世界の幸せをカタチにする。
Creating Peace & Happiness for the World

Musashino University

Improvement of Thai NER and the Corpus

Thatsanee Charoenporn[1], Virach Sornlertlamvanich[1,2]

[1]Asia AI Institute, Faculty of Data Science, Musashino University, Japan

[2]SIIT, Thammasat University, Thailand

{thatsane, virach}@musashino-u.ac.jp

SIIT SINCE 1992

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%.

บทคัดย่อ

การพัฒนาคลังข้อความภาษาไทยสำหรับการประมวลผลภาษาธรรมชาตินั้น มีประเภทและปริมาณเพิ่มมากขึ้น แต่คลังข้อความชื่อเฉพาะภาษาไทย หรือ Thai Name Entity Corpus ยังคงมีจำนวนจำกัด แม้ว่างานวิจัยด้านการรู้จำชื่อเฉพาะ (Name Entity Recognition: NER) จะส่งผลต่อความถูกต้องของการประมวลผลขอความเป็นอย่างมากก็ตาม งานวิจัยนี้เสนอวิธีการปรับแต่ง NER แบบวนซ้ำ โดยใช้แบบจำลอง BiLSTM-CNN-CRF ประกอบกับ คำแวดล้อม หน้าที่ของคำ และกลุ่มอักขระข้างเคียง เพื่อปรับปรุงคลังขอคู่วามชื่อเฉพาะภาษาไทย จากเดิม จำนวน 172,232 ชื่อ ให้มีความถูกต้อง แม่นยำและสอดคล้องกัน ผลการวิจัยพบว่า คลังขอความชื่อเฉพาะภาษาไทย ที่ปรับปรุงขึ้น ประกอบด้วยคำและป้ายระบุชื่อเฉพาะ (Tags) จำนวนถึง 639,335 ชื่อ ทั้งนี้ ผลการปรับปรุงคลังข้อความชื่อเฉพาะด้วยแบบจำลองที่นำเสนอนี้ สามารถกำกับชื่อเฉพาะภาษาไทยได้ถูกต้อง วัดด้วยค่า F1-score ได้ที่ 89.22 เปอร์เซ็นต์ ซึ่งให้ผลที่ดีกว่าแบบจำลองที่สร้างด้วยคลังขอความเดิมถึง 16.21 เปอร์เซ็นต์

MOTIVATION

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 focused texts automatically.

Continuously, there have been researched 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.

CHALLENGES IN THAI NE CORPUS CONSTRUCTION

Incorrectness of Word Segmentation and NE Tag Assignment

นายก/NCMN/ O <space> /PUNC/ O อบ /VACT/ O จ. /NTTL/ O อุตรดิตถ์ /NPRP/ O	ร้อยตำรวจเอก/NTTL/ B-PER เฉลิม/NPRP/ I-PER <space>/PUNC/ O อยู่/XVAE/ O บำรุง/VACT/ O

Inconsistency of NE Tagging (LOC)

ราคา/NCMN/ O ทองคำ/NCMN/ O ใน/RPRE/ O ประเทศไทย/NPRPE/ B-LOC ที่/PREL/ O ปรับตัว/VACT/ O สูงขึ้น/ADVN/ O	นายก/NCMN/ O สมาคม/NCMN/ O ลูกจ้าง/NCMN/ O ส่วน/NCMN/ B-LOC ราชการ/NCMN/ O แห่ง/NPRP/ O ประเทศไทย/NPRP/ O

Inconsistency of NE Tagging (NAM)

พรีเมียร์ลีก/NCMN/ O <space>/PUNC/ O อังกฤษ/NCMN/ O <space>/PUNC/ O ฤดูกาล/NCMN/ O <space>/PUNC/ O 2008/NCMN/ 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

NE CORPUS CLEANING PROCESS

We combined to create the BiLSTM-CNN-CRF model for predicting named entity tags. (Fig. 1) The character-level representation of each word is calculated by the CNN, as shown in Fig. 2. 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. Fig.3 illustrates the corpus cleaning up process.

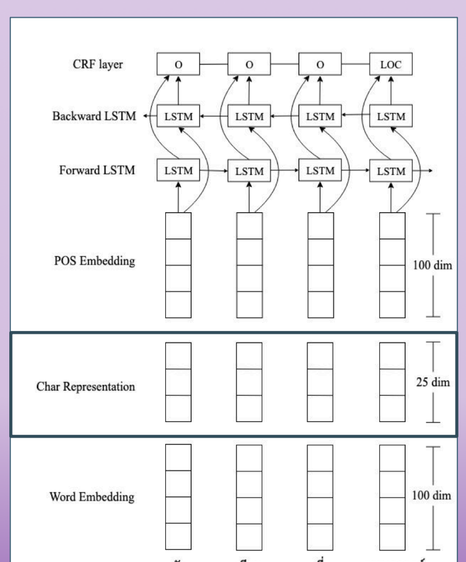

Figure 1. Architecture of the Proposed NER Model

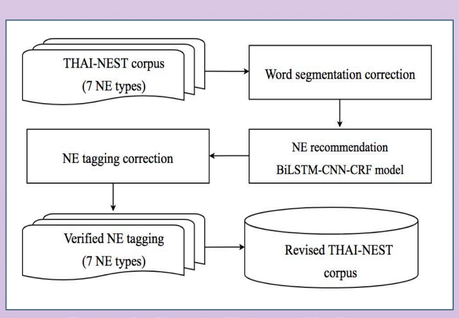

Figure 3. Corpus Cleaning Up Process

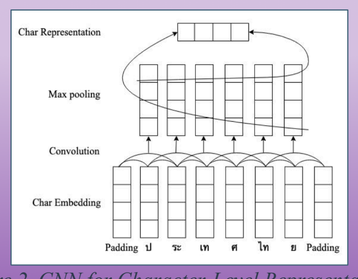

Figure 2. CNN for Character-Level Representation

PROPOSED THAI NE TAGSET

CAT.	TAG	DESCRIPTION	EXAMPLE
DATE	B-DAT	Beginning of Date Name	วันที่ (Date)
	I-DAT	Inside of Date Name	14 กุมภาพันธ์ (February 14)
LOCATION	B-LOC	Beginning of Location Name	เมือง (City)
	I-LOC	Inside of Location Name	นิวยอร์ค (New York)
MEASUREMENT	B-MEA	Beginning of Measurement Name	ห้า (Five)
	I-MEA	Inside of Measurement Name	เล่ม (Book)
NAME	B-NAM	Beginning of Proper Name, except Location, Person and Organization Name	ศึก (League)
	I-NAM	Inside of Proper Name	ลาลีกา (La Liga)
ORGANIZATION	B-ORG	Beginning of Organization Name	บริษัท (Corp.)
	I-ORG	Inside of Organization Name	โตโยต้า มอเตอร์ (Toyota Motor)
PERSON	B-PER	Beginning of Person Name	นาย (Mr.)
	I-PER	Inside of Person Name	ณัฐวุฒิ สะกิดใจ (Natthawut Sakidjai)
TIME	B-TIM	Beginning of Time	สิบ (Ten)
	I-TIM	Inside of Time	นาฬิกา (O'clock)
OTHER	O	Does not belong any types	

RESULT OF COMBINED BKD NE TAGGED CORPUS

ORIGINAL 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
//
```

COMBINED CORPUS

```
%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/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
มี.ค. 52/NPRP/I-DAT
)/PUNC/O
.
.
ยาเสพติด/NCMN/O
ร่วมกัน/ADVN/O
//
```

COMBINED CORPUS

CONCLUSION

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 with the best BiLSTM-CNN-CRF model with word, part-of-speech and character embedding approach improves the NER F1-score 16.21% to mark 89.22%.

REFERENCES

- 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, September 17-21, 2015, Portugal.
- Suriyachay, K., and Sornlertlamvanich, V. **Named Entity Recognition Modeling for the Thai Language from a Disjointedly Labeled Corpus.** The 5th International Conference on Advanced Informatics: Concept Theory and Applications.
- Theeramunkong, T., Boriboon, M., Haruechaiyasak, C., Kittiphattanabawon, N., Kosawat, K., Onsuwan, C., Siriwat, I., Suwa napong, T., and Tongtep, N. **THAI-NEST: A Framework for Thai Named Entity Tagging Specification and Tools**. Proceedings of the 2nd International Conference on Corpus Linguistics, May 13-15, 2010, University of A Coruña, Spain.
- X, Ma., and E, Hovy. **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). August 7-12, 2016. Berlin, Germany.