Online Temporal Language Model Adaptation for a Thai Broadcast News Transcription System

Kwanchiva Saykham[1,2], Ananlada Chotimongkol[1], Chai Wutiwiwatchai[1]

[1]National Electronics and Computer Technology Center (NECTEC)
112 Phahonyothin Rd., Klong Nueng, Klonng Luang, Pathumthani 12120, Thailand

[2]School of Information and Computer Technology,
Sirindhorn International Institute of Technology
Thammasat University, Pathumthani, 12000, Thailand
E-mail: kwanchiva.say@nectec.or.th, ananlada.cho@nectec.or.th, chai.wut@nectec.or.th

Abstract

This paper investigates the effectiveness of online temporal language model adaptation when applied to a Thai broadcast news transcription task. Our adaptation scheme works as follow: first an initial language model is trained with broadcast news transcription available during the development period. Then the language model is adapted over time with more recent broadcast news transcription and online news articles available during deployment especially the data from the same time period as the broadcast news speech being recognized. We found that the data that are closer in time are more similar in terms of perplexity and are more suitable for language model adaptation. The LMs that are adapted over time with more recent news data are better, both in terms of perplexity and WER, than the static LM trained from only the initial set of broadcast news data. Adaptation data from broadcast news transcription improved perplexity by 38.3% and WER by 7.1% relatively. Though, online news articles achieved less improvement, it is still a useful resource as it can be obtained automatically. Better data pre-processing techniques and data selection techniques based on text similarity could be applied to the news articles to obtain further improvement from this promising result.

1. Introduction

Broadcast news transcription is a challenging task due to the spontaneous nature of the input speech and the richness of the language in the news domain. In this paper, we are interested in the second issue, the richness of the news language, specifically, how to create an efficient language model for a Thai broadcast news domain.

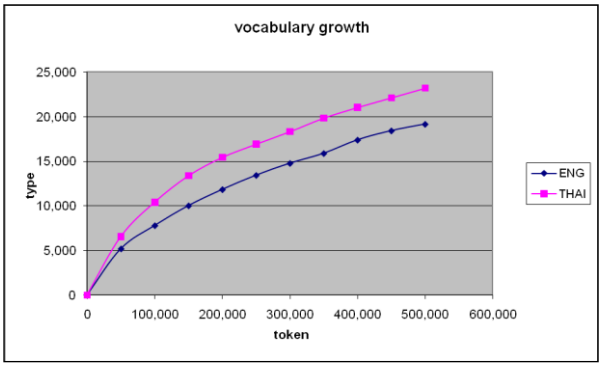

Figure 1: Type-token curves of English and Thai broadcast news data

In Figure1, the vocabulary growth in broadcast news data is represented by a type-token curve. We can see that the vocabulary of the Thai broadcast news, taken from the LOTUS-BN corpus (Chotimongkol et al., 2009), grows slightly faster than the vocabulary of the English broadcast news, taken from the LDC-1996 English Broadcast News Transcripts (HUB4) corpus (Graff & Alabiso, 1997). Nevertheless, the vocabularies of both languages continue to grow even when the sizes of the corpora reach half a million words. In the news domain, new names (e.g. Barack Obama and Port-au-Prince) and new terms (e.g. Facebook and Twitter) are introduced every day. This poses a problem even for a language with rich resources such as English. For a resource-limited language such as Thai, this problem becomes more severe. The number of language corpora, both speech and text, in Thai is quite small. Apart from the high cost in developing a corpus, the lack of word and sentence boundaries is another obstacle as additional pre-processing steps of sentence segmentation and word segmentation are required when dealing with Thai texts.

Our goal is to build an effective language model for a speech recognizer in a broadcast news transcription task. Automatic broadcast news transcription can be applied in many applications such as closed captioning and audio indexing for news search. As new words are introduced every day, as shown in Figure 1, it is quite challenging to develop an efficient language model in a broadcast news domain.

Let consider the situation where a speech recognizer is trained from broadcast news data available during the development period and then deployed for transcribing the upcoming broadcast news programs. Since news stories are time-varying data, i.e. their word distributions change over time, the language model of the recognizer should be modified to reflect the language in the more recent news programs. This modification can be done through *online temporal language model adaptation*. The best resource for the temporal adaptation would be the transcription of the news programs broadcasted during the same time period as the news program being recognized.

However, the correct transcription of the most recent news programs is not easy to obtain. A practical alternative resource would be online news articles which can be obtained automatically.

Our online temporal adaptation scheme works as follow: first an initial language model is trained with broadcast news transcription available during the development period. Then the language model is adapted over time with more recent broadcast news transcription and online news articles available during deployment especially the data from the same time period as the broadcast news speech being recognized.

This paper is organized as follows: the following section discusses existing works on temporal language model adaptation. Section 3 then describes the news data utilized in our experiments both the broadcast news transcription and online news articles in more detail. A set of experiments in Section 4 are conducted to examine the similarity between news stories over time and identify an appropriate amount of adaptation data. The results of our temporal language model adaptation are reported in Section 5. Finally, we conclude our paper in Section 6.

2. Related works

Since news stories are time-varying data, there have been several research works on temporal language model adaptation on broadcast news data in various languages such as English, Japanese, Chinese and Italian (Whittaker, 2001; Matsui et. al, 2001; Chen et. al, 2004; Federico and Bertoldi, 2004).

In (Whittaker, 2001) both temporal vocabulary adaptation and language model adaptation were applied for the task of recognizing an English broadcast news show. The adaptation was done on a daily basis with the transcription of the previous day shows and newspaper texst from the same time period. The adapted vocabulary comprises of the most frequent 65K words in the accumulated data while the adapted language model is a linear interpolation of the base language model, trained from 150M words of news paper text, and ten daily language models from current and nine previous days. In (Federico & Bertoldi, 2004) similar adaptation techniques for Italian broadcast news transcription system were utilized. However, Federico and Bertoldi also proposed a new scheme for updating the unigram probabilities of the extended set of lexicons which takes into account the time period that each lexicon was found.

In (Matsui et. al, 2001; Chen et. al, 2004), only temporal language model adaptation technique was used for modelling Japanese and Chinese broadcast news transcriptions respectively. Matsui et. al used a linear interpolation method and utilized only the transcription of the news broadcasted twelve hours before the show as adaptation data while Chen et. al used a minimum discrimination information (MDI) adaptation technique on a block of several months of adaptation data.

Nevertheless, word error rate (WER) reduction obtained when applying the adapted language model to a speech recognizer varies. Similar to the observation made by Whiitaker (2001), we found that if the initial model is well trained, only small improvement can be obtained from the adapted model, whereas, for a model with less initial training data, more improvement can be achieved (Matsui et. al, 2001). Since the available Thai broadcast news corpus (LOTUS-BN) is quite small, less than 1M words of transcribed speech, we could expect to see some improvement on WER with temporal language model adaptation in a Thai broadcast news transcription task

3. Language model adaptation resources

The best resource for temporal adaptation would be the transcription of the news programs broadcasted during the same time period as the news program being recognized. However, the correct transcription of the most recent news programs is quite costly to obtain. One practical alternative resource would be online news articles which can be collected automatically.

We utilize news data from two sources: the LOTUS-BN corpus (Chotimongkol et al., 2009) and online news articles excerpted from the National News Bureau of Thailand website (http://thainews.prd.go.th/). The LOTUS-BN corpus was collected in two phases that are about six months apart. We separate the LOTUS-BN transcription into three sets:

- *TIT-BN* is the first phase of the corpus collected by Tokyo Institute of Technology (TIT) and is used as initial training data.
- *NECTEC-BN* is the second phase of the corpus collected by the National Electronics and Computer Technology Center (NECTEC) and is used as temporal adaptation data.
- *Test-BN* is the most recent news programs in the corpus reserved solely as testing data.

The news articles (*NewsText*) are collected from the same time period as the NECTEC-BN set. The NewsText consists of four news topics: politics, economics, royal news, and general news. The statistics of all data sets are summarized in Table1.

Data set	Time period	Amount of data	Tokens	Token types
TIT-BN	01/2007 - 03/2007	17 hours	223,993	10,303
NECTEC-BN	10/2007 - 03/2008	60 hours	529,136	23,993
NewsText	10/2007 - 03/2008	6,074 articles	1,357,702	38,485
Test-BN	04/2008	0.5 hour	4,150	1,300

Table 1: Statistic of four news data sets

4. Language similarity over time

Base on the assumption that news stories that are closer in time are more similar than new stories that are many months apart, we believe that the transcription of the more recent news programs is more useful as language model adaptation data than the transcription of the news

programs broadcasted many months ago. To verify this assumption, we separated the adaptation data (NECTEC-BN) into small blocks according to their broadcast dates and trained a small language model (LM) from each block. The adapted LM was then created by linearly interpolating this LM with the base LM trained from the TIT-BN set. For simplicity, we assigned both LMs equal interpolation weights. The effectiveness of the adaptation data was measured by the perplexity of each adapted language model with respect to the Test-BN set as shown in Figure2. All the LMs are a trigram model and are estimated by the CMU-Cambridge Statistical Language Modeling Toolkit (Clarkson & Rosenfeld, 1997).

Since the news programs in LOTUS-BN are not evenly distributed over time, the time boundaries of each block were adjusted so that it contains about the same amount of data (4,000 utterances). Block1 is a set of adaptation data that is furthest away (in time) from the test set while Block11 is the closest one.

The base LM (TIT-BN) has the highest perplexity as it was trained from the news programs broadcasted more than one year prior to the test set. The graph in Figure2 shows that more perplexity reduction can be achieved when the base LM was adapted with a data set that is closer to the test set. Relative perplexity reduction of 21.2% is achieved when the transcription of the news programs broadcasted just before the test set was used. Perplexity reduction is the result of both the reduction in the out of vocabulary (OOV) rate and the increase in the 3-gram hit rate as shown in Figure 3. The LM that was adapted with more recent data has a lower OOV rate and a higher 3-gram hit rate, thus has a lower perplexity.

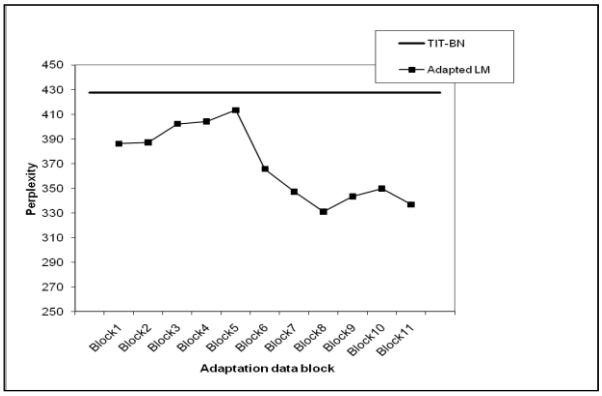

Figure 2: The perplexities of the base LM (TIT-BN) and the adapted LMs

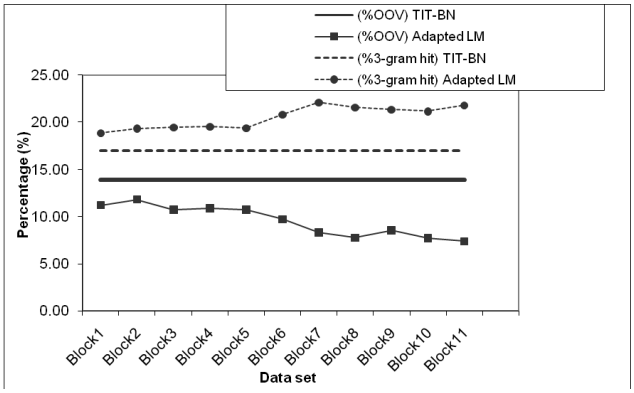

Figure 3: The OOV rates and the 3-gram hit rates of the base LM (TIT-BN) and the adapted LMs

Since a larger amount of adaptation data might contribute to more perplexity reduction, we conduct another experiment to identify an appropriate amount of adaptation data. In this experiment, we accumulated data from consecutive blocks. Since the data that is closer to the test set is better adaptation data, we started with Block11. Figure4 shows the perplexities of the adapted LMs when more data from earlier news programs were utilized. The last data point presents the perplexity when all data in the NECTEC-BN set was used.

More adaptation data can further reduce perplexity up to one point. We found that the optimal number of adaptation data is four blocks (200K words) which can achieve 30.3% relative perplexity reduction from the base, uninterpolated, LM. After this point more adaptation data did not seem to help and even slightly increased the perplexity. This finding also confirms that the data that are much further away are less useful for temporal adaptation which makes us consider adjusting the base LM over time in the future.

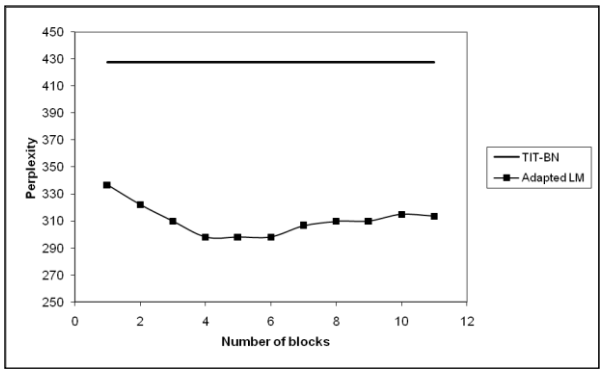

Figure 4: The effect of the amount of adaptation data on the perplexities of adapted LMs

5. Temporal adaptation results

We applied our online temporal adaptation technique to adapt a language mode in a broadcast news transcription task. Three LMs are compared: the base LM trained from the TIT-BN set (baseline), the LM adapted with four blocks of the most recent broadcast news transcription (Adapted-BN) and the LM adapted with four blocks of the most recent online news texts (Adapted-NT). For an acoustic model, we adopted the one used in (Jongtaveesataporn et al., 2008). This acoustic model was trained from TIT-BN and Thai read speech data.

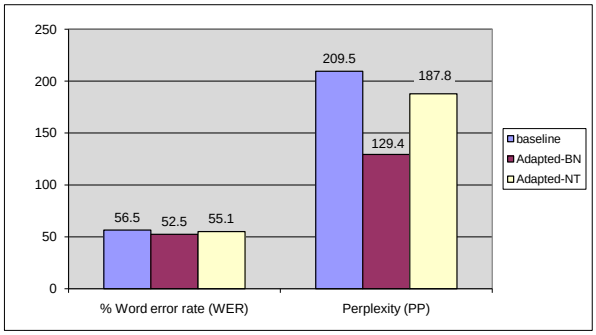

Figure 5: WERs and perplexities of the base LM and two temporally adapted LMs

The NewsText data was divided into blocks in the same way as the NECTEC-BN data. Each block contains about 500 articles; hence, four block of adaptation data is equivalent to 2,000 articles (450K words). Since there is no word boundary marker in Thai texts, we used TLex (Haruechaiyasak & Kongyoung, 2009), a Thai word segmentation tool based on the Conditional Random Fields, to automatically identify word boundaries in NewsText. For new lexicons that are not in the LOTUS-BN pronunciation dictionary, we generated their pronunciations using an automatic grapheme-to-phoneme conversion (G2P) (Thangthai et al., 2007).

In the main experiment, the interpolation weights were determined from a held out set. We used the last block before the test set (Block11) as the held out set and used data from Block7 to Block 11 as adaptation data.

Both adapted LMs achieved better WERs and perplexities than the baseline as shown in Figure5. In terms of WER, the Adapted-BN model gained more relative improvement (7.1%) over the baseline compared to the Adapted-NT model (2.4%). Both adapted LMs can reduce the out of vocabulary (OOV) rate from 5% to 3% approximately. However, the Adapted-BN model has a better 3-gram hit rate. This might due to the difference between spoken and written styles even both sets of data are from a news domain. Furthermore, the online news articles contain some symbols and numbers which required an additional pre-processing step.

A set of topics covered by the NewsText data might be another reason that makes the Adapted-NT model less efficient than the Adapted-BN model. The NewsText data consists of only four news topics while the NECTEC-BN data consists of as much as eighteen topics. There are many topics in the broadcast news data that are not covered by the online news articles such as foreign news and sport. Therefore, the part of the language model that involves these news topics is not adapted in the Adapted-NT model. The interpolation weights seem to confirm with this argument as more weight was given to the baseline model than to the adaptation data (i.e. 0.7 for the baseline model vs. 0.3 for the adaptation data). In the case of the Adapted-BN model, on the other hand, more weight was give to the adaptation data than to the baseline model (i.e. 0.35 for the baseline model vs. 0.65 for the adaptation data).

We note that the WER on our test set is quite higher than the WER of the Thai broadcast news transcription task reported in (Jongtaveesataporn et al., 2008). We found that the OOV rate in our test set is higher (3% versus 1.2%). Another possible reason is acoustic characteristic mismatch between the acoustic model training data and the test data as the TIT-BN set was recorded from an analog TV antenna while the Test-BN set was recorded from a satellite. Both of them were also recorded from different channels. Acoustic model adaptation is also required in order to get a better result.

The perplexities in this experiment are lower than those in previous experiments. We found that the automatic word segmentation produces smaller word units than the manual segmentation in the transcription. The smaller unit size leads to lower OOV rate and higher 3-gram hit rate which decrease perplexity and result in a slight improvement on WER. This might indicate that a small word unit is more suitable for a recognition task. Nevertheless, more investigation needs to be done in order to properly conclude this issue.

6. Conclusion

This paper investigates the effectiveness of online temporal language model adaptation when applied to a Thai broadcast news transcription task. The LMs that are adapted over time with more recent news data are better, both in terms of perplexity and WER, than the static LM trained from only the initial set of broadcast news data. Adaptation data from broadcast news transcription improved perplexity by 38.3% and WER by 7.1% relatively. Though, online news articles achieved less improvement, it is still a useful resource as it can be obtained automatically.

In the future, we plan to investigate on how to extract information in written news texts that is useful for modeling broadcast news language. Data selection based on text similarity is one of the techniques that would be worth to investigate. We are also interested in exploring more sophisticated interpolation techniques and weighting schemes.

7. Acknowledgements

We would like to thank Mr. Markpong Jongtaveesataporn, a Ph.D. candidate at Department of Computer Science, Tokyo Institute of Technology, for his generous help and advice on acoustic model training.

8. References

Chen, L., Gauvain, J., Lamel, L., and Adda, G. (2004). Dynamic Language Modeling for Broadcast News. In *Proceedings of INTERSPEECH 2004-ICSLP*.

Chotimongkol, A., Saykhum, K., Chootrakool, P., Thatphithakkul, N., and Wutiwiwatchai, C. (2009). LOTUS-BN: A Thai Broadcast News Corpus and Its Research Applications. In *Proceedings of Oriental-COCOSDA'09*.

Clarkson, P. and Rosenfeld, R. (1997). Statistical Language Modeling Using the CMU-Cambridge Toolkit. In *Proceedings of EuroSpeech'97*.

Federico, M., and Bertoldi, N. (2004). Broadcast news LM adaptation over time. *Computer Speech and Language*, 18(4), pp. 417—435.

Graff, D. and Alabiso, J. (1997). 1996 English Broadcast News Transcripts (HUB4). Linguistic Data Consortium, Philadelphia

Haruechaiyasak, C., and Kongyoung, S. (2009). TLex: Thai Lexeme Analyser Based on the Conditional Random Fields. In *Proceedings of SNLP'09*.

Jongtaveesataporn, M., Wutiwiwatchai, C., Iwano, K., Furui, S. (2008). Thai broadcast news construction and evaluation. In *Proceedings of International Conference on Large Resources and Evaluation*.

Matsui, A., Segi, H., Kobayashi, A., Imai, T., and Ando, A. (2001). Speech Recognition of Broadcast Sports News. Laboratories Note No. 472. NHK Laboratories.

Thangthai, A., Wutiwiwatchai, C., Rugchatjaroen, A., and Saichum, S. (2007). A Learning Method for Thai Phonetization of English Words. In *Proceedings of INTERSPEECH'07*.

Whittaker, E. (2001). Temporal Adaptation of Language Models. In *Proceedings of ISCA Adaptation Methods for Speech Recognition Workshop*.