論文 / 著書情報
Article / Book Information

論題(和文)	
Title(English)	Dictionary Generation Using Multiple Segmentation Approaches for Thai LVCSR
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2006年春季講演論文集, Vol. , No. , pp. 85-86
Citation(English)	, Vol. , No. , pp. 85-86
発行日 / Pub. date	2006, 3

Dictionary Generation Using Multiple Segmentation Approaches for Thai LVCSR*

○ 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 a basis of various speech technology applications. In order to build an LVCSR system, high-accuracy acoustic models as well as large-scale language models are essential. In writing Thai, there is no marker for separating sentences and words, and even word definition is ambiguous. Therefore, building a text corpus is very difficult, and an efficient word segmentation approach is required. Although dictionary-based segmentation is effective, it causes serious problems if unknown words appear. This paper introduces a technique to automatically build the dictionary for Thai LVCSR based on a dictionary-based word segmentation approach using Thai pseudo-morpheme units.

2. Thai language

In Thai, no explicit marker is used to separate between sentences, phrases, and words. Sometimes a space is used to indicate end of phrases or sentences as shown in Figure 1. There are usually two types of words in Thai: 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 syllable; and 2) compound words, composed of two or more simple words.

พ.ต.ท.ทักษิณ ชินวัตร นายกรัฐมนตรี เปิดเผยถึงกรณีการชุมนุมประท้วงการเจรจา เขตการค้าเสรี (เอฟทีเอ) ไทย-สหรัฐฯ ครั้งที่ 6 วันที่ 9-13 ม.ค.นี้ ที่ จ.เชียงใหม่ และเรียกร้องให้นำรายละเอียดการเจรจาเข้าพิจารณาในสภาผู้แทนราษฎรก่อนว่า ไม่จำเป็น สภาไม่มีผู้เชี่ยวชาญ และไม่มีกฎหมายกำหนด

Figure 1: A Thai text example

3. Segmentation in Thai language

3.1 Dictionary-based segmentation

Existing Thai language processing systems mostly rely on the hand-coded dictionaries in acquiring the information about words. By employing several sophisticated techniques, such as maximal matching, part-of-speech n-grams, and dictionary-based segmentation, this method yields often good results. However, this approach has many disadvantages. First, it cannot efficiently deal with words that are not registered in the dictionary. Second, building a large scale dictionary is a very difficult and time-consuming task.

3.2 Pseudo-morpheme segmentation (PMSEG)

One big problem for Thai word segmentation is caused by the lack of clear definition of words. Aroonmanakun[1] solved the segmentation ambiguities problem by pseudo-morpheme segmentation, based on the fact that pseudo-morphemes are more well-defined units and more consistent in analysis than words. The 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.

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)

3.3 Proposed method

There have not been many reports on the study of Thai LVCSR. Automatic approaches for building Thai language model for LVCSR are expected to make future advances in Thai speech processing. A totally automatic approach was proposed by Thienlikit et al.[2]. However, it cannot be applied to a large text corpus, since the vocabulary size generated by this approach grows too fast and easily exceeds the limit of tools.

As mentioned above, the dictionary-based segmentation works quite well if all words in a sentence are in its dictionary. Errors occur when unknown words are included in a sentence. Table 2 shows an example of errors. ("-" indicates a word boundary.)

Table 2: Example of segmentation result

Perfect	-นาย-โรเบิร์ต-ผู้นำ-คอมมูนิตี้-เชื่อว่า-
Dictionary-based	-นา-ยโรเบิร์ต-ผู้นำ-คอ-มมูนิตี้-เชื่อว่า-
PMSEG	-นาย-โร-เบิร์ต-ผู้-นำ-คอม-มู-นิ-ตี้-เชื่อ-ว่า-
Meaning	Mr. Robert, the chief of the community, believes that

Since โรเบิร์ต (Robert) and คอมมูนิตี้ (community) are foreign words, they may not be included in the dictionary. The segmentation tool tries to find matching words and group unknown segments together. Then, นายโรเบิร์ต is divided into นา (field) + ยโรเบิร์ต (unknown) while คอมมูนิตี้ is segmented into คอ (neck) + มมูนิตี้ (unknown). These errors cause unpronounceable segments or weird pronunciation.

On the other hand, if the same sentence is processed by pseudo-morpheme segmentation, perfect segmentation can be made. Therefore, we utilize the benefits of both approaches in the following way:

Step 1: Apply dictionary-based segmentation. In this study, we use the SWATH[4] word segmentation tool developed by NECTEC of Thailand

Step 2: Apply pseudo-morpheme segmentation (PMSEG).

Step 3: Each word in the result that does not exist in the dictionary ("ยโรเบิร์ต" and "มมูนิตี้" in this case) is processed as follows:

(1) Find common word boundaries for each unknown word between dictionary-based and pseudo-morpheme segmentation methods and merge all tokens between a pair of common boundaries to form a new word. For example, the unknown word "ยโรเบิร์ต" (SWATH) has characters spanning in

---

* タイ語の LVCSR のための多重セグメンテーションによる辞書生成
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

"นาย-โร-เบิร์ต" by PMSEG. A common word boundary can be found before "นา"(SWATH) and "นาย"(PMSEG). The other one is located after "ยโรเบิร์ต"(SWATH) and "เบิร์ต"(PMSEG). Therefore, the token between these boundaries are merged to form "นายโรเบิร์ต". We named this method PM-SWATH1.

(2) Or find common word boundaries for each unknown word and replace the dictionary-based segmentation result by the pseudo-morpheme segmentation result between each pair of common word boundaries. For example, "นา-ยโรเบิร์ต" (SWATH) is replaced by "นาย-โร-เบิร์ต"(PMSEG). We name this method PM-SWATH2.

The final processed results of these approaches are:

PM-SWATH1: -นายโรเบิร์ต-ผู้นำ-คอมมูนิตี้-เชื่อว่า-

PM-SWATH2: -นาย-โร-เบิร์ต-ผู้นำ-คอม-มู-นิ-ตี้-เชื่อว่า-

With this approach, a dictionary for ASR with a better list of vocabularies in terms of word boundaries and generated pronunciation can be created.

4. Experiments

4.1 Experimental conditions

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. A 91MB text corpus was gathered from 53 months of Thai newspaper website in several specific columns and used to train LMs by the CMU SLM toolkit. Tone information was not used. JULIUS was used as a speech decoder. Evaluation speech data consisted of 1,000 utterances from 5 male speakers.

After the whole text corpus was segmented into words by SWATH, sentences that contain unknown words for SWATH were selected and processed by the proposed methods. A dictionary to be used by ASR was then enumerated from the list of words in the final processed text corpus.

A pronunciation for each entry in the dictionary was obtained from 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), we use a normalized perplexity:

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

where N_b is the length of the test set and N is the length of the original text made by pure SWATH word segmentation. The perplexities and OOV rate of various approaches are shown in Table 3.

Table 3: Normalized test-set perplexity and OOV rate comparison

Method	Normalized test-set PP	%OOV
SWATH	294.93	0.62%
PMSEG	302.08	0.18%
PM-SWATH1	257.35	0.09%
PM-SWATH2	256.58	0.08%

As shown in Table 3, the OOV rate of SWATH is considerably high comparing with that of PMSEG, while the perplexity of PMSEG is slightly higher than that of SWATH. When applying our proposed methods, we could successfully reduce the OOV rate significantly to 0.09% (PM-SWATH1) and 0.08% (PM-SWATH2), which were lower than those of SWATH and PMSEG. This is because our proposed methods can fix the incorrect word boundaries into better ones. In the same way, the perplexity also decreased from 294.93 to 257.35 and 256.58 for PM-SWATH1 and PM-SWATH2, respectively.

Since the vocabulary lists of each method were not the same, due to different segmentation techniques, we chose the character error rate (CER) as the measure for the comparison of ASR performance.

Table 4: CER comparison

Method	CER
SWATH	11.89
PMSEG	15.43
PM-SWATH1	11.68
PM-SWATH2	11.49

The best result of CER was obtained by PM-SWATH2. Both of CERs from our proposed methods were lower than SWATH and PMSEG. However, the improved performance gained by them was not much.

Table 5: Vocabulary size

Method	Vocabulary size
SWATH	34244
PMSEG	17788
PM-SWATH1	55076
PM-SWATH2	24699

The vocabulary size of the dictionary built from PM-SWATH1 increased enormously and may finally reach the limit of vocabulary size of CMU SLM toolkit if the text corpus gets bigger. On the other hand, the vocabulary size from PM-SWATH2 reduced comparing with that of SWATH. This is due to the fact that PM-SWATH2 generalizes the word unit by concatenating shorter simple units using pseudo-morphemes.

5. Conclusion

In this paper, we proposed a quick approach to create a language model for Thai to overcome manual work caused by the difficulty of Thai word segmentation. Dictionary-based segmentation approach works well when there is no unknown word included in input sentences. However, when unknown words are involved in the text, segmentation result becomes unreliable. The incorrect word boundaries, in most cases, make segmented units unpronounceable. In our proposed method, pseudo-morpheme segmentation is used to analyze word boundaries in the area where unknown words appear in the dictionary-based segmentation. The first approach combines unknown words with neighboring words having common boundaries given by the both segmentations. The other approach replaces the segmentation result by pseudo-morpheme segmentation in the areas where unknown words exist. Both approaches give better test-set perplexity as well as OOV rate, and also improve CER.

6. Acknowledgement

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

7. References

[1] W. Aroonmanakun, "Collocation and Thai word segmentation", Proc. SNLP and Oriental COCOSDA Workshop, pp.68-75., 2002.

[2] 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.

[3] V. Sornlertlamvanich, P. Tarsaku, R. Thongprasirt, "Thai grapheme-to-phoneme using probabilistic GLR parser", Proc. Eurospeech-01, pp.1057-1060, 2001.

[4] "SWATH – Smart Word Analysis for Thai", http://www.links.nectec.or.th/, NECTEC.