RELIABLE FEATURE SELECTION FOR LANGUAGE MODEL ADAPTATION

Chuang-Hua Chueh and Jen-Tzung Chien

Department of Computer Science and Information Engineering
National Cheng Kung University, Tainan, Taiwan, ROC
{chchueh, chien}@chien.csie.ncku.edu.tw

ABSTRACT

Language model adaptation aims to adapt a general model to a domain-specific model so that the adapted model can match the lexical information in test data. The minimum discrimination information (MDI) is a popular mechanism for language model adaptation through minimizing the Kullback-Leibler distance to the background model where the constraints found in adaptation data are satisfied. MDI adaptation with unigram constraints has been successfully applied for speech recognition owing to its computational efficiency. However, the unigram features only contain low-level information of adaptation articles which are too rough to attain precise adaptation performance. Accordingly, it is desirable to induce high-order features and explore delicate information for language model adaptation if the adaptation data is abundant. In this study, we focus on adaptively select the reliable features based on re-sampling and calculating the statistical confidence interval. We identify the reliable regions and build the inequality constraints for MDI adaptation. In this way, the reliable intervals can be used for adaptation so that interval estimation is achieved rather than point estimation. Also, the features can be selected automatically in the whole procedure. In the experiments, we carry out the proposed method for broadcast news transcription. We obtain significant improvement compared to conventional MDI adaptation with unigram features for different amount of adaptation data.

Index Terms- Feature Selection, Confidence Interval, Minimum Discrimination Information, Language Model, Speech Recognition

1. INTRODUCTION

Language modeling plays a critical role in automatic speech recognition and many other applications. Traditionally, language models are estimated from a large corpus containing different domains. Usually, the test environment for speech recognition is changed by the speakers with their speaking styles and conversation topics, etc, which causes the environmental mismatch problems in acoustic models as well as language models. In this study, we focus on the issue of compensating domain mismatch through the language model adaptation [1]. However, it is difficult to collect sufficient in-domain articles in advance. There is only a limited amount of adaptation data available for adapting a general domain model to a specific domain. In [6][9], some language model adaptation methods have been established. One popular approach was to do model interpolation, where the background model was interpolated with the model trained on adaptation data using a fixed interpolation weight. Also, the minimum discrimination information (MDI) has been successfully employed in language model adaptation [6][12] and popularly known to be superior to model interpolation. MDI method started from the maximum entropy (ME) based language model [1]. MDI is more powerful than model interpolation since an adaptive weight is assigned for individual feature. Using MDI, the adapted model is consistent with the statistics extracted from the adaptation data. The Kullback-Leibler divergence is minimized in MDI procedure. When adaptation data is small, only unigram features are reliable to represent the statistics of adaptation documents. In [6], the constraints of unigram features were involved. In [13], the topic unigrams extracted by Latent Dirichlet Allocation was used for MDI adaptation. When adaptation data is increased, the higher-order features, e.g. bigrams or trigrams, become reliable for MDI adaptation. In general, the lower-order features are reliable but rough for model adaptation. In contrast, the higher-order features are delicate but the sufficient adaptation data are required. Similar to acoustic model adaptation [4], there is a tradeoff between reliability and effectiveness with different size of adaptation data. Accordingly, how to properly balance and select suitable features for MDI adaptation is critical for speech recognition and understanding. Traditionally, the feature selection in ME language modeling was done following the likelihood gain [2]. This method selected good features in the same feature order. It was not suitable for the case of covering different orders of features because the high-order features always attain higher likelihood gain than the low-order features. The overtraining phenomenon happens and deteriorates the system performance. In this paper, we adopt the re-sampling technique and calculate the confidence interval statistically to determine the reliable regions of linguistic features. These regions represent the reliable intervals that features can be extracted from the available adaptation data. Based on the inequalities in ME approach [10], we merge the regions and relax the constraints for MDI adaptation. Namely, we perform *interval estimation* for each linguistic feature rather than point estimation. The advantage of using inequality constraints is that we can remove the unreliable features and induce the salient features in MDI estimation procedure. In the experiments, we evaluate the proposed method for large vocabulary continuous speech recognition using different number of adaptation data.

2. MDI ADAPTATION

Let $p_B(h,w)$ denote the background language model of history h and word w trained from a large corpus in general domain and $p_A(h,w)$ represent the target or the adapted model estimated from an adaptation corpus in new domain. Given features $f_1,\cdots,f_F$ from adaptation data, we determine the expectation of f_i with respect to the empirical distribution $\tilde{p}_A(h,w)$ by

$$E_{\tilde{p}_A}[f_i] = \sum_{h,w} \tilde{p}_A(h,w) f_i(h,w) , \qquad (1)$$

where $f_i(\cdot)$ is a binary-valued feature function. Also, using the conditional probabilities in language model, we can calculate the expectation with respect to the target conditional distribution $p_A(w|h)$ by

$$E_{p_A}[f_i] = \sum_{h,w} \widetilde{p}_A(h) p_A(w|h) f_i(h,w) . \quad (2)$$

Thus, we specify the constraints by

$$E_{p_A}[f_i] = E_{\widetilde{p}_A}[f_i], \quad i = 1, \cdots, F . \quad (3)$$

Under these constraints, we minimize the Kullback-Leibler divergence (KLD) between $p_A(w|h)$ and $p_B(w|h)$ [12]

$$\mathrm{KLD}(p_A(w|h), p_B(w|h)) = \sum_{h,w} \widetilde{p}_A(h) p_A(w|h) \log \frac{p_A(w|h)}{p_B(w|h)} . \quad (4)$$

MDI procedure is implemented by introducing Lagrange multipliers and solving the constrained optimization through minimizing

$$\mathrm{KLD}(p_A(w|h), p_B(w|h)) + \sum_{i=1}^{F} \lambda_i (E_{p_A}[f_i] - E_{\widetilde{p}_A}[f_i]) . \quad (5)$$

Accordingly, we obtain MDI model which is expressed as a log-linear or Gibbs distribution

$$p_A(w|h) = \frac{1}{Z(h)} p_B(w|h) \exp\left(\sum_{i=1}^{F} \lambda_i f_i(h,w) \right), \quad (6)$$

with the normalization term

$$Z(h) = \sum_{w} \left[p_B(w|h) \exp\left(\sum_{i=1}^{F} \lambda_i f_i(h,w) \right) \right] . \quad (7)$$

MDI parameters $\{\lambda_1, \cdots, \lambda_F\}$ can be estimated by the generalized iterative scaling (GIS) or the improved iterative scaling (IIS) algorithm [2][5].

In MDI approach, feature functions are incorporated to extract the helpful information from adaptation data. When the amount of adaptation samples is small, one may assume that only unigram features [6]

$$f_{w_i}(h,w) = \begin{cases} 1 & \text{if } w = w_i \\ 0 & \text{otherwise} \end{cases}, \quad (8)$$

can be reliably extracted. However, unigram features only represent simple lexical characteristics. It is desirable to exploit delicate features reflecting the properties of adaptation domain. For this consideration, we should develop a solution to reliable feature selection for language model adaptation in case of different amount of adaptation data. In [12], a cutoff threshold method was proposed to induce high-occurrence features. Here, we apply re-sampling technique and hypothesis test principle for adaptive feature selection. The reliable regions of features are adopted in building inequality constraints for MDI adaptation. We can select features automatically in a reliable estimation procedure.

3. RELIABLE SELECTION FOR MDI ADAPTATION

Assuming that there exists a true in-domain *n*-gram distribution for target domain, this distribution should sufficiently represent the characteristics of domain knowledge. Given a set of in-domain training data, we adapt a general language model to a domain-specific language model. When the amount of adaptation data is increased, the distribution of adaptation data tends to target in-domain distribution and the feature extraction goes stable. However, we don't know whether the observed data are located in a suitable region. Some instable features need to be pruned for reliable adaptation. In what follows, we address the solution to statistically analyze the reliability of features and adapt the language model to the target domain. We determine the reliable region for each feature by performing re-sampling and calculating the confidence intervals so as to attain interval estimation for each feature rather than point estimation. We define the inequality constraints using these regions for MDI adaptation.

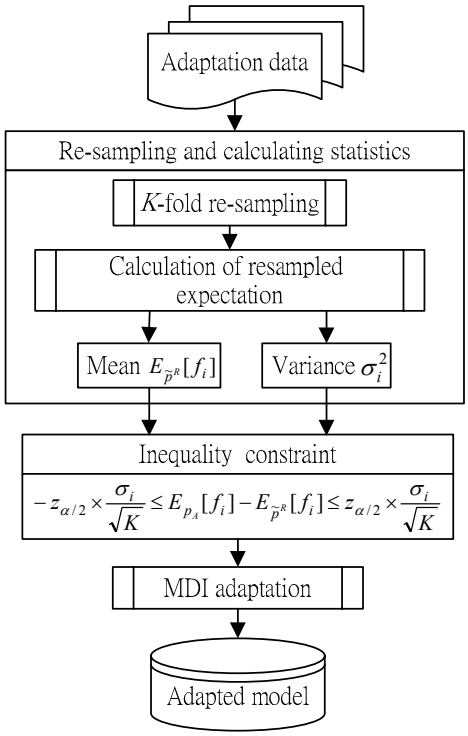

Figure 1 Procedure of reliable feature selection for MDI adaptation

3.1. Determination of reliable region

Re-sampling technique has been successfully used to estimate robust mutual information for feature selection [7] as well as to construct random forest language models [14]. Here, we adopt the re-sampling technique to extract information for analyzing the reliability of features. Then, we calculate the confidence interval to identify the reliable region individually for each feature from adaptation data. The features in these regions are statistically stable. The reliable features are applied for language model adaptation. The whole procedure is shown in Figure 1. We use *K*-fold re-sampling method which divides the adaptation data into *K* subsets and build the distributions, $p_k^R(h,w) \quad k = 1 \sim K$ by using the data excluding subset k. Then, the empirical expectations are calculated *K* times based on different distributions. We compute the sample mean $E_{\widetilde{p}^R}[f_i]$ and the sample variance σ_i^2. According to the Central Limit Theorem, when the number of subsets is sufficient, the re-sampled expectation tends to be a Gaussian distribution with mean $E_{\widetilde{p}^R}[f_i]$ and variance σ_i^2. Under a significance level α, we have $1-\alpha$ confidence to believe that the reliable expectation $E_{\hat{p}_A}[f_i]$ falls into the confidence interval

$$E_{\widetilde{p}^R}[f_i] - z_{\alpha/2} \times \frac{\sigma_i}{\sqrt{K}} \le E_{\hat{p}_A}[f_i] \le E_{\widetilde{p}^R}[f_i] + z_{\alpha/2} \times \frac{\sigma_i}{\sqrt{K}} , \quad (9)$$

where $z_{\alpha/2}$ is the statistic value determined by referring a standard Gaussian distribution. Thus, confidence interval makes the probability of extracting reliable expectation based on sampled data satisfy the following condition

$$p(E_{\tilde{p}_A^R}[f_i]-z_{\alpha/2}\times\frac{\sigma_i}{\sqrt{K}}\le E_{\hat{p}_A}[f_i]\le E_{\tilde{p}^R}[f_i]+z_{\alpha/2}\times\frac{\sigma_i}{\sqrt{K}})\ge 1-\alpha\,. \tag{10}$$

The empirical expectation given adaptation data should satisfy the inequality in (9). In MDI, we strive to characterize the expectations as reliable as the empirical expectations. Namely, we calculate the model expectation by considering its reliable region or confidence interval. Therefore, we determine the lower bound A_i and the upper bound B_i for the inequality constraint

$$A_i=-z_{\alpha/2}\times\frac{\sigma_i}{\sqrt{K}}\le E_{p_A}[f_i]-E_{\tilde{p}^R}[f_i]\le z_{\alpha/2}\times\frac{\sigma_i}{\sqrt{K}}=B_i\,. \tag{11}$$

In this procedure, we use the re-sampled empirical distribution $\tilde{p}^R(h,w)=\sum_{k=1}^{K}\tilde{p}_k^R(h,w)\Big/K$ instead of the original empirical distribution $\tilde{p}(h,w)$ for calculation of expectation function.

3.2. Inequality constraints for MDI

With the reliable region in (11), we establish the inequality constraints in MDI criterion which is consisted of a KLD term and two sets of constraints corresponding to the lower bound and the upper bound

$$\begin{aligned}\mathrm{KLD}(p_A(w|h),p_B(w|h))&+\sum_{i=1}^{F}a_i\left(E_{\tilde{p}^R}[f_i]-E_{p_A}[f_i]+A_i\right)\\&+\sum_{i=1}^{F}b_i\left(E_{p_A}[f_i]-E_{\tilde{p}^R}[f_i]-B_i\right).\end{aligned} \tag{12}$$

MDI solution to the adapted language model is then given by

$$p_A(w|h)=\frac{1}{Z(h)}p_B(w|h)\exp\left(\sum_{i=1}^{F}(a_i-b_i)f_i(h,w)\right), \tag{13}$$

with parameter set $\lambda=\{a_i,b_i,i=1\sim F\}$. Based on the Karush-Kuhn-Tucker (KKT) condition [11] the optimal parameters must satisfy the constraints of $a_i\ge 0, b_i\ge 0$ and

$$\begin{aligned}a_i\left(E_{\tilde{p}^R}[f_i]-E_{p_A}[f_i]+A_i\right)&=0,\\ b_i\left(E_{p_A}[f_i]-E_{\tilde{p}^R}[f_i]-B_i\right)&=0.\end{aligned} \tag{14}$$

GIS or IIS algorithm is applied to estimate MDI parameters. Because the lower and upper bound constraints can not be satisfied at the same time, one or both of the parameters a_i and b_i must be zero. Namely, there are three conditions for parameters a_i and b_i:

1. $a_i>0, b_i=0$ (lower bound is active),
2. $a_i=0, b_i>0$ (upper bound is active),
3. $a_i=0, b_i=0$ (inactive).

In the first and the second cases, the feature f_i is active. Only active features have influence on the adapted model. The inactive features in the third case are explicitly removed.

Also, it is interesting to explore the relation between the inequalities of MDI and maximum likelihood (ML) methods. By substituting (13) into (12), the minimization of MDI criterion is equivalent to the maximization of ML criterion

$$\sum_{h,w}\tilde{p}^R(h,w)\log p_A(w|h)-\sum_{i=1}^{F}\left[(a_i+b_i)\cdot z_{\alpha/2}\cdot\frac{\sigma_i}{\sqrt{K}}\right], \tag{15}$$

which is composed of the log-likelihood of re-sampled adaptation data and the penalty term. The penalty is affected by the standard variance σ_i, the number of subsets K and the significant level α. When σ_i is large, the expectation has a great variation and the corresponding feature f_i is not stable so that the penalty is increased. Also, in case of a larger number K, the re-sampled estimation is robust so that the penalty can be decreased. Also, a larger K makes the Gaussian assumption of Central Limit Theorem more reasonable so that the confidence interval can be estimated more robustly. Therefore, it is meaningful that the inequalities in MDI can balance the tradeoff between the likelihood function and the model penalty for model adaptation.

4. EXPERIMENTS

In the experiments, the proposed algorithm was evaluated by the broadcast news database MATBN. We estimated the speaker-independent HMM models using the benchmark Mandarin speech corpus TCC300, which was recorded using closed-talking microphone in office environments. Each Mandarin syllable was modeled by initial and final models using right context-dependent states with 32 mixture components at most. The initial and final models were specified by three and five states, respectively. Speech feature vector consisted of twelve Mel-frequency cepstral coefficients, one log-energy and their first derivation. The baseline HMM was adapted using MAP adaptation [8] with 500 sampled conversations in MATBN. The baseline trigram model was trained using the Academic Sinica CKIP balanced corpus and the CIRB corpus with web news stories [3]. There were totally forty-five million words. Our lexicon contained 32,909 Chinese words with one to four characters. To conduct a consistent evaluation, Jelinek-Mercer smoothing method was used to build baseline trigram model by interpolating with unigram and bigram. To evaluate language model adaptation, we randomly sampled different sizes of adaptation data from MATBN corpus with 500, 1000, 2000, 3000 and 4000 conversations. The significant level α was fixed to be 0.05 and the number of subsets K was 50, which was sufficient to apply Gaussian assumption in this method. For parameter estimation, 20 iterations were performed in IIS algorithm. The language model scaling factor was set as 10 by optimizing the recognition accuracy of baseline model. MDI models with different order features were carried out for comparison. The test data of MATBN contained 22 minutes with totally 7473 characters.

First, we analyze the relation between number of adapted utterances and the width of reliable regions. We randomly selected a trigram feature and calculated the width of reliable region using (11). The results in Table 1 show that the more data is available, the smaller the variation of a feature is. Therefore, when adapted data is fewer, the features are easier to be removed because a larger width of reliability implies a larger penalty.

Table 1 Relation between the number of adapted utterances and the region width of a trigram feature

	Number of adapted utterances				
	500	1000	2000	3000	4000
Width ($\times 10^{-6}$)	1.72	0.81	0.38	0.24	0.18

Next, we evaluate the proposed method using perplexity measure. Perplexity can be interpreted as the average number of branches in the text. The higher perplexity, the more branches the speech recognition system should consider. In the results of Figure 2, when using 500 adapted utterances, the data is too sparse to estimate robust trigram statistics so that the adapted model by using trigram features greatly degrade the system performance. By increasing the amount of data, the perplexities of all methods are improved. Also, the proposed method attained lower perplexities than MDI models with unigram features as well as trigram features. If the data is sufficient to train trigrams for target domain, i.e. the case that the number of adapted utterances larger than 3000, the proposed method and the conventional MDI model with trigram features achieve comparable performance. Namely, reliable features can be adaptively induced for adaptation based on available data.

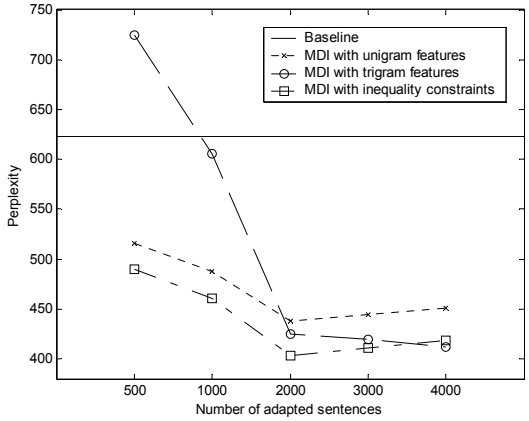

Figure 2 Perplexities using different adapted models

Also, we applied the proposed model in speech recognition system. Figure 3 shows the character error rates using different adapted models. The inconsistence in perplexity performance is caused by the other factors, e.g. the effects of acoustic likelihood and competing hypotheses. Clearly, using inequality constraints for MDI adaptation attains better accuracy than using unigram and trigram features. In conclusion, the proposed method can reliably select the suitable features in different cases using various number of adaptation data for language model adaptation.

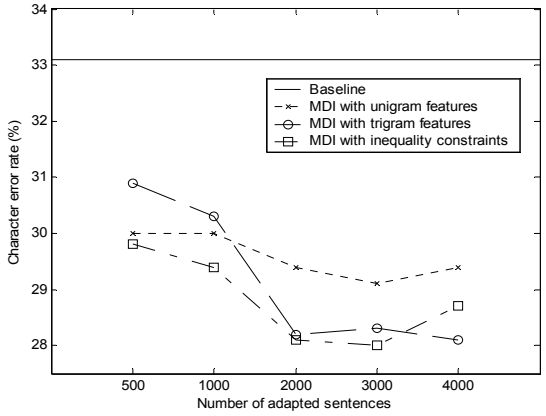

Figure 3 Character error rates using different adapted models

5. CONCLUSIONS

We have presented a new feature selection method for language model adaptation. Given the available adaptation data, we adopted re-sampling method to extract information for evaluating the reliability of features. Then, we calculated the confidence interval statistically and obtained the reliable region for each feature to calculate the confidential empirical expectation. A larger reliable region reflected a higher variation of feature statistics in the data. The corresponding features were more unstable and should be removed. We defined inequality constraints by merging the reliable regions for MDI adaptation. We attained the interval estimate rather than the point estimate. Meaningfully, we obtained the MDI parameters by considering the likelihood function as well as the penalty terms. The penalty was proportional to the width of the reliable region. Therefore, the feature selection was done by balancing the likelihood and penalty. In the experiments, we carried the proposed method using Mandarin broadcast news data. Different sizes of adaptation data were analyzed. The proposed feature selection for MDI adaptation attained significant reduction of perplexity. The recognition performance was desirable as well.

6. REFERENCES

[1] J. R. Bellegarda, "Statistical language model adaptation: review and perspectives", *Speech Communication*, vol. 42, pp. 93-108, 2004.

[2] A. Berger, S. Della Pietra and V. Della Pietra, "A maximum entropy approach to natural language processing", *Computational Linguistics*, vol. 22, no. 1, pp. 39-71, 1996.

[3] K. Chen and H.-H. Chen, "Cross-language Chinese text retrieval in NTCIR workshop – towards cross-language multilingual text retrieval", *ACM SIGIR Forum*, vol. 35, no. 2, pp. 12–19, 2001.

[4] J.-T. Chien, C.-H. Lee and H.-C. Wang, "A hybrid algorithm for speaker adaptation using MAP transformation and adaptation", *IEEE Signal Processing Letters*, vol. 4, no. 6, pp. 167-169, 1997.

[5] J. Darroch and D. Ratcliff, "Generalized iterative scaling for log-linear models", *The Annals of Mathematical Statistics*, vol. 43, pp. 1470-1480, 1972.

[6] M. Federico, "Efficient language model adaptation through MDI estimation", in *Proc. EUROSPEECH*, pp. 1583-1586, 1999.

[7] D. Francois, F. Rossi, V. Wertz and M. Verleysen, "Resampling methods for parameter-free and robust feature selection with mutual information", *Neurocomputing*, vol. 70, 1276-1288, 2007.

[8] J.-L Gauvain and C.-H. Lee, "Maximum *a posteriori* estimation for multivariate Gaussian mixture observation of Markov chain", *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 4, pp. 291-298, 1994.

[9] D. Janiszek, F. Bechet, R. de Mori, "Integrating MAP and linear transformation for language model adaptation", in *Proc. ICSLP*, vol. 2, pp 895-898, 2000.

[10] J. Kazama and J. Tsujii, "Maximum entropy models with inequality constraints: A case of study on text categorization", *Machine Learning*, vol. 60, pp. 159-194, 2005.

[11] H. W. Kuhn and A. W. Tucker, "Nonlinear programming", in *Proc. of the 2th Berkeley Symposium on Mathematical Statistics and Probabilities*, pp. 481-492, 1951.

[12] P. S. Rao, S. Dharanipragada and S. Roukos, "MDI adaptation of language models across corpora", in *Proc. ICASSP*, vol. 1, pp. 161-164, 1995.

[13] Y. C. Tam and T. Schultz, "Unsupervised language model adaptation using latent semantic marginals", in *Proc. ICSLP*, pp. 2206-2209, 2006.

[14] P. Xu and F. Jelineck, "Random forests and the data sparseness problem in language modeling", *Computer Speech and Language*, vol. 21, pp. 105-152, 2007.