Segmenting Words in Thai Language Using Minimum Text Units and Conditional Random Field

Kannikar Paripremkul and Ohm Sornil
Graduate School of Applied Statistics, National Institute of Development Administration (NIDA), Thailand
Email: kannikar.par@stu.nida.ac.th, osornil@as.nida.ac.th

***Abstract*—Word segmentation is important to natural language processing tasks. Thai language as well as many Asian languages does not have word delimiter. Word segmentation in Thai language does not only require to focus on dividing a sequence of characters into meaningful words, but the word must also be divided correctly and relevant to the context of a sentence. With the popularity of social media, unknown, informal and slang words are widely used, in addition to words adopted from other languages. Word segmentation methods, generally trained from formal corpuses or dictionaries, do not yield good performance. This research proposes a novel technique to Thai word segmentation where the smallest units constituting words are first extracted, then combined into syllables using Conditional Random Field. Words are then segmented by merging the syllables together with a set of rules learned from language characteristics. The technique is evaluated on both formal and informal datasets against a method based on a convolutional neural network, currently giving the best performance for Thai word segmentation. The results show that the proposed method outperforms the comparing system and gives F-score of 0.9965 and 0.9857 for formal and informal text, respectively.**

***Index Terms*—word segmentation, syllable segmentation, minimum text unit, conditional random field**

I. Introduction

A sentence without word delimiters can be segmented into words in different manners, giving different meanings. In languages with no word delimiter, such as Chinese, Japanese, Korea and Thai, the problem of word boundary ambiguity may lead to incorrect segmentation. For example, "นั่นมือถืออะไร" can be segmented into "นั่น|มือ|ถือ|อะไร" (What is in your hand?) and "นั่น|มือถือ|อะไร" (What brand is that mobile phone?) or "ผ้าไหมลายสวยมาก" can be segmented into "ผ้าไหม|ลาย|สวย|มาก" (This silk has a gorgeous pattern) and "ผ้า|ไห|มลาย|สวย|มาก" (This fabric and jar are destroyed very nicely). The first sentence without preceding or following sentences, the segmentation result can be either "นั่น|มือ|ถือ|อะไร" or "นั่น|มือถือ|อะไร" because lacking of context. In the second sentence, the result should be "ผ้าไหม|ลาย|สวย|มาก". Since "ผ้าไหม" (Silk cloth) is a compound word of "ผ้า" (Fabric) and "ไหม" (Silk), if it is segmented incorrectly, then true meaning of the sentence and words may not be achieved.

The problem today is more difficult with the styles of writing in social media which contains unknown, informal, slangs and words adopted from other languages. These words cannot be found in dictionaries but are understood among social media users while new words are invented in a short time. Word segmentation methods that rely on dictionaries or are trained on formal corpus will not be able to handle these words correctly.

This research proposes a novel technique to Thai word segmentation by extracting Minimum Text Units which are the smallest units that constitute words. These units are then used by Conditional Random Field to identify syllables. Finally, words are segmented from merging syllables together with a set of rules from analyzing language characteristics.

II. Literature Review

This section discusses the problem, techniques and tools previously proposed for word segmentation mainly for Asian languages which have no explicit word boundary delimiter.

A. Word Segmentation

In languages with no clear boundaries between words, word segmentation is considered a necessary step for many text processing tasks such as name tagging, part-of-speech tagging, question answering. In Chinese [1], Word segmentation is also used to investigate clinical note. Words in the notes can be a cue for health speculation. Since these clinical notes contain with many specific words then the existed word segmenter may not working well. To speculation detect on clinical notes, Zhang *et al.* [1] proposed CRF with annotation rules to identify the boundary of cues. Annotation rules contains 31 rules where 16 rules are speculation cues with constraints and another 15 rules are the cues without constraints. The experiment result on twelve systems shows that CRF outperforms all other systems. The best performance method achieves f-score at 92.2%. For word segmentation by Stanford segmenter and CRF segmenter, the performance to handle the clinical notes yields 83.1% and 69.0% for CRF and Stanford segmenter. However,

Manuscript received August 7, 2020; revised February 3, 2021.

doi: 10.12720/jait.12.2.135-141

CRF segmenter was trained on the annotated admission notes when Stanford segmenter was trained on Chinese news.

The study by Xiong *et al.* [2] also investigated Chinese Word Segmentation (CWS) on clinical text. This study compares the performance of two machine learning techniques, CRF and Bi-LSTM with CRF layer. The Electronic Health Record (EHR) system including admission notes and discharge summaries are used for experimenting for word segmentation and POS tagging. The corpus is divided into three parts for train, validation, and test. CCKS2017 is the corpus with Named Entity Recognition (NER). It is also separated into training set and test set. The features for CRF model are unigram, bigram, and trigram. Bi-LSTM-CRF used 50-dimensional embeddings, 10-dimensional embeddings, and 20-dimensional embeddings to Chinese character. Overall, the experiment results showed that CRF score higher than Bi-LSTM-CRF in all experiments. In CWS task, CRF yields 96.94% on f-score when Bi-LSTM-CRF scores at 96.61%. In CWS and POS tagging task, CRF outperforms Bi-LSTM-CRF by 0.14% on CWS and 0.34% on POS tagging.

Tibetan language, one of languages used in China, also has unclear boundary delimiter. Tibetan texts are divided into syllables by a marker called 'tsheg', and words are made by one or more syllables. Lui *et al.* [3] proposed a CRF model and tagged each syllable with position tag to identify its position within each word. For training data, corpus A contained 64,419 sentences and corpus B had 67,484 sentences. These training sets were generated by machine while the test set contained 1,000 Tibetan sentences prepared manually. From the experiments, the quantity of training data had an impact on the performance of word segmentations. The results of training corpus A combined with corpus B, was 95.12% while the results of training corpus A and corpus B were at 93.22% and 93.77%, respectively.

In Korean, research in morpheme segmentation and POS tagging by Na [4] applied CRF technique to the segmentation and tagging. This study separated the proposed method into three process: (1) morpheme segmentation, (2) POS tagging, and (3) Post-processing compound morphemes. For third process, the morpheme in compound unit will be decomposed into atomic compound using pre-analyzed patterns of the compound morphemes or by using lattice HMM. The features of morpheme segmentation consist of three types which are uni-syllable, bi-syllable, and tri-syllable. The BI tag is use for labelling each syllable when 'B' label represents as the beginning of a morpheme and 'I' label represents as the inside of a morpheme. For POS tagging, the feature uses unigram in form of W_{-1}, W_0, W_1 when W is morpheme. In morpheme segmentation result by training on SEJONG and ETRI corpus, SEJONG scores greater than ETRI at 98.41% and 94.65%, respectively, in f-score.

Another non-boundary language, Myanmar, represented by Thet, Na, and Ko [5] proposed word segmentation in two phases. The first phase is syllable segmentation which syllables are formed by rules. Syllable patterns in Myanmar are limited and it is also unambiguous form. The second phase is syllable merging. It is to combine syllable from the first phase into word using dictionary-based statistical approach. First, the input sentences with syllable units are merged into all possible word based on dictionary for each sentence. Normally, the sentence with the least number of words will be selected but sometimes it is bias to follow the longest word from the dictionary. To solve the bias problem, the statistical approach is used to calculate the strength of the sentence. In the evaluation result shows that syllable segmentation is very successful without incorrect segmentation. For word segmentation, the average scores are 99.05%, 98.94%, and 98.99% on recall, precision, and f-score, respectively.

Kudo *et al.* [6] studied morphological analysis in Japanese and compared performance of Hidden Markov Models (HMMs), Maximum Entropy Markov Models (MEMMs) and CRFs. F-score results of Kyoto University corpus in word segmentation of HMMs, MEMMs and CRFs were at 96.22%, 96.44% and 98.96%, respectively. As a result, it confirmed that CRF can solve the problem of word boundary ambiguity. Since CRFs can include related features, while HMMs cannot, without label bias problem.

For Thai word segmentation, Theeramunkong *et al.* [7] presented a concept of Thai Character Cluster (TCC) for retrieving Thai information. The researcher claimed that TCC does not form an ambiguous group of characters. TCC was created from Thai writing rules to segment a sequence of characters into inseparable units that smaller than words, but it cannot be divided further. In Theeramunkong, and Usanavasin [8], TCC are used with a decision tree to deal with unknown word problem and to propose a dictionary-independent method. This study compared the proposed method with a dictionary-based algorithm. As the results, maximum matching algorithm and longest matching algorithm accuracy were 86.21% and 82.60%, respectively which were slightly lower than the proposed method of 87.41%. In conclusion, the researcher suggested applying a dictionary into the method to improve the accuracy of segmenting unknown words.

Aroonmanakun [9] studied Thai word segmentation and found that ambiguity in word segmentation could be solved by inserting a syllable process. The segmentation was then separated into two processes. The first process is syllable segmentation. The process of syllable segmentation applied a trigram model with syllable patterns which there were about 200 patterns. The second process is merging syllables into word units. This step used collocation strength to merge syllables with a dictionary to determine a sequence of syllables. Thus, if unknown words existed in the input sentence, then the segmentation could be incorrect.

Haruechaiyasak *et al.* [10] compare two approaches: dictionary-based and machine learning-based. The results showed that CRF, machine learning based algorithm, achieved the highest score in comparison with other algorithms. The input for algorithm is a character. It was

predicted as the beginning of word or Intra-word. The features used for this study were 10 Thai character patterns. In addition, it was claimed that CRF can handle unknown words and word ambiguity by learning the patterns of word from the ORCHID corpus. The F-score evaluated on 11-gram was 95.38 percentage.

Haruechaiyasak and Kongyoung [11] proposed word segmentation using CRF with three feature sets: character of Thai language, character type and combined features of character and character type. The experiments used InterBEST2009 which contained a larger number of words than ORCHID to ensure that the algorithm could learn the word patterns as much as possible. The model learned to predict each input character as: word-beginning character and intra-word character. This study also contained a post-process to improve the performance of segmentation by included Named Entities (NEs) to merge text as a single word. CRF with the combined feature yields the best performance (F-score) at 93.90% followed by character and character type, respectively.

According to the recent research, Thai study on word segmentation [12] presented the technique to improving the performance due to the problem of compound word ambiguity. There are two sub-processes which are word segmentation and post-processing. CRF model is used for word segmentation applying the features introduced by the previous study [11]. To improve the performance of accuracy, the original feature templates are used and reconsidered from individual to the combination of character and category. The individual feature template is called 'Single' when the combination of character and category are called ‘Combined-1’ and ‘Combined-2’, respectively. The BEST2009 corpus with all five million words is used for this experiment. For post-processing, words in corpus are relabeled and merged to compound word. The merging process using the longest sequence technique followed words in six corpora. For evaluation, training data and test data are split into 80:20. The result of word segmentation shows that the feature template ‘Combined-2’ yields the highest f-score at 0.99% followed by ‘Combined-1’ and ‘Single’ at 0.96% and 0.93%, respectively. However, the accuracy in word merging process was decreased by 0.1%. The merging errors are incorrect chucked word combination and incorrect of word segmentation.

Currently, texts in social network consist of words from foreign languages and new words which do not exist in a dictionary or any previous Thai word corpus. With these texts, dictionary-based methods generally result in low accuracy. To deal with the problems, this research divides segmentation into three parts: Minimum Text Unit (MTU) extraction, syllable identification and word segmentation. MTU and syllable are constructed by Conditional Random Field. Then, words are segmented by using a rule-based longest matching approach.

B. Thai Word Characteristics

Thai characters consisted of 44 consonants, 18 basic vowels (6 combined vowel), 4 tones and others symbol. Thai characters are shown in Table I. Thai text has not explicit stop word or syllable.

TABLE I. THAI CHARACTERS

Consonant	ก ข ฃ ค ฅ ฆ ง จ ฉ ช ซ ฌ ญ ฎ ฏ ฐ ฑ ฒ ณ ด ต ถ ท ธ น บ ป ผ ฝ พ ฟ ภ ม ย ร ล ว ศ ษ ส ห ฬ อ ฮ
Vowel	ั ะ า ๅ ิ ี ึ ื ุ เ แ โ ใ ไ ุ ู ฺ ฤ ฦ
Tone	่ ้ ๊ ๋
Others symbol	ฯ ๆ ์ ๎
Numerals	๐ ๑ ๒ ๓ ๔ ๕ ๖ ๗ ๘ ๙

Many previous approaches studied Thai word segmentation on word level but the problem of ambiguities exists. Aroonmanakun [9] noticed that Thai words normally created from one or more syllables and proposed a method to reduce word ambiguity by first segmenting syllables using 200 patterns of syllable and then putting the syllables into words using mutual information.

Jucksriporn and Sornil [13] proposed the syllable segmentation method to resolve unknown word and ambiguous words. Thai Minimum Clusters (TMCs) is a technique using to solve the previous problems. It creates smaller units than syllable and then combines the unit into syllable using trigram statistical model. TMCs is adjusted from the research’s experience and Thai writing. For the rules, Thai minimum clustering creates a strong sequence of unit such as a word รักษา (Heal) was generated into รัก - ษา since ‘ั’ must be located before and always followed by another character and ‘า’ needs a character ahead. Therefore, this research will modify TMC to suit the pattern rules and use TMCs to identify Minimum Text.

However, the method proposed in [13] is mainly for Thai speech, but this research segments syllable following by writing system. The reason that Thai speech system cannot be used for word segmentation is the method of syllable segmentation. Syllable possibly pronounces more than the character of word. For example, ‘ประวัติศาสตร์’ (history) is pronounced in four syllables as ‘ประ|หวัด|ดิ|ศาสตร์’ (pra-wat-ti-sard). It is not suitable with the objective of research. Instead, the research suggested to separate following by writing system (no characters are added). Then the given word will be segmented less than speech system as ‘ประ|วัติ|ศาสตร์’ with three syllables.

III. PROPOSED METHOD

To cope with the ambiguity in word segmentation, this research divides the word segmentation into three sub-processes: MTU extraction, syllable identification and word construction. An MTU is the smallest unit that constitutes Thai words. For example, ‘เปลี่ยน’ is a word formed by consonants, vowels and tones. In Thai, a vowel and a tone cannot be standalone. A vowel must be combined with at least one consonant, and a tone must be combined with at least one consonant or vowel. So, the MTUs ‘เป’, ‘ลี่’, ‘ย’ and ‘น’ can be extracted from that word.

A syllable is formed by merging MTUs. This research focuses on Thai writing system. Thus, some words that can be divided into three syllables in speech are divided

into two syllables in writing system. For example, 'สนามบิน' (airport) in writing system will be divided as 'สนาม|บิน' but in speech system it will be 'สะ|หนาม|บิน'.

Words are combined by syllables or syllable can be a word by itself depending on a context of sentence. For example, as the previous example, syllable is 'สนาม' (field) and 'บิน' (fly). Using both MTUs and syllables reduces an ambiguity of merging error. Also, both can limit a scope of error for the next step of segmentation method (MTU for syllable segmentation and syllable for word segmentation).

This research proposes a machine learning based algorithms to MTU extraction and syllable identification where unknown word boundaries can be predicted from language structure. Words are then constructed from combining syllables by a rule-based longest matching technique.

A. Conditional Random Field

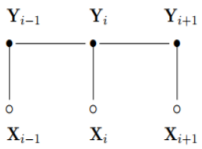

Figure 1. Undirected graph of CRF (Lafferty *et al.*, 2001).

CRF is an undirected graphical model proposed by Lafferty *et al.* [14] (Fig. 1). It uses a global normalization to avoid the label bias problem. A linear-chain CRF is suitable for sequence labeling. The model of CRF can be explained in (1) where X is input sequence and Y is the output label sequence. The conditional probability distribution is P(X|Y) that can be written as follows:

$$P(Y|X) = \frac{1}{Z(x)} \exp\left(\sum_{t=1}^{T}\sum_{k=1}^{K} \lambda_k f_k(y_{t-1}, y_t, x, t)\right) \quad (1)$$

where λ_k is a learned weight of feature function (f) $\lambda_1, \ldots, \lambda_k$ then f_k is a set of feature function $f_1, \ldots, f_k$ when k is the index of feature function, and K is the number of weight index.

In this research, Y is boundary markers, Y = {B, M, E, S}. Example sentence (X) is 'ฉันไม่สบาย' (I am ill). Table II shows an example of sequence label (y_t) and input data (x_t) for syllable segmentation. Feature function (f_k) of syllable segmentation will be described in the next section.

TABLE II. EXAMPLE OF INPUT AND BOUNDARY MARKERS

t	x_t	y_t
0	ฉัน	S
1	ไม่	S
2	ส	B
3	บา	M
4	ย	E

B. Minimum Text Unit Extraction

In this step, MTUs are extracted from input text. Boundaries of MTUs are determined by a CRF model. The features used in this step include two consonant types: consonant (C) and non-suffix consonant (N), a consonant that cannot be placed at the last, six vowels: front vowel (F), special vowel (S), upper vowel (U), rear vowel (B), lower vowel (L) and others vowel (O), tone (T), number (D), space (G) and symbol (Q). A boundary marker is the answer tag to identify boundary of an MTU. There are 4 markers to indicate a border: the beginning token labeled as 'B', middle token labeled as 'M', the ending token labeled as 'E' and the standalone token labeled as 'S'. Example of character features are shown in Table III. The template details for this character sequence are shows in Table IV.

TABLE III. EXAMPLE FEATURES FOR MTU EXTRACTION

Character	ร	ี้	า	น	อ	า	ห	า	ร
Character feature	C	T	B	C	C	B	N	B	C

TABLE IV. FEATURE TEMPLATE FOR MINIMUM TEXT UNIT EXTRACTION

Type	Feature	Description
Unigram	C_n, n = −3, −2, −1, 0, 1, 2, 3	The third previous character, the second previous character, the previous character, the current character, the next character, the second next character, the third next character
Bigram	$C_{-1}C_1$	The previous character and current character

C. Syllable Identification

Syllable identification uses MTUs extracted in the previous step as input to a CRF model. The features are proposed from characteristics of Thai characters which consist of consonant, vowel, tone (T), space (S), number of input character (D), the first character (FC) and last character (LC) of an MTU.

Consonants are categorized into 6 types: single consonant (C), non-suffix consonant (N), combined consonant (CC) which is a consonant that can be combined with the previous consonant and when pronounce it still have one syllable, prefix combined consonant (PC) which is a consonant that can be combined with combined consonant (CC), character as vowel (CV) which is a consonant representing a vowel, and 'Aor' (A) 'อ'. Vowels are categorized into 12 types: leading vowel_1 (LV1) which is leading vowel that cannot be placed by any character in front of it, leading vowel_2 (LV2) which is leading vowel that can be placed by character in front of it, special vowel (SV) which is upper vowel that must be following by a consonant, vowel 'Maiyamok' 'ๆ', vowel 'Garund' 'ุ', vowel 'Maitaikoo' 'ี', vowel 'Paiyarn' 'ๆ', upper vowel (UV), lower vowel (LV), rear vowel_1 (RV1) which is a rear vowel that cannot have any consonant come after, rear vowel_2 which is a rear vowel that can have consonant behind it, and combined vowel (CV) which is any vowel that can combine with other vowels in one syllable. An

example of features is shown in Table V and Table VI. Table VII shows Unigram template and Bigram template used for syllable identification.

TABLE V. MTU WITH CHARACTER FEATURES FOR SYLLABLE SEGMENTATION

	C	N	CC	PC	VC	A	LV1	LV2	SV	UV	LV
ร้า	N	N	R	N	N	N	N	N	N	N	N
น	Y	N	N	N	N	N	N	N	N	N	N
อา	N	N	N	N	N	Y	N	N	N	N	N
หา	N	N	N	N	N	N	N	N	N	N	N
ร	Y	N	R	N	N	N	N	N	N	N	N

TABLE VI. MTU WITH CHARACTER FEATURES FOR SYLLABLE SEGMENTATION (CONTINUED)

	RV1	RV2	M	G	M	P	CV	T	S	FC	LC	D
ร้า	N	Y	N	N	N	N	N	Y	N	ร	า	3
น	N	N	N	N	N	N	N	N	N	N	N	1
อา	N	Y	N	N	N	N	N	N	N	อ	า	2
หา	N	Y	N	N	N	N	N	N	N	ห	า	2
ร	N	N	N	N	N	N	N	N	N	N	N	1

TABLE VII. FEATURE TEMPLATE FOR SYLLABLE IDENTIFICATION

Type	Feature	Description
Unigram	C_n, n = −4, −3, −2, −1, 0, 1, 2, 3, 4	The fourth previous MTU, The third previous MTU, the second previous MTU, the previous MTU, the current MTU, the next MTU, the second next MTU, the third next MTU, the fourth next MTU
Bigram	$C_{-1}C_1$	The previous MTU and current MTU

D. Word Construction

Once syllables are identified, words are constructed from merging nearby syllables together. A combination of longest matching and pattern rules is employed for this task. Pattern rules are constructed from Thai language structure; some are shown in Table VIII. There are 18 rules with 8 types of characters. The rules are found to enhance the accuracy and avoid ambiguities from unknown words, such as, an informal text 'เป็นมาส์ก' (is facial mask). Without the pattern rules, it will be segmented as 'เป็นมา | ส์ก' (occur | no meaning for 'ส์ก'). The pattern rules will group this input as 'เป็น' (is) 'มาส์' (no meaning for 'มาส์') 'ก' (consonant 'ก'). (Fig. 2)

TABLE VIII. EXAMPLE PATTERN RULES OF THAI STRUCTURE

	Pattern Rules
1	< Consonant + Tone >
2	< Consonant + Upper vowel >
3	< Consonant + Upper vowel + Tone >
4	< Front vowel + Consonant >
5	< Front vowel + Consonant + Tone >
6	< Front vowel + Consonant + Upper vowel + Consonant >
7	< Front vowel + Consonant + Upper vowel + Tone + Consonant >

```
charList = null
while s_i is a syllable in a sentence S
  block = s_i-5 s_i-4 ... s_i
  foreach character c in the block
    charList = charList + c
    if charList matches a rule or a word in dictionary and longer
      than current word
       word = charList
       charList = null
    end if
  end for
  i = i + 1
end while
```

Figure 2. Pseudocode for word construction.

Six syllables are processed at a time from left to right. From a sequence of six syllables, one character is considered at a time by checking against rules and then comparing with a self-gathered dictionary which includes official dictionaries, abbreviations and slang words, implemented as a trie, for matching with the longest word in the dictionary. For example, a character sequence 'ธรรมชาติ', the first entry is 'ธรร' (Dharma) which has no exact match found in the dictionary. The next entry is character 'ม' when combined with the previous entry becomes 'ธรรม', and it is found in dictionary. However, when the word 'ธรรม' is combined with the next two entries, it becomes a new word 'ธรรมชาติ' (nature). The previous word is discarded and replaced by the longer one. The process is repeated until the last entry. When the longest sequence of characters is selected as a word, the remaining characters will become the new syllable at the beginning of the next syllable sequence to be processed. The process continues until it reaches the end of the text.

IV. EVALUATIONS

In this section, the proposed method is evaluated using actual data collected from the web and social networks. The training data is part of the BEST2010 corpus [15], a Thai corpus published by National Electronics and Computer Technology Center (NECTEC). It consists of 132,836 characters. The test data are collected from several sources to represent both formal and information texts. Formal texts are collected from news. Informal texts are collected from social media which includes unknown words, specific words, slangs and informal ways of expressing opinions. The two test datasets comprise 10,023 and 13,464 characters, respectively.

This study evaluated performance with standard measures, as follows:

$$\text{F} - \text{score} = \frac{2 \times \text{Recall} \times \text{Precision Recall}}{\text{Recall} + \text{Precision}} \quad (2)$$

$$\text{Recall} = \frac{\text{number of correct tokens}}{\text{number of tokens in test corpus}} \quad (3)$$

$$Precision = \frac{number\ of\ correct\ tokens}{number\ of\ tokens\ in\ system\ output} \quad (4)$$

The baseline model to compare with is a system available on the web which uses Convolutional Neural Network (CNN) to segment words [16].

The results are shown in Table IX and Table X. For formal texts, the proposed method yields the F-scores between 0.9784 to 0.9965 while the baseline method gives F-scores between 0.9428 to 0.9925. For informal texts, the proposed method yields the F-scores between 0.9797 and 0.9857; and the baseline model gives F-scores between 0.9196 and 0.9414. The results are summarized in Fig. 3 and Fig. 4.

TABLE IX. WORD SEGMENTATION PRECISION, RECALL AND F-SCORE OF FORMAL TEXTS

Formal	Algorithm	P(%)	R(%)	F-score
1	Baseline	0.9933	0.9917	0.9925
	Proposed Method	0.9944	0.9981	0.9963
2	Baseline	0.9877	0.9938	0.9907
	Proposed Method	0.9954	0.9977	0.9965
3	Baseline	0.9343	0.9514	0.9428
	Proposed Method	0.9877	0.9922	0.9899
4	Baseline	0.9639	0.9639	0.9639
	Proposed Method	0.9936	0.9968	0.9952
5	Baseline	0.9717	0.9818	0.9767
	Proposed Method	0.9731	0.9837	0.9784

TABLE X. WORD SEGMENTATION PRECISION, RECALL AND F-SCORE OF INFORMAL TEXTS

Informal	Algorithm	P(%)	R(%)	F-score
1	Baseline	0.9275	0.9558	0.9414
	Proposed Method	0.9834	0.9880	0.9857
2	Baseline	0.9159	0.9459	0.9306
	Proposed Method	0.9857	0.9810	0.9833
3	Baseline	0.9230	0.9411	0.9320
	Proposed Method	0.9815	0.9847	0.9831
4	Baseline	0.9005	0.9396	0.9196
	Proposed Method	0.9819	0.9775	0.9797
5	Baseline	0.9149	0.9526	0.9333
	Proposed Method	0.9803	0.9881	0.9842

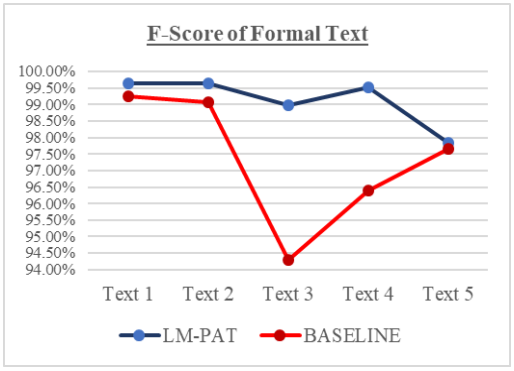

Figure 3. The results of formal texts.

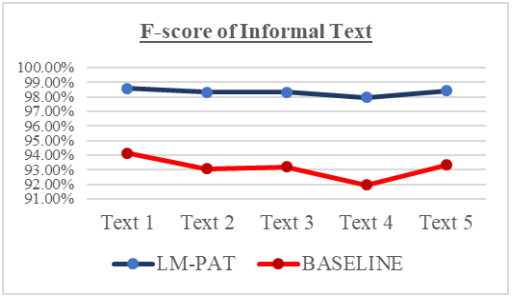

Figure 4. The results of informal texts.

We can see that the proposed method outperforms the baseline model in both types of environments. In precision, the results of formal texts show that the baseline method scored of 0.9933 slightly lower than the proposed method of 0.9954. The results of informal texts show the highest score of the proposed method is 0.9857 when the highest score of the baseline is only 0.9275. Similarly, the recalls of formal texts show that the baseline method scores lower than the proposed method. However, the results of informal texts show that the highest score of the proposed method is 0.9881 while the highest score of the baseline is only 0.9558.

The baseline model often merges two or more words into a single word. For example, 'สั่ง|กาแฟ' (order coffee), the baseline model combines two words into one word 'สั่งกาแฟ'. This segmentation is incorrect because 'สั่ง' (order) and 'กาแฟ' (coffee) cannot be combined into a compound word. Not only two words are combined, there are 'กินคีโตอนุโลม' which should be separated into 'กิน' (eat), 'คีโต' (keto), and 'อนุโลม' (allow). The error of the proposed method mostly occurs as 'รถ (car)|ติดใจ (impress)|กลาง (center)|แมนฮัตตัน (Manhattan)' which every word has its own meaning. Even so, it is incorrect because the context of the sentence incompatible with the segmented words. The expected result of the given sentence should be segmented as 'รถ (car)|ติด (jam)|ใจกลาง (center)|แมนฮัตตัน (Manhattan)'.

Clearly, the proposed technique outperforms the baseline model, especially in informal texts since the baseline model was constructed from a formal corpus. Therefore, the proposed technique is more applicable to segmenting words in Thai language in both formal and informal environments.

V. CONCLUSION

In this paper, we presented a novel technique for Thai word segmentation which is effective in handling formal words found in dictionaries and formal writing, as well as informal words used in social media. Minimum Text Units (MTUs), the smallest unit constituting Thai words are extracted. Characteristics of MTUs are then used to identify syllables. MTU extraction and syllable identification are proposed to reduce the main problems of Thai word segmentation which are word boundary ambiguities and unknown words, and both units are accomplished by using Conditional Random Field.

Finally, syllables are merged into words using rule-based longest matching. The proposed technique is evaluated on both formal and informal datasets against a method based on a convolutional neural network, currently giving the best performance for Thai word segmentation. The results show that the proposed method outperforms the comparing system and gives F-scores of 0.9965 and 0.9875 for formal and informal texts, respectively.

In the future, by exploring more features or techniques to correct an error of word segmentation, the accuracy of word segmentation can be improved. We are also experimenting on Part-of-Speech tag processing for using with our word segmentation. Furthermore, with more available data, deep learning approach for Thai word segmentation can be deployed.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

All authors contributed equally to this work; all authors had approved the final version.

REFERENCES

[1] S. Zhang, T. Kang, X. Zhang, D. Wen, N. Elhadad, and J. Lei, "Speculation detection for Chinese clinical notes: Impacts of word segmentation and embedding models," *Journal of Biomedical Informatics*, vol. 60, pp. 334-341, 2016.

[2] Y. Xiong, Z. Wang, D. Jiang, X. Wang, Q. Chen, H. Xu, J. Yan, and B. Tang, "A fine-grained Chinese word segmentation and part-of-speech tagging corpus for clinical text," *BMC Medical Informatics and Decision Making*, vol. 19, no. 2, pp. 179-184, 2019.

[3] H. Liu, M. Nuo, L. Ma, J. Wu, and Y. He, "Tibetan word segmentation as syllable tagging using conditional random field," in *Proc. 25th Pacific Asia Conference on Language, Information and Computation*, 2011, pp. 168-177.

[4] S. H. Na, "Conditional random fields for Korean morpheme segmentation and POS tagging," *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, vol. 14, no. 3, pp. 1-16, 2015.

[5] T. T. Thet, J. C. Na, and W. K. Ko, "Word segmentation for the Myanmar language," *Journal of Information Science*, vol. 34, no. 5, pp. 688-704, 2008.

[6] T. Kudo, K. Yamamoto, and Y. Matsumoto, "Applying conditional random fields to Japanese morphological analysis," in *Proc. Conference on Empirical Methods in Natural Language Processing*, 2004, pp. 230-237.

[7] T. Theeramunkong, V. Sornlertlamvanich, T. Tanhermhong, and W. Chinnan, "Character cluster based Thai information retrieval," in *Proc. 4th International Workshop on Information Retrieval with Asian Languages*, 2000, pp. 75-80.

[8] T. Theeramunkong and S. Usanavasin, "Non-dictionary-based Thai word segmentation using decision trees," in *Proc. 1st International Conference on Human Language Technology Research*, 2001, pp. 1-5.

[9] W. Aroonmanakun, "Collocation and Thai word segmentation," in *Proc. 5th SNLP & 5th Oriental COCOSDA Workshop*, 2002, pp. 68-75.

[10] C. Haruechaiyasak, S. Kongyoung, and M. Dailey, "A comparative study on Thai word segmentation approaches," in *Proc. 5th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology*, May 2008, vol. 1, pp. 125-128.

[11] C. Haruechaiyasak and S. Kongyoung, "Tlex: Thai lexeme analyser based on the conditional random fields," in *Proc. 8th International Symposium on Natural Language Processing*, 2009.

[12] R. Nararatwong, N. Kertkeidkachorn, N. Cooharojananone, and H. Okada, "Improving Thai word and sentence segmentation using linguistic knowledge," *IEICE Transactions on Information and Systems*, vol. 101, no. 12, pp. 3218-3225, 2018.

[13] C. Jucksriporn and O. Sornil, "A minimum cluster-based trigram statistical model for Thai syllabification," in *Proc. International Conference on Intelligent Text Processing and Computational Linguistics*, Berlin, Heidelberg, February 2011, pp. 493-505.

[14] J. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. 18th International Conf. on Machine Learning*, 2001.

[15] Thailand National Electronics and Computer Technology Center (NECTEC). (2010). Benchmark for Enhancing the Standard of Thai language processing 2010, BEST 2010. [Online]. Available: http://www.hlt.nectec.or.th/best/?q=node/10

[16] R. Kittinaradorn, K. Chaovavanich, T. Achakulvisut, and C. Kaewkasi, *DeepCut: A Thai Word Tokenization Library Using Deep Neural Network*, v1.0, Zenodo, 23 Sept. 2019.

Kannikar Paripremkul received the B.S. degree in computer science from Kasetsart University, Thailand and the M.S. degree in information technology management from Graduate School of Applied Statistics, National Institute of Development Administration (NIDA), Thailand. She is currently studying Ph.D. at NIDA in computer science. Her research interests include natural language processing and machine learning.

Ohm Sornil is an associate professor of computer science at School of Applied Statistics, National Institute of Development Administration, Thailand. He received Ph.D. in Computer Science from Virginia Tech, M.S. in Computer Science from Syracuse University, and B.Eng. (Electrical Engineering) from Kasetsart University. He has been working actively on artificial intelligence, data analytics, and computer security.