論文 / 著書情報
Article / Book Information

論題(和文)	
Title(English)	Language Modeling for Thai Broadcast News LVCSR
著者(和文)	ﾁｮﾝﾀｳｨｰｻﾀｰﾎﾟｰﾝ, 古井 貞熙
Authors(English)	Markpong Jongtaveesataporn, Sadaoki Furui
出典(和文)	日本音響学会2008年秋季講演論文集, Vol. , No. 3-1-4, pp. 85-86
Citation(English)	, Vol. , No. 3-1-4, pp. 85-86
発行日 / Pub. date	2008, 9

Language Modeling for Thai Broadcast News LVCSR *

◎ Markpong Jongtaveesataporn, Sadaoki Furui
Tokyo Institute of Technology

1. Introduction

Tokyo Institute of Technology has started developing a Thai broadcast news transcription system and constructed the first Thai broadcast news speech and language corpora [1]. These corpora are still very small compared to other major languages. A language model which is one of the key components in an LVCSR system requires a large text corpus for training. This paper illustrates our approach to build a language model for Thai broadcast news LVCSR by using a small transcript text corpus and a newspaper text corpus which can be constructed more easily.

2. Characteristics of the Thai language

2.1. General background

The phonological system of Thai consists of 38 initial consonants, 18 vowels, 6 diphthongs, 12 final consonants, and 5 tones. Thai syllable structure is /C_i V C_f/ where C_i, V, and C_f represent an initial consonant, a vowel, and a final consonant, respectively. Table 1 shows the phonetic symbols used in our system.

Table 1: Thai phonetic symbols

C_i	Single: ph th ch kh p t c k h b d m n ng r l j w s f z Cluster: phr phl thr khr khl khw pr pl tr kr kl kw br bl fr fl dr
V	Single: i i: e e: x x: v v: q q: u u: @ @: Diphthong: ia i:a va v:a ua u:a
C_f	p^ t^ k^ m^ n^ ng^ w^ j^ f^ s^ ch^ l^

There is no inflection in the Thai language. Tenses, subject-verb agreements, and levels of politeness are represented by adding words to a sentence. Thai is written from left to right without sentence or word boundary markers. A space is sometimes used to separate phrases and sentences for aesthetic reasons; however there is no rule or convention requiring the space. Since a word unit is not well defined in the Thai language, word segmentation is not a trivial task.

2.2. Spoken and written style

One significant difference between Thai spoken and written style texts is the level of politeness of a sentence. In a formal conversation as well as a TV news report, a speaker needs to make speech polite to the other party. For a man, “ครับ” (khrap^3) is used and for a woman, “คะ” (kha3) or “ค่ะ” (kha1) are used. These words are added at the end of a sentence but sometimes they are also inserted within a sentence when the speaker tries to make a pause.

Some other words are used together with the words described above to express additional meaning or feeling. For example, “นะ” (na3) which, in most cases, holds no special meaning but sometimes it expresses emphasis on the sentence or it is used when the speaker requests for something. Another word that appears occasionally is “ล่ะ” (la1) which is used in questions. In fact, in addition to “นะ” and “ล่ะ”, there are many other words being used but they are not usually spoken in formal speech or in a TV news report. These words are always placed in front of words indicating politeness. As a result, they are formed to create “นะครับ”, “นะคะ”, “ล่ะครับ”, “ล่ะคะ”, etc. We refer to these words as ending words for the rest of this article.

There are other differences between spoken and written style texts such as spontaneity, which are out of the scope of this article.

3. Language modeling

3.1 Text resources

Since the domain of our application is TV broadcast news, a text corpus that is perfectly suitable for training a language model is a TV broadcast news transcript text corpus. The construction of a broadcast news transcript text corpus is extremely time-consuming and expensive. The only available corpus in Thai was transcribed from around 52 hours of TV broadcast news, which was rather small. A part of the corpus was selected to compose a test set. The rest of the corpus was then used for language modeling.

There is possibility that a text corpus collected from newspaper texts can be effectively used for training a language model for the TV broadcast news LVCSR system since it also comprises texts of the target news domain. A newspaper text corpus can be constructed much more easily than a transcript text corpus. Therefore, a newspaper text corpus is very helpful if a newspaper language model can improve the performance of the LVCSR system.

3.2 Word segmentation

Since a statistical language model requires the concept of words, a Thai text corpus needs to be segmented into words before it can be used for training. In [2], the concept of pseudo-morpheme (PM) for Thai was proposed. It also showed that a language model trained from a text corpus segmented into compound pseudo-morphemes (CPM) yielded better performance than one trained from a text corpus segmented into traditional words. Text corpora used in all experiments in this paper were segmented into compound pseudo-morphemes before they were used to train language models.

3.3 Language model interpolation

Our transcript text corpus is much smaller than a newspaper text corpus. An *n*-gram language model trained from a newspaper text corpus should be reliable to predict general news text; however, the newspaper model cannot predict spoken style text well because there is not much spoken style text in the newspaper corpus. Interpolating the newspaper model with the transcript model should help improve the performance of the newspaper model. The combined model can be thought of as the newspaper model that was integrated with better *n*-gram estimation on spoken style text. The overall likelihood $P(w|h)$ of a word w occurring after the history h is computed as linear interpolation of $P(w|h)$ for each of the models as,

$$P(w|h) = \lambda P_1(w|h) + (1-\lambda)P_2(w|h) \quad (1),$$

where λ is the weight of the newspaper text model, $P_1(w|h)$ and $P_2(w|h)$ are *n*-gram probabilities of the newspaper language model and the transcript language model, respectively.

4. Experimental conditions

Two newspaper read-speech corpora, LOTUS [3], and a phonetically balanced sentence corpus collected by Tokyo Institute of Technology were employed for training the acoustic model. The total amount of acoustic training data was 40.3 hours from 68 male and 68 female speakers. Gender-dependent acoustic models (AM) of 1000-tied-state triphones with 8 Gaussian mixtures were trained by HTK [4]. 25-dimensional feature vectors consisted of 12 MFCCs, their delta, and a delta energy.

＊タイ語放送ニュース音声を対象とした大語彙連続音声認識システムのための言語モデル構築
マッカポン・チョンタウィーサターポーン、古井貞熙
東京工業大学

A newspaper text corpus covering about 5 years of news (2003-2007) was used in the experiments. The corpus contained about 139 million PMs. The size of the transcript text corpus used in the experiments was around 962k PMs. 3-gram language models were trained by SRILM toolkit [5].

The test set contained clean speech utterances randomly selected from the Thai broadcast news speech corpus. In total, 1033 speech utterances (626 male and 407 female utterances) were used for evaluation. The test set contained 24507 PMs. JULIUS [6] version 4.0.1 was used as a speech decoder.

5. Experimental results

The newspaper and transcript language models were trained from the two text corpora. Then various models were created by interpolating the two models, varying the newspaper language model weight from 0.0 to 1.0. The dictionary size was fixed to around 64k. These language models were employed in the LVCSR system. Evaluation on test set perplexity (PP) and speech recognition accuracy was performed. Here, we used PM error rate (PER) as the measure for ASR performance. PPs and PERs are displayed in Figure 1 as a function of the newspaper language model weight.

The result shows that the transcript language model trained from a relatively small text corpus outperformed the newspaper language model both in terms of PP and PER. This implies that a language model trained from an in-domain text corpus is much more powerful than one trained from an out-domain text corpus. The PER and PP obtained from the newspaper language model were 21.7% and 451.0 respectively while those from the transcript language model were 19.9% and 318.8 respectively. Additionally, better performance could be achieved from the interpolated models. The best result was obtained when the newspaper language model weight was 0.5. The PER was 17.7% and the PP was 183.8. The improvement over the newspaper model was probably due to the enhanced estimation on spoken style speech. Conversely, the improvement over the transcript model was resulted from better *n*-gram estimation on general text.

Further analysis shows that the improvement was partly gained from better *n*-gram estimation on ending words. PPs of each sentence in the test set were calculated for the newspaper, transcript, and best interpolated models. Table 2 shows the ratio of sentences that match each case indicated in the Case column. N_PP, T_PP, I_PP refer to the perplexity calculated for the newspaper, transcript, and best interpolated models respectively. For each case, the matched sentences were listed. Then, the number of ending words appearing in these sentences was counted. The ratio of ending words to all words in each case is shown in Table 2. There were about 52.5% of the sentences of which the PPs for the newspaper model were larger than ones for the transcript model. Within these sentences, the occurrence ratio of ending words was 1.9% which was much higher than the opposite case (0.3%). This implies that the newspaper model can predict written style sentences better than spoken style sentences. After interpolation, the ratio of sentences of which the PPs for the newspaper model were larger than the ones for the interpolated model rose to 83.5% and the ratio of ending words in these sentences was 2.6% which was much higher than the opposite case (0.1%). This signifies that interpolating the transcript model to the newspaper model help improve the prediction of spoken style sentences.

A recognition error analysis focusing on the ending words is shown in Table 3. The symbol "→" indicates the change in recognition correctness when using the newspaper and interpolated models. The interpolated models repaired 164 incorrectly recognized PMs, plus 99 PMs that surround ending words. There are only 5 PMs newly misrecognized by the interpolated model.

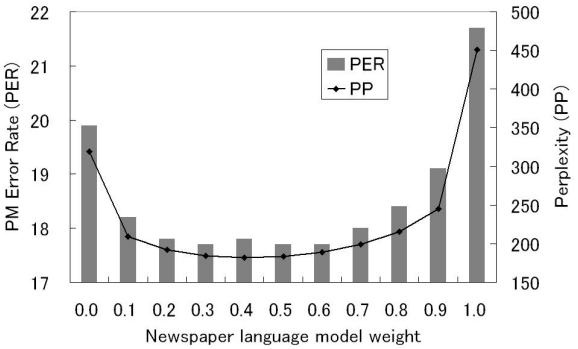

Figure 1: Test set PP and PER based on interpolated models

Table 2: Sentence PP comparison

Case	Ratio of sentences that match the case	Ratio of ending words to all words in sentences that match the case
N_PP > T_PP	52.5%	1.9%
N_PP < T_PP	47.5%	0.3%
N_PP > I_PP	83.5%	2.6%
N_PP < I_PP	16.5%	0.1%

Table 3: Recognition error analysis on recognition result

Case	Number of PMs
Correct → Correct	128
Incorrect → Incorrect	100
Incorrect → Correct	164
Correct → Incorrect	2
Surrounding PM: Incorrect → Correct	99
Surrounding PM: Correct → Incorrect	3

6. Conclusion

This paper described the construction of a language model for Thai broadcast news LVCSR system. Since an in-domain text corpus was rather small, we combined a large newspaper text corpus to train a language model. A better language model was obtained by interpolating the two models. The analysis showed that improvement on perplexity and recognition accuracy was obtained by better *n*-gram estimations for spoken style speech.

7. Acknowledgement

The speech corpus used for training the acoustic model was funded by the METI Project "Development of Fundamental Speech Recognition Technology".

8. References

[1] M. Jongtaveesataporn, C. Wutiwiwatchai, K. Iwano and S. Furui, "Thai broadcast news corpus construction and evaluation," Proc. LREC 2008, 2008.

[2] M. Jongtaveesataporn, I. Thienlikit, C. Wutiwiwatchai and S. Furui, "Towards better language modeling for Thai LVCSR," Proc. Interspeech 2007, pp. 1553-1556, 2007.

[3] S. Kasuriya, V. Sornlertlamvanich, P. Cotsomrong, S. Kanokphara and N. Thatphithakkul, "Thai speech corpus for Thai speech recognition," Proc. The Oriental COCOSDA 2003, pp. 54-61, 2003.

[4] http://htk.eng.cam.ac.uk

[5] A. Stolcke, "SRILM -- An Extensible Language Modeling Toolkit," Proc. ICSLP, vol. 2, pp. 901-904, 2002.

[6] http://julius.sourceforge.jp/en/julius.html