






Thai Named Entity Corpus Annotation Scheme and Self Verification by BiLSTM-CNN-CRF

Virach Sornlertlamvanich^{1,3} , Kitiya Suriyachay² ,
and Thatsanee Charoenporn¹ 

¹ Faculty of Data Science, Musashino University, Tokyo, Japan
{virach, thatsanee}@ds.musashino-u.ac.jp

² School of ICT, Sirindhorn International Institute of Technology,
Thammasat University, Pathumthani, Thailand
m5922040075@g.sii.tu.ac.th

³ Faculty of Engineering, Thammasat University, Pathumthani, Thailand

Abstract. Corpus is one of the essential parts of language research, especially for the low resource language. To ensure the researching result to be most effective, the corpus that has been used also requires effectiveness and accuracy. The Thai language has some special characteristics that cause difficulty in building the corpus and affect the error of those corpora. Therefore, this paper proposes an effective and efficient approach to clean up the existing Named Entity corpus before using it in any language research. The THAI-NEST corpus is adopted to verify the consistency and integrity of the data and re-design with our proposed model. The revised corpus is verified by the BiLSTM-CNN-CRF model that combined the features among word, POS, and Thai character clusters (TCCs). Experimental results show the effectiveness of the verification, which increased the accuracy by up to 12%, and the model can effectively detect and handle errors of word segmentation and NE tag consistency.

Keywords: Corpus annotation · Named entity recognition · Thai named entity · Thai corpus

1 Introduction

Methodical and accurate use of this enormous amount of data is the key to the success of many businesses. Many organizations have invested in developing information extraction and retrieval to use the data and information efficiently. Information Extraction (IE) is an automated process of extracting specific information from unstructured data, for example, the process of extracting names, addresses, and phone numbers from a web page. Named Entity Recognition (NER) is a subtask of IE that aims to recognize and classify specific entities in focused texts such as person names, locations, and organizations.

NER has gained popularity over several years. NER in the Thai language is more challenging and complex than English or other European languages. Due to there is no capitalization or special characters to identifying named entities. In addition, there is no space or word boundary in a sentence causing difficulties in word segmentation and the

ambiguity between common nouns and proper nouns. Moreover, incorrect word segmentation causes problems for named entity recognition and directly affects the accuracy of the NER process. The ambiguity of homographs can also provide different meanings depending on the context.

For example, “พันตำรวจเอกทวี สอดส่อง อธิบดีกรมสอบสวนคดีพิเศษ”(Colonel Tawee Sordsong Director-General of the Department of Special Investigation) the word “สอดส่อง” in this context refers to a surname of the person. However, it also means “to observe” in the general context. Thai writing styles are also a problem for NER. In general, a proper noun is written with a common noun or prefix that specifies the type of the proper noun, but Thais often use a shorter name, abbreviation, or cut its prefix. “ธรรมศาสตร์ร่วมกับเอกชนสร้างสรรค์ไอเดีย”(Thammasat collaborates with private companies to create IT ideas.), “มหาวิทยาลัยธรรมศาสตร์ก่อตั้งขึ้นในปี ค.ศ. 1934”(Thammasat University was established in 1934), and “ม.ธ. ประกาศกำหนดการในภาคการศึกษาแรก”(TU announces the schedule for the first semester.), from the examples, Thammasat University is used in three different ways including “ธรรมศาสตร์”(Thammasat) “มหาวิทยาลัยธรรมศาสตร์”(Thammasat University) and “ม.ธ.”(the abbreviation of Thammasat University). Another problem of named entity in the Thai language is that named entity in different types can be the same word, such as “สุรินทร์”(Surin), which can be both a name of the person or Thailand province (location name), depending on the context of the word.

Furthermore, the consistency of named entity tags in the corpus is also an important issue because inconsistency leads to wrong name entity recognition. Corpus which is used in this paper is the THAI- NEST corpus [18]. The corpus is disjointedly generated into seven files, and each file is exclusively for each main category. In addition, the corpus contains some inconsistency of named entity tags due to the word segmentation process on the corpus. To solve these problems, we propose a method to clean up and verify the existing corpus for the Thai NER. We also performed cross annotation among the separate seven named entity tagged files to enhance the number of NE tags and to prepare for further developing the NER model.

The previous related researches are explained in Sect. 2. Section 3 describes the corpus that has been used in this research. The methodology to improve the consistency and correctness of the corpus is showed in Sect. 4. Then, in Sects. 5 and 6, we presented the result of our experiment and combined corpus approach, respectively. The conclusion is described in Sect. 7.

2 Related Work

Research on NER is widely popular in many languages, so the NER tools have been gaining attention and continually improving. However, in many languages, a small number of the corpus is used for NLP tasks, while the study of named entities is quite limited. Therefore, having a small number of the corpus is not enough [4]. Several approaches can be used to solve these problems, but one of the most influential and popular approaches is based on machine learning techniques [5].

Nevertheless, there is a pretty limited Thai NE corpus. The famous Thai corpus is the Orchid corpus created in 1996 by collecting Thai text for more than two million

words and splitting all of the words with their part of speech. However, Thai did not have any clear boundary or punctuation. To build the Orchid corpus, they need to separate the paragraph into sentences and then from sentences to words before tagging each word with POS using trigram. Their trigram model will consider the word segmentation and POS tagging together within the model [15].

Many previous research papers show an Error Correction for many languages, such as Chinese, use a transformation-based error-driven machine learning technique to found error positions and produce error repairing rules [20]. The dictionary-based approach is easy, fast, and widely used, but this approach can only be used with known and unambiguous words. In Vietnamese [10], they use two entropy-based methods to detect error and inconsistency in the Vietnamese word-segmented and POS-tagged corpus. The first method is to rank the order of error candidates using a scoring function that depends on conditional entropy. The second method uses beam search to find a subset of error candidates that has been changed its label and finally leads to the decreasing of conditional entropy.

Some traditional machine learning techniques have been applied for NLP tasks over the past year. [1] introduced the NER model of the Hindi language by using Hidden Markov Model (HMM). [3] presented the Support Vector Machine (SVM) with word shape, and POS is used as features to recognize named entities in Biomedical Text such as genes, DNA, and protein.

[12] proposed Malay Named Entity Recognition using the CRF model. Some characteristics of Malay are employed for training models such as capitalization, lowercase, previous and neighboring word, word suffix, digit, word shape, and POS. [6] introduced the CRF model for Chinese electronic medical records recognition with bag-of-characters, part-of-speech, dictionary feature, and word clustering as features.

Deep Learning architectures have recently made significant advances in various fields. BiLSTM is a type of deep learning model used in many research studies. [11] conducted the NER model for recognizing Indonesian information on Twitter using Bi-LSTM with word embedding and POS tag and showed that both features provided the most F1-score. [19] presented the Deep Learning model for Chinese telecommunication information recognition using character embedding instead of word embedding. [16] applied word with POS as an input of the Bi-LSTM model to recognize the named entity in the Thai language.

Furthermore, there are several researches on NER that use a hybrid approach. [2] proposed a Bi-LSTM and CRF model for Chinese NER based on character and word embedding. [7] also presented Bi-LSTM-CNN-CRF model, which achieved performance on NER and POS tagging and successfully employed CNN to extract more useful character-level features. More recent work used the CRF as the last layer of the pipeline to handle the classification and provided the satisfactory results [17].

3 Corpus

The original corpus used in our research is the THAI-NEST corpus, which is collected from Thai online news published on the Internet including politics, economic, foreign, crime, sports, entertainment, education, and technology news [18]. The THAI-NEST corpus is already tokenized text into words and punctuation. The corpus is separated

into seven files based on named entity categories, consisting of date (DAT), time (TIM), measurement (MEA), name (NAM), location (LOC), person (PER), and organization (ORG). The first three characters abbreviate each category. The number of words and named entity tags in each file is listed in Table 1.

Table 1. Number of sentences, words, and named entity tags in each file

	No. of sentence	No. of word	No. of NE tag
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

3.1 Structure of the Corpus

In this experiment, the THAI-NEST corpus was designed and restructured based on the structural format of the Orchid corpus [15] as shown in Fig. 1. There are two types of mark-ups to differentiate text information line and numbering line in the corpus. Text information line beginning with “%” symbol, which is used to describe the additional information of the corpus as shown in Table 2. The numbering line begins with a “#” symbol, which is used to order the sequence of text in the corpus as shown in Table 3, and also there are four special mark-up characters 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 line

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

Table 4. Special mark-up characters

Mark-up	Description
\\	Line break symbol for the long text
//	End of sentence
/[POS]	Tag marker for POS annotation of a word
/[NE]	Tag marker for NE annotation of a word

<pre>%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 Somlertlamvanich %Email: m6922040075@siit.tu.ac.th and virach@siit.tu.ac.th %Affiliation: Sirindhorn International Institute of Technology, Thammasat University #S1 นายสุเทพ เทือกสุบรรณ รองนายกรัฐมนตรี กล่าวในวันพุธนี้ (18 มี.ค.52) รัฐบาลโดย\\ นายอภิสิทธิ์ เวชชาชีวะ นายกรัฐมนตรี จะมอบนโยบายและแนวทางในการป้องกันและ\\ ปราบปรามยาเสพติดให้กับส่วนราชการต่าง ๆ เพื่อบูรณาการแผนปฏิบัติการป้องกันและ\\ ปราบปรามยาเสพติดร่วมกัน// 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 วันพุธนี้/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 //</pre>	<pre>%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 Somlertlamvanich %Email: m6922040075@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/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 //</pre>
--	--

(a)

(b)

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

For the named entity tags, BIO annotation format is used for all seven categories of the named entity as shown in Table 5, and we use 47 types of part-of-speech (POS) as defined in the Orchid corpus.

Table 5. NE tags in each corpus file

Category	Format	Description	Example
Date	B-DAT	Beginning of a date	วันที่ (Date)
	I-DAT	Inside of a date	14 กุมภาพันธ์ (Feb, 14)
Location	B-LOC	Beginning of a location name	เมือง (City)
	I-LOC	Inside of a location name	นิวยอร์ก (New York)
Measurement	B-MEA	Beginning of a measurement unit	ห้า (Five)
	I-LOC	Inside of a measurement unit	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.	ลีก (League)
	I-NAM	Inside of any proper name	ลา ลีกา (La Liga)
Organization	B-ORG	Beginning of an organization name	บริษัท (Corp.)
	I-ORG	Inside of an organization name	โตโยต้า มอเตอร์ (Toyota Motor)
Person	B-PER	Beginning of a person name	นาย (Mister)
	I-PER	Inside of a person name	ณัฐวุฒิ (Natthawut)
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	

3.2 Corpus Challenges

As we mentioned above, some Thai language problems need to be solved as they cause some defects and limitations in the corpus. Thus, this paper aims to solve the difficulties of the existing corpus to improve the consistency and efficiency of the NE corpus in subsequent research. The defects of the corpus are as follows:

The Error of Word Segmentation. The major challenge of this corpus is the error of word segmentation. Since the Thai language has no clear word boundary or space between words, it is challenging to segment words. If the word segmentation is incorrect, it will affect the following process, especially in the process of named entity labeling. In Fig. 2, each picture represents an example of mistakes in word segmentation. The errors of segmenting words also occur to the abbreviation such as “มี.ค.”(Mar) this word means the abbreviation of March in the Thai language, but it was cut separately into different tokens as can be seen in Fig. 2(a). These wrong word segmentations also result in incorrect POS and named entity tagging.

<div> มี/VSTA/O ./PUNC/O ค/NLBL/O ./PUNC/O </div>	<div> นายก/NCMN/O <space>/PUNC/O อบ/VACT/O จ./NTTL/O อุตรดิตถ์/NPRP/O </div>
(a)	(b)

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

The Inconsistency of NE Tagging. The problem of inconsistency also happens with the named entity tag. Some words that can only belong to one category are labeled as the main category at some places, while they are labeled as “Other” in some other sentences in the same file. For example, “ประเทศ” is labeled as a location name (B-LOC) and it is labeled as other (O) in another place as shown in Fig. 3. Figure 4 also present an example of inconsistency NE tag labeling, “พรีเมียร์ลีก อังกฤษ”(Premier League) is annotated both name of football league (NAM) and other (O) in the same file.

<div> ราคา/NCMN/O ทองคำ/NCMN/O ใน/RPRE/O ประเทศไทย/NPRP/B-LOC ที่/PREL/O ปรับตัว/VACT/O สูงขึ้น/ADVN/O </div>	<div> นายก/NCMN/O สมาคม/NCMN/O ลูกจ้าง/NCMN/O ส่วน/NCMN/O ราชการ/NCMN/O แห่ง/NPRP/O ประเทศไทย/NPRP/O </div>
(a)	(b)

Fig. 3. Inconsistency of named entity tagging in Location corpus file

<div> พรีเมียร์ลีก/NCMN/O <space>/PUNC/O อังกฤษ/NCMN/O <space>/PUNC/O ฤดูกาล/NCMN/O <space>/PUNC/O 2008/NCMN/O </div>	<div> แชมป์/NCMN/O พรีเมียร์ลีก/NCMN/B-NAM <space>/PUNC/I-NAM อังกฤษ/NCMN/I-NAM <space>/PUNC/O ฤดูกาล/NCMN/O นี้/DDAC/O </div>
(a)	(b)

Fig. 4. Inconsistency of named entity tagging in Name corpus file

The Error of Named Entity Tag Assignment. The mistake in word segmentation will affect the POS of the word. Moreover, the incorrect word segmentation and POS annotation lead to wrong named entity tagging. Figure 5 shows incorrect name entity tags of the person’s surname due to incorrect word segmentation and POS assignment.

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

Fig. 5. False NE tagging of surname in Person corpus file

4 Corpus Revision Methodology

4.1 Approaches to Clean the Corpus

There are three main steps for cleaning and verifying the existing THAI-NEST corpus shown in Fig. 6.

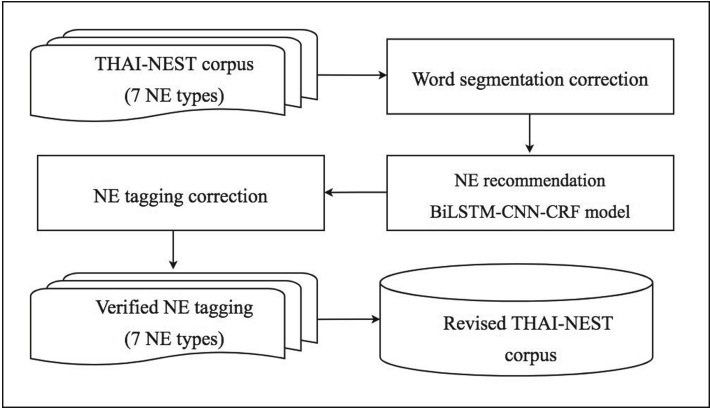


Fig. 6. The process for cleaning up the corpus

For the first step, we deal with the error of word segmentation in the corpus by searching incorrect named entities tags and manually correcting word segmenting and its POS. Next, NER models were trained for each of the main category files. Our proposed model is described in the following paragraph. Each of these models was trained separately using its own training set and validation set from each file and use these seven models to classify the entity of words. The most appropriate NE tag of each word is predicted and selected by the model. In the final step, we proceed with the cross annotation among the disjoint seven NE tagged files.

4.2 Named Entity Recognition Model

In this section, we describe the components of our model for improving and correcting the named entity tagging of the THAI-NEST corpus. The Thai NER model is presented, which was inspired by the research of [7]. Our proposed model consists of five important layers: Word Embedding, Character-level Representation, Part of Speech Embedding, Bi- LSTM layer, and the last is CRF layer. The architecture of the model is shown in Fig. 7.

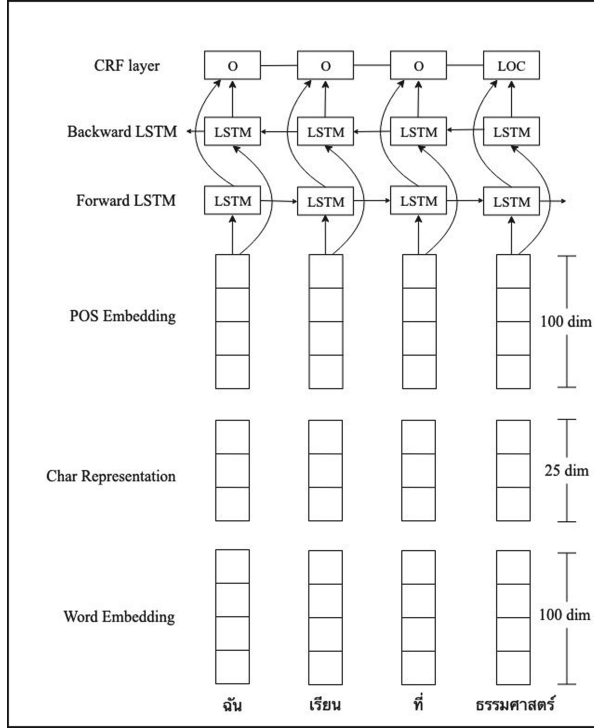


Fig. 7. The architecture of the proposed model

Word Embedding Layer. In our experiment, the Word2Vec tool of the Gensim library was used to pre-train word embedding by using the skip-gram model with 300 dimensions and window size of three words, three words before and three words after.

Thai Character Cluster (TCC)-Level Representation. Character-level representation can extract morphological information from the word and is extremely helpful in particular for languages with complicated structures or a rich morphological language, i.e., Hindi [8], Korean [9], and Thai. However, the Thai language has various characters such as vowels, consonants, tones, and special characters. In addition, a tone mark and a vowel sign cannot stand alone and they must be placed with the character only. Hence, if we use only word embedding may not significantly improve the

performance of our NER model. For this reason, we used the Thai Character Cluster (TCC) technique in character-level representation, which is an unambiguous unit that is smaller than a word and larger than a character and cannot be further divided based on Thai rules to group these characters [13, 14]. For example, “ป|ระ|เ|ท|ศ|ไ|ท|ย”(Thailand) or “น|า|ย|ก|ร|ั|ฐ|ม|น|า|ต|ร|ี”(The prime minister). In addition, CNN has good performance for NLP and character embedding. So, the CNN layer is applied to create the character vectors of the model. Detail of the CNN layer is as shown in Fig. 8.

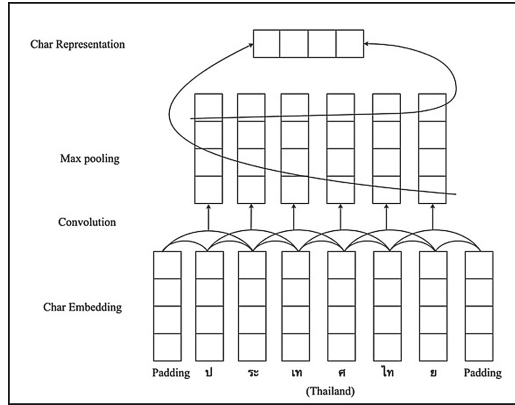


Fig. 8. Convolution Neural Network for character-level representation

POS Embedding: Part-of-speech (POS) can help to optimize NER model performance because most Thai named entities tend to be adjacent to or attached to part-of-speech such as verbs or prepositions. Several prior studies supported the use of POS in the NER model, such as Indonesian and Chinese language, so POS is introduced to be used with our model. POS of each word is encoded into the one-hot vector format in the embedding layer.

Bi-LSTM Layer. LSTM is capable of effectively capturing long-term dependencies and can retrieving rich global information. Furthermore, information from previous words is useful for prediction, but information from words coming after is also helpful. This can be done by having a second LSTM running backward, and this pair of forwarding and backward LSTMs is referred to as a bidirectional LSTM.

Conditional Random Field (CRF): The CRF is widely used model to predict the sequence of labels with the most likely tendency that corresponds to the sequence of given input sentences. The CRF model will take advantage of the neighbor tag information and consider the previous context in predicting current tags. We thus consider that the CRF is the last layer to predict the NE of each word and followed [14] to create our linear-chain CRF layer. The linear-chain CRF will find the highest scoring path through an input sequence and gives the best tags and final score.

Each layer is combined to construct the Bi-LSTM-CNN-CRF model for predicting named entity tags. Word, POS, and character-cluster vectors from each embedding

layer are concatenated before being fed into the Bi-LSTM layer. Then, the outputs of Bi-LSTM are transferred to the CRF layer and decoded by the Viterbi algorithm (part of the CRF layer) to determine the most possible entity tags. The model has been able to enhance the capacity to predict target words from the vector of the surrounding context.

5 Experiment

5.1 Pre-processing

As mentioned above in Sect. 3, the format of the named entity tag in this corpus is BIO format. However, the Thai writing system usually does not have a prefix or an indication of a name. It means that we cannot measure the score of B-tag and I-tag separately because some words in the corpus do not have a prefix. Therefore, the format of the NE tag needs to change from BIO to IO format instead to solve this problem.

5.2 Experiment Setup

Each NE tagged file is divided into three parts: 80% of all sentences for the training set, 10% for the validation set, and the last 10% for the testing set.

For our neural network model, a list of all parameters needs to be set. The various parameters are adjusted on the development set to get the most suitable final parameters. All parameters and settings are displayed in Table 6.

Table 6. Parameters for all experiments

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

6 Result and Discussion

In this experiment, the performances of models are compared based on F1-score. Seven models predict only their own testing set and are measured their F1-score separately. Table 7 shows the results of each corpus before and after the segmenting correction. Due to the prediction of the NER model, we can also manually resolve inconsistencies and invalid NE tags problems based on the results from the model. Each revised NE tagged file was retrained by using the proposed model. All F1-score after correcting the NE tags is higher than one before correcting up to an average of 12%, and the results are listed in Table 8.

Table 7. Comparison results between before and after correcting word segmentation

NE	F1-score	
	Before	After
DAT	85.04	89.14
LOC	69.27	73.68
MEA	77.58	80.45
NAM	42.76	46.91
ORG	70.19	75.03
PER	81.71	85.64
TIM	84.53	88.55

Table 8. The results after correcting NE tags in each corpus file

NE	F1-score
DAT	93.21
LOC	88.93
MEA	86.52
NAM	84.96
ORG	87.31
PER	88.90
TIM	94.76

As shown in Table 7, the f1-score of each category after solving word segmentation is higher than before correcting. The correction of word segmentation errors improves F1-score and affects the named entity prediction of the model. Once words are resolved, their entity is also correctly predicted. For example, words in the sentence “วันที่ 1 มี.ค.2551”(March 1, 2008), the NE tag is labeled as “Other” before correcting the error of segmentation but the model can correctly predict the named entity tags after correction as shown in Fig. 9. The fourth column is the predicted NE tag from the model.

ตั้งแต่/JSBR/O O	ตั้งแต่/JSBR/O O
วันที่/NCMN/O O	วันที่/NCMN/O DAT
1/DONM/O O	1/DONM/O DAT
<space>/PUNC/O O	<space>/PUNC/O DAT
มี/VSTA/O O	มี.ค./NPRP/O DAT
./PUNC/O O	<space>/PUNC/O DAT
ค/NLBL/O O	2551/NCNM/O DAT
./PUNC/O O	
<space>/PUNC/O O	
2551/NCNM/O O	

(a)

(b)

Fig. 9. The predicted named entity tag (a) before and (b) after editing word segmentation in Date corpus file

Furthermore, the use of POS can improve the NER model efficiency. For instance, the word “ใน”(in) is a preposition, and “ไป”(go) is a verb, once these words occur before a noun, the noun will be a named entity as shown in the following example. In Fig. 10(a), “ความน่าเชื่อถือของธนาคารในประเทศไทย”(The reliability of the banks in Thailand), the word “ใน”(in) is placed before the word “ประเทศไทย”(Thailand), therefore, “ประเทศไทย”(Thailand) is location (LOC) not person (PER). The sentence “นายกรัฐมนตรีจะเดินทางไปเมืองปุตราจายา”(The prime minister will go to Putrajaya), the word “ไป”(go) preceding the word “เมืองปุตราจายา”(Putrajaya) which is a city in Malaysia, “เมืองปุตราจายา” is a location as shown in Fig. 10(b).

ความ/FIXN/O O	นายกรัฐมนตรี/NCMN/O O
น่า/XVAM/O O	จะ/XVBM/O O
เชื่อถือ/VSTA/O O	เดินทาง/VACT/O O
ของ/RPRE/O O	ไปยัง/RPRE/O O
ธนาคาร/NCMN/O O	เมือง/NCMN/O LOC
ใน/RPRE/O O	ปุตราจายา/NPRP/O LOC
ไทย/NPRP/O LOC	

(a)

(b)

Fig. 10. Examples of NE tags that close to (a) preposition, and (b) verb

However, even POS is beneficial for the named entity recognition, but there are some errors in the prediction process. As can be seen in Fig. 11, “มีกิจกรรมพิเศษค้นหาอัจฉริยะไอที เพื่อแข่งขันทักษะไอทีของนักเรียนป ระถมและมัธยมทั่วประเทศ”(There is a special event called **Search for IT genius** to compete for IT skills of a nationwide elementary and high school student.), “ค้นหาอัจฉริยะไอที”(Search for IT genius) is an event name, but the word “ค้นหา”(search) in the event name is a verb, so “ค้นหาอัจฉริยะไอที”(Search for IT genius) seems like an activity not the name of the event. Thus, the model cannot predict the named entity correctly.

กิจกรรม/NCMN/O	O
พิเศษ/VATT/O	O
<space>/PUNC/O	O
"/PUNC/O	O
ค้นหา/VACT/B-NAM	O
อัจฉริยะ/NCMN/I-NAM	O
ไอที/NCMN/I-NAM	O
"/PUNC/O	O

Fig. 11. Incorrect NE prediction in Name corpus file

In the sentence as shown in Fig. 12, “ความรู้การพัฒนาซอฟต์แวร์มาตรฐานสากลCMMI จากมหาวิทยาลัยซอฟต์แวร์ในประเทศสหรัฐอเมริกา”(CMMI International Software Development Knowledge from software university in the United States), “มหาวิทยาลัยซอฟต์แวร์”(software university) mean a university that offers software discipline instruction. This word should be predicted as other but because the word “จาก”(from) is a verb and it makes the model predict inaccurately, so “มหาวิทยาลัยซอฟต์แวร์”(software university) is incorrectly predicted as a location.

มาตรฐานสากล/NCMN/O	O
<space>/PUNC/O	O
CMMI/NPRP/O	O
<space>/PUNC/O	O
/PUNC/O	O
Capability<space>Maturity<space>Model<space>Integration/NPRP/O	O
/PUNC/O	O
<space>/PUNC/O	O
จาก/RPRE/O	O
มหาวิทยาลัย/NCMN/O	LOC
ซอฟต์แวร์/NCMN/O	LOC
<space>/PUNC/O	O
ประเทศสหรัฐอเมริกา/NPRP/B-LOC	LOC

Fig. 12. Incorrect NE prediction in Location corpus file

We also performed experiments with other baseline models using the same revised NE tagged file to prove the effectiveness of consistency verifying and named entity selection of the proposed models and show the importance of using POS and TCC in dealing with wrong word segmentation and NE assignment. Table 9 presents the comparison result of the performance of each model.

According to the result shown in Table 9, our Bi-LSTM-CNN-CRF model outperforms other baseline models, especially a Date and Time category file in which the F1-score was around 94% and the Time category file in which the F1-score was about 93%.

Table 9. Performance of our model and other baseline models

NE	F1-score		
	BiLSTM (Word)	BiLSTM-CNN-CRF (Word+TCC)	BiLSTM-CNN-CRF (Word+POS+Char)
DAT	79.20	84.65	92.43
LOC	75.33	79.26	87.07
MEA	72.41	77.53	85.66
NAM	67.18	73.84	82.25
ORG	73.56	77.90	85.79
PER	74.29	81.72	87.32
TIM	82.77	86.05	93.88
NE	F1-score		
	BiLSTM (Word+POS)	BiLSTM-CNN (Word+POS+TCC)	BiLSTM-CNN-CRF (Word+POS+TCC)
DAT	86.14	89.72	93.21
LOC	83.67	86.24	88.93
MEA	80.22	82.67	86.52
NAM	75.48	80.13	84.92
ORG	81.17	84.75	87.31
PER	82.35	85.07	88.90
TIM	88.12	91.36	94.76

7 Combined Corpus

Another major issue of this corpus is that it is disjointedly generated into seven files, and each file is exclusively for each main category. Due to this structure, the corpus cannot use directly to create such a model for named entity recognition which can classify all categories. We thus conduct cross annotation among the seven NE tagged files and combine all named entity types into the same file. The brief idea of this approach is to use the seven trained models obtained from our proposed model training, applying cross annotation of every named entity category for labeling named entity tags in one category file until all seven original category files are complete. Table 10 shows an example of a combined corpus.

Table 10. Named entity tags in the combined corpus (a) in Thai original text, and (b) in English translated text

<pre> %Title: BKD19-1 (Thai NE Corpus) %Description: Based on THAINEST corpus %Number of sentence: 2,783 %Number of word: 272,753 %Date: March 17, 2019 %Creator: Kitiya Suriyachay and Virach Somlertlamvanich %Email: m6922040075@siit.tu.ac.th and virach@siit.tu.ac.th %Affiliation: Sirindhorn International Institute of Technology, Thammasat University #S1 นายสุเทพ เพื่ออุบลวรรณ รองนายกรัฐมนตรี กล่าวว่า ในวันพรุ่งนี้ (18 มี.ค.52) รัฐบาลไทย\ นายอภิสิทธิ์ เวชชาชีวะ นายกรัฐมนตรี จะมอบนโยบายและแนวทางการป้องกัน\ และปราบปรามยาเสพติดให้กับส่วนราชการต่าง ๆ เพื่อบูรณาการแผนปฏิบัติการป้องกัน\ และปราบปรามยาเสพติดร่วมกัน// นาย/NTTL/B-PER สุเทพ/NPRP/I-PER <space>/PUNC/I-PER เพื่ออุบลวรรณ/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 มี.ค. 52/NPRP/I-DAT)/PUNC/O . . ยาเสพติด/NCMN/O ร่วมกัน/ADVN/O // </pre>	<pre> %Title: BKD19-1 (Thai NE Corpus) %Description: Based on THAINEST corpus %Number of sentence: 2,783 %Number of word: 272,753 %Date: March 17, 2019 %Creator: Kitiya Suriyachay and Virach Somlertlamvanich %Email: m6922040075@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 Mar 09/NPRP/I-DAT)/PUNC/O . . plan/NCMN/O together/ADVN/O // </pre>
---	--

(a)

(b)

Finally, we train the proposed model on the Combined corpus. The F1-score of the Combined model on each main category is listed in Table 11. As being shown, all F1-scores of Combined corpus are quite similar to the F1-score of models trained by each file. Nevertheless, using the Combined-Corpus approach provides better results, but the F1-Scores in some categories drop slightly (e.g., Location and Organization). In addition, the number of named entities in each corpus file is dramatically increasing after combining all seven named entity categories.

Table 11. Results of the combined corpus

NE	F1-score
DAT	94.02
LOC	87.15
MEA	87.36
NAM	86.17
ORG	85.84
PER	89.27
TIM	96.44

8 Conclusion

This paper presented an approach by generating a NER model to clean the existing named entity corpus in the Thai language and verify its consistency. We use the THAI-NEST corpus to verify the consistency of NE tags and re-annotate the named entities by the proposed model. Our model can deal with the named entity tag inconsistency problem, including word segmentation mistakes, and yield impressive results. In order to enhance the amount of NE tags and prepare for further NE tag context captures in NER model development, we have performed a cross-annotation technique of all the seven NE tagged files. The Bi-LSTM-CNN-CRF model with the word, part-of-speech, and TCC features is used to verify the revised NE tagged corpus. Furthermore, POS and TCC play an important role in solving the problems related to word segmentation errors and inconsistency of NE tags. The model provides the performance of the verification, which increases the accuracy up to 12%.

Acknowledgements. The project is financial support provided by Thammasat University Research fund under the TSRI, Contract No. TUFF19/2564 and TUFF24/2565, for the project of “AI Ready City Networking in RUN”, based on the RUN Digital Cluster collaboration scheme.

References

1. Chopra, D., Joshi, N., Mathur, I.: Named entity recognition in Hindi using hidden Markov model. In: 2016 Second International Conference on Computational Intelligence & Communication Technology (CICIT), pp. 581–586 (2016)
2. Shijia, E., Xiang, Y.: Chinese named entity recognition with character-word mixed embedding. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pp. 2055–2058 (2017)
3. Ju, Z., Wang, J., Zhu, F.: Named entity recognition from biomedical text using SVM. In: 5th International Conference on Bioinformatics and Biomedical Engineering, pp. 1–4 (2011)
4. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 260–270 (2016)

5. Limsopatham, N., Collier, N.: Bidirectional LSTM for named entity recognition in Twitter messages. In: *Proceedings of the 2nd Workshop on Noisy User-generated Text*, pp. 145–152 (2016)
6. Liu, K., Hu, Q., Liu, J., Xing, C.: Named entity recognition in chinese electronic medical records based on CRF. In: *14th Web Information Systems and Applications Conference (WISA)*, pp. 105–110 (2017)
7. Ma, X., Hovy, E.: End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (vol. 1: Long Papers)*, pp. 1064–1074 (2016)
8. 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* **8**(4), 157 (2017)
9. Na, S., Kim, H., Min, J., Kim, K.: Improving LSTM CRFs using character-based compositions for Korean named entity recognition. *Comput. Speech Lang.* **54**, 106–121 (2019)
10. Nguyen, P., Le, A., Ho, T., Do, T.: Two entropy-based methods for detecting errors in POS-tagged treebank. In: *3th International Conference on Knowledge and Systems Engineering*, pp. 150–156 (2011)
11. Rachman, V., Savitri, S., Augustianti, F., Mahendra, R.: Named entity recognition on Indonesian Twitter posts using long short-term memory networks. In: *International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pp. 228–232 (2017)
12. Salleh, M.S., Asmai, S.A., Basiron, H., Ahmad, S.: A Malay named entity recognition using conditional random fields. In: *5th International Conference on Information and Communication Technology (ICoICT)*, pp.1–6 (2017)
13. Sornlertlamvanich, V., Tanaka, H.: The automatic extraction of open compounds from text corpora. In: *Proceedings of the 16th Conference on Computational Linguistics (COLING-1996)*, pp. 1143–1146 (1996)
14. Sornlertlamvanich, V., Tanaka, H.: Extracting open compounds from text corpora. In: *Proceedings of the 2nd Annual Meetings of the Association for Natural Language Processing*, pp. 213–216 (1996)
15. Sornlertlamvanich, V., Takahashi, N., Isahara, H.: Thai part-of-speech tagged corpus: ORCHID. In: *Proceedings of Oriental COCOSDA Workshop*, pp. 131–138 (1998)
16. Suriyachay, K., Sornlertlamvanich, V.: Named entity recognition modeling for the Thai language from a disjointedly labeled corpus. In: *5th International Conference on Advanced Informatics: Concept Theory and Applications (ICAICTA)*, pp. 30–35 (2018)
17. Suriyachay, K., Sornlertlamvanich, V., Charoenporn, T.: Thai named entity tagged corpus annotation scheme and self verification. In: *Proceedings of the 9th Language & Technology Conference (LTC2019)*, pp.131–137 (2019)
18. Theeramunkong, T., et al.: THAI-NEST: a framework for Thai named entity tagging specification and tools. In: *Proceedings of the 2nd International Conference on Corpus Linguistics (CILC10)*, pp. 895–908 (2010)
19. Wang, Y., Xia, B., Liu, Z., Li, Y., Li, T.: Named entity recognition for Chinese telecommunications field based on Char2Vec and Bi-LSTMs. In: *12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*, pp. 1–7 (2017)
20. Yao, T., Ding, W., Erbach, G.: Repairing errors for Chinese word segmentation and part-of-speech tagging. In: *Proceedings of the International Conference on Machine Learning and Cybernetics*, pp. 1881–1886 (2002)