MAXIMUM ENTROPY THAI SENTENCE SEGMENTATION COMBINED WITH THAI GRAMMAR RULES CORRECTION

Hongbin WANG[1], Jianxiong WANG[2], Qiang SHEN[3], Yantuan XIAN[4], Yafei ZHANG[5 *]

Sentence segmentation or sentence boundary detection is a basic task of natural language processing research. In Thai language writing, the end of a sentence is often simply represented by a "space" character. However, the "space" character in Thai occurs not only at the end of a sentence but also in other positions. In order to resolve the Thai sentence boundary detection, we propose a maximum entropy Thai sentence segmentation method which integrates the correction of Thai grammar rules. This method combines the syntax rules of Thai sentence boundary recognition into the maximum entropy model which integrates Thai context features and transforms the task of Thai sentence segmentation into the problem of classifying Thai language "space" characters, so as to realize Thai sentence segmentation. We experiment on the ORCHID 1997 Thai corpus, our method's space-correct rate, false-break rate, and recall rate were 94.16%, 1.71%, and 86.16%, respectively. The experimental results show that our method can perform Thai sentence segmentation better than existing models.

Keywords: Thai language; Grammar rules; Context feature; Sentence boundary detection; Maximum entropy.

1. Introduction

Sentence segmentation or sentence boundary detection is a basic task of natural language processing research [1]. Most types of natural language processing, such as machine translation, named entity recognition, sentence similarity calculation, and rapid construction techniques of large corpora, require language input or output as a sentence rather than an entire paragraph. The study of sentence segmentation in natural language processing can be divided into two

[1] Associate Professor, Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, 650500, China

[2] Master, Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, 650500, China

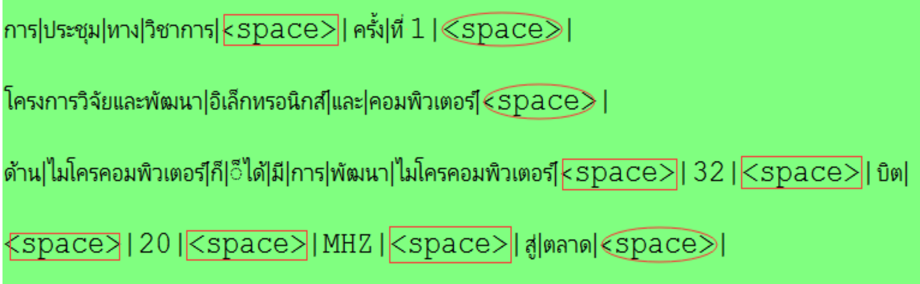
[3] Master, Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, 650500, China

[4] Associate Professor, Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, 650500, China

[5] PhD, Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, 650500, China (*Corresponding author)

areas. One area is the detection of sentence boundaries for languages without sentence ending identification or with weak sentence ending identification, such as Uygur, Tibetan, and Thai. The other area is the elimination of the ambiguity of sentence boundary detection for languages with sentence ending identification, such as Chinese [2-4] and English [5]. The study of sentence segmentation or sentence boundary detection is of equal importance to research in areas such as word segmentation and part-of-speech tagging, and it can bring a great value to the research on follow-up natural language processing.

Thai sentence segmentation research uses computers to automatically divide the Thai language block, paragraph, or chapter into a collection of Thai sentences. In Thai language writing, the "space" character is normally used to represent the ending of a sentence, rather than using a punctuation mark, as in languages like Chinese or English with clear sentence ending identification [6]. Thus, when reading Thai language texts, the reader is often required to do sentence segmentation according to the semantics. This feature has brought great difficulties to machine intelligent sentence segmentation. Additionally, the "space" character in Thai appears not only at the ending of the sentence, but also in the other locations—for example, after Thai numerals or Arabic numerals and between personal honorific words and personal names. Fig. 1 shows a partial example of the location of the "space" character in Thai language. In Fig. 1, the <space> string is a "space" character, the red box is a "space" symbol in the sentence, and the red ellipse is a "space" character at the ending of the sentence.


การ|ประชุม|ทาง|วิชาการ|<space>| ครั้ง|ที่ 1 |<space>|

โครงการวิจัยและพัฒนา|อิเล็กทรอนิกส์|และ|คอมพิวเตอร์|<space> |

ด้าน|ไมโครคอมพิวเตอร์|ก็|ีได้|มี|การ|พัฒนา|ไมโครคอมพิวเตอร์|<space>| 32 |<space>| บิต|

<space>| 20 |<space>| MHZ |<space>| สู่|ตลาด|<space>|

Fig. 1. Positions of "space" Characters in Thai

Therefore, we can transform the segmentation of Thai sentences into the problem of classifying the "space" character in Thai. The "space" characters in Thai are divided into sentence break (sb) "space" character and non-sentence break (nsb) "space" character. The natural language processing technology is used in automatic sentence segmentation of Thai. It is of great significance to the study of lexical analysis, syntactic analysis and machine translation of Thai.

The rest of this paper is organized as follows: Section 2 describes some of the recent related works. Section 3 describes The Maximum Entropy Classification Model. The detailed description of the proposed Maximum Entropy

Thai sentence segmentation method combined with Thai grammar rule correction has been made in Section 4. In Section 5, the results and discussions on the dataset are given. Finally, conclusions are drawn in Section 6.

2. Related Works

The sentence segmentation research started with the early rule-based research method, progressed to the research method based on statistics, and developed to today's research method combining the advantages of both rules and statistics. For example, in 2006, Yu Zhong-hua and his colleagues put forward the method for sentence boundary detection based on context morphological features and teacher-assisted learning, which achieved good experimental performance in the biomedical literature. However, this method relies largely on medical professionals in the training process [7]. In 2007, Chen et al. treated ancient Chinese sentence punctuation as a classification problem and put forward a context-based n-gram model for ancient Chinese punctuation [8]. In 2009, Wang et al. achieved a 96.48% F value in comprehensive experimental performance of ancient Chinese sentence segmentation by using the cascading conditional random field model; however, the applicable field of this model is limited [9]. In 2013, Zhang Zhinan et al. proposed a high accuracy automatic sentence segmentation method based on the combination of forced alignment technology and semi-supervised learning method in the field of speech and achieved good experimental performance [10]. In 2015, Chen Hong et al. selected the candidate sentence segmentation node through statistical features in the online commodity comment text, and then used logical regression to transform the segmentation of a long product review sentence into the classification of the candidate sentence segmentation nodes to achieve the segmentation of the long review sentences and it achieved good experimental results [11].

In the study of Thai sentence segmentation, we can draw on some research results of sentence segmentation for Uygur and Tibetan, which have language characteristics similar to those of Thai. For example, in 2010, Ahsan et al. combined with the characteristics of Uyghur context language and used the maximum entropy model to study Uyghur sentence boundary detection; they achieved a good sentence segmentation effect. They also conducted a further performance optimization study based on Uyghur language rules [12,13]. In 2011, Li et al. used the Tibetan boundary word list and the maximum entropy model to conduct a Tibetan sentence boundary detection experiment combining maximum entropy and rules [14]. In 2012, Cai et al. realized the task of sentence segmentation of Tibetan texts by constructing special rules and Thesaurus related to Tibetan sentence boundary, and then further identifying Tibetan sentences with ambiguous sentence boundary by combining the maximum entropy classification model [15]. In 2013, Ma constructed a practical Tibetan sentence boundary rule library by analyzing Tibetan grammar in detail and achieved an accuracy rate of 96.37% in Tibetan sentence boundary detection [16]. In 2013, Zhao et al. considered the common phenomenon of the Tibetan auxiliary verb as the end of sentences in the modern written Tibetan language, and then constructed the Tibetan auxiliary verb sentence boundary library, Tibetan verb lexicon, and the isomorphic heterogeneous constituent library, achieved excellent performance in Tibetan sentence boundary detection [17]. In 2016, Uliniansyah et al. proposed an Indonesian sentence segmentation method in the study of the Indonesian text-to-speech system [18]. In 2016, Wanjari et al. achieved Marathi sentence boundary detection by constructing Marathi language rules [19]. In 2017, Lengzhi et al. realized Tibetan sentence boundary detection by using the statistical characteristics of the parts of speech of the sentence ending words in Tibetan and achieved a very good accuracy [20].

At present, there are relatively few research results on Thai sentence segmentation. The research methods are mainly divided into rule-based and statistic-based Thai sentence boundary detection methods. The rule-based approach involves constructing rules by considering the linguistic phenomena of the main verbs or conjunctions [21], as well as constructing rules through some fixed Thai sentence-ending language phenomena and non-sentence-ending language phenomena. The statistical method is mainly used to extract the language features in the Thai corpus and to train the specific statistical model to identify Thai sentence boundaries [22,23]. In 2002, Aroonmanakun achieved Thai sentence segmentation with good results through the use of the British and Thai parallel corpus [24]. However, the bilingual parallel corpus is difficult to construct, and it requires bilingual linguistics experts to ensure that the structure of the corpus is consistent. In 2010, Slayden et al. conducted feature-based Thai sentence boundary detection using the Thai word feature in a study of large-scale statistical machine translation systems [25], but they only achieved a certain effect in a single field of the test corpus. In 2013, Tangsirirat et al. studied Thai sentence boundary detection with Thai grammar rules [26], but they only used simple category grammatical features in feature selection. In addition, the construction of Thai grammar rules has relied on professional linguists, and the coverage of the rules is limited. Increasing rules also bring the problems of rule conflict, rule priority division, and slow identification.

Based on the analysis of Thai writing habits and sentence structure, this paper proposes a maximum entropy Thai sentence segmentation method which integrates the correction of Thai grammar rules according to the language phenomenon of Thai sentences.

According to Thai language writing habits, sentence structure and the language phenomenon of Thai sentences, we propose a maximum entropy Thai sentence segmentation method which integrates the correction of Thai grammar rules. This method combines the Thai sentence boundary recognition syntax rules into the maximum entropy model which integrates Thai context features, and transforms the Thai language sentence segmentation task into the problem of classifying Thai language “space” characters, so as to realize Thai sentence segmentation. When the experiments were conducted on the ORCHID 1997 Thai corpus, the space-correct rate, the false-break rate, and the recall rate were 94.16%, 1.71%, and 86.16% respectively, which is a good Thai sentence segmentation result.

3. Maximum Entropy Classification Model

The maximum entropy classification model uses a unified framework to count prior knowledge from different sources. It is a relatively mature mathematical model and widely used in statistical classification problems. The main idea is that it does not provide any subjective assumptions about the unknown situation, and when it makes a prediction of the probability distribution of a random event, the prediction should satisfy all known conditions. This assures the fairest forecast and the most uniform probability distribution. In this case, the entropy of the obtained model is maximized and can satisfy all the constraints, so that the model is “the maximum entropy classification model,” as shown in Formula (1). When the fitness of the model to the known data or the fitness to the unknown data needs to be adjusted, it is only necessary to adjust the constraint conditions flexibly, and the problem of smoothing the parameters in the statistical model can be solved naturally.

$$\max_{p\in P} H\left(Y \mid X\right) = \sum_{(x,y)} p\left(x,y\right) \log \frac{1}{p\left(y \mid x\right)} \tag{1}$$

This model shows that for a given input set X, the probability of output set Y is $P(Y \mid X)$, then select the maximum probability p from the probability set P. $p\left(x,y\right)$ is the probability that the output is y when the input is x. In the actual classification task, all the problem spaces can be represented by feature engineering. Regardless of the complexity of the features described by feature engineering, it is necessary to express the prior knowledge related to the classification task. Each feature in the feature set corresponds to a constraint in the model, and its mathematical essence is a binary function f_i, the characteristic function $f_i(x,y)$, $f_i(x,y)$ is used to describe the relationship between input x and output y. It is defined as:

$$f_i(x,y) = \begin{cases} \text{1, if x and y satisfy some conditions;} \\ \text{0, others.} \end{cases} \tag{2}$$

where $1 \le i \le k$, and k is the number of feature rules in the feature rule set. Then, a probability model that can maximize the entropy value is chosen from the probability distribution models $p \in P$ that satisfy all the constraints. This model is the final maximum entropy classification model, $\hat{p}$ is the probability of entropy value maximum i.e.

$$\hat{p} = \arg\max_{P} H(p) \tag{3}$$

4. Maximum Entropy Thai Sentence Segmentation Method combined with Thai Grammar Rules Correction

Based on the Thai corpus, this paper analyzes the writing habits and sentence structure of Thai. It is found that punctuation is rarely used in Thai texts to break sentences, but only the “space” character is used to indicate the separation in the text. It requires Thai people to judge whether the position of the “space” character needs to break sentences based on semantics. That is to say, the main problem of Thai sentence segmentation is to distinguish the type of “space” character in Thai text. Therefore, we proposed a maximum entropy Thai sentence segmentation method combined with Thai grammar rule correction. We train the maximum entropy classification model by using the “space” character context features in Thai corpus. Then, we use the constructed grammar rules related to Thai sentence boundaries to optimize the maximum entropy classification results and achieved Thai sentence segmentation. The detailed research ideas are shown in Fig. 2.

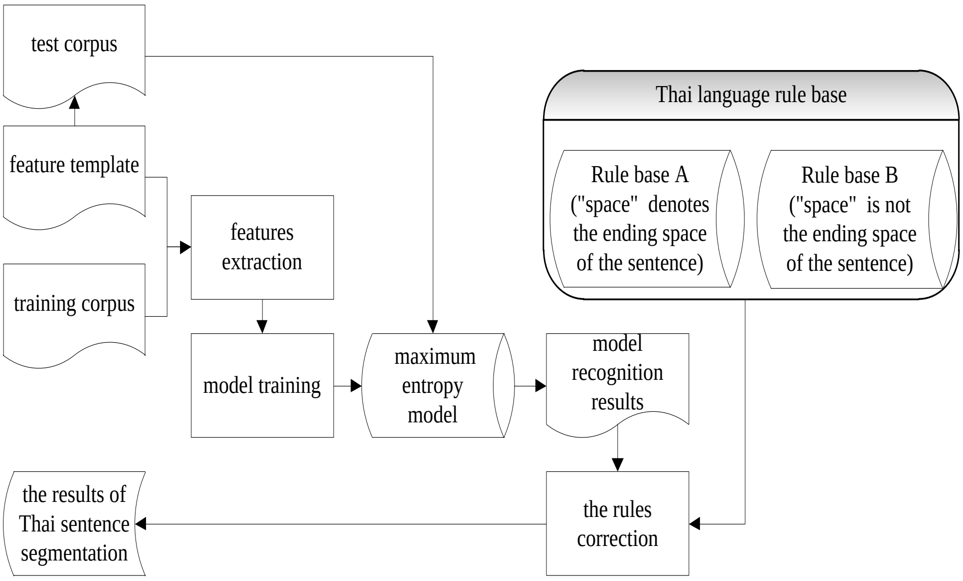

test corpus
feature template
training corpus
features extraction
model training
maximum entropy model
model recognition results
Thai language rule base
Rule base A ("space" denotes the ending space of the sentence)
Rule base B ("space" is not the ending space of the sentence)
the rules correction
the results of Thai sentence segmentation

Fig. 2. Flow Chart of Thai language Sentence Segmentation

4.1 Constructing the Thai Language Rule Base of Thai Sentence Boundaries

Thai is a special language that rarely uses explicit punctuation marks (such as “.”, “?”, “!”) at the end of sentences, instead using the “space” character at the ending of the sentence for separation of most sentences. The Royal College of Thailand also defined some of the rules on use of “space” character in its Thai language writing system research [6], dividing the “space” character into “space” is the ending of sentence (sentence-break, sb) and “space” is not the ending of sentence (non-sentence-break, nsb). We made a statistical analysis of the sentence end combination and the “space” characters in Thai corpus, and identified many special linguistic phenomena of Thai and the fixed Thai combination. So, after summarizing these findings, we constructed the two types of “space” character classification rules (Thai language rule base) that were useful for Thai sentence boundary detection, we used “sb” label to indicate that the “space” character is the ending of the sentence, and used “nsb” label to indicate that the “space” character is not the ending of sentence. The Thai language rule bases are as follows:

Rule base A: the space at the end of sentence rules (sb is short for sentence-break, sb denotes the “space” is the ending space of the sentence.)

(1) The “space” character follows the obvious punctuation at end of the sentence, such as the “space” character follows “.”, “?” and “!”;

(2) The “space” character is located after the specific Thai vocabulary in the interrogative sentence, such as the “space” character follows เหรอ, ไหม, and มั๊ย;

(3) The “space” character follows a specific Thai word in an affirmative sentence, such as the “space” character follows จ๊ะ, จ้ะ, ค่ะ, ครับ, น, นะ, น่า and เถอะ.

Rule base B: space not at the end of sentence rules (nsb is short for non-sentence-break, nsb denotes the “space” is not the ending space of the sentence.)

(1) The “space” character follows the comma character, such as the “space” character follows “,”;

(2) The “space” character before or after the quotes character, such as the “space” character before or after (‘’ or “““);

(3) The “space” character before or after the paired parentheses character, such as “space” character before or after (());

(4) The “space” character before or after the Thai inherent overlap symbol (ๆ);

(5) The “space” character before or after mathematical symbols;

(6) The “space” character before or after Arabic numerals or time;

(7) The “space” character before or after Thai quantifiers (ลักษณะนาม);

(8) The “space” character after a small ellipsis in Thai (ฯ);

(9) The “space” character between Thai titles นาย (Mr.), นาง (Mrs.), นางสาว (Miss) and names.

4.2 Maximum Entropy Modeling

Let us define $B=\{sb,nsb\}$ as the category set of each “space” character in Thai and $C=\{c_1,c_2,\cdots,c_i,\cdots,c_n\}$ as the set of contextual information around each “space” character that can be observed in the Thai language training data set, and C is from Thai language rule base. The set of contextual features around the “space” character is constructed through the binary function $f_j(b,c)$, with $f_j(b,c)$ defined in Formula (2), where $b\in B, c\in C$, $1\le j\le k$. This paper mainly studies three feature rules, so k is 3. The binary function f_j is as follows:

When $j=1$,

$$f_1(b,c)=\begin{cases}1, \text{ if the preceding word of the space character is English(c) and b=rsb;}\\0, \text{ others.}\end{cases} \tag{4}$$

This feature can help us learn the phenomenon that the “space” character after English is usually a non-sentence ending “space” character (non-sentence break space).

When $j=2$,

$$f_i(b,c)=\begin{cases}1, \text{ if the space character is preceded by a quantifier or a number(c) and b=nsb;}\\0, \text{ others.}\end{cases} \tag{5}$$

This feature learns that after Arabic numerals or quantifiers in Thai, there is usually a non-sentence ending “space” character (non-sentence-break space).

When $j=3$,

$$f_i(b,c)=\begin{cases}1, \text{ if the space character is within a pair of punctuation characters(c) and b=nsp;}\\0, \text{ others.}\end{cases} \tag{6}$$

This feature helps to learn the “space” characters in pairs of punctuation such as quotation marks or brackets, there is more likely to be a non-sentence ending “space” character (non-sentence-break space).

Then, the probability $p\left(b|c\right)$ that satisfies the condition of maximum entropy can be expressed as follows:

$$p\left(b|c\right)=\frac{1}{Z_\lambda\left(c\right)}\exp\sum_{j=1}^{k}\lambda_j f_j\left(b,c\right) \tag{7}$$

$$Z_\lambda\left(c\right)=\sum_b\exp\sum_{j=1}^{k}\lambda_j f_j\left(b,c\right) \tag{8}$$

Formula (8) is the normalization factor from Formula (7), and each characteristic function f_j corresponds to a weight value λ_j. Therefore, the purpose of maximum entropy modeling is to find the model parameter weight λ_j with the maximum entropy value in the probability model set satisfying all constraints.

If there are k features, then the binary function corresponding to each feature is f_j, and the constraint of f_j on probability $p(b,c)$ can be expressed as follows:

$$E_p f_j = E_{\tilde{p}} f_j \tag{9}$$

where $E_p f_j$ is the expected value of the feature function f_j when the probability distribution is p, and $E_{\tilde{p}} f_j$ represents the expected value of the empirical probability of feature f_j in the training sample. i.e.

$$E_p f_j = \sum_{c \in C} p(b|c) f_j(b,c) \tag{10}$$

$$E_{\tilde{p}} f_j = \sum_{c \in C} \tilde{p}(b|c) f_j(b,c) \tag{11}$$

Therefore, the meaning of Formula (9) is that when the probability distribution p, the expected value of the feature should be consistent with the expected value of the probability obtained from the training sample data. Then according to Formula (9), a set of λ_j required by the maximum entropy model is calculated by using the GIS (generalized iterative scaling) algorithm.

4.3 Tag Set and Feature Selection

In this paper, the tag set used for the “space” character in a Thai sentence is a Thai tag set expanded on the basis of the ORCHID Thai part-of-speech tag set. The main purpose of this is to use the classified tag set B of the “space” character to replace the “space” tags in the original ORCHID tag set, as shown in Fig. 3.

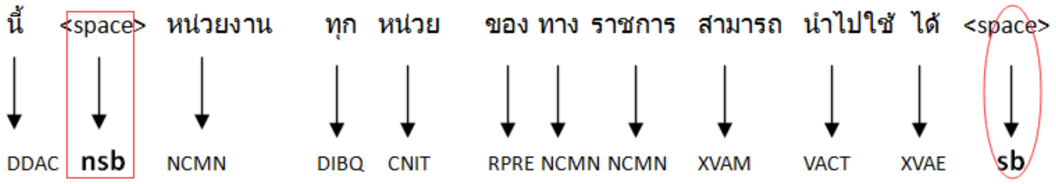


Fig. 3. Extended ORCHID Thai Tag Set

According to Thai language features, the context window selected by the maximum entropy model in this paper is $\{l_{-3}, l_{-2}, l_{-1}, r_1, r_2, r_3, p, n\}$, where l_{-3}, l_{-2}, and l_{-1} are the context tags of the three symbols to the left side of the “space” character and r_1, r_1, and r_3 are the context tags of the three symbols to the right side of the “space” character. p is the distance between the current “space” character and the previous “space” character, and n is the distance between the current “space” character and the next “space” character.

The possible values for each of the context features in the context window during feature extraction are shown in Table 1.

Table 1

Feature Description of the Context Window of Thai Space Characters

Context feature	Thai feature description
YK	Yamok overlapping symbols (ๆ) in Thai
SP	Space character (<space>)
NUM	Thai or Arabic numerals
ASCII	Non-Thai texts or ASCII symbol sequences
(Part of speech)	ORCHID Thai Part-of-Speech tag set

The matching value of each context feature is based mainly on the first match item in Table 1. To ensure that the last entered “space” character can extract the context feature on the right; it is added to the beginning of the input at the same time, as shown in the shaded part of Table 2.

Table 2

Composition of Thai Input Sequences

Type	Content
Original text	ข้อมูลที่ได้ขนาด 6 บิท จะวิ่งเข้า Video Data Bus ซึ่งมีขนาด 8 บิท “run 6-bit data on 8-bit video data bus”
Word sequences	<space>, ข้อมูล, ที่, ได้, ขนาด, <space>, 6, <space>, บิท, <space>, จะ, วิ่ง, เข้า, <space>, Video, <space>, Data, <space>, Bus, <space>, ซึ่ง, มี, ขนาด, <space>, 8, <space>, บิท, <space>
Tag sequences	sb, NCMN, PREL, VSTA, NCMN, nsb, DCNM, nsb, CMTR, nsb, XVBM, VACT, RPRE, nsb, NCMN, nsb, NCMN, nsb, NCMN, nsb, JSBR, VSTA, NCMN, nsb, DCNM, nsb, CMTR, sb

The tags used by the tag sequence in the table are the extended Orchid Thai part-of-speech tag set. NCMN is the common noun tag, PREL is the relation pronominal tag, VSTA is the state verb tag, DCNM is the cardinal qualifier tag, CMTR is the measurement unit tag, XVBM is the front auxiliary verb tag, VACT is the active verb tag, RPRE is the preposition tag, and JSBR is the subordinate conjunction tag. The results of the extracted specific features are shown in Table 3.

Table 3

Contextual Features of Thai Space Characters

b	**$c = l_{-3}$**	**$c = l_{-2}$**	**$c = l_{-1}$**	**$c = r_1$**	**$c = r_2$**	**$c = r_3$**	**$c = p$**	**$c = n$**
nsb	PREL	VSTA	NCMN	NUM	SP	CMTR	4	1
nsb	NCMN	SP	NUM	CMTR	SP	XVBM	1	1
nsb	NUM	SP	CMTR	XVBM	VACT	RPRE	1	3
nsb	XVBM	VACT	RPRE	NCMN	SP	NCMN	3	1
nsb	RPRE	SP	NCMN	NCMN	SP	NCMN	1	1

nsb	NCMN	SP	NCMN	NCMN	SP	JSBR	1	1
nsb	NCMN	SP	NCMN	JSBR	VSTA	NCMN	1	3
nsb	JSBR	VSTA	NCMN	NUM	SP	CMTR	3	1
nsb	NCMN	SP	NUM	CMTR	SP	NCMN	1	1
sb	NUM	SP	CMTR	NCMN	PREL	VSTA	1	4

5 Experiments and Analysis

5.1 Experimental Corpus

The experiment used the Orchid1997 corpus [27], which were developed and constructed by the National Center for Computing Technology in Thailand. There are more than 23,000 Thai sentences that have been accurately calibrated. After the experimental corpus was preprocessed and manually calibrated again, the Thai corpus was standardized and stored in paragraphs according to the extended tag set in this paper, so as to obtain the Thai sentence segmentation corpus needed for our experiment. In the experiment, the whole corpus was divided into the training corpus and test corpus according to the proportion of 9:1, as shown in Table 4.

Table 4

Corpus Data Set Composition

Type	Number of paragraphs	Number of sentences
Training corpus	8870	20927
Test corpus	986	2198
Total	9856	23125

5.2 Experimental Evaluation Index

The evaluation indexes are used in the Thai sentence segmentation experiment, the recognition accuracy rate of the “space” character in Thai input sequence is (space-correct), the recognition error rate of the “space” character at the ending of sentence in Thai input sequence is (false-break), and the recall rate of the “space” character at the ending of sentence in Thai input sequence is (sb-recall). We set variable name in the test corpus as:

(1) The total number of all spaces is T;

(2) The sum of the correct identifications of the sentence-break spaces and non-sentence-break spaces is TC;

(3) The number of false identifications of sentence-break spaces is FSB;

(4) The number of correct identifications of sentence-break spaces is TCB;

(5) The total number of sentence-break spaces is TSB.

So, the specific definition of the evaluation index and the formula are as follows:

$$\text{space-correct}=TC/T \tag{12}$$

$$\text{false-break}=FSB/T \tag{13}$$

$$\text{sb-recall}=TCB/TSB \tag{14}$$

5.3 Experiment and Analysis

For the Thai sentence segmentation, we compared the experimental performance of three Thai sentence segmentation methods according to the current research status of Thai sentence segmentation and thereby verified the effectiveness of the method proposed in this paper. First, we used the n-gram language model [8], which is commonly used in predicting letters, words, or symbolic labels in natural language processing and continuous speech recognition. This method is simple, practical, and easy to implement, and it was used as the reference for the comparative experiment; we call this comparative experiment method “Method 1”. In 2010, The Thai sentence segmentation method realized by Slayden et al., used as a comparative experiment [25], we call this comparative experiment method “Method 2”. In the specific experimental verification, we used the Orchid1997 experimental corpus described in section 5.1. Nine-tenths of the 10.6-MB corpus—a total of 8870 paragraphs—was used as the model training corpus. The remaining one-tenth was the model test corpus. The selected comparison experiment methods are shown in Table 5.

Table 5

Contrast Experiment Settings of Thai Sentence Segmentation

Experimental method	Experimental description
Method 1	Thai sentence segmentation method based on n-gram model [8]
Method 2	Maximum entropy Thai sentence segmentation method developed by Slayden et al. (2010) [25]
Method 3	Our proposed method

Experiment 1: without using the Thai sentence boundary rule base constructed in this paper, we use the maximum entropy classification model to classify the “space” characters in the experimental corpus, and compare it (our method, “Method 3”) with the simple n-gram model [8] (“Method 1”) and the maximum entropy classification model based on context features [25] (“Method 2”). The experimental results of the models of each method on the Thai test corpus are shown in Table 6 below:

Table 6

Experimental Performance Comparison without Rule Correction

Method	Space-correct (%)	False-break (%)	Sb-recall (%)
Method 1	85.43	11.39	62.47
Method 2	**91.19**	3.94	83.50

Method 3	90.87	**2.95**	**84.31**

As can be seen from the experimental performance comparison of Table 6, after analyzing the Thai language features in depth, the experimental performance of Thai sentence segmentation by our method is slightly better than that those by other methods in the same category due to its selection of larger feature windows and more suitable context features. Therefore, we considered a combination of the grammatical rule matching corrections related to Thai sentence boundaries based on the statistical method to improve our method performance and domain applicability in the Thai language sentence segmentation.

Experiment 2: the maximum entropy classification model is used to classify the “space” characters in Thai test corpus. The regular expressions of two kinds of Thai sentence boundary rules constructed in this paper are used to match the type and context of the “space” characters after the maximum entropy classification, so as to correct the classification results of the maximum entropy “space” characters in this paper. In the process of rule correction, first, the rule base B was used to correct the sentence-break “space” characters of the maximum entropy classification model, and rule base A was used to correct the non-sentence-break “space” characters. The correction example is shown in Table 7.

Table 7

Rule Correction Process of Thai Space Character Classification

<table>
<tr><th>Process</th><th>Content</th></tr>
<tr><td>The maximum entropy classification results of this paper</td><td>(ใน,RPRE)|(ปลายปี,NCMN)|(<space>,sb)|(2529,NCNM)|(<space>,nsb)|…|(<space>,nsb)|(โดย,RPRE)|(มติ,NCMN)|(คณะ,NCMN)|(รัฐมนตรี,NCMN)|(<space>,nsb)|(ได้,XVAM)|(จัดตั้ง,VACT)|…|(ขึ้น,XVAE)|(<space>,nsb)</td></tr>
<tr><td>Rule set B correction</td><td>(ใน,RPRE)|(ปลายปี,NCMN)|(<space>,nsb)|(2529,NCNM)|(<space>,nsb)|…|(<space>,nsb)|(โดย,RPRE)|(มติ,NCMN)|(คณะ,NCMN)|(รัฐมนตรี,NCMN)|(<space>,nsb)|(ได้,XVAM)|(จัดตั้ง,VACT)|…|(ขึ้น,XVAE)|(<space>,nsb)</td></tr>
<tr><td>Rule set A correction</td><td>(ใน,RPRE)|(ปลายปี,NCMN)|(<space>,nsb)|(2529,NCNM)|(<space>,nsb)|…|(<space>,nsb)|(โดย,RPRE)|(มติ,NCMN)|(คณะ,NCMN)|(รัฐมนตรี,NCMN)|(<space>,nsb)|(ได้,XVAM)|(จัดตั้ง,VACT)|…|(ขึ้น,XVAE)|(<space>,sb)</td></tr>
<tr><td>Output</td><td>(ใน,RPRE)|(ปลายปี,NCMN)|(<space>,nsb)|(2529,NCNM)|(<space>,nsb)|…|(<space>,nsb)|(โดย,RPRE)|(มติ,NCMN)|(คณะ,NCMN)|(รัฐมนตรี,NCMN)|(<space>,nsb)|(ได้,XVAM)|(จัดตั้ง,VACT)|…|(ขึ้น,XVAE)|(<space>,sb)</td></tr>
</table>

When the Thai sentence boundary rule base constructed in section 4.1 was used, the maximum entropy “space” character classification result was corrected to realize the Thai sentence segmentation, and our method ("Method 3") was compared experimentally with "Method 1" and "Method 2". The comparison results of the experimental performance obtained from the Thai test corpus are shown in Table 8.

Table 8

Thai Sentence Segmentation Performance Comparison after Rule Correction

Method	Space-correct (%)	False-break (%)	Sb-recall (%)
Method 1	85.43	11.39	62.47
Method 2	91.19	3.94	83.50
Method 3	**94.16**	**1.71**	**86.16**

The results showed that the three methods used for the Thai sentence segmentation task all achieved Thai sentence segmentation to varying degrees. The simple n-gram model ("Method 1") was the least effective, with recognition error rate being 11.39%. With the maximum entropy model based on context features ("Method 2"), Thai sentence segmentation was greatly improved, and the recognition error rate dropped to 3.94%. In addition, we choose the advantages of comprehensive rules and statistical methods, and propose a method combining the correction of Thai sentence boundary grammar rules and the maximum entropy Thai sentence segmentation model based on context features (our method, "Method 3"), so that the experimental effect of Thai sentence segmentation can be better optimized. The error rate of Thai sentence ending "space" character recognition is only 1.71%, and in the construction of Thai rules only for the Thai language sentence boundary relative rules, it simplifies the construction work of a large number of complex Thai grammar rules.

6. Conclusions

In this paper, we proposed a method for maximum entropy Thai sentence segmentation combined with Thai grammar rules correction. Firstly, the task of Thai sentence segmentation is analyzed. Secondly, the characteristics of Thai grammar are analyzed, and according to the knowledge of Thai sentence boundary recognition, the Thai grammar rule base is constructed. The method transformed the Thai language sentence segmentation task into the problem of classifying Thai language "space" characters, so as to realize Thai sentence segmentation. On experimental verification our approach has shown a good performance in Thai sentence segmentation on Orchid1997 corpus. The experimental results have demonstrated the superiority by our method over the simple n-gram method and the maximum entropy model based on context features. In the next work, we will consider semantic content and use deep neural network to realize Thai sentence segmentation, in order to improve the accuracy and recall rate of sentence segmentation.

Acknowledgement

This work is supported by the National Nature Science Foundation of China (61462054, 61562052); the Science and Technology Plan Projects of Yunnan province (2015FB135).

REFERENCES

[1]. Ketui N, Theeramunkong T, and Onsuwan C, A Rule-Based Method for Thai Elementary Discourse Unit Segmentation (TED-Seg), Proceedings of International Conference on Knowledge, Information and Creativity Support Systems. Melbourne, **vol. 7710**, 2012, pp. 195-202.

[2]. Liu Lin, Shi Hongmei, and Zhang Yanjun, “A new method of phrase segmentation in statistical machine translation”, Electronic Test, **vol.** 2017, no. 2, 2007, pp. 26-27.

[3]. Wang Boli, Shi Xiaodong, and Su Jinsong, “A Sentence Segmentation Method for Ancient Chinese Texts Based on Recurrent Neural Network”, Acta Scientiarum Naturalium Universitatis Pekinensis, **vol. 2017**, no. 2, 2007, pp. 255-261.

[4]. Wang Boli, Shi Xiaodong, and Tan Zhixing, et al, A Sentence Segmentation Method for Ancient Chinese Texts Based on NNLM, Springer International Publishing, 2016, pp. 3870-396.

[5]. Yang Shu-li, The Study on Sentence Segmentation for English-Chinese Machine Translation, Master Thesis, Beijing Institute of Technology, 2016.

[6]. N. Danvivathana, The Thai Writing System, Master Thesis, Helmut Buske Verlag Hamburg, 1987.

[7]. Yu Zhong-Hua, Zhang Rong, Tang Chang-Jie, Zuo Jie, and Zhang Tian-qing, “Sentence Boundary Detection in Biomedical Texts Using Context Morphological Features”, Mini-Micro Systems, **vol. 27**, no. 1, 2006, pp. 180-184.

[8]. Chen Tianying, CHEN Rong, and PAN Lulu, et al, “Archaic Chinese Punctuating Sentences Based on Context N-gram Model”, Computer Engineering, **vol. 33**, no. 3, 2007, pp.192-193.

[9]. Wang Chuan, Zhang Xiao-Hong, and Han Cai-Hua, “Research on Sentence Segmentation and Punctuation in Ancient Chinese”, Journal of Henan University (Natural Science), **vol. 39**, no. 5, 2009, pp. 525-529.

[10]. Zhang Z., Li L., and Zhang W, Zero-labeling Automatic Sentence Segmentation Algorithm with High Accuracy, Proceedings of National Conference on Man-Machine Speech Communication, NCMMSC2013, Guiyang, China, 2013.

[11]. Chen Hong, Jin Pei-Quan, Yue Li-Hua, Hu Yu-Juan, and Yin Feng-mei, “Comment Long Sentence Segmentation Method Based on Contextual Feature Classification”, Computer Engineering, **vol. 41**, no. 9, 2015, pp.233-237.

[12]. Aishan Wumaier, Tuergen Yibulayin, “Uyghur Sentence Boundary Identification Model Based on Maximum Entropy”, Computer Engineering, **vol. 36**, no. 6, 2010, pp. 24-26.

[13]. Aishan Wumaier, Tuergen Yibulayin, “Sentence boundary detection of Uyghur based on rules and statistics”, Computer Engineering and Applications, **vol. 46**, no. 14, 2010, pp. 162-165.

[14]. Li Xiang, Cai Zangtai, Jiang Wenbin, Lv Yajuan, and Liu Qun, “A Maximum Entropy and Rules Approach to Identifying Tibetan Sentence Boundaries”, Journal of Chinese Information Processing, vol. 25, no. 4, 2011, pp.39-44.

[15]. Cai Zangtai, “Research on the automatic Identification of Tibetan Sentence Boundaries with Maximum Entropy Classifier”, Computer Engineering &. Science, **vol.34**, no. 6, 2012, pp. 187-190.

[16]. Ma Wei-Zhen, Wanme Zhaxi, and Nima Zhaxi, “Method of Identification of Tibetan Sentence Boundary”, Journal of Tibet University, **vol. 2012**, no.2, 2012, pp. 70-76.

[17]. Zhao Wei-Na, Yu Xin, and Liu Hui-dan, et al, “Modern Tibetan Auxiliary Ending Sentence Boundary Detection”, Journal of Chinese Information Processing, **vol. 27,** no.1, 2013, pp.115-119.

[18]. MT Uliniansyah, Gunarso, E Nurfadhilah, et al, A Tool to Solve Sentence Segmentation Problem on Preparing Speech Database for Indonesian Text-to-speech System, Procedia Computer Science, **vol. 81**, 2016, pp. 188-193.

[19]. N Wanjari, GM Dhopavkar, NB Zungre, Sentence Boundary Detection For Marathi Language, Procedia Computer Science, **vol. 78**, 2016, pp. 550-555.

[20]. Wanma Lengzhi, Tibetan sentence boundary identification method based on sentence-final part of speech, Master Thesis, Qinghai Normal University, 2016.

[21]. L. Sungkornsaran, Thai Syntactical Analysis System by Method of Splitting Sentences from Paragraph for Machine Translation, Master thesis, King Mongkut’s Institute of Technology Ladkrabang, Thailand, 1995.

[22]. P. Mittrapiyanuruk and V. Sornlertlamvanich, The Automatic Thai Sentence Extraction, Proceedings of the Fourth Symposium on Natural Language Processing. Soochow, 2000, pp. 23-28.

[23]. P. Charoenpornsawat and V. Sornlertlamvanich, Automatic Sentence Break Disambiguation for Thai, Proceedings of International Conference on Computer Processing of Oriental Languages (ICCPOL), 2001. Seoul, 2001, pp. 231-235.

[24]. Wirote Aroonmankun, Thoughts on word and sentence segmentation in Thai, Proceedings of the Seventh Symposium on Natural language Processing, Pattaya, Thailand, December 13–15, 2007, ed. A. Kawtrakul & M. Zock, Kasetsart University, 2007, pp. 85–90.

[25]. Glenn Slayden, Mei-Yuh Hwang, and Lee Schwartz, Thai sentence-breaking for large-scale SMT, Proceedings of the 1st Workshop on South and Southeast Asian Natural Language Processing (WSSANLP), the 23rd International Conference on Computational Linguistics (COLING), Beijing, August 2010, pp. 8-16.

[26]. Nathcha Tangsirirat, Atiwong Suchato, and Proadpran Punyabukkana, et al, Contextual behaviour features and grammar rules for Thai sentence-breaking, Proceedings of the 10th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology. Krabi, 2013, pp. 1-4.

[27]. Virach Sornlertlamvanich, Naoto Takahashi, and Hitoshi Isahara, “Building a Thai part-of-speech tagged corpus (ORCHID)”, Journal of the Acoustical Society of Japan, **vol. 20**, no.3, 2000, pp. 189-198.