A LANGUAGE MODEL ADAPTATION USING MULTIPLE VARIED CORPORA

Hirofumi Yamamoto †*and Yoshinori Sagisaka* †,‡

†ATR Spoken Language Translation Research Labs
2–2, Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0288 Japan
yama@slt.atr.co.jp
‡Graduate School of Global Information and Telecommunication Studies
1-3-10, Nishi-Waseda, Shinjuku-ku, Tokyo-to, 169-0051 Japan
sagisaka@giti.waseda.ac.jp

ABSTRACT

A new language model adaptation scheme is proposed to cope with multiple varied speech recognition tasks. Both topic difference and sentence style difference resulting from the speaker's role are reflected in the proposed language model adaptation. An adaptation is carried out using two different language corpora where only the topic or speaker's style is matched.

New word clustering techniques are introduced to extract the topic or style dependency separately. Word neighboring characteristics in the two adaptation source data regarded as different features in this clustering. All words are classified into commonly use word classes and topic or style dependent classes. Furthermore, target topic and sentence style dependent words and their neighboring characteristics are emphasized according to their frequency in the adaptation target data.

In the evaluation experiment, the proposed method shows a 13% lower perplexity and a 9% lower word error rate in continuous speech recognition compared with the conventional adaptation method.

1. INTRODUCTION

Language model adaptation has been efficiently applied to a varied task where only a small amount of language corpora is available. However, in spoken language, there exist multiple variance factors which are not so commonly seen in written language corpora. For speech-to-speech translation, we have been studying the speech recognition task where multiple varied adapting factors exist. For example, in conversations between customers and hotel employees in various situations, not only is there a task difference between reservations and trouble-shooting, but also the speaker's role cause a huge difference in language use. Though these multiple varied adaptation factors are widely seen in conversational speech, little attention has been paid to them until now.

In many cases when multiple variance factors should be taken into account, it is quite difficult to collect enough of an adaptation corpus that matches to target conditions. For example, in the above described instance, you may be able to collect only a very small amount of conversation data in trouble-shooting. To cope with this data sparseness problem caused by multiple variance factors, an adaptation using multiple source corpora can be carried out. We can expect the use of polite and kind language by hotel employee in the trouble-shooting domain, and these language statistical properties can be partially obtained from other conversation corpora in different topics. On the other hand, task specific vocabularies used in trouble-shooting can be obtained from another language corpora not necessarily collected from employee conversations.

In this paper, a new adaptation scheme is proposed to cope with multiple varied factors. Factor-dependent, large-sized multi corpora are used in stead of a exactly matched, very small corpus.

2. MULTIPLE DOMAIN ELEMENTS AS TOPIC AND SENTENCE STYLE VARIANCE

2.1. Data sparseness problem in the multiple varied source corpora

When multiple varied corpora are used in adaptation, the adaptation source and target corpora shown in Table 1 are used. In the adaptation source corpora, the topic is matched and the speech style is mismatched, or the topic is mismatched and the speech style is matched. We must recover these mismatches using very small amount of adaptation target corpora. In the conventional word N-gram based adaptation, word neighboring characteristics that are observed in the target corpora are reflected on the adapted model. However, only the observed word sequence in the adaptation target data is reflected. For a word sequence that is not observed in the target data, the probabilities in the source data are directly used. Therefore, mismatched language characteristics between source and target corpora cause a very serious data sparseness problem.

Table 1. Combination of topic and speaker's role in adaptation source and target corpora

Topic Speaker	Hotel Reservation	Trouble-shooting
Employee		Source Corpora 1
Customer	Source Corpora 2	Target Corpora

2.2. Class N-gram based adaptation

To avoid data sparseness in the previous subsection, a class N-gram based adaptation approach is proposed [1]. In class N-grams, word transition probabilities are given by the next formula.

$$P(c(w_i)|c(w_{i-1}))P(w_i|c(w_i)) \tag{1}$$

Where, $c(w)$ represents the class in which word w belongs. Adaptation for the first term is performed for all of the class pairs to which the observed word pairs belong. The coverage of class pairs is wider than that of word pairs. Therefore, effective adaptation can be expected even if the adaptation data is insufficient. However, good results cannot be obtained with unmatched word classes for the adaptation target.

2.3. The class mismatched problem in the class N-gram based adaptation

Conventional word clustering is performed based on only one set of corpora. Therefore, a mixture of source corpora 1 and 2 in the Table 1 is used. In combinations of source and target corpora like those in Table 1, these word clusters cannot maintain the style or topic dependent word characteristics in the following case.

Suppose our hotel reservation corpus consists of the following two sentences.

- What will be my **room number** ?
- Is my **room size wide** ?

Similarly, the trouble-shooting corpus has the following three sentences.

- This is **room number** 1234.
- The **room noise** is intolerable.
- The **room temperature** is too hot.

In hotel reservation, the word sequence "room number" and "room size" are used. In trouble-shooting, "room number," "room noise" and "room temperature" are used. If we combine these corpora, then "number," "size," "noise" and "temperature" will be classified to the same cluster, since these four words share the same preceding word "room." In this word class, the observed word sequence "room noise" in the adaptation target corpora will emphasize from "room" to all of the four words, however only "room noise" and "room temperature" must be emphasized. The loss of the topic dependent word neighboring characteristics makes this unreasonable emphasis. Therefore, the neighboring characteristics in each topic must be handled separately.

3. WORD CLUSTER CONSIDERING MULTIPLE DOMAIN ELEMENTS

In the usual class N-gram, the following features for representing word neighboring characteristics are used for clustering [2].

$$\begin{aligned} V(X) \quad = \quad & [\{P_f(w_1|X), P_f(w_2|X), ..., P_f(w_N|X)\}, \\ & \{P_b(w_1|X), P_b(w_2|X), ..., P_b(w_N|X)\}] \end{aligned} \tag{2}$$

where, $P_f(w|X)$ expresses the forward word 2-gram from word X to w, and $P_b(w|X)$ expresses the backward word 2-gram.

When Multi-Classes [3] are used for word classes, the features of neighboring characteristic are given by the following equation, since forward and backward classes are separately clustered.

$$V(X) = [\{P(w_1|X), P(w_2|X), ..., P(w_N|X)\}] \tag{3}$$

where, $P(w|X)$ expresses the forward or backward word 2-gram from word X to w in mixed corpora 1 and 2.

Furthermore, to maintain the corpora 1 and 2 dependent (speaker and topic dependent) characteristics, the word 2-gram in corpora 1 and 2 must be regarded as different characteristics. The features for representing these separated characteristics are expressed by the following equation.

$$\begin{aligned} V(X) \quad = \quad & [\{P_1(w_1|X), P_1(w_2|X), ..., P_1(w_N|X)\}, \\ & \{P_2(w_1|X), P_2(w_2|X), ..., P_2(w_N|X)\}] \end{aligned} \tag{4}$$

Where, $P_1(w|X)$ expresses the word 2-gram in corpora 1, and $P_2(w|X)$ expresses the word 2-gram in corpora 2. The corpora 1 and 2 dependent characteristics are considered in this feature.

In fact, different features are assigned to "number," "size" and "noise," and the same features are assigned to "noise" and "temperature" in the previous subsection.

$$\begin{aligned} V(size) &= [\{P_{Reservation}(room|size)\}, \\ &\quad \{P_{Trouble}(room|size)\}] \\ &= [\{1\}, \{0\}] \\ V(number) &= [\{1\}, \{1\}] \\ V(noise) &= [\{0\}, \{1\}] \\ V(temperature) &= [\{1\}, \{1\}] \end{aligned} \tag{5}$$

These features give the same word class to only "noise" and "temperature". The observed word sequence "room noise" in the adaptation target corpora will only emphasize "room noise" and "room temperature."

We call this word clustering method multi-dimensional word clustering (MDWC). MDWC is performed in the following manner.

1. Assign one class per word.
2. Assign a feature vector $V(X)$ to each class or to each word X according to equation 4.
3. Merge the two classes. We choose classes that result in the lowest merge cost $U_{new} - U_{old}$, and merge these two classes:

$$U_{new} = \sum_X (p(X)D(V(c_{new}(X)), V(X))) \tag{6}$$

$$U_{old} = \sum_X (p(X)D(V(c_{old}(X)), V(X))) \tag{7}$$

   Where, $D(V_c, V_X)$ represents the square of the Euclidean distance between vectors V_c and V_X, c_{old} represents the classes before the merging, and c_{new} represents the classes after the merging. $p(X)$ represents the word 1-gram of X in mixed corpora 1 and 2.

4. Repeat step 2 until the number of classes is reduced to the desired number.

4. ADAPTATION BASED ON MDWC

After MDWC, class based adaptation is performed. In this adaptation, MAP estimation [4] is employed with the adaptation source corpora as a-priori knowledge and the target as a-posteriori knowledge. The transition probability in the word N-gram after the MAP estimation is represented by the following equation.

$$P_{adapt}(X|Y) = \frac{\lambda \times C^T(Y,X) + C^S(Y,X)}{\lambda \times C^T(Y) + C^S(Y)} \quad (8)$$

where, $C^T(A)$ represents the occurrence of word A in the adaptation source corpora, and $C^S(A)$ represents the same in the target corpora. λ is constant, i.e., it is fixed in our experiment.

Next, this estimation is applied to a class N-gram based on Multi-Classes. In the Multi-Class based N-gram, the transition probability is represented by the following equation.

$$P(X|Y) = P(c_t(X)|c_f(Y))P(X|c_t(X)) \quad (9)$$

where, $c_t(X)$ represents the backward class of word X, and $c_f(X)$ represents the forward class.

The probabilities $P(c_t(X)|c_f(Y))$ and $P(X|c_t(X))$ after the MAP estimation are given by the following equations by applying equation 8 to 9.

$$P_{adapt}(c_t(X)|c_f(Y)) = \frac{\lambda \times C^T(c_f(Y), c_t(X)) + C^S(c_f(Y), c_t(X))}{\lambda \times C^T(c_f(Y)) + C^S(c_f(Y))} \quad (10)$$

$$P_{adapt}(X|c_t(X)) = \frac{\lambda \times C^T(X) + C^S(W)}{\lambda \times C^T(C) + C^S(c_t(X))} \quad (11)$$

In equations 10 and 11, the numerator is 0 when no data is observed in both the source and target corpora. For this case, back-off smoothing based on the Good-Turing discount is used. Discounting is applied when the occurrence is smaller than k. In contrast, if $\lambda > k$, any data observed in the target corpora at least one time is not discounted.

5. EVALUATION

5.1. Evaluation of Perplexity

We evaluated the proposed scheme in terms of perplexity. The experiment data is a Japanese conversational corpora selected from the ATR Spoken Language Database [5]. These conversations including two different topics, hotel reservation and conversations with front desk, including room service and trouble-shooting. There are two talkers, a customer and a clerk. These conversations include two different speech styles since the speech style of the clerk is more polite than that of the customer.

First, we evaluate speaker and topic dependency in perplexity. The adaptation target corpora is a clerk in a conversation with the front desk. Source corpora 1 is a clerk in a hotel reservation, where the speaker is matched and the topic is unmatched. Source corpora 2 is a customer in a conversation with the front desk, where the speaker is unmatched and the topic is matched. The target corpora is separated into three parts. The first 40 conversations are used for adaptation, and the next 195 are used for evaluation. The last 401 are used to create a target task dependent model for comparison, since this model gives the upper limit of the adapted model.

Table 2. Experiment Conditions

	Topic	Sentence Style	Number of Conversations
Source 1	Reservation	Clerk	544
Source 2	Front Desk	Customer	485
Target	Front Desk	Clerk	40
Evaluation	Front Desk	Clerk	195
Task Dependent	Front Desk	Clerk	401

Table 3. Style and Topic Dependency

Source 1	69.01
Source 2	150.70
Task Dependent	25.51

The size of each corpora is shown in Table 2. The perplexity in evaluation data, using word 2-grams created from source corpora 1, source corpora 2, and the comparison data are shown in Table 3. This table shows that the style and the topic dependency is quite heavy, especially in the style dependency.

We evaluate the adapted models. We compared non-adapted model using mixed corpora with source corpora 1 and 2, and three adapted models, consisting of the conventional word 2-gram based adaptation, the class 2-gram based adaptation, and the proposed MDWC based adaptation. Through all of the method, the Good-Turing discount with $k = 5$ is used and λ is 30. In the class 2-gram based model and MDWC based model, the number of word classed is 600, and this provides the lowest perplexity in evaluation data. Perplexity in each model are shown in Table **??**. The proposed MDWC based method results in a 13% lower perplexity than the conventional word 2-gram based adaptation.

5.2. Evaluation of Continuous Speech Recognition

Next, we evaluate each model in continuous speech recognition. The conditions of the experiment were as follows.

- Evaluation set
  - The same 195 conversations as used in the evaluation of perplexity
- Acoustic features
  - Sampling rate of 16 kHz
  - Frame shift of 10 msec
  - Mel-cepstrum of 12 + power and their delta, for a total of 26
- Acoustic models
  - 800-state 5-mixture HMnet model based on ML-SSS [6]
  - Automatic selection of gender dependent models

Table 4. Comparison with Conventional Method

Adaptation Method	Perplexity
Word 2-gram Non-Adaptation	49.34
Word 2-gram Adaptation	42.26
Class 2-gram Adaptation	37.72
MDWC Base Adaptation	36.02

Table 5. Evaluation in Continuous Speech Recognition

Adaptation Method	Acc.	%Corr.
Word 2-gram Non-Adaptation	72.7	77.8
Word 2-gram Adaptation	75.8	80.5
Class 2-gram Adaptation	77.1	81.0
Upper Limit (Task Dependent)	81.7	84.7
MDWC Base Adaptation	77.9	81.7

- Decoder [7]
  - 1st pass: frame-synchronized viterbi search
  - 2nd pass: full search after changing the language model and LM scale

The evaluation measures are conventional word accuracy and %correct cluculated as follows.

$$WordAccuracy = \frac{W - D - I - S}{W} \times 100$$

$$\%Correct = \frac{W - D - S}{W} \times 100$$

(W: Number of correct words, D: Deletion error, I : Insertion error, S: Substitution error)

The word accuracy and %correct are shown in Table 5. The proposed MDWC based adapted model resulted in the highest performance in all of the adapted models with respect to perplexity, and the error reduction compared with the conventional word 2-gram based adaptation is about 9%.

6. CONCLUSION

A new language model adaptation scheme is proposed for a multiple varied recognition task. In the proposed method, not only topics but also sentence styles reflecting speaker's difference are regarded as adaptation factors. In the conventional adaptation, only a topic independent corpora is used as the adaptation source corpora. In the proposed scheme, a combination of two sets of adaptation source corpora are used. In one, only the topic is matched and the speech style is unmatched. In the other, only the style is matched and the topic is unmatched.

These two different corpora cause the following two new problems.

- A data-sparseness problem caused by mismatch between the adaptation source and target corpora.
- A new method is required to extract topic and speech style dependent word neighboring characteristics separately.

Class N-gram based adaptation is applied to resolve the data sparseness problem. Next, word neighboring characteristics features are introduced to carry out multi-dimensional word clustering (MDWC). MDWC can classify all words into commonly used word classes and topic or style dependent classes. Furthermore, target topic and sentence style dependent word neighboring characteristics are emphasized according to their frequency in the adaptation target corpora.

In the evaluation experiment, the proposed method showed a 13% lower perplexity and 9% lower word error rate in continuous speech recognition compared with the conventional word 2-gram base adaptation method.

7. REFERENCES

[1] G. Moore, S. Young: "Class-based language model adaptation using mixture of word-class weight," Proc. ICSLP-2000, vol. 4, pp. 512-515 (2000).

[2] S. Bai, H. Li, B. Yuan: "Building Class-based Language Models with Contextual Statistics," Proc. ICASSP-98, pp. 173-176 (1998).

[3] H. Yamamoto, Y. Sagisaka: "Multi-Class Composite N-gram Based on Connection Direction," Proc. ICASSP-99, pp. 533-536 (1999).

[4] H. Masataki, Y. Sagisaka, K. Hisaki, T. Kawahara: "Task Adaptation Using MAP Estimation in N-gram Language Modeling," Proc. ICASSP-97, pp. 783-786 (1997).

[5] T. Takezawa, T. Morimoto, Y. Sagisaka: "Speech and Language Databases for Speech Translation Research in ATR," Proc. of the 1st International Workshop on East-Asian Language Resource and Evaluation (1998).

[6] M. Ostendorf, H. Singer: "HMM topology design using maximum likelihood successive state splitting," Computer Speech and Language, vol. 11, pp. 17-41 (1997).

[7] Tohru Shimizu, Hirofumi Yamamoto, Hirokazu Masataki, Shoichi Matsunaga, and Yoshinori Sagisaka: "Spontaneous sialog speech recognition using cross-word context constrained word graphs," Proc. ICASSP-96, pp. 145-148 (1996).