UNSUPERVISED CV LANGUAGE MODEL ADAPTATION BASED ON DIRECT LIKELIHOOD MAXIMIZATION SENTENCE SELECTION

Takahiro Shinozaki, Yasuo Horiuchi, Shingo Kuroiwa

Division of Information Sciences, Graduate School of Advanced Integration Science, Chiba University

ABSTRACT

Direct likelihood maximization selection (DLMS) selects a subset of language model training data so that likelihood of in-domain development data is maximized. By using recognition hypothesis instead of the in-domain development data, it can be used for unsupervised adaptation. We apply DLMS to iterative unsupervised adaptation for presentation speech recognition. A problem of the iterative unsupervised adaptation is that adapted models are estimated including recognition errors and it limits the adaptation performance. To solve the problem, we introduce the framework of unsupervised cross-validation (CV) adaptation that has originally been proposed for acoustic model adaptation. Large vocabulary speech recognition experiments show that the CV approach is effective for DLMS based adaptation reducing 19.3% of error rate by an initial model to 18.0%.

Index Terms— Cross-validation, language model, sentence selection, relative entropy, unsupervised adaptation

1. INTRODUCTION

In deployment of speech recognition systems, it is often required to train a high performance language model using an existing domain independent data. Since such training data contains sentences that are not relevant to the recognition task, it is important to select a subset of the data to make a domain-matched language model. When in-domain development data is available, a strategy is to make a model based on the development data, and select sentences in the training data that gives higher likelihood with that model [1]. However, this strategy has an essential problem that the most frequent patterns in the development data are excessively emphasized in the selected training subset [2].

To avoid the problem, another selection strategy has been proposed that selects a subset of training data so as to directly maximize development set likelihood [3, 4]. We refer to this method as Direct Likelihood Maximization Selection (DLMS). A closely related method is minimum relative entropy based selective adaptation [2] that minimizes relative entropy or KL divergence from a language model estimated on development data to a language model estimated on selected training data.

While these methods had originally been proposed for domain adaptation using development data, they can be applied to unsupervised adaptation by first running a speech recognizer to generate a recognition hypothesis using an initial model and making an adapted model using that hypothesis as adaptation data. This two-pass approach has been tested in a speech-to-speech automatic translation system using a machine translation output as adaptation data and has been shown to be effective [5].

The two-pass unsupervised adaptation approach can be easily extended to multi-pass adaptation using the adapted model as an initial model for the next pass for higher recognition performance. However, a problem is that recognition errors are unavoidable in a recognition hypothesis and the errors are reinforced during the iteration. Therefore, the adaptation performance is limited. To solve the problem, we introduce the unsupervised cross-validation (CV) adaptation framework that has originally been proposed for acoustic model adaptation [6] to selective language model adaptation.

The organization of the rest of this paper is as follows. In Section 2, DLMS is reviewed and relation to the minimum relative entropy method is explained. In Section 3, the framework of unsupervised CV adaptation is shown. Experimental conditions are described in Section 4 and the results are shown in Section 5. Conclusions are given in Section 6.

2. DIRECT LIKELIHOOD MAXIMIZATION SELECTION

Direct likelihood maximization selection (DLMS) has been proposed by Klakow [3]. It selects articles in a training set so as to maximize adaptation data likelihood based on an objective function shown in Equation (1).

$$L = \sum_{w} C_A(w) \log P_T(w), \tag{1}$$

where $C_A(w)$ is count of word w in an adaptation set and P_T is a language model estimated on a selected training subset. Since testing all the possible subsets of a training data is not feasible, greedy strategies are adopted. In [3], first an importance of an article is scored by the objective function (1) making a model using a whole training set removing that article. Lower score means that article is important for the adaptation

set. After all the articles are scored independently, N articles with the lowest scores are selected and an adapted language model is made.

The relative entropy method proposed by Sethy et al. [2] is based on minimizing relative entropy shown in Equation (2).

$$R = \sum_w P_A(w) \log \frac{P_A(w)}{P_T(w)}, \quad (2)$$

$$= \sum_w P_A(w) \log (P_A(w)) - \sum_w P_A(w) \log (P_T(w)), \quad (3)$$

where P_A is a language model estimated on an adaptation subset. Equation (2) is rewritten as shown in Equation (3) and the first term is independent of the sentence selection in the training set. By multiplying a total number of words in the adaptation set, the second term of Equation (3) becomes the same as the negation of Equation (1) if the differences such as N-gram discounting are ignored. Therefore, DLMS and the relative entropy method are basically based on the same objective function.

By using speech recognition output as adaptation data, these selective adaptation methods can be used for unsupervised adaptation. In this study, we apply DLMS to iterative unsupervised adaptation using a sentence as a selection unit. The greedy approach is adopted with a zero threshold criterion. That is, a sentence is selected if removing it from whole training data decreases the likelihood compared to using all the training data.

3. ITERATIVE UNSUPERVISED ADAPTATIONS

3.1. Unsupervised self adaptation

A widely used unsupervised adaptation framework is to first decode recognition data and use that output to estimate an adapted model. For the selective adaptation, the adapted model is made by using the recognition output to select a subset of training data. For higher performance, the adaptation process is iterated several times [7] as shown in Figure 1. The final recognition result is obtained by outputting the hypothesis made in the last decoding step. Since decoder output is used as adaptation data, it has an advantage that it does not require a separate in-domain development data. We refer to this conventional framework as unsupervised self adaptation.

A disadvantage of this procedure is that recognition errors in a decoding output are reinforced during the iteration, since the same data is used both in the decoding step and the model update step. This problem decreases the efficiency of the adaptation.

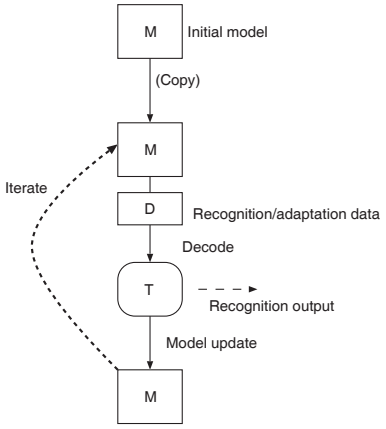

Fig. 1. Unsupervised self adaptation. M is model, T is recognition hypothesis, and D is recognition/adaptation data. Hypothesis T is made from data D using initial model M. Using that hypothesis, adapted model M is made. The adapted model is used to recognize the same data in the next iteration.

3.2. Unsupervised CV adaptation

Unsupervised CV adaptation has originally been proposed for acoustic model adaptation [6]. In this paper, we apply it to unsupervised language model adaptation. Unsupervised CV adaptation reduces the problem of the self adaptation by effectively separating the data used in the decoding step and in the model update step based on K-fold CV, as shown in Figure 2. In the procedure, recognition utterances are divided into K exclusive subsets $(D(1), D(2), \cdots, D(K))$ so that each subset has roughly the same size. The first decoding step is basically the same as the self adaptation and the K subsets are processed using the same initial model. Then, given the K recognition hypotheses $(T(1), T(2) \cdots, T(K))$, K CV models $(M(1), M(2), \cdots, M(K))$ are made by excluding one of the recognition hypotheses, instead of making a single model. Each model is used in the next decoding step to make a new hypothesis for a data subset that has been excluded from the estimation of that model. The decoding step and the model update step are repeated as in the conventional self adaptation.

The final recognition output is obtained by gathering hypotheses of the K subsets made in the last decoding step. Alternatively, a global CV model $(M(0))$ is made in the last update step together with the CV models using all recognition hypotheses, and the final output is obtained from that model.

With this procedure, the data used for the decoding step and for the model update step are effectively separated minimizing the undesired effect of reinforcing the errors. Because the utterances used for model estimation are not decoded by that model, it is unlikely that the same recognition error is repeated. The data fragmentation problem is minimal for large K, since $(K-1)/K$ of the data is used for the parameter estimation of each CV model.

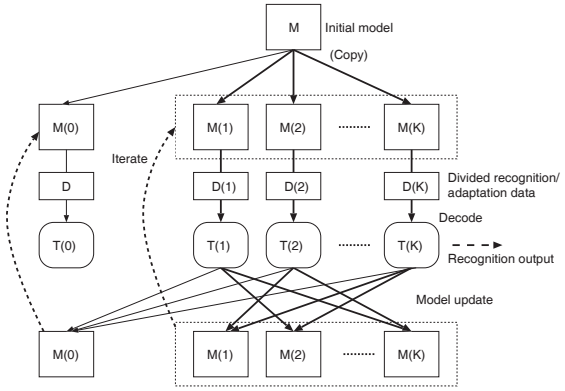

Fig. 2. Unsupervised CV adaptation. M is model and D is recognition/adaptation data. $M(k)$ is k-th CV model, $D(k)$ is k-th exclusive data subset, and $T(k)$ is recognition hypothesis of k-th subset using $M(k)$. The data used for the decoding step and for the model update step are separated. $M(0)$ denotes a global CV model and $T(0)$ denotes a hypothesis by $M(0)$.

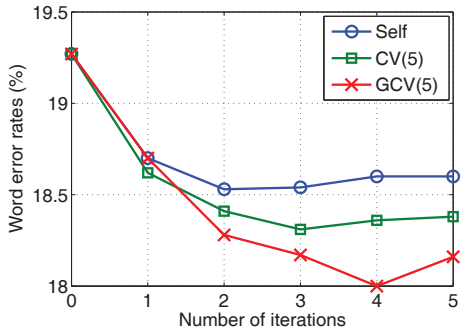

Fig. 3. Number of iterations and word error rates. Zero-th iteration indicates initial model. CV-folds K is five for CV adaptation. GCV indicates global CV model.

4. EXPERIMENTAL SETUP

The test set was the official evaluation set of the Corpus of Spontaneous Japanese (CSJ) [8] that consisted of 10 academic presentations given by different speakers. The length of each presentation was about 10 to 20 minutes and the total duration was 2.3 hours. The unsupervised self and CV adaptations were performed for each of these presentations. For CV adaptation, recognition utterances were randomized before the CV partitioning. The speech recognition system was based on the T^3 WFST decoder [9]. The training set for language models was the official CSJ training set consisted of 6.8M words of academic and extemporaneous presentations. The initial language model used in the recognition system was a trigram model estimated using all the training data. The dictionary size was 30k where the vocabulary was selected based on frequency. For the unsupervised adaptations, subsets of the training set were extracted using unigram probabilities. Given the selected subsets, trigram models were estimated and used for speech recognition. Because the training and test data were from the same domain, the evaluation was focused on adaptation performance for each presentation. The acoustic model was a tied-state Gaussian mixture triphone HMM estimated by MPE discriminative training using 254 hours of academic presentations. It had 3000 states and 32 mixtures per state. Feature vectors had 39 elements comprising 12 MFCCs and log energy, their delta, and delta delta values.

5. EXPERIMENTAL RESULTS

Tables 1 and 2 show properties of language models made by the self adaptation and five fold CV adaptation, respectively. About $1/4$ to $1/3$ sentences were selected by the selective adaptations. Words that do not appear in selected subsets were removed when adapted language models were made. However, increases of OOV rates were less than 0.2%. Test-set perplexities by the CV models were generally higher than the self adapted models because the CV models were estimated excluding a CV subset and were used to evaluate a reference transcript that corresponded to that CV subset. Even though they gave higher test-set perplexity, it is expected that they have less confusion with errors included in recognition hypotheses used for their estimation. By construction, the global CV model made in the first iteration is the same as the first model made by self adaptation. When more than two iterations were applied, the global CV models gave lower perplexity than self adaptation.

Figure 3 shows number of iterations and word error rates. The first iteration by self adaptation corresponds to the conventional two-pass method. The CV models gave lower word error rates than the models adapted by self adaptation. After the first iteration, the global CV models gave better performance than the CV models. Since the global CV models are estimated using all CV hypotheses in the last adaptation step, they are directly adapted to the target sentences while minimizing the risk of repeating recognition errors that occurred in the adaptation process. The word error rate by the initial model was 19.3%. The lowest word error rates by self adaptation based models, CV models, and global CV models, were 18.5%, 18.3% and 18.0%, respectively. The difference between the self adapted model and the global CV model was statistically significant by the MAPSSWE test.

Figure 4 shows the number of CV-folds K and word error rates. $K = 1$ indicates self adaptation. When $K = 2$, CV model gave worse result than self adapted model because in this case only a half of adaptation data is effectively used to estimate CV models. The global CV models generally gave better performance than the CV models.

6. CONCLUSION

Direct likelihood maximization selection (DLMS) has been applied to unsupervised language model adaptation for pre-

Table 1. Properties of language models estimated by conventional self adaptation. Zero-th iteration shows results of a presentation independent initial model estimated using all training data. Unigram and trigram test-set perplexities are evaluated using a reference transcript excluding OOVs. The values are averages for the test set presentations.

# of iterations	0	1	2	3	4	5
# selected sentences	384k	127k	103k	144k	111k	158k
Vocabulary	30.0k	23.1k	21.6k	23.6k	22.4k	24.1k
OOV (%)	1.90	2.00	2.03	2.01	2.02	1.99
1gram PP	485.0	359.8	345.7	361.2	348.9	370.4
3gram PP	71.6	59.5	57.6	59.4	58.2	60.7

Table 2. Properties of language models estimated based on proposed CV adaptation. Zero-th iteration shows results of a presentation independent initial model estimated using all training data. Test-set perplexities by the CV models are evaluated according to CV partitioning and are averaged in log domain.

# of iterations	0	1	2	3	4	5
CV model (5-fold)						
# selected sentences	384k	106k	108k	111k	122k	114k
Vocabulary	30.0k	23.3k	22.9k	23.3k	23.8k	23.5k
OOV (%)	1.90	2.07	2.09	2.09	2.08	2.06
1gram PP	485.0	375.4	376.7	368.7	370.5	365.8
3gram PP	71.6	63.5	64.0	62.5	62.1	62.1
global CV model (5-fold)						
# selected sentences	384k	127k	129k	91k	96k	101k
Vocabulary	30.0k	23.1k	22.9k	21.7k	23.4k	22.3k
OOV (%)	1.90	2.00	2.03	2.04	2.08	2.03
1gram PP	485.0	359.8	352.2	342.4	338.7	346.6
3gram PP	71.6	59.5	59.1	57.9	57.2	58.2

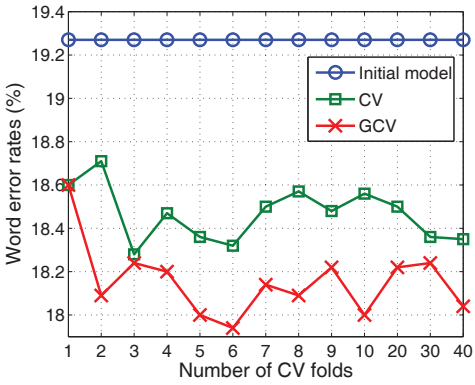

Fig. 4. Number of CV-folds K and word error rates at fourth iteration. CV-fold 1 indicates self adaptation.

sentation speech recognition. In order to reduce the disadvantage of the self adaptation strategy, unsupervised CV adaptation framework has been introduced that has originally been proposed for acoustic model adaptation. Experimental results show that adaptation performance by DLMS is further improved by incorporating the CV adaptation framework. Future work includes trying unsupervised aggregated adaptation [6] instead of CV adaptation.

7. ACKNOWLEDGMENT

This research has been partially supported by KAKENHI (21300066) and by KAKENHI (21300062).

8. REFERENCES

[1] R. Nisimura, K. Komatsu, Y. Kuroda, K. Nagatomo, A. Lee, H. Saruwatari, and K. Shikano, "Automatic n-gram language model creation from web resources," in *Proc. Eurospeech*, 2001, pp. 2127–2130.

[2] A. Sethy, P. Georgiou, and S. Narayanan, "Text data acquisition for domain-specific language models," in *Proc. EMNLP*, 2006, pp. 382–389.

[3] D. Klakow, "Selecting articles from the language model training corpus," in *Proc. ICASSP*, 2000, vol. 3, pp. 1695 –1698.

[4] T. Shinozaki, Y. Kubota, S. Furui, E. Utsunomiya, and Y. Shindoh, "Sentence selection by direct likelihood maximization for language model adaptation," in *Proc. Interspeech*, 2011, pp. 613–616.

[5] S. Maskey and A. Sethy, "Resampling auxiliary data for language model adaptation in machine translation for speech," in *Proc. ICASSP*, 2009, pp. 4817–4820.

[6] T. Shinozaki, Y. Kubota, and S. Furui, "Unsupervised acoustic model adaptation based on ensemble methods," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 6, pp. 1007 –1015, Dec. 2010.

[7] P. C. Woodland, D. Pye, and M. J. F. Gales, "Iterative unsupervised adaptation using maximum likelihood linear regression," in *Proc. ICSLP*, oct 1996, vol. 2, pp. 1133–1136.

[8] T. Kawahara, H. Nanjo, T. Shinozaki, and S. Furui, "Benchmark test for speech recognition using the Corpus of Spontaneous Japanese," in *Proc. SSPR2003*, 2003, pp. 135–138.

[9] P. R. Dixon, D. A. Caseiro, T. Oonishi, and S. Furui, "The titech large vocabulary wfst speech recognition system," in *Proc. IEEE ASRU*, 2007, pp. 443–448.