Thai Named Entity Tagged Corpus Annotation Scheme and Self Verification

Kitiya Suriyachay[*], Thatsanee Charoenporn[†] and Virach Sornlertlamvanich[*†]

[*]School of ICT, Sirindhorn International Institute of Technology,
Thammasat University, Thailand
[†]Faculty of Data Science, Musashino University, Japan
m5922040075@g.siit.tu.ac.th and {thatsanee, virach}@ds.musashino-u.ac.jp

Abstract

In this paper, we propose a method to clean up the existing named entity (NE) corpus and verify its consistency in creating a model for the named entity recognition (NER) task, especially for the low resource language such as Thai. To create a task specific Thai language corpus, we heavily rely on many fundamental language tools in morphological analysis pre-processing, which can propagate the errors affecting the preciseness of the acquire model. We adopted a collection for NE corpus prepared by THAI-NEST, verified its annotation consistency and iteratively re-annotated it with the created model. We found that most of the errors are caused by human errors in word segmentation and NE tag interpretation. Each cycle in applying our new model, the higher accuracy can be measured and resulted in a list of errors for correction. Moreover, the original THAI-NEST provides only one NE tag for a file. Therefore, only one model can be created at a time. We extensively conducted the cross annotation among the seven NE tagged files to increase the number of NE tags and to prepare for additional NE tag context capturing in NER model development. The revised NE tagged corpus is finally verified with the best BiLSTM-CNN-CRF model with word, part-of-speech and character embedding approach. The results of the experiment show the effectiveness of the self verification which increases the accuracy up to 12%.

Keywords: Corpus Annotation, Named Entity Recognition, Recurrent Neural Network, Thai Named Entity, Thai language

1. Introduction

Making use of this huge amount of data methodically is a key of organizations to a success and they have strong ambition to use that information as well. Many services related to information extraction and analytics are provided dramatically and broadly. Research on the topic of information extraction is therefore continuously undertaken in various approaches.

Information extraction (IE) refers to automated extraction of specific information from unstructured natural language data into a structured representation. The performance of IE depends on many NLP preprocessing subtasks including word segmentation, POS tagging, and especially, named entity recognition (NER). NER is a key success of information extraction, since information extracted from the body of the text is information related to entity categories. NER task is to identify and classify the particular proper nouns in focused texts automatically.

Continuously, there have been researches on NER for many languages with various approaches. But NER for Thai language were still limited. There are several challenges in Thai NER. Firstly, unlike English or other European languages, there is no word boundary in Thai language. Thai words are implicitly recognized and some depend on the individual judgement. Incorrect word identification certainly affects other upper recognition than word level. As well as in NER, incorrect word segmentation will lead to false named entity recognition. Secondly, there is no capitalization in writing system to identify named entities. Even though, there are some markers in some cases identifying proper nouns like person name or institution name. For example, in general writing system of name and surname with/without title, first name will be preceded by title without any space as "พลเอกประยุทธ์" ("พลเอก" General [Title] "ประยุทธ์" Prayut [Name]), while there is at least one space between first name and surname as "พลเอกประยุทธ์ จันทร์โอชา" ("จันทร์โอชา" Chanocha [Surname]). In general, a proper name usually occurs with a common noun which indicates the type of the proper name (common noun + proper name) but, in a discourse, this pattern can be changed in various ways, for example, "ธรรมศาสตร์ร่วมกับเอกชนสร้างสรรค์ไอเดีย ... มหาวิทยาลัยธรรมศาสตร์เปิดเวทีดึงศิษย์เก่า ม.ธ. นำไอเดียใหม่สดไปต่อยอด" (Thammasat together with the private sector to create ideas ... Thammasat University opens the stage to draw alumni of M.T. bringing top ideas to business.). From the sample, Thammasat University occurs in three different ways, these are "ธรรมศาสตร์" (Thammasat) "มหาวิทยาลัยธรรมศาสตร์" (Thammasat University) and "ม.ธ." (the abbreviation of Thammasat University). Ambiguity of homographs is also a big challenge. Name entities are able to construct like common nouns which occasionally causes an ambiguous analysis. For example, "พันตำรวจเอกทวี สอดส่อง อธิบดีกรมสอบสวนคดีพิเศษ" (Colonel Tawee Sordsong Director-General of the Department of Special Investigation), "สอดส่อง" occurs here as his surname which also conveys a meaning of an activity of monitoring in general. The problems on distinguishing between named entity and other types of noun then frequently occur.

Moreover, once words are segmented and marked with named entity tags, consistency of NE tags throughout the corpus is also the important considerable issue. Since inconsistency is going to cause the failure in further processes. This paper aims to propose a method to clean up the existing named entity (NE) corpus and verify its consistency in creating a model for the Thai named entity recognition (NER) task. The experiment has been done with THAI-NEST NE corpus. We also conduct the cross

annotation among the disjoint seven NE tagged files to increase the number of NE tags and to prepare for additional NE tag context capturing in NER model development. The revised NE tagged corpus is finally verified with the best BiLSTM-CNN-CRF model.

The remainder of paper is related work in Section 2. Section 3 explains about how to construct and verify the NE corpus. In Section 4, we explain the methodology to improve the correctness and consistency of the corpus. The results, discussion and conclusion will be described in Section 5 and Section 6 respectively.

2. Related Work

The named entity recognition is one of the most broadly researched topics for information extraction. NE corpus and tools have been constructed and openly provided for researching over the past years, especially for the most spoken languages like English and Chinese. MUC 7, for example, is an English dataset provided by the Message Understanding Conference which was the first one to broadly introduce NER task. MUC 7 contains many types of name entities including person, organization, location, dates, times, percentages, and monetary amount (Chinchor, 1998). CoNLL-2003 shared task data (Tjong *et al.*, 2003) is an English and German named entity corpus with four types of name entities: person, location, organization and name of other miscellaneous entity (MISC). Dataset of ACE is also one another corpus provided by the Linguistic Data Consortium with the aim to develop extraction technology to support automatic processing for entities, relations and events of source language data. ACE 6 consists of five main types of entities: Person, Organization, Location, Facility, and Geographical/ Social/Political. The ACE corpus also contains HEAD and EXTENT annotations of entities as well (Augenstein et al., 2017; LDC, 2008).

However, it is quite limited for Thai NE corpus. THAI-NEST (THAI-Named Entities Specification and Tools, Theeramunkong *et al.*, 2010) is only the open general Thai corpus with named entity tags. Over 300,000 Thai online news articles on seven major categories from twenty-one publishers are word-segmented and tagged with seven named entity categories including person name, organization name, place name, date, time, measurement, and name. We conduct our experiment on the THAI-NEST corpus, verify and improve its tag consistency for a proper release.

There are various approaches proposed in order to improve NER. Some classical machine learning approaches are applied over the past years. For example, Hidden Markov Model (HMM) (Chopra *et al.*, 2016), Support Vector Machine (Ju *et al.*, 2011), Conditional Random Fields (CRF) (Tirasaroj and Aroonmanakun, 2009), pattern-based approach (Tongtep and Theeramung kong, 2008) and Singular Value Decomposition (Suwanapong and Theeramungkong, 2009).

Some techniques of deep learning have been adapted to NER recently, such as Recurrent Neural Network model (RNN), LSTM and its extension, BiLSTM with character embedding for better performance (Wang *et al.*, 2017; Rachman *et al.* (2017); (Suriyachay and Sornlertlam vanich, 2018)

3. Named entity Corpus Construction

NE tagged corpus presented in this paper is designed and constructed based on the annotation scheme proposed in ORCHID corpus construction, which is the first open online Thai POS Tagged corpus (Sornlertlamvanich *et al.*, 1997).

3.1 Structure of the NE Corpus

The original corpus is marked up with two types of markers in order to give some additional information about text information line and numbering line. The text information line which is a line beginning with a "%" is marked to give some information in addition to the text. The mark-ups for text information lines and their description are shown in Table 1. The string of characters attached with a colon (:) behind the "%" is interpreted as a comment line for providing the additional information only. Neither types of markers nor the string behind is referred as any parts of the original texts. Text information in the line is given in English language. The numbering line which is a line beginning with a "#" is used to index the sequence of line in the text. There are two numbering lines, those are "#P" and "#S" as shown in Table 2.

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 1: Mark-up for text information 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 2: Mark-up for numbering line

As for the characteristics of Thai language, there is no explicit word breaking character applied in the common real text, a paragraph or a sentence is generally wrapped at either a space character or at any breakable syllable construction. Three special mark-ups characters, therefore, are introduced to identify line break, sentence break and NE Tag marker, as shown in Table 3. In addition, BIO format is introduced to identify the chunk or component of NE. As "B" indicates the beginning of the chunk, "I" is within the chunk and "O" denotes that the word does not involve in any types of NE.

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

Table 3: Special characters for Mark-up

Furthermore, all special characters other than alphanumeric characters are replaced by the internal defined strings enclosed by a pair of "<" and ">" bracket. The structure of the corpus is illustrated in Fig. 1.

3.2 Challenges in NE corpus construction

There is limitation of an open Thai NE tagged corpus as previously mentioned. Moreover, there are some Thai language related problems that need to be solved to improve the NE corpus. One objective of this paper is to overcome the challenges that still occur in the existing corpus, THAI-NEST, for making the NE corpus more consistent and efficient in further research. The challenges occurred during the process of NE tagging includes the correctness of word segmentation, the correctness of NE tag assigning, and the consistency of NE tag assigned along the corpus.

3.2.1 The correctness of word segmentation

The difficulties of Thai Natural Language Processing begin with word level recognition as the Thai language has no boundary indicator among words such as space or white space in English. If the word segmentation is not correctly done, it can affect to all next processes, especially in the process of NE recognition. Fig. 2 shows the samples of false word segmentation. As a result, the NE words are recognized as some other types of noun.

3.2.2 The correctness of NE tag assignment

As a result of the mistakes in word segmentation and POS assignment, some NE tags are assigned inaccurately in Fig. 3.

3.3.3 The consistency of NE tagging

One another challenge in the NE tagged corpus construction is the consistency of the tagging throughout the corpus. From the original corpus, we found that there are some inconsistencies of tag labelling between NE and Proper Noun (NPRP) or other categories. For example, “ประเทศไทย” (Thailand) is tagged as NE-LOC and NPRP, “บาท” (Baht) as NE-MEA and Classifier, as shown in Fig. 4 and Fig. 5. In addition, Fig. 6 and Fig. 7 present an example of incorrect NE tag labelling between Proper Noun, NE-NAM and NE-ORG, respectively. For handling these challenges, we, therefore, propose BiLSTM-CNN-CRF model to improve the consistency of NE tagging in the corpus.

```
%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 จะมอบนโยบายและแนวทางในการป้องกันและปราบปรามยา\\
เสพติดให้กับส่วนราชการต่างๆ เพื่อบูรณาการแผนปฏิบัติการป้องกันและปราบปรามยาเสพติดร่วมกัน//

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
//
```

(a)

```
%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/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
//
```

(b)

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

```
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 word segmentation in different corpus file (a) Date corpus file and (b) Name corpus file

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

Fig. 3: False NE tagging of surname in person file

Moreover, the original THAI-NEST provides only one NE tag for a file. Therefore, only one model can be created at a time. We extensively conducted the cross annotation among the seven NE tagged files to increase the number of NE tags and to prepare for additional NE tag context capturing in NER model development.

```
  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 named entity tagging in location file

```
  ทอง/NCMN/O
  หนัก/VATT/O
  <space>/PUNC/O
* 3/DCNM/O
* <space>/PUNC/O
* บาท/CMTR/O
  และ/JCRG/O
  <space>/PUNC/O
* 5/DCNM/B-MEA
* <space>/PUNC/I-MEA
* บาท/CMTR/I-MEA
```

Fig. 5: Inconsistency of named entity tagging in measurement file

```
* พรีเมียร์ลีก/NCMN/O
* <space>/PUNC/O
* อังกฤษ/NCMN/O
  <space>/PUNC/O
  ฤดูกาล/NCMN/O
  <space>/PUNC/O
  2008/NCNM/O
```

```
  แชมป์/NCMN/O
* พรีเมียร์ลีก/NCMN/B-NAM
* <space>/PUNC/I-NAM
* อังกฤษ/NCMN/I-NAM
  <space>/PUNC/O
  ฤดูกาล/NCMN/O
  นี้/DDAC/O
```

Fig. 6: Inconsistency of named entity tag in name file

```
  หัวหน้าส่วน/NCMN/O
  ราชการ/NCMN/O
  <space>/PUNC/O
  ที่/RPRE/O
  <space>/PUNC/O
* กระทรวงพาณิชย์/NPRP/B-ORG
```

```
  ที่ประชุม/NCMN/O
  จึง/XVBM/O
  มอบหมาย/VACT/O
* กระทรวงพาณิชย์/NPRP/O
  จัดทำ/VACT/O
  แผน/NCMN/O
  ปฏิบัติการ/VACT/O
```

Fig. 7: Inconsistency of named entity tag in organization file

4. Corpus Annotation Revision

In order to clean up the existing named entity (NE) corpus and verify its consistency, we followed the steps shown in Fig. 8. There were main three steps in cleaning up process, starting from word segmentation correction, named entity recommendation and named entity tagging correction.

Regarding to the word segmentation correction, we searched for the errors of named entity tags, then manually corrected word segmenting and POS tags of the words as well as their neighboring words. Next, we trained our proposed models on all seven files of THAI-NEST. As a result, all NE tag candidates were returned for the process of selecting an only appropriate one of each word. The across annotation among the seven NE tagged files were then conducted as the final step. The proposed NER model architecture is summarily illustrated in Fig. 9.

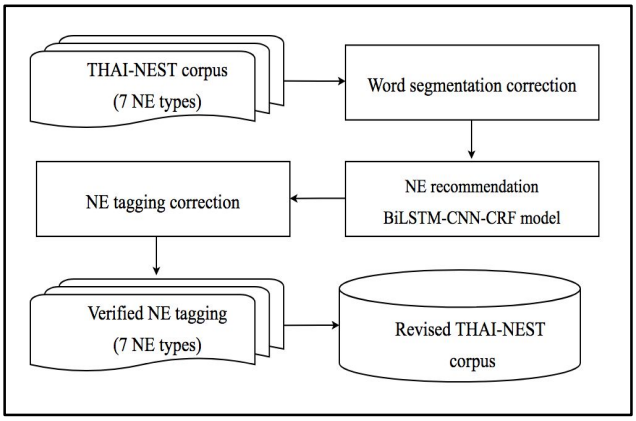

Fig. 8: Corpus Cleaning up Process

The architecture in Fig. 9 shows that firstly, words and POS are embedded to obtain their vector representations. For character embedding, we use Thai Character Cluster (TCC) (Sornlertlamvanich, V. and Tanaka H., 1996a and 1996b) of words and composing the resulting vectors with a max-pooled CNN. The words, POS, and character vectors are concatenated before feeding into BiLSTM layer. Then, the output vectors from BiLSTM are passed through a CRF layer. Dropout layer will be used in both input and output vectors of BiLSTM. Lastly, the CRF layer will predict named entity tag with the highest possible tendency that followed MA and Hovy (2016).

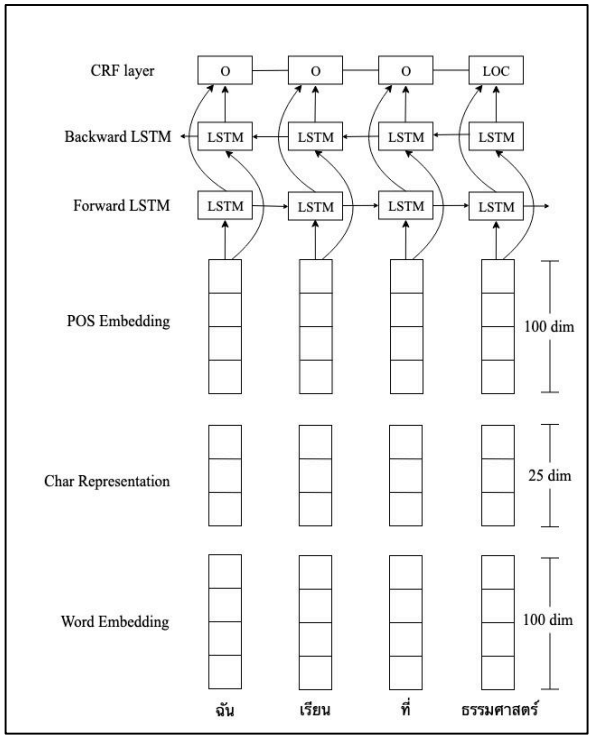

Fig. 9: Architecture of our NER model

5. Result and Discussion

In the experiment, we calculate the F1-score to evaluate the performance of the model. The result of each corpus before and after correcting error from word segmentation displays in Table 4.

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 4: The results before and after correcting word segmentation

All F1-score after correcting is higher than the one before. This score implied that correcting the errors of word segmentation in each corpus file increased F1-score and has an effect on the prediction of named entity of the model. The words in each corpus are fixed and can predict the named entity correctly as shown in Fig. 10, the fourth column is the predicted named entity tag from the model.

In addition, part-of-speech is able to enhances the efficiency of the model. For example, "ใน" (in) is a preposition when preceding a noun, the noun will be an NE, as shown in these following samples. The sentence "ความน่าเชื่อถือของธนาคารในไทย" (The reliability of the banks in Thailand), "ไทย" is LOC not PER, or "นายกรัฐมนตรีจะเดินทางไปปุตราจายา" (The prime minister will go to Putrajaya), that "ไป" (go) is a verb indicating a LOC of the following noun "ปุตราจายา" (Putrajaya) which is one of city in Malaysia.

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

(a)

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

(b)

Fig. 10: The predicted named entity tag (a) before and (b) after editing word segmentation in date corpus file

Nevertheless, the prediction of the named entity still found some errors, occurring by the preposition. For example, in the sentence as shown in Fig. 11, "ความรู้การพัฒนาซอฟต์แวร์มาตรฐานสากล CMMI จากมหาวิทยาลัยซอฟต์แวร์ในประเทศสหรัฐอเมริกา" (CMMI International Software Development Knowledge from the software university in the United States), here "มหาวิทยาลัยซอฟต์แวร์" (software university) refers to a university that offers software discipline instruction, which should be predicted as Other, But the word "จาก" (from) makes the model predicted as LOC, which is incorrect.

```
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. 11: Incorrect prediction of location corpus file

Due to the prediction of the model, we can solve some inconsistency and false named entity tag problems by human. The result is listed on the Table 5.

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

Table 5. The result after solving named entity tag

As being shown, all F1-score in Table 5 is higher than those in Table 4 as expected. The F-1 score after correcting named entity in each corpus file is dramatically increased as an average 12 percent compared to the result after editing word segmentation.

However, one of the major problems of this corpus is that the corpus is disjointedly managed in seven files according to the type of named entity. Therefore, one of another task in this research is to combine every named entity tag into the same file. We use the trained model derived from the training BiLSTM-CNN-CRF model of each named entity type to train and label named entity tag on one corpus file by using cross tagging method until completing all seven named entity types. Example of the corpus combining every named entity tag is shown in Fig. 12. Lastly, we trained the model with a corpus including all types of named entity. The result of the experiment is shown in Table 6.

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

Table 6. Result of combined corpus

Comparing the results with training models on each file, Table 5 presents that results of the model of combined corpus are similar to the model of the disjoint corpus tagged by human. This may imply that the combined corpus model has no effect on each named entity tagging.

```
%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 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 จะมอบนโยบายและแนวทางในการป้องกันและปราบปรามยา\\
เสพติดให้กับส่วนราชการต่างๆ เพื่อบูรณาการแผนปฏิบัติการป้องกันและปราบปรามยาเสพติดร่วมกัน//

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
//
```

(a)

```
%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 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
Mar 09/NPRP/I-DAT
)/PUNC/O
.
.
plan/NCMN/O
together/ADVN/O
//
```

(b)

Fig. 12: Named entity tags in the combined corpus (a) in Thai original text, and (b) in English translated text

6. Conclusion

This paper proposes a method to clean up the existing named entity (NE) corpus and verify its consistency in creating a model for the named entity recognition (NER) task, especially for the low resource language such as Thai. We adopted a collection for NE corpus prepared by THAI-NEST, verified the annotation consistency and iteratively re-annotated it with the created model. We extensively conducted the cross annotation among the seven NE tagged files of THAI-NEST to increase the number of NE tags and to prepare for additional NE tag context capturing in NER model development. The revised NE tagged corpus is verified with the best BiLSTM-CNN-CRF model with word, part-of-speech and character embedding approach. The results of the experiment show the effectiveness of the self verification which increases the accuracy up to 12%.

References

Augenstein, I., Derczynski, L., and Bontcheva, K. (2017). Generalisation in named entity recognition: A quantitative analysis. *Computer Speech & Language, 44*, pp. 61-83.

Chinchor, N. (1998). MUC-7 named entity task definition. *Proceedings of the Seventh Message Understanding Conference, Virginia, April 29 - May 1, 1998*.

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

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

Linguistic Data Consortium, ACE (Automatic Content Extraction) English Annotation Guidelines for Entities Version 6.6 2008.06.13

Rachman, V., Savitri, S., Augustianti, F., & Mahendra, R. (2017). Named entity recognition on Indonesian Twitter posts using long short-term memory networks. *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*.

Sang, E. F., and Meulder, F. D. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, 4*, pp. 142-147.

Sornlertlamvanich, V., Charoenporn, T., and Isahara, H. (1997). ORCHID: Thai Part-Of-Speech Tagged Corpus. Tech. Rep. TR-NECTEC-1997-001, National Electronics and Computer Technology Center, Thailand, pp. 5-19.

Sornlertlamvanich, V., and Tanaka H. (1996)a. The Automatic Extraction of Open Compounds from Text Corpora. *The 16th International Conference on Computational Linguistics (COLING-96)*, pp. 1143-1146.

Sornlertlamvanich, V., and Tanaka H. (1996)b. Extracting Open Compounds from Text Corpora. *The Second Annual Meetings of the Association for Natural Language Processing*, pp 213-216.

Suriyachay, K., and Sornlertlamvanich, V. (2018). 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)*.

Suwannapong, T., and Theeramunkong, T. (2009). Aliases Discovered in Thai Sports News Articles. *Proceedings of the 8th International Symposium on Natural Language Processing*, pp. 63-66.

Theeramunkong, T., Boriboon, M., Haruechaiyasak, C., Kittiphattanabawon, N., Kosawat, K., Onsuwan, C., Siriwat, I., Suwanapong, T., and Tongtep, N. (2010). THAI-NEST: A framework for Thai named entity tagging specification and tools. *Proceedings of the 2nd International Conference on Corpus Linguistics (CILC10), May 13-15, 2010, University of A Coruña, Spain.*

Tirasaroj, N., and Aroonmanakun, W. (2009). Thai named entity recognition based on conditional random fields. *2009 Eighth International Symposium on Natural Language Processing*.

Tongtep, N., and Theeramunkong, T. (2008). Pattern-based Extraction of Named Entities in Thai News Documents. *Proceedings of 3rd International Conference on Knowledge, Information and Creativity Support Systems (KICSS'08)*, pp. 82-89.

Wang, Y., Xia, B., Liu, Z., Li, Y., and Li, T. (2017). Named entity recognition for Chinese telecommunications field based on Char2Vec and BiLSTMs. *2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*.

X, Ma., and E, Hovy. (2016). "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)*.