Thai Word Segmentation with Hidden Markov Model and Decision Tree

Poramin Bheganan, Richi Nayak, and Yue Xu

Faculty of Science and Technology, Queensland University of Technology, Australia
poramin.bheganan@student.qut.edu.au, {r.nayak,y.xu}@qut.edu.au

Abstract. The Thai written language is one of the languages that does not have word boundaries. In order to discover the meaning of the document, all texts must be separated into syllables, words, sentences, and paragraphs. This paper develops a novel method to segment the Thai text by combining a non-dictionary based technique with a dictionary-based technique. This method first applies the Thai language grammar rules to the text for identifying syllables. The hidden Markov model is then used for merging possible syllables into words. The identified words are verified with a lexical dictionary and a decision tree is employed to discover the words unidentified by the lexical dictionary. Documents used in the litigation process of Thai court proceedings have been used in experiments. The results which are segmented words, obtained by the proposed method outperform the results obtained by other existing methods.

Keywords: Hidden Markov Model, Thai Word segmentation, Decision tree.

1 Introduction

In civil court proceedings, judges need to read the documents submitted for a litigation case by litigants to settle the points in dispute by finding facts and relevant information. These documents, called plaint and defend, lodged by the litigants, referring to plaintiff and defendant respectively are approximately 20 – 45 pages long. Judges read and analyse these documents to find points in dispute for which the plaintiff and defendant do not agree and take action. There are approximately 13,000 cases in the civil court of Thailand to be proceeded in each year and there are only 100 judges responsible for these cases [4]. It is an overwhelming task for a judge to read the documents, identify the points in dispute, and reach a decision. With the recent advancements in computing resources and data mining techniques, it is feasible to automate some of these tasks for judges. However, the first step is computationally processing Thai documents.

The computational processing of Thai written language is different and more complicated than the English language. Thai text does not have word boundaries and there is no delimiter to separate two sentences. The spaces found in documents are for making the readers comfortable with reading the passages rather than separating words or sentences. The task includes identification of syllables, words and then sentences. Some research has been done in Thai word segmentation [1, 9, 12] but a

T. Theeramunkong et al. (Eds.): PAKDD 2009, LNAI 5476, pp. 74–85, 2009.

higher accuracy can still be achieved. Most of the conventional Thai word segmentation methods use a dictionary to segment the texts [1, 5, 9]. If the dictionary does not contain the matching word, it leads to error. There is some research that uses a non-dictionary word segmentation method using decision trees [12]. The decision tree is trained to learn the basic grammar rules and to represent the principles or general rules about the interested information [12]. Similarly, the work of [9] used the decision tree to extract Thai words from a number of Thai corpora to build a corpus-based Thai dictionary. However, no grammar rules were used.

To increase the accuracy of the dictionary-based methods, there must be a sophisticated process to accurately segment the Thai texts into words. There can be ambiguous words formed by combining possible syllables. There are some shortcomings with the sole use of the decision tree for Thai language processing as the tree needs to use the grammar rules to construct it. If the grammar-rules are incorrectly defined, especially for the unknown words or words derived from other languages, the tree cannot generate suitable output. On the other hand, a dictionary-based word segmentation method needs no grammar rules but it needs all words to exist in the dictionary. If the potential words are not in the dictionary, the method cannot generate the output either.

This paper proposes a novel method of Thai word segmentation by combining a probability-based model using the six-state left-right Hidden Markov model (HMM) with the dictionary-based model using the decision tree. The HMM model is used to extract possible words combining several possibilities of syllables. If the probability-based model is used alone, the quality of the word segmentation heavily relies on the quality and quantity of the training data set. After words are formed as the output of HMM, the TCL's lexicon dictionary [11] is used to verify those words. The identified words are tagged according to their part of speech using this dictionary. If a word formulated from HMM cannot be identified by the dictionary, it is sent as a group of syllables to the decision tree to be identified as a possible word. The decision tree is trained to learn the patterns of combining syllables into words, by using the words that exist in the lexicon dictionary as a training set.

In this paper, we study the use of the combination of two different word segmentation techniques. We compared the results obtained by the proposed method with the results obtained by using the decision tree only, by using the six-state left-right HMM only, by using the three-state left-right HMM only, and by using the lexicon dictionary only. The proposed method significantly outperforms all other methods. The results show improvement in precision and recall for Thai language processing when applying the proposed method to a real-world data set.

The novel contribution of this paper is to develop a Thai word segmentation and processing method which combines the benefit of probabilistic and dictionary-based word segmentation together.

2 Problem Definition: Automatic Thai Language Processing

Thai characters are members of the Brahmic family of characters descended and modified from Brahmi [13]. The Thai written language does not use the conjunct

consonant mechanism and independent vowel letters that are found in most Brahmic writing [13]. This makes Thai language processing more difficult [1, 9, 10, 12].

There are 44 consonants and 32 vowel forms which can be written in front of, on top of, at the bottom of, and at the end of the consonants. In addition, four different tonal marks and special characters are used to form a Thai word. The smallest unit of Thai word is defined as a syllable which consists of at least one consonant with or without vowel(s). There is no space to separate among words or even sentences. Figure 1 shows an example of Thai text containing three paragraphs, four sentences, 211 words, and 282 syllables.

การนำรถยนต์คันที่เช่าซื้อไปขาย ถ้าโจทก์แจ้งให้จำเลยทราบก่อน จำเลยก็จะซื้อคืนหรือหาผู้ซื้อ ได้ราคาสูงกว่าที่โจทก์ขาย เพราะรถยนต์คันที่โจทก์ยึดคืน อยู่ในสภาพเรียบร้อย ใช้การได้ดีและมีอุปกรณ์ตกแต่งครบถ้วนอยู่ในสภาพใหม่ การขายของโจทก์ที่ขายได้ราคาเพียง ๑๘๖,๐๐๐ บาท จึงไม่สุจริตและไม่ถูกต้อง

ตามฟ้องของโจทก์ โจทก์เรียกค่าขาดประโยชน์ โดยกล่าวอ้างว่าสามารถนำรถยนต์ที่เช่าซื้อให้บุคคลอื่นเช่าได้ไม่ต่ำกว่าเดือนละ ๗,๐๐๐ บาท ซึ่งโจทก์ขอคิดนับแต่วันที่ ๑๔ สิงหาคม ๒๕๓๑ อันเป็นวันผิดสัญญาถึงวันยึดรถคืน เป็นเวลา ๑๓ เดือน คิดเป็นค่าเสียหาย ๙๑,๐๐๐ บาท

จำเลยที่ ๒ ขอกราบเรียนว่า การคิดคำนวณค่าเสียหาย ค่าขาดประโยชน์ของโจทก์ ดังกล่าวในข้อนี้ เป็นการคิดโดยความคาดหมายของโจทก์เองทั้งสิ้น ตามวิสัยและพฤติการณ์ของพ่อค้า ซึ่งเป็นการเอาเปรียบผู้บริโภค ความจริงแล้ว รถยนต์คันที่เช่าซื้อไม่สามารถนำออกให้เช่าได้ในราคา ๗,๐๐๐ บาทต่อเดือน ตามที่โจทก์กล่าวอ้าง แต่จะได้ราคาต่ำกว่าที่โจทก์เรียกร้อง ซึ่งอย่างมากก็ไม่เกินเดือนละ ๒,๐๐๐ บาท แต่อย่างไรก็ตามจำเลยที่ ๒ ก็ไม่ต้องรับผิดในส่วนนี้

Fig. 1. An example of Thai paragraphs

Information processing of documents in Thai language is considered more difficult in comparison to English language. The reasons are manifold: (1) there is no simple way to determine where a syllable starts and ends; (2) once a symbol is recognised as a syllable, the problem becomes apparent as to where a word starts and ends; (3) the next problems is how words in a sentence can be segmented correctly according to their context; (4) what words make a sentence due to the absence of word boundary in a sentence; and (5) what sentences form a paragraph.

3 The Proposed Method Combining the HMM and Decision Tree Models

The proposed method of Thai word segmentation, as depicted in Figure 2, combines the techniques of probabilistic (that is, non-dictionary based) and decision-tree (that is, dictionary based) modeling. The integration of these two techniques allows overcoming of the shortcomings of each individual technique. Both are different in detail or at the level of the analysis which makes the integrated process and result accurate. The process starts from syllable segmentation in Phase 1 by applying the standard language rules. The process is then followed by word segmentation in Phase 2. The initial word segmentation phase uses the HMM based non-dictionary word segmentation technique. When the first set of identified words are extracted using the HMM, the dictionary-based word segmentation technique with the lexical dictionary [11] is used in revising and combining the possible list of words before identifying the sentence. Any syllables unidentified by the dictionary are processed by the decision tree to turn into a tagged word.

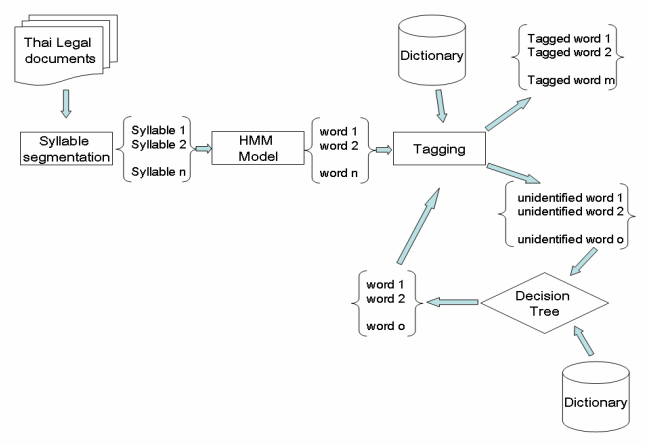

Fig. 2. Thai word segmentation process using HMM and Decision Tree

3.1 Phase 1: Syllable Segmentation

This task includes the analysis of input strings in a document to separate syllables. The process uses the standard Thai syllable formation rules to extract possible syllables, using the necessary grammar rules to suit all possible Thai syllables. The input string is analysed by applying the grammar rules for identifying the possible pattern of the syllables. The intermediate result may have some ambiguous syllables. Due to this, the segmented syllables are then combined and formulated for the possible words using the Hidden Markov model in the next phase.

This research implemented the syllable rules to formulate syllables from the Thai strings existing in a document. The syllable rules include fifteen (15) Thai consonant rules and Thirty Three (33) syllable structure rules. Before applying these syllable rules, a string is parsed though the Thai syllable structure to identify the syllables present in the string. This formula is able to recognise a syllable formed by the combinations of 44 consonants, 21 vowels, and four different tonal marks. It can also recognise the Thai vowels located in three levels: upper, middle and lower levels. The Thai syllable structure is shown in the equation (1):

$$S = [Vi] + Ci + [Cm] + [Vc] + [Tn] + [Vf] + [Cm] + [Cf] \quad (1)$$

S is a syllable. *Vi* is an initial vowel. *Ci* is initial consonant. *Cm* is initial-compound consonant. *Vc* is centre vowel. *Tn* is tone. *Vf* is final vowel. *Cf* is final consonant. Terms in brackets *[ ]* are optional.

The above formula stands true for every syllable formation. For example, a string "ซื้อขาย" consists of two syllables which are "ซื้อ" and "ขาย". The first syllable contains value for the following variables Ci ="ซ", Vc="ื", Tn="้", and Vf = "อ". When the next consonant "ข"is detected in the next position of string, the next possible syllable is started. The second syllable consists of Ci="ข", Vc="า", and Cf="ย".

In order to form a syllable, a string is parsed using the above formula. Then, the syllable rules are applied to verify a syllable. The composition of a syllable in Thai language is unambiguous and can be defined by a set of fifteen (15) Thai consonant rules and thirty three (33) syllable structure rules. These are some examples of rules that are implemented to process a string:

- The initial vowel must not be in the last position of a syllable.
- The initial compound consonant must not be in the last position of a syllable.
- A final vowel and the previous character have to be grouped into a same unit.
- A Final vowel must be in the last position of a syllable.
- A Garun (a particular vowel) must not be in the initial vowel position.
- An Upper vowel must not be in the initial vowel.
- An Upper vowel must not be in the final vowel.
- A Lower vowel must not be in the initial vowel.
- A Lower vowel must not be in the final vowel.
- There must be only one Tonal mark in a syllable.

3.2 Phase 2: Word Segmentation

The previous phase merges and groups the characters in the input string into syllables. Once the syllables are extracted, the next task is to identify combinations and form the possible words.

3.2.1 Word Segmentation Using HMM

The theory of a Hidden Markov Model (HMM) was derived from the Markov chains theory [2]. In an HMM, a state corresponds to each possibility and is not directly observable (i.e. it is hidden). Given a sequence of observation symbols, the HMM will be able to predict the probability of the sequence based on previous observations used to build the HMM. Given two sequences of observation symbols, the model communicates the probabilities of these two sequences. Therefore, it is possible to decide which one is better. In terms of applying HMM to the Thai word segmentation problem, if each state corresponds to a character position and the character values are the observation symbols, the HMM will be able to tell the best sequence of characters among a number of sequences given to the model.

In this paper, a left-right Hidden Markov model of six states is developed for Thai word segmentation. A six-state HMM model has been chosen after finding that the majority of words are the size of one to six syllables in the Thai TCL's lexicon dictionary developed by the Thai Computational Linguistics Laboratory [11]. In addition, the left-right model has been chosen because the Thai language is written from left to right [5]. In this model, a one-syllable word is discovered when the loop transition is at state 1, whereas, a multi-syllable word is discovered when the transition is from state 4 to state 1. A one-syllable word will pass only state 1, while a two-syllable word passes states 1 and 6. A three-syllable word passes states 1, 5, and 6, whereas, a four-syllable word passes states 1, 2, 5, and 6. A five-syllable long word passes states 1, 2, 3, 5, and 6, while, a six-syllable word passes states 1, 2, 3, 4, 5, 6.

In order to determine the most suitable words in the document, a Hidden Markov Model can be represented by equation 2:

$$\lambda = (A, B, \pi) \tag{2}$$

A is the state transition probability distribution matrix where a_{ij} is the transition probability from state i to state j. B is a matrix in which each row contains the probability distribution of observation symbols in particular state. π is the initial state distribution in the vector of i column, which indicates how likely the whole sequence of the trials start from state i.

With the model in Figure 3, the transition probability for the corresponding HMM model can be formulated as matrix A. This matrix shows the possibility of each word when in transit from one state to another state. Each word will pass a different group of states. The value will be calculated according to matrix A, B, and π.

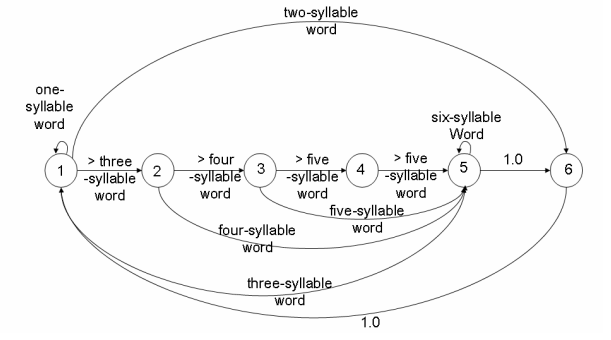

Fig. 3. A six-state left-right Hidden Markov model for the Thai word segmentation problem

$$A = \begin{bmatrix} a_{11} & a_{12} & 0.0 & 0.0 & a_{15} & a_{16} \\ a_{11} & a_{12} & 0.0 & 0.0 & a_{15} & a_{16} \\ 0.0 & 0.0 & 0.0 & a_{34} & a_{35} & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & a_{45} & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & a_{55} & 1.0 \\ 1.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \end{bmatrix}$$

Where a_{11}, a_{12}, a_{15}, a_{16}, a_{23}, a_{25}, a_{34}, a_{35}, a_{45}, and a_{55} are the ratios between the number of words within one syllable, with more than three syllables, with three syllables, with two syllables, with more than four syllables, with four syllables, with more than five syllables, with five syllables, with more than five syllables, and with more than six syllables respectively over the number of words in the training data set.

For initial state distribution π, every word must start from state 1. Therefore,

$$\pi = \begin{bmatrix} 1.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \end{bmatrix}$$

To formulate B as the matrix of 6X6, an observation symbol will be counted in a particular state when each possible word is parsed into HMM. The counted values will be divided by the number of times the specific state is passed through [8].

Once the matrices A, B and π are obtained, the HMM for the given data can be executed. If the HMM is used alone, the quality of the word segmentation heavily relies on the quality and quantity of the training data set. However, many words are inaccurately formed (as shown in experiments) and there remains a need to parse through the words with a dictionary to ensure the accuracy of the word and tag the words with its part-of-speech.

3.2.2 Dictionary-Based Word Segmentation Using Decision Tree

After each word is identified by the HMM, the lexicon dictionary is used to tag the part-of-speech to each word. If a word is not identified by the lexicon, it is processed by the decision tree. The decision tree is trained with various information that a word in the lexicon dictionary has. All the words that exist in the lexicon dictionary form the training set for decision tree modeling. During the iteration, the tree is branched according to the frequency of the occurrence of the words. The attributes in the decision tree include the following information gathered from each word:

1 Left Mutual information and Right Mutual Information

The mutual information of the co-occurrence of the word [3] is calculated as one of the attributes in the tree. The mutual information on the left and the right [9] will be calculated as followed in equations (3) and (4):

$$L_m(xyz) = \frac{p(xyz)}{p(x)\,p(yz)} \tag{3}$$

$$R_m(xyz) = \frac{p(yz)}{p(xy)\,p(z)} \tag{4}$$

where

- Lm is the left mutual information of the noticing string
- Rm is the right mutual information of the noticing string
- x is the leftmost syllable of xyz
- y is the middle syllable of xyz
- z is the rightmost syllable of xyz
- $p(\ )$ is the probability of the occurrence

2 Length of word

The number of syllables to form a word is counted as an attribute in the decision tree. The possible lengths of common Thai words are between one to six syllables.

3 Functional words

In the Thai language, there are many functional words such as “then” and “will”. The appearance of these functional words may cause the grouping of syllables to form a word wrongly [9]. Therefore, a function *Func(s)* is considered as an attribute to eliminate these functional words from the string pattern.

$$Func(s) = \begin{cases} 1 & \text{if string contains a functional word,} \\ 0 & \text{if otherwise.} \end{cases}$$

These attributes values are prepared for a total of 30,000 words that exist in the lexicon dictionary. The C4.5 learning algorithm [7] is used to train the decision tree with this data set. The unidentified words from the previous phase are tested with this trained decision tree using the patterns of combining syllables that exist in the words of training data.

4 The Proposed Method Combining the HMM and Decision Tree Models Empirical Analysis and Discussion

The proposed method starts by applying the Thai language grammar rules to identify the syllables. Syllables are combined into possible words with HMM. The lexicon dictionary is then used in verifying the words and the words that are not verified by the dictionary are checked in the decision tree for identification as suitable words.

Data Set: There are two sets of input data used in the experiment. The first set of data is the legal documents that need to be processed for word segmentation. These legal documents are collected from the archive centre of the Court of Justice, Thailand [4]. All documents are civil cases which were pleaded between the years 2000 to 2003. There are a total of 300 cases collected that are used in experiments. In each case, there are about 5 – 15 pages of the plaint document and about 10 - 20 pages of the defend document. Each case also contains about 5 – 10 pages of judgment verdict.

The second set of data is the TCL's Thai lexicon dictionary [11]. As stated earlier, there are 30000 unique words in the lexicon. The lexicon dictionary is used to verify and tag words from HMM. It is also used for training the decision tree.

Evaluation Criteria: Precision, recall, and F-measure are used as the evaluation criteria in experiments. Precision is the ratio of words retrieved compared to all of the words in the legal documents. Recall is the ratio of the relevant words retrieved compared to all relevant words in the legal documents. F-measure is the weighted harmonic mean of precision and recall.

10-fold CV experiments are used throughout the experiments and the average performance is reported along with standard deviation.

4.1 Empirical Analysis and Discussion

Results obtained by the proposed method are compared with the results obtained by the decision tree only, by the six-state left-right HMM, by the three-state left-right HMM and by lexicon only.

Five (5) sets of comparative experiments are conducted to evaluate the proposed method. They are as follows:

- The six-state left-right HMM in the proposed method is replaced with the three-state left-right HMM to check whether the six-state left-right HMM is suitable for Thai processing. Results of the three-state left-right HMM are integrated with the Lexicon and C 4.5 decision tree as described in our method.

- The six-state left-right HMM is used alone for the word segmentation problem. It means that the decision tree and dictionary are not used to further enhance the quality of word segmentation. The output words produced from the HMM become the final word set.
- The three-state left-right HMM is also used alone for the word segmentation problem. It means that the decision tree and dictionary are not used to further enhance the quality of word segmentation.
- The lexicon dictionary is used alone to identify the words. The process of this experiment is as follows. The Thai language grammar rules are used to identify the syllables, then syllables are segmented. The first combination a group of syllables form becomes the word for lexicon if this word exists. Even though this may not be the correct interpretation.
- The decision tree is used alone to identify the words. The process of this experiment is as follows.

The training dataset from each fold of 270 legal documents consists of an average of 534330 words for HMM. The decision tree training dataset consists of 30,000 records, each record representing a word in the lexicon dictionary. Each record is made of four (4) input attributes, namely Left mutual information, Right mutual information, length of words, and functional words, and one (1) binary output attribute that is a word formed or not. The test dataset in our experiments from each fold of 30 legal cases consists of an average of 57360 words.

Table 1 shows the numbers of words and syllables found in the data set. This table also includes the number of words identified by the proposed method and other methods used for comparison. The numbers of words identified by the lexicon are the highest, 677846 for training dataset, and 62413 in test dataset. It can also be observed that the numbers of words identified by the lexicon are even more than the number of words in the documents (that are 534330 in training dataset and 57360 in test dataset).

The reason is that when a group of syllables in a string is presented to the lexicon dictionary, it accepts the first possible combination of syllables found in the lexicon as a word. It ignores any other combination that may be formed as a word by adding more syllables next in the string. Consequently, most of the syllables in the input set matched with one-syllable words in the lexicon. It can also be observed from Table 2 that it has the least accuracy as a result.

Table 1. Numbers of words and syllables in the experiments

Method/ Information	Dataset	
	Training (270 documents)	Test (30 documents)
Total No. of words	534330	57360
Total No. of syllables	1870155	216821
No. of words identified by 6-state left-right HMM only	341139	37159
No. of words identified by 6-state left-right HMM + Lexicon + Decision Tree (the proposed method)	348623	38862
No. of words identified by 3- state left-right HMM only	316445	36881
No. of words identified by 3-state left-right HMM + Lexicon + Decision Tree	306285	31285
No. of words identified by Lexicon Only	677846	62413
No. of words identified by Decision Tree Only	356173	39458

Table 2. Precision, Recall, and F-score from the experiments

Method	Precision		Recall		F-Score		Elapsed time	
	Training	Test	Training	Test	Training	Test	Training	Test
6-state HMM + Lexicon + DT	82.65% (0.530)	79.96% (1.272)	53.92% (1.348)	54.17% (0.987)	65.26% (0.959)	64.59% (1.012)	56.51s	55.16s
6-state HMM only	65.51% (0.707)	64.34% (1.254)	41.83% (0.871)	41.68% (0.896)	51.05% (0.775)	50.59% (0.609)	53.80s	52.42s
3-State HMM + Lexicon +DT	59.81% (0.800)	59.72% (0.919)	34.28% (1.413)	32.57% (0.769)	43.58% (1.139)	42.15% (0.576)	35.18s	33.96s
3-state HMM only	50.73% (1.305)	50.06% (1.194)	30.04% (1.429)	32.18% (1.183)	37.74% (1.093)	39.18% (0.908)	32.07s	31.23s
DT only	73.03% (1.009)	72.84% (0.847)	48.68 % (1.335)	50.11% (0.873)	58.42% (0.968)	59.37% (0.721)	10.06s	9.95s
Lexicon only	28.83% (0.736)	24.73% (0.927)	36.58% (0.751)	26.90% (1.007)	32.25% (0.501)	25.77% (0.862)	10.15s	10.03s

* Values in parentheses are Standard Deviation.

The number of words identified by the six-state left-right HMM only is the lowest (34119 in training dataset and 37159 in test dataset). As HMM is a statistic-based method, the quality of the output depends upon the quality of training dataset. The number of identified words can be improved if there are more training datasets with good quality. There are approximately 20,000 words that HMM is unable to correctly identify. It can be seen that when we included the lexicon dictionary and decision tree model to further identify the words produced from HMM, the number is increased to 348623 for the training data set and 38862 for the test data set.

Besides, the experiment results also reveal that numbers of identified words by decision tree (356173 in training and 39458 in test dataset) are more in quantity than when only using HMM. This is because the DT is trained on the entire lexicon words. However, it can be seen that their accuracy is lower in comparison to our results.

Table 2 reports the precision, recall, F-score measures, and elapsed times of the experiments. The precision, recall, F-Score values gained by our proposed method are the highest among other methods, whereas those of the lexicon dictionary only method give the lowest performance among other methods. This is because, when the syllables are compared with the lexicon, most syllables were matched with one-syllable words in the dictionary. The improvement is 55.23% for precision value and 27.27% for recall value for the test data set.

The precision and recall ratio on the test set is 63.34% and 41.68% respectively when the six-state left-right HMM alone is used to identify the words. The improvement is 15.62% for precision value and 12.49% for recall value when the outputs by HMM are combined with lexicon and DT.

The experiment also reports that if we change the HMM from six-state to be three-state left-right HMM [6] without using the decision tree, the precision ratio is dropped to 50.06% and the recall ratio is dropped to 32.18% in the test dataset. This number reveals that Thai words tend to be better segmented by using the six-state left-right HMM rather than using the three-state left-right HMM. When we combine the outputs

of the three-state left-right HMM with the lexicon and the decision tree part, the precision ratio is increased to 59.72% and the recall ratio is slightly increased to 32.57%. It is, however, significantly lower than the proposed method.

Considering elapsed times in experiments, the proposed method consumed the maximum time of 56.51 seconds on training and 55.16 seconds on testing whereas the decision tree took the least time. The training dataset used only 10.06 seconds and the testing consumed 9.95 seconds. This is because the proposed method consumes the first period of time to retrieve the intermediate results using the six-state left-right HMM. The remaining time consumed are counted when unknown words from the HMM is parsed to the lexicon and the rest of unsegmented words parsed to the decision tree. The time consumed in the experiment on the decision tree only is the minimal since there is no intermediate result to pass to another process.

It can be ascertained based on the results reported in Table 1 and Table 2 that the proposed method produces the best results as it utilizes the strength of each individual method used. The words unidentified by the HMM only are further processed using the decision tree and the overall accuracy is enhanced.

5 Conclusion

In this paper, we combine the techniques of probabilistic method word segmentation as well as dictionary-based word segmentation together. Both of the techniques provide different advantages and drawbacks. The integration of these two techniques allows overcoming the shortcomings of each individual technique. Most are different in detail or the level of the analysis, which makes the results different. The initial word segmentation technique is developed by using the non-dictionary word segmentation. When the first set of words is extracted, the dictionary-based word segmentation is used to help revise and combine the possible list of words before identifying the sentence.

The proposed method is extensively evaluated against other methods. Results show that the proposed method significantly outperforms all other methods. The word segmentation accuracy (in terms of precision and recall ratio) can be further improved. Firstly, more training and testing data will help with improving the accuracy of the experiment. The data set used in this paper was collected from real legal cases. The number of cases lodged was limited in the specified period. If there were more cases, the accuracy would have been increased. Secondly, the nature of the words in these data sets is different from the nature of the words that exist in the lexicon dictionary used. The domain of collected data is in the legal domain, while, the dictionary used in refining the words is the general lexicon. Therefore, some legal- specific terms might not be present in the dictionary.

Future work includes experimentation with some more datasets and the construction of the domain-specific vocabulary that can be used with the dictionary, to help with segmenting the more suitable words according to the domain. We also plan to use the tagged words to discover the relationships of particular paragraphs in the plaint and the defend.

References

1. Aroonmanakul, W.: Collocation and Thai Word Segmentation. In: Joint International Conference of SNLP-Oriental COCOSDA, Thailand, pp. 68–75 (2002)
2. Christen, P., Churches, T., Hegland, M., Lim, K., Nielsen, O.M., Roberts, S., Zhu, J.: High-Performance Computing Techniques for Record Linkage. In: Australian Health Outcomes Conference, Canberra, Australia, pp.1–14 (2002)
3. Church, K.W., Robert, L., Mark, L.Y.: A Status Report on ACL/DCL. In: 7th Annual Conference of the UW Centre New OED and Text Research: Using Corpora, Canada, pp. 84—91 (1991)
4. Civil court of Thailand, http://www.cvcourt.com
5. Kawtrakul, A., Thumkanon, C., Poovarawan, Y., Varasrai, P., Suktarachan, M.: Automatic Thai Unknown Word Recognition. In: Natural Language Processing Pacific Rim Symposium, Phuket, Thailand, pp. 341–346 (1997)
6. Nagata, M.: Context-based spelling correction for Japanese OCR. In: 16th conference on Computational linguistics, New Jersey, USA, pp. 806–811 (1996)
7. Quinlan, R.: C4.5: Programs for Machine Learning. Morgan Kaufman, USA (1993)
8. Rabiner, L.R.: A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. IEEE 77(2), 257–285 (1989)
9. Sornlertlamvanich, V., Potipiti, T., Charoenporn, T.: Automatic corpus-based Thai word extraction with the C4.5 learning algorithm. In: 18th conference on Computational linguistics. Saarbrücken, Germany, pp. 802–807 (2000)
10. Sudprasert, S., Kawtrakul, A.: Thai word segmentation based on Global and Local Unsupervised learning. In: NCSEC, Chonburi, Thailand (2003)
11. Thai Computational Linguistics Laboratory.: TCL's Computational Lexicon, http://www.tcllab.org/tcllex/
12. Theeramunkong, T., Usanavasin, S.: Non-dictionary-based Thai word segmentation using decision trees. In: The first international conference on Human language technology research, New Jersey, USA, pp. 1–5 (2001)
13. Unicode Consortium.: The Unicode Standard 4.0: Southeast Asian Scripts. Addison Westley, California (2004)