N-GRAM LANGUAGE MODEL ADAPTATION USING SMALL CORPUS FOR SPOKEN DIALOG RECOGNITION

Akinori Ito, Hideyuki Saitoh, Masaharu Katoh and Masaki Kohda
Faculty of Engineering, Yamagata University
Jonan 4-3-16, Yonezawa, Yamagata 992 Japan
TEL&FAX +81 238 26 3369 Email: aito@ei5sun.yz.yamagata-u.ac.jp

10.21437/Eurospeech.1997-690

ABSTRACT

This paper describes an N-gram language model adaptation technique. As an N-gram model requires a large size sample corpus for probability estimation, it is difficult to utilize N-gram model for a specific small task. In this paper, *N-gram task adaptation* is proposed using large corpus of the general task (TI text) and small corpus of the specific task (AD text). A simple weighting is employed to mix TI and AD text. In addition to mix two texts, the effect of vocabulary is also investigated. The experimental results show that adapted N-gram model with proper vocabulary size has significantly lower perplexity than the task independent models.

1.INTRODUCTION

N-gram based language model is popular for continuous speech recognition. N-gram shows good performance as a language model for speech recognition as far as a large corpus of the task domain is available. When you make CSR system with N-gram LM for a specific task domain, you have to gather large number of sentences belonging to the domain, which is not always easy. It is desired to make good LM of the domain from small number of sentences.

One way to estimate an LM from small data is to adapt general LM to the task using the data. This technique is called `task adaptation', which is comparable with `speaker adaptation' in speech recognition. There are several types of task adaptation. We investigated the task adaptation which mixes an LM from small corpus of the specific task to an LM from large general corpus[1,2].

2.TASK ADAPTATION

2.1 Task adaptation using weighted mixture of N-gram

We use three corpora for adaptation and evaluation of N-gram. These are a task independent (TI) text , a text for adaptation (AD text) and a text for evaluation. When a word w occurs $N_I(w)$ times in the TI text and $N_A(w)$ times in the AD text, unigram probability $P(w)$ is calculated as follows[2].

$$P(w) = \frac{N_I(w) + \gamma N_A(w)}{\sum_w \left(N_I(w) + \gamma N_A(w)\right)}$$

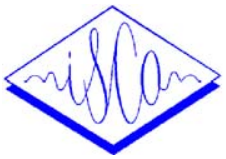
γ is a weight to mix the TI and AD texts. A word with m occurrences in the AD text is treated to be equivalent to a word with γm occurrences in the TI text. Bigram and trigram probabilities are calculated in the similar way.

In addition to mix two texts, we investigated the effect of vocabulary restriction. We examined following two vocabulary restriction methods.

2.2 Weight-and-restrict adaptation

Weight-and-restrict (WR) adaptation procedure is as follows.

1. Add texts in the adaptation corpus to the TI corpus with weight γ.
2. Replace low frequency words in the mixed texts by a unique symbol <UNK>. (Vocabulary restriction)



EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

3. Create N-gram model from the mixed text with vocabulary restriction.

FIGURE 1 shows the block diagram of these procedure.

2.3 Restrict-and-weight adaptation

Restrict-and-weight (RW) adaptation procedure is as follows.

1. Replace low frequency words in TI text and AD text by a unique symbol <UNK>. The thresholds of the vocabulary cut-off for each text are determined individually. As for the AD text, words in the restricted vocabulary of TI text are not eliminated even if their occurrences are lower than the threshold.
2. Add texts in the adaptation corpus to the TI corpus with weight γ.
3. Create N-gram model from the mixed text with vocabulary restriction.

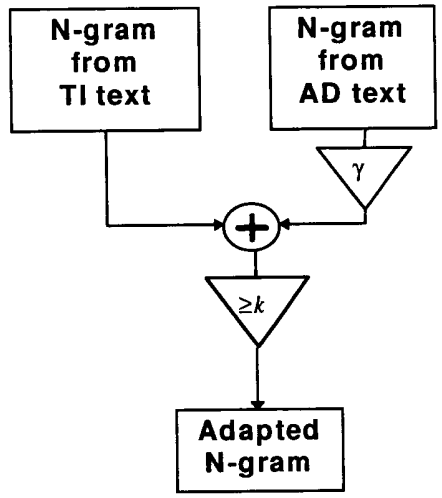


FIGURE 1: Block diagram of WR adaptation.

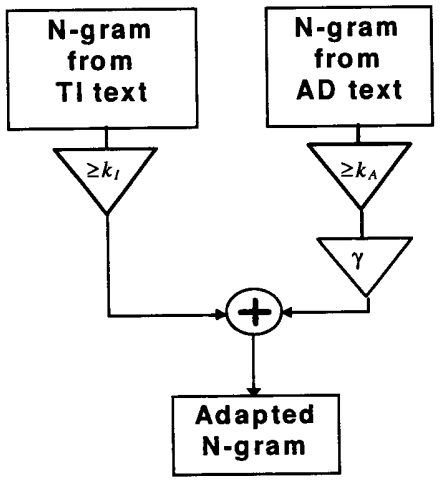


FIGURE 2: Block diagram of RW adaptation

TABLE 1: Task independent corpora

corpus	*#sentence*	*#word*
ASJ	3000	96779
ATR	5000	69677
EDR	5000	125960

TABLE 2: Corpora for adaptation and evaluation

dialog	*#sentence*	*#word*
Kyoto 1	117	1542
Kyoto 2	117	1254
Kyoto 3	131	1598
Kyoto 4	97	1120

FIGURE 2 shows the block diagram of these procedure.

2.4 What does vocabulary restriction means?

In the previous works[1,2], effect of vocabulary wa not investigated. If the TI text has almost nothing to do with the adaptation and evaluation text, most of words in the TI text are not only unnecessary but harmful for the speech recognition of the target task. In our work, most of unnecessary words in the TI text are eliminated by the vocabulary restriction.

Another effect of vocabulary restriction is automatic classification of words. In our vocabulary restriction scheme, <UNK> stands for a low-frequency word class. For example, a sentence in the TI text

What kind of restaurant do you like? (a)

becomes

What <UNK> *of* <UNK> *do you like?* (b)

by vocabulary restriction[1]. In the AD text, words *grade, kind, hotel, temple, shrine, ...* are classified into <UNK> class. Therefore, sentence (b) in TI text is equivalent to a sentence

[1] In the later experiment, we used Japanese corpus.

What kind of temple do you like?
in the AD text.

3.CORPORA FOR TASK ADAPTATION

We used following corpora for task adaptation experiment.

3.1 Task independent text

We used three corpora as TI texts. First is a subset of ASJ corpus, transcription of human to human spoken dialogue. Second is ATR dialogue corpus, human to human keyboard conversation about an international conference. Third is a subset of EDR corpus, sentences from newspapers and magazines. Number of sentences and words are shown in TABLE 1.

3.2 Adaptation and evaluation text

Adaptation and evaluation texts are four dialogues from ASJ corpus (they are not included in the TI texts). Their task is sightseeing information of Kyoto city. Three dialogues are used for adaptation (dialog "Kyoto 2", "Kyoto 3" and "Kyoto 4") and one dialogue is used for evaluation (dialog "Kyoto 1"). Number of sentences and words in each dialogue are shown in TABLE 2.

4.TASK ADAPTATION EXPERIMENTS

Task adaptation experiments were carried out. In the experiment, trigram model with linear back-off [3] was employed. To evaluate the effect of task adaptation, we compared adjusted perplexity (APP) [4] of adapted and non-adapted models.

4.1 N-grams from TI text or AD text

First, we measured APP of models created from the TI texts. The results for optimum vocabulary size are shown in TABLE 3. "≥15" in the table means that words with less than 15 occurrences are replaced with <UNK>.

Second, APP was calculated using trigram models created from adaptation texts. The results are shown in TABLE 4.

TABLE 3: Perplexity of trigram from TI texts

TI text	*vocabulary*	*APP*
ASJ	≥15	375.0
ATR	≥40	647.6
EDR	≥200	2035.5

TABLE 4: Perplexity of trigram from AD texts

AD text	*vocabulary*	*APP*
Kyoto 2	≥3	85.51
Kyoto 2+3	≥3	54.32
Kyoto 2+3+4	≥4	45.69

TABLE 5: Perplexity of WR adapted models

TI text	*AD text*	*γ*	*vocab*	*APP*
ASJ	Kyoto 2	4	≥20	44.0
ASJ	Kyoto 2+3	4	≥20	41.3
ASJ	Kyoto 2+3+4	4	≥20	38.1
ATR	Kyoto 2	16	≥70	50.5
ATR	Kyoto 2+3	14	≥100	44.2
ATR	Kyoto 2+3+4	12	≥100	40.6
EDR	Kyoto 2	30	≥200	87.5
EDR	Kyoto 2+3	14	≥100	65.3
EDR	Kyoto 2+3+4	8	≥40	57.1

4.2 Adapted N-gram by WR adaptation

TI text and adaptation text were mixed according to the WR adaptation procedure. The results for optimum weight and vocabulary are shown in TABLE 5.

When we used ASJ or ATR corpus as a TI text, APP of the model was improved compared with models in TABLE 3 and 4. On the contrary, when we used EDR corpus as TI text, the result was worse than the model created from adaptation texts. This results seem to be caused by the difference between TI text and adaptation text. Sentences in ASJ and ATR are dialog (ATR is a keyboard dialog corpus) while sentences in EDR corpus are written language. The difference of speaking (writing) style affected the performance of adapted model.

4.3 Adapted N-gram by RW adaptation

Finally, task adaptation by RW adaptation

TABLE 6: Perplexity of RW adapted models

AD text	kyoto 2	kyoto 2+3	kyoto 2+3+4
TI vocab	≥18	≥18	≥18
AD vocab	≥5	≥8	≥8
weight	14	10	8
APP	42.3	39.3	36.5

procedure was investigated. In this experiment, only ASJ corpus was used for TI text. The results for optimum TI vocabulary, AD vocabulary and weight are shown in TABLE 6. This experiment proved that perplexity of a RW model was lower than that of a WR model. From the result, it is found that the vocabulary sizes of TI and AD text have to be determined individually.

Figure 3 and 4 shows APP of WR and RW models when TI vocabulary ≥18 and AD text is kyoto2. APP of WR model draws irregular curve against weight. It is caused by the vocabulary of AD text. In the WR model, when TI vocabulary is $\geq n$ and weight is γ, words in AD text with less than $\lfloor n/\gamma \rfloor$ occurrences are replaced to <UNK>. As occurrence of a word is discrete, APP of WR model jumps when $\lfloor n/\gamma \rfloor$ changes.

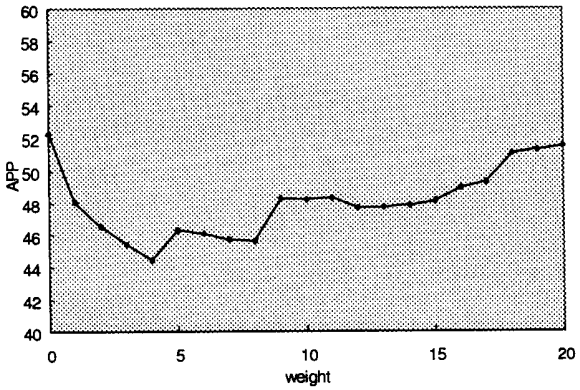


FIGURE 3: Weight and APP of WR model when TI vocabulary ≥18

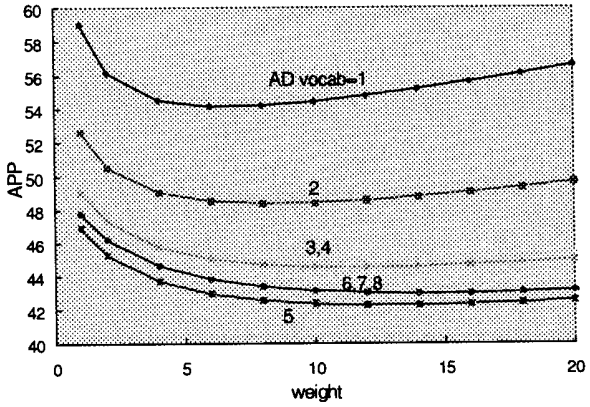


FIGURE 4. Weight and APP of RW model when TI vocabulary ≥18

5.CONCLUSION

We investigated task adaptation using task independent corpus and small task dependent text. From the experimental results, it is shown that

- The task adaptation using a task independent corpus and an adaptation corpus improves the performance of the model when the style of TI corpus is similar to the target.
- Vocabulary restriction improves the perplexity of the model.
- Vocabulary size of TI and AD text must be determined individually.

We are going to prove the effect of task adaptation by speech recognition experiment.

6.REFERENCES

[1] A. I. Rudnicky: "Language modeling with limited domain data", Proc. ARPA Spoken Language Systems Technology Workshop, pp.66-69 (1995-1)

[2] M. Federico: "Bayesian Estimation Methods for N-gram Language Model Adaptation", Proc. ICSLP-96, pp.240-243 (1996)

[3] P. Placeway et al. :"The estimation of powerful language models from small and large corpora", Proc. ICASSP-93, vol. II, pp.33-36 (1993)

[4] J. Ueberla: "Analyzing a simple language model - some general conclusion for language models for speech recognition", Computer Speech and Language, vol.8, no.2,pp.153-176 (1994-4)