2011 International Conference on Modeling, Simulation and Control
IPCSIT vol.10 (2011) © (2011) IACSIT Press, Singapore

Thai-Word Segmentation through Thai Writing Structure Matching

Vuttichai Vichianchai

Informatics of Faculty Mahasarakham University, Thailand
Email: onizuka.p@gmail.com

Abstract. This research proposed the approach of Thai-Word Segmentation through Thai-Writing Structure Matching, with an aim to create Thai writing structure for word segmentation. Indeed, the writing structure was originated from the words stored in the 1999 Royal Institute Dictionary and Thai-writing levels. This research also aimed to improve word segmentation by adding the number of the rules for the matching, decreasing the waste of the storage space, and reducing the time-consumption of the word segmentation processing. The documents used for the performance test of word segmentation contained a variety of data, in terms of the writing patterns used for communicating with the readers under the identical understanding. These documents should provide neither illustrations nor equations. In fact, the documents used in this research were concerned with news papers, articles, Buddhism, encyclopedia, laws, non-fictions, the Royal Family's news, interviewing, and general news. Theses papers contained a variety of word use. Our approach works successful with about 94% accuracy.

Keywords: word segmentation, Thai-word segmentation, Thai writing structure, Thai-word processing

1. Introduction

Writing Thai sentence was a continual writing pattern without any punctuation marks but with word segmentation by occasionally leaving an empty space, and this differentiated Thai sentences from English sentences which apparently separated a single word with an empty space. With this distinctive feature, to process Thai language via the computer necessarily required the approach that assisted the computer in recognizing the extent of Thai words, in order to find the solution for the word segmentation in Thai language processing, under the principles of linguistics. In this regard, the previous studies revealed that the researches on Thai-word segmentation have been continuously developed for many years, so that many useful approaches were proposed. These approaches could be categorized based on the word-segmentation models as followings: 1) Rule-based Approach; 2) Algorithm Approach; 3) Dictionary-based Approach; and 4) Corpus-based Approach. The 4 mentioned approaches provided the highly-effective word segmentation but still contained some problematic gaps. In particular, Rule-based Approach and Algorithm Approach were time-consuming for matching; Dictionary-based approach spent too much space for the storage; and Corpus Approach took much time for the processing. As a consequence, to solve these problems, the researchers initiated the approach of Word Segmentation through Thai-Writing Structure Matching.

2. Related Studies and Theories

2.1 The studies of Thai syllable deletion by the invented rule, in which the rule was established under Thai grammar and to consider the features of Thai syllable. Indeed, the rule could be divided into 2 categories including: 1) Front Boundary Recognition Rule; and 2) Tail Boundary Recognition Rule. In further, each of these rules could be divided into 2 sub-categories, which were 1) Group A categorized by vowel, consonants, and tone marks; and Group B categorized by the use of a single vowel [2].

2.2 The studies of the syllable deletion by the rule established from Thai grammar, in which various exceptional syllables were stored in the data files, since these syllables did not match the invented rule.

Features of the rule were considered from the alphabets existing in the syllables or words, in which Thai characters could be categorized into 5 groups [3].

2.3 The studies of the syllable segmentation by using dictionary which were considered the primary research regarding the word deletion with an application of dictionary; the syllables were stored in the dictionary. The grammatical rules would be brought in to help in case that some syllables were not available in the dictionary. That is, there would be an investigation of the Alphabet String from the left to the right on the syllables stored in the dictionary. In case that more than 1 syllable were found in the dictionary, it would be the syllable segmentation by selecting the longest syllable first and then put a reversal mark on the rest of the syllables. This process would be rerun until it met the end of the String. On the contrary, in case that the longest syllable was selected but there still were some unavailable word existing in the dictionary, it would be a backtracking to the latest reversal mark and the second-longest syllable would be selected instead. After all, this process was called the Longest Matching [4].

2.4 This research used the statistics to solve the problems in word deletion and word-functional determination or the word sub-category. Trigram model was employed in word deleting which meant that there was the use of the statistical value accumulated from the continuality of the word function or the word sub-category. To be exact, the most proper word deletion could be done by seeking out the sentence that offered the highest possibility [5].

2.5 The studies of Thai romanization is the way to write Thai language using roman alphabets. It could be performed on the basis of orthographic form (transliteration) or pronunciation (transcription) or both. As a result, many systems of romanization are in use. The Royal Institute has established the standard by proposing the principle of romanization on the basis of transcription. To ensure the standard, a fully automatic Thai romanization system should be publicly made available. In this paper, we discuss the problems of Thai Romanization. We argue that automatic Thai romanization is difficult because the ambiguities of pronunciation are caused not only by the ambiguities of syllable segmentation [7].

2.6 The studies on a pioneer Thai LVCSR system. An automatic data-driven technique for buildinga language model for Thai LVCSR has been proposed, which is expected to accelerate Thai text corpus development in the future [8].

2.7 The studies of Thai speech recognition, investigating pronunciation changes such as syllable and phoneme elisions as well as phoneme shifts in Thai spontaneous speech. We compare several approaches to model these effects in large vocabulary continuous speech recognition across multiple domains. This work includes experiments on two new speech databases that significantly alleviate the data sparseness problem of earlier publications. We found that given sufficient training data, a fully data driven approach using an allophone cluster tree yields the best results. For the experiments in this paper we segmented the training text first using a segmented from a previous project [6] and built a statistical language model on the output [10].

3. Research Methodology

3.1. Thai-Writing Structure

Thai language contained 72 characters and 4 levels of structure as the followings: Level 1 was 4 tone marks including เอก, โท, ตรี, จัตวา, and การันต์; Level 2 was the lower line from the first one consisting of 7 tone marks: ไม้หันอากาศ, สระอิ, สระอี, สระอึ, สระอือ, ไม้ไต่คู้, and หยาดน้ำค้าง; Level 3 was lower line from the second one and was the major line of Thai language containing 44 consonants and ฤ, ฦ, สระอะ, สระอา, สระอำ, สระเอ, สระแอ, สระโอ, สระไอ, สระใอ, etc.; Level 4 was the lowest line consisting of 3 สระ: สระอุ, สระอู, and พินทุอิ [1]. For that reason, the writing level must usually start from Level 3; meanwhile, Level 2 and 4 would be next (if needed). However, when Level 2 was written on, Level 4 would normally not be needed. In return, when Level 4 was written on, Level 2 would be excluded. Level 1 would be written on as the last (if needed).

Figure 1: Thai-Writing Structure: Pra-Tet-Yee-Poon (Japan)

Thai-writing structure illustrated in Figure 1 was the reason pushing the researchers to find out the writing structure of the word in the dictionary, as demonstrated in Figure 2.

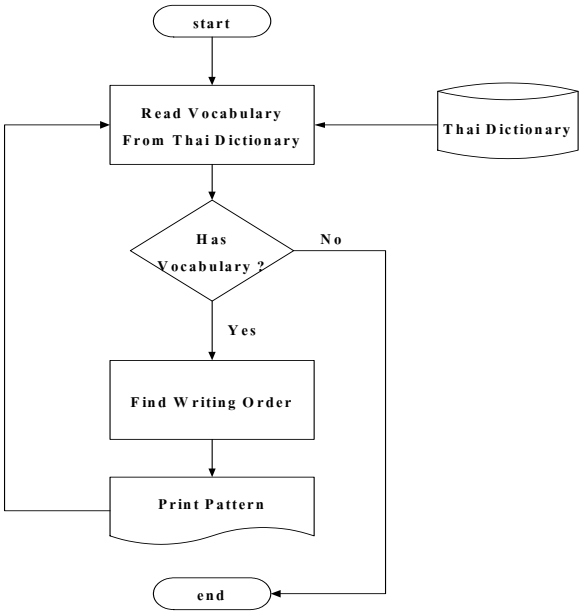

Figure 2: Steps of Finding Thai-Writing Structure Based on Thai dictionary Matching

3.2. Word Segmentation

After finding Thai-writing structure as required, there would be the deletion of the repeated words in order to gain the smallest number of the structures. After that, the leftover structures would be used for the word segmentation following the procedure in Figure 3.

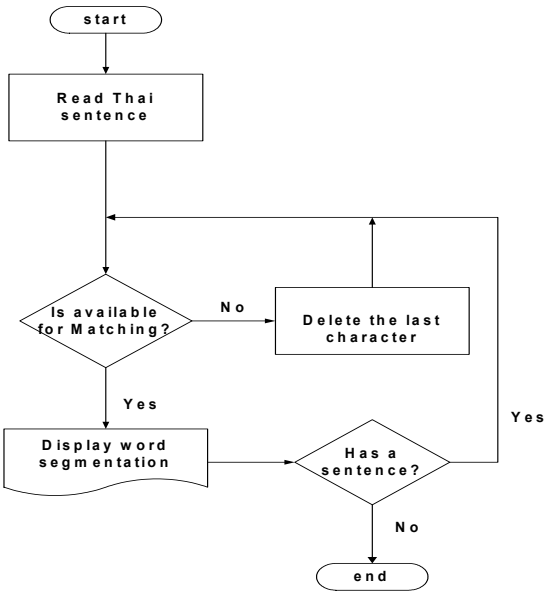

Figure 3: Steps of Word Segmentation Matching with Thai-Writing Structure

From Figure 3, the procedure could be clearly explained through the demonstration of word deletion in the sentence "เขาหาเพื่อน" as presented in Table 1.

Table 1: Demonstration of Word Segmentation Matching with the Found Structure

Sentence	Sentence Structure	Available for Matching	After deleting the latter consonant	Segmented word
เขาหาเพื่อน	3-3-3-3-3-3-3-2-1-3-3	Unavailable	เขาหาเพื่อ / (น)	
เขาหาเพื่อ	3-3-3-3-3-3-3-2-1-3	Unavailable	เขาหาเพื่ / (อน)	
เขาหาเพื่	3-3-3-3-3-3-3-2-1	Unavailable	เขาหา / (เพื่อน)	
เขาหา	3-3-3-3-3	Unavailable	เขา / (หาเพื่อน)	
เขา	3-3-3	Available	หาเพื่อน	เขา
หาเพื่อน	3-3-3-3-2-1-3-3	Unavailable	หาเพื่อ / (น)	
หาเพื่อ	3-3-3-3-2-1-3	Unavailable	หาเพื่ / (อน)	
หาเพื่	3-3-3-3-2-1	Unavailable	หา / (เพื่อน)	
หา	3-3	Available	หา / (เพื่อน)	หา
เพื่อน	3-3-2-1-3-3	Available		เพื่อน

4. Research Outcome

For the performance test of word segmentation, this research applied Confusion Matrix [9], so that the result could be shown in the table (Table 2).

Table 2: Demonstration of Confusion Matrix from the Performance Test of Word Segmentation through Thai-Writing Structure Matching

Word Status in Dictionary	Status of Word Segmentation after the Processing		
	Correct	Incorrect	Totals
Available	93,294	5,703	98,997
Unavailable	213	790	1,003
Totals	93,507	6,493	100,000

Based on Table 2, it could be explained as the followings: number of word in the dictionary that the program could correctly carried out word segmentation was 93,294; number of word in the dictionary that the program could not complete word segmentation was 5,703; number of word unavailable in the dictionary that the program could correctly complete word segmentation was 213; and number of the word unavailable in the dictionary that the program could not finish word segmentation was 1,003.

5. Conclusion

Thai-Word Segmentation through Thai-Writing Structure Matching was conducted by creating the program based on the steps proposed in this research. According to the performance test of the word segmentation program, it was revealed that the percentage of the word-segmentation accuracy was 94, in which it indicated that Word Segmentation through Thai-Writing Structure Matching, resulted from the writing structure of the word in the dictionary, could practically be applied into word segmentation for processing either Thai sentences or words, decreasing number of the word-segmentation rules, sparing the space of the storage, as well as reducing the time-duration of the word-segmentation processing.

6. Acknowledgements

I would like to thank Software and Laboratory in Mahasarakham University.Many thanks to Miss Piyarat Hematulin for providing Copus of testing.

7. References

[1] Panupong Vijin. 1981, **Thai Structure : Grammatical system**. Ramkhamhang University, Bangkok, Thailand.

[2] Thairatananond Yupin. 1981. **Towards the design of a Thai text syllable analyzer**. Master Thesis. Asian Institute of Technology, Thailand.

[3] Charnyapornpong Surin, 1983. **A thai Syllable Separation Algorithm**. Master Thesis. Asian Institute of Technology, Thailand.

[4] PuVaravarn Yuen and Imarrom Wiwat, 1986. **Thai Syllable Separater by Dictionary.** The report of the 9th International Conference on Electrical Engineering. Thailand.

[5] Kawtrakul Asanee Thumkanon Chalathip and Seriburi Sapon, 1995. **A Statistical Approach to Thai Word Filtering**. In Proceedings of the Symposium on Natural Language Processing in Thailand'95.

[6] Charoenpornsawat Paisarn, 2003 SWATH: **Thai Word Segmentation Program**.
http://www.cs.cmu.edu/~paisarn/software.htm

[7] Aroonmanakun Wirote and Rivepiboon Wanchai, 2004. **A Unified Model of Thai Romanization and Word Segmentation**. PACLIC 18, December 8th-10th, 2004, Waseda University, Tokyo, Japan.

[8] Issara Thienlikit, Chai Wutiwiwatchai, Sadaoki Furui, 2004. **Language Model Construction for Thai LVCSR**. Tokyo Institute of Technology, Japan. NECTEC of Thailand.

[9] Lorchirachunkul Vichit and Jitawech Jirawarn, 2004. **Application of data mining, and automobile insurance** : Documentation Research and Applied Statistics at the national level, Merchant Court Holtel, Bangkok, Thailand.

[10] Woszczyna Monika, Charoenpornsawat Paisarn and Schultz Tanja, 2006. **Spontaneous Thai Speech Recognition**. Multimodal Technologies, Inc. Carnegie Mellon University, Australia. In The Ninth International Conference on Spoken Language Processing.