NAMED ENTITY RECOGNITION MODELING FOR THE THAI LANGUAGE FROM A DISJOINTEDLY LABELED CORPUS

BY

MS. KITIYA SURIYACHAY

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE (ENGINEERING AND TECHNOLOGY)
SIRINDHORN INTERNATIONAL INSTITUTE OF TECHNOLOGY
THAMMASAT UNIVERSITY
ACADEMIC YEAR 2019

NAMED ENTITY RECOGNITION MODELING FOR THE THAI LANGUAGE FROM A DISJOINTEDLY LABELED CORPUS

BY

MS. KITIYA SURIYACHAY

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE
(ENGINEERING AND TECHNOLOGY)
SIRINDHORN INTERNATIONAL INSTITUTE OF TECHNOLOGY
THAMMASAT UNIVERSITY
ACADEMIC YEAR 2019

THAMMASAT UNIVERSITY
SIRINDHORN INTERNATIONAL INSTITUTE OF TECHNOLOGY

THESIS

BY

MS. KITIYA SURIYACHAY

ENTITLED

NAMED ENTITY RECOGNITION MODELING FOR THE THAI LANGUAGE
FROM A DISJOINTEDLY LABELED CORPUS

was approved as partial fulfillment of the requirements for
the degree of Master of Science (Engineering and Technology)

on December 12, 2019

Chairperson

(Choochart Haruechaiyasak, Ph.D.)

Member and Advisor

Natsuda K.

(Natsuda Kaothanthong, Ph.D.)

Member and Co-advisor

(Virach Sornlertlamvanich, D.Eng.)

Member

(Assistant Professor Teerayut Horanont, Ph.D.)

Director

(Professor Pruettha Nanakorn, D.Eng.)

Thesis Title	NAMED ENTITY RECOGNITION MODELING FOR THE THAI LANGUAGE FROM A DISJOINTEDLY LABELED CORPUS
Author	Ms. Kitiya Suriyachay
Degree	Master of Science (Engineering and Technology)
Faculty/University	Sirindhorn International Institute of Technology/ Thammasat University
Thesis Advisor	Dr. Natsuda Kaothanthong, Ph.D.
Thesis Co-Advisor	Dr. Virach Sornlerdlumvanich, Ph.D.
Academic Years	2019

ABSTRACT

The main purpose of this research is to develop an efficient and effective Thai Named Entity Recognition by using deep learning approach and corpus in Thai language which name entity in each type is labeled only in each different file. Named Entity Recognition (NER) in the Thai language is a challenge task as the Thai language does not have definite word boundary markers. Sometimes, the problem of incorrect word segmentation can occur and affecting the efficiency in processing of NER task. In addition, the ambiguity between common nouns and named entities are another important problem because named entities and some common nouns have the same to spelling. According to the Thai language, most of named entities are usually close to a verb or a preposition. This implies that the POS can be one of the good feature to determine the type of named entity. Therefore, for these reasons, this research presents the Bi-LSTM-CNN-CRF model with the combination feature among word, POS and Thai character clusters (TCCs). TCCs is used instead of single character to deal with the problem of word segmentation error in the corpora and increase the performance of the model. The experiment results show that the proposed model outperforms the other baseline models in all seven named entity types. The TCC is a suitable feature and gives better results than single character embedding.

Keywords: Named Entity Recognition, Thai NER, Recurrent Neural Network, Bidirectional LSTM, CNN, CRF, Thai language, Thai named entity, TC

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to Dr. Virach Sornlertlamvanich for his invaluable help, constant support and patience throughout the course of this research. I am most grateful for all teaching and guidance not only research methodologies but there are still other principles in life. This research would not have been completed without these valuable support and guidance.

I would like to thank Dr. Natsuda Kaothantho, Asst. Prof. Dr. Teerayut Horanon, and Dr. Choochart Haruechaiyasak, for the suggestions and the observations that made me develop the concepts and ponder the problems more carefully until the research was successful.

In addition, thank you to all laboratory members, my friends and everyone else who helped contribute to this research for the encouragement and assistance from the beginning to completion.

Last of all, I am extremely grateful to my parents for all their support, especially for education and for the encouragement throughout the period of this research.

Ms. Kitiya Suriyachay

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS/ABBREVIATIONS

Symbols/Abbreviations	**Terms**
NLP	Natural Language Processing
NER	Named Entity Recognition
POS	Part-of-speech
TCC	Thai Character Cluster
Bi-LSTM	Bidirectional Long Short-Term Memory
CRF	Conditional Random Field
CNN	Convolutional Neural Network
NE	Named entity

CHAPTER 1
INTRODUCTION

1.1 Background

In Natural Language Processing (NLP), Named Entity Recognition (NER) is one of the first and most important stage which has been continuously researched for many years in variety of approaches. The task of NER is to automatically identify and classify an entity of each word in the sentences such as location names, organization names, person names, dates and times. NER is a key success and plays an important role for many fields in NLP like question answering, information retrieval, information extraction and machine translation.

However, dealing with NER in some languages can be even more difficult than English or other European languages (e.g. Chinese, Japanese and Thai language). In English, NER is broadly used and successful in many NLP applications. Due to the characteristics of English are beneficial to NER, whether it is the capital letter at the beginning of a proper noun and there is a space to separate each word from its neighboring words. Unfortunately, Thai NER is still limited since these characteristics do not exist in Thai language. There are several challenges in Thai NER. Firstly, unlike European languages, there is no explicit word boundary and even writing also has no stopping mark at the end of the sentence. Secondly, Thai words are implicitly recognized and some depend on the individual judgment. In addition, there is no capitalization to identify named entities. Even though, there are some markers in some cases identifying proper nouns like person name, institution or location name but sometimes Thai often use abbreviation or cut a prefix or an indication off. For example, “**ห้างสรรพสินค้าเซ็นทรัล**เริ่มให้บริการซื้อสินค้าบนสังคมออนไลน์ในปีที่ผ่านมาทางโมบายแอปพลิเคชัน ในปัจจุบัน **เซ็นทรัล**มีการให้บริการบนเฟสบุ๊ค ทวิตเตอร์ ไลน์ และอินสตาแกรม” (**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). These things lead to incorrect word boundary identification and segmentation, certainly affects the recognition of named entity in NER process. These problems are considered as the challenges of Thai NER.

With all these reasons, the study and research about Thai Natural Language Processing is a challenge and quite difficult. 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. In addition, the THAI-NEST corpus, which is used 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. 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. Therefore, the main purpose of this study is to develop Named Entity Recognition model for Thai language and clean up the existing NE corpus.

This study aims to propose a bi-directional LSTM model with a convolution neural network (CNN) and a conditional random field layer (CRF) using the word, part-of-speech of each word and Thai character cluster as input features. Due to most of Thai named entities can likely be determined by the nearby verbs or prepositions such as ไป (go), ใน (in), ที่ (at), and จาก (from), so POS of each word is introduced to be used with the model to predict named entity type of word. In addition, there are several previous studies that support the use of POS with word in the NER model and provides better performance than the model that does not use POS (Rachman, Savitri, Augustianti, & Mahendra, 2017).

As well as in NER, incorrect word segmentation will lead to false named entity recognition. Therefore, some research use only character instead of word as the input of NER model such as Wang, Xia, Liu, Li and Li (2017). However, much research on NER use both word and character in the model and provide good results (Wang, Bao, & Gao, 2016; Chiu & Nichols, 2016; Ma & Hovy, 2016). So, the part-of-speech and Thai character cluster (TCC) are considered that they can improve the performance of the NER model. In addition, TCC can handle words that have been segmented incorrect.

1.2 Objective

1. To propose an efficient method to generate an NER model from a noisy NE corpus.
2. To compose an NE corpus from the existing disjointedly annotated corpus.
3. To improve NE annotation accuracy by introducing TCC in character-level representation.

1.3 Scope

Develop Thai Named Entity Recognition model for recognizing seven types of entities including person name, location name, organization name, measurement, date, time, and any proper name using the existing NE corpus called THAI-NEST corpus which name entity in each type is labeled only in each different file.

1.4 Expected result

1. High consistency and fewer morphological errors in NE corpus.
2. High accuracy NE annotation model.
3. Large scale NE tagged corpus.

CHAPTER 2
REVIEW OF LITERATURE

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 (Lample, Ballesteros, Subramanian, Kawakami & Dyer, 2016). There are several approaches that can be used to solve the problem of Named Entity Recognition, but one of the most effective and popular approaches are usually based on machine learning techniques (Limsopathan & Collier, 2016).

2.1 Meaning of Named Entity

In general, nouns usually divide into 2 types which are common nouns and proper nouns. A Common noun is a word used to refer to place, people, animal, or things in general, not specifically such as university, river, dog, man, woman, etc. A proper noun is the name given to something to make it more specific such as “แม่น้ำเจ้าพระยา” (the Chao Phraya River), “นายสมชาย” (Mr. Somchai) and “มหาวิทยาลัยธรรมศาสตร์” (Thammasat University). Normally, in English, proper nouns begin with uppercase letters and easier to classify common and proper nouns, which is a characteristic that Thai language does not have.

Name entity is a real-world object, such as persons, places, and organizations that can be referenced or represented by a proper name. Named entity can be either meaningful or meaningless or abstract. Thus, this can be said that the named entity is a proper noun.

In general, in the Thai language, a proper name usually occurs with a common noun which indicates the type of the proper name (common noun + proper name)

2.2 Machine Learning

Machine learning is a type of artificial intelligence that provides systems with the ability to learn and recognize new patterns in data from experience without being explicitly programmed. In several previous researches, traditional machine learning approaches is widely used in NLP. The term “traditional” means the things which have been doing for many years and is often the foundation for more cutting-edge machine learning. For example, Support Vector Machine (SVM), Decision Tree, Hidden Markov Model (HMM), Maximum Entropy, Conditional Random Field (CRF).

Chopra, Joshi and Mathur (2016) presented the Hidden Markov Model to recognize the named entity in Hindi language. The model performed training and testing on 2,343 tokens and 105 tokens respectively and their model achieved F1-measures of 97.14%. But the limitation of HMM is this model requires large amount of training data and it cannot be used for large dependencies.

Ju, Wang and Zhu (2011) performed Named Entity Recognition for Biomedical Text such as protein, gene and DNA using Support Vector Machine. They use word shape and POS as features in the model to recognize biomedical named entity.

In addition, CRF is another method used for NLP in many languages. For example, Salleh, Asmai, Basiron and Ahmad (2017) proposed Malay Named Entity Recognition using CRF model. Some features of Malay language like capitalization, lowercase, previous and neighboring word, word suffix, digit, word shape and POS are used for training the model. The model provided F1-measure of 70% since training the model with small dataset. Liu, Hu, Liu and Xing (2017) used bag-of-characters, part-of-speech, dictionary feature and word clustering as features in the CRF model for Chinese electronic medical records recognition. Yang and Huang (2018) also developed a CRF model with character embedding, POS, radical, PinYin, dictionary and rule features to recognize named entities from Chinese clinical text and achieved F1-measure of 89.26%. They also compared results between CRF and Bi-LSTM-CRF model and CRF provided a better performance because the amount of dataset is not large enough and cause overfitting in Bi-LSTM-CRF. So, the result of Bi-LSTM-CRF will be better, if the amount of dataset become larger.

These researches showed that the important thing of machine learning approach is the scale of corpus and the selection of appropriate features because it directly affects the performance of the NER model. 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 (Li, Jin, Jiang, Song & H, 2015), and CRF model hardly yields the result of a word that has never met in a model training.

2.3 Bidirectional Long Short-Term Memory (Bi-LSTM)

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 effectively capture the problem of long-term dependencies better. Furthermore, not only information from previous words is useful for prediction, but information from words coming after is useful also. This can be done by having a second LSTM running backward with another set of parameters, and this pair of forward and backward LSTMs is referred as a bidirectional LSTM (Bi-LSTM). The bi-directional approach in Bi-LSTM can also capture the context from both left and right-hand sides of the sentence.

Rachman, Savitri, Auguatianti and Mehendra (2017) created Bi-LSTM model for Indonesian information on Twitter and proved that using both word embedding and POS tag provides the most F1-score. Wang, Xia, Liu, Li and Li (2017) researched Named Entity Recognition for Chinese telecommunication information using the Deep Learning model. They use Bi-LSTM with character embedding instead of word embedding. Their results show that character embedding is more suitable and useful for Chinese NER than word embedding and the model performs better than other traditional baselines. In Chinese, a word usually contains several characters, for example, the word “智能” (intelligence). The semantic meaning of this word can be learned from the context around the word in the sentence. Meanwhile, its meaning can be deduced from the meaning of the characters in the word “智” (intelligent) and “能” (ability). Therefore, the semantic meaning of internal characters may play an important role in modeling (Chen, Xu, Liu, Sun, & Luan, 2015; Wang, Xia, Liu, Li & Li, 2017).

2.4 A hybrid approach for Named Entity Recognition

There are several previous researches that use the word together with the character and provide better results than using either word or character. Wang, Bao and Gao (2016) conducted NER model for Mongolian language using word embedding with character embedding as input features in the Bi-LSTM with CRF layer. The experiment results showed that Bi-LSTM-CRF improved the recall score that make F1-measure higher and its performance outperform traditional CRF model. Chiu and Nichols (2016) introduced a combined model of Bi-LSTM and CNN which utilized both word and character-level features for two major English datasets. They obtained F1-measure of 91.62% on CoNLL-2003 dataset and 86.28% on OntoNotes dataset.

There are several researches on Chinese NER that use a hybrid approach. For example, E and Xiang (2017) proposed a Bi-LSTM-CRF model for Chinese NER based on both character and word embedding. Their experiments showed that the incorporation of character and word can improve the performance of NER for Chinese corpus and the F1-measure increases almost 9% from previously experiments.

Furthermore, Ma and Hovy (2016) also presented a powerful Bi-LSTM-CNN-CRF model which achieved state-of-the-art performance on NER and POS tagging and successfully employed CNN to extract more useful character-level features.

2.5 Named Entity Recognition in Thai language

NER in Thai language has been researched over the years. For example, Thai proper name identification using feature-based approach with context words and collocations (Charoenpornsawat, Kijsirikul & Meknavin, 1998) and Thai person name recognition using Likelihood Probability (Saetiew, Achalakul & Prom-on, 2017). Suwanno, Suzuki and Yamazaki (2007) presented that the combination of word, semantic concept, and orthographic yielded the best performance for Thai NE extraction based on SVM algorithm. Tirasaroj and Aroonmanakun (2009) compared the performance of NER model based on word and character levels by using the CRF model. The experimental result shows that the performance of both are not much different. The efficiency of character segmented is slightly higher because the model cannot recognize words or syllables that never appear in the list such as abbreviation and part of the problem came from not using context clues as a feature.

For deep learning approach in Thai NER, Suriyachay and Sornlertlamvanich (2018) applied POS with word as input of the Bi-LSTM model to recognize and classify the named entity in Thai and the model outperform CRF model. Variational LSTM with CRF also used for That NER and provided a satisfactory result (Udomcharoenchaikit, Vateekul & Boonkwan, 2017).

CHAPTER 3
METHODOLOGY

3.1 Corpus

The corpus used in this research is THAI-NEST corpus. This corpus collected from Thai online news articles which are published on the internet such as political news, foreign news, economic news, crime news, sport news, entertainment news, educational news, and technological news (Theeramunkong et al., 2010). The corpus is disjointedly managed in seven files according to the type of named entity, including Date (DAT), Time (TIM), Measurement (MEA), Name (NAM), Location (LOC), Person (PER), and Organization (ORG). The statistics of each corpus are listed in Table 3.1. Each category is abbreviated by the first three characters. In addition, the original THAI-NEST corpus is also designed and constructed based on the structure of the Orchid corpus as shown in Figure 3.1.

Table 3.1 Number of sentences, words and named entity tags in each file.

	No. of sentences	No. of words	No. of NE tags
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

```
%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
//
```

Figure 3.1 Example of Date corpus

From the figure, there are two types of marker in this corpus. The information line which is a line beginning with the "%" is used to display additional information of the corpus (Description are shown in Table 3.2). The numbering line which is a line beginning with the "#" is used to sequence the line in the corpus as shown in Table 3.3, and also there are three special mark-ups characters as shown in Table 3.4.

Table 3.2 Mark-up for text information line.

Mark-up	Description
%Title:	Title of the corpus
%Description:	Detail of the corpus or reference
%Number of sentences:	Total number of sentences in the file
%Number of words:	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 3.3 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 3.4 Special characters for Mark-up

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

Mark-up	Description
/[NE]	Tag marker for appropriate NE annotation of a word

There are 47 types of POS in the corpus. These POS tags are referenced from Orchid Corpus (Sornlertlamvanich, Charoenporn, Isahara, 1997). Table 3.5 shows each POS that appeared in the corpus.

Table 3.5 All types of part-of-speech in THAI-NEST corpus

No.	POS	Description	No.		Description
1	NPRP	Proper noun	25	RPRE	Preposition
2	NCMN	Cardinal number	26	INT	Interjection
3	NONM	Ordinal number	27	FIXN	Nominal prefix
4	NLBL	Label noun	28	FIXV	Adverbial prefix
5	NCMN	Common noun	29	VACT	Active verb
6	NCMN	Title noun	30	VSTA	Stative verb
7	PPRS	Personal noun	31	VATT	Attributive verb
8	PDMN	Demonstration pronoun	32	ADVN	Adverb with normal form
9	PNTR	Interrogative pronoun	33	ADVI	Adverb with iterative form
10	PREL	Relative Pronoun	34	ADVP	Adverb with prefixed form
11	XVBM	Pre-verb auxiliary, before negator "ไม่"	35	ADVS	Sentential adverb
12	XVAM	Pre-verb auxiliary, after negator "ไม่"	36	JCRG	Coordinating conjunction
13	XVMM	Pre-verb auxiliary, after negator "ไม่"	37	JCMP	Comparative conjunction
14	XVBB	Pre-verb auxiliary, in imperative mood	38	JSBP	Subordinating conjunction
15	XVAE	Post-verb auxiliary	39	CNIT	Unit classifier

No.	POS	Description	No.		Description
16	DDAN	Definite determiner, after noun without classifier in between	40	CLTV	Collective classifier
17	DDAC	Definite determiner, allowing classifier in between	41	CMTR	Measurement classifier
18	DDBQ	Definite determiner, between noun, and classifier or preceding quantitative expression	42	CFQC	Frequency classifier
19	DDAQ	Definite determiner, between noun and classifier or preceding quantitative expression	43	CVBL	Verbal classifier
20	DIAC	Definite determiner, following quantitative expression	44	EAFF	Ending for affirmative sentence
21	DIBQ	Indefinite determiner, between noun and classifier or preceding quantitative expression	45	EITT	Ending for interrogative sentence
22	DIAQ	Indefinite determiner, following quantitative expression	46	NEG	Negator
23	DCNM	Determiner, cardinal number expression	47	PUNC	Punctuation
24	DONM	Determiner, ordinal number expression			

For the format of NE tags, BIO annotation scheme is used for all types of named entity as shown in Table 3.6.

Table 3.6 Format of named entity types in each category

Category	Format	Description	Example
Date	B-DAT	The beginning of date	วันที่ (Date)
	I-DAT	The inside of date	14 กุมภาพันธ์ (February 14)
Location	B-LOC	The beginning of a location name	เมือง (City)
	I-LOC	The inside of a location name	นิวยอร์ค (New York)
Measurement	B-MEA	The beginning of a measurement name	ห้า (Five)
	I-MEA	The inside of a measurement name	เล่ม (Books)
Name	B-NAM	The beginning of any proper name except location, person and organization name e.g. name of competition, position	ศึก (League)
	I-NAM	The inside of any proper name	ลาลีกา (La Liga)
Organization	B-ORG	The beginning of an organization name	บริษัท (Corp.)
	I-ORG	The inside of an organization name	โตโยต้า มอเตอร์ (Toyota Motor)
Person	B-PER	The beginning of a person name	นาย (Mister)
	I-PER	The inside of a person name	ณัฐวุฒิ สะกิดใจ (Natthawut Sakidjai)
Time	B-TIM	The beginning of a time	สิบ (Ten)
	I-TIM	The inside of a time	นาฬิกา (O’clock)

Category	Format	Description	Example
Other	O	The word does not belong to any type of entities	

3.1.1 Corpus Challenges

Unfortunately, The THAI-NEST corpus has some limitations and mistakes that directly affect the prediction process of NER. These defects are as follows:

3.1.1.1 The error of word segmentation

One of the important issues in the corpus is the mistakes in word segmentation. Some characteristics of the Thai language have a profound effect on word segmentation due to this language does not have spaces between words, making it difficult to identify the boundaries of words and may make mistakes in word segmentation. If word segmentation in the corpus is incorrect, it will affect the NER process. Figure 3.2 shows an example of wrong word segmentation.

```
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)

Figure 3.2 Example of mistakes word segmentation in (a) Date and (b) Name corpus

3.1.1.2 The error of named entity tag assignment

Due to incorrect word segmentation, it affects the POS of the word. Moreover, the error of word segmentation and POS causes the NE tag to be incorrectly defined as shown in Figure 3.3.

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

Figure 3.3 Wrong named entity tagging of surname

3.1.1.2 Named entity tagging inconsistency

The inconsistency problem also occurs with named entity tag. Some words can be only one category are labeled as a main category at some places, and labeled as “Other” category at some other places in the same file. Figure 3.4 presents an example of named entity tag inconsistency in the same file.

```
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
```

Figure 3.4 Inconsistency of NE tagging in Location file

3.2 Model

In this study, the NER model is presented for Thai language which was inspired by the research of MA and Hovy (2016). I will describe the details of the model from bottom to top.

The proposed model consists of five important layers as follows: 1) Word Embedding, 2) Character-level Representation, 3) Part of Speech Embedding, 4) Bi-LSTM layer, and 5) CRF layer. The architecture of the model is shown in Figure 3.5.

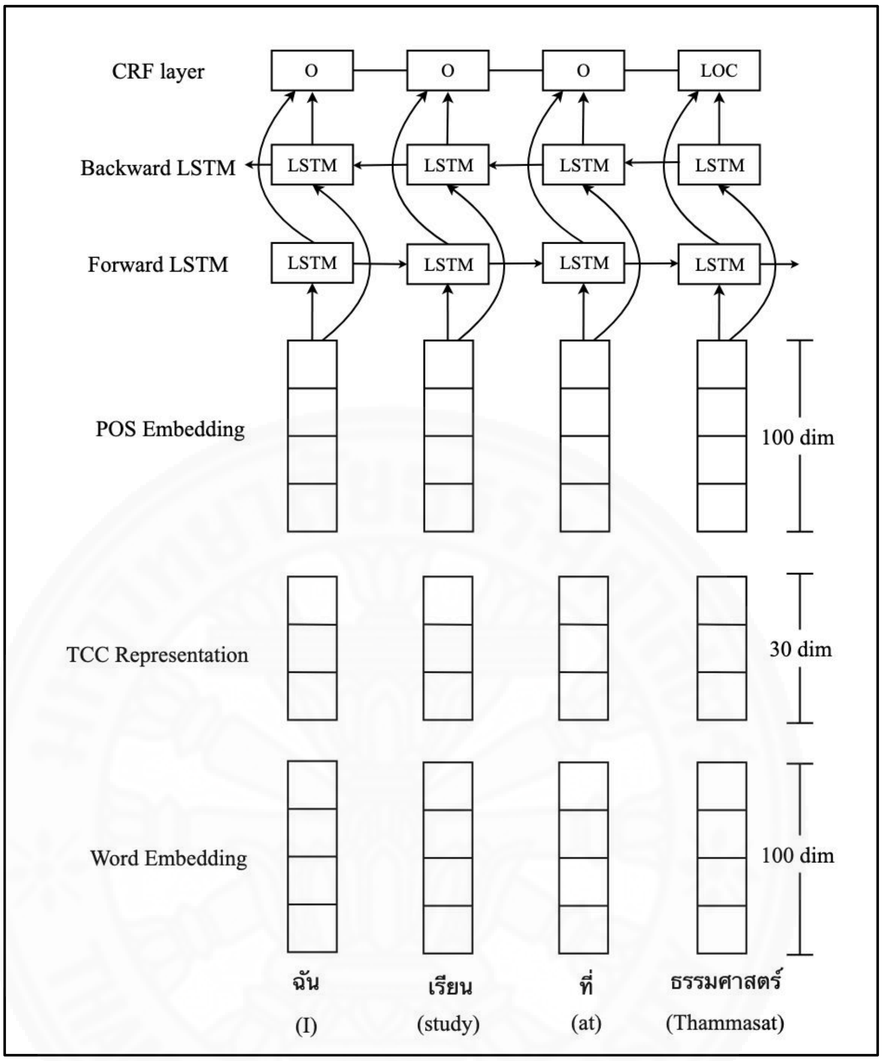

Figure 3.5 The architecture of the proposed NER model

3.2.1 Word embedding

Word Embedding is a type of word representation that allows words with similar meaning to be understood by machine learning. It is a mapping of words into real number vector. The word vector can be calculated from the context around that word. Word embedding could effectively extract semantic and syntactic information among words (Limsopatham & Collier, 2016). In addition, there are many previous studies, e.g. (Mikolov, Chen, Corrado & Dean, 2013), that support this advantage of word embedding. In this experiment, the skip-gram model with 300 dimensions and a window size of three words (three words before and three words after) from the Word2Vec library was used to pre-trained the word embedding.

3.2.2 Thai Character Cluster-level Representation

Character-level representation can extract morphological information from words and very useful especially for languages that have a complex word structure or morphologically rich language, i.e. the Hindi (Maimaiti, Wumaier, Abiderexiti & Yibulayin, 2017), Chinese, Korean, and Thai languages. In Korean, a word usually contains several syllables (Na, Kim, Min & Kim, 2019). Each Korean syllable consists of three parts: a consonant, a vowel and a final consonant (if any) similar to the Thai language, for example, the word “한글” (Korean language). This word can be separated into two syllables “한” and “글”. The first syllable “한” is composed of three characters: the consonant “ㅎ”, the vowel “ㅏ” and the final consonant “ㄴ”. The second syllable “글” also contains three characters: the consonant “ㄱ”, the vowel “ㅡ” and the final consonant “ㄹ”. There are several Korean NER research works that use syllables (Kwon, Ko & Seo, 2018) or morphemes as an input for NER tasks. Therefore, the internal syllables of the word play an important role in modelling. The characteristics of the Thai language are quite similar to Korean, such as the word structure. Therefore, each cluster of the TCCs is rather equivalent to each syllable in Korean and can also preserve more complex information. Hence, I have hypothesized that using only character embedding may not greatly improve the performance of our NER model, so, TCCs can handle this problem and provide better results. The TCC is an unambiguous and inseparable unit that is smaller than a word but larger than a character and cannot be further divided based on Thai language spelling rules to group these characters (Sornlertlamvanich & Tanaka, 1996a and 1996b). For example, a vowel sign and a tone mark cannot stand alone and they must be placed with the character only. This also solves the problem of the tone mark and vowel sign separation. For example, “ป|ระ|เท|ศ|ไท|ย” (Thailand) and “นา|ย|ก|รัฐ|ม|น|ต|รี” (Prime minister). Therefore, I consider that each cluster of TCC is rather equivalent to each character in Chinese and can preserve more complete information. Kim (2016) and Zhang et al. (2016) have proposed CNN model for sentence classification and text classification respectively, allowing me know that CNN has good performance for NLP and character embedding. CNN layer is used to create character cluster representations of the model. The details of the CNN layer are shown in Figure 3.6.

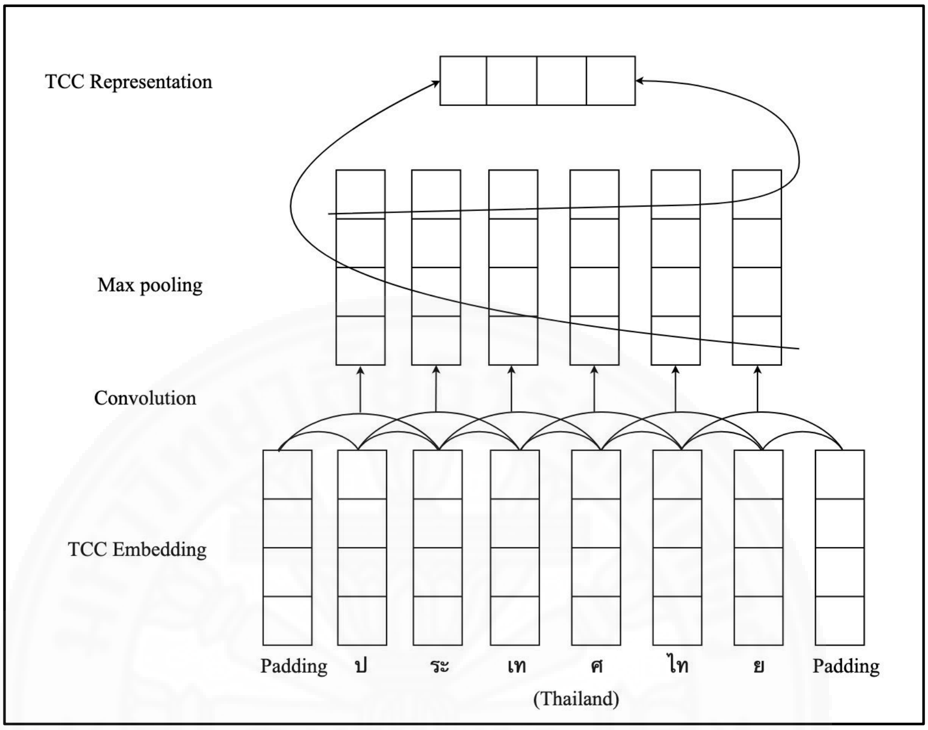

Figure 3.6 The Convolution Neural Network for character-level representation

3.2.3 POS embedding

Part-of-speech (POS) is considered that it can help increase the efficiency of the model because most of Thai named entities will be close to or adjacent to part-of-speech in the type of verb or preposition. In addition, there are many previous studies supporting the use of the POS, both Indonesian and Chinese. POS of each word is encoded into the one-hot vector format in embedding layer.

3.2.4 Bi-directional LSTM

An RNN has an effective performance in many NLP tasks. LSTM is a one kind of an RNN that has the ability to capture long-term dependencies efficiently and can retrieve rich global information. In addition, the bi-directional approach in forward LSTM and backward LSTM can also capture the context from both left and right-hand sides of the sentence. Therefore, I can effectively utilize the previous states and future states to learn a sequence of words.

3.2.5 Conditional Random Field (CRF)

CRF is a standard model and widely used in NLP for predicting the sequence of labels with the most likely tendency that corresponds to the given label sequence. The CRF model takes advantage of the neighbor tag information and takes into account the previous context in predicting the current tag. For example, I-LOC cannot be followed by I-PER in a sentence. Therefore, CRF is considered that appropriate to be the last layer to predict the named entity tag of each word. This study followed MA and Hovy (2016) to generate our linear-chain CRF layer. The linear CRF will trying to find the highest scoring path through a sequence and gives the best tags and final score as shown in Figure 3.7. The transition matrix values are represented by the arrows.

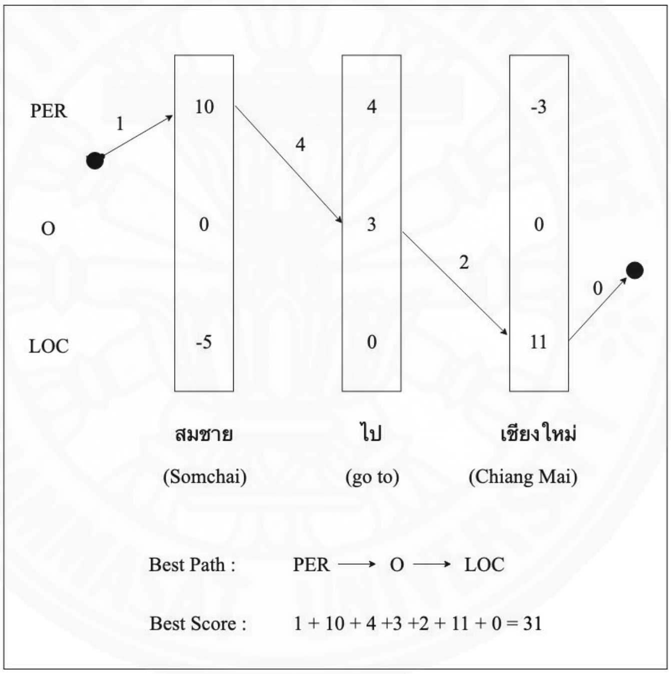

Figure 3.7 Linear-chain CRF decoded

Finally, these layers are combined together to create the Bi-LSTM-CNN-CRF model for predicting named entity tags. For the characters-level representation of each word is calculated by the CNN in Figure 3.6. From each embedding step, it gives vector representations of the words, POS tags, and character clusters. Then, these vectors are concatenated before fed into Bi-LSTM layer. Dropout layers are applied on both the input and output vectors of Bi-LSTM to prevent overfitting and regularize the model. The dropout works by randomly dropping out nodes from the network during training.

Finally, the output vectors of Bi-LSTM layer are passed through the CRF layer and decoded via the Viterbi algorithm (part of CRF layer) to select the most possible sequence of named entity tag. The model learned these vectors to improve the ability to predict target words from the vector of surrounding context.

3.3 Experiment

3.3.1 Pre-processing data

As mentioned above in Section 3.1, the corpus used in this research is the THAI-NEST corpus, in which the form of the named entity tag is the BIO annotation scheme. But unfortunately, due to the written style in Thai is often omitted the identifier or prefix of the named entity, making it impossible to measure the score of the B-tag and I-tag separately. Therefore, the format of the named entity tag has been changed from the BIO to IO to solve this problem. In addition, each word in the corpus is converted to an integer ID is using the key:value pair, also known as the python dictionary, and does the same for character clusters, POS, and named entity tags. Representing these things with integers can save a lot of memory in the training of model.

In this experiment, each corpus file is divided into three parts: 80% of the sentences in the files for the training set, 10% for validation set, and the final 10% for testing set.

3.3.2 Experiment setup

For the proposed NER model, the list of parameters need to be set for training the model. The various parameters are adjusted on the development set to get the most suitable final parameters. Parameters and all were settings are shown in Table 3.7.

Table 3.7 All parameter setting for the model

Parameter	Setting
Char_dim	30
Character-level CNN filters	30
Character-level CNN window size	3
Word_dim	100

Parameter	Setting
Word_LSTM_dim	200
Word_bidirection	TRUE
POS_dim	100
Dropout_rate	0.5
Batch size	10
Learning rate (initial)	0.001
Decay rate	0.05
Gradient clipping	5.0
Learning method	SGD
Training epoch	60

Learning rate is an important hyper-parameter used in the training of neural networks that controls how quickly the model learns or adapts a problem. A large learning rate allows the model to learn faster. That is, it will adapt quickly to the new input data and abandoning or ignoring the previous information. On the other hand, if the learning rate is small, the model learns more slowly and take more time but the advantage is that the model gradually adapts to the new input data and do not abandon things that learned from the previous information. The initial value of the learning rate is set to 0.001.

3.4 Evaluation

To measure the performance on named entity prediction, the model is evaluated in terms of Precision (P), Recall (R), and F-measure (F1-score). Precision is the ratio of the number of correct named entity divided by the total number of named entities that recognized by the model. This can be calculated from the following formula:

$$P = \frac{Correctly\ recognized\ NEs}{All\ recognized\ NEs}$$

Recall is a value that how many of the number of correct named entities that model can be recognize divided by the total number of named entities in the corpus. Recall is calculated as follows:

$$R = \frac{Correctly\ recognized\ NEs}{All\ NEs\ in\ the\ corpus}$$

F-measure or F1-score is the harmonic mean of precision and recall used to measure model capability that can be calculated by the following formula:

$$F = \frac{2 \times P \times R}{P + R}$$

In this experiment, the focus is only on the predictive performance that is related to the main category in each file and ignored the performance of “Other” category.

CHAPTER 4
RESULTS AND DISCUSSION

This chapter will summarize all is findings and describe the problem of recognizing the named entity of the model found in the research. To prove the performance and efficiency of the proposed model, it is compared with the other five baseline models using the same dataset. The learning method of the model is supervised learning, which is the answers are provided to the model in the training dataset. The answer format used is the IO annotation scheme as described above in the section 3.2.5 that it has been changed from BIO annotation to IO.

After the model testing has been completed, the precision, recall, F1-score of each model is shown in Table 4.1.

Table 4.1. Performance of the proposed model and other baseline models

NE	F1-score					
	Bi-LSTM (Word)		Bi-LSTM-CNN-CRF (Word+TCC)		Bi-LSTM-CNN-CRF (Word+POS+Char)	
	No W2V	W2V	No W2V	W2V	No W2V	W2V
DAT	77.51	79.20	82.18	84.65	90.02	92.43
LOC	72.00	75.33	76.90	79.26	84.51	87.07
MEA	69.74	72.41	75.04	77.53	82.94	85.66
NAM	65.29	67.18	71.26	73.84	80.40	82.25
ORG	70.41	73.56	75.63	77.90	82.28	85.79
PER	69.88	74.29	78.57	81.72	83.55	87.32
TIM	79.25	82.77	83.13	86.05	90.16	93.88

NE	F1-score					
	Bi-LSTM (Word+POS)		Bi-LSTM-CNN (Word+POS+TCC)		Bi-LSTM-CNN-CRF (Word+POS+TCC)	
	No W2V	W2V	No W2V	W2V	No W2V	W2V
DAT	84.07	86.14	87.35	89.72	90.77	93.21
LOC	80.22	83.67	83.10	86.24	85.63	88.93
MEA	77.52	80.22	79.85	82.67	83.88	86.52
NAM	72.16	75.48	76.51	80.13	81.29	84.92
ORG	78.54	81.17	81.44	84.75	85.76	87.31
PER	76.08	82.35	80.94	85.07	84.51	88.90
TIM	84.51	88.12	88.59	91.36	91.35	94.76

According to the results shown in the table, the F1-score of Bi-LSTM model that uses only word as input feature is the lowest in every corpus when compared to other models. When the Bi-LSTM model is used word and POS together, the predictive performance of named entity of each corpus is increased significantly. Moreover, the combination of word, POS and TCC in Bi-LSTM-CNN model also provides a higher F1-score. This means that POS and TCC are good features that play an important role in the prediction of the named entity type in the Thai NER task. At last, adding the CRF layer to the last layer in the Bi-LSTM-CNN-CRF (Word+POS+TCC) model enormously improves the model performance and gives the most evaluated F1-score in all corpus when comparing to the other baseline models, especially, the Date file where the F1-score approximately 94 percent and followed by Time file where the F1-score around 93 percent. This indicates that jointly decoding label sequences in the CRF layer is very useful for the final stage of the NER model. A comparison of the performance between the model with TCC and the one with single character is explained in the next section. Additionally, in order to test the importance of pre-trained word embedding, each model is trained without using the pre-trained word embedding. It shows that the F1-score of each model is reduced compared to the one used the pre-trained word embedding.

Next, the examples of the results of each model are explained step by step. The second column in each figure is the named entity tag obtained from the prediction of the model. Therefore, the first step will start with the example of Bi-LSTM model with word as shown in Figure 4.1.

```
จาก/RPRE/O O
การ/FIXN/O O
จับสลาก/VACT/O O
มี/VSTA/O O
ดังนี้/JSBR/O O
ตาก/NPRP/B-LOC O
,/PUNC/O O
กรุงเก่า/NCMN/O O
,/PUNC/O O
ภูเก็ต/NPRP/B-LOC LOC
,/PUNC/O O
อสช./NPRP/O O
ธนบุรี/NPRP/B-LOC LOC
,/PUNC/O O
ฉะเชิงเทรา/NPRP/B-LOC LOC
,/PUNC/O O
กรุงเทพ/NPRP/B-LOC LOC
ตริสเตียน/NPRP/I-LOC O
```

(a)

```
ลงชิงชัย/VACT/O O
ตำแหน่ง/NCMN/O O
นาย/NTTL/O  O
กสมาคม/NCMN/O O
กอล์ฟอาชี/NCMN/O  O
พ/PDMN/O  O
แห่ง/RPRE/O  O
ประเทศไทย์/NPRP/O  O
หรือ/JCRG/O O
สกอ./NPRP/NAM O
ที่จะ/JSBR/O O
มี/VSTA/O O
การ/FIXN/O O
เลือกตั้ง/VACT/O O
นายกฯ/NCMN/O O
```

(b)

Figure 4.1 Predicted named entity types of Bi-LSTM with the word model of (a) location corpus file, and (b) name corpus file

From Figure 4.1(a), this can be seen that the model is not able to correctly predict the words “ตาก”, “กรุงเก่า”, “อสช.”, and “ตริสเตียน”. Actually, the word “ตริสเตียน” is misspelled from the word “คริสเตียน”. For the word “ตาก”, it is difficult for the model to distinguish between “ตาก” that is a noun (the name of a province in Thailand) and “ตาก” which is a verb (to dry in the air). In addition, the word “กรุงเก่า” refers to Ayutthaya, but the POS type of this word is incorrect, in fact, it should be NPRP (proper noun) not NCMN (common noun). For Figure 4.1(b), the name file, “นายกสมาคมกอล์ฟอาชีพแห่งประเทศไทย” (President of the Professional Golf Association of Thailand) has been segmented incorrectly as “นาย-กสมาคม-กอล์ฟอาชี-พ-แห่ง-ประเทศไทย”. Therefore, the model cannot predict the named entity of these words, includes the abbreviation “สกอ.”.

In Figure 4.2, when the POS is added to the Bi-LSTM model with word and POS, the words “ตาก”, “อสช.”, and “สกอ.” are labelled correctly because the POS of each word is considered as well. This means that POS has a high influence in predicting the named entity type of the model and also helps to solve the problem of abbreviation and category ambiguity.

```
จาก/RPRE/O O
การ/FIXN/O O
จับสลาก/VACT/O O
มี/VSTA/O O
ดังนี้/JSBR/O O
ตาก/NPRP/B-LOC LOC
,/PUNC/O O
กรุงเก่า/NCMN/O O
,/PUNC/O O
ภูเก็ต/NPRP/B-LOC LOC
,/PUNC/O O
อสช./NPRP/O LOC
ธนบุรี/NPRP/B-LOC LOC
,/PUNC/O O
ฉะเชิงเทรา/NPRP/B-LOC LOC
,/PUNC/O O
กรุงเทพ/NPRP/B-LOC LOC
คริสเตียน/NPRP/I-LOC O
```

```
ลงชิงชัย/VACT/O O
ตำแหน่ง/NCMN/O O
นาย/NTTL/O  O
กสมาคม/NCMN/O O
กอล์ฟอาชี/NCMN/O  O
พ/PDMN/O  O
แห่ง/RPRE/O  O
ประเทศไทย/NPRP/O  O
หรือ/JCRG/O O
สกอ./NPRP/NAM NAM
ที่จะ/JSBR/O O
มี/VSTA/O O
การ/FIXN/O O
เลือกตั้ง/VACT/O O
นายกฯ/NCMN/O O
```

(a) (b)

Figure 4.2. Predicted named entity types of Bi-LSTM with the word and POS model of (a) location corpus file, and (b) name corpus file

In addition, when considering the results of using word and POS together in the model, it is found that the surrounding context words help to identify the named entity because in the Thai language, named entities are usually adjacent or close to the preposition or verb (e.g. ไป (go), จาก (from), and ที่ (at)) that indicates the location and can sometimes be used to identify the organization as well. Unfortunately, this model is still unable to predict the named entity of the word that is segmented incorrectly.

However, once TCC are applied in the Bi-LSTM-CNN (Word+POS+TCC) model, this model provides better results and is able to handle misspelling and mistake word segmentation problems. The word "ตริสเตียน" is correctly labeled as the location name in Figure 4.3(a). The model can also predict the word นายกสมาคมกอล์ฟอาชีพแห่งประเทศไทย" is more accurate. Although it is not possible to correctly predict the named entity of certain words, which are "พ", "แห่ง", and "ประเทศไทย" as shown in Figure 4.3(b).

```
จาก/RPRE/O O
การ/FIXN/O O
จับสลาก/VACT/O O
มี/VSTA/O O
ดังนี้/JSBR/O O
ตาก/NPRP/B-LOC LOC
,/PUNC/O O
กรุงเก่า/NCMN/O O
,/PUNC/O O
ภูเก็ต/NPRP/B-LOC LOC
,/PUNC/O O
อสช./NPRP/O LOC
ธนบุรี/NPRP/B-LOC LOC
,/PUNC/O O
ฉะเชิงเทรา/NPRP/B-LOC LOC
,/PUNC/O O
กรุงเทพ/NPRP/B-LOC LOC
ตริสเตียน/NPRP/I-LOC LOC
```

```
ลงชิงชัย/VACT/O O
ตำแหน่ง/NCMN/O O
นาย/NTTL/O  NAM
กสมาคม/NCMN/O  NAM
กอล์ฟอาชี/NCMN/O NAM
พ/PDMN/O  O
แห่ง/RPRE/O  O
ประเทศไทย́/NPRP/O  O
หรือ/JCRG/O O
สกอ./NPRP/NAM NAM
ที่จะ/JSBR/O O
มี/VSTA/O O
การ/FIXN/O O
เลือกตั้ง/VACT/O O
นายกฯ/NCMN/O O
```

(a) (b)

Figure 4.3. Predicted named entity types of Bi-LSTM-CNN with the word, POS, and TCC model of (a) location corpus file and, (b) name corpus file

Last of all, as shown in Figure 4.4, Bi-LSTM-CNN-CRF (Word+POS+TCC) model can accurately predict "กรุงเก่า" in the location file and both "แห่ง" and "ประเทศไทย" in the name file because the CRF layer can predict the sequence of the labels with the most possible tendency that corresponds to the sequence of the given input sentence. In addition, Bi-LSTM can effectively capture the sequence of relationships between words in the sentence and CRF will calculate the joint probability distribution and allow the prediction of all the labels in the sentence appropriately, capturing the relationships at label level.

จาก/RPRE/O O การ/FIXN/O O จับสลาก/VACT/O O มี/VSTA/O O ดังนี้/JSBR/O O ตาก/NPRP/B-LOC LOC ,/PUNC/O O กรุงเก่า/NCMN/O LOC ,/PUNC/O O ภูเก็ต/NPRP/B-LOC LOC ,/PUNC/O O อสช./NPRP/O LOC ธนบุรี/NPRP/B-LOC LOC ,/PUNC/O O ฉะเชิงเทรา/NPRP/B-LOC LOC ,/PUNC/O O กรุงเทพ/NPRP/B-LOC LOC คริสเตียน/NPRP/I-LOC LOC	ลงชิงชัย/VACT/O O ตำแหน่ง/NCMN/O O นาย/NTTL/O NAM กสมาคม/NCMN/O NAM กอล์ฟอาชี/NCMN/O NAM พ/PDMN/O NAM แห่ง/RPRE/O NAM ประเทศไทย์/NPRP/O NAM หรือ/JCRG/O O สกอ./NPRP/NAM NAM ที่จะ/JSBR/O O มี/VSTA/O O การ/FIXN/O O เลือกตั้ง/VACT/O O นายกฯ/NCMN/O O
(a)	(b)

Figure 4.4 Predicted named entity type of Bi-LSTM-CNN-CRF with the word, POS, and TCC model of (a) location corpus file, and (b) name corpus file

4.1 TCC vs. Single Character-level Representation

To prove the hypothesis that TCC provides superior results and more efficient than a single character-level, therefore, another experiment is conducted by using the single Thai character instead of TCC on the same data set. The F1-score from each model is shown in Table 4.1. As being shown, all F1-score of the model using TCC is higher than the one that uses only single character.

In Figure 4.5, shows a comparison of the predictions of both models. From the example result, the term “บริษัทอิมแพ็ค แมนเนจเม้นท์ จำกัด” (IMPACT Management Company Limited) with the problem of mistake word segmentation, TCC-based model succeeded in dealing with this problem. On the other hand, the use of single characters cannot correctly predict this word. Due to the model may not be able to learn this word because incorrect word segmentation and take only a single character into account for word embedding is not enough to help the model to make accurate predictions.

```
บริษัท/NCMN/B-ORG ORG
อิมแพ็ค/NCMN/I-ORG ORG
เอ็กซิบิชั่น/NTTL/I-ORG ORG
แมเนจเม้/NCMN/I-ORG O
นท์ จำกัด/NCMN/I-ORG O
ได้/XVAM/O O
เปิด/VACT/O O
ตัว/CNIT/O O
```

(a)

```
บริษัท/NCMN/B-ORG ORG
อิมแพ็ค/NCMN/I-ORG ORG
เอ็กซิบิชั่น/NTTL/I-ORG ORG
แมเนจเม้/NCMN/I-ORG ORG
นท์ จำกัด/NCMN/I-ORG ORG
ได้/XVAM/O O
เปิด/VACT/O O
ตัว/CNIT/O O
```

(b)

Figure 4.5 Prediction results from (a) the model with single character and (b) with TCC

For each corpus file, both models are trained on the GPU server and found that besides the TCC-based model provides higher F1-scores in all corpus, it also consumed less processing time and save memory usage. For example, in the case of Location file, the TCC model took 13 hours for run the model and used 1,689 MB of memory, while the model uses single character took longer processing time, which is 15 hours and used the memory up to 3,371 MB. Moreover, TCC helps reduce the training loss value and performs better than single character. Although the results of both models are not much different but this demonstrates the efficiency and capability of the TCC-level representation over the Character-level representation.

4.2 Problems with the named entity prediction

The main problem encountered in recognizing the named entity of the model is that the model recognizes the wrong named entity type due to the confusion between common nouns and named entities. This case often happens with organization and location names because the most prefixes use with organizations and locations such as มหาวิทยาลัย, องค์กร and บริษัท can be as common nouns and no need to follow by a proper noun every time. For example, in Figure 4.6, it is an error of the model prediction. The proposed model predicts "มหาวิทยาลัยซอฟต์แวร์" as a location. In fact, this word should be predicted to be "Other" category because it is a general noun and not a university name but the word “มหาวิทยาลัยซอฟต์แวร์” refers to a university that offers software discipline instruction. The model predicts this word as a university name since most of the words “มหาวิทยาลัย” (university) is in front of the university name or proper name and the model may highly consider the possible context preposition “จาก” (from), causing the model to be confused and misunderstood.

```
ความรู้/NCMN/O   O
การ/FIXNO     O
พัฒนา/VACT/O    O
ซอฟต์แวร์/NCMN/O     O
มาตรฐานสากล/NCMN/O     O
CMMI/NCMN/O    O
(/PUNC/O
Capability Maturity Model Integration/NCMN/O  O
)/PUNC/O     O
จาก/RPRE/O    O
มหาวิทยาลัย/NCMN/O    LOC
ซอฟต์แวร์/NCMN/O    LOC
ใน/RPRE/O     O
ประเทศสหรัฐอเมริกา NPRP  B-LOC     LOC
```

Figure 4.6 Sample of prediction error

4.3 Combined Corpus

However, another major issue of this corpus is that the corpus is disjointedly managed into seven files based on the named entity type, so the model is not able to learn all types of named entities at once. Therefore, I recognized the importance of this issue and combine all named entity types in one file. As for the method of combining the corpus, the trained model obtained from training the Bi-LSTM-CNN-CRF model of each named entity type is used to train and label the named entity tag on one corpus file until completing all seven types by using cross-tagging technique. The example of the corpus that combines all types of named entities is shown in Figure 4.7.

```
%Title: BKD-1 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: 2,785
%Number of word: 272,749
%Number of named entity tag: 59,596
%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

#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/B-NAM
<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
//
```

Figure 4.7 Sample of combined corpus

Finally, after combining all types of named entities into one file, the number of named entities in each file is greatly increased as shown in Table 8-14.

Table 4.2. Number of each named entity type in the Date corpus file

NE	DAT		
	B-x	I-x	B-x with I-x
DAT	4,523	9,804	3,779
LOC	2,235	1,487	707
MEA	6,000	8,031	4,444
NAM	4,296	4,541	1,597
ORG	5,448	4,072	1,624
PER	3,523	5,157	2,399
TIM	410	397	204

Table 4.3. Number of each named entity type in the Location corpus file

NE	LOC		
	B-x	I-x	B-x with I-x
DAT	5,261	8,519	3,479
LOC	20,527	8,329	4,467
MEA	14,704	19,328	10,993
NAM	10,047	10,315	3,712
ORG	17,731	12,219	5,179
PER	11,585	14,757	7,478
TIM	632	599	360

Table 4.4. Number of each named entity type in the Measurement corpus file

NE	MEA		
	B-x	I-x	B-x with I-x
DAT	1,088	1,820	670
LOC	1,383	1,541	655
MEA	5,244	12,127	4,558
NAM	3,289	2,714	1,183
ORG	2,760	2,057	869
PER	2,975	4,482	2,150
TIM	288	298	139

Table 4.5. Number of each named entity type in the Name corpus file

NE	NAM		
	B-x	I-x	B-x with I-x
DAT	4,530	7,636	3,084
LOC	5,016	3,028	1,457
MEA	13,395	17,632	9,864
NAM	16,759	23,766	9,507
ORG	14,029	10,501	4,334
PER	10,343	13,841	6,778
TIM	539	511	266

Table 4.6. Number of each named entity type in the Organization corpus file

NE	ORG		
	B-x	I-x	B-x with I-x
DAT	11,564	18,614	7,710
LOC	14,030	6,047	3,199
MEA	30,923	40,175	22,733
NAM	26,848	24,942	9,113
ORG	49,397	46,204	20,607
PER	26,614	35,723	17,471
TIM	1,318	1,232	702

Table 4.7. Number of each named entity type in the Person corpus file

NE	PER		
	B-x	I-x	B-x with I-x
DAT	20,215	31,270	12,554
LOC	20,034	14,487	6,760
MEA	52,394	66,552	38,412
NAM	53,391	51,128	19,456
ORG	60,213	47,224	20,535
PER	75,287	146,789	63,508
TIM	3,535	3,532	1,933

Table 4.8. Number of each named entity type in the Time corpus file

NE	TIM		
	B-x	I-x	B-x with I-x
DAT	537	852	305
LOC	459	569	219
MEA	993	1,270	747
NAM	870	936	374
ORG	663	642	268
PER	995	1,519	686
TIM	495	1,048	459

4.4 Conclusion

This research presents the Bi-LSTM-CNN-CRF model with the word, POS and Thai character cluster (TCC) features for Named Entity Recognition in Thai language. The proposed model provides a best and impressive performance in Thai NER task. It can handle both the problem of inconsistency of the named entity tag in same file, misspelling word and especially, the problem of mistake word segmentation. The experimental results show that the key factor influencing the decision of the NE type is the surrounding context such as ไป (go), จาก (from), and ที่ (at), so, POS is one of the important features that helps the model to be more efficient in predicting named entity. Furthermore, TCCs plays an important role in solving the problems related to word segmentation errors, allowing the model to accurately predict the types of NEs even if there is an error of word segmentation in the corpus. In addition, TCCs also improves the effectiveness of creating the word embedding and are useful for morphologically rich languages rather than using only word embedding.

Finally, in the future, the TCC-based model can improve and useful for NER task in other languages, such as Lao, Burmese and Cambodian. Since the characteristics of these languages are similar to Thai which has vowels and tones. Therefore, the use of the TCC principles in the model may also benefit the NER of these languages.

REFERENCES

Charoenpornsawat, P., Kijsirikul, B., & Meknavin, S. (1998). Feature-based Proper Name Identification in Thai. *Proceeding of National Computer Science and Engineering Conference: NCSEC'98*

Chen, X., Xu, L., Liu, Z., Sun, M., & Luan, H. (2015). Joint Learning of Character and Word Embeddings. *Proceeding of the 24th International Joint Conference on Artificial Intelligence (IJCAI 2015)*

Chiu, J. P., & Nichols, E. (2016). Named Entity Recognition with Bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, *4*, 357–370. doi: 10.1162/tacl_a_00104

Chopra, D., Joshi, N., & Mathur, I. (2016). Named Entity Recognition in Hindi Using Hidden Markov Model. *2016 Second International Conference on Computational Intelligence & Communication Technology (CICT)*. doi: 10.1109/cict.2016.121

E, S., & Xiang, Y. (2017). Chinese Named Entity Recognition with Character-Word Mixed Embedding. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management - CIKM 17*. doi: 10.1145/3132847.3133088

Ju, Z., Wang, J., & Zhu, F. (2011). Named Entity Recognition from Biomedical Text Using SVM. *2011 5th International Conference on Bioinformatics and Biomedical Engineering*. doi: 10.1109/icbbe.2011.5779984

Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. doi: 10.3115/v1/d14-1181

Kwon, S., Ko, Y., & Seo, J. (2019). Effective vector representation for the Korean named-entity recognition. *Pattern Recognition Letters*, *117*, 52–57. doi: 10.1016/j.patrec.2018.11.019

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural Architectures for Named Entity Recognition. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. doi: 10.18653/v1/n16-1030

Limsopathan, N., & Collier, N. (2016). Bidirectional LSTM for Named Entity Recognition in Twitter Messages. *Proceedings of the 2nd Workshop on Noisy User-generated Text*

Li, L., Jin, L., Jiang, Z., Song, D., & Huang, D. (2015). Biomedical Named Entity Recognition based on Extended Recurrent Neural Networks. *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. doi: 10.1109/bibm.2015.7359761

Li, L., Jin, L., Jiang, Z., Song, D., & Huang, D. (2015). Biomedical named entity recognition based on extended Recurrent Neural Networks. *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. doi: 10.1109/bibm.2015.7359761

Liu, K., Hu, Q., Liu, J., & Xing, C. (2017). Named Entity Recognition in Chinese Electronic Medical Records Based on CRF. *2017 14th Web Information Systems and Applications Conference (WISA)*. doi: 10.1109/wisa.2017.8

Ma, X., & Hovy, E. (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)*. doi: 10.18653/v1/p16-1101

Maimaiti, M., Wumaier, A., Abiderexiti, K., & Yibulayin, T. (2017). Bidirectional Long Short-Term Memory Network with a Conditional Random Field Layer for Uyghur Part-Of-Speech Tagging. *Information*, *8*(4), 157. doi: 10.3390/info8040157

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. arXiv preprint arXiv:1301.3781.

Na, S.-H., Kim, H., Min, J., & Kim, K. (2019). Improving LSTM CRFs using character-based compositions for Korean named entity recognition. *Computer Speech & Language*, *54*, 106–121. doi: 10.1016/j.csl.2018.09.005

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)*. doi: 10.1109/icacsis.2017.8355038

Saetiew, N., Achalakul, T., & Prom-On, S. (2017). Thai Person Name Recognition (PNR) using Likelihood Probability of Tokenized Words. *2017 International Electrical Engineering Congress (IEECON)*. doi: 10.1109/ieecon.2017.8075816

Salleh, M. S., Asmai, S. A., Basiron, H., & Ahmad, S. (2017). A Malay Named Entity Recognition using Conditional Random Fields. *2017 5th International Conference on Information and Communication Technology (ICoIC7)*. doi: 10.1109/icoict.2017.8074647

Sornlertlamvanich, V., Charoenporn, T., & Isahara, H. (1997). ORCHID: Thai Part-Of-Speech Tagged Corpus: Technical Report. TR-NECTEC-1997-001, National Electronics and Computer Technology Center, Thailand.

Sornlertlamvanich, V., & Tanaka, H. (1997)a. The Automatic Extraction of Open Compounds from Text Corpora. *Proceeding of the 16th International Conference on Computational Linguistics (COLING-96)*. doi: 10.3115/993268.993386.

Sornlertlamvanich, V., & Tanaka, H. (1997)b. Extracting Open Compounds from Text Corpora. *Proceeding of the 2nd Annual Meetings of the Association for Natural Language Processing*

Suriyachay, K., & 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)*. doi: 10.1109/icaicta.2018.8541344

Suwanno, N., Suzuki Y., & Yamazaki H. (2007). Selecting the Optimal Feature Sets for Thai Named Entity Extraction. *Proceedings of ICEE-2007 & PEC*

Theeramunkong, T., Boriboon, M., Haruechaiyasak, C., Kittiphattanabawon, N., Kosawat, K., Onsuwan, C., Siriwat, I., Suwanapong, T., & Tongtep, N. (2010). THAI-NEST: A framework for Thai named entity tagging specification and tools

Tirasaroj, N., & Aroonmanakun, W. (2009). Thai Named Entity Recognition based on Conditional Random Fields. *2009 Eighth International Symposium on Natural Language Processing*. doi: 10.1109/snlp.2009.5340913

Wang, W., Bao, F., & Gao, G. (2016). Mongolian Named Entity Recognition with Bidirectional Recurrent Neural Networks. *2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI)*. doi: 10.1109/ictai.2016.0082

Wang, Y., Xia, B., Liu, Z., Li, Y., & Li, T. (2017). Named Entity Recognition for Chinese Telecommunications Field based on Char2Vec and Bi-LSTMs. *2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*. doi: 10.1109/iske.2017.8258773

Yang, X., & Huang, W. (2018). A Conditional Random Fields Approach to Clinical Name Entity Recognition

Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level Convolutional Networks for Text Classification. *Proceeding of International Conference on Neural Information Processing Systems*

BIOGRAPHY

Name	Ms. Kitiya Suriyachay
Date of Birth	February 07, 1994
Education	2016: Bachelor of Science (Software Engineering) Burapha University

Publications

Suriyachay, K., & Sornlertlamvanich, V. (2018). Named Entity Recognition Modeling for the Thai Language from a Disjointedly Labeled Corpus. *5th International Conference on Advanced Informatics: Concept Theory and Applications (ICAICTA)*, 30-35.

Suriyachay, K., Charoenporn, T., & Sornlertlamvanich, V. (2019). Thai Named Entity Tagged Corpus Annotation Scheme and Self Verification. *9th Language and Technology Conference (LTC 2019)*, 131-137.