

Language Model Construction for Thai LVCSR*

Issara Thienlikit, Chai Wutiwivatchai, Sadaoki Furui (Tokyo Institute of Technology)

1. Introduction

Large-scale language modeling is essential in building an LVCSR system. However, in Thai writing, there is no marker for separating words, and the definition of words is also ambiguous. Therefore, building a text corpus is very difficult. Moreover, automatic word segmentation is still far from perfect and manual correction is unavoidable. We propose an automatic data-driven technique for language modeling without using any dictionary. It is first segmenting text into pseudo-morpheme units, then merging these units using forward and reverse bigram statistics.

2. Thai language

In written Thai, there is no explicit marker for separating sentences, phrases, and words. Spaces sometimes indicate ends of phrases or sentences, as seen in Fig. 1. There are usually two types of words in Thai: simple words, which can have one or more syllables, each syllable may have a meaning, but the meaning of the word is not related to the meaning of any syllable; and compound words, which are composed of two or more simple words.

ตามที่รัฐบาลไทยได้เจรจาและลงนามเปิดเขตการค้าเสรีกับต่างประเทศ โดยเฉพาะออสเตรเลีย และนิวซีแลนด์ ที่ทั้ง 2 ประเทศ เป็นผู้ผลิตและส่งออกสินค้าปศุสัตว์รายใหญ่

Figure 1: A Thai text example

We define the term “pseudo-morpheme” to represent a written form of syllable-like unit, in order to avoid confusion with the definition of “syllable” in the sound system, which consists of an initial consonant, a vowel and an optional final consonant. Table 1 shows the difference between pseudo-morphemes and syllables.

3. Data-driven language modeling technique

The accuracy of the automatic word segmentation method is not so high. Lots of manual correction is needed. In order to improve the accuracy, either a large dictionary or a large text corpus is required. Therefore, increasing the vocabulary size of the LVCSR system is not trivial. We propose an automatic data-driven approach. First, a raw text is segmented into pseudo-morpheme-based units. Consequently, a data-driven unit compounding technique is applied iteratively to merge each two frequent units into a larger unit. The new set of compounded units is used as the dictionary for LVCSR.

3.1 Pseudo-morpheme segmentation

The lack of clear definition of Thai words causes a problem for word segmentation. Segmentation ambiguities can be resolved by performing pseudo-morpheme segmentation proposed by Aroonmanakun[1] instead of word segmentation. Since there is no consistent rule on how to indicate word boundaries, a pseudo-morpheme is a more well-defined unit than a word. This approach uses a number of pseudo-morpheme patterns to match an input string, and trigram statistics of pseudo-morphemes to determine the best segmentation from the ambiguous results after matching the patterns.

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

automatic grapheme-to-phoneme (G2P) conversion for pseudo-morpheme cannot ensure the correctness of the pronunciation. For example, the G2P results of the pseudo-morphemes “วิท” and “ยา” are /wit3/ and /jaa0/, instead of the correct pronunciation /wit3 ta3 jaa0/ of the word “วิทยา”. Therefore unit merging is required to create a longer unit close to a word.

Table 1: Word, pseudo-morpheme, and syllable

Word	Pseudo-morpheme	Syllables (pronunciation)
หน้าต่าง	หน้า ต่าง	หน้า-ต่าง (naa2 taang1)
วิทยา	วิท ยา	วิท-ทะ-ยา (wit3 ta3 jaa0)
พลศึกษา	พล ศึก ษา	พะ-ละ-ศึก-สา (pha3 la3 svk1 saa4)

3.2 Data-driven unit-merging

Although the pseudo-morpheme segmentation significantly reduces OOV rates, it yields many short lexical units with higher acoustic confusion, and the span of an N -gram language model (LM) is significantly shorter given a fixed size of N . Therefore, it is highly likely to lose a part of performance gained from reducing the OOV rate. The loss can be partially recovered by compounding frequently co-occurring lexical units and adding them to the lexicon as new lexical units.

There have been several attempts to automatically build compound words based on statistical-based methods[2][3]. A measure that we use for compounding two lexical units w_i and w_j is the geometrical average of the direct and reverse bigrams:

$$M(w_i, w_j) = \sqrt{P_f(w_j | w_i)P_r(w_i | w_j)} = \frac{P(w_i, w_j)}{\sqrt{P(w_i)P(w_j)}}$$

The measure M is between 0 and 1. A high value for M means that both the direct and the reverse bigram values are high for (w_i, w_j) . This makes the pair a good candidate for compounding, since the co-occurrence probability of (w_i, w_j) is high.

In our implementation, first, we train a bigram LM using the initial set of pseudo-morpheme units and compute M for all the bigrams. The bigrams that have an M value higher than a threshold are merged and added to the lexicon. Then we modify the training text using the new lexicon, train a new bigram LM and choose another subset of bigrams to merge. This process can be repeated a number of times to create longer units.

The pronunciation of each entry in the lexicon is automatically generated using a G2P conversion tool from NECTEC[4]. The words that failed the G2P conversion process are excluded from the lexicon. Then the new training text containing only lexical units that passed the G2P conversion is 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 of 18 male-speakers' voices using HTK Toolkit. 25-dimensional feature vectors consisted of 12 MFCCs, their delta, and a delta power. A 10M-word text corpus was collected from 21 months of Thai newspaper website and used to train LMs by the CMU SLM toolkit. Tone information was neglected. JULIUS was used as a speech decoder. Evaluation

* タイ語の大語彙連続音声認識のための言語モデル作成

イッサラー・ティアンリキット、チャイ・ウツィウィワッチャイ、古井貞熙（東京工業大学）

speech data consists of 1,000 utterances from 5 male speakers. A baseline LVCSR model was built and achieved the best word error rate (WER) of 11.6%.

We experimented various LMs built by the proposed data-driven technique in different conditions of the threshold value (log M) and the number of iterations. We also compared our method to

- the baseline system, in which manual correction was performed after automatic word segmentation by a tool from Thailand’s NECTEC called SWATH[5] and the pronunciations were manually typed in, denoted as “Baseline”
- the automatic word segmentation only, also by using SWATH tool, no manual correction was performed, the pronunciations were obtained from the automatic G2P conversion, denoted as “SWATH”
- the pseudo-morpheme segmentation before unit merging is applied, the pronunciations were obtained from the automatic G2P conversion, denoted as “PMSEG”.

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 with word segmentation and manual correction. The perplexities of various configurations are shown in Table 2. It can be seen that in most cases perplexities decrease as iterations increase except for the threshold value of -0.8 and -1.2 where the perplexities increase after 2 iterations.

Table 2: Normalized test-set perplexity comparison

Baseline	139.96	Thres.					
		Iter.	-0.4	-0.8	-1.2	-1.6	-2.0
SWATH	229.03	1	156.24	139.11	119.15	117.01	137.80
PMSEG	174.41	2	149.76	119.73	95.44	105.65	
		3	148.71	124.40	102.88	104.75	

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

Table 3: CER comparison

Baseline	8.65	Thres.					
		Iter.	-0.4	-0.8	-1.2	-1.6	-2.0
SWATH	9.88	1	11.93	10.86	10.05	9.80	10.02
PMSEG	13.42	2	11.41	10.15	9.27	9.27	
		3	11.18	9.94	9.40	9.50	

As shown in Table 3, the best result is obtained at the threshold value of -1.2 or -1.6 with 2 iterations. Generally, decreasing threshold value yields better recognition result. The proposed method outperforms the SWATH word segmentation, though the best CER is still higher than the manually-corrected baseline model by 0.62%.

Fig. 2 shows the effect of the number of iterations to the CER and perplexity. A strong correlation between the CER and perplexity occurs at a few first iterations. Longer lexical units obtained at high iterations overcome the effect of high perplexities in speech recognition. The lowest CER of 9.24% can be obtained with 4 iterations.

Table 3 and 4 show that by decreasing the threshold or increasing the iteration, the recognition accuracy and the vocabulary size increase. Nevertheless, Fig. 3 shows that the vocabulary size seems to be saturated after 5 iterations.

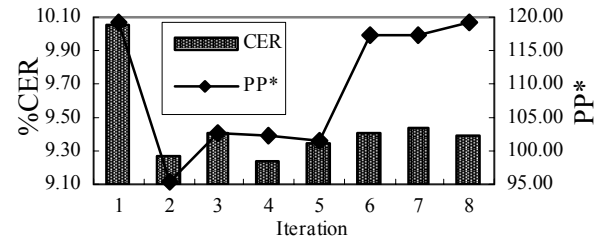


Figure 2: Relationship of iteration, CER and perplexity (threshold = -1.2)

Table 4: Vocabulary size

Baseline	5,000	Thres.					
		Iter.	-0.4	-0.8	-1.2	-1.6	-2.0
SWATH	5,252	1	4,425	5,177	6,745	10,054	16,740
PMSEG	4,198	2	4,524	5,916	10,007	23,230	
		3	4,570	6,395	13,209	43,300	

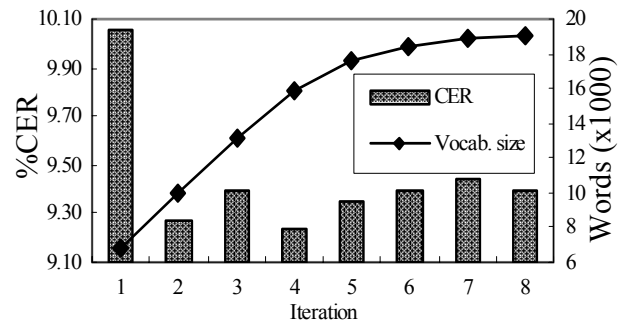


Figure 3: Iteration, CER and vocabulary size (threshold = -1.2)

5. Conclusion

This paper has briefly explained the research on a pioneer Thai LVCSR system. An automatic data-driven technique for building a language model for Thai LVCSR has been proposed, which is expected to accelerate Thai text corpus development in the future. Although more iterations yield better CERs, lexical units become extremely long in higher iterations. Long units might rarely occur in other domains; this is minus in the aspect of generality. The trade-off between the vocabulary size and accuracy should also be considered. Counts of the unit pairs to be merged should also be taken into account, in order to provide generality and avoid fast expansion of vocabulary size.

6. Acknowledgement

The authors would like to thank Dr. Wirote Aroonmanakun, who has given permission to use his segmentation tool, and NECTEC, which has provided many Thai-related useful tools used in this research.

7. References

- [1] W. Aroonmanakun, “Collocation and Thai word segmentation”, Proc. SNLP and Oriental COCOSA Workshop, pp.68-75., 2002.
- [2] G. Saon, M. Padmanabhan, “Data-driven approach to designing compound words for continuous speech recognition”, IEEE Trans. on Speech and Audio Processing, 9(4), May, 2001.
- [3] K. Hacioglu, B. Pellom, T. Ciloglu, O. Ozturk, M. Kurimo, M. Creutz, “On lexicon creation for Turkish LVCSR”, Proc. Eurospeech-03. Sep., 2003.
- [4] V. Sornlertlamvanich, P. Tarsaku, R. Thongprasirt, “Thai grapheme-to-phoneme using probabilistic GLR parser”, Proc. Eurospeech-01, pp.1057-1060, 2001.
- [5] “SWATH – Smart Word Analysis for Thai”, <http://www.links.nectec.or.th/>, NECTEC.