A review of Khmer word segmentation and part-of-speech tagging and an experimental study using bidirectional long short-term memory

Sreyteav Sry[1*], Amrudee Sukpan Nguyen[1]

[1]Paragon International University, Phnom Penh, Cambodia Phnom Penh, Cambodia
[*]Corresponding author: ssry@paragoniu.edu.kh

ARTICLE INFO

DOI:10.46223/HCMCOUJS.tech.en.12.1.2219.2022

Received: March 28th, 2022

Revised: April 12th, 2022

Accepted: April 15th, 2022

Keywords:

Khmer natural language processing; long short-term memory; part-of-speech tagging; word segmentation

ABSTRACT

Large contiguous blocks of unsegmented Khmer words can cause major problems for natural language processing applications such as machine translation, speech synthesis, information extraction, etc. Thus, word segmentation and part-of-speech tagging are two important prior tasks. Since the Khmer language does not always use explicit separators to split words, the definition of words is not a natural concept. Hence, tokenization and part-of-speech tagging of these languages are inseparable because the definition and principle of one task unavoidably affect the other. In this study, different approaches used in Khmer word segmentation and part-of-speech are reviewed, and an experimental study using a single long short-term memory network is described. Dataset from Asia Language Treebank is used to train and test the model. The preliminary experimental model achieved a 95% accuracy rate. However, more testing to evaluate the model and compare it with different models is needed to conduct to select the higher accuracy model.

1. Introduction

Natural Language Processing is the interdisciplinary field of computer science and linguistics which aims to provide the ability for computers to understand human languages. Natural language processing applications are very important and useful in daily lives, so recently, many researchers have focused on the natural language processing field in different human languages. According to Ding, Utiyama, and Sumita (2019), some basic tasks should be applied to raw textual data before developing sophisticated natural language processing applications such as tokenization and part-of-speech tagging.

Khmer language, also called Cambodian, is the official language of the Kingdom of Cambodia. Khmer is categorized as a low-resource, analytic, and isolating language in the Natural Language Processing (NLP) field (Ding, Kaing, Utiyama, Chea, & Sumita, 2016). According to Magueresse, Carles, and Heetderks (2020) low resource language can be understood as less studied, resource-scarce, less computerized, less privileged, less commonly taught, or low density, among other denominations. And as Analytic and isolating language, morphemes refer to a smallest meaningful unit of a language that can be combined freely (Ding et al., 2019). In the Khmer writing system, Khmer is typically head-initial where modifiers follow the word they modify. The common structure of Khmer sentences is subject-verb-object and topic-comment (Ding et al., 2016). Khmer writing is written from left to right with optional spaces used for

readability or other functional purposes (Ding et al., 2019). There are no word separators or delimiters to show boundaries between words. Since there is no word boundary and there are no exact rules for word combination, this causes the issue and error in identifying the words or word segmentation in Khmer text, which is a prior task for developing further NLP applications. Moreover, word segmentation, which is a task that segments text into words, is a prior step in Khmer text processing tasks such as part-of-speech tagging (POS tagging), and thus, the robustness of POS tagging is highly depends on word segmentation (Ding et al., 2019). POS tagging is one of the sequence labeling tasks in which a word is assigned to one of a predefined tag set according to its syntactic function. There are no standard rules for using spaces in the Khmer writing system. These large contiguous blocks of unsegmented words can cause major problems for natural language processing applications such as machine translation, speech synthesis, and information extraction, and therefore word segmentation techniques need to be developed. According to Chea et al. (2015), determining the positions of word boundaries is easy for Khmer native speakers. However, developing an automatic word segmentation is not a trivial task.

This study is based on the NOVA annotation system which was developed by Ding et al. (2019). NOVA was designed with three motivating features such as a compact tag set for cross-lingual word classes, extensibility to language-specific word classes, and flexibility to the concept of words. To represent fundamental word classes, NOVA uses 04 basic tags such as n (general noun), v (general verb), a (adjective), o (for which n, v, and tags are not applicable), and 03 auxiliary tags such as "1", ".", and "+" to define numbers, punctuation marks, and tokens with the weak syntactic role (Ding et al., 2019). Because of its limitation in expressiveness, the NOVA's researchers further modify and combine schemes in NOVA to get more flexible and expressive in the annotation. They modify by adding a minus sign "-" to represent functional word classes that play a similar role to the corresponding original tags. For example, prepositions can be defined as "v-" since they take nouns as arguments, so it is similar to verbs. Finally, the tag combination was used to form more complex and expressive tags and tag sequences to adapt to tokenization ambiguities. A pair of brackets "[" and "]" to illustrate "working(together) as" one or multiple tags and a slash "/" to show non-detachable components within one token.

There are two objectives for this study. Firstly, the aim of this study is to review the different approaches used to solve word segmentation and part-of-speech tagging in the Khmer language. And the second objective is to experiment with automatic joint Khmer word segmentation and part-of-speech tagging using a bidirectional long short-term memory neural network. However, in this study, it will not be focused on long tokens, which means the pair of brackets "[" and "]" will be removed.

This paper is structured as follows: literature reviews part, which discusses different methods and frameworks in recent research related to Khmer word segmentation and POS tagging, Bidirectional long short-term memory section, which describes the experiment of this study and result, and finally is future work section.

2. Literature reviews

2.1. Word segmentation

There is no explicit word boundary in the Khmer writing system; thus, to build natural language processing applications, the researchers have to consider the automatic word segmentation tools. Different word segmentation approaches and their results in five recent types of research are discussed below to provide a clear understanding and identify each method's strengths and weaknesses.

One of the early research projects was written by Seng, Sam, Besacier, Bigi, and Castelli

(2008), titled "First broadcast news transcription system for the Khmer language." The authors of Seng et al. (2008) aimed to develop a Large Vocabulary Continuous Speech Recognition (LVCSR) system for the Khmer language. To do the speech recognition in the Khmer language, one of the challenges is the writing system without explicit word boundaries, which calls for automatic segmentation approaches to make statistical language modeling feasible. For automatic segmentation, Seng et al. (2008) developed and studied segmentation tools to segment text into sentences, words, and character-cluster. However, since this literature focuses only on the word segmentation method so, the automatic segmentation tools to segment text into sentences and character-cluster will not be discussed. In the Khmer writing system, the Khmer Character Cluster (KCC) refers to the segmentation of Khmer text into an inseparable combination of characters units (Huor, Hemy, & Navy, 2004). A vowel cannot be by itself; a vowel must be placed after a consonant (Chea et al., 2015). Word segmentation is a primary and crucial task to build a system such as speech recognition. Seng et al. (2008) proposed the longest matching algorithm, which is a greedy algorithm that matches the longest word based on the official Khmer dictionary (Chhoun Nat dictionary) with a list of vocabulary of 18,000 words to develop automatic segmentation for their speech recognition system. However, correctly segmenting sentences into words requires the full knowledge of the vocabulary and the semantics of the sentence. Thus, the automatic segmentation method, which is generally based on a vocabulary in a dictionary, cannot give 100% of a correct segmentation because of the ambiguities and performs worst when the out-of-vocabulary rate is high. Out-of-vocabulary is a word that is not included in the dictionary or training set, such as personal name, names of places, and new words. According to Chea et al. (2015), there are two main causes of word boundary ambiguities. The first cause is lexical semantics which means a single sentence can be segmented in several ways based on its meaning in context. The second cause is unknown words which are words that are not found in dictionaries or training data and are often named entities such as personal names and locations. The Word segmentation tool of Seng et al. (2008) achieved 95% of correct word segmentation on some held data.

Another method was presented in research called "Detection and Correction of Homophonous Error Word for Khmer Language," a word segmentation. Huor et al. (2004) proposed a word segmentation method by using combination approaches between dictionary matching and bigram. This method leads to a decrease in the number of dictionary accesses in the process of searching for possible segmentations, which is time-consuming because KCC is an inseparable unit and well designed so that the boundary of KCC might be the boundary of words as well. Khmer Common Expression (KCE) is an expression created to make Khmer strings with the same pronunciation similar. A simple character-to-character mapping method could not be applied here because there are many cases of sound changes in the Khmer pronunciation system. Then, dictionary lookup is done using KCE to produce the possible word segmentation hypothesis for the original input text. Finally, a bigram model is applied to resolve segmentation ambiguities. Bigram models over KCC units and words were studied. Word bigrams proved to be the most effective, achieving 91.56% precision, 92.14% recall, and 91.85% F-score. Longest Matching and bigram were two early proposed methods to solve word segmentation in Khmer. Since each research used different training and test set, thus the accuracy rates are not comparable. However, by looking at the results of methods such as longest matching and bigram described above, it was good already to implement at that time, even with some issues.

Since there is still a lot of room to improve in Khmer word segmentation and different approaches available, Bi and Taing (2014) proposed a method called "Bi-directional Maximal Matching" for Khmer word segmentation in plaintext and Microsoft word documents (Ding et al.,

2019) try to improve the level of accuracy and performance in Khmer word segmentation by studying Bi-directional Maximal Matching (BiMM) with Khmer Clusters, Khmer Unicode character order correction, corpus list optimization to reduce the frequency of dictionary lookup, and Khmer text manipulation tweaks. Bi-directional Maximum Matching (BiMM), the supervised segmentation approach, is fundamentally the combination of Forward Maximal Matching (FMM) and Backward Maximal Matching (BMM) algorithm, which solves ambiguity problems caused by a greedy characteristic of the Maximal Matching algorithm when using alone; either FMM or BMM. The result of the study is 98.3% accuracy with time spent of 2.581 seconds for Khmer contents of 1,110,809 characters which are about 160,000 Khmer words (Bi & Taing, 2014). Even Bi and Taing (2014) achieved higher accuracy compared to previous methods; it seems difficult to define which one is the best method to implement in natural language processing applications to obtain a better performance.

After that year, Chea et al. (2015) discussed and compared two methods for Khmer word segmentation which are conditional random fields and maximum matching as a baseline. The research problem being addressed is whether the conditional random fields technique used in Khmer word segmentation can achieve higher accuracy and provide large improvements in statistic machine translation tasks. Maximum matching method segments using words chosen from a dictionary. In the maximum matching experiments, the researchers ran them with the dictionary, which contained 27,070 unique words extracted from the manually annotated corpus used to train the CRF models. Conditional random fields are models that consider dependencies among the predicted segmentation labels that are inherent in the state transitions of finite-state sequence models and can incorporate domain knowledge effectively into the segmentation process. 12,468 sentence test set was randomly selected from the full corpus. Chea et al. (2015) had put a lot of effort into developing manually annotated training sets for use in their experiments. However, their dataset was not open to the publics. Maximum matching and CRF methods were also evaluated as a pre-processing step in a statistical machine translation system. Chea et al. (2015) reported in their paper that the CRF model outperformed the baseline in terms of precision, recall, and F-score by a wide margin. CRF-based word segmentation provides significant results because it considers dependencies among the predicted segmentation labels that are inherent in the state transitions of finite-state sequence models and can incorporate domain knowledge effectively into the segmentation process (Chea et al., 2015). The evaluation method used was the Edit Distance of the Word Separator (EDWS). There were two common errors in word segmentation which are OOV errors and compound word errors. The authors found out that the maximum matching method was unable to segment any of the OOVs correctly; this was because if a word is not in the dictionary, the default character segmentation was used. The CRF model had an OOV segmentation accuracy of 0.44. The CRF also shows much better performance on segmenting compound words; here, the accuracy was 0.88, compared to only 0.57 with the maximum matching method.

The primary limitation of a linear CRF model is that the model is unable to handle long sequences effectively. Bouy, Taing, and Kor (2020) proposed a deep learning approach to build automatic word segmentation. In their paper, an RNN-based approach was studied. Many-to-many BiLSTM networks, to be precise, are trained to determine the word boundaries. Unlike the CRF model, there is no need for feature engineering, which means inputs can be passed directly into the networks which learn feature representation during training. Two BiLSTM networks are proposed, depending on the input types, including character-level, which is the network reads a sequence of one-hot encoded characters and passes the inputs through stacked BiLSTM cells. The outputs from the last BiLSTM cell are fed to a Fully Connected (FC) layer to produce binary outputs, indicating whether an input character is the starting part of a word and otherwise, and the KCC level, which

is the network reads a sequence of embedded KCCs and passes the inputs through stacked BiLSTM cells. The processed corpus has 188,043 sentences, 80% of which are used as a train set, and the rest is used to validate performance. The networks are implemented in PyTorch, which is a modern deep learning library. The same evaluation metrics as Chea et al. (2015) were used to evaluate the segmentation performance. The performance of the BiLSTM networks at both KCC and Character levels is comparable with state-of-the-art for word segmentation on the Khmer document using Conditional Random Field (CRF) by Chea et al. (2015). Unlike a CRF model, the BiLSTM networks can, however, learn the feature representation from the training corpus directly; the CRF model requires hand-picked feature engineering. The encouraging results of the BiLSTM network for Khmer word segmentation will encourage more deep learning applications in the Khmer NLP community.

In Conclusion, there were 04 main approaches used in the previous studies, such as maximum matching and its variances, Bigram, conditional random field, and deep learning. Maximum matching method is depended on the dictionary to perform the word segmentation; it might not be able to handle the out-of-vocabulary. Whereas the conditional random field requires to do the hand-picked feature engineering, we believe the researcher might need to have deep knowledge of linguistics. Deep learning is state-of-the-art in doing word segmentation in the Khmer language; thus, we will explore this approach in our study.

2.2. Part-of-speech tagging

Part-of-speech tagging is one of the sequence labeling tasks in which a word is assigned to one of a predefined tag set according to its syntactic function. POS tagging is required for various downstream tasks such as spelling check, parsing, grammar induction, word sense disambiguation, and information retrieval. In this section, we will discuss approaches from recent studies of POS tagging in Khmer language, which includes rule-based, conditional random fields, and six different approaches from Nou and Kameyama (2007) have presented initiative research on Khmer part-of-speech tagging. The researchers proposed a rule-based transformation approach with some modifications to adapt to the Khmer language. Moreover, the hybrid approach of rule-based and trigram models was studied. The dataset was manually annotated and contained about 32,000 words. For known words, the hybrid model could achieve up to 95.55% and 94.15% on the training and test set, respectively. For unknown words, 90.60% and 69.84% on the training ad test set were obtained. Nou and Kameyama (2007) have defined 27 tags; these tags also include small classes such as time expression, possibility expression, action expression, etc. Using these tags, it provides more comprehensive information about a word or a sentence, such as the tense, active, or passive forms. Nou and Kameyama (2007) used the word segmentation tool created by Huor et al. (2014), Ding et al. (2016) to segment a sentence into words. There are two main processes of the transformation-based approach, which are the learning process to learn the rules automatically from an annotated corpus based on a learning algorithm and predefined templates, and the transformation process refers to when tagging new texts and learning rules are applied to correct errors caused by the initial tagging process.

Sangvat and Pluempitiwiriyawej (2018) proposed an alternative approach to Khmer POS tagging using Conditional Random Fields (CRF). Sangvat and Pluempitiwiriyawej (2018) investigated different templates of contextual information and used them as the baseline model. As a result, they achieved a testing accuracy of 96.39%.

Thu, Chea, and Sagisaka (2017) developed manually annotated twelve thousand sentences POS tagged corpus for a general domain. The researchers discussed and evaluated six POS tagging approaches such as the Hidden Markov Model (HMM), Maximum Entropy (ME), Support Vector Machine (SVM), Conditional Random Fields (CRF), Ripple Down Rules-based (RDR), and Two

Hours of Annotation Approach (combination of HMM and Maximum Entropy Markov Model) on their developed POS tagged corpus. The result shows that RDR and HMM approach is the best performance for the open test set. The SVM approaches achieved the highest performance on the closed test set and also compared results with CRF on the open test set.

2.3. Joint word segmentation and part-of-speech tagging

The robustness of POS tagging highly depends on word segmentation because word segmentation is a prior step in Khmer text processing tasks (Bouy, Taing, & Kor, 2021; Ding et al., 2019). The Khmer language does not use explicit word separators, so the definition of words is not a natural concept and therefore, segmentation and part-of-speech tagging cannot be separated as both tasks unavoidably affect one another.

The research article "Joint Khmer Word Segmentation and Part-of-Speech Tagging Using Deep Learning" by Bouy et al. (2021) proposed a joint word segmentation and POS tagging using a single deep learning network to remove the adverse dependency effect. The proposed model is a bidirectional Long Short-Term Memory (LSTM) recurrent network with one-to-one architecture at a character level. The network takes inputs as a sequence of characters and outputs a sequence of POS tags. The researchers use the publicly available Khmer POS dataset by Thu et al. (2017) to train and validate the model. The authors revised the POS tag set by grouping similar tags together from 24 POS tags that were derived from Choun Nat Dictionary. After the revision, the modified tag set consists of 15 tags. The 12,000 sentence dataset was used as a training set. Space denotes word boundary, and the POS tag of a word is right after a slash. If a character is the starting character of a word, its label is the POS tag of the word, and it also means the beginning of a new word. Otherwise, the No-Space (NS) tag is assigned. The proposed network processes starting from the network take a sequence of characters as inputs, encode an input character using one-hot encoding, process an input sequence in both directions forward and backward, LSTM stack is concatenated to form a single hidden vector, the concatenated hidden vector is fed into a feed-forward layer to produce an output vector, and finally, Softmax activation is applied to the output vector to produce probabilistic outputs. The network was implemented in the Pytorch framework and trained on Google Colab Pro. Training utilized mini-batch on GPU. The overall accuracy of the proposed model is on par with the conventional two-stage Khmer POS tagging. The training dataset available is limited in size, and a significantly larger dataset is required to train a more robust joint POS tagging model with greater generalization ability. This article focused on the necessary research topic within natural language processing, especially for the Khmer language, as other researchers need these two tasks to achieve creating useful real-life applications. Bouy et al. (2021) have provided proof of high accuracy in using a deep learning model on the Khmer language on joint word segmentation and part-of-speech tagging. Bouy et al. (2021) presented their process of building in a simple and easy-to-understand whole processing of building their model. The paper was not published in any journal yet; thus, it might be concerned with its credibility. Future research on join word segmentation and part-of-speech tagging by using more training datasets and evaluating by using that model in other tasks such as information retrieval, machine translation, is necessary to provide proof of the high accuracy and feasibility of the model.

The research article "NOVA: A Feasible and Flexible Annotation System for Joint Tokenization and Part-of-Speech Tagging" by Ding et al. (2019) aimed to solve word segmentation or tokenization and part-of-speech tagging issues of East and Southeast Asia languages using a proposed system called NOVA. NOVA framework was designed with three motivating features such as a compact tag set for cross-lingual word classes, extensibility to language-specific word classes, and flexibility to the concept of words. To represent fundamental

word classes, NOVA uses 04 basic tags such as n (general noun), v (general verb), a (adjective), o (for which n, v, and tags are not applicable), and 3 auxiliary tags such as "1", ".", and "+" to define numbers, punctuation marks, and tokens with the weak syntactic role. After annotating using basic tags of NOVA, it is still difficult to define whether some small parts such as "t", "s", or "wake up" should be segmented. Because of its limitation in expressiveness, the researchers further modify and combine schemes in NOVA to get more flexible and expressive in the annotation. They modify by adding a minus sign "-" to represent functional word classes that play a similar role to the corresponding original tags. For example, prepositions can be defined as "v-" since they take noun as arguments, so it is similar to verbs. Finally, the tag combination was used to form more complex and expressive tags and tag sequences to adapt to tokenization ambiguities. A pair of brackets "[" and "]" to illustrate "working(together) as" one or multiple tags and a slash "/" to show non-detachable components within one token. The researchers of this article have been in the Asia Language Treebank (ALT) project which involving in tokenization, part-of-speech tagging, phrase structure tree building, and token alignment with English translations over some low-resource Southeast Asian languages. They pick up two languages, Burmese and Khmer, to practice NOVA since both are high analytic in morphology, which is a common feature in Southeast Asian languages. Almost all languages in Southeast Asia have a pure head-initial structure, such as Khmer, Laotian, Thai, and Vietnamese, except Burmese, which is a strong head-final language. Consequently, the NOVA's authors believe that these two languages can represent Southeast Asian languages to apply the NOVA annotation system; if it is applicable in both languages, it is also applicable to Laotian, Thai, and Vietnamese. However, the NOVA's authors should provide more linguistics discussion and guidelines on other east and southeast Asia languages besides Khmer and Burmese so it would be proof that the NOVA system is flexible and feasible enough for these languages. In addition, Ding et al. (2019) show statistics on the Khmer and Burmese ALT data annotated using NOVA. The statistics were conducted in two ways, short units (tags and tokens) and long units. Based on these statistics, n, v, and o tags cover nearly 90% of the short units, which is a good fit for the nova motivation is to provide basic word classes. The article also shows that 90% of the long tokens in both languages are formed by the 12 most frequent basic tags and combination patterns. Moreover, the comparison demonstrates that NOVA's descriptive ability is approximately similar to G-UNI's, with the added benefit of the ability to switch between a generalized and modified version. The generalized version refers to using the basic NOVA basic tag set, and the modified version is used for the specificity of each language. Thus, the ability to switch these two versions can be consistently applied to languages with different features. Since the Khmer text is translated text from English, the author should also observe the result of the original text of these two languages using the NOVA annotating system.

In the same framework, an article titled "Towards Tokenization and Part-of-Speech Tagging for Khmer: Data and Discussion" by Kaing et al. (2021) aimed to establish a benchmark for automatic processing of tokenization and part-of-speech tagging for the Khmer language. To successfully build automatic processing of Khmer tokenization and part-of-speech tagging, there are important steps to consider, such as data collection, experiments on different models, and discussion on the achievement and limitation of the data and approaches. The experiments were conducted using a Support Vector Machine (SVM), a Conditional Random Field (CRF), a Long Short-Term Memory (LSTM), and an integrated LSTM-CRF model. The authors of Kaing et al. (2021) presented the analysis and discussions based on the experimental results to show the achievements and limitations of the data and approaches. The source data comprise 20,000 sentences translated from English Wikinews and fully tokenized and POS-tagged by native speakers. The team for Khmer annotation comprised only around ten people and, at most times,

fewer than five people. All the team members are researchers and students proficient in NLP. It took around 04 years to prepare the released data of this work from 2016 to 2019. This dataset is open source for research use. The authors mentioned the manual process for annotating the data that they used in doing their experiments; however, the timeline for preparing the data for this work is around 04 years which seems to take a long time. Thus, based on the experience, the authors should include a suggestion on how the annotation should be processed. And why they decided to translate from English rather than using the original Khmer text for doing the annotating and used as the data for the experiment. For tokenization, two types of tokens, short ones, and long ones, were designed to accommodate the morpheme and word units, respectively. Spaces were inserted to separate short tokens; long tokens were bracket-wrapped in short token sequences. Smaller than a short token, the basic atom in Khmer scripts is referred to as a writing unit. It comprised one or more staked consonant characters, with one or more optional modifying diacritics. Several conventional Khmer dictionaries are referred to in the annotation for the identification of long tokens. A sequence composed of one or more short tokens is identified as a long token is listed as an entry in dictionaries. The annotation related to foreign names and loanwords was based on the original languages. Every single word in the original languages was treated as a short token. Each short token is tagged with one POS tag, and each long token composed of multiple short tokens is tagged with a second-layer POS ag outside a pair of wrapping brackets.

The division of the training/development/test sets was performed according to the unified setting under the Asia Language Treebank project. The SVM performs point-wise predication for each writing unit with a sliding window of contextual features. Bi, tri, and 4-gram features were experimented with. As a tradeoff of the fastness, the point-wise estimation is intrinsically weak in capturing the sequential information. A CRF model implemented by the CRF++ toolkit was set as a standard sequence labeling baseline. The settings and feature template followed the previous work. An LSM- based RNN was implemented using DyNet. The network was essentially configured according to Ding et al. (2020), whereas the dimensions of layers were enlarged to achieve better performance. The model ensemble was also conducted, up to 100 models, as a large-scale model ensemble was found to gradually increase the performance. An LSTM-CRF was implemented using NCRF++. The network configuration followed that in MA and Hovy, with the dimension of layers adjusted to fit our task. For the evaluation, the F-score, which is the harmonic mean of the precision and recall, was used. For tokenization, precision is the ratio between the number of correct segmented tokens and the total number of tokens obtained by automatic processing, and recall is the ratio between correct segmented and the total number of manually segmented tokens in the reference. For POS-tagging, the F-score was calculated jointly on correct tokens with correct POS tags. The result shows that the differences between the RNN and the LSTM-CRF are gradually reduced when more training data are provided. However, RNN always outperformed the CRF, regardless of the size of the dataset. This suggests that performance can be further boosted by increasing the training data, especially on long tokens. Regarding the POS tagging, there are intrinsic ambiguities for certain morphemes because of the highly analytic features of Khmer. A solution may rely on a more informative POS-tag set than the nova scheme used in this study. This requires a more insightful, linguistically oriented investigation of the Khmer language. Experiment-based analysis showed that automatic processing up to morpheme level was satisfactory, but for larger constituents, for example, complex compounds and phrases, joining processing was required. The models that the authors of Kaing et al. (2021) used in their experiments are grounded in previous literature; thus, there are enough reasons to convince that those models are the current state of technology to use.

4. Experimental study using bidirectional Long Short-Term memory

In this research, we will follow the same approach as Bouy et al. (2021) since deep learning can handle the long sequences of text effectively in word segmentation and POS tagging. Moreover, there are not many studies using the Deep learning approach in Khmer NLP tasks; thus, it is very interesting to implement and test the similar approach from Bouy et al. (2021) with a different dataset to observe the efficiency of the method. With the help of the PyTorch framework, we believe it will help the implementation process a lot simpler. "Bidirectional Long Short-Term Memory network (LSTM)" is used to do joint word segmentation and part-of-speech tagging in this research. We will be utilizing neural networks already implemented in the libraries provided by a framework called PyTorch, which is an open-source machine learning framework that accelerates the path from research prototyping to production deployment (https://pytorch.org/). PyTorch is primarily developed by Facebook's AI Research lab. It is free and open-source software released under the Modified BSD license. PyTorch provides a few different styles of models like Recurrent Neural Networks and Convolutional Neural Networks, which makes it easy for models to be developed and modified with as many different layers of neural networks as necessary. A Recurrent Neural Network (RNN) is a type of neural network with a cycle in its connections. So, the value of an RNN cell depends on both inputs and their previous outputs. In NLP tasks, simple recurrent networks have been proven to be very effective (Dan & James, 2021). Moreover, we will utilize Google Collaboratory (https://colab.research.google.com/notebooks/intro.ipynb) as an online notebook to implement the data cleaning phase, training, and testing of the neural network.

Language consists of an ordered sequence of words in which the context of surrounding words is important.

A Recurrent Neural Network (RNN) is designed to process time-series data, one datum at a time. This suggests that RNNs might be useful for processing language, one word or character at a time (Bouy et al., 2021). An RNN cell struggles to carry forward critical information when processing a long sequence because the weight matrices must provide information for both the current and future outputs, and back-propagation through time suffers from vanishing gradients due to repeated multiplications along a long sequence (Bouy et al., 2021). Long Short-Term Memory (LSTM) networks were devised to address the above issues by introducing sophisticated gate mechanisms and an additional context vector to control the flow of information into and out of the units. The model of the study is a bidirectional Long Short-Term Memory (Bi-LSTM) network at the character level for word segmentation and POS tagging, as shown in Figure 4. It is bidirectional since the model has access to an entire input sequence during the forward run.

The descriptions of the Bi-LSTM model from (Bouy et al., 2021):

• Input: the network takes a sequence of characters as inputs;

• Input Encoding: one-hot encoding is used to encode as an input character;

• LSTM layers: the network processes an input sequence in both directions (forward and backward). The forward and backward hidden vectors in the final LSTM stack are concatenated to form a single hidden vector;

• Feed-forward layer: The concatenated hidden vector is ed into a feed-forward layer to produce an output vector. The size of the output vector is equal to the number of POS tags plus one as an additional no-space(ns) tag is introduced;

• Softmax: Softmax activation is applied to the output vector to produce probabilistic outputs;

• Multi-class cross-entropy is used to train the proposed model.

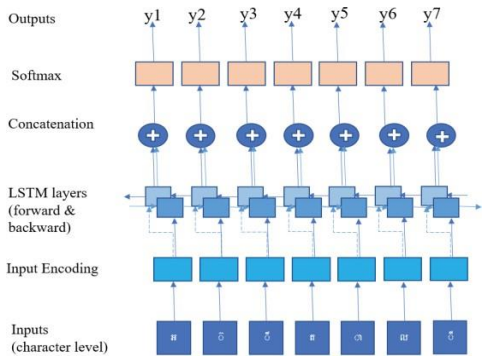

Figure 1. The proposed Bi-STM network for joint segmentation and POS tagging at character level (Bouy et al., 2021)

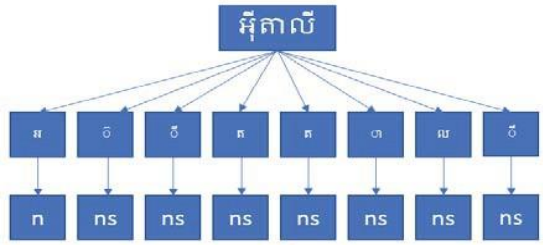

Figure 2. Input and Output Sequence representation example

4. Experimental setup and results

4.1. Dataset

Deep learning model performance is based on the input data; thus, the quality of data is very important. Collecting Khmer text data is very challenging and time-consuming. Because of these reasons, we decided to use secondary data from Asian Language Treebank (ALT) project conducted by Riza et al. (2016). ALT project is a project which aims to advance the start-of-art Asian natural language processing technique through the open collaboration for developing and using ALT. Khmer ALT has been developed by the National Institute of Information and Communications Technology, Japan, and the National Institute of Posts, Telecoms & ICT, Cambodia (Currently known as Cambodia Academy of Digital Technology). The license of Khmer ALT is Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) License (https://creativecommons.org/licenses/by-nc-sa/4.0/). The dataset contains a sampling of 20,106 sentences from English Wikinews under a Creative Commons Attribution 2.5 License, and then these sentences were translated into the other six languages, including Khmer. The current Khmer ALT dataset has word segmentation and part-of-speech tags. There are three files in the dataset such as "data_km.km-tag.nova" which contains Khmer tokens separate by space to indicate the boundary of each token, "data_km.km-tok.nova" which contains the assigned part-of-speech tags for each token, and README.txt that describe the guideline of the dataset.

SNT.80188.1 អ៊ីតាលី បាន ឈ្នះ លើ ព័រទុយហ្គាល់ 31-5 ក្នុង ប៉ូល C នៃ ពិធី ប្រកួត ពាន រង្វាន់ ពិភព លោក នៃ កីឡា ប៉ាល់ ឱប ឆ្នាំ 2007 ដែល ប្រព្រឹត្ត នៅ ប៉ាស ឌេស ប្រិន ក្រុង ប៉ារីស បារាំង ។

Figure 3. An example sentence in the Token file of the Khmer ALT dataset

SNT.80188.1 n v- v o n n o n[n n]n o n[n v]n n[n n]n n[n n]
o n[n n v]n n[n 1]n n v v- n[n n n]n n[n n]n n .

Figure 4. Corresponding sentence of tag file to an above token sentence of Khmer ALT dataset

4.2. Experimental setup and result

The model of this study is implemented in PyTorch and trained using mini-batch and GPU to speed up. The dataset is divided into a train set (80%) and a test set (20%). There were 16,084 and 4,022 sentences in the train and test set, respectively.

These are hyper-parameters randomly selected and used in the network:

- Number of stacks = 5;
- Hidden dimension = 150;
- Batch size = 128;
- Optimizer = Adam with a learning rate of 0.001;
- Epochs = 100.

4.3. Evaluation protocols and result

To evaluate the proposed model, the accuracy will be implemented.

The overall accuracy of POS tagging is calculated as below:

$$Accuracy = \frac{\sum_{i=1}^{tag} NumCorrect(i)}{\sum_{i=1}^{tag} NumWord(i)} \quad (1)$$

Where:

NumCorrect(i) is the number of correctly predicted POS tag*(i)*;

NumCorrect(i) is the number of POS tags (i) in the corpus;

tag is the number of POS tags.

The accuracy of the model is approximately 95%.

4. Conclusion

In this work, we discussed and reviewed previous studies and did the experimental study using a bidirectional long short-term memory network. The joint word segmentation and POS tagging using a bidirectional LSTM network that takes inputs at the character level and outputs a sequence of POS tagging is implemented. Due to the time constraint of this study, it will be future work of this study to do more testing to evaluate the model and compare it with different models to choose the best model with higher accuracy and extend the network of this study with the ability to identify and assign POS tags of long tokens or compounds.

References

Bi, N., & Taing, N. (2014). *Khmer word segmentation based on Bi-directional maximal matching for plaintext and Microsoft Word document.* Paper presented at the Signal and Information Processing Association Annual Summit and Conference (APSIPA), Chiang Mai, Thailand.

Buoy, R., Taing, N., & Kor, S. (2020). *Khmer word segmentation using BiLSTM networks.* Paper presented at the 4th Regional Conference on OCR and NLP for ASEAN Languages, Phnom Penh, Cambodia.

Buoy, R., Taing, N., & Kor, S. (2021). *Joint Khmer word segmentation and part-of-speech tagging using deep learning.* Retrieved October 10, 2021, from https://arxiv.org/ftp/arxiv/papers/2103/2103.16801.pdf

Chea, V., Thu, Y. K., Ding, C., Utiyama, M., Finch, A., & Sumita, E. (2015). *Khmer word segmentation using conditional random fields.* Retrieved October 10, 2021, from https://www2.nict.go.jp/astrec-att/member/ding/KhNLP2015-SEG.pdf

Dan, J., & James, H. M. (2021, December 29). *Speech and language processing*. Retrieved March 02, 2022, from https://web.stanford.edu/~jurafsky/slp3/

Ding, C., Aye, H. T., Pa, W. P., Nwet, K. T., Soe, K. M., Utiyama, M., & Sumita, E. (2020). Towards Burmese (Myanmar) morphological analysis. *ACM Transactions on Asian and Low-Resource Language Information Processing, 19*(1), 1-34.

Ding, C., Kaing, H., Utiyama, M., Chea, V., & Sumita, E. (2016). *Tokenization and part-of-speech annotation guidelines for Khmer (Cambodian).* Retrieved March 02, 2022, from https://att-astrec.nict.go.jp/member/mutiyama/ALT/Khmer-annotation-guideline.pdf

Ding, C., Utiyama, M., & Sumita, E. (2019). NOVA. *ACM Transactions on Asian and Low-Resource Language Information Processing, 18*(2), 1-18.

Huor, C. S., Hemy, R. P., & Navy, V. (2004). *Detection and correction of homophonous error word for Khmer language.* Retrieved March 02, 2022, from https://www.yumpu.com/en/document/read/25135741/detection-and-correction-of-homophonous-error-word-for-khmer-

Kaing, H., Ding, C., Utiyama, M., Sumita, E., Sam, S., Seng, S., . . . Nakamura, S. (2021). Towards tokenization and part-of-speech tagging for Khmer: Data and discussion. *ACM Transactions on Asian and Low-Resource Language Information Processing, 20*(6), 1-16.

Magueresse, A., Carles, V., & Heetderks, E. (2020). *Low-resource languages: A review of past work and future challenges.* Retrieved March 02, 2022, from https://arxiv.org/pdf/2006.07264.pdf

Nou, C., & Kameyama, W. (2007). *Khmer POS tagger: A transformation-based approach with hybrid unknown word handling.* Paper presented at the International Conference on Semantic Computing (ICSC 2007), Irvine, CA, USA.

Riza, H., Purwoadi, M., Gunarso, Uliniansyah, T., Ti, A. A., Aljunied, S. M., . . . Ding, C. (2016). *Introduction of the Asian language treebank.* Paper presented at the 2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA), Bali, Indonesia.

Sangvat, S., & Pluempitiwiriyawej, C. (2018). Khmer POS tagging using conditional random fields. *Communications in Computer and Information Science*, 169-178. doi:10.1007/978-981-10-8438-6_14

Seng, S., Sam, S., Besacier, L., Bigi, B., & Castelli, E. (2008). *First broadcast news transcription system for khmer language.* Retrieved March 02, 2022, from https://hal.archives-ouvertes.fr/hal-01392538/document

Thu, Y. K., Chea, V., & Sagisaka, Y. (2017). Comparison of six POS tagging methods on 12K sentences Khmer language POS tagged corpus. *Proceedings 1st Regional Conference Optical Character Recognition and Natural Language Processing Technologies for ASEAN Languages*, 1-12.