19th International Computer Science and Engineering Conference (ICSEC)
Chiang Mai, Thailand, 23-26 November, 2015

Development of Thai Word Segmentation Technique for Solving Problems with Unknown Words

Chanin Mahatthanachai
Asian Development College for Community Economy and Technology, Chiang Mai Rajabhat University
E-mail: chaninm2000@gmail.com

Nuttiya Tantranont
Asian Development College for Community Economy and Technology, Chiang Mai Rajabhat University
E-mail: nuttiya18@gmail.com

Kanchit Malaivongs
Royal Society of Thailand
E-mail: kanchit.ma@gmail.com

Ekkarat Boonchieng
Faculty of Science, Department of Computer Science, Chiang Mai University
E-mail: ekkarat@boonchieng.net

***Abstract*—This research has an objective to develop an efficient technique for Thai word segmentation, especially those nonexistent in dictionaries. The researchers developed a model for Thai word segmentation by relying on grammar and rules to solve the problems with words not found in dictionaries. The model was intended to be used as the best approach of word segmentation, which applied the segmentation technique developed by the researchers called PTTSF (Parsing Thai Text with Syntax and Feature of Word). The system of this technique operates by starting from finding the boundary of each word in Thai sentences. If the system finds a word that does not exist in the dictionary or a meaningless word, it would not be able to solve the problem with the method of longest-matching algorithm. Therefore, rules need to be specified to solve such problems. In this study, 28 rules were created and Digraph method was used to find a pattern of word segmentation with the highest probability based on the grammatical principle. After the procedure of finding boundary of the word, the result from correct word segmentation can be used for further processes. In analyzing efficiency of the system, its accuracy in word segmentation was the main point of concern. The results revealed that the derived mapping technique could solve the problem concerned with segmentation words that do not exist in the dictionary with an average accuracy over 90% of the whole document. However, the researchers encountered with ambiguous words problem. Although this problem rarely occurs, it could affect accuracy of word segmentation.**

Keywords—Thai Word Segmentation; Parsing Thai Text with Syntax and Feature of Word; Unknown Words

I. INTRODUCTION

Word segmentation is a very important process because it is the very first step of word processing [1]-[3], which includes word finding, language translation, [4], [5] text-to-speech conversion, and so on. There are different segmentation methods for different languages [6] depending on complexity and unique characteristics of each language [7]-[16]. If there is incorrect word segmentation in any language, other processes following the word segmentation will also be wrong. For Thai language, there are two major problems [17] that have been commonly found, namely ambiguous words and not-in-dictionary words. These words are usually found in specific names, technical terms, or names of places. Since these words appear commonly in Thai documents, they can lead to defects of word segmentation, which will be less accurate. At present, solutions for word segmentation can be categorized into 3 approaches. 1) Using the rule of Thai syllable creation [18], [19]. The rule is that a syllable must consist of a consonant, a vowel, a tone mark, a final consonant, and a silent symbol or Karan mark (the mark placed over the final consonant of a word in Thai language to indicate that it is mute). This approach is the easiest one, and can work fastest. However, there are still some problems with it. Firstly, it can separate only a word with one syllable, not able to handle a multi-syllable word. Secondly, it cannot solve the complicate problem of Thai consonants, which can be both initial consonants and final consonants. 2) Using a dictionary [3]-[5]. This approach requires making a list of words in advance. In order to separate words in a sentence, the sentence's words will be compared with the list of words in the dictionary. This approach can solve the problem regarding multi-syllable words. However, there is still a problem concerning with words not found in the dictionary. 3) Using the technique of machine learning [20], [21]. With this approach, the machine is trained with a large warehouse of texts [22], [23], of which words have been separated correctly. This approach is highly efficient. However, it depends mainly on accuracy and size of the text warehouse [24]. Segmentation techniques for Thai words have been developed in various forms for maximizing the benefit in applying Thai language on computer. The techniques also play important roles in language processing, especially processing at the word level, to be more efficient. This research was conducted to study algorithms for Thai word segmentation and to find the best approach for word segmentation. This approach should also be able to solve the word segmentation problem concerning unknown words. Therefore, the researchers proposed the technique of PTTSF, which relies on using a rule base, so that the segmentation result has accuracy over 90% of the whole document. The new word segmentation algorithm had been presented in detail. The model of the PTTSF (Parsing Thai

978-1-4673-7825-3/15/$31.00 ©2015 IEEE

Text with Syntax and Feature of Word) technique had been developed and tested. The segmentation results were presented and assessed. This study also gives recommendations for further improvement on the technique.

II. RESEARCH METHODOLOGY

The researchers had developed a new algorithm for word segmentation by building a module to be used for cutting the entered texts or sentences into separated words to be used for other processes. This new method of word segmentation is called PTTSF, which can be explained with Figure 1. PTTSF technique consists of 3 main components (steps) namely: A) word segmentation based on the dictionary; B) solving unknown words in the dictionary; and C) analyzing probability of the grammar by using Digraph.

A. Word segmentation based on the dictionary

Word segmentation based on the dictionary is the first step of PTTSF. This step separates words in each sentence into individual words based on a basic method such as the longest matching method. Such method can be applied together with some simple rules. The rules applied in this research included the use of a blank and the beginning of a new paragraph. After segmentation in this step, the following result is obtained.

เขา/ไป/[ช้]/อป/ปิ้ง/ที่/[คาร์ฟูร์]/แผนก/[ซูเปอร์มาร์เก็]/ต/และ/ไป/กิน/ข้าว/ที่/[แมค]/โค/[ร]

Khao/Pai/[Sh]/Op/Ping/Tee/[Carrefour]/Phanaek/[Superm arke]/T/Lae/Pai/Kin/Khao/Tee/ [Mak]/ko/[r].

(He went shopping at the supermarket department of Carrefour Mall and had a meal at Makro.)

After using the word segmentation technique based on the dictionary, problems still remain with words in the [ ] brackets, which are words unknown in the dictionary. Hence, the result became inaccurate. However, this problem can be solved by using the 28 rules that the researchers had synthesized from Principles of Thai Language [25], which were aimed to solve problems with unknown words.

B. Solving the problem with unknown words in the dictionary

The solution of solving the problem with unknown words involves with finding boundary of words that do not exist in the dictionary. Initially, the problem of unknown words has to be figured out, and then some rules need to be created to solve the problem. The researchers collected some rules to be a part of the PTTSF system. There were 28 rules altogether. These rules were divided into 2 sets, namely the rules for combination of words before and after the unknown words, and the rules for combination of characters to become a new word. For the part of using the rule base for solving the problem, these rules were obtained from studying sentences written in general documents or in online articles. Most of the words appeared to be nouns. According to the statistical analysis on occurrence of word types in sentences, 40% of words in Thai sentences were found to be nouns [26]-[28]. After taking the unknown words to be solved by the two sets of rules, the new words that were created from the rules had been defined for their type to belong to the type of noun.

- The rules for combination of words before and after the unknown words. This set of rules will check combination of words that do not exist in the dictionary, as well as nearby words (before and after the unknown word) in order to build a new word.

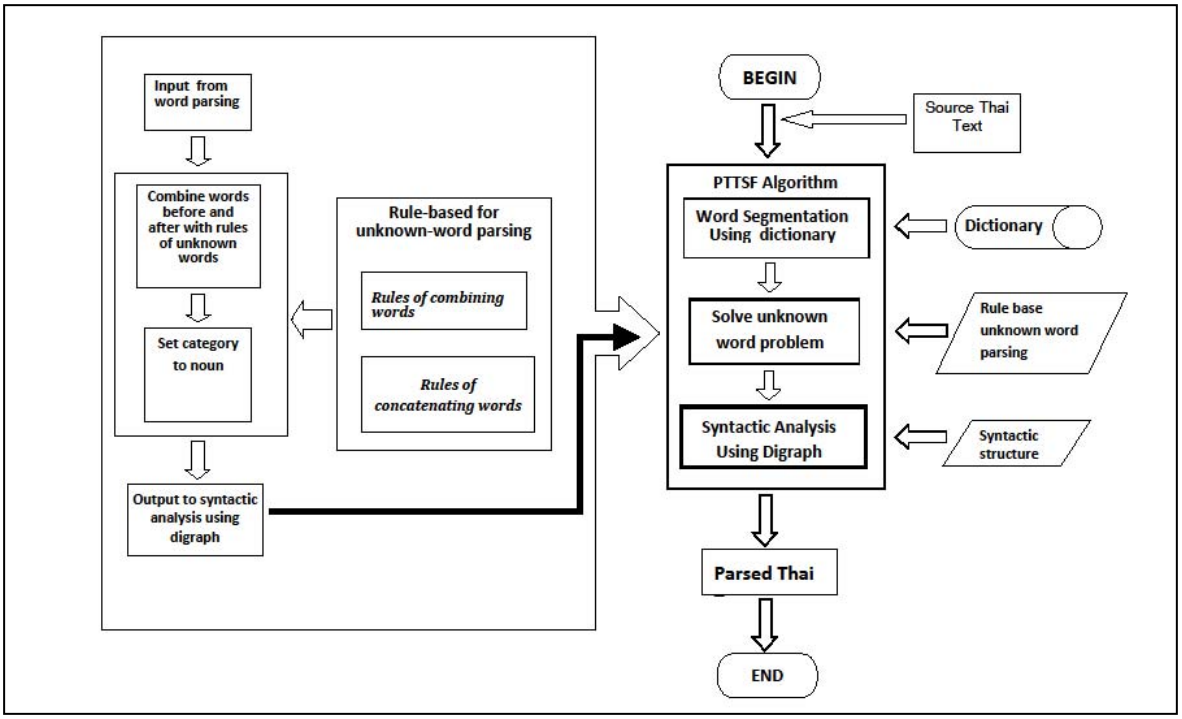

Fig. 1. The model representing operation of PTTSF technique and the application of rule base

Types of surrounding words were taken for consideration in order to check whether they comply with the established rules or not. If so, the nearby words will be combined with the unknown word and define the type of the newly created word as "noun." Therefore, this problem was solved by using 7 rules. The words to be combined will be checked with the rules.

If their combination complies with all the 7 rules, these words can be combined into a new word. The words to be combined must have these following characteristics.

Rule No. 1: The words must not be verbs, such as กิน, เดิน, นั่ง, etc.
Rule No. 2: The words must not be prepositions, such as ที่, บน, etc.
Rule No. 3: The words must not be conjunctions, such as กับ, และ, etc.
Rule No. 4: The words must not be adjectives, such as เร็ว, ช้า, etc.
Rule No. 5: The words must not be articles, such as ความ, การ, etc.
Rule No. 6: The words must not be negative words, such as ไม่ได้, มิได้, etc.
Rule No. 7: The words must not be pronouns, such as ผม, เขา, ใคร, นี่, etc.

- The rules for combination of characters to become a new word consist of rules No. 8-28. These rules were taken for processing with each character. Before combining these characters together, they need to be checked with the rules for combination first. In summarize, the rule 8-28 were as of follows:

  Rule No. 8: The words must not be English.

  Rule No. 9: The words must not be numeric.

  Rule No. 10: If there are 2 characters and a silent symbol (์) , such as, นต์ , อร์ , etc, these characters can be combined with the characters in front of them.

  Rule No. 11: If there is 1 character that is not (ฌ), the character must be combined before and after the words or characters, for example, a character ร can be combined with the word แมคโค and become แมคโคร.

  Rule No. 12: If the words are not "มี" "การ" and "ความ" they are allowed to be combined.

  Rules No. 13 to 28 is the rules for a combination of words. For example, if the first character is "เ", the second character is a consonant "ข", the third character is a cluster "ล", and the last character is "า", then the words can be combined together as เขลา.

  These rules were derived from both analysis and synthesis methods. They are rules that originated from various sentences found in daily life. For instance, อินเท + อร์ will become a new word "อินเทอร์เน็ต", The source of this rule is from words commonly found in Thai articles in journals, textbooks, the internet, etc. Other rules were derived from the Principles of Thai Language Textbook [25].

When taking the two sets of rules (28 rules in total) for the process of word synthesis, the nearby contextual words will be combined with the unknown word. Combination of words will rely on the surrounding context as the contextual words are combined with the unknown word. This combination starts with the first word right before the unknown word and continues to find words that do not follow the two sets of rules. This procedure is repeated until no unknown word is found. Then the type of the words will be determined as noun. The sentence that has been regressed will then be entered to find different patterns of the segmentation. Finally, probabilities of the grammar will be analyzed. The pattern with the highest probability will be used as the best pattern for word segmentation.

C. Analyzing probability of the grammar by using Digraph

From the steps of word segmentation by using the dictionary and solving the problem with unknown words, a sentence is derived. This sentence is represented with the type of words as 'noun' (NP) . The next step is to find probability (P) of each type of sentence occurrence by using Diagraph. The results from using Diagraph are words and their types. In addition, probability of the grammar can be determined correctly because the type of the unknown words has been set as noun. The patterns of word segmentation will be analyzed to select only one pattern that is most accurate according to the grammar. The principle of Digraph can be explained as follow.

Assume A = Node
R= Edge
SENT = Sentence
NP= Noun
VP= Verb

For example

SENT $\longrightarrow$ (NP*) VP* (NP*) (SENT)

A = {SENT , NP1 , VP, NP2}

R={(SENT,NP1),(SENT,VP),NP1,NP1),(NP1,VP), (VP,VP),(VP,NP2),(NP2,NP2),(NP2,SENT)}

From the principle of Digraph, SENT can be linked to the Node NP, VP, and SENT if A consists of Node SENT, NP1,

VP, NP2, and R, which is a relationship between Nodes that occurs. For examples, (SENT, NP1), (SENT, VP), (NP1, NP1), (NP1, VP), (VP, VP), (VP, NP2), (NP2, NP2), and (NP2, SENT). The highest probability of an accurate sentence can be calculated by the probability of relationship (R) that is linked to each Node (A).

From the step of using Digraph to find probability of each pattern of word segmentation, we derive the pattern of word segmentation with the highest probability, which is also the one with the highest accuracy. The result from this technique of word segmentation can be implemented in real application as shown in the following example.

เขา/ไป/ช้อปปิ้ง/ที่/คาร์ฟูร์/แผนกซูเปอร์มาร์เก็ต/และ/ไป/กิน/ข้าว/ที่/แมคโคร

Khao/Pai/Shopping/Tee/Carrefour/PhanaekSupermarket/ Lae/Pai/Kin/Khaw/Tee/Makro

(He went shopping at the supermarket department of Carrefour Mall and had a meal at Makro).

For the operation of word segmentation with the 3 steps of the PTTSF algorithm (namely segmentation words based on the dictionary, solving the problem with unknown words in the dictionary, and analyzing probability of the grammar by using Digraph), the researchers has tested efficiency of the system to find accuracy of its word segmentation results. The tests were taken two times. The first test was conducted with the initial rules. The second test was conducted in order to improve the original rules to be more accurate and to solve the problems related to implementation of the prototype rules. The word segmentation was expected to have an accuracy rate over 90% of the whole document. The researchers tested the system with words from 30 examples of documents and articles from various sources including academic articles, general articles, announcements, computer advertisements, news, poems, and computer articles as shown in TABLE I.

TABLE I. Types of documents

No	Type	Number of Document
1	Ambiguous text	1
2	Announcement	2
3	Article	6
4	Computer advertisement	6
5	Computer article	3
6	Communication advertisement	2
7	Doctrine	1
8	Lesson	2
9	News	4
10	Technical announcement	1
11	Verse	2
Total		30

III. TEST RESULTS AND DISCUSSION

This article used PTTSA word segmentation method by Chimpipop [[29 as a benchmark data for a comparision of accuracy. The results shown that PTTSA could not be able to wrap unknown words that do not exis in the dictionary, therefore, the average of accuracy of word segmentaion using PTTSA were 86.62%. As a result, the authors have developed PTTSF techniques to the solve unknown word problems. The results of comparison on accuracy of word segmentation can be shown in Table II

TABLE II. Comparison on accuracy of word segmentation

Article No.	Accuracy of word segmentation using PTTSA (%)	Accuracy of word segmentation using PTTSF (%)		% of increased accuracy
		Experiment 1	*Experiment 2*	
1	81.22	67.87	84.16	24.00
2	86.35	87.37	90.10	3.12
3	91.37	84.17	91.61	8.84
4	89.58	82.81	93.23	12.58
5	85.92	81.59	96.03	17.70
6	85.27	83.72	92.25	10.19
Average	86.62	81.26	91.23	12.74

A. The test results from the 1st word segmentation experiment

From conducting the first word segmentation test with the PTTSF technique by using 30 articles to find accuracy of the technique, there were 24 articles that had accuracy over 90%. The total average of accuracy for these 24 articles were 95.07%. However, there were 6 articles as shown in Table II that had accuracy less than 90%. The total average of accuracy using PTTSF technique was only 81.26% (which had an average accuracy less than PTTSA). Since the algorithm of PTTSF cannot make segmentation of English-native words that are written in Thai and mix with Thai words, a special technique is required for a sentence that contains both Thai words and English-native words. In this case, the rules of word segmentation need to be adjusted. Words must first be connected. Then the rule of word combination and the rule of character combination will be applied to derive a new word as follow.

- Add parameter : (checkeng(ssbuf)) in procedure TmdParsing.WordParsing1.

The function developed by the researchers will have a characteristic of being used for checking words with English-native words and numbers. However, this technique of PTTSF can also wrap transliterated words such as ซอฟต์แวร์ (software) and อีเมล์ (e-mail). Therefore, this technique is more efficient than the technique in segmentation words based only on the dictionary, which is not capable to wrap such these words if they do not exist in the dictionary.

B. The test results from the 2nd word segmentation experiment

The results from the 2nd experiment were obtained after adjusting the word segmentation rules, connecting the words, and using the word combination rule and also the character combination rule. It was found that the results from the 30 examples of articles became more accurate than the original word segmentation from the 1st test. The total average of accuracy for 24 articles were increased from 95.07% to

95.92%. Moreover, the 6 articles that had accuracy less than 90% (which had an average accuracy of 81.26%) now have an average accuracy of 91.23% (TABLE II).

In comparing accuracy rates from the first experiment and the second experiment, it was found that the overall accuracy of word segmentation increased in the second experiment. However, there was only one article that still had accuracy lower than 90% (by having an accuracy of only 84.16%). This problem happened because there were some verbs appear between the unknown words as shown below.

- The word "บร/อด/แบ/นด์" (broadband) had a problem of incorrect word segmentation because the derived words "อด" (miss) and "แบ" (open) are words that exist in the dictionary as verbs.
- The word "ออพติไมส์" (optimize) had a problem because the derived word "ติ" (blame) is a verb in the dictionary, and appears during the two unknown words.
- The word "โปรโตคอล" (protocol) had a problem because the derived word "โต" (big) is an adjective in the dictionary, and appears during the word "โปร" and the word "คอล"

This problem was found to involve with ambiguity of words. Ambiguous words in Thai language have a characteristic of being able to be wrapped correctly for more than one form. These ambiguous words appear commonly in Thai dictionaries. This problem cannot be solved by using the rule-base approach because, after analyzing these words grammatically and wrap them correctly according to the established rules, they can be wrapped correctly into more than one form. Therefore, we need to consider the context around the interested word as well. This approach is called Feature Based Approach [30]. The rules for this approach are built from conducting 'data training' with nearby words to find rules that can help to wrap such ambiguous words correctly. For example, in the sentence "เขายืนทำตากลมอยู่หน้าบ้าน" [Khao-Yuen-Ta-Klom-Yoo-Na-Ban] (He stands and makes round eyes in front of the house), there is an ambiguous word "ตากลม" (which can be pronounced as either Ta-Klom or Tak Lom). According to the data training, it was found that the important conditional word is "ทำ"(make). Therefore, the ambiguous word should be wrapped as "ตา กลม" (Ta-Klom). Next, from the sentence saying "วันนี้อากาศดีมากเขาจึงออกมานั่งตากลมอยู่ที่ชานบ้าน" [Wannee-Agard-Dee-Mak-Khao-Jueng-Ork-Ma-Nang-Tak-Lom-Yoo-Tee-Charn-Barn] (Today the weather is very nice so he came out to sit and expose to the wind at his house's terrace). The ambiguous word in this sentence is "ตากลม", which can be pronounced as either Ta-Klom or Tak-Lom. However, from conducting the data training, it was found that there are two important conditional words namely "อากาศ" (weather) and "นั่ง" (sit). Therefore, the pattern in segmentation this ambiguous word would be "ตาก ลม" (Tak-Lom). The researchers will use the method of building a dictionary to collect the knowledge base regarding ambiguous words, which is called ambiguous knowledge. This method is an algorithm for building a rule based on the data training [23]. The derived rule can be presented in a form of continuous conditions as follow.

If T1 and T2 and …. Tn then class Cx

When T1 and T2 and…Tn refer to the example rules. For instance, in the case where a word like "ตากลม" appears in the sentence, if conditional words like "ทำ" (make) or "ปลา" (fish) are found in the sentence, the pattern of word segmentation would be Ta-Klom. That means T1 and T2 refers to "ทำ" and "ปลา" respectively. These conditions can be regarded as rule conditions. Cx is the output, which can be either ตา/กลม (Ta-Klom) or ตาก/ลม (Tak-Lom).

IV. Conclusion

The researchers has studied problems regarding word segmentation at present and found that a major problem concerns with unknown words in the selected dictionary. Therefore, the researchers developed a new algorithm for word segmentation by creating a new prototype of words. The technique with this algorithm is called PTTSF, which consists of 3 main components namely: 1) word segmentation based on the dictionary; 2) solving the problem with unknown words; and 3) analyzing probability of the grammar by using Digraph. After the algorithm was applied to wrap some texts, it was found that the results had an average accuracy over 90%. However, the results appeared to be lower than 90% accurate in some tasks. Therefore, the researchers developed a solution to solve the word segmentation problem with unknown words by using a rule base. To use the rule base, two sets of rules were established, namely the rules for combination of unknown words and the rules for combination of characters. However, the researchers still encountered some problems of the segmentation system based on the PTTSF and the Rule Base models. The problems happened when there were verbs appear between two unknown words. This case remains to be a problem. The method of using the PTTSF algorithm together with the Rule base was found to be unable to wrap words correctly. This cause also affects accuracy of the results. From the articles that were used for the experiments in applying this word segmentation approach, it was found that ambiguous words can also be an important problem. Although this problem is not found so commonly, there are chances that the problem can occur. Therefore, we must find a solution to this problem by considering the context of nearby words [31]-[33]. This solution can be further studied by conducting a research on Feature base approach. This PTTSF technique could be used for the data warehouse and data mining as the implementation. The new technique could also be applied to an information retrieval systems.

ACKNOWLEDGMENT

We would like to thank Mr. Anukul Chimpipop for the PTTSA Framework so that we could be able to compare and contrast the accuracy of word segmentation. Special thanks also to NECTEC for the word segmentation techniques and VAJA TTS in order to test the results of word pharsing.

REFERENCES

[1] V. Sornlertlamvanich, et al., "The state of the art in thai language processing," in *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics,* Association for Computational inguistics, 2000. pp. 1-2.

[2] W. Aroonmanakun, "Collocation and thai word segmentation," in *Proceedings of the 5th SNLP & 5th Oriental COCOSDA Workshop*, 2002. pp. 68-75.

[3] Y. Poowarawan, "Dictionary-based thai syllable separation," in *Proceedings of the Ninth Electronics Engineering Conference*, 1986.

[4] W. Aroonmanakun, "Thoughts on word and sentence segmentation in Thai," in *Proceedings of the Seventh Symposium on Natural language Processing*, Pattaya, Thailand, 2007. pp. 85-90.

[5] S. Klaithin, et al., "Thai Word Segmentation Verification Tool," in *Proceedings of the 2nd Workshop on South and Southeast Asian Natural Language Processing*, IJCNLP 2011, Chiang Mai, Thailand, 2011. pp. 16-22.

[6] A. Kawtrakul, C. Thumkanon, "A Statistical Approach to Thai Morphological Analyzer," in *Proceeding of the 5th Workshop on Very Large Corpora*, 1997. pp. 289-286.

[7] T. Kudo, K. Yamamoto, and Y. Matsumoto, "Applying conditional random fields to japanese morphological analysis," *EMNLP*, vol. 4, pp. 230-237, Jul. 2004.

[8] F. Peng, F. Feng, and A. McCallum, "Chinese segmentation and new word detection using conditional random fields," in *Proceedings of the 20th international conference on Computational Linguistics*, Association for Computational Linguistics, 2004. pp. 562.

[9] J. Gao, et al., "Chinese word segmentation and named entity recognition: A pragmatic approach," *Computational Linguistics*, vol. 31, no. 4, pp. 531-574, Dec. 2005.

[10] Z. Wang, T. Liu, "Chinese unknown word identification based on local bigram model," *International journal of computer processing of oriental languages*, vol. 18, no. 3, pp. 185-196, Sep. 2005.

[11] N. Xue, "Chinese word segmentation as character tagging," *Computational Linguistics and Chinese Language Processing*, vol. 8, no. 1, pp. 29-48, Feb. 2003.

[12] Q. T. Dinh, et al., "Word segmentation of Vietnamese texts: a comparison of approaches," in *6th international conference on Language Resources and Evaluation*, Marrakech, Morocco, 2008.

[13] W. Liu, L. Lin, "Probabilistic ensemble learning for vietnamese word segmentation," in *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, ACM, 2014. pp. 931-934.

[14] P. C. Chang, M. Galley, and C. D. Manning, "Optimizing Chinese word segmentation for machine translation performance," in *Proceedings of the third workshop on statistical machine translation,* Association for Computational Linguistics, 2008. pp. 224-232.

[15] C. K. Fan, W. H. Tsai, "Automatic word identification in Chinese sentences by the relaxation technique," *Computer Processing of Chinese & Oriental Languages*, vol. 4, no. 2, pp. 33-56, Nov. 1988.

[16] C. S. Khoo, Y. Dai, and T. E. Loh, "Using statistical and contextual information to identify two-and three-character words in Chinese text," *Journal of the American Society for Information Science and Technology*, vol. 53, no. 5, pp. 365-377, Jan. 2002.

[17] C. Haruechaiyasak, S. Kongyoung, and M. Dailey, "A Comparative Study on Thai Word Segmentation Approaches," in *Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, 2008. ECTI-CON 2008,* 2008. pp. 125-128.

[18] D. D. Palmer, "A trainable rule-based algorithm for word segmentation," in *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, 1997. pp. 321-328.

[19] A. Kawtrakul, et al., "Automatic thai unknown word recognition," in *Proceedings of the Natural Language Processing Pacific Rim Symposium*, Phuket, Thailand, 1997. pp. 341-348.

[20] C. Haruechaiyasak, S. Kongyoung, and C. Damrongrat, "LearnLexTo: a machine-learning based word segmentation for indexing Thai texts," in *Proceedings of the 2nd ACM workshop on Improving non english web searching*, ACM, 2008. pp. 85-88.

[21] P. Charoenpornsawat, B. Kijsirikul, and S. Meknavin, "Feature-based thai unknown word boundary identification using winnow," in *Circuits and Systems, 1998. IEEE APCCAS 1998*, Chiangmai, Thailand: IEEE, 1998. pp. 547-550.

[22] V. Sornlertlamvanich, "Word segmentation for Thai in machine translation system," *Machine Translation*, Jan. 1993.

[23] C. Haruechaiyasak, S. Kongyoung, "TLex: Thai Lexeme Analyser Based on the Conditional Random Fields", in *Proceedings 8th International Symposium on Natural Language Processing*, 2009.

[24] C. Haruechaiyasak, et al., "A collaborative framework for collecting Thai unknown words from the web," in *Proceedings of the COLING/ACL on Main Conference Poster Sessions*, Association for Computational Linguistics, 2006. pp. 345-352.

[25] K. Thonglor, *Principles of Thai Language*, 3rd Edition, Bangkok: Ruamsarn, 2002.

[26] T. Theeramunkong, S. Usanavasin, "Non-Dictionary-Based Thai Word Segmentation Using Decision Trees," in *Proceedings of the first international conference on Human language technology research*, Association for Computational Linguistics, 2001. pp. 1-5.

[27] C. Kruengkrai, V. Somlertlamvanich and H. Isahara, "A conditional random field framework for thai morphological analysis," in *Proceedings of LREC*, 2006. pp. 2419-2424.

[28] M. Boriboon, et al., "Best corpus development and analysis," in *Asian Language Processing, 2009. IALP'09*, Singapore: IEEE, 2009. pp. 322-327.

[29] A. Chimpipop, "Parsing thai text with syntactic analysis using digraph representation," Diss. MS Thesis in Computer Science. Faculty of Graduate Studies., Mahidol University, Bangkok, 1999.

[30] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the Eighteenth International Conference on Machine Learning*, C. E. Bradley, Eds. San Francisco, USA: Morgan Kaufmann Publishers Inc, 2001. pp 282-289.

[31] K. Khankasikam, N. Muansuwan, "Thai word segmentation a lexical semantic approach," in *the Proceedings of the Tenth Machine Translation Summit,* 2005. pp. 331-338.

[32] P. Mittrapiyanuruk, V. Sornlertlamvanich, "The automatic Thai sentence extraction," in *Proceedings of the fourth symposium on Natural Language Processing*, 2000. pp 23-28.

[33] A. Islam, D. Inkpen, and I. Kiringa, "Applications of corpus-based semantic similarity and word segmentation to database schema matching," *The VLDB Journal*, vol. 17, no. 5, pp. 1293-1320, Oct. 2008.