Improvement of Thai NER and the Corpus

Thatsanee Charoenporn[1], Virach Sornlertlamvanich[1,2], Kitiya Suriyachay[2]
[1]AAII, Faculty of Data Science, Musashino University
3-3-3 Ariake, Koto-ku, Tokyo 135-8181, Japan
[2]SIIT, Thammasat University
131 Moo5, Tiwanond Road, Bangkadi, Mueang, Pathumthani, 12000, Thailand
{thatsane, virach}@musashino-u.ac.jp, m5922040075@g.siit.tu.ac.th

Abstract

Thai named entity (NE) corpus is rarely found though the named entity recognition (NER) task can make a big contribution in processing the huge amount of available texts. We propose an iterative NER refinement method using BiLSTM-CNN-CRF model with word, part-of-speech, and character cluster embedding to clean up the existing NE tagged corpus due to its inconsistent and disjointed annotation. As a result, in the newly generated corpus, we obtain 639,335 NE tags, much larger than the original size of 172,232 NE tags. The generated model by the newly generated corpus also improves the NER F1-score 16.21% to mark 89.22%.

Keywords: name entity, Thai language, corpus, NE corpus

Résumé

การพัฒนาคลังข้อความภาษาไทยสำหรับการประมวลผลภาษาธรรมชาตินั้น มีประเภทและปริมาณเพิ่มมากขึ้น แต่คลังข้อความชื่อเฉพาะภาษา ไทย หรือ Thai Name Entity Corpus ยังคงมีจำนวนจำกัด แม้ว่างานวิจัยด้านการรู้จำชื่อเฉพาะ (Name Entity Recognition: NER) จะส่งผลต่อความถูกต้องของการประมวลผลข้อความเป็นอย่างมากก็ตาม งานวิจัยนี้ เสนอวิธีการปรับแต่ง NER แบบวนซ้ำ โดยใช้แบบจำลอง BiLSTM-CNN-CRF ประกอบกับ คำแวดล้อม หน้าที่ของคำ และกลุ่มอักขระข้างเคียง เพื่อปรับปรุงคลังข้อความชื่อเฉพาะภาษาไทย จากเดิม จำนวน 172,232 ชื่อ ให้มีความถูกต้อง แม่นยำ และสอดคล้องกัน ผลการวิจัยพบว่า คลังข้อความชื่อเฉพาะภาษาไทย ที่ปรับปรุงขึ้น ประกอบด้วยคำและป้ายระบุชื่อเฉพาะ (Tags) จำนวนถึง 639,335 ชื่อ ทั้งนี้ ผลการปรับปรุงคลังข้อความชื่อเฉพาะด้วยแบบจำลองที่นำเสนอนี้ สามารถกำกับชื่อเฉพาะภาษาไทยได้ถูกต้อง วัดด้วยค่า F1-score ได้ที่ 89.22 เปอร์เซ็นต์ ซึ่งให้ผลที่ดีกว่าแบบ จำลองที่สร้างด้วยคลังข้อความเดิมถึง 16.21 เปอร์เซ็นต์

1. Introduction

The performance of IE depends on many NLP preprocessing subtasks including word segmentation, POS tagging, and especially, named entity recognition (NER). NER task is to identify and classify the particular proper nouns in focus 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.

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 proposes 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. As a result, the BKD (Bangkok Data) NE corpus is newly released as an NE silver standard corpus.

2. Thai Computational Corpus

Thai is an isolate and analytic language which is spoken and written by 65 million of population inside and outside Thailand. Thai language has 44 consonant characters and 18 vowel characters with 5 tones (4 tonal symbols). Word order in sentences is the essential way to illustrate syntactic role and convey the meaning. In the writing system, Thai has no explicit word and sentence boundary as occurred in English nor the capital letter used for beginning the sentence or name entity.

Although the Thai language has a limited number of users, there is a continuous development of corpus and tools for analysis and research in the field of Thai Linguistics and Natural Language Processing. But there is still a limited number of releases for public usage. Starting from ORCHID in 1997, the first Thai POS tagged corpus which is the collaborative research between the Communications Research Laboratory and the Electrotechnical Laboratory of Japan and NECTEC of Thailand (Thatsanee et al., 1997). Now, Thai-English Parallel Corpus, Thai Speech Corpus, Thai Character Image Corpus are also publicized by NECTEC under Creative Commons Attribute 3.0 License. Thai National Corpus, initiated by Chulalongkorn University in 2006, provides the collection of current written Thai text marked up with TEI P4 standard as well as word boundaries and romanized transcription. At present, some Thai Natural Language Processing Resources including corpus, lexicons and software libraries, have been collected by Thai NLP group and can be accessed online (Kobkrit, 2019). However, the size of the corpus provided is still limited.

3. Challenges in Thai Name Entity Construction

Conforming to the Thai Name Entity (NE) Corpus, it is still limited either in number or size. THAI-NEST (THAI-Named Entities Specification and Tools) is only the open general Thai corpus with word segmentation and name entity tags. The corpus consists of over 300,000 Thai online news articles on seven major categories from twenty-one publishers which were word-segmented and tagged with seven name entity categories including person name, organization name, place name, date, time, measurement, and name (Theeramunkong et al, 2010). Figure 1 depicts NE corpus provided by THAI-NEST.

Figure 1 : THAI-NEST Corpus

From the corpus, some attempts are undertaken for NE tagged annotation. Nevertheless, some challenges can be found from the original tagged NE corpus version 0.1 (N/A). 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. Incorrectness of word segmentation and POS assignment sometimes have been found on the abbreviation such as "มี.ค." [March] or "อบจ." [Provincial Administration Organization]. Once they occur, as a result, some NE tags are assigned inaccurately too. Figure 2(a) and 2(c) illustrate examples of incorrectness of NE tag assignment. สอดส่อง/VACT/O in Figure 2(a), and แมนเชสเตอร์/NPRP/O and ยูไนเต็ด/NPRP/O in Figure 2(c) should be tagged as NE-PER and NE-NAM respectively instead, as shown in Figure 2(b) and (d).

พ.ต.อ./NCMN/B-PER
<space>/PUNC/I-PER
ทวี/NPRP/I-PER
<space>/PUNC/O
สอดส่อง/VACT/O

(a)

พ.ต.อ./NCMN/B-PER
<space>/PUNC/I-PER
ทวี/NPRP/I-PER
<space>/PUNC/O
สอดส่อง/NPRP/I-PER

(b)

แมนเชสเตอร์/NPRP/O
<space>/PUNC/O
ยูไนเต็ด/NPRP/O
<space>/PUNC/O
แชมป์ลีก/NCMN/B-NAM
ฤดูกาล/NCMN/O
ล่าสุด/VATT/O

(c)

แมนเชสเตอร์/NPRP/B-NAM
<space>/PUNC/O
ยูไนเต็ด/NPRP/I-NAM
<space>/PUNC/O
แชมป์ลีก/NCMN/O
ฤดูกาล/NCMN/O
ล่าสุด/VATT/O

(d)

Figure 2 : Incorrect and correct tag assignment.

บังคับ/VACT/O
ใช้/VACT/O
แล้ว/XVAE/O
ตั้งแต่/RPRE/O
เที่ยง/NCMN/O
คืน/VACT/O

(a)

จนถึง/RPRE/O
เที่ยง/NCMN/B-TIM
คืน/VACT/I-TIM
<space>/PUNC/O

(b)

Figure 3 : Inconsistency of Common Noun and NE-TIM tag assignment.

The consistency of the tagging throughout the corpus is one another challenge. From the original corpus, we found that there are some inconsistencies of tag assignment between NE and other categories, as shown in Figure 3 (a) and (b), "เที่ยง" is tagged as Common Noun (NCMN) and NE-TIM. In addition, Figure 4 (a) and (b) illustrate the examples of inconsistent and incorrect NE tag assignment between Number (DONM), Proper Name (NPRP) and NE-DAT.

วันที่/NCMN/O
23/DONM/O
<space>/PUNC/O
ถึง/RPRE/O
<space>/PUNC/O
31/NCNM/O
<space>/PUNC/O
พฤษภาคม/NPRP/O

(a)

วันที่/NCMN/B-DAT
1/DONM/I-DAT
<space>/PUNC/I-DAT
พฤษภาคม/NPRP/I-DAT

(b)

Figure 4 : Inconsistency of Number, Proper Noun and NE-DAT tag assignment.

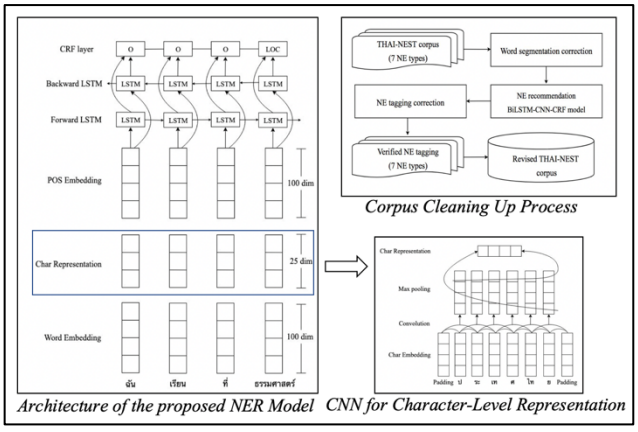

Figure 5 : Thai Name Entity Tagging Cleaning up Process

4. Thai Name Entity Tagging

We combined to create the BiLSTM-CNN-CRF model (Ma and Hovy, 2016) for predicting named entity tags. The character-level representation of each word is calculated by the CNN, as shown in Figure 5. From each embedding step, we obtain the vector representations of the words, POS tags, and character clusters. 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 to select the most possible sequence of the NE tag.

5. Bangkok Data Name Entity Corpus

In this part, the proposed schema and the characteristics of Bangkok Data Name Entity Corpus 2019 are described.

5.1 Corpus Mark-up Schema

There are 2 types of markers proposed to the original NE corpus, these are 1) file information markers, and 2) line number and special markers. Table 1 below displays the mark-up schema of the file information.

Mark-up	Description
%Title:	Title of the document/file
%Description:	Detail/Information of the original Text
%Number of sentence:	Total amount of sentences in the file
%Number of word:	Total amount of segmented words in the file
%Number of named entity tag:	Total of NE tags found in the file
%Date:	Date of running the NE tagger
%Creator:	Name of person who create or run the tagger
%Email:	Email address of person who create or run the tagger
%Affiliation:	Organization of the creator

Table 1: The Mark-up Schema of File Information

5.2 Number of Line and Other Special Mark-up Schema

There are 2 types of number of line marker, number of paragraph and number of sentences, as shown in Table 2 And Table 3 shows the special markers used in the corpus to convey the additional information and tags.

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

Table 2: The Mark-up Schema of Line Number

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

Table 3: The Special Mark-up Delimiters

5.3 NE Tag Annotation

According to part of speech annotation, we follow the POS tagset provided by ORCHID. NE tagset provided in the corpus consists of Date, Location, Measurement, Name, Organization, Person and Time, as displays in the Table 4. Additionally, BIO format is brought out to identify the chunk or component of NE in the sentence. “B” indicates the beginning of the chunk, and “I” presents the position of NE occurred within the chunk. While “O” is marked to identify that the word does not involve in any types of NE. Figure 6 and 7 illustrate the current Thai NE corpus in Thai and English translation.

Category	Tag	Description	Example
Date	B-DAT	Beginning of Date Name	วันที่ (Date)
	I-DAT	Inside of Date Name	1 มกราคม (January 1)
Location	B-LOC	Beginning of Location Name	จังหวัด (province)
	I-LOC	Inside of Location Name	ปทุมธานี (Pathumthani)
Measure-ment	B-MEA	Beginning of Measurement Name	สาม (Three)
	I-MEA	Inside of Measurement Name	คัน (Car)
Name	B-NAM	Beginning of Proper Name, except Location, Person and Organization Name	ศึก (League)
	I-NAM	Inside of Proper Name	ลาลีกา (La Liga)
Organiza-tion	B-ORG	Beginning of Organization Name	บริษัท (Corp.)
	I-ORG	Inside of Organization Name	เสริมสุข (Sermsuk)
Person	B-PER	Beginning of Person Name	นาง (Mrs.)
	I-PER	Inside of Person Name	สุเทพ เทือกสุบรรณ (Suthep Thaugsuban)
Time	B-TIM	Beginning of Time	เก้า (Ten)
	I-TIM	Inside of Time	โมง (O’clock)
Other	O	Does not belong any types	

Table 4: The Thai Name Entity Tagset

6. Conclusion

We adopted a collection for NE corpus originally prepared by THAI-NEST and undertook POS and NE tagged by N/A, by verifying 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 newly generated corpus consists of 639,335 NE tags. The revised NE tagged corpus with the BiLSTM-CNN-CRF model with word, part-of-speech and character embedding approach improves the NER F1-score 16.21% to mark 89.22%. Our next step is to undertake the NE tagging to ORDHID and distribute for research purpose.

%Title: BKD-7 corpus
%Description: This corpus based on the original THAI-NEST corpus and combined seven types of entities: DATe, LOCation, MEAsurement, NAMe, ORGanization, PERson, TIMe
%Number of sentence: 419
%Number of word: 53,319
%Number of named entity tag: 11,891
%Date: July 31, 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

#S14

เมื่อเวลา 13.00 น. วันที่ 6 พฤษภาคม พ.ต.ท.ไชยศ มุกดาหาญ รอง ผกก.ป.สภ.นครชัยศรี จ.นครปฐม ได้รับแจ้งมีอุบัติเหตุรถบรรทุกสิบล้อ\\ชนกับรถยนต์ มีผู้ได้รับบาดเจ็บ 2 ราย เหตุเกิดที่บริเวณ ถ.เพชรเกษมขาเข้า หมู่ 5 ต.ศีรษะทอง อ.นครชัยศรี จ.นครปฐม …//

เมื่อ/JSBR/O
เวลา/NCMN/O
<space>/PUNC/O
13.00/DCNM/B-TIM
<space>/PUNC/I-TIM
น./CMTR/I-TIM
<space>/PUNC/O
วัน/NCMN/B-DAT
ที่<space>6/DONM/I-DAT
<space>/PUNC/I-DAT
พฤษภาคม/NCMN/I-DAT
<space>/PUNC/O
พ.ต.ท./NTTL/B-PER
ไชยศ/NPRP/I-PER
<space>/PUNC/I-PER
มุกดาหาญ/NPRP/I-PER
…
ผู้/PPRS/O
ได้รับ/VSTA/O
บาดเจ็บ/VSTA/O
<space>/PUNC/O
2/DCNM/B-MEA
<space>/PUNC/I-MEA
ราย/CNIT/I-MEA
<space>/PUNC/O
เหตุ/NCMN/O
เกิด/VSTA/O
ที่/RPRE/O
บริเวณ/NCMN/O
…
//

Figure 6: Thai Name Entity Corpus (Thai)

%Title: BKD-7 corpus
%Description: This corpus based on the original THAI-NEST corpus and combined seven types of entities: DATe, LOCation, MEAsurement, NAMe, ORGanization, PERson, TIMe
%Number of sentence: 419
%Number of word: 53,319
%Number of named entity tag: 11,891
%Date: July 31, 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

#S14

At 13.00 hrs., on May 6, Pol. Col. Chaiyos Mukdahan, Deputy Director of Nakhon Chai Si Police Station, Nakhon Pathom, was informed that a ten-wheel truck accident collided with a car. There are 2 people injured in the accident. The accident occurred at the Inbound of Petchkasem Rd., Village No. 5, Sisa Thong Subdistrict, Nakhon Chai Si District, Nakhon Pathom Province. //

at/JSBR/O
time/NCMN/O
<space>/PUNC/O
13.00/DCNM/B-TIM
<space>/PUNC/I-TIM
o'clock/CMTR/I-TIM
<space>/PUNC/O
day/NCMN/B-DAT
<space>sixth/DONM/I-DAT
<space>/PUNC/I-DAT
May/NCMN/I-DAT
<space>/PUNC/O
Pol.Col./NTTL/B-PER
Chaiyos/NPRP/I-PER
<space>/PUNC/I-PER
Mukdahan/NPRP/I-PER
…
man/PPRS/O
was/VSTA/O
injured/VSTA/O
<space>/PUNC/O
2/DCNM/B-MEA
<space>/PUNC/I-MEA
person/CNIT/I-MEA
<space>/PUNC/O
accident/NCMN/O
occur/VSTA/O
at/RPRE/O
area/NCMN/O
…
//

Figure 7: Thai Name Entity Corpus (English Translation)

7. Bibliographical References

Kobkrit (2019). Thai NLP Resource. Retrieved from https://github.com/kobkrit/nlp_thai_resources.

Thatsanee Charoenporn, Virach Sornlertlamvanich and Hitoshi IsaharaCastor, Building A Large Thai Text Corpus---Part-Of-Speech Tagged Corpus: ORCHID. Proceedings of the Natural Language Processing Pacific Rim Symposium, 1997.

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), Spain.*

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)*.