LANGUAGE MODEL ADAPTATION VIA MINIMUM DISCRIMINATION INFORMATION

P. Srinivasa Rao, Michael D. Monkowski, *and Salim Roukos*

IBM Thomas J. Watson Research Center
P.O. Box 704, Yorktown Heights, NY 10598

ABSTRACT

Statistical language models improve the performance of speech recognition systems by providing estimates of a priori probabilities of word sequences. The commonly used trigram language models obtain the conditional probability estimate of a word given the previous two words, from a large corpus of text. The text corpus is often a collection of several small diverse segments such as newspaper articles, or conversations on different topics. Knowledge of the current topic could be utilized to adapt the general trigram language models to match that topic closely. For example, an interpolation of the general language model with one built on the topic data could be used. We first discuss the adaptation of general trigram language models to a known topic using the minimum discrimination information (MDI) method. We then present results on the Switchboard corpus which consists of telephone conversations on several topics.

1. ADAPTATION USING MDI

The goal of language model adaptation is to use a large amount of mixed or general text data from a portion of a corpus and a smaller amount of data from a homogeneous sub-corpus for building a language model that performs well on the sub-corpus. Similarly adaptation can be done across different corpora, i.e., from a large corpus to a smaller one. One of the ways adaptation can be achieved is by building separate language models on the two available portions of training data and interpolating between them [5]. Another approach is to find a model that fits important features of the topic or homogeneous data and that is closest (in the discrimination information distance) to an initial model estimated from the large general data. We discuss here the adaptation of general trigram language models to a known topic using this minimum discrimination information (MDI) approach [8, 6].

Denote by Q_0 the initial distribution over the finite sample space $X = \{x_1, x_2, \cdots, x_n\}$ estimated using a large amount of data. In order to adapt this distribution to the known topic, we require that the final distribution match important features of the topic data. These features are selected in the form of a set of m real-valued constraint functions

$$\{f_i : X \rightarrow \Re,\ i = 1, 2, \cdots, m,\ m < n\},$$

and a set of constants $\{a_i,\ i = 1, \cdots, m\}$. The class $\mathcal{C}$ of topic matching distributions contains all distributions $P = \{p_1, \cdots, p_n\}$ that satisfy the linear constraints

$$E_P[f_i(x)] \triangleq \sum_{j=1}^{n} f_i(x_j)\, p_j = a_i. \qquad (1)$$

The topic adapted model is obtained by selecting a distribution $P^* \in \mathcal{C}$ that minimizes the discrimination information (Kullback-Liebler distance)

$$D_X(P, Q_0) \triangleq \sum_{j=1}^{n} p_j \log \frac{p_j}{q_{0j}}$$

with respect to the initial distribution. If the initial distribution Q_0 is uniform,

$$D_X(P, Q_0) = \sum_{j=1}^{n} p_j \log \frac{p_j}{1/n} = -H(P) + \log n,$$

where $H(P)$ is the entropy of the distribution P. Hence the MDI solution in this case is the one with maximum entropy in the constraint class $\mathcal{C}$.

Using Lagrange multipliers, the solution of this constrained optimization problem can be shown to be of the form

$$p_j^* = \frac{q_{0j}}{Z(\lambda_1, \cdots, \lambda_m)}\, e^{-\lambda_1 f_1(x_j) - \cdots - \lambda_m f_m(x_j)},$$

leading to an exponential family of distributions with the constraint functions as sufficient statistics. The normalizing constant Z and the natural parameters λ_i

0-7803-2431-5/95 $4.00

of this family can be found using the generalized iterative scaling algorithm [3, 2]. The constants a_i are usually chosen to be the sample means of the constraint functions in the topic data, i.e. $a_i = E_{\tilde{P}}[f_i(x)]$, where $\tilde{P}$ is the empirical distribution of the topic data. In such a case the MDI solution is also the maximum likelihood estimate of a member of the exponential family given the data.

When dealing with trigram language models, we are interested in conditional distributions of the form $\{p(w_3|w_1, w_2)\}$ instead of the marginal distributions. The above discussion can easily be applied to joint distributions $\{p(w_1, w_2, w_3)\}$ and the conditional distributions then derived from these. However due to the large number of parameters in $X = V \times V \times V$, where V is the vocabulary, the few constraints obtained from the small topic data may not provide an effective adaptation [6]. Instead of first adapting joint distributions and then obtaining the required conditional distributions, the above formulation can be rephrased directly in terms of conditional distributions [7]. The constraint equations (1) can be written as

$$\sum_{w_1,w_2} p(w_1, w_2) \cdot \sum_{w_3} p(w_3|w_1, w_2)\, f_i(w_1, w_2, w_3) \;=\; a_i$$

and approximated as

$$\sum_{w_1,w_2} \tilde{p}(w_1, w_2) \cdot \sum_{w_3} p(w_3|w_1, w_2)\, f_i(w_1, w_2, w_3) \;=\; a_i$$

where $\{\tilde{p}(w_1, w_2)\}$ is the empirical joint distribution on the history. Given an initial trigram model $\{q_0(w_3|w_1, w_2)\}$, a unique MDI model $\{p^*(w_3|w_1, w_2)\}$ exists and can be found using the generalized iterative scaling procedure.

2. EXPERIMENTS ON SWITCHBOARD

Switchboard corpus contains acoustic data and transcriptions of about 2300 telephone conversations on nearly 70 topics [4]. The transcriptions have a total of about 3 million words. Because of factors such as channel and background noise and the conversational style of speech, recognition accuracy on the Switchboard corpus is quite poor compared to other corpora such as the Wall Street Journal task.

In this section we describe some experiments on the use of the MDI technique for adapting a general language model on Switchboard to the particular topic being recognized. We selected three topics for adaptation; *credit card use, care of the elderly, and buying a car.*

Features that are characteristic of the topic to be adapted are used to define the constraint functions. In the unigram case this implies the use of most frequent (also most infrequent) words in the topic data. We used words that contribute most to the Kullback-Liebler distance between the unigram distributions of the topic data and the rest of Switchboard data to define the unigram constraints. The set of unigram feature words is

$$S_1 \;=\; \{w^{(i)} :\; q_t(w^{(i)}) \log \frac{q_t(w^{(i)})}{q_{0x}(w^{(i)})} \;>\; \alpha\},$$

where $q_t(w^{(i)})$ and $q_{0x}(w^{(i)})$ are relative frequencies of $w^{(i)}$ in the topic and the training data of LM0x respectively, and α is a threshold value. For example, in the case of credit card topic the set S_1 contains about 250 words such as *credit, charge, I,* and *if* etc. The unigram constraints are given by

$$f_i(w_1, w_2, w_3) \;=\; \begin{cases} 1 & \text{if } w_3 = w^{(i)} \\ 0 & \text{otherwise}, \end{cases}$$

and the corresponding constant a_i is simply $q_t(w^{(i)})$.

Bigram constraint functions are similarly defined to be indicator functions of bigrams occuring frequently in the topic data, i.e. those that belong to the set

$$S_{23} \;=\; \{w^{(j)}w^{(k)} :\; q_t(w^{(j)}, w^{(k)}) \;>\; \beta\}$$

Typically there are a few hundred unigram features and a few thousand bigram features.

As the initial distribution, we use a trigram language model made using the entire Switchboard corpus training data (denoted below as LM0). For each of the three topics selected, we derive two MDI models, one using unigram constraints (LM1) and the other using both unigram and bigram constraints (LM2). In order to compare the performance of these models with one that has not been trained on any topic data, we also made another general trigram model (LM0x) with all of Switchboard training data excluding the portion from the three topics. All language models have a vocabulary of about 17000 words which provides a good coverage of the Switchboard corpus. From the set S_1 of unigram feature words, we obtain a list of *topic words* by eliminating those that are unrelated to the topic (e.g. *I, if,* etc. in the case of credit card). Recognition accuracies for all the language models are presented on the entire vocabulary as well as the subset of topic words. Tabulated below are the number of words in the training data, test data, and the topic words set for each of the topics.

Topic	Training size	Test size	Topic words set size
Credit Card	35000	2700	130
Care of Elderly	51000	2000	190
Buying a Car	61000	2000	470

Following are the perplexities and recognition accuracies for the various language models. Percentage error on the topic words includes insertions in the decoded text of topic words that do not appear in the original script, in addition to deletions and substitutions of those that do. Acoustic training was done on about 5300 sentences from the three topics using 9 dimensional Mel frequency cepstral coefficients, delta and delta-delta parameters and context-dependent continuous parameter hidden Markov models. These models use both the left and right contexts in order to better model the phonetic variation with context. The recognition system uses a rank based stack algorithm [1].

Topic: Credit Card

LM	Perplexity	% error overall	% error topic words
LM0	104	54.0	46.3
LM0x	116	55.0	53.7
LM1_cc	89	51.8	41.8
LM2_cc	89	51.7	43.5

Topic: Care of Elderly

LM	Perplexity	% error overall	% error topic words
LM0	107	55.7	58.8
LM0x	118	57.1	67.0
LM1_eld	95	53.8	53.6
LM2_eld	96	55.0	59.3

Topic: Buying a Car

LM	Perplexity	% error overall	% error topic words
LM0	124	65.7	80.7
LM0x	140	66.4	83.0
LM1_car	110	63.4	81.7
LM2_car	110	63.6	80.3

As can be seen from the tables, the topic adapted models provide about 10 - 15 % reduction in overall perplexity and 3 - 5 % improvement in recognition accuracy. One reason for this performance could be the fact that topic words account for less than 10 % of the test data. As we go from model LM1 to LM2, the large number of additional bigram constraints worsens the performance in most cases. This may be due to over-training of the models.

For the credit card topic, we also used models adapted using interpolation of the general model LM0 with a trigram model LMCC made on the credit card training data. The conditional probabilities for these models are given by

$$p(w_3 \mid w_1, w_2) = (1-\gamma)\, p_{\mathrm{LM0}}(w_3 \mid w_1, w_2) + \gamma\, p_{\mathrm{LMCC}}(w_3 \mid w_1, w_2),$$

where γ is the interpolation constant. The recognition performance of the interpolated model is given below for three values of γ.

Credit Card Interpolated model

γ	% error overall
0.1	52.7
0.25	51.9
0.4	51.0

The performance improvement obtained for the interpolated model in this case is quite similar to that with the MDI model.

3. CONCLUSION

Adaptation using the MDI approach provides a reasonable improvement in performance on the Switchboard corpus. Various methods of defining constraints needs to be explored further.

REFERENCES

[1] L. R. Bahl, P. V. de Souza, P.S. Gopalakrishnan, D. Nahamoo, and M. Picheny. Robust methods for using context-dependent features and models in a continuous speech recognizer. In *Proc. ICASSP-94*, 1994.

[2] I. Csiszar and G. Tusnady. Information geometry and alternating minimization procedures. *Statistics & Decisions*, Supplement Isuue(1):205–237, 1984.

[3] J. N. Darroch and D. Ratcliff. Generalized iterative scaling for log-linear models. *Ann. Math. Stat.*, 43(5):1470–1480, 1972.

[4] J. Godfrey, E. Holliman, and J. McDaniel. SWITCHBOARD: Telephone speech corpus for research development. In *Proc. ICASSP-92*, pages I–517–520, 1992.

[5] F. Jelinek, B. Merialdo, S. Roukos, and M. Strauss. A dynamic language model for speech recognition. In *Proc. Speech and Natural Language DARPA Workshop*, pages 293–295, February 1991.

[6] R. Lau, R. Rosenfeld, and S. Roukos. Trigger-based language models: a maximum entropy approach. In *Proc. ICASSP-93*, pages II–45–48, Minnesota, 1993.

[7] S. Della Pietra and V. Della Pietra. Statistical modelling by maximum entropy. IBM Technical Report in preparation.

[8] S. Della Pietra, V. Della Pietra, R. L. Mercer, and S. Roukos. Adaptive language modeling using minimum discrimination estimation. In *Proc. ICASSP-92*, pages I–633–636, San Fransisco, March 1992.