IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 28, 2022, accepted May 8, 2022, date of publication May 16, 2022, date of current version May 23, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3175201

Thai Named Entity Recognition Using BiLSTM-CNN-CRF Enhanced by TCC

VIRACH SORNLERTLAMVANICH[1,2] AND SUMETH YUENYONG[3]

[1]Faculty of Data Science, Asia AI Institute, Musashino University, Tokyo 135-8181, Japan
[2]Faculty of Engineering, Thammasat University, Pathumthani 12120, Thailand
[3]Faculty of Engineering, Mahidol University, Nakorn Pathom 73170, Thailand

Corresponding author: Virach Sornlertlamvanich (virach@musashino-u.ac.jp)

This work was supported by the Thailand Science Research and Innovation Fundamental Fund under Contract TUFF19/2564 and Contract TUFF24/2565.

ABSTRACT The languages spoken in Asia share common morphological analysis errors in word segmentation which normally propagate to higher-level processing, i.e., part-of-speech (POS) tagging, syntactic parsing, word extraction, and named entity recognition (NER), as we discuss in this research. We introduce the Thai character cluster (TCC) to reduce the errors propagated from word segmentation and POS tagging by incorporating it into the character representation layer of bidirectional long short-term memory (BiLSTM) for NER. The initial NER model is created from the original THAI-NEST named-entity (NE) tagged corpus by applying the best performing BiLSTM-CNN-CRF model (the combination of BiLSTM, convolutional neural network (CNN), and conditional random field (CRF)) with the word, POS, and TCC embedding. We determine the errors and improve the consistency of the NE annotation through our holdout method by retraining the model with the corrected training set. After the iteration, the overall result of the annotation F1-score has been improved to reach 89.22%, which improves 16.21% from the model trained on the original corpus. The result of our iterative verification is a promising method for low resource language modeling. As a result, The NE silver standard corpus is newly generated for the Thai NER task, called Bangkok Data NE tagged Corpus (BKD). The consistency of annotation is checked and revised according to the improvement of the scope of NE detection by TCC which can recover the errors in word segmentation.

INDEX TERMS Low-resource languages, named entity recognition, Thai language, Thai named entity annotation corpus.

I. INTRODUCTION

This paper proposes a novel method of iterative NE tagging refinement that can be applied to a noisy NE corpus, to generate a silver standard Bangkok Data NE tagged Corpus (BKD) to solve the problem of the limited resources of the Thai language. Some difficulties of language processing in the fundamental issues are also a high barrier to overcome, to improve language processing. The Thai language is an alphabetic language, with no explicit word or sentence boundary, and it is an isolated language, without grammatical markers. These issues can induce a vast amount of ambiguities in morpho-syntactic analysis. Therefore, the consistency problem in word segmentation and grammatical tag annotation is not trivial. The errors are always propagated to consecutive tasks such as word dependency, parse tree annotation, word sense disambiguation, and certainly in the current task of named-entity (NE) annotation. Since word segmentation is not in the scope of this paper, we manually correct the result when necessary and apply the state-of-art word segmentation based on the trigram part-of-speech (POS) tagging model [17]. The POS tagset is introduced from the ORCHID POS tagged corpus [20].

The THAI-NEST corpus [27] is currently the largest Thai NE corpus, collected from 21 Thai online newspaper publishers from January to December 2009. The collection contains a good balance in the variation of the text domain. In the corpus, word segmentation is applied and annotated with the POS tagset. On top of that, the seven types of NE tags, namely, date (DAT), location (LOC), measure (MEA), name (NAM), organization (ORG), person (PER), and time (TIM) are annotated to the corresponding words. The corpus is

The associate editor coordinating the review of this manuscript and approving it for publication was Essam A. Rashed.

manually annotated with one type of NE for a particular file. The size of the corpus is significantly large. It is a corpus of approximately seven million words or about 80 thousand sentences, as shown in Table 1. However, it needs a proper data cleansing process, especially for the word segmentation errors and inconsistency in NE tagging, as we preliminarily conducted a consistency test on the corpus and found that the errors have a significant effect on the accuracy of the named entity recognition (NER) task.

The accuracy of word segmentation has a strong impact on the quality of the corpus. The error normally propagates to higher-level processing in producing the features of word spelling and its POS labeling. To recover the errors, TCC is used instead of a character in many cases. [19] proposed TCC, the smallest standalone character unit according to the spelling rules, to represent the character to reduce the errors in determining the breakable positions in the string. For example, the next breakable position in the string after "กระทรวงการค" is "กระทรวงการคลัง", not "กระทรวงการคล" because the vowel sign " ั" has to be combined with a base consonant like a diacritical sign. "ลัง" is called a character cluster or Thai Character Cluster (TCC). TCC can be defined by a set of spelling rules. There is no ambiguity in forming a cluster, therefore, there is no error in clustering and it can provide a better context to represent a character-level feature of a word.

To improve the performance of NER, especially for the non-segmented language such as Thai, we found that utilizing the advantages of TCC representation instead of a character in the character embedding layer of bidirectional long short-term memory (BiLSTM) can mitigate the NER errors according to the inaccurate word segmentation results. The approach is also viable for other non-segmented languages having similar character composition clues such as Lao, Myanmar, and Cambodian. Though it is out of the scope of this research, the larger unit of character composition can reduce the perplexity at the character representation level.

In this paper, we propose an efficient method to clean up a noisy corpus with language difficulties by state-of-the-art NER using the combination of BiLSTM, convolutional neural network (CNN), and conditional random field (CRF) (BiLSTM-CNN-CRF) [13]. Our novel approach of applying the Thai character cluster (TCC) proposed by [18] for character-level representation in the character-embedding layer performs better in BiLSTM-CNN-CRF with POS and word embedding [22], [23].

The main contribution of this research is to overcome the problem of the shortage and the quality of annotated corpus for model training and evaluation. The existing corpora, though there are not many, still have a big problem in the consistency of word segmentation and annotation. Using a noisy corpus for training certainly cannot expect a high precision out of the model. With the limitations of the availability of the NE annotated corpus, we propose an efficient method to refine the existing corpus though it is full of errors because developing a new large corpus is labor-intensive and costly. We refine the noisy corpus of THAI-NEST automatically to construct a so-called silver standard corpus (automatically constructed corpus with comparable quality to the gold standard corpus) [7] for the NER study.

The paper is structured as follows. Section II provides the grounded works on NE dataset development and the proposed tagset for preparing the NE corpus. Section III summarizes the previous works on some effective approaches for the Thai language NER, and our proposed solution. Section IV gives the information of the THAI-NEST corpus which is used in this study. The size and the annotation scheme are elaborated. Section V discusses how CNN-TCC for character-level representation in the character embedding level can capture the NE spelling pattern to enhance the performance of BiLSTM-CNN-CRF for Thai NER. Section VI describes the performance and the comparison results of each model when applied to the same corpora. Section VII proposes a method of iterative NE tagging refinement to improve the existing noisy corpus and analysis results of the detected annotation errors. Section VIII concludes the achievement of our proposed approach of using TCC in character-level representation and applying the iterative refinement method to achieve the BKD silver standard Thai NE corpus.

II. RELATED WORKS

Many types of NE tagsets have been proposed. The types and number of tags are defined according to the groups and the tasks they are used in. The following are some examples of the representative tagsets used in the NER task: question-answering, information extraction, text summarization, and machine translation.

- CONLL-2003, reported in NER shared task dataset [29], is a well-known collection of Reuters 1,393 newswire articles that contains a large portion of sports news. It is annotated with four entity types (person (PER), location (LOC), organization (ORG), and miscellaneous (MISC)).
- MUC-6 [5] is a dataset consisting of newswire articles from the Wall Street Journal annotated with person (PER), location (LOC), organization (ORG), as well as several temporal and numerical entities.
- OntoNotes 5.0 dataset [8] is annotated with 18 fine-grained NE categories. Those categories are PERSON, NORP (Nationalities or religious or political groups), FACILITY, ORGANIZATION, GPE (Countries, cities, states), LOCATION, PRODUCT, EVENT, WORK OF ART, LAW, LANGUAGE, DATE, TIME, PERCENT, MONEY, QUANTITY, ORDINAL, and CARDINAL.
- IJCNLP-08 NERSSEAL shared task tagset is a dataset consisting of 12 fine-grained NE tags.[1] Those tags are person (NEP), designation (NED), organization (NEO), abbreviation (NEA), brand (NEB), title-person (NETP), title-object (NETO), location (NEL), time (NETI), number (NEN), measure (NEM), and terms (NETE).

[1] available at http://ltrc.iiit.ac.in/ner-ssea-08

For the Thai language, there is a THAI-NEST corpus which is word-segmented and annotated with POS and seven types of NE tags, namely, date (DAT), location (LOC), measure (MEA), name (NAM), organization (ORG), person (PER), and time (TIM). The tagset is detailed enough for common tasks but due to Thai language difficulties in word segmentation and POS tagging, it is difficult to find common agreement in the annotation. These morphological errors cause difficulties in higher levels of NE annotation.

[12] has exhaustively surveyed NER research and classified the approaches into (i) Rule-based approach, which does not need annotated data as it relies on hand-crafted rules; (ii) Unsupervised learning approach, which relies on unsupervised algorithms without hand-tagged training examples; (iii) Feature-based supervised learning approach, which relies on supervised learning algorithms with careful feature engineering; (iv) Deep-learning-based approach, which automatically discovers representations needed for the classification and/or detection from raw input in an end-to-end manner. The state-of-the-art NER in the deep-learning-based approach has been proposed by [13], BiLSTM-CNN-CRF. An experiment has been conducted on the CoNLL-2003 corpus, obtaining 91.21% F1 for the NER task.

III. THAI NAMED ENTITY RECOGNITION

Up to the present, several approaches have been applied to the Thai language NER task.

In the rule-based approach, [3] surveyed to show that the NE lexicon and clue words for NE can be used to create a rule set for extracting and annotating the class. For example, province name, person name, or company name usually follow a particular word, such as "จังหวัด" (province), "นาย" (Mr.), and "บริษัท" (company), respectively. In the case of the name without a clue word, the frequency of word co-occurrence is used to give a threshold for selecting the NE. They combined the heuristic rule set and the frequency of word co-occurrence threshold to annotate PER, ORG, and LOC of 200 articles from Kinnaree Magazine and newspapers. The results showed an average precision of 78.8% and an average recall of 66%. This study has shown that it is possible to use clue words to extract and classify the NE. [24] proposed a method to extract Thai personal named entity without relying on word segmentation or POS tagging to avoid the errors of the resulting words and POS tags. Instead, the gazette of 1,487 Thai personal names is created from a 900 news article collection. The variable length of character n-gram of the front and rear contexts of NE were extracted to generate a set of patterns to evaluate the F1-score of the personal name extraction. Though the average result of the F1-score was reported as 91.58% for the context of 7-character, it was not clear how the context character n-gram was trained and matched to the patterns. [30] prepared 15,077 patterns to map the three types of NE tags (DAT, LOC, and PER). Patterns of the NE contexts and clue words are the keys to creating a rule set to extract the NE. The experiment was conducted on a very small set of corpus and the F1-score widely ranged between 68% to 100%.

In the feature-based approach, Winnow [1] was introduced to extract proper nouns from Thai texts [4]. It used the surrounding words and their POS as the features for Winnow to predict the POS of the target unknown word with NPRP (proper noun) as its POS. It is assumed that the POS of the word with NE type of PER, LOC, and ORG is likely to be NPRP. The authors reported that 92.17% of the test set from 5,000 sentences are correctly annotated.

[2] avoided using POS as a feature in extracting NE because of the unreliable result of word segmentation and POS tagging. Instead, a combination of a heuristic rule set with a word co-occurrence approach for detecting NE, and a maximum entropy model of word features from its orthography and the surrounding context was used to extract PER, LOC, and ORG from a political news corpus of 110,000 words. The F1-score of using plus-minus one-word context (87.70%) was higher than plus-minus two words context (79.78%). The comparison results varied according to the type of NE. It was hard to make a conclusion about the suitable features for their approach.

[25] investigated support vector machine (SVM) in selecting the features among word, POS, word concept, and orthography (types of character) for NER. The experiment was conducted on a collection of 500 articles of Thai business news from Krungthep Turakij news site.[2] The combination of word, word concept, and orthography features yielded the best F1-score for all PER, ORG, and LOC evaluations with an average of 86.31%. There is no doubt at all about the word concept feature because normally the class of the concept can make a better contribution than others. However, the paper did not discuss how the word concept was assigned. Word sense disambiguation is not trivial in this evaluation.

[28] proposed the syllable-segmented input rather than the general word-segmented input to avoid word segmentation errors. The experiment was conducted on the BEST2009 corpus[3] by using the CRF approach. The size of the corpus was about 80,000 words for training. The study evaluated the effectiveness of the CRF model based on a syllable-segmented training set against a word-segmented training set. As expected, the average F1-score of the syllable-segmented model is 80.80%, which is higher than the 80.39% of the word-segmented model for the test on PER, ORG, and LOC annotation. The features used in the word-segmented model are word dictionary, keyword list, and word uni-gram and bi-gram, while the features used in the syllable-segmented model are syllable list, and syllable uni-gram and bi-gram.

[9] utilized k-character prefix and suffix of a word in addition to its word n-gram and POS n-gram context to train MIRA [6] for PER, ORG, LOC, and DAT annotation. The results showed that the k-character prefix and suffix played an

[2] http://www.bangkokbiznews.com

[3] Prepared by National Electronics and Computer Technology Center (NECTEC) for the Thai word segmentation algorithm contest in 2009.

important role in the NE annotation task. The overall F1-score was 82.71% when testing on the THAI-NEST corpus.

In the deep-learning-based approach, the Variational BiLSTM with CRF (V-BiLSTM-CRF) provided a variational inference-based dropout technique to regularize the model [31]. The experiment was conducted on the BEST2010 corpus[4] with 5,238 text files (2,924,433 words) for training and 249 text files (227,302 words) for testing. There were twelve types of NE tags annotated. An 83.7% F1-score can be achieved with POS embedding.

Many particular Thai NE characteristics have been raised and studied. The accuracy of word segmentation and POS tagging are still a big barrier to improving the NER performance. The effective features proposed in the studies can be summed up to a list of clue words, character prefix/suffix, POS, syllable, TCC, and character type. These features are introduced in the above three types of approaches. Unfortunately, the reported F1-score are based on various types and sizes of the corpus due to the lack of gold standard corpus. It is not fair to compare the approaches to the reported results. However, we conduct the comparison experiment of some recent approaches with the same THAI-NEST and BKD corpus to show the improvement when applied to the revised corpus, and also to show the contribution of TCC embedding to the model.

In this paper, we propose a novel method to improve the quality of the existing THAI-NEST NE corpus by applying the BiLSTM-CNN-CRF method, iteratively. The model is also enhanced by the TCC embedding scheme for generating the character-level representation. In contrast to word and syllable segmentation, TCC is a string unit defined in-between the unit of word and syllable, which can correctly be used to separate a Thai string according to the spelling rules. Handling a Thai string in the TCC manner is reported to perform better than other string units in many experiments, e.g., Thai word boundary estimation for open compound extraction in [19], and Thai word indexing for information retrieval in [26]. We are also preparing to release a large enough standard Thai NE corpus for future study and evaluation.

IV. THAI NAMED ENTITY CORPUS

The THAI-NEST corpus is a collection of news articles collected from 21 Thai online newspaper publishers from January to December 2009. There are more than 300,000 news articles covering seven major categories of crimes, politics, foreign affairs, sports, education, entertainment, and economy. Table 1 shows the statistics of the THAI-NEST corpus with additional information on the number of characters and TCC. There are more than 7 million words in 83,248 sentences.

The total size of the text collection is large enough for training a model. However, the corpus is manually tagged.

[4]Prepared by National Electronics and Computer Technology Center (NECTEC) for the Thai word segmentation algorithm contest in 2010.

TABLE 1. Size of the original THAI-NEST corpus.

File	NE type	Sentence	Word	Character	TCC
DAT	DAT	2,979	275,668	1,053,364	574,575
LOC	LOC	10,094	719,594	2,783,544	1,510,245
MEA	MEA	2,084	201,650	756,661	419,346
NAM	NAM	8,541	694,922	2,681,076	1,462,526
ORG	ORG	23,584	1,743,558	6,809,757	3,726,951
PER	PER	35,454	3,418,018	13,063,267	7,232,569
TIM	TIM	512	53,795	196,923	110,658
	Total	83,248	7,107,205	27,344,592	15,036,870

Therefore, there are a lot of problems with the tag consistency. The detail is reported in Subsection VII-A. Moreover, a single file is tagged with only one NE type, and the same file is not used to tag with another NE types at all. This means that each type of NE tagging has been exclusively done on each file. This is proper for evaluating the performance of the model for each NE tagging. However, in the end, we need an algorithm to merge the results from each model. Also, the number of NE tags can be increased by conducting cross-tagging among the files. The corpus is archived in seven files, and each file is exclusively tagged by each type of NE.

A. NE ANNOTATION SCHEME

The corpus is tagged by BIO (aka IOB2) formatting, namely, the Begin-Inside-Outside tagging format proposed by [16]. It is the same as IOB formatting which is proposed by [15] except that the B- tag is used at the beginning of every chunk (i.e. all chunks start with the B- tag). Each type of NE tag is fully expressed in the example, but in general, it is occasionally found that the clue words of each expression are omitted, i.e., "date" in the date expression, "university" in the organization expression, or "Mr." in the name expression. This can cause some difficulties in capturing the pattern model of the NE tags. The orthography feature does not help in most cases when common nouns are used to name an organization or person.

In Section V, our approach has shown that CNN-TCC for character-level representation in the character embedding level can capture the NE spelling pattern even though some clue words are omitted, or some parts of the word are wrongly segmented. As a result, the errors of the absence of the clue words and the wrong-segmented words can be recovered by the TCC embedding.

V. TCC BASED THAI NAMED ENTITY RECOGNITION

We apply the BiLSTM, the state-of-the-art NER using BiLSTM-CNN-CRF) [13]. According to our baseline THAI-NEST corpus analysis, we found that word segmentation errors seriously affect the next coming tasks such as POS tagging, NER, and more. The errors are propagated to cause misjudging in further modeling. Some typical errors are discussed in Subsection VII-A2. However, we are going to improve word segmentation in this paper but we are going to see how we can recover such errors in the NER task. It is reported in [19] that TCC is a larger chunk of characters

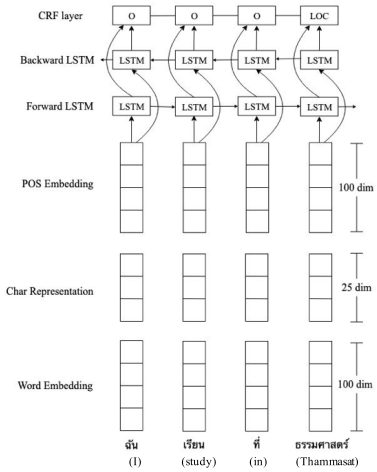

FIGURE 1. BiLSTM-CNN-CRF with POS and CNN-TCC for character-level representation.

TABLE 2. Parameter setting for the model.

Parameter	Setting
Character dimension	25
Character-level CNN filters	30
Character-level CNN window size	3
Word dimension	100
Word LSTM dimension	200
Word bidirection	TRUE
POS dimension	100
Dropout rate	0.5
Batch size	10
Learning rate (initial)	0.01
Decay rate	0.5
Gradient clipping	5.0
Learning method	SGD
Training epoch	60

that can be unambiguously segmented. It also contains more information about word components compared to a single character.

Figure 1 shows the architecture of the full combination of BiLSTM with a word vector from Word2Vec (W2V) for the word-embedding level, CNN encoded TCC for character-level representation, and POS embedding.

The proposed NER model consists of five layers as follows: (i) Word Embedding, (ii) Character-level Representation, (iii) POS Embedding, (iv) BiLSTM layer, and (v) CRF layer. Word, POS, and TCC vectors are concatenated before being fed to the BiLSTM layer. Table 2 shows the hyper-parameters setting for all experiments.

In the word embedding layer, we use GloVe 100-dimensional embeddings trained on the corpus of 5.7 million words as shown in Table 1. For the comparison between character-based and TCC-based performance, 25-dimensional character and TCC embeddings are prepared to extract the character-level representation of words. In the POS embedding layer, we use 47 different POS tags from the ORCHID tagset [20] to extract the 100-dimensional

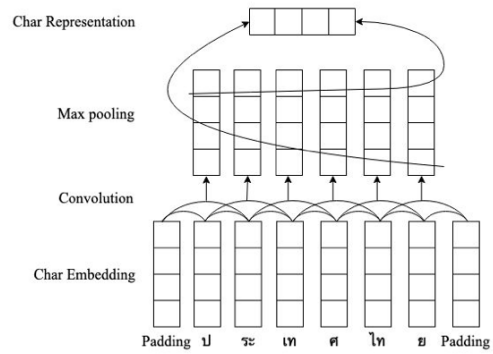

FIGURE 2. CNN encoding TCC for character-level representation of "ประเทศไทย". (Thailand.)

POS embeddings. In addition, to reduce model overfitting, we apply the dropout method [21] to regularize our model.

The output vector from BiLSTM is fed into the CRF layer for sequence labeling. The previous word features and labels are included in the CRF model to predict the current word label. CRF is formally defined in Equations (1a) and (1b). $z = \{z_1, \ldots, z_n\}$ is the input sequence, and z_i is the vector of ith word. $y = \{y_1, \ldots, y_n\}$ is the sequence label for z. $\nu(z)$ is the set of possible sequences for z. $W^T_{y',y}$ and $b_{y',y}$ are the weight and bias vectors corresponding to the label pair (y', y), respectively. CRF is used to determine the weights of different feature functions that maximize the likelihood of the labels in the training data.

$$p(y|z; W, b) = \frac{\prod_{i=1}^{n} \psi_i(y_{i-1}, y_i, z)}{\sum_{y' \in \nu(z)} \prod_{i=1}^{n} \psi_i(y'_{i-1}, y'_i, z)} \tag{1a}$$

where

$$\psi_i(y', y, z) = exp(W^T_{y',y} z_i + b_{y',y}) \tag{1b}$$

In the character-level representation, we apply CNN to encode TCC rather than character because it can present a larger unit of a string. The TCC representation is a non-ambiguously segmentable unit. It is used to capture a larger character pattern to reduce the errors from word segmentation. The result of word segmentation has been improved. Furthermore, to make the character-level representation more meaningful, TCC plays an important role to represent the character unit in a larger pattern. Analogically explaining TCC in the English alphabet, in the case of "th" in the "think" string, it is represented as a unit of "th" rather "t" and "h". Therefore, the representation at the character-level of the word containing "th" can be distinguished from the one containing "t" or "h".

Figure 2 shows the convolution neural network that encodes TCC in the form of character embedding. In the character embedding layer, for the word "ประเทศไทย" (Thailand), TCCs (ป | ระ | เท | ศ | ไท | ย, analogically Th | a | i | l | a | n | d) are fed into CNN rather characters (ป | ร | ะ | เ | ท | ศ | ไ | ท | ย, analogically T | h | a | i | l | a | n | d) in general approaches.

TABLE 3. Comparison of annotation result from CNN encoding character (CNN-CHAR) and TCC (CNN-TCC) for the sentence. (Impact exhibition management Co., Ltd. has launched ...)

No.	BiLSTM POS-CNN-CHAR-CRF	BiLSTM POS-CNN-TCC-CRF	Meaning
1	บริษัท/NCMN/B-ORG	บริษัท/NCMN/B-ORG	company
2	อิมแพ็ค/NCMN/I-ORG	อิมแพ็ค/NCMN/I-ORG	Impact
3	เอ็กซิบิชั่น/NTTL/I-ORG	เอ็กซิบิชั่น/NTTL/I-ORG	Exhibition
4	*แมเนจเม้/NCMN/O	*แมเนจเม้/NCMN/I-ORG	Management
5	*นท์ จำกัด/NCMN/O	*นท์ จำกัด/NCMN/I-ORG	Limited
6	ได้/XVAM/O	ได้/XVAM/O	get
7	เปิด/VACT/O	เปิด/VACT/O	open
8	ตัว/CNIT/O	ตัว/CNIT/O	itself

* Word segmentation error

TABLE 4. Performance of the models on each NE file against the baseline BiLSTM model.

	F1-score							
	BiLSTM		BiLSTM POS		BiLSTM POS-CNN-TCC		BiLSTM POS-CNN-TCC-CRF	
NE Type	Word	W2V	Word	W2V	Word	W2V	Word	W2V
DAT	77.51	79.20	84.07	86.14	87.35	89.72	90.77	**93.21**
LOC	72.00	75.33	80.22	83.67	83.10	86.24	85.63	**88.93**
MEA	69.74	72.41	77.52	80.22	79.85	82.67	83.88	**86.52**
NAM	65.29	67.18	72.16	75.48	76.51	80.13	81.29	**84.92**
ORG	70.41	73.56	78.54	81.17	81.44	84.75	85.76	**87.31**
PER	69.88	74.29	76.08	82.35	80.94	85.07	84.51	**88.90**
TIM	79.25	82.77	84.51	88.12	88.59	91.36	91.35	**94.76**
Average	72.01	74.96	79.01	82.45	82.54	85.71	86.17	**89.22**

The effect of CNN-TCC compared to CNN-CHAR is shown in Table 3. Word nos. 1, 2, and 3 are a pattern of an organization name and all are correctly segmented and POS tagged. Therefore, both CNN-CHAR and CNN-TCC can annotate the correct ORG tag. The problem occurs when the string is an unregistered word for word segmentation and POS tagging. Word nos. 4 and 5 are a part of the name of the organization but they are wrongly segmented. The character embedding by CNN-CHAR is not enough to capture the pattern of the word orthography compared to CNN-TCC. The NE for the word with word segmentation error can then be correctly annotated by the CNN-TCC, as shown in the correct ORG annotation in the word nos. 4 and 5. Therefore, CNN-TCC is tolerant of the inputs with word segmentation errors.

We investigate the effect of W2V vector embedding compared to the original word in the word embedding layer. The experiments are conducted under the same environments of the combination of BiLSTM, with POS, with CNN-TCC, and with CRF. The corpus is randomly divided into 80% for training and 20% for testing.

Table 4 shows the best F1-score of 89.22% in the total evaluation when applying the full combination of BiLSTM-POS-CNN-TCC-CRF with the W2V vector for word representation. Adding the features of POS, CNN-TCC, and CRF improves the F1-score in all types of NE tags. Compared to baseline (BiLSTM), the average F1-score is improved 17.21% from 72.01%, while compared to the performance of the model generated from the original corpus (before cleaning), the average F1-score is improved by 16.21% from 73.01% as shown in Table 5.

TABLE 5. Performance of our best F1-score model trained on the refined corpus (BKD) against the original corpus. (THAI-NEST.)

	F1-score on BiLSTM-POS-CNN-TCC-CRF with W2V embedding	
NE Type	Trained on original corpus (THAI-NEST)	Trained on refined corpus (BKD)
DAT	85.04	**93.21**
LOC	69.27	**88.93**
MEA	77.58	**86.52**
NAM	42.76	**84.92**
ORG	70.19	**87.31**
PER	81.71	**88.90**
TIM	84.53	**94.76**
Average	73.01	**89.22**

TABLE 6. Evidence for additional feature-wise improvement for the sentence (...to compete for the position of Chairperson of the Professional Golf Association of Thailand or PGA which will have a Chairperson election...).

No.	BiLSTM	BiLSTM POS	BiLSTM POS-CNN-TCC	BiLSTM POS-CNN-TCC-CRF
1	ลงชิงชัย/VACT/O	ลงชิงชัย/VACT/O	ลงชิงชัย/VACT/O	ลงชิงชัย/VACT/O
2	ตำแหน่ง/NCMN/O	ตำแหน่ง/NCMN/O	ตำแหน่ง/NCMN/O	ตำแหน่ง/NCMN/O
3	*นาย/NTTL/O	*นาย/NTTL/O	*นาย/NTTL/B-NAM	*นาย/NTTL/B-NAM
4	*กสมาคม/NCMN/O	*กสมาคม/NCMN/O	*กสมาคม/NCMN/I-NAM	*กสมาคม/NCMN/I-NAM
5	*กอล์ฟอาชี/NCMN/O	*กอล์ฟอาชี/NCMN/O	*กอล์ฟอาชี/NCMN/I-NAM	*กอล์ฟอาชี/NCMN/I-NAM
6	*พ/PDMN/O	*พ/PDMN/O	*พ/PDMN/O	*พ/PDMN/I-NAM
7	แห่ง/RPRE/O	แห่ง/RPRE/O	แห่ง/RPRE/O	แห่ง/RPRE/I-NAM
8	ประเทศไทย/NPRP/O	ประเทศไทย/NPRP/O	ประเทศไทย/NPRP/O	ประเทศไทย/NPRP/I-NAM
9	หรือ/JCRG/O	หรือ/JCRG/O	หรือ/JCRG/O	หรือ/JCRG/O
10	สกอ./NPRP/O	สกอ./NPRP/B-NAM	สกอ./NPRP/B-NAM	สกอ./NPRP/B-NAM
11	ที่จะ/JSBR/O	ที่จะ/JSBR/O	ที่จะ/JSBR/O	ที่จะ/JSBR/O
12	มี/VSTA/O	มี/VSTA/O	มี/VSTA/O	มี/VSTA/O
13	การ/FIXN/O	การ/FIXN/O	การ/FIXN/O	การ/FIXN/O
14	เลือกตั้ง/VACT/O	เลือกตั้ง/VACT/O	เลือกตั้ง/VACT/O	เลือกตั้ง/VACT/O
15	นายกฯ/NCMN/O	นายกฯ/NCMN/O	นายกฯ/NCMN/O	นายกฯ/NCMN/O

* Word segmentation error

Word meaning: (1) compete (2) position (3) chairman (4) association (5) golf (6) NON (7) of (8) Thailand (9) or (10) PGA (11) which will (12) have (13) NON (14) election (15) chairman

Table 6 shows evidence of F1-score improvement (step by step) when adding a new feature. The baseline BiLSTM cannot annotate any NAM at all. When POS is additionally applied, word no. 10, which is an abbreviation, is correctly annotated. This is because the POS feature shows that NPRP is more likely to be a NAM. CNN-TCC can significantly annotate word nos. 3, 4, 5, and 6 though these words are wrongly segmented. In this context, the word no. 3 must be "นายก" (Chairperson) rather than "นาย" (Mr.) This means that TCC provides better spelling information to the word than the character within the word itself, especially to the word nos. 5 and 6 which are the unregistered words. In the last column, CRF shows the effectiveness of the sequential context in additionally annotating word nos. 7, and 8 because these words frequently occur at the end of NAM phrases. As a result, the consecutive word nos. 3-8 are annotated as a phrase of NAM.

As a result of applying our final model (BiLSTM-POS-CNN-TCC-CRF), the annotation errors between ORG and LOC, NAM and ORG, and PER and common word have been improved. Especially, in the case of Table 6 where the word segmentation errors confuse the results in POS tagging and NE tagging. After adding CNN-TCC to BiLSTM-POS as shown in column 4, the model can recover the error of NE tagging even if the input string is still wrongly segmented. It shows that TCC can successfully represent a more informative unit than a single character.

TABLE 7. Comparison of model performance.

No.	Model	Approach	Corpus	Size	# NE	F1-score
1	[3]	Pattern & rule	Kinnaree Mag. & news.	200 articles	3	71.83%
2	[24]	Pattern & rule	News article	900 articles	1	91.58%
3	[30]	Pattern & rule	News article	N/A	3	84.00%
4	[4]	Winnow	N/A	5,000 sentences	3	92.17%
5	[2]	Maximum entropy	Political news	110,000 words	3	87.70%
6	[25]	SVM	Krungthep Turakij news	500 articles	3	86.31%
7	[28]	CRF	BEST2009	80,000 words	3	80.80%
8	[9]	MIRA	THAI-NEST	7m words	4	82.71%
9	[31]	V-BiLSTM-CRF	BEST2010	3m words	12	83.70%
10	**Ours**	**BiLSTM-CNN-CRF**	**BKD (refined THAI-NEST)**	**7m words**	**7**	**89.22%**

TABLE 8. Model performance on the BKD refined corpus compared to THAI-NEST original corpus.

NE Type	F1-score							
	SVM [25]		CRF [28]		V-BiLSTM-CRF [31]		BiLSTM-CNN-CRF [Ours]	
	THAI-NEST	BKD	THAI-NEST	BKD	THAI-NEST	BKD	THAI-NEST	BKD
DAT	88.66	91.53	**93.30**	93.00	91.18	92.20	85.04	93.21
LOC	73.32	73.57	80.12	82.23	71.87	86.19	69.27	**88.93**
MEA	83.86	84.10	80.12	85.95	73.79	78.59	77.58	**86.52**
NAM	67.50	67.74	74.03	74.27	80.35	81.08	42.76	**84.92**
ORG	73.65	74.21	79.60	79.81	77.70	78.21	70.19	**87.31**
PER	90.28	91.87	95.22	95.47	95.17	**96.22**	81.71	88.90
TIM	91.21	93.87	95.84	**96.09**	89.93	92.73	84.53	94.76
Avg	81.21	82.41	85.46	86.69	82.86	86.46	73.01	**89.22**

VI. MODEL COMPARISON

Table 7 shows the performance of each approach and its evaluation environment. It is hard to compare the performance of our model to the first three models, which have been done by using pattern and rule-based approaches because their rules and corpora cannot be reproduced at all. The corpora used in models 4-6 are personally collected and are not available. Some studies are applied to the relatively small sizes of corpora. Especially, models 2 and 4 are reported with a very high F1-score but the corpora used in the evaluation are very small compared to others.

Compared to models 7-9, which use a similar size of the corpus with a comparable number of NE types, our model significantly outperforms them in terms of F1-score measurement. Model 8 applies the MIRA approach to the same THAI-NEST corpus as our model. We can confirm that our model can improve the F1-score by 6.51%. Lastly, our model shows the improvement of the similar approaches using CRF in model 7 by 8.42%, and the combination of V-BiLSTM and CRF in model 9 by 5.52%. The table shows that the deep-learning-based approaches still can be improved by using a larger and cleaner corpus as well as the appropriate combination of features.

To compare the performance of the recent models in the feature-based and deep-learning-based approaches, we evaluate on the same corpus to see how the BKD refined corpus can improve their performance, and to confirm the contribution of TCC in the character embedding. F1-score has been improved in almost all types of NE when applying to the BKD refined corpus. The detailed sizes of the THAI-NEST original corpus and BKD refined corpus are elaborated in Table 1 and Table 10, respectively. For each type of NE, the corpora are randomly divided into 80% for training and 20% for testing.

On average F1-score, our proposed model outperforms the other three models trained in SVM, CRF and V-BiLSTM-CRF based approaches on the same corpus, as shown in Table 8. The results show that the refined corpus has been improved in terms of the annotation consistency. Furthermore, the results of our approach with the advantages of using TCC to mitigate the word segmentation errors can achieve the highest F1-score on the average measure.

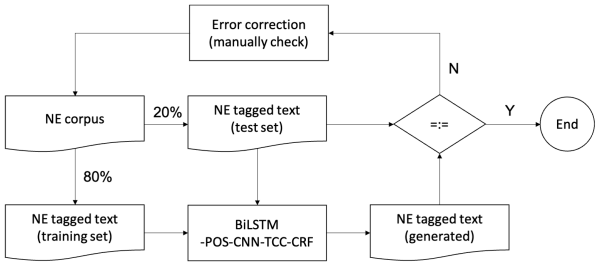

FIGURE 3. Iterative verification NE tagging model.

VII. ITERATIVE NE TAGGING REFINEMENT

Since the original corpus is disjointedly annotated by the NE tags, we train each NE dataset to create a model separately. As a result, we obtain seven models and use them to evaluate the performance, one by one. BiLSTM is our baseline to evaluate any significant improvement of the combination of W2V [14] to encode the word-level representation, CNN [11] to encode TCC in character-level representation, and CRF [10] for including the transition score after the output emission score from BiLSTM. The CRF emission score for each NE tag is used to decide in the case that there is more than one NE tag given to a particular word in the results merging state.

We apply the full combination of BiLSTM-POS-CNN-TCC-CRF with the W2V vector for word representation iteratively to improve the errors in the original corpus. The proposed iterative verification method works as supervised learning in the manner of adjusting the trained model by the corrected training set resulting from the errors detected when comparing the test set and the generated tagged text. It is applied across the seven files of different NE annotations, as shown in Figure 3. We perform a repeated holdout method to find the difference between the test set and the generated tagged text. The corpus is divided into two random disjoint subsets i.e. training set and test set. The model is retrained by the corrected training set until there is no difference, or the error is less than a threshold and becomes steady. Most of the errors described in Subsection VII-A are manually removed by comparing the generated result with the original tagged text.

After a certain cycle of retraining the model, the accuracy has been improved as well as the number of the proper training sets. The total correction of words, POSs, and NE taggings of each file are shown in Table 9, and the total statistics of the refined corpus (BKD) are shown in Table 10. The major errors are from the result of word segmentation (14,527 corrections) which causes the errors in POS (14,693 corrections)

TABLE 9. Result of corpus refinement in number of correction.

File	Word	POS	NE Type
DAT	832	3,581	2,472
LOC	1,242	6,027	905
MEA	962	1,784	2,416
NAM	773	759	1,461
ORG	1,338	1,300	302
PER	9,329	1,210	557
TIM	51	32	8
Total	14,527	14,693	8,121

TABLE 10. Statistics of the refined corpus. (BKD.)

File	Sentence	Word	B-x	I-x
DAT	2,782	272,565	4,133	8,681
LOC	8,584	709,953	21,172	8,324
MEA	1,967	199,305	5,205	11,142
NAM	7,551	199,305	16,911	23,468
ORG	20,397	1,722,063	49,573	45,844
PER	33,231	3,388,885	75,287	146,789
TIM	419	53,330	495	1,048
Total	74,931	7,033,809	172,776	245,296

and NE tagging (8,121 corrections). Correction of word segmentation errors in PER (9,329 corrections) is highly detected because it contains the highest tags (75,287 as shown in Table 10) compared to others, and person names are not normally defined in the dictionary.

A. ERRORS IN THE THAI-NEST CORPUS

The difficulties in the Thai language can cause many problems in the pre-processing step of morphological analysis. Word segmentation and POS tagging have been large issues in Thai language processing. Some errors can be reduced but state-of-art word segmentation and POS tagging have not been able to eliminate the errors.

In terms of word segmentation errors, we generate a list of words for both THAI-NEST and BKD to measure the similarity between them. We calculate the Levenshtein distance between words from the two-word lists and convert them into a similarity score by normalizing the score by the longer word of the pair. The overall mean of the similarity score is 0.9865. The errors can be found in all types of NE when faced the ambiguity of expression. For example, "ทะเล (sea)" and "สาป (curse)" are combined to be the correct word for "ทะเลสาป (lake)". An example of a proper noun is "ศาลเข (unknown)" and "ตเกาลูน (unknown)" are combined to be a correct word of "ศาลเขตเกาลูน (Kowloon District Court)".

By observing the difference between the input and the generated NE tagged text in the process of iterative verification shown in Figure 3, we found the following three types of significant errors in the incorrect and inconsistent annotation.

1) ABBREVIATION TAGGING ERRORS

Abbreviations are not obvious. In many cases they result in the same form as a common word i.e. "กก" which has the meaning of "a reed" or "to embrace", while it is also a shortened form of กิโลกรัม (kilogram), กรรมการ (committee); "บก" which has the meaning of "a land" or "terrestrial", while it is also a shortened form of กองบัญชาการ (headquarter), กรมบัญชีกลาง (Comptroller General's Department). Sometimes, they result in a meaningless string if they are not registered. It is difficult for word segmentation to determine the word boundary and POS of these types of strings. The errors can affect the NE tagging of the surrounding words.

2) WORD SEGMENTATION ERROR

The typical error in word segmentation can be found in the string of "นาย" or "นายก" as shown in the result of นาย/NTTL or นายก/NCMN. This kind of error frequently occurs when the string contains a part of an abbreviation, proper noun, or out-of-vocabulary word (OOV).

3) NE AND POS ANNOTATION ERROR

Most of the cases are from POS tagging errors. Normally, the digits in the date expression must be tagged as a DONM (determiner, ordinal number expression). With inconsistency in POS tagging, digits are sometimes tagged as NCNM (cardinal number) and DCNM (determiner, cardinal number expression). The NE model then cannot capture the pattern of the date expression.

4) ANNOTATION ERROR

It is reported that the corpus is manually revised but some pairs of NE tags can confuse the annotators in making decisions. We found that there are many errors in the confusion cases between PER and common words, ORG and LOC, and NAM and ORG.

The tags of PER, ORG, and LOC are confusing because they are, for one reason, named by using a proper name which is often an OOV, unknown to the word segmentation. For example, the name "วอชิงตัน (Washington)" which is correctly tagged as "วอชิงตัน/NPRP/B-LOC", but it is segmented and tagged as "วอ/NCMN/B-LOC (palanquin)", "ชิง/NCMN/I-LOC (fight for)", and "ตัน/VATT/I-LOC (stuck)". For another reason, they can be found in a compound word or phrase expression, which can be composed of common words. For example, the name of a political party "ภูมิใจไทย (Bhumjaithai)" which is correctly tagged as "ภูมิใจไทย/NPRP/B-ORG", but it is segmented and tagged as "ภูมิใจ/VSTA/B-ORG (proud of)", and "ไทย/NPRP/I-ORG (Thai)". In the case of PER, it is found in many examples, such as "อยู่บำรุง (Yubamrung)" which is correctly tagged as "อยู่บำรุง/NPRP/B-PER", but it is segmented and tagged as "อยู่/XVAE/O (stay)", and "บำรุง/VACT/O (maintain)".

Once the word segmentation and POS tagging are corrected, we do not pass the input string to the word segmentation again because it will produce the same errors, and we do not want to make any changes in word segmentation. Instead, the word segmented with POS tagged string is passed directly

to the NER module, and the new words together with the tags are registered.

The correction in the tags of DAT, MEA, NAM, and TIM is relatively low. The expression of DAT and TIM is quite straightforward with a common format and a closed set of the name of the months. The errors in NAM are similar to the case of PER, ORG, and LOC because of the way of naming the entities.

The errors in MEA are an interesting case for the Thai language because of the expression of the classifier. There is a particular set of classifiers that is designated by the head noun of the noun phrase. For example, "รถยนต์ (car)" always takes a classifier of "คัน (classifier of a car)". However, the difficulty occurs in the case of words having themselves as the classifier. For example, the classifier for "คน (person)" is the same as itself "คน (classifier of a person)". So, the expression of "1 person" is "คน 1 คน". Some examples of errors found in the case of MEA are, "1 วัน (1 day)" is tagged as "1/DCNM/O วัน/NCMN/O (day)" instead of "1/DCNM/B-MEA วัน/CMTR/I-MEA (classifier of a day)", or "90 คน (90 persons)" is tagged as "90/DCNM/O คน/NCMN/O (person)" instead of "90/DCNM/B-MEA คน/CMTR/I-MEA (classifier of a person)".

VIII. CONCLUSION

Since there are many difficulties in the Thai language, it is very cost-intensive to prepare a high-quality consistently-annotated corpus. Our proposed methods are effective in detecting the errors which occur in word segmentation, POS tagging, and the inconsistently NE tagging processes. CNN encoding of TCC for the character-level representation can also recover the word segmentation and POS tagging errors in many cases. Following our proposed NE tagging method (BiLSTM-POS-CNN-TCC-CRF) which can achieve 89.22% in F1-score measure when trained on the refined corpus, compared to 74.96% of the baseline model BiLSTM, 73.01% of the same model but trained on the original noisy corpus, and 87.77% of the same model with CNN-CHAR embedding. Iteratively, the corpus is refined and used to re-trained the model. As a result, the performance of the NE tagging is improved along with having the NE tags in the corpus refined. This means that our iterative NE tagging refinement method is effective in constructing a silver standard NE corpus. The proposed iterative NE tagging refinement method is general. This can benefit corpus development, especially for low-resource languages. The BKD corpus is a result of the refinement of an existing noisy corpus. It is a silver standard NE corpus which is available for NER model training and evaluation. The proposed iterative refinement method works well to correct the inconsistent tags. However, the iteration stops when there is no difference between the test set and the tagged results. Randomization of the test set can somehow help to avoid the local maximum problem. Optimization of the refinement process and automatic tag correction are left for future work.

REFERENCES

[1] A. Blum, "Empirical support for winnow and weighted-majority algorithms: Results on a calendar scheduling domain," *Mach. Learn.*, vol. 26, no. 1, pp. 5–23, 1997.

[2] H. Chanlekha and A. Kawtrakul, "Thai named entity extraction by incorporating maximum entropy model with simple heuristic information," in *Proc. IJCNLP*, 2004, pp. 49–55.

[3] H. Chanlekha, A. Kawtrakul, P. Varasrai, and I. Mulasas, "Statistical and heuristic rule based model for Thai named entity recognition," in *Proc. SNLP-Oriental COCOSDA*, 2002.

[4] P. Charenpornsawat, B. Kijsirikul, and S. Meknavin, "Feature-based proper noun identification in Thai," in *Proc. NCSEC*, 1998.

[5] N. Chinchor and B. Sundheim, "Message understanding conference (MUC) 6," in *Proc. LDC*. Philadelphia, PA, USA: Linguistic Data Consortium, 2003.

[6] K. Crammer, R. McDonald, and F. Pereira, "Scalable large-margin online learning for structured classification," in *Proc. NIPS Workshop Learn. Struct. Outputs*, 2005, pp. 1–7.

[7] Y. Hahm, J. Park, K. Lim, Y. Kim, D. Hwang, and K. S. Choi, "Named entity corpus construction using Wikipedia and DBpedia ontology," in *Proc. LREC*, 2014, pp. 2565–2569.

[8] E. Hovy, M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel, "OntoNotes: The 90% solution," in *Proc. Hum. Lang. Technol. Conf. NAACL*, 2006, pp. 57–60.

[9] C. Kruengkrai, V. Sornlertlamvanich, W. Buranasing, and T. Charoenporn, "Semantic relation extraction from a cultural database," in *Proc. SANLP, COLING*, 2012, pp. 15–24.

[10] J. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. ICML*, vol. 951, 2001, pp. 282–289.

[11] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, 1989.

[12] J. Li, A. Sun, J. Han, and C. Li, "A survey on deep learning for named entity recognition," *CoRR*, vol. abs/1812.09449, pp. 1–20, Dec. 2018.

[13] X. Ma and E. Hovy, "End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF," in *Proc. Annu. Meeting ACL*, 2016, pp. 1064–1074.

[14] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.

[15] L. A. Ramshaw and M. P. Marcus, "Text chunking using transformation-based learning," in *Proc. 3rd Workshop Very Large Corpora*, 1995, pp. 82–94.

[16] T. K. Sang, "Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition," in *Proc. CoNLL*, 2002, pp. 155–158.

[17] V. Sornlertlamvanich, "Word segmentation for Thai in machine translation system," (in Thai), in *Machine Translation*. Bangkok, Thailand: NECTEC, 1993, pp. 50–56.

[18] V. Sornlertlamvanich and H. Tanaka, "The automatic extraction of open compounds from text corpora," in *Proc. COLING*, 1996, pp. 1143–1146.

[19] V. Sornlertlamvanich and H. Tanaka, "Extracting open compounds from text corpora," in *Proc. 2nd Annu. Meetings ANLP*, 1996, pp. 213–216.

[20] V. Sornlertlamvanich, N. Takahashi, and H. Isahara, "Building a Thai part-of-speech tagged corpus (ORCHID)," *J. Acoust. Soc. Jpn. E*, vol. 20, no. 3, pp. 189–198, 1999.

[21] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, Jan. 2014.

[22] K. Suriyachay, T. Charoenporn, and V. Sornlertlamvanich, "Thai named entity tagged corpus annotation scheme and self verification," in *Proc. LTC*, 2019, pp. 131–137.

[23] K. Suriyachay, T. Charoenporn, V. Sornlertlamvanich, and N. Kaothanthong, "Enhancement of character-level representation in bi-LSTM model for Thai NER," *J. Sci. Technol. Asia*, vol. 26, no. 2, pp. 61–78, 2021.

[24] P. Sutheebanjard and W. Premchaiswadi, "Thai personal named entity extraction without using word segmentation or POS tagging," in *Proc. 8th Int. Symp. Natural Lang. Process.*, Oct. 2009, pp. 221–226.

[25] N. Suwanno, Y. Suzuki, and H. Yamazaki, "Selecting the optimal feature sets for Thai named entity extraction," in *Proc. ICEE&PEC*, vol. 5, 2007.

[26] T. Theeramunkong, V. Sornlertlamvanich, T. Tanhermhong, and W. Chinnan, "Character cluster based thai information retrieval," in *Proc. IRAL*, New York, NY, USA, 2000, pp. 75–80.
[27] T. Theeramunkong, M. Boriboon, C. Haruechaiyasak, N. Kittiphattanabawon, K. Kosawat, C. Onsuwan, I. Siriwat, T. Suwanapong, and N. Tongtep, "THAI-NEST: A framework for Thai named entity tagging specification and tools," in *Proc. CILC*, 2010, pp. 895–908.
[28] N. Tirasaroj and W. Aroonmanakun, "Thai named entity recognition based on conditional random fields," in *Proc. 8th Int. Symp. NLP*, Oct. 2009, pp. 216–220.
[29] K. S. Tjong, F. Erik, and F. De Meulder, "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition," in *Proc. 7th Conf. Natural Lang. Learn. HLT-NAACL*, 2003, pp. 142–147.
[30] N. Tongtep and T. Theeramunkong, "Pattern-based extraction of named entities in Thai news documents," *J. Sci. Technol. Asia*, vol. 15, no. 1, pp. 70–81, 2010.
[31] C. Udomcharoenchaikit, P. Vateekul, and P. Boonkwan, "Thai named-entity recognition using variational long short-term memory with conditional random field," in *Proc. iSAI-NLP)*, 2017, pp. 288–298.

VIRACH SORNLERTLAMVANICH received the B.Eng. and M.Eng. degrees in mechanical engineering from Kyoto University, Kyoto, Japan, in 1984 and 1986, and the Ph.D. degree in computer engineering from the Tokyo Institute of Technology, Tokyo, Japan, in 1998. He was a recipient of the "National Distinguished Researcher Award" in Information Technology and Communication from the National Research Council of Thailand, in 2003, followed by the "ASEAN Outstanding Engineering Achievement Award" from the ASEAN Federation of Engineering Organizations (AFEO), in 2011, and was esteemed as "The Researcher of the Year 2001" by the Nation Newspaper, in 2001. He was also recognized with an "Outstanding Alumni Award" by Tokyo Tech Alumni Association (Thailand Chapter), in 2021.

SUMETH YUENYONG received the B.Eng. degree in electrical engineering and the M.Eng. degree in information and communication technology for embedded systems from the Sirindhorn International Institute of Technology (SIIT), in 2004 and 2010, respectively, and the M.Eng. and D.Eng. degrees in communication and integrated systems from the Tokyo Institute of Technology, in 2011 and 2014, respectively. He received the Japanese Government Scholarship (Monbukagakusho) to study in Japan. Currently, he is with the Department of Computer Engineering, Faculty of Engineering, Mahidol University. He received a New Faculty Member grant from the Thailand Research Fund, in 2016. His research interests include signal and image processing, deep learning, and embedded systems.