# 論文 / 著書情報
# Article / Book Information

| | |
|---|---|
| 論題(和文) | |
| Title(English) | Automatic Dictionary Generation for Thai LVCSR with the C4.5 Learning Algorithm |
| 著者(和文) | 古井 貞熙 |
| Authors(English) | Markpong Jongtaveesataporn, Chai Wutiwiwatchai, Sadaoki Furui |
| 出典(和文) | 日本音響学会講演論文集, Vol. , No. , pp. 59-60 |
| Citation(English) | , Vol. , No. , pp. 59-60 |
| 発行日 / Pub. date | 2006, 9 |

# Automatic Dictionary Generation for Thai LVCSR with the C4.5 Learning Algorithm*

◎ Markpong Jongtaveesataporn[1)], Chai Wutiwiwatchai[2)], Sadaoki Furui[1)]
1) Tokyo Institute of Technology 2) NECTEC of Thailand

## 1. Introduction

Large vocabulary continuous speech recognition (LVCSR) is one of the basic applications of speech technology. High-quality acoustic models as well as large-scale language models are necessary to create an LVCSR system. In written Thai, there is no marker for separating words and even the definition of a word is ambiguous. Words are usually separated into two types: 1) simple words, consisting of one or more syllables, each of which may have a meaning, but the meaning of the word is not related to the meaning of any one syllable; and 2) compound words, composed of two or more simple words. With this criterion, the determination of word boundary on the real given data is still ambiguous and may not be consistent for different persons. Therefore, building a Thai text corpus is not a trivial task, and an efficient word segmentation approach is required. Generally, the dictionary-based word segmentation is employed to automatically segment a text corpus. However, the performance of this approach depends heavily on the dictionary's coverage of vocabulary in the target corpus. In this paper, an automatic corpus segmentation method using no dictionary is proposed. The C4.5 learning algorithm is employed here to help arrange a text corpus for building a language model.

## 2. Previous works

There have been a few studies on automatic language modeling for Thai LVCSR. An automatic data-driven language modeling approach was presented by Thienlikit [1]. A text corpus was first segmented into pseudo-morpheme units, and then these units were merged on the condition of forward and reverse bigram statistics. An approach to generate a dictionary for Thai LVCSR by using multiple segmentation approaches was also introduced [2]. The results of dictionary-based and pseudo-morpheme segmentations were used together to evaluate better segmentation results. However, this technique still relied on the performance of dictionary-based word segmentation.

## 3. Proposed method

### 3.1 Pseudo-morpheme segmentation (PMSEG)

The term "pseudo-morpheme" is defined here to represent a written form of a syllable-like unit in order to avoid confusion with the definition of syllable in the sound system. Table 1 shows the relation between words, pseudo-morphemes, and syllables.

Table 1: Word, pseudo-morpheme, and syllable

| Word | Pseudo-morpheme (pronunciation for each pseudo-morpheme) | Syllables (pronunciation) |
|---|---|---|
| หน้าต่าง | หน้า-ต่าง (naa2 taang1) | หน้า-ต่าง (naa2 taang1) |
| วิทยา | วิท-ยา (wit3 jaaz0) | วิด-ทะ-ยา (wit3 ta3 jaaz0) |
| พลศึกษา | พล-ศึก-ษา (phon0 svk1 saaz4) | พะ-ละ-สึก-สา (pha3 la3 svk1 saaz4) |

One significant problem for Thai word segmentation is caused by the lack of a clear definition of a word. The segmentation problem can be solved by pseudo-morpheme segmentation, based on the fact that pseudo-morphemes are more well-defined units and can be more consistently analyzed than words. Pseudo-morpheme segmentation is therefore more reliable. Once a number of pseudo-morpheme patterns are defined, every input string can be matched to these patterns. Trigram statistics of pseudo-morphemes can be used to determine the best segmentation result.

Table 1 shows that pseudo-morphemes are similar to syllables, but some pseudo-morphemes may contain more than one syllable, or have many pronunciations.

### 3.2 The C4.5 Learning Algorithm

A decision tree is a tree that classifies instances. Leaves of a decision tree represent classifications of given instances, and branches represent conjunctions of attributes that lead to those classifications. An attribute describes a characteristic of an instance. Therefore, instances can be classified based on the results of successive attribute tests. In this paper, we employ the C4.5 decision tree induction program [3] as the learning algorithm.

The C4.5 is a greedy algorithm that chooses the next attribute to test based on the information gain associated with this attribute. At each step of learning procedure, the information gain that would be obtained by using each unused attribute to classify the particular set of instances will be calculated. The attribute with the greatest information gain will be selected as a new branch. Branches will be inserted until all instances in the training set are classified. The C4.5 tries to eliminate overfitting data by pruning the tree that has been created. Pruning involves replacing a subtree with a classification to see whether this reduces the expected error rate.. This pruning technique helps make the decision tree work with novel data better.

### 3.3 Applying C4.5 to pseudo-morpheme unit merging

Although pseudo-morpheme segmentation accuracy is very high, the pseudo-morpheme is not suitable to be used as a unit for LVCSR system due to several reasons. Firstly, pseudo-morpheme segmentation produces many short lexical units which generate high acoustic confusion. In addition, the span of an N-gram language model is significantly smaller since the unit is short. Therefore, the language model cannot perform efficiently. Moreover, since some pseudo-morphemes may have variation in pronunciation depending on the context as we can observe from the table 1, the G2P conversion process may not give a correct pronunciation. Hence, some pseudo-morpheme units should be merged together and added to the lexicon as a new lexical unit.

We utilize the C4.5 decision tree to consider whether any pairs of consecutive pseudo-morphemes should be merged together. In order to train the decision tree, a text corpus is segmented into words manually while it is also segmented into pseudo-morphemes. The attributes for each pair of consecutive pseudo-morphemes are computed and they are also compared with the manually segmented text corpus to see whether they are parts of the same word. If they are parts of the same word, it indicates that these two pseudo-morphemes should be merged together. With this approach, the goal attribute, i.e. "merge" and "notmerge" can be retrieved.

The following attributes are used for the learning algorithm.

---

* C4.5学習アルゴリズムを用いたタイ語大語彙連続音声認識のための辞書の自動作成
マッカポン・チョンタウィーサターポーン[1)]、チャイ・ウッティウィワッチャイ[2)]、古井貞熙[1)]
1) 東京工業大学 2) NECTEC of Thailand

1. Geometric average of direct and reverse bi-grams

$$M(w_i, w_{i+1}) = \sqrt{P_f(w_{i+1} \mid w_i) P_r(w_i \mid w_{i+1})}$$

The value of M varies from 0 to 1. The higher the value, the higher the probability these two pseudo-morphemes are found together.

2. Whether or not the first consonants of two consecutive pseudo-morphemes are the same. We include this attribute because, in Thai, many compound words happen to have the same consonant for each starting syllable.

3. Length of combined pseudo-morphemes

4. Length of the preceding pseudo-morpheme

5. Length of the following pseudo-morpheme

6. Number of pseudo-morphemes from the last proper nouns' prefixes

7. Number of pseudo-morphemes from the last preposition

8. Number of pseudo-morphemes from the last nominalizer

9. Number of pseudo-morphemes from the last conjunction

10. The pseudo-morphemes that are being considered contain words identifying numerical value or not.

In order to calculate the attributes 6, 7, 8, 9, and 10, the words of proper nouns' prefixes, preposition, nominalizer, conjunction, and numerical value in Thai are provided in advance. With all these attributes and a goal attribute, a decision tree is then constructed.

Another text corpus is used to test the decision tree. It is first segmented into pseudo-morphemes. Then, all attributes will be calculated for each pair of consecutive pseudo-morphemes by the same process done in the training set. Finally, the decision tree is used to classify "merge" and "notmerge" pseudo-morphemes. All pairs of pseudo-morphemes with the "merge" attribute will then be compounded together and a newly modified text corpus can be achieved. Each pair of consecutive tokens in this new corpus will be merged again through an iterative procedure.

## 4. Experiments

### 4.1 Experimental conditions

A 45MB text corpus collected from a Thai newspaper website was used for training the C4.5 decision tree. All 691,527 sentences were segmented into words manually.

A gender-dependent acoustic model (AM) of 1000-tied-state triphones with 8 Gaussian mixtures was trained by a phonetically-balanced speech corpus with 18 male-speakers' voices using HTK Toolkit. 25-dimensional feature vectors consisted of 12 MFCCs, their delta, and a delta energy. Tone information was not used. A 62MB text corpus was gathered from a Thai newspaper website in several specific columns. It was first segmented into pseudo-morphemes and pseudo-morphemes were merged together iteratively based on the decision tree. The text corpus in each round was used to train LMs by the CMU SLM toolkit. JULIUS was used as a speech decoder. Evaluation speech data consisted of 1,000 utterances from 5 male speakers.

A pronunciation for each entry in the dictionary was obtained by the automatic G2P conversion. In order to avoid manual work, the words that failed in the G2P conversion process were excluded from the dictionary. Then the sentences containing excluded words were also removed from the corpus.

### 4.2 Results

In order to compare LM perplexities of the different versions of the test set with different text lengths (the number of units in the text), a normalized perplexity is used:

$$PP^* = PP^{\frac{N_b}{N}},$$

where $N_b$ is the length of the test set and $N$ is the length of the original text made by pseudo-morpheme segmentation.

To compare the performance of ASR, character error rate (CER) is chosen since the lexical entries are different for each method. As shown in figure 1, CER decreases iteratively while PP* also roughly declines. The best CER of 11.47% can be obtained from iteration 10. In contrast, the vocabulary size grows rapidly as shown in figure 2.

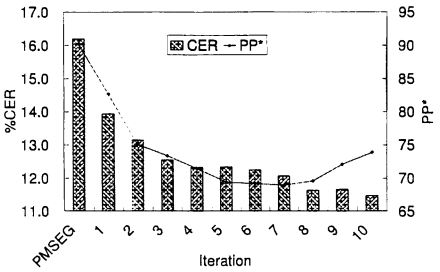


Figure 1: Relationship of iteration, CER and perplexity including CER and perplexity of PMSEG

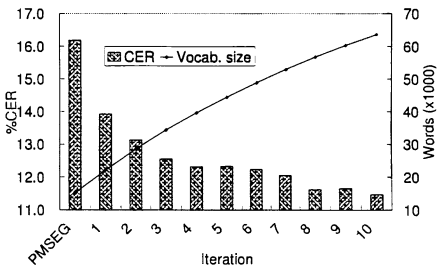


Figure 2: Relationship of iteration, CER and vocabulary size including CER and vocabulary size of PMSEG

Table 3 shows that the performance of ASR built from the best iteration is improved enormously from the system built from the base PMSEG. Furthermore, it outperforms the system developed from SWATH, a Thai dictionary-based word segmentation tool, in the aspect of CER, PP*, and OOV rate. However, no dictionary is required in our proposed method.

Table 3: CER, PP*, OOV rate, and Vocabulary size comparison between PMSEG, SWATH, and the best iteration

| Method | CER | PP* | %OOV | Vocab. size |
|---|---|---|---|---|
| PMSEG | 16.19 | 90.28 | 0.18 | 15286 |
| SWATH | 12.25 | 83.38 | 0.64 | 25806 |
| Best iter. | 11.47 | 73.88 | 0.08 | 63755 |

## 5. Conclusion

This paper has introduced an automatic approach to build a language model for Thai LVCSR by using the C4.5 learning algorithm. The performance of ASR built by this proposed method is quite good as indicated by CER. Nevertheless, in higher iterations, lexical units are very long and the vocabulary sizes are greatly increased. Longer units mean that they are more specific words. Therefore, some modification should be done in order to help stop the increment of vocabulary size while maintaining the performance of ASR.


## 6. Acknowledgements

We would like to thank Dr. Wirote Aroonmanakun for allowing us to use his pseudo-morpheme segmentation tool.


## 7. References


[1] I. Thienlikit, C. Wutiwiwatchai, S. Furui, "Language Model Construction for Thai LVCSR", Autumn Meeting of Acoustic Society of Japan, 3-1-11, pp.131-132, Sep, 2004.

[2] M. Jongtaveesataporn, C. Wutiwiwatchai, S. Furui, "Dictionary Generation Using Multiple Segmentation Approaches for Thai LVCSR", Spring Meeting of Acoustic Society of Japan, 2-1-8, pp. 85-86, Mar, 2006.

[3] J.R. Quinlan, "C4.5 Programs for Machine Learning", Morgan Publishers San Mated, California, 302p.