P-ISSN 2586-9000
E-ISSN 2586-9027
Homepage : https://tci-thaijo.org/index.php/SciTechAsia
Science & Technology Asia
Vol.26 No.2 April - June 2021
Page: [61-78]
Original research article

Enhancement of Character-Level Representation in Bi-LSTM model for Thai NER

Kitiya Suriyachay[1], Thatsanee Charoenporn[3], Virach Sornlertlamvanich [1,3,*], Natsuda Kaothanthong[2]

[1]*School of ICT, Sirindhorn International Institute of Technology, Thammasat University, Pathum Thani, 12120, Thailand*
[2]*School of Management Technology, Sirindhorn International Institute of Technology, Thammasat University, Pathum Thani, 12120, Thailand*
[3]*Asia AI Institute, Faculty of Data Science, Musashino University, Tokyo, 202-0023, Japan*

Received 20 December 2019; Received in revised form 20 November 2020
Accepted 29 December 2020; Available online 25 June 2021

ABSTRACT

Named Entity Recognition (NER) in the Thai language is a relatively challenging task because the Thai language does not have an explicit word boundary. This normally can cause difficulties in word segmentation, which affects the efficiency in NLP post-processing such as NER tasks. Moreover, one of the important problems is the ambiguity in using common nouns to express named entities. According to the Thai language, most named entities are usually placed close to a verb or a preposition with a specific pattern. This means that the part of speech (POS) can be effectively used as a feature to consider the type of named entity. For these reasons, in this paper, we generate the BiLSTM-CNN-CRF model to investigate the effectiveness of a combination of the features among word, POS, and Thai character clusters (TCCs). We use TCCs instead of characters to minimize word segmentation errors in the corpora and increase the efficiency in generating the model. Experimental results show that our proposed model outperforms other models. The TCC is a suitable unit for character embedding, providing better results than single character embedding.

Keywords: Named Entity Recognition; Recurrent Neural Network; Bidirectional LSTM; CNN; CRF; Thai language; Thai named entity; TCC

1. Introduction

The amount of global digital information is growing at a high rate. International Data Corporation (IDC) predicts that the global digital data will increase from 33 Zettabytes (ZB) in 2018 to 175 ZB by 2025. Accurately making use of this massive amount of data is important for

***Corresponding author:** virach@musashino-u.ac.jp **doi:** 10.14456/scitechasia.2021.26

success in all kinds of businesses. Many large companies, therefore, have invested in information extraction and retrieval techniques, to efficiently extract the essence of the data and information for understanding the current situation and future prediction.

Some approaches of information extraction and retrieval rely on NLP preprocessing subtasks including word segmentation, POS tagging, named entity recognition (NER), etc.

NLP has been conducted in various languages for several years for finding the effective guidelines and methods to identify the NER in texts. NER is responsible for identifying and classifying the named entity of each word, such as a person's name, location, and organization.

NER in English has been continuously developed to get better results [1-4]. However, NER in some isolated or agglutinative languages, such as Thai, Japanese, or Chinese, still has additional challenges according to their linguistic characteristics. For the Thai language, the most challenging for NLP research is that there is no word boundary or space between words, as explicitly occurs in English sentences. Secondly, there is no capitalization or special characters used to provide a hint for identifying the named entity. Additionally, incorrect word segmentation problems cause difficulties and have a direct effect on the NER accuracy. The ambiguity of homographs, in which the same spelled word can represent different meanings depending on its context, is also unavoidable. For example, the word "ปทุม" (Pa-Thum) can have at least three meanings, which can be recognized in three different types of named entities: 1) a province name in Thailand (Pathum Thani), 2) a person name, 3) literally, the name of the lotus flower. Sometimes, named entities and common nouns occur in similar phrases which cause ambiguity. For example, "พันตำรวจเอกทวี สอดส่อง อธิบดีกรมสอบสวนพิเศษ" (Phan-Tam-Ruad-Eak-Ta-Wee-Sod-Song-A-Thi-Bor-Dee-Krom-Sob-Suan-Phi-Set; Colonel Tawee Sordsong, Director of the Department of Special Investigation). In this sentence, the word "สอดส่อง" (Sod-Song) is used as a surname and it also conveys a meaning of "to observe, or to monitor" in common use. Homographs can be recognized by interpreting the meaning. Incorrect homograph word interpretation also leads to wrong named-entity recognition.

To deal with the aforementioned challenges, we propose a bi-directional Long Short-Term Memory (LSTM) model with a convolutional neural network (CNN) and a conditional random field (CRF) using the word, part-of-speech (POS), and Thai character cluster (TCC) as the input features. Many previous studies support the use of the POS with the word in the NER model, which can provide better performance than a model that does not use the POS [18]. Therefore, the POS tag may improve the performance of NER in the Thai language, as most Thai named entities are located close to or adjacent to verbs or prepositions, such as ที่ (Thee-at), จาก (Jak; from), ไป (Pai; go), and ใน (Nai; in). The bi-directional LSTM can learn and remember the context from both the left and the right of a given word. In addition, words in the Thai language usually consist of several characters and contain rich internal information. Hence, using the internal character together with the word can make better use of the word information. Moreover, the POS and TCC can help in recovering the errors in word segmentation and NE tagging in the training corpus. In this paper, we propose an effective and efficient approach for Thai NER tasks. The remainder of the paper is organized as follows. The related works are described in Section 2. Section 3 describes the characteristics of a Thai named entity, while the corpus used in the experiment is explained in Section 4. Section 5 clarifies our proposed model and NER architecture for better accuracy and

consistency. The results and discussion are described in Section 6. A comparison between TCCs and single characters in character-level representation is described in Section 7 and the results of our model with other models are compared in Section 8. Section 9 presents an approach to combine all seven named entity types. The conclusions of this paper are explained in Section 10.

2. Related Work

Research on NER has already been conducted in many languages, and NER tools have been developed continuously. In many languages, there are a comparatively small number of NE-tagged corpora for supervised training. The study of named entities of words for NER modeling is limited [5]. One of the most effective and popular ways to deal with NER is the use of machine learning techniques [6].

Previously, traditional machine learning was widely used in NLP tasks. The Hidden Markov Model [7] recognizes the named entities in the Indian language, which is similar to the Thai language. The Thai and Indian languages have no capitalization, and they are free word-order languages. NER for biomedical text can use a Support Vector Machine (SVM) [8]. The CRF is another method that is used for NLP in many languages, such as Chinese [9] and Malaysian [10].

NER in the Thai language has been researched for many years. Examples are Thai proper name identification using a feature-based approach with context words and collocations [11], and Thai person name recognition using Likelihood Probability [12]. [13] reported that a combination of the word, semantics, and orthography yielded the best performance for Thai NE extraction, based on an SVM algorithm. [14] compared the performance of the NER model based on the word and character levels by using the CRF model. The experimental result shows that the performance of both is similar. The efficiency of a character-segmented model is slightly higher because the model cannot recognize words or syllables that never occur in the list, such as abbreviations. A part of the problem came from the absence of context clues (as a feature). Although these methods provide reliable results, there are some defects and drawbacks that can affect the performance of NER. The process to reconstruct a set of system features is difficult when changing the corpus or language [15]. In addition, the CRF model does not work well with words that have never occurred in the training dataset.

Recently, LSTM, which is a type of RNN, has shown great success and efficiency for NER tasks. [16] applied the POS with the word as input in the Bi-LSTM model, to recognize and classify the named entity in Thai, and the model outperforms the CRF model. Variational LSTM with CRF is also used for Thai NER and provides a satisfactory result [17]. In addition, [18] created the Bi-LSTM model for Indonesian information on Twitter and showed that using both word embedding and POS tags provides the highest F1-score. In Chinese, a word usually contains several characters, for example, the word “智能” (intelligence). The meaning of the word can be learned from the context surrounding the word in the sentence. Its meaning can also be deduced from the meaning of the characters in the word, i.e., “智” (intelligent) and “能” (ability). Therefore, the meaning of component characters plays an important role in modeling [19-20]. In addition, Bi-LSTM is used with character embedding instead of word embedding for Chinese NER research. Their results indicate that character embedding is more suitable and useful for Chinese NER than word embedding, and their model performs much better than the traditional baselines [20].

There are several research works that use the word together with the character, to provide better results than using either the word or character. [21] conducted the NER

model for the Mongolian language using word embedding with character embedding as input features in the Bi-LSTM with a CRF layer. [2] introduced a combined model of Bi-LSTM and CNN which utilized both the word and character-level features. Furthermore, [4] also presented a powerful BiLSTM-CNN-CRF model that achieved state-of-the-art performance on NER and successfully employed CNN to extract more useful character-level features. These research works show that character and POS features are very useful for the efficiency of the model in deciding on the right named entity tag.

3. Thai Named Entity

This section describes the characteristics of the Thai language that make it difficult to identify the named entity. As discussed above, Thai NER is a challenging task since the Thai language does not have rich character features, such as uppercase characters that can help to specify a named entity (distinguishing between named entities and common nouns). In addition, there is no character or space to separate each word from its neighboring words in a sentence, as can be found in English. Moreover, there is no symbol such as a period to indicate the end of a sentence. The variety of Thai writing styles is also a problem for NER as well. In general, in Thai, a proper name usually appears with a common noun that can be used to indicate the type of proper noun, such as in "นายสมชาย" (Nai-Som-Chai), where the word "นาย" (Nai; Mr.) is a title noun indicating that "สมชาย" (Somchai) is a proper noun and a person's name. Sometimes, a shorter name or abbreviation of the named entity is used in the Thai writing system instead or its prefix is cut off. Examples are, "มหาวิทยาลัยธรรมศาสตร์ได้รับรางวัลในเวที ประกวดสิ่งประดิษฐ์โลก ณ กรุงเจนีวา" (Thammasat University received an award at the International Exhibition of Inventions in Geneva) and "ธรรมศาสตร์มุ่งพัฒนาด้านวิทยาศาสตร์และเทคโนโลยี โดยมีนักศึกษา มธ. ร่วมมือผลักดันนวัตกรรมให้ทัดเทียม นานาชาติ" (Thammasat aims to develop science and technology with the cooperation of TU students to drive innovation to international standards.). In both sentences, there are three forms for "มหาวิทยาลัยธรรมศาสตร์" (Ma-Ha-Wit-Ta-Ya-Lai-Tham-Ma-Sat; Thammasat University), namely, "มหาวิทยาลัยธรรมศาสตร์" (Thammasat University) itself, "ธรรมศาสตร์" (Thammasat), and "มธ." (TU, the abbreviation of Thammasat University).

In addition, there is a problem of the ambiguity between common words and named entities, such as "น้องพี่ที่รักเป็นภาพยนต์ ที่ได้รับความนิยมสูงในประเทศไทย" (Nong Pee Teerak or "Brother of the Year" is a commercial hit in Thailand). The term "น้อง พี่ ที่รัก" (Nong, Pee, Teerak or Sister, Brother, Lover) in this sentence is the name of a Thai movie, which is a named entity, and it is also a phrase that is composed of common nouns.

Because of the omission of expressions in Thai, another problem in named entity recognition is that the same expression can be assigned to different types of named entities, such as "ลำพูน" (Lamphun), which can be both the name of a province in Thailand and a person's name. It depends on the context of the word.

4. Corpus

The initial corpus used in this study is the THAI-NEST corpus. The texts in the corpus are collected from Thai online news articles that are published on the internet, such as political news, foreign news, economic news, crime news, sport news, entertainment news, educational news, and technological news [22]. The corpus is disjointedly managed in seven files according to the type of named entity, which are Date (DAT), Time (TIM), Measurement (MEA), Name (NAM), Location (LOC), Person (PER), and Organization (ORG). Each category is

abbreviated by the first three characters. To be precisely annotated, the original THAI-NEST corpus in this study was redesigned and constructed based on the structure of the Orchid corpus [23-24], as shown in Fig. 1. The statistics of each corpus are listed in Table 1.

Table 1. Statistics of each corpus.

	No. of sentences	No. of words	No. of NE tags
DAT	2,784	214,467	14,334
LOC	8,585	569,292	33,596
MEA	1,969	157,788	17,371
NAM	7,553	547,489	40,537
ORG	20,399	1,386,824	95,566
PER	33,233	2,705,218	222,075
TIM	419	41,493	3,362

We introduce two types of mark-ups to separate the part of text information from the text itself. The text information line, which is a line beginning with "%", is used to describe the information of the corpus, as shown in Table 2. The numbering information line, which is a line beginning with the "#" symbol, is used to sequence the lines in the corpus as shown in Table 3. There are also three special mark-up delimiters as shown in Table 4.

Table 2. Mark-up for text information line.

Mark-up	Description
%Title:	Title of the corpus
%Description:	Detail of the corpus or reference
%Number of sentence:	Total number of sentences in the file
%Number of word:	Total number of words in the file
%Number of NE tag:	Total number of named entity tags in the file
%Date:	Date of creating the corpus
%Creator:	Name of the creator(s)
%Email:	Email Address(es) of the creator(s)
%Affiliation:	Affiliation(s) of the creators

Table 3. Mark-up for numbering information line.

Mark-up	Description
#P[number]	Paragraph number of the text. The number in the bracket presents the sequence of the paragraphs within a text.
#S[number]	Sentence number of the paragraph. The number in the bracket presents the sequence of the sentences within a paragraph.

Table 4. Special mark-up delimiters.

Mark-up	Description
\\	Line break symbol
//	Sentence break symbol
/[POS]	Tag marker for appropriate POS annotation of a word
/[NE]	Tag marker for appropriate NE annotation of a word

We used 47 types of the POS, defined in the Orchid Corpus [23].

For the format of NE tags, the BIO annotation scheme is used for all types of named entities, as shown in Table 5.

```
%Title: Date corpus
%Description: Date in any format
%Number of sentence: 2,783
%Number of word: 272,753
%Number of named entity tag: 14,330
%Date: January 6, 2019
%Creator: Kitiya Suriyachay and Virach
Sornlertlamvanich
%Email: m5922040075@g.siit.tu.ac.th and
virach@siit.tu.ac.th
%Affiliation: Sirindhorn International
Institute of Technology, Thammasat University

#S1
นายสุเทพ เทือกสุบรรณ รองนายกรัฐมนตรี กล่าวว่า ในวันพรุ่งนี้ (18 มี.ค.52) รัฐบาลโดย\\
นายอภิสิทธิ์ เวชชาชีวะ นายกรัฐมนตรี จะมอบนโยบายและแนวทางในการป้องกันและปราบปราม\\ ยา
เสพติดให้กับส่วนราชการต่างๆเพื่อบูรณาการแผนปฏิบัติการป้องกันและปราบปรามยาเสพติดร่วมกัน//
```

```
%Title: Date corpus
%Description: Date in any format
%Number of sentence: 2,783
%Number of word: 272,753
%Number of named entity tag: 14,330
%Date: January 6, 2019
%Creator: Kitiya Suriyachay and Virach
Sornlertlamvanich
%Email: m5922040075@g.siit.tu.ac.th and
virach@siit.tu.ac.th
%Affiliation: Sirindhorn International
Institute of Technology, Thammasat University

#S1
Mr. Suthep Thaugsuban, Deputy Prime Minister, said that tomorrow (18 Mar 2009) the\ \
government by Prime Minister Abhisit Vejjajiva will give policies and guidelines for\ \ prevention
and suppression of drugs to government agencies to integrate the drug\ \ prevention and
suppression action plan together//
```

นาย/NTTL/O สุเทพ/NPRP/O <space>/PUNC/O เทือกสุบรรณ/NPRP/O <space>/PUNC/O รองนายกรัฐมนตรี/NCMN/O <space>/PUNC/O กล่าว/VACT/O ว่า/JSBR/O <space>/PUNC/O ใน/RPRE/O วันพรุ่งนี้/ADVS/B-DAT <space>/PUNC/O (/PUNC/O 18/DONM/B-DAT <space>/PUNC/I-DAT มี.ค. 52/NPRP/I-DAT)/PUNC/O . . ยาเสพติด/NCMN/O ร่วมกัน/ADVN/O //	Mr./NTTL/O Suthep/NPRP/O <space>/PUNC/O Thaugsuban/NPRP/O <space>/PUNC/O Deputy Prime Minister/NCMN/O <space>/PUNC/O said/VACT/O that/JSBR/O <space>/PUNC/O tomorrow/ADVS/B-DAT <space>/PUNC/O (/PUNC/O 18/DONM/B-DAT <space>/PUNC/I-DAT Mar 09/NPRP/I-DAT)/PUNC/O . . plan/NCMN/O together/ADVN/O //
(a)	(b)

Fig. 1. Example of Date corpus in (a) original and (b) English translated text.

Table 5. Format of named entity types in each category.

Category	Format	Description	Example
Date	B-DAT	Beginning of a date	วันที่ (Date)
	I-DAT	Inside of a date	14 กุมภาพันธ์ (February 14)
Location	B-LOC	Beginning of a location name	เมือง (City)
	I-LOC	Inside of a location name	นิวยอร์ค (New York)
Measurement	B-MEA	Beginning of a measurement name	ห้า (Five)
	I-MEA	Inside of a measurement name	เล่ม (Books)
Name	B-NAM	Beginning of any proper name except location, person, and organization names, e.g., name of competition, name of position, etc.	ศึก (League)
	I-NAM	Inside of any proper name	ลาลีกา (La Liga)
Organization	B-ORG	Beginning of an organization name	บริษัท (Corp.)
	I-ORG	Inside of an organization's name	โตโยต้า มอเตอร์ (Toyota Motor)
Person	B-PER	Beginning of a person name	นาย (Mister)
	I-PER	Inside of a person's name	ณัฐวุฒิ สะกิดใจ (Natthawut Sakidjai)
Time	B-TIM	Beginning of a time	สิบ (Ten)

	I-TIM	Inside of a time	นาฬิกา (O'clock)
Other	O	Word does not belong to any type of entity	

4.1 Corpus Errors

Unfortunately, the THAI-NEST corpus has some limitations and is full of mistakes that directly affect the modeling process of NER. These defects are described as follows.

4.1.1 Errors of word segmentation

Important issues in the corpus are mistakes in word segmentation. Some characteristics of the Thai language have a profound effect on word segmentation because the language does not have any space between words. This makes it difficult to identify the boundary of each word. Furthermore, if the word segmentation in the corpus is incorrectly annotated, this will affect the NER modeling. Fig. 2 shows an example of mistakes in word segmentation. They result in errors of POS tagging and NE annotation.

```
มี/VSTA/O
./PUNC/O
ค/NLBL/O
./PUNC/O
```

(a)

```
นายก/NCMN/O
<space>/PUNC/O
อบ/VACT/O
จ./NTTL/O
อุตรดิตถ์/NPRP/O
```

(b)

Fig. 2. Example of mistakes in word segmentation, in (a) Date and (b) Name corpus.

4.1.2 Errors of named entity tag assignment

Due to incorrect word segmentation, the POS of the word can also be affected. Moreover, the errors of word segmentation and POS will cause the NE tags to be incorrectly defined, as shown in Fig. 3.

```
ร้อยตำรวจเอก/NTTL/B-PER
เฉลิม/NPRP/I-PER
<space>/PUNC/O
อยู่/XVAE/O
บำรุง/VACT/O
```

Fig. 3. Wrong named entity tagging of surname.

4.1.3 Named entity tagging inconsistency

The inconsistency problem also occurs in named entity tagging. Fig. 4 presents an example of named entity tag inconsistency in the same file. The word "ประเทศไทย" is annotated as a location (B-LOC) in one place while it is annotated as other (O) in another place.

```
  ราคา/NCMN/O
  ทองคำ/NCMN/O
  ใน/RPRE/O
* ประเทศไทย/NPRP/B-LOC
  ที่/PREL/O
  ปรับตัว/VACT/O
  สูงขึ้น/ADVN/O
```

```
  นายก/NCMN/O
  สมาคม/NCMN/O
  ลูกจ้าง/NCMN/O
  ส่วน/NCMN/O
  ราชการ/NCMN/O
  แห่ง/NPRP/O
* ประเทศไทย/NPRP/O
```

Fig. 4. Inconsistency of NE tagging in Location file.

5. Methodology

In this experiment, we present the NER model for the Thai language, which was inspired by the research of [13]. In this section, we describe the process of preparing data. We also describe the details of our proposed model in a bottom to top level manner.

5.1 Model

Our model consists of five important layers, as follows: 1) Word Embedding, 2) Character-level Representation, 3) Part-of-Speech Embedding, 4) Bi-LSTM layer, and 5) CRF layer. The architecture of the model is shown in Fig. 5.

5.1.1 Word embedding

Word embedding is a type of word representation that allows words with similar meanings to be understood by machine learning. It is a mapping of words into a real number vector. The word vector can be calculated from the context around that word.

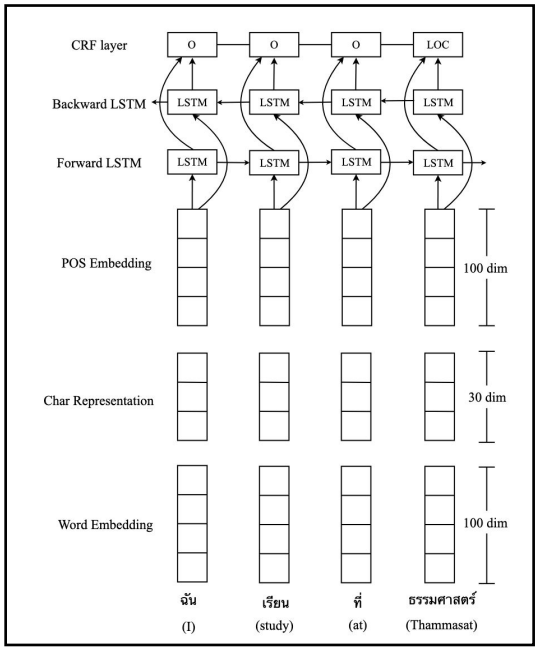


Fig. 5. Architecture of our NER model.

Word embedding can effectively extract semantic and syntactic information among words [6]. In addition, many previous studies, e.g. [25], support the advantages of word embedding. In our experiments, the skip-gram model with 300 dimensions and a window size of three words (three words before and three words after) from Word2Vec was used to pre-train the word embedding.

5.1.2 Thai Character Cluster (TCC)-level Representation

The character-level representation can extract morphological information from words and is very useful, especially for languages that have a complex word structure or morphologically rich languages, for example, the Hindi [15], Korean, and Thai languages. In Korean, a word usually contains several syllables [26]. Each Korean syllable consists of three parts: a consonant, a vowel, and a final consonant (if any), similar to the Thai language, for example, the word "한글" (Korean language). This word can be separated into two syllables "한" and "글". The first syllable "한" is composed of three characters: the consonant "ㅎ", the vowel "ㅏ" and the final consonant "ㄴ". The second syllable "글" also contains three characters: the consonant "ㄱ", the vowel "ㅡ" and the final consonant "ㄹ". There are several Korean NER research works that use syllables [27] or morphemes as input for NER tasks. Therefore, the internal syllables of the word play an important role in modeling. The characteristics of the Thai language, such as the word structure, are similar to the Korean language. Therefore, we consider that each cluster of the TCCs is equivalent to each syllable in Korean and can also preserve more complex information. The TCC is an unambiguous and inseparable unit that is smaller than a word but larger than a character and cannot be further divided, according to the Thai language spelling rules regarding the grouping of these characters [28-29]. For example, a vowel sign and a tone mark cannot stand alone. They must be placed with a base consonant character. This also solves the problem of the tone mark and vowel sign separation. For example, the TCC expression of a word can be "ป|ระ|เท|ศ|ไท|ย" (Pra-Thed-Thai; Thailand) and "นา|ย|ก|รัฐ|มน|ตรี" (Na-Yok-Rad-Tha-Mon-tri; Prime minister). Hence, we have hypothesized that using only character embedding may not greatly improve the performance of our NER model; however, TCCs can successfully handle this problem and provide better results. [30] and [31] have proposed a CNN model for sentence classification and text classification, respectively. This allows us to realize that CNN can provide good performance for NLP and character embedding. We thus use CNN layers to create the character cluster-level representations of the model. The details of the CNN layer are shown in Fig. 6.

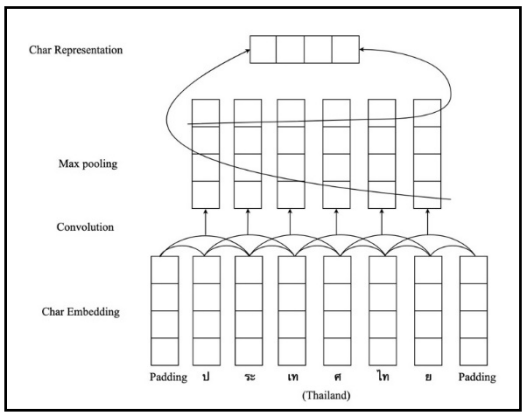


Fig. 6. Convolution Neural Network for character-level representation.

5.1.3 POS embedding

Researchers have considered that the Part-of-Speech (POS) can help increase the efficiency of the model because most of the Thai named entities will be close to or adjacent to the part-of-speech (verb or preposition). In addition, many previous studies support the use of the POS with both Indonesian and Chinese. The POS of each word is encoded into the one-hot vector format in the embedding layer.

5.1.4 Bi-directional LSTM

An RNN has effective performance in many NLP tasks. LSTM is an RNN that has the ability to capture long-term dependencies efficiently and can retrieve rich global information. Furthermore, the information from previous words is useful for prediction, and the information from words coming after is also useful. The previous features are extracted by a forward LSTM layer, and the future features are captured by a backward LSTM layer. This pair of forward and backward LSTM layers is referred to as a bidirectional LSTM. In this way, we can effectively utilize the previous states and future states to learn a sequence of words.

5.1.5 Conditional Random Field (CRF)

The CRF model is widely used in NLP for predicting the sequence of labels with the most likely tendency, corresponding to the given label sequence. The CRF model takes advantage of the neighboring tag information and considers the previous context in predicting the current tag. For example, I-LOC cannot be followed by I-PER in a sentence. Therefore, we consider that it is appropriate for CRF to be the last layer to predict the named entity tag of each word. We followed [4] to generate our linear-chain CRF layer.

We combined each layer to create the BiLSTM-CNN-CRF model for predicting named entity tags. The character-level representation of each word is calculated by the CNN, as shown in Figure 6. From each embedding step, we obtain the vector representations of the words, POS tags, and character clusters. Then, these vectors are concatenated before being fed into the Bi-LSTM layer. We apply dropout layers to both the input and output vectors of Bi-LSTM to prevent overfitting and to regularize the model. The dropout works by randomly dropping out nodes from the network during training. Finally, the output vectors of the Bi-LSTM layer are passed through the CRF layer and decoded via the Viterbi algorithm (part of the CRF layer) to select the most possible sequence of the named entity tag. The model has learned these vectors in order to improve its ability to predict target words from the vector of the surrounding context.

5.2 Experiment

5.2.1 Pre-processing data

As mentioned in Section 4, the corpus that we use is the modified THAI-NEST corpus. The format of the named entity tags in this corpus is the BIO annotation scheme. The writing style of the Thai language often omits the indicators or prefixes of the named entity, which means that some words in the corpus have a prefix but some words do not. As a result, it is not possible to separately measure the score of the B-tag or I-tag. Therefore, the format of the named entity tag must be changed from the BIO to the IO format, to solve this problem.

In addition, we use a Python dictionary to map each word to a corresponding integer ID and to do the same with the character cluster, POS, and named entity tag. Representing these things as unique integers saves a large amount of memory and is faster.

5.2.2 Experiment Setup

In this experiment, each corpus is divided into three parts: 80% of all sentences in the file for the training set, 10% for the validation set and the last 10% for the testing set.

Regarding our neural network model, all parameters need to be set. We tune the value of parameters on the development set, to get the most suitable final parameters. All parameters are displayed in Table 6.

Table 6. All parameters for the model.

Parameter	Setting
Char_dim	30
Character-level CNN filters	30
Character-level CNN window size	3
Word_dim	100
Word_LSTM_dim	200
Word_bidirection	TRUE
POS_dim	100
Dropout_rate	0.5
Batch size	10
Learning rate (initial)	0.01
Decay rate	0.5
Gradient clipping	5.0
Learning method	SGD
Training epoch	60

5.3 Evaluation

The performance of the models is evaluated in terms of the F1-score. In this experiment, we focus only on the predictive performance for the main category in each file. The performance for the "Other" category in each file was ignored. Some words are not actually in the Other category, but are in some other main categories.

6. Results and Discussion

To show the efficiency of our proposed model, we experimented with other baseline models on the same dataset. Table 7 shows a comparison of the performance of each model.

According to the results shown in Table 7, the F1-scores of the BiLSTM (Word+POS) model and BiLSTM-CNN (Word+POS+TCC) model are significantly higher than the BiLSTM (Word) model, showing that POS and TCC both play an important role in the named entity prediction for Thai NER tasks. Finally, adding a CRF layer to the BiLSTM-CNN-CRF (Word+POS+TCC) model greatly improves the model performance and provides the highest F1-scores compared to the other baseline models, especially the F1-scores of the Date and Time corpus (approximately 90 percent). This indicates that jointly decoding the label sequences is very useful for the final step of the model.

A comparison of the performance between the model with TCC and the one without TCC is explained in the following paragraph.

Table 7. Performance of our model and other baseline models.

NE	F1-score					
	BiLSTM (Word)		BiLSTM-CNN-CRF (Word+TCC)		BiLSTM-CNN-CRF (Word+POS+Char)	
	No W2V	W2V	No W2V	W2V	No W2V	W2V
DAT	77.51	79.20	82.18	84.65	90.02	92.43
LOC	72.00	75.33	76.90	79.26	84.51	87.07
MEA	69.74	72.41	75.04	77.53	82.94	85.66
NAM	65.29	67.18	71.26	73.84	80.40	82.25
ORG	70.41	73.56	75.63	77.90	82.28	85.79
PER	69.88	74.29	78.57	81.72	83.55	87.32
TIM	79.25	82.77	83.13	86.05	90.16	93.88

NE	F1-score					
	BiLSTM (Word+POS)		BiLSTM-CNN (Word+POS+TCC)		BiLSTM-CNN-CRF (Word+POS+TCC)	
	No W2V	W2V	No W2V	W2V	No W2V	W2V
DAT	84.07	86.14	87.35	89.72	90.77	93.21
LOC	80.22	83.67	83.10	86.24	85.63	88.93
MEA	77.52	80.22	79.85	82.67	83.88	86.52
NAM	72.16	75.48	76.51	80.13	81.29	84.92
ORG	78.54	81.17	81.44	84.75	85.76	87.31
PER	76.08	82.35	80.94	85.07	84.51	88.90
TIM	84.51	88.12	88.59	91.36	91.35	94.76

W2V = Word2Vec library

Several examples from the prediction of each model are considered step by step. The second column is the named entity type derived from the model prediction. First, we start with an example of BiLSTM as shown in Fig. 7.

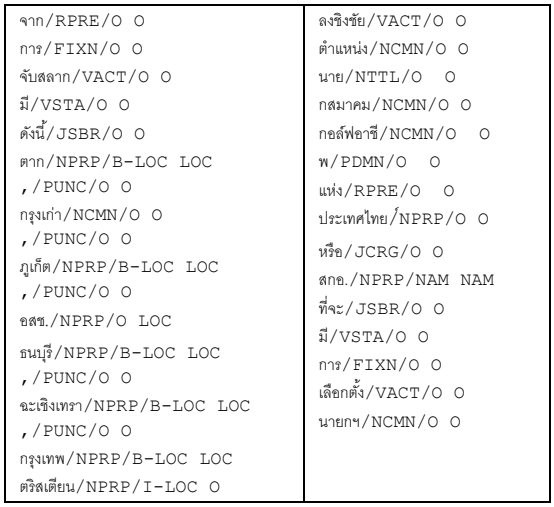
จาก/RPRE/O O การ/FIXN/O O จับสลาก/VACT/O O มี/VSTA/O O ดังนี้/JSBR/O O ตาก/NPRP/B-LOC O ,/PUNC/O O กรุงเก่า/NCMN/O O ,/PUNC/O O ภูเก็ต/NPRP/B-LOC LOC ,/PUNC/O O อสช./NPRP/O O ธนบุรี/NPRP/B-LOC LOC ,/PUNC/O O ฉะเชิงเทรา/NPRP/B-LOC LOC ,/PUNC/O O กรุงเทพ/NPRP/B-LOC LOC ตริสเตียน/NPRP/I-LOC O	ลงชิงชัย/VACT/O O ตำแหน่ง/NCMN/O O นาย/NTTL/O O กสมาคม/NCMN/O O กอล์ฟอาชี/NCMN/O O พ/PDMN/O O แห่ง/RPRE/O O ประเทศไทย/NPRP/O O หรือ/JCRG/O O สกอ./NPRP/NAM O ที่จะ/JSBR/O O มี/VSTA/O O การ/FIXN/O O เลือกตั้ง/VACT/O O นายกฯ/NCMN/O O
(a)	(b)

Fig. 7. Predicted named entity type of BiLSTM with the word model of (a) location corpus file, and (b) name corpus file.

As shown in Fig. 7(a), for the location file, the model cannot predict the words “ตาก” (Tak), “กรุงเก่า” (Krung-Kao), “อสช.” (Aor-Sor-Chor), and “ตริสเตียน” (Tris-Tien) correctly. Actually, the word “ตริสเตียน” (Tris-Tien) is a typographical error, as this word should be “คริสเตียน” (Kris-Tien). For the word “ตาก” (Tak), it is difficult to distinguish between a proper noun (the name of a province in Thailand) and a verb (to dry in the air). In this case, it correctly labels the word as a location. In addition, the word “กรุงเก่า” (Krung-Kao) refers to Ayutthaya, but it has the wrong POS type. For the Name file in Fig. 7(b)., “นายกสมาคมกอล์ฟอาชีพแห่งประเทศไทย” (President of the Professional Golf Association of Thailand) has been segmented incorrectly as “นาย-กสมาคม-กอล์ฟอาชี-พ-แห่ง-ประเทศไทย”. Thus, the model cannot predict the named entity type of each word, including the abbreviation, “สกอ.” (Sor-Kor-Aor).

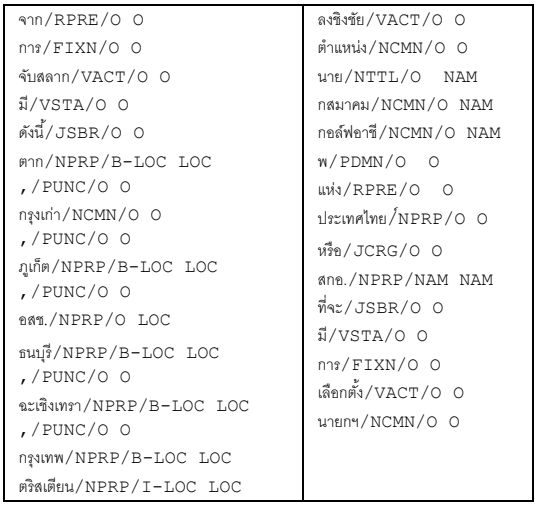
In Fig. 8, the terms “ตาก” (Tak), “อสช.” (Aor-Sor-Chor), and “สกอ.” (Sor-Kor-Aor) are labeled correctly by the BiLSTM model with POS. This means that POS influences the named entity prediction of the model and solves the problems of category ambiguity and abbreviations. Unfortunately, in the case of mistaken word segmentation, the model still cannot predict the named entity type correctly.

Once we applied TCC to the BiLSTM in the BiLSTM (Word + POS + TCC) model, the model provides better results and can cope with the problems of misspelled words and mistaken word segmentation. The word “ตริสเตียน” (Tris - Tien) is labeled as a location, as



(a) (b)

Fig. 8. Predicted named entity type of BiLSTM with the word and POS model of (a) location corpus file, and (b) name corpus file.

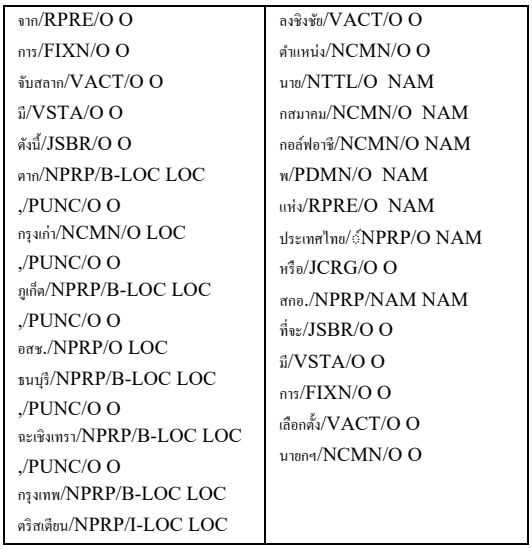


(a) (b)

Fig. 9. Predicted named entity type of BiLSTM-CNN with the word, POS, and TCC model of (a) location corpus file, and (b) name corpus file.

shown in Fig. 9(a) and the model also predicts the word "นายกสมาคมกอล์ฟอาชีพแห่งประเทศไทย" (President of the Professional Golf Association of Thailand) more accurately, although it cannot predict some of the words correctly, which are "พ" (Phor), "แห่ง" (Heang), and "ประเทศไทย" (Pra-Thed-Thai), as shown in Fig. 9(b).

Finally, as shown in Fig. 10., the BiLSTM-CNN-CRF (Word+POS+TCC) model predicts "กรุงเก่า" (Krung-Kao) in the location file and "แห่ง" (Heang) and "ประเทศไทย" (Pra-Thed-Thai) in the name file correctly. Since CRF predicts the sequence of labels with the most likely tendency that corresponds to the sequence of the given input sentences, BiLSTM can effectively capture the sequence of relationships between words. Then, CRF calculates the joint probability distribution and allows for



(a) (b)

Fig. 10. Predicted named entity type of BiLSTM-CNN-CRF with the word, POS, and TCC model of (a) location corpus file, and (b) name corpus file.

optimal prediction of all the labels in the sentence, capturing the relationships at the label level.

7. TCC vs. Single Character in Character-level Representation

To prove the hypothesis that TCCs provide better results and performance than using single characters at the character-level, we conducted another experiment using single Thai characters instead of TCCs in our proposed model on the same dataset. As shown in Table 7, the F1-scores of the model with TCCs are significantly higher than the model that uses single Thai characters. Fig. 11. shows the comparison of the sample results of both models. The example, "บริษัทอิมแพ็ค แมนเนจเม้นท์ จำกัด" (Impact Management Co., Ltd.), is incorrectly predicted by the model using single Thai characters. The model may not be able to learn this word because of the incorrect word segmentation. The model considers only a single internal character for word embedding, which is not enough to help the model predict correctly. In contrast, the model that applies TCCs can successfully handle this problem.

We also investigate the efficiency in processing time and memory usage. We trained each model on a GPU server and found that the model trained with TCCs consumed less processing time with less memory usage. For example, the model trained with TCCs took 13 hours and used 1,689 MB of memory in the case of the Location file. The model with single Thai characters took 15 hours and used more than double the memory, which was 3,371 MB. Furthermore, TCCs also helped in reducing the training loss value and yielded improvements over single character embedding. The experimental results have shown that the efficiency and ability of the TCC-level representation are superior to those of the character-level representation.

บริษัท/NCMN/B-ORG ORG
อิมแพ็ค/NCMN/I-ORG ORG
เอ็กซิบิชั่น/NTTL/I-ORG ORG
แมเนจเม้/NCMN/I-ORG O
นท์ จำกัด/NCMN/I-ORG O
ได้/XVAM/O O
เปิด/VACT/O O
ตัว/CNIT/O O

(a)

บริษัท/NCMN/B-ORG ORG
อิมแพ็ค/NCMN/I-ORG ORG
เอ็กซิบิชั่น/NTTL/I-ORG ORG
แมเนจเม้/NCMN/I-ORG ORG
นท์ จำกัด/NCMN/I-ORG ORG
ได้/XVAM/O O
เปิด/VACT/O O
ตัว/CNIT/O O

(b)

Fig. 11. Prediction results from (a) model with single character, and (b) with TCCs.

8. Comparison with other methods

To evaluate the performance of Thai named entity recognition of our method and to compare the results with other methods, we also use F1-score as our evaluation metrics.

We compare our proposed method with other methods on all the NE types in the THAI-NEST corpus to show our considerable performance advantage over the other methods. The results of each method are provided in Table 8.

Table 8. Comparison of our proposed method with state-of-the-art methods on THAI-NEST corpus.

NE	Methods (F1-score)		
	CRF	HMM	**BiLSTM-CNN-CRF**
DAT	81.54	84.37	**93.21**
LOC	72.68	76.25	**88.93**
MEA	69.13	72.04	**86.52**
NAM	65.81	68.70	**84.92**
ORG	76.79	80.62	**87.31**
PER	73.08	78.19	**88.90**
TIM	75.42	81.03	**94.76**

For the CRF model used in this comparison is the CRF++ which was developed by Taku Kudo and is free for research proposes (Available: https://taku910.github.io/crfpp/). The results of the proposed method outperform the other methods. The BiLSTM is more effective than the CRF and HMM trigram models in the Thai NER task due to the BiLSTM being able to store the data to long-term memory and learn what to keep or ignore while the CRF uses only word lists, unigram, and bigram. For the HMM model, we predict NE tag for input words by using the Viterbi algorithm to find the tagging sequence with the highest probability. Both CRF and HMM learn only from the current word and adjacent words. Besides, adding a CNN and CRF layer to BiLSTM increase the predictability of the type of named entities.

9. Combined Corpus

Another important issue of this corpus is that it is disjointedly managed into seven files according to the type of named entity. As a result, the model cannot be trained with all types of named entities at once. For this reason, we recognize the importance of this issue and conduct cross annotation among the seven single-labeled files. As a result, the number of named entity tags in each file is greatly increased. A sample of the corpus that combines all named entity types is shown in Fig. 12.

%Title: Date corpus %Description: Date in any format %Number of sentence: 2,783 %Number of word: 272,753 %Number of named entity tag: 14,330 %Date: January 6, 2019 %Creator: Kitiya Suriyachay and Virach Sornlertlamvanich %Email: m5922040075@g.siit.tu.ac.th and virach@siit.tu.ac.th %Affiliation: Sirindhorn International Institute of Technology, Thammasat University #S1 นายสุเทพ เทือกสุบรรณ รองนายกรัฐมนตรี กล่าวว่า ในวันพรุ่งนี้ (18 มี.ค.52) รัฐบาลโดย\\ นายอภิสิทธิ์ เวชชาชีวะ นายกรัฐมนตรี จะมอบนโยบายและแนวทางในการป้องกันและปราบปราม\\ ยาเสพติดให้กับส่วนราชการต่างๆเพื่อบูรณาการแผนปฏิบัติการป้องกันและปราบปรามยาเสพติดร่วมกัน// นาย/NTTL/B-PER สุเทพ/NPRP/I-PER <space>/PUNC/I-PER เทือกสุบรรณ/NPRP/I-PER <space>/PUNC/O รองนายกรัฐมนตรี/NCMN/O <space>/PUNC/O กล่าว/VACT/O ว่า/JSBR/O <space>/PUNC/O ใน/RPRE/O วันพรุ่งนี้/ADVS/B-DAT <space>/PUNC/O (/PUNC/O 18/DONM/B-DAT <space>/PUNC/I-DAT	%Title: Date corpus %Description: Date in any format %Number of sentence: 2,783 %Number of word: 272,753 %Number of named entity tag: 14,330 %Date: January 6, 2019 %Creator: Kitiya Suriyachay and Virach Sornlertlamvanich %Email: m5922040075@g.siit.tu.ac.th and virach@siit.tu.ac.th %Affiliation: Sirindhorn International Institute of Technology, Thammasat University #S1 Mr. Suthep Thaugsuban, Deputy Prime Minister, said that tomorrow (18 Mar 2009) the\\ government by Prime Minister Abhisit Vejjajiva will give policies and guidelines for\\ prevention and suppression of drugs to government agencies to integrate the drug\\ prevention and suppression action plan together// Mr./NTTL/B-PER Suthep/NPRP/I-PER <space>/PUNC/I-PER Thaugsuban/NPRP/I-PER <space>/PUNC/O Deputy Prime Minister/NCMN/O <space>/PUNC/O said/VACT/O that/JSBR/O <space>/PUNC/O tomorrow/ADVS/B-DAT <space>/PUNC/O (/PUNC/O 18/DONM/B-DAT <space>/PUNC/I-DAT

มี.ค. 52/NPRP/I-DAT	Mar 09/NPRP/I-DAT
)/PUNC/O	)/PUNC/O
.	.
.	.
ยาเสพติด/NCMN/O	plan/NCMN/O
ร่วมกัน/ADVN/O	together/ADVN/O
//	//
(a)	(b)

Fig. 12. Sample of combined corpus in (a) original, and (b) English translated text.

After combining all seven named entity types, the number of named entities in each corpus file increases significantly. The number of named entity tags in each corpus is listed in Tables 9-15.

Table 9. Number of each named entity type in the Date corpus file.

NE	DAT		
	B-x	I-x	B-x with I-x
DAT	4,523	9,804	3,779
LOC	2,235	1,487	707
MEA	6,000	8,031	4,444
NAM	4,296	4,541	1,597
ORG	5,448	4,072	1,624
PER	3,523	5,157	2,399
TIM	410	397	204

Table 10. Number of each named entity type in the Location corpus file.

NE	LOC		
	B-x	I-x	B-x with I-x
DAT	5,261	8,519	3,479
LOC	20,527	8,329	4,467
MEA	14,704	19,328	10,993
NAM	10,047	10,315	3,712
ORG	17,731	12,219	5,179
PER	11,585	14,757	7,478
TIM	632	599	360

Table 11. Number of each named entity type in the Measurement corpus file.

NE	MEA		
	B-x	I-x	B-x with I-x
DAT	1,088	1,820	670
LOC	1,383	1,541	655
MEA	5,244	12,127	4,558
NAM	3,289	2,714	1,183
ORG	2,760	2,057	869
PER	2,975	4,482	2,150
TIM	288	298	139

Table 12. Number of each named entity type in the Name corpus file.

NE	NAM		
	B-x	I-x	B-x with I-x
DAT	4,530	7,636	3,084
LOC	5,016	3,028	1,457
MEA	13,395	17,632	9,864
NAM	16,759	23,766	9,507
ORG	14,029	10,501	4,334
PER	10,343	13,841	6,778
TIM	539	511	266

Table 13. Number of each named entity type in the Organization corpus file.

NE	ORG		
	B-x	I-x	B-x with I-x
DAT	11,564	18,614	7,710
LOC	14,030	6,047	3,199
MEA	30,923	40,175	22,733
NAM	26,848	24,942	9,113
ORG	49,397	46,204	20,607
PER	26,614	35,723	17,471
TIM	1,318	1,232	702

Table 14. Number of each named entity type in the Person corpus file.

NE	PER		
	B-x	I-x	B-x with I-x
DAT	20,215	31,270	12,554
LOC	20,034	14,487	6,760
MEA	52,394	66,552	38,412
NAM	53,391	51,128	19,456
ORG	60,213	47,224	20,535
PER	75,287	146,789	63,508
TIM	3,535	3,532	1,933

Table 15. Number of each named entity type in the Time corpus file.

NE	TIM		
	B-x	I-x	B-x with I-x
DAT	537	852	305
LOC	459	569	219
MEA	993	1,270	747
NAM	870	936	374
ORG	663	642	268
PER	995	1,519	686
TIM	495	1,048	459

10. Conclusions

In this paper, we present an approach of NER for the Thai language by using our proposed BiLSTM-CNN-CRF model with the word, POS, and Thai character cluster (TCC) features. Our model provides the best performance of the past reported approaches. It can deal with problems of named entity tagging inconsistency in the same file, especially errors of word segmentation. The experimental results show that the major factor influencing the decision-making for the types of named entities is the surrounding context, such as verbs and prepositions. Furthermore, TCCs play an important role in solving the problems related to word segmentation errors, allowing the model to accurately predict the types of NEs even if the word is not correctly segmented. TCCs also improve the efficiency in generating the word embedding layer and are useful for morphologically rich languages instead of using only word embedding. We expect to extend our TCC model to improve the NER tasks in other orthographically similar languages, such as Lao, Burmese, and Cambodian.

References

[1] Chieu HL, Ng HT. Named Entity Recognition: A Maximum Entropy Approach Using Global Information. Proceedings of CoNLL-2003. 2003.

[2] Chiu JP, Nichols E. Named Entity Recognition with Bidirectional LSTM-CNNs. Transactions of the Association for Computational Linguistics. 2016;4:357-70.

[3] Luo G, Huang X, Lin C-Y, Nie Z. Joint Entity Recognition and Disambiguation. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015;879-88.

[4] Ma X, Hovy E. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016;1:1064-74.

[5] Rachman V, Savitri S, Augustianti F, Mahendra R. Named Entity Recognition on Indonesian Twitter Posts Using Long Short-Term Memory Networks. 2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS). 2017.

[6] Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. Neural Architectures for Named Entity Recognition. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 2016.

[7] Limsopathan N, Collier N. Bidirectional LSTM for Named Entity Recognition in Twitter Messages. Proceedings of the 2nd Workshop on Noisy User-generated Text. 2016;145-52.

[8] Chopra D, Joshi N, Mathur I. Named Entity Recognition in Hindi Using Hidden Markov Model. 2016 Second International Conference on Computational Intelligence & Communication Technology (CICT). 2016.

[9] Ju Z, Wang J, Zhu F. Named Entity Recognition from Biomedical Text Using SVM. 2011 5th International Conference on Bioinformatics and Biomedical Engineering. 2011.

[10] Liu K, Hu Q, Liu J, Xing C. Named Entity Recognition in Chinese Electronic Medical Records Based on CRF. 2017 14th Web Information Systems and Applications Conference (WISA). 2017.

[11] Salleh MS, Asmai SA, Basiron H, Ahmad SA. Malay Named Entity Recognition Using Conditional Random Fields. 2017 5th International Conference on Information and Communication Technology (ICoIC7). 2017.

[12] Charoenpornsawat P, Kijsirikul B, Meknavin S. Feature-based Proper Name Identification in Thai. Proceeding of

National Computer Science and Engineering Conference: NCSEC'98. 1998.

[13] Saetiew N, Achalakul T, Prom-On S. Thai Person Name Recognition (PNR) Using Likelihood Probability of Tokenized Words. 2017 International Electrical Engineering Congress (iEECON). 2017.

[14] Suwanno N, Suzuki Y, Yamazaki H. Selecting the Optimal Feature Sets for Thai Named Entity Extraction. Proceedings of ICEE-2007 & PEC. 2007.

[15] Tirasaroj N, Aroonmanakun W. Thai Named Entity Recognition Based on Conditional Random Fields. 2009 8th International Symposium on Natural Language Processing. 2009.

[16] Li L, Jin L, Jiang Z, Song D, Huang D. Biomedical Named Entity Recognition Based on Extended Recurrent Neural Networks. 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2015;649-52.

[17] Suriyachay K, Sornlertlamvanich V. Named Entity Recognition Modeling for the Thai Language from a Disjointedly Labeled Corpus. 2018 5th International Conference on Advanced Informatics: Concept Theory and Applications (ICAICTA). 2018;30-5.

[18] Udomcharoenchaikit C, Vateekul P, Boonkwan P. Thai Named-Entity Recognition Using Variational Long Short-Term Memory with Conditional Random Field. Advances in Intelligent Systems and Computing Advances in Intelligent Informatics, Smart Technology and Natural Language Processing. 2018;82-92.

[19] Chen X, Xu L, Liu Z, Sun M, Luan H. Joint Learning of Character and Word Embeddings. Proceeding of the 24th International Joint Conference on Artificial Intelligence (IJCAI 2015). 2015.

[20] Wang Y, Xia B, Liu Z, Li Y, Li T. Named Entity Recognition for Chinese Telecommunications Field Based on Char2Vec and Bi-LSTMs. 2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE). 2017;1-7.

[21] Wang W, Bao F, Gao G. Mongolian Named Entity Recognition with Bidirectional Recurrent Neural Networks. 2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI). 2016.

[22] Theeramunkong T, Boriboon M, Haruechaiyasak C, Kittiphattanabawon N, Kosawat K, Onsuwan C, Siriwat I, Suwanapong T, Tongtep N. THAI-NEST: A Framework for Thai Named Entity Tagging Specification and Tools. 2010.

[23] Sornlertlamvanich V, Charoenporn T, Isahara H. ORCHID: Thai Part-Of-Speech Tagged Corpus: Technical Report. TR-NECTEC-1997-001, National Electronics and Computer Technology Center, Thailand. 1997.

[24] Suriyachay K, Charoenporn T, Sornlertlamvanich V. Thai Named Entity Tagged Corpus Annotation Scheme and Self Verification. Proceedings of the 9th Language & Technology Conference (LTC2019). 2019;131-7.

[25] Maimaiti M, Wumaier A, Abiderexiti K, Yibulayin T. Bidirectional Long Short-Term Memory Network with a Conditional Random Field Layer for Uyghur Part-Of-Speech Tagging. Information. 2017;8(4):157.

[26] Na S-H, Kim H, Min J, Kim K. Improving LSTM CRFs Using Character-Based Compositions for Korean Named Entity Recognition. Computer Speech & Language. 2019;54:106-21.

[27] Kwon S, Ko Y, Seo J. Effective Vector Representation for the Korean Named-Entity Recognition. Pattern Recognition Letters. 2019;117:52-57.

[28] Sornlertlamvanich V, Tanaka H. The Automatic Extraction of Open Compounds from Text Corpora. Proceedings of the 16th conference on Computational linguistics. 1996.

[29] Sornlertlamvanich V, Tanaka H. Extracting Open Compounds from Text Corpora. Proceeding of the 2nd Annual Meetings of the Association for Natural Language Processing. 1996.

[30] Kim Y. Convolutional Neural Networks for Sentence Classification. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014.

[31] Zhang X, Zhao J, LeCun Y. Character-Level Convolutional Networks for Text Classification. Proceeding of International Conference on Neural Information Processing Systems. 2015.

[32] Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space. arXiv preprint arXiv:1301.3781. 2013.