Named Entity Recognition Modeling for the Thai Language from a Disjointedly Labeled Corpus

Kitiya Suriyachay
School of ICT,
Sirindhorn International Institute of Technology,
Thammasat University, Thailand
Kitiya55160175@gmail.com

Virach Sornlertlamvanich
School of ICT,
Sirindhorn International Institute of Technology,
Thammasat University, Thailand
virach@siit.tu.ac.th

Abstract— **In the Thai language, named entity can be used with or without a prefix or an indication of word. This may cause confusion between named entity and other types of noun. However, a named entity is likely to be used in adjacent to verbs or prepositions. This means that the adjacent verbs or prepositions to a noun can be as a good feature to determiner the type of named entity. There are some studies on named entity recognition (NER) task in other languages such as Indonesian showing that combination of word embedding and part-of-speech (POS) tag can improve the performance of the NER model. In this paper, we investigate the Thai Named Entity Recognition task using Bi-LSTM model with word embedding and POS embedding for dealing with the relatively small and disjointedly labeled corpus. We compare our model with the one without POS tag, and the baseline model of CRF with the similar set of feature. The experiment results show that our proposed model outperforms the other two in all F1-score measures. Especially, in the case of location file, the F1-score is increased by 14 percent.**

Keywords-Named Entity Recognition; Recurrent Neural Network; Thai language; Bi-LSTM

I. INTRODUCTION

In Natural language processing, Named Entity Recognition (NER) is one of the most important tasks which is very popular and has been researched continuously. Dealing with NER in some languages is difficult (e.g. Chinese, Indian, Japanese and Thai). These languages do not contain character like capitalization in English to identify a named entity and there are no spaces to separate each word in the sentence.

Furthermore, Thai language is more complicated than English because of Thai writing has a variety of styles. For example, sometimes writing in Thai use the full name of named entity for the first time and when referring to that named entity again, Thais often use abbreviation or cut a prefix or an indication off such as "**ห้างสรรพสินค้าเซ็นทรัล**เริ่มให้บริการซื้อสินค้าบนสังคมออนไลน์ในปีที่ผ่านมาทางโมบายแอปพลิเคชัน ในปัจจุบัน**เซ็นทรัล**มีการให้บริการบนเฟสบุ๊ค ทวิตเตอร์ ไลน์และอินสตาแกรม" (**Central Department Store** began providing shopping on social media platforms last year via mobile application. At present, **Central** offers services on Facebook, Twitter, Line and Instagram). For this reason, it is ambiguous and difficult to distinguish between named entity and the other types of noun. However, one of the problems or limitations for Thai Named Entity Recognition is having a small amount of corpus, which is not enough capture the accurate model for the Thai NER. Furthermore, the THAI-NEST corpus, which we use in this task, having been labeled with only one type of named entity in each file. This causes the task not being able to use other named entity context in training process.

To solve these problems, in this paper, we present a Named Entity Recognition for Thai language corpus using Bi-LSTM with word embedding and POS embedding. Since most of named entities can likely be determined by the nearby verbs or prepositions such as ไป (go), ใน (in), ที่ (at), and จาก (from), so we explicitly prepare part-of-speech such as verb or preposition of each word in our model to predict named entity type of word. The bi-directional approach in Bi-LSTM can also capture the context from both left and right hand sides of the sentence.

Previous studies that are related to NER are reviewed and described in section 2. Section 3 explains about the corpus we use in the experiment. In section 4, we explain the details of our methods, procedures, and the model architecture. The results of the experiments are discussed in section 5 and section 6, showing the comparison results between the proposed Bi-LSTM model and CRF model. Finally, section 7 is the summary.

II. RELATED WORK

Over the past years, Named Entity Recognition has been quite popular and has been being developed to improve system performance. However, Named Entity Recognition is one of the most challenging problems since there is only a small number of supervised training data available for many languages, while to name the entity of words, there are quite some constraints. Thus, having only a small amount of data is insufficient [3]. There are several approaches that can be used to solve the problem of Named Entity Recognition, but most of the effective NER approaches are usually based on machine learning techniques [1].

In several previous researches, traditional machine learning approaches is widely used in NLP, for example, Named Entity Recognition for Hindi language using Hidden Markov Model (HMM) [6], use Support Vector Machine (SVM) to recognize Biomedical Named Entity [7]. In additional, CRF is used for NER in many languages such as Chinese [8], Manipuri language [9], Malay language [10], and Thai language [11]. Although these approaches are robust and reliable, but they have some shortcoming that may affect system performance, such as the process to reconstruct a set of system features is difficult when changing the corpus or language [4], and CRF model hardly yields the result of a word that has never met in a model training.

For the Thai Named Entity Recognition conducted by CRF model [11], they compare the performance of their model during the use of word-segmented data and syllable-segmented data as features of the system. The comparison result shows that the system using a syllable-segmented data is a bit more effective than word-segmented data. One of the problems in this model is the model cannot recognize words or syllables that have never appeared in the training dataset such as abbreviations.

Recently, the Deep Learning architectures have impressive advances in various field. As for the NLP task, it provides better results than traditional approaches. Recurrent Neural Network (RNN) model is a type of Deep Learning that is suitable for Named Entity Recognition, however, such RNN has a problem of long-term dependencies. Thus, LSTM is the most used type of RNN, because LSTM can handle the problem of long-term dependencies better. Reference [12] researched Named Entity Recognition for Chinese telecommunication information using the Deep Learning model. They use Bi-LSTM together with character embedding instead of using word embedding. The results of the research showed that their model provides better performance than the traditional machine learning, and character embedding was more suitable for NER in Chinese language more than word embedding.

In addition, reference [5] also created NER model to recognize information on Twitter in the Indonesian language which the model they used were Bi-LSTM. The best result of the model was to combine word embedding with the POS tag, which provides the most F1-score value. So, POS tag was a useful feature that allowed the system to decide named entity tag correctly.

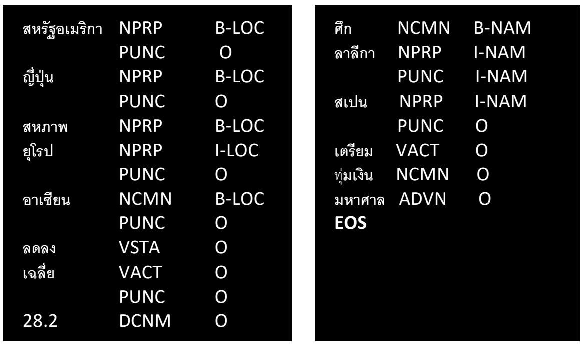
III. Corpus

The corpus used in our experiment is the THAI-NEST corpus. This corpus is collected from the Thai online news articles, which are published on the Internet such as political news, economic news, crime news, sport news, entertainment news, educational news and technological news [2]. The THAI-NEST corpus is already tokenized the sentence into words and punctuation. The corpus is disjointedly managed in seven categories by type of named entity including DATe, TIMe, MEAsurement, NAMe, LOCation, PERson, and ORGanization, where each category is abbreviated by the first three characters. The number of words and all named entity tags in each file are shown in table 1.

TABLE I. NUMBER OF SENTENCES, WORDS AND NAMED ENTITY TAG 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

The format of the data in the corpus is shown in Fig 1, consisting of three components separated by a tab for each line. The first component is a word, the second component is part-of-speech (POS) of the word, and the last component is the category or tag of the word in the same line. Some lines consist of only one part which is EOS, indicating the end of sentence. As for the named entity format, we use the BIO format ("B-" is beginning of a chuck. "I-" is inside a chuck and O denotes that the word does not belong to any type of entities).



```
สหรัฐอเมริกา  NPRP    B-LOC
              PUNC     O
ญี่ปุ่น          NPRP    B-LOC
              PUNC    O
สหภาพ         NPRP    B-LOC
ยุโรป          NPRP    I-LOC
              PUNC    O
อาเซียน        NCMN    B-LOC
              PUNC    O
ลดลง          VSTA    O
เฉลี่ย          VACT    O
              PUNC    O
28.2          DCNM    O
```

(a)

```
ศึก       NCMN    B-NAM
ลาลีกา    NPRP    I-NAM
          PUNC    I-NAM
สเปน     NPRP    I-NAM
          PUNC    O
เตรียม    VACT    O
ทุ่มเงิน    NCMN    O
มหาศาล   ADVN     O
EOS
```

(b)

Figure 1. Example labeled data in different corpus file (a) Location file and (b) Name file

IV. Methodology

This section describes the procedure of data preparation and the architecture of the model used in research.

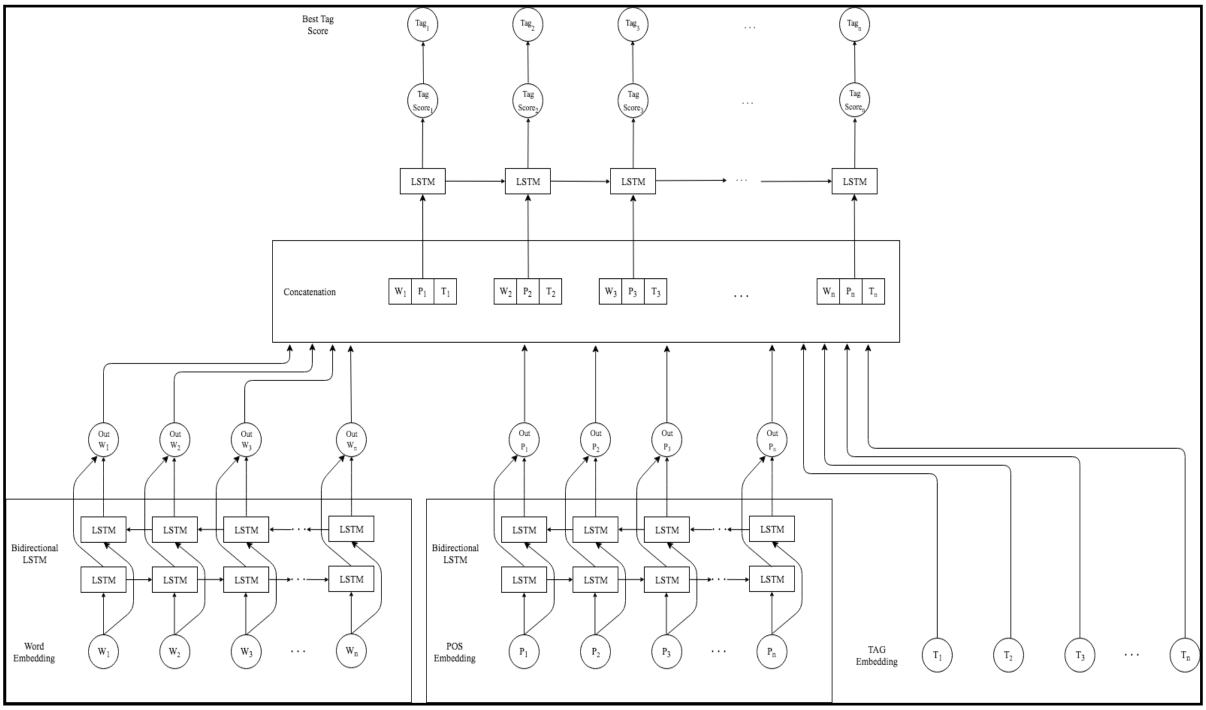


Figure 2. Architecture of the proposed NER model

A. *Pre-processing Data*

As mentioned above, in section 3, the corpus we use consists of three components: word, part-of-speech, and named entity tag. Each component is separated by a tab, so, firstly, we split tab of each line to separate words, POS and named entity tags off from each other. We change the format of named entity tag from BIO format to IO format since Thai language often cut indication or prefix of named entity off, thus we cannot measure score of separate B and I tag. Then, we store the extracted data in form of word lists, part-of-speech lists and named entities tag lists respectively and we need three dictionaries with a word, POS and named entity tag as a key and set of index of these keys as a value. Each corpus is divided into two parts: 80% of all the sentences in the file for training set and 20% for the test set.

B. *Model*

When we successfully complete all pre-processing data procedures. The data is fed into the Bi-LSTM model. This model composed of five important layers as follows: Word Embedding layer, POS Embedding layer, Tag Embedding layer, Bidirectional Long Short-Term Memory (Bi-LSTM) layer, and the last layer is Long Short-Term Memory (LSTM) layer.

Our model architecture is shown in Fig 2, the input to the embedding layer is the list of indices where W is the indices list of word, P is the indices list of part-of-speech list and T is the indices list of named entity tag list and the output is their corresponding embedding. Next, the output of each embedding layer is fed to the Bi-LSTM layer of themselves, except the output of the tag embedding layer. Then, the last LSTM layer brings the sequence of the vector caused by the concatenation of each word, part-of-speech and tag as input to predict named entity tag of each word. The output of last LSTM layer is in the form of vector and named entity tag that comes from the prediction of model is the highest value tag in vector of each word.

V. RESULT AND DISCUSSION

In the experiment, we measure precision, recall and F1-score to evaluate model performance. Additionally, we want to prove that the POS has an effect on the prediction of named entity of the model. Thus, to do this, we use the same Bi-LSTM model but without POS.

From the results shown in table 2, all F1-score values are higher than the one in table 3. It shows that the use of POS in a model makes the F1-score significantly increased when comparing to the one without POS, especially, the location file where the F1-score increases by approximately 14 percent, followed by organization file where the F1-score increases by 8.5 percent.

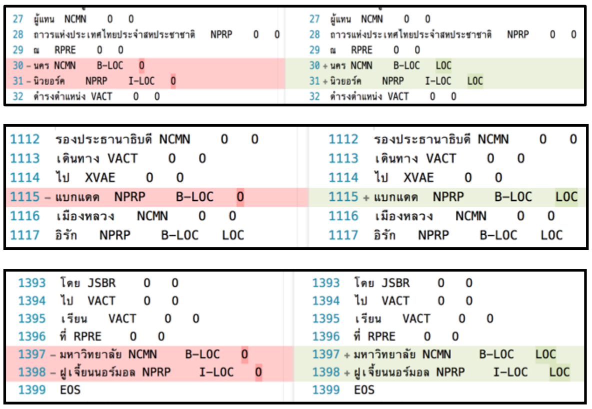
TABLE II. THE RESULT OF BI-LSTM MODEL WITH POS

	Precision	Recall	F1-score
DAT	91	88	90
LOC	89	81	85
MEA	83	77	80
NAM	75	72	73
PER	80	76	80
ORG	86	78	82
TIM	93	94	94

TABLE III. THE RESULT OF BI-LSTM MODEL WITHOUT POS

	Precision	Recall	F1-score
DAT	94	85	89
LOC	71	74	73
MEA	80	78	79
NAM	72	70	71
PER	77	75	76
ORG	79	72	75
TIM	90	92	91

We took some samples of the results of the Bi-LSTM model of the location file comparing between the one using POS and not using POS, as shown Fig 3. The location file has the most difference in F1-score between these two models.



(a)

```
27   ผู้แทน  NCMN    0   0
28   ถาวรแห่งประเทศไทยประจำสหประชาชาติ   NPRP    0   0
29   ณ   RPRE    0   0
30 - นคร NCMN    B-LOC   0
31 - นิวยอร์ค    NPRP    I-LOC   0
32   ดำรงตำแหน่ง VACT    0   0
```

```
1112   รองประธานาธิบดี NCMN     0    0
1113   เดินทาง VACT     0    0
1114   ไป   XVAE     0    0
1115 - แบกแดด   NPRP     B-LOC    0
1116   เมืองหลวง    NCMN     0    0
1117   อิรัก    NPRP     B-LOC    LOC
```

```
1393   โดย JSBR     0    0
1394   ไป   VACT     0    0
1395   เรียน    VACT     0    0
1396   ที่ RPRE     0    0
1397 - มหาวิทยาลัย NCMN     B-LOC    0
1398 - ฝูเจี้ยนนอร์มอล NPRP     I-LOC    0
1399   EOS
```

(b)

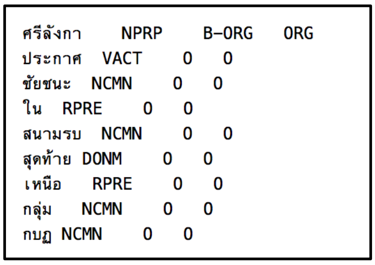
```
27   ผู้แทน  NCMN    0   0
28   ถาวรแห่งประเทศไทยประจำสหประชาชาติ   NPRP    0   0
29   ณ   RPRE    0   0
30 + นคร NCMN    B-LOC   LOC
31 + นิวยอร์ค    NPRP    I-LOC   LOC
32   ดำรงตำแหน่ง VACT    0   0
```

```
1112   รองประธานาธิบดี NCMN     0    0
1113   เดินทาง VACT     0    0
1114   ไป   XVAE     0    0
1115 + แบกแดด   NPRP     B-LOC    LOC
1116   เมืองหลวง    NCMN     0    0
1117   อิรัก    NPRP     B-LOC    LOC
```

```
1393   โดย JSBR     0    0
1394   ไป   VACT     0    0
1395   เรียน    VACT     0    0
1396   ที่ RPRE     0    0
1397 + มหาวิทยาลัย NCMN     B-LOC    LOC
1398 + ฝูเจี้ยนนอร์มอล NPRP     I-LOC    LOC
1399   EOS
```

Figure 3. Sample of result between Bi-LSTM model (a) not using POS and (b) using POS of location file

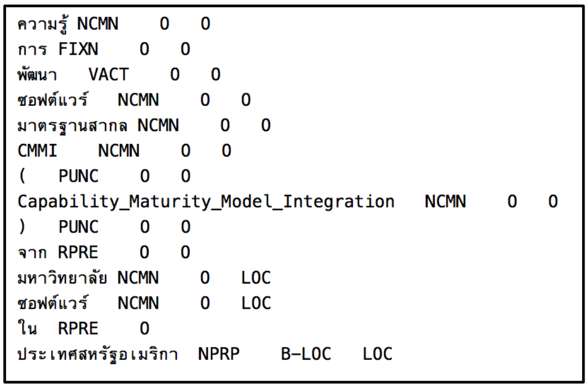
In Fig 3, the third column shows the correct answer, and the fourth column is the predicted named entity tag from the model. It is apparent that Bi-LSTM model which does not have part-of-speech cannot predict some types of named entity correctly, while the model that uses POS provides a better result. For the Thai language, the named entity is often adjacent or close to a preposition or a verb that determines a location, and sometimes can be used to identify the organization name. For example, "ฉันกำลังจะไปกรุงเทพ" (I am going to go to Bangkok), the word "ไป" (go) is a verb next to "กรุงเทพ" (Bangkok) which is the name of Thailand's capital. In this case, with the verb "ไป" (go), the name "กรุงเทพ" (Bangkok) is correctly recognized as a location, not an organization. On contrary, in the sentence "ศรีลังกาประกาศชัยชนะในสนามรบสุดท้ายเหนือกลุ่มกบฏ" (Sri Lanka announces victory in the final battle with rebels), the verb "ประกาศ" (announce) on the right hand side of the name "ศรีลังกา" (Sri Lanka) shows its agentive role in the sentence. Therefore, the "ศรีลังกา" (Sri Lanka) is well recognized as an organization, not the location, as shown in Fig 4. One more example concerning the usage of preposition, in the sentence "ฉันเรียนอยู่ที่มหาวิทยาลัยธรรมศาสตร์" (I study at Thammasat University), the word "ที่" (at) is a preposition adjacent to the name "มหาวิทยาลัยธรรมศาสตร์" (Thammasat University) which is recognized as a location. Both preposition and verb have a high influence in predicting the type of named entity. Bi-LSTM model learns to predict named entity from the surrounding words and POSs. When it finds the named entity that is near or adjacent to the POS like preposition and verb, the model will be able to predict named entity more accurately.



```
ศรีลังกา     NPRP     B-ORG    ORG
ประกาศ   VACT    0    0
ชัยชนะ   NCMN    0    0
ใน   RPRE    0    0
สนามรบ   NCMN    0    0
สุดท้าย DONM    0    0
เหนือ    RPRE    0    0
กลุ่ม    NCMN    0    0
กบฏ NCMN    0    0
```

Figure 4. Sample of result of organization file

Nevertheless, there are some mistakes in named entity prediction as shown in Fig 5. Our model predicts some named entity incorrectly "มหาวิทยาลัยซอฟต์แวร์" (University for software development) as a location, while it should be labeled as other. The model may highly consider the possible context preposition "จาก" (from).



```
ความรู้ NCMN    0    0
การ FIXN    0    0
พัฒนา    VACT    0    0
ซอฟต์แวร์    NCMN    0    0
มาตรฐานสากล NCMN    0    0
CMMI     NCMN    0    0
(    PUNC    0    0
Capability_Maturity_Model_Integration    NCMN    0    0
)    PUNC    0    0
จาก RPRE    0    0
มหาวิทยาลัย NCMN    0    LOC
ซอฟต์แวร์    NCMN    0    LOC
ใน   RPRE    0
ประเทศสหรัฐอเมริกา   NPRP    B-LOC    LOC
```

Figure 5. Sample of prediction error

VI. COMPARISON OF BI-LSTM MODEL AND CRF MODEL

The CRF model was used in this research is the CRF++ 0.58, which developed by Taku Kudo and free for research proposes [13]. The Corpus used for this model is the same set as Bi-LSTM model and is divided into two parts as well, which are 80% for training and 20% for the testing.

For this CRF model, the feature used in the model is the words and POS of each word. We create the template file of this CRF model from these features. We have concluded the results of CRF model training of every corpus as shown in Table 4. Comparison the results of CRF model with F1-score of the Bi-LSTM model in Table 1, the F1-score of Bi-LSTM model is quite much higher than that of CRF model. The Bi-LSTM model is more effective in NER task over CRF model because the Bi-LSTM model can memory the data coming into a model as long-term information and learn what should be kept or what should be discarded while the CRF model only learn from the current word and adjacent words which we set to use three previous words and the next three words.

TABLE IV. THE RESULT OF CRF MODEL WITH POS

	Precision	Recall	F1-score
DAT	84	75	79
LOC	72	75	73
MEA	74	61	67
NAM	69	52	59
PER	75	54	63
ORG	89	74	81
TIM	95	88	92

VII. CONCLUSION

This paper proposes a new method for Named Entity Recognition using Bi-LSTM model with POS tag which is tolerant to the small and disjointedly labeled THAI-NEST corpus. Our model provides the best performance for the NER of Thai language. The results of experiments show that using word together with POS in the model helps improve the performance of our model in named entity prediction even though some named entities have no indication or prefix. We have also compared the results between the Bi-LSTM model and the CRF model using the same corpus. As expected, Bi-LSTM model provides much better results, whose average F1-score is more than 10 percent better than the one of CRF model.

However, one of the limitations of this research is that the sizes of each type of the named entity are quite different. There are only 3,362 time labeled words in the total files of 41,493 words, while there are 222,075 person labeled words in the total files of 2,705,218 words. If the experiment can be conducted on the more data and comparable size of data in every file, the results may change. We can also directly compare the results between each other.

ACKNOWLEDGMENT

We would like to express our gratitude to Mr. Panuwat Assawinjaipetch and Mr. Songwong Tasneeyapant for their support and good guidance about the implementation and tackling the research problems.

REFERENCES

[1] Limsopathan. N, and Collier. N, “Bidirectional LSTM for Named Entity Recognition in Twitter Messages” in Proceedings of the 2nd Workshop on Noisy User-generated Text, 2016, pp. 145-152.

[2] Theeramunkong. T et al, “THAI-NEST: A framework for Thai named entity tagging specification and tools”, 2010.

[3] Lample. G, Ballesteros. M, Subramanian. S, Kawakami. K, and Dyer. C, (2016). Neural Architectures for Named Entity Recognition. arXiv:1603.01360v3.

[4] L. Li, L. Jin, Z. Jiang, D. Song, and D. H, “Biomedical named entity recognition based on Extended Recurrent Neural Networks,” in *IEEE International Conference on Bioinformatics and Biomedicine,* Washington DC, USA, Nov. 2015, pp. 649-652.

[5] V. Rachman, S. Savitri, F. Augustianti, and R. Mahendra, “Named entity recognition on Indonesian Twitter posts using long short-term memory networks,” in 2017 International Conference on Advanced Computer Science and Information Systems, 2017, Indonesia, pp. 228-232.

[6] D. Chopra, N. Joshi,and I. Mathur, “ Named Entity Recognition in Hindi Using Hidden Markov Model,” in 2016 Second International Conference on Computational Intelligence & Communication Technology, India, 2016, pp. 581-586.

[7] Z. Ju, J. Wang, and F. Zhu, “Named Entity Recognition from Biomedical Text Using SVM,” in 5th International Conference on Bioinformatics and Biomedical Engineering, 2011, Wuhan, China, pp. 1-4.

[8] K. Liu, Q. Hu, J. Liu, and C. Xing, “Named Entity Recognition in Chinese Electronic Medical Records Based on CRF,” in 14th Web Information Systems and Applications Conference, 2017, China, pp. 105-110.

[9] K. Nongmeikapam, T. Shangkhunem, N. M. Chanu, L. N. Singh, B. Salam, S. Bandyopadhyay, "CRF Based Name Entity Recognition (NER) in Manipuri: A Highly Agglutinative Indian Language," in 2nd National Conference Emerging Trends and Applications in Computer Science, 2011, *India*, pp. 1-6.

[10] M. Sharilazlan Salleh, S. Azirah Asmai, H. Basiron, S. Ahmad, “A Malay Named Entity Recognition using conditional random fields,” in 2017 5th International Conference on Information and Communication Technology, 2017, Malaysia, pp. 1-6.

[11] N. Tirasaroj, W. Aroonmanakun, “Thai Named Entity Recognition Based on Conditional Random Fields,” in 2009 Eighth International Symposium on Natural Language Processing, 2009, Bangkok, Thailand, pp. 216-220

[12] Y. Wang, B. Xia, Z. Liu, Y. Li, and T. Li, “Named entity recognition for Chinese telecommunications field based on Char2Vec and Bi-LSTMs,” in 12th International Conference on Intelligent Systems and Knowledge Engineering, 2017, Nanjing, China, pp. 1-7.

[13] T. Kudo, “CRF++: Yet Another CRF toolkit”.[Online]. Available: https://taku910.github.io/crfpp/. [Accessed: Feb. 17, 2017] .