



Topic and style-adapted language modeling for Thai broadcast news ASR

Markpong Jongtaveesataporn, Sadaoki Furui

Department of Computer Science, Tokyo Institute of Technology

marky@furui.cs.titech.ac.jp, furui@cs.titech.ac.jp

Abstract

The amount of available Thai broadcast news transcribed text for training a language model is still very limited, comparing to other major languages. Since the construction of a broadcast news corpus is very costly and time-consuming, newspaper text is often used to increase the size of training text data. This paper proposes a language model topic and style adaptation approach for a Thai broadcast news ASR system, using broadcast news and newspaper text. A rule-based speaking style classification method based on the existence of some specific words is applied to classify training text. Various kinds of language models adapted to topics and styles are studied and shown to successfully reduce test set perplexity and recognition error rate. The results also show that written style text from newspaper can be employed to alleviate the sparseness of the broadcast news corpus while spoken style text from the broadcast news corpus is still essential for building a reliable language model.

Index Terms: Thai broadcast news, language model adaptation, topic and style

1. Introduction

Broadcast news (BN) recognition systems for several languages have been advanced greatly. A large amount of text corpora for those languages has been collected in order to train reliable language models (LM). In case of resource deficient languages such as Thai, corpora are not widely constructed. In addition to the first Thai BN speech and language corpora we have developed [1], a collaborative work with NECTEC [2] was established to increase the size of the corpus. However, the amount of available BN transcript text is still less than 100 hours while there are more than 1000 hours of transcribed text in other major languages.

Since the construction of a BN corpus takes a lot of resources, labor, time and money, it would be favorable if an alternative text resource can be employed to train a LM for BN speech. Text resources of which the content seems to be similar to BN text are text in newspapers (NP). An NP text corpus can be constructed much more easily than a BN transcript text corpus. Hence, an NP text corpus is very attractive if it can improve the performance of an ASR system.

In fact, not only resource deficient languages suffer from the lack of training resources, but also there are general difficulties in constructing language models matched to a target speech because the amount of well-matched data is usually limited. A research group in [3] tried to perform LM style adaptation and achieved a reduction in recognition error rate on a broadcast conversation recognition task using an LDA (latent Dirichlet allocation) based approach to combine multiple LMs trained from two corpora with different styles: BN and broadcast conversation corpora. An LM adaptation method based on both topics and speaker characteristics, which can also be considered as styles, was proposed in [4]. They performed probabilistic latent semantic analysis (PLSA) on two corpora covering various topics and speakers, and used

initial recognition hypotheses to produce unigram probabilities for LM adaptation.

The focus of our work is to investigate LM topic and style adaptation for Thai BN ASR, using two text resources, BN and NP text corpora. Styles, here, refer to the differences of text styles used in BN and NP, and specific speaking styles used in the Thai language. Based on the characteristics of the Thai language, a rule based speaking style classification approach is used to classify text into spoken and written styles. LMs for different topics and styles are trained and then combined together by linear interpolation.

2. LM training and adaptation methods

2.1. Lexical unit

There is no word boundary marker such as a space in Thai written text. Since a statistical LM requires lexical units defined as entities separated from each other, a Thai text corpus needs to be segmented into units prior to LM training. The lexical unit used in this work is compound pseudo-morpheme (CPM) [5]. CPM is created by combining several pseudo-morphemes (PM), syllable-like units in Thai written form.

2.2. Thai speaking styles and BN speech

One significant difference between Thai spoken and written style text is the level of politeness of a sentence. In a formal conversation as well as a BN report, a news announcer needs to speak politely to the other party. For a man, “ครັบ” (khrap³) 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 indicating politeness to express additional meaning or feeling. For example, one of the most frequent words found in Thai BN is “นะ” (na3) which, in most cases, holds no special meaning but sometimes emphasizes the content of the sentence or is used in imperative sentences. Another word that appears occasionally is “ละ” (la1) which is used mostly in questions. The above words are always placed in front of words indicating politeness, forming spoken style words such as “นะครັบ”, “นะคะ”, “ละครັบ”, and “ละคะ”. Since these spoken style words are often (but not always) put at the end of a sentence, we refer to these words as spoken style ending words (SSEW) for the rest of this paper.

2.3. Rule-based speaking style classification

With the different characteristics of Thai spoken and written style sentences, we propose a rule-based classification method to indicate the speaking style of a sentence. A sentence containing SSEWs is considered as a spoken style sentence, and a sentence without such SSEW is classified as a written style sentence. In this work, 11 words were defined as SSEWs.

Since CPM is used as the lexical unit, a list of SSEW must be taken in a form of CPM. CPM is usually longer than the word unit and SSEWs can be combined with some other PMs. Therefore, CPMs including SSEWs are considered as spoken style CPMs and there were 132 spoken style CPMs in our system.

Thai BN normally comprises written style speech and spoken style speech. Written style speech is often found when news announcers read news script narrating detailed news reports. On the other hand, spoken style speech is mostly found at introductions, transitions, conclusions of news stories. In our transcribed corpus, written style and spoken style sentences cover 57% and 43%, respectively. On the contrary, NP normally uses written style text. For example, our NP text corpus contains mostly written style sentences (99.3%).

2.4. Text clustering

Training text data can be clustered based on three aspects as follows:

1. Text source: Text is grouped based on its source, which results in distinguishing BN and NP style text.
2. Speaking style clustering: Text is clustered to spoken style (SP) and written style (WR). Here, the rule-based classification is employed.
3. Topic clustering: Text can also be clustered into topics. We use the TF-IDF (term frequency-inverse document frequency) vectors to represent sentences and the cosine function to measure the similarity between sentences. A cluster of text is constructed based on the similarity scores.

2.5. Adaptation of n-gram model

All training text is clustered by sources, topics and styles. Each specialized LM is trained from a text cluster. Interpolation weights of the models are optimized with EM algorithm on hypotheses derived from a previous pass of a multi-pass recognition system. The following model types are investigated for our LM adaptation scheme.

● Model Type I: A specialized model is trained from text in a specific source. The adapted model is obtained by the following formula:

$$P(w|h) = \lambda \cdot P_s(w|h, T_{NP}) + (1 - \lambda) \cdot P_s(w|h, T_{BN}) \quad (1)$$

where w is the current word for which the probability is calculated, h is the history, λ is the weight assigned to an LM trained from T_{NP} , P_s is a specialized model built from training text T_x which refers to all text from source x .

● Model Type II: Irrespective of text source, a specialized model is trained from text with a specific speaking style. The adapted model is obtained by the following formula:

$$P(w|h) = \lambda \cdot P_s(w|h, T_{WR}) + (1 - \lambda) \cdot P_s(w|h, T_{SP}) \quad (2)$$

where λ is the weight assigned to an LM trained from T_{WR} , and T_y refers to all y style text.

● Model Type III: A specialized model is trained from text in a specific source and speaking style. The adapted model is obtained by the following formula:

$$P(w|h) = \lambda_{NP,WR} \cdot P_s(w|h, T_{NP,WR}) + \lambda_{NP,SP} \cdot P_s(w|h, T_{NP,SP}) + \lambda_{BN,WR} \cdot P_s(w|h, T_{BN,WR}) + \lambda_{BN,SP} \cdot P_s(w|h, T_{BN,SP}) \quad (3)$$

where $\lambda_{x,y}$ is the weight assigned to a specialized model trained

from source x with style y , such that $\sum \lambda = 1$, and $T_{x,y}$ refers to text from source x with style y .

● Model Type IV: A specialized model is trained from each topic cluster without considering its source and style. The adapted model is obtained by the following formula:

$$P(w|h) = \sum_{i=1}^C \lambda_i \cdot P_s(w|h, T_i) \quad (4)$$

where λ_i is the weight assigned to each specialized model component such that $\sum \lambda = 1$, C gives the number of topic clusters, P_s is a specialized language model built from training text T_i , and T_i refers to text in cluster i .

● Model Type V: A specialized model is trained from text in a topic cluster from a single source. The adapted model is obtained by the following formula:

$$P(w|h) = \sum_{i=1}^C [\lambda_{i,NP} \cdot P_s(w|h, T_{i,NP}) + \lambda_{i,BN} \cdot P_s(w|h, T_{i,BN})] \quad (5)$$

where $T_{i,x}$ refers to text from source x in topic cluster i .

● Model Type VI: A specialized model is trained from text in a topic cluster with a speaking style. The adapted model is obtained by the following formula:

$$P(w|h) = \sum_{i=1}^C [\lambda_{i,WR} \cdot P_s(w|h, T_{i,WR}) + \lambda_{i,SP} \cdot P_s(w|h, T_{i,SP})] \quad (6)$$

where $T_{i,x}$ refers to style y text in topic cluster i .

● Model Type VII: A specialized model is trained from text in a cluster with a source and a style. The adapted model is obtained by the following formula:

$$P(w|h) = \sum_{i=1}^C \left[\begin{array}{l} \lambda_{i,NP,WR} \cdot P_s(w|h, T_{i,NP,WR}) \\ + \lambda_{i,NP,SP} \cdot P_s(w|h, T_{i,NP,SP}) \\ + \lambda_{i,BN,WR} \cdot P_s(w|h, T_{i,BN,WR}) \\ + \lambda_{i,BN,SP} \cdot P_s(w|h, T_{i,BN,SP}) \end{array} \right] \quad (7)$$

where $T_{i,x,y}$ refers to text from source x with style y in topic cluster i .

In summary, Model Type I, II, and III can be considered as LMs adapted to styles. Model Type IV is a topic adapted LM. Model Type V, VI, VII, and VIII are LMs adapted to both topics and styles.

3. Experimental conditions

Gender-dependent acoustic models were trained from newspaper read speech corpora (LOTUS [6] and a phonetically balanced sentence speech corpus collected by Tokyo Institute of Technology) using HTK [7]. The total amount of acoustic training data was 40.3 hours from 68 male and 68 female speakers. 25-dimensional feature vectors consisting of 12 MFCCs, their delta, and a delta energy were used for acoustic model training. The HMM states were clustered by a phonetic decision tree. The number of leaves was 1,000. Each state of the HMM was modeled by a mixture of eight Gaussians. No special tone information was incorporated.

An NP text corpus covering about five years of news (2003-2007) was used in the experiments. The corpus contained about 139 million PMs. The size of our BN text corpus was around 1.8 million PMs. All experiments used CPM as a lexical unit for LVCSR. Tri-gram backoff LMs were trained by SRILM toolkit [8]. The dictionary having around 60k CPMs was used for all experiments.

Text from the BN and NP text corpora was clustered by topics as described in Subsection 2.4. Around 350 function words (in CPM forms) were defined and excluded in the calculation of TF-IDF vectors. Two-phase bisecting K-means algorithm was employed to cluster the text. In this work, the software named CLUTO [10] was used to perform the clustering.

A BN test set contained clean speech utterances randomly selected from the Thai BN speech corpus. In total, 1033 speech utterances (626 male and 407 female utterances) were used for the evaluation. The OOV rate was 0.2%. JULIUS [9] version 4.1.2 was used as a speech decoder. The recognition result was evaluated by PM error rate (PER).

4. Experiments

4.1. Experiments on various LM adaptation schemes

LVCSR experiments were performed based on the source, topic and style LM adaptation schemes proposed in Subsection 2.5. For Model Types IV, V, VI, and VII, the number of topics in the text corpora was decided first in order to reduce computation time required by varying the number of topic clusters for each model type. The number of topic clusters was varied from 2 to 20, and Model Type IV was used to test the performance of adapted models. PERs of the systems ranged from 19.5% to 20.7%, and the system with 8 topic clusters performed the best. Therefore, the rest of the experiments were conducted with 8 topic clusters. It is worth noting that, for other model types, the best system may be constructed by a different number of topic clusters since the amount of training text data changed from one model type to another. However, as the results in the next experiments show that Model Type IV gave the worst PER result among above model types, we can ensure that the performance of other model types with the best condition of topic numbers will be at least equal to or better than those reported in this paper.

First of all, the performances of all model types were evaluated when supervised adaptation was applied. PPs and PERs of all model types are shown in Table 1. The baseline refers to the system using one LM trained with combined BN and NP text. LM adaptation to text source and speaking styles performed in Model Type I and II could successfully reduce PP and PER. Moreover, Model Type III which performed adaptation to both text source and speaking styles further lowered PP and PER significantly. Topic adaptation achieved by Model Type IV was able to decrease PP and PER to the level of Model Type I but still worse than Model Type III. The rest of the model types performing topic and style adaptation gave better results than previous model types except for Model Type VI which yielded similar results to Model Type III. The best result was achieved by Model Type VII yielding better PP and PER than models that performed only style or topic adaptation. The best PP and PER results of Model Type VII is partly attributed to the fact that there are more parameters that can be adjusted to fit the hypothesis set than other model types.

Even though the supervised evaluation seems to give impressive improvement on PP and PER, a real adaptation process needs to be done by using a recognition hypothesis set with recognition errors. In this paper, hypotheses from the baseline with 20.2% PER were used. Table 2 shows PP and PER results for all model types obtained from the best pass of recognition. Similar to the results from the supervised adaptation, Model Type I and II could reduce PP and PER,

Table 1: PP and PER (%) obtained from supervised adaptation

Adaptation Type	Model Type	PP	PER (%)
No adaptation	Baseline	207.8	20.2
Source and style adaptation	I	147.5	18.4
	II	160.1	18.8
	III	124.5	16.9
Topic adaptation	IV	150.3	18.4
Source, topic and style adaptation	V	103.3	16.3
	VI	124.8	16.9
	VII	99.1	15.9

Table 2: PP and PER (%) obtained from unsupervised adaptation, and percentages of changes in PERs from supervised to unsupervised adaptation

Adaptation Type	Model Type	PP	PER (%)	%Change in PER
Source and style adaptation	I	157.5	19.0	3.3
	II	168.1	19.2	2.1
	III	138.7	18.2	7.7
Topic adaptation	IV	171.6	19.5	6.0
Source, topic and style adaptation	V	127.1	18.0	10.4
	VI	146.1	18.5	9.5
	VII	129.5	18.3	15.1

compared to the baseline, and Model Type III could further decreased PP and PER to 138.7 and 18.2% respectively. Model Type IV performing topic adaptation gave PP and PER of 171.6 and 19.5%, respectively. Unlike the results of supervised adaptation, Model Type IV gave worse PP and PER results than all models adapted to styles. For topic and style adaptation, the best result was obtained from Model Type V instead of Model Type VII which was the best in the supervised adaptation experiment. Model Type V achieved PP and PER of 127.1 and 18.0%, respectively, which were the best results among all of the model types.

4.2. Discussion

One of the first observations that can be seen from the experimental results is the performance degradation of adapted models when performing unsupervised adaptation. By using hypotheses with 20.2% PER, PP and PER increased drastically, compared to the case of supervised adaptation. Table 2 presents the percentages of changes in PERs for the different cases. The percentages of changes in PER seems to vary by how complicated the text corpus was clustered. These percentages are rather low for Model Type I and II which are the models adapted to only text source and speaking styles, respectively. The percentage rises in Model Type IV of which the mixture of LMs was trained from text clustered into topics. The percentage gradually increases when text are clustered by both text sources and speaking styles (Model Type III), by topics and text sources/speaking styles (Model Type V and VI), and by all aspects (Model Type VII and VIII). This occurs because LM weights are not appropriately estimated when hypotheses contain recognition errors. Furthermore, since training text for Model Type VII is repeatedly clustered into many small clusters, the models might not be robust. This should explain why Model Type VII cannot outperform Model Type V in unsupervised adaptation.

Another point of discussion is the weight distribution assigned to LM components. The weight distributions for Model Type III and VII are shown in Figures 1 and 2. Numbers on the x-axis in Figure 2 refer to the 8 topic-dependent LM components. Since there are more written style utterances than spoken style utterances in the test set, written style models are more dominant than spoken style models. Moreover, it is quite clear that the models trained from NP written style text are more important than ones from BN written style text. Regarding the spoken style models, the models from BN are considerably more important than ones from NP. This suggests that spoken style text from BN text is crucial for training a LM for BN recognition task. Thus, the most desirable information in the BN text corpus is spoken style text. Written style models can be trained by using written style text in the NP text corpus. Therefore, the transcription of spoken style speech in a collection of BN database has a higher priority than the transcription of written speech. This information must be a useful tip for managing BN language resources for resource deficient languages like Thai.

We finally discuss the performance of Model Types III and V. Model Type III is a model adapted to text source and speaking styles for which the LM training process is simple and takes less time than Model Type V which performs adaptation based on topics and text sources. The reduction in PP given by Model Type V over Model Type III is not significant enough to reduce PER, gaining only 0.1% improvement in PER. A study on topic clustering technique needs to be conducted to improve the performance of the recognition system. So far, Model Type III is adequate enough to be used in LM adaptation for Thai BN. Hence, the proposed rule-based speaking style classification is practical in classifying text for Thai BN language modeling.

5. Conclusion

This paper has presented a language modeling approach for a Thai broadcast news ASR system. Since the amount of broadcast news transcript text for language model training is rather limited, newspaper text was used to increase the size of training data. We proposed a simple rule-based speaking style classification to categorize spoken and written style text, based on the existence of specific spoken style words. Various kinds of n-gram models adapted to topics and styles were investigated, and could successfully reduce test set perplexity and recognition error rate. An analysis of experimental results showed that we could employ written style text from newspaper to alleviate the sparseness of the broadcast news transcript text. However, spoken style text from the broadcast news corpus was still essential for building a reliable language model. Therefore, for a resource deficient language like Thai, a broadcast news corpus including a large number of spoken style utterances should be constructed with the highest priority, to which newspaper text can be added to model written style speech in the broadcast news speech recognition. Enhancing the available data with web corpora such as blogs and twitters which contain a large amount of spoken style text is an interesting topic for future research. It is also interesting to try other interpolation schemes to improve the performance of the adapted LM.

6. Acknowledgement

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

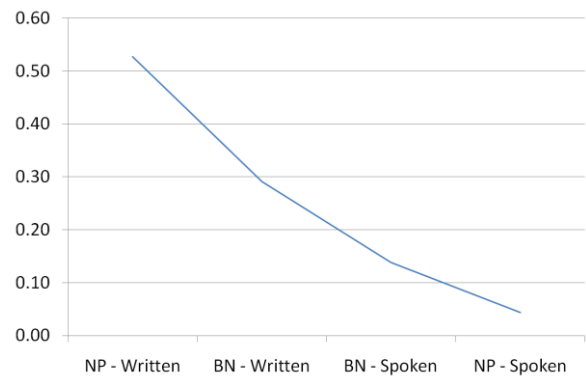


Figure 1: Average LM weight distribution for Model Type III

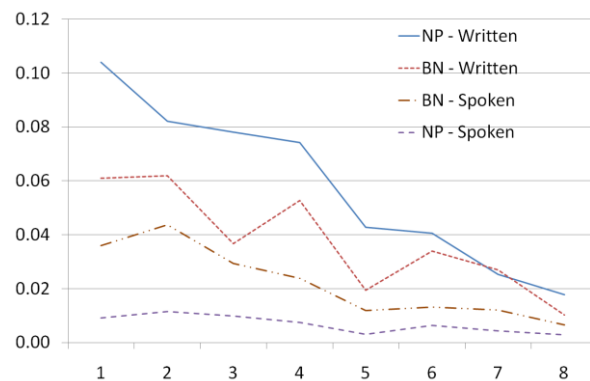


Figure 2: Average LM weight distribution for Model Type VII

NECTEC for providing us additional part of Thai broadcast news corpus.

7. References

- [1] M. Jongtaveesataporn, C. Wutiwiwatchai, K. Iwano and S. Furui, "Thai broadcast news corpus construction and evaluation," in Proc. LREC, 2008.
- [2] A. Chotimongkol, K. Saykhum, P. Chootrakool, N. Thatphithakkul, C. Wutiwiwatchai, "LOTUS-BN: A Thai broadcast news corpus and its research applications," in Proc. Oriental COCODSA, 2009, pp. 44-50.
- [3] D. Mrva and P. Woodland, "Unsupervised language model adaptation for Mandarin broadcast conversation transcription," in Proc. ICSLP, 2006, pp. 2206-2209.
- [4] Y. Akita and T. Kawahara, "Language model adaptation based on PLSA of topics and speakers for automatic transcription of panel discussions," IEICE transactions on information and systems, E88-D, no. 3, 2005, pp. 439-44.
- [5] M. Jongtaveesataporn, I. Thienlikit, C. Wutiwiwatchai and S. Furui, "Lexical units for Thai LVCSR," Speech Communication, vol. 51, no. 4, 2009, pp. 369-389.
- [6] S. Kasuriya, V. Sornlertlamvanich, P. Cotsomrong, S. Kanokphara and N. Thatphithakkul, "Thai speech corpus for Thai speech recognition," in Proc. The Oriental COCODSA, 2003, pp. 54-61.
- [7] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, The HTK Book Version 3.0, Cambridge University Press, 2000.
- [8] A. Stolcke, "SRILM -- An Extensible Language Modeling Toolkit," in Proc. ICSLP, vol. 2, 2002, pp. 901-904.
- [9] A. Lee, T. Kawahara and K. Shikano, "Julius -- an open source real-time large vocabulary recognition engine," in Proc. EUROSPEECH, 2001, pp. 1691-1694.
- [10] G. Karypis, "CLUTO-A Clustering Toolkit," University of Minnesota - Computer Science and Engineering Technical Report, 2002.