain both positive and negative values, thus we have applied the *tanh* as the activation function [11] of the hidden layers. The mathematical function ot *tanh* can be expressed as,

$$tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{1}$$

The output of this function ranges from -1 to 1. To minimize the losses we have applied the optimization algorithm named Adam with a learning rate of 0.001.

5.2 BERT

This phase starts with pre-training the BERT model with Khmer followed by fine-tuning and feature-based approaches for the downstream tasks. The extensive discussion of this subsections are as follows:

Table 4. Configurations of BERT model

Parameters	Values
attention_probs_dropout_prob	0.1
hidden_act	gelu
hidden_dropout_prob	0.1
hidden_size	768
initializer_range	0.02
intermediate_size	307
max_position_embeddings	512
num_attention_heads	12
num_hidden_layers	12
type_vocab_size	2
vocab_size	167896

Pre-training Model Architecture. In this phase, we have built a pre-trained model for the Khmer language using BERT. BERT's model architecture is a multi-layer bidirectional transformer encoder based on [10]. In the configurations that we have applied in our experiment, the number of hidden layers (i.e.Transformer blocks) is 12, the size of each of the hidden layer is 768, and the self-attention heads are 12. Table 4 shows the configurations of our BERT model.

Fine-Tuning Model Architecture. Fine-tuning is a supervised learning process where the weights of the pre-trained model are used as the initial weights for a new model which is being trained on a similar task. This process not only speeds up the training but also creates a state-of-the-art model for a wide range of NLP tasks.

In this study, we have fine-tuned our pre-trained BERT model for two different applications such as Sentiment Analysis and News Classification. Firstly, we have initialized the fine-tuned model with the same pre-trained model parameters for both of the downstream tasks, such as Sentiment Analysis and News Classification. Then, we have fine-tuned all of the parameters end-to-end using the corresponding task-specific labeled data. Eventually, we have incorporated an additional output layer according to the target classes. Thus, the fine-tuned models for both of the downstream tasks are different, even though they have been initialized with the same pre-trained parameters. Figure 2 represents the architecture of the models.

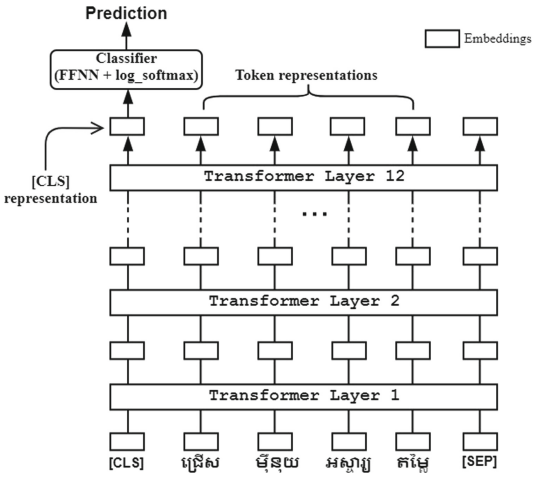

Fig. 2. Architecture of the BERT Fine-tuning Model

Feature-Based Approach. In this phase, we have extracted the feature embeddings from our pre-trained BERT model. Then, by applying these extracted features, we have designed two Deep Neural Network (DNN) models: one is for sentiment analysis and the other one is for news classification.

As the length of the BERT feature embeddings is 768, we have specified 768 neurons in the input layer for both of the models. However, we have applied 4 hidden layers for the sentiment analysis model while 5 hidden layers for the news classification model. For both of the BERT Feature-based models, the number of neurons in each of the hidden layers are shown in Table 5. According to aforementioned terminology in FastText, we have specified the number of neurons in the output layer 2 and 8 for the sentiment analysis and news classification, respectively. Like FastText, we have applied the tanh as the activation function and the Adam as the optimization algorithm in our featured-based model architectures of BERT.

Table 5. Parameters of DNN Architectures.

Sentiment analysis			News classification		
Layers	Number of neurons	Activation function	Layers	Number of neurons	Activation function
Input Layer	768		Input Layer	768	
Dense 1	1024	tanh	Dense 1	2048	tanh
Dense 2	512	tanh	Dense 2	1024	tanh
Dense 3	256	tanh	Dense 3	512	tanh
Dense 4	256	tanh	Dense 4	256	tanh
Output Layer	2	softmax	Dense 5	128	tanh
			Output Layer	8	softmax

6 Results and Analysis

In this section, we have evaluated the performance of our models, both sentiment analysis and news classification, using the classification metrics namely accuracy, precision, recall, and f1-score.

6.1 Sentiment Analysis

After the training and validation, we have tested our models with a test dataset that contains 400 instances. The obtained results for the three models of the sentiment analysis have been presented in Table 6 and Table 7.

Table 6. Accuracy of sentiment analysis models.

Accuracy			
	FastText	BERT (Feature-based)	BERT (Fine-tuning)
Training	0.79	0.73	0.83
Validation	0.78	0.71	0.83
Test	0.77	0.70	**0.81**

Table 6 shows the training, validation, and testing accuracies for the three models: FastText, BERT Feature-based, and BERT Fine-tuning of sentiment analysis. It is apparent from this table that, for each of the models, our obtained validation accuracy is very close to the training accuracy and slightly lower which indicates that each of the models has learned the underlying patterns very well from the data without overfitting. Sometimes, solely accuracy is not a good measure for the evaluation of a classification model. For this reason, we have also analyzed other metrics such as precision, recall, and f1-score during the testing phase of the sentiment analysis models that presents in Table 7.

Table 7. Precision, recall, and f1-score of sentiment analysis models.

	FastText			BERT feature-based			BERT fine-tuning		
	precision	recall	f1-score	precision	recall	f1-score	precision	recall	f1-score
Negative	0.86	0.67	0.75	0.77	0.61	0.68	0.86	0.77	0.81
Positive	0.71	0.88	0.78	0.65	0.80	0.72	0.77	0.86	0.81
macro avg	0.78	0.77	0.77	0.71	0.70	0.70	0.81	0.81	0.81
weighted avg	0.79	0.77	0.77	0.71	0.70	0.70	0.81	0.81	0.81

From both of the Table 6 and Table 7, it can be observed that the BERT Fine-tuning model outperforming the other two by the accuracy and f1-score. Another important finding is that, in the case of sentiment analysis, surprisingly, the FastText is performing better than the BERT Feature-based model.

6.2 News Classification

Similar to sentiment analysis, after the training and validation of our news classification models, we have tested our models with a dataset that contains 668 samples. The training, validation, and testing accuracies for each of the news classification models are presented in Table 8.

Table 8. Accuracy of News Classification models.

Accuracy			
	FastText	BERT (Feature-based)	BERT (Fine-tuning)
Training	0.83	0.85	0.89
Validation	0.82	0.84	0.85
Test	0.83	0.82	**0.85**

From Table 8, it can be observed that the validation accuracy is slightly lower than the training accuracy for each of the model's output. This lower differences

in training and validation accuracies indicate that each of the models has learned the underlying patterns very well from the news data without overfitting. We have obtained a good accuracy for each of the news classification models while BERT Fine-tuning model outperforms the others. Unlike sentiment analysis, BERT Feature-based model also achieved higher training and validation accuracies compared to FastText. The other metrics such as precision, recall, and f1-score have also been analyzed during the testing phase of news classification models are shown in Table 9.

Table 9. Precision, recall, and f1-score of News Classification models.

	FastText			BERT Feature-based			BERT Fine-tuning		
Class Labels	precision	recall	f1-score	precision	recall	f1-score	precision	recall	f1-score
arts-and-culture	0.90	0.93	0.92	0.93	0.91	0.92	0.95	0.91	0.93
business	0.80	0.66	0.72	0.70	0.73	0.72	0.86	0.83	0.84
health	0.80	0.90	0.85	0.77	0.88	0.82	0.74	0.92	0.82
international	0.75	0.83	0.79	0.80	0.83	0.82	0.86	0.83	0.85
national	0.86	0.88	0.87	0.84	0.87	0.85	0.90	0.89	0.89
research	0.73	0.59	0.65	0.68	0.59	0.63	0.67	0.69	0.68
service	0.86	0.89	0.87	0.90	0.78	0.83	0.89	0.76	0.82
sports-news	0.97	0.94	0.96	0.90	0.96	0.93	0.96	0.96	0.96
macro avg	0.83	0.83	0.83	0.82	0.82	0.81	0.86	0.85	0.85
weighted avg	0.83	0.83	0.83	0.82	0.82	0.82	0.85	0.85	0.85

From Table 9, it is apparent that, like sentiment analysis, BERT Fine-tuning outperforms the other two models of news classification by precision, recall, and f1-score. Interestingly, for both of the applications, the test scores of the FastText models are slightly higher than the BERT feature-based models.

7 Conclusion

The purpose of this study was to employ the state-of-the-art natural language processing techniques for Khmer language, one of the most low-resourced languages currently available. In this study, we have defined two widely used application scopes such as news category classification and sentiment analysis. Three different type of experiments such as FastText (feature-based), BERT (feature-based), and BERT (fine-tuning-based) have been conducted for both of the aforementioned downstream tasks. The experimental results show that in terms of Sentiment Analysis, BERT fine-tuning based approach outperformed the other approaches with a test accuracy of 81%. Similarly, in terms of News Classification, again BERT fine-tuning based approach stood out as the best performer with a test accuracy of 85%. In future, we would like to investigate other state-of-the-art variants of BERT such as RoBERT, DistilBERT, XLM-RoBERTa and the new giant GPT-3 for Khmer language.

References

1. Devlin, J., Chang, M. W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR, abs/1810.04805 (2018). http://arxiv.org/abs/1810.04805
2. Valy, D., Verleysen, M., Chhun, S., Burie, J.C.: Character and text recognition of Khmer historical palm leaf manuscripts. In: 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), pp. 13–18. IEEE (2018). https://doi.org/10.1109/ICFHR-2018.2018.00012
3. Sangvat, S., Pluempitiwiriyawej, C.: Khmer POS tagging using conditional random fields. In: Hasida, K., Pa, W.P. (eds.) PACLING 2017. CCIS, vol. 781, pp. 169–178. Springer, Singapore (2018). https://doi.org/10.1007/978-981-10-8438-6_14
4. Nou, C., Kameyama, W.: Khmer POS tagger: a transformation-based approach with hybrid unknown word handling. In: International Conference on Semantic Computing (ICSC), pp. 482–492. IEEE (2007). https://doi.org/10.1109/ICSC.2007.104
5. Long, P., Boonjing, V.: Longest matching and rule-based techniques for Khmer word segmentation. In: 10th International Conference on Knowledge and Smart Technology (KST), pp. 80–83. IEEE (2018). https://doi.org/10.1109/KST.2018.8426109
6. Bi, N., Taing, N.: Khmer word segmentation based on bi-directional maximal matching for plaintext and microsoft word document. In: Signal and Information Processing Association Annual Summit and Conference (APSIPA), Asia-Pacific, pp. 1–9. IEEE (2014). https://doi.org/10.1109/APSIPA.2014.7041822
7. Chea, V., Thu, Y. K., Ding, C., Utiyama, M., Finch, A., Sumita, E.: Khmer word segmentation using conditional random fields. In: Khmer Natural Language Processing, pp. 62–69 (2015)
8. Ning, S., Yan, X., Nuo, Y., Zhou, F., Xie, Q., Zhang, J.P.: Chinese-Khmer parallel fragments extraction from comparable corpus based on Dirichlet process. Procedia Comput. Sci. **166**, 213–221 (2020)
9. Koh Santepheap Daily. https://kohsantepheapdaily.com.kh/. Accessed 28 Aug 2020
10. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008. Curran Associates, Inc. (2017). http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf
11. Nwankpa, C., Ijomah, W., Gachagan, A., Marshall, S.: Activation functions: comparison of trends in practice and research for deep learning. CoRR, abs/1811.03378 (2018). http://arxiv.org/abs/1811.03378