MDI ADAPTATION OF LANGUAGE MODELS ACROSS CORPORA

P. S. Rao *S. Dharanipragada* *S. Roukos*

IBM Thomas J. Watson Research Center
P. O. Box 218, Yorktown Heights, NY 10598

ABSTRACT

The amount of text data available from a corpus for training language models is usually limited. Data from larger general or related corpora can be utilized to improve the performance of the language model on the corpus of interest. We explore one method of adapting a prior model from a large corpus to a smaller one of interest. Perplexity results of adapting a prior model constructed using the NAB corpus to the Switchboard and ATIS corpora are presented and compared with those of interpolated models.

1. INTRODUCTION

Language models for speech recognition are usually estimated using text from the corpus to be recognized. However the amount of such data available is often limited. Well known estimation techniques such as deleted interpolation and back-off modeling address the problem of estimating language model parameters from sparse data [1, 2]. Further improvement of language models could be obtained by utilizing additional data from other similar corpora for which larger amounts of data are available. Adaptation of a language model built on a large corpus to the corpus of interest could for example be achieved by interpolating the former model with one built on the smaller corpus of interest [3]. We discuss here adaptation of trigram language models using the minimum discrimination information (MDI) method [4]. We describe adaptation of models from a large corpus, the North American Business (NAB) news corpus, to two small corpora: the Switchboard corpus which consists of transcriptions of telephone conversations, and the ATIS corpus consisting of airline travel queries. Previous work on improving the recognition accuracy on Switchboard by MDI adaptation of a general Switchboard language model to a specific topic has been reported in [5].

2. BACKGROUND

Suppose $Q(x,y)$ is an initial probability distribution, with x and y belonging to finite spaces X and Y respectively. We are interested in constructing a distribution $P(x,y)$ that models the given data while being close to the distribution Q. The distribution of interest P is required to match important features of the data. This requirement is enforced by a set of linear constraint equations involving the ensemble averages of certain *feature functions* $\{f_i, i = 1, \cdots, m\}$:

$$E_P\left[f_i(x,y)\right] \triangleq \sum_{X,Y} f_i(x,y)\, P(x,y) \;=\; a_i. \quad (1)$$

The feature functions $\{f_i\}$ and the constants $\{a_i\}$ are chosen using the data. Suppose for example that a specific sample (x_0, y_0) occurs frequently in the data and we wish $P(x_0,y_0)$ to match it's relative frequency $\tilde{P}(x_0,y_0)$. The feature function is chosen as the indicator function of this sample, i.e.,

$$f_i(x,y) \;=\; \begin{cases} 1 & \text{if } x = x_0, y = y_0 \\ 0 & \text{otherwise}, \end{cases}$$

and equation (1) reduces to $P(x_0,y_0) = \tilde{P}(x_0,y_0)$.

In order to select P that is closest to Q from the set of all distributions that satisfy equation (1), the minimum discrimination information (MDI) method uses as a distance measure the discrimination information (also known as the Kullback-Leibler distance) $D(P,Q)$ defined by

$$D(P,Q) \triangleq \sum_{X,Y} P(x,y) \log \frac{P(x,y)}{Q(x,y)}$$

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

$$P^*(x,y) \;=\; \frac{Q(x,y)\, e^{\lambda_1 f_1(x,y) + \cdots + \lambda_m f_m(x,y)}}{Z(\lambda_1, \cdots, \lambda_m)}, \quad (2)$$

a member of an exponential family of distributions with the constraint functions as sufficient statistics. The normalization constant Z and the parameters λ_i can be found using the *generalized iterative scaling algorithm* [6, 7]. If the initial distribution Q is uniform,

$$\begin{aligned} D(P,Q) &= \sum_{X,Y} P(x,y) \log \frac{P(x,y)}{1/n} \\ &= -H(P) + \log n, \end{aligned}$$

where n is the number of elements in $X \times Y$ and $H(P)$ is the entropy of the distribution P. Hence the MDI solution in this case is the one with maximum entropy that satisfies all the constraints.

3. MDI ADAPTATION OF LANGUAGE MODELS

In language modeling, we are interested in the conditional distribution $P(w|h)$ of a word w given history h. For example, in the case of trigram models, h is of the form (w_1, w_2), and the sample spaces of the previous section become $X = V \times V$, $Y = V$, where V is the vocabulary. The problem of estimating the joint distribution $P(h, w) = P(h)\,P(w|h)$ can be reduced to that of estimating $P(w|h)$ by imposing additional constraints of the form $P(h) = \tilde{P}(h)$, $\tilde{P}(h)$ being the empirical distribution of histories in the data. The constraint equations (1) then become

$$\sum_h \tilde{P}(h) \cdot \sum_w P(w|h)\, f_i(h, w) \;=\; a_i. \qquad (3)$$

The constants a_i are usually chosen to be the empirical averages of the feature functions f_i.

Given the marginal distribution $\tilde{P}(h)$ and language models P and Q, we now use

$$D(P, Q|\tilde{P}) \;\triangleq\; \sum_h \tilde{P(h)} \sum_w P(w|h) \log \frac{P(w|h)}{Q(w|h)}$$

as a distance measure. The adaptation problem can now be posed as that of finding a distribution P^* that minimizes $D(P, Q|\tilde{P})$ and satisfies the constraints (3). As before, P^* is of the form

$$P^*(w|h) \;=\; \frac{Q(w|h)\, e^{\lambda_1 f_1(h,w) + \cdots + \lambda_m f_m(h,w)}}{Z(h, \lambda_1, \cdots, \lambda_m)}. \qquad (4)$$

The parameters λ_i and the normalization constants Z are again obtained using the generalized iterative scaling algorithm.

4. EXPERIMENTS

Our goal is to adapt a general English language model built from a large corpus to smaller corpora using the MDI method described above. This model referred to as the prior model was trained using nearly 260 million words of text from the North American Business (NAB) news corpus which contains primarily financial newspaper articles. Since we are interested in benefiting from the most frequent n-grams which are likely to appear in other corpora, infrequent n-grams were discarded before building the model. Count thresholds of 2,4,6 were used for unigrams, bigrams and trigrams respectively, i.e. trigrams with count less than 6, bigrams with count less than 4 and singleton unigrams were dropped. This prior model was then adapted to Switchboard and ATIS, using the training text from these two corpora.

4.1. Adaptation to Switchboard Corpus

The Switchboard corpus consists of transcriptions of telephone conversations on various topics and has approximately 2.1 million words available for training. A baseline deleted-interpolation (DI) type model was built using the entire Switchboard training data. This model has a perplexity of 100 on a subset of nearly 5350 words of the development test corpus, while the prior NAB model has a perplexity of 268 on the same test set.

Using different feature sets derived from the Switchboard corpus we constructed three MDI adapted models. Model 1 uses unigram, bigram and trigram features with count thresholds of 5, 3, and 2 respectively, i.e., all unigrams with counts 5 or more, bigrams with counts 3 or more and trigrams with a count of 2 or more, are selected as features (we refer to this as the 5-3-2 feature MDI model). This adapted model gives a reasonable 6% reduction in perplexity compared to the baseline model. Model 2 uses the same 5-3-2 features but the bigram and trigram counts below a threshold value are discounted using the Good-Turing method as is often done in n-gram language model estimation [2]. Here we used a discounting threshold of 6. The discounted adapted model 2 improves slightly over the model 1 in terms of perplexity. Discounting not only helps in smoothing the language model but also permits us to include singleton trigram features. Inclusion of all trigrams without discounting would otherwise lead to the trivial maximum likelihood solution. Model 3 includes singleton trigrams with 4-2-1 thresholds and discounting. As seen from the table, inclusion of singleton trigrams increases the trigram features by more than a factor of 4, and this perhaps leads to an overtrained model because the perplexity increased in this case.

An alternate approach to the MDI adaptation is to interpolate the NAB prior model with the smaller Switchboard baseline model. Several values of the static interpolating constant were tried and Model 4 uses the value (0.2) that resulted in the lowest test set perplexity. This model gives a 10% reduction in perplexity over the baseline model, better than the three MDI adapted models.

Model	Description	Perplexity
Baseline	DI model	100
Prior	NAB model	268
Model 1	5-3-2	94
Model 2	5-3-2 discounted	93
Model 3	4-2-1 discounted	96
Model 4	0.2 NAB prior + 0.8 baseline	90

Count Thresholds	Number of Features
5-3-2	8200 unigram 67300 bigram 191700 trigram
4-2-1	9400 unigram 108000 bigram 885000 trigram

4.2. Adaptation to ATIS Corpus

The second corpus we experimented with is the Airline Travel Information System (ATIS) corpus consisting of airline travel queries with about 175,000 words of training text. It is a very constrained task, as seen from the low perplexity (25) of the baseline DI model. Also the style is not very natural and the prior general English model has a very high perplexity of 427 on the ATIS task.

We constructed three MDI models with different number of features. Model 1 with 4-2-1 features provides a small reduction in perplexity compared to the baseline model. Models 2 and 3 with 4-3-2 and 5-3-2 features respectively have slightly higher perplexity than baseline model. In all three cases, bigram and trigram counts below 6 were discounted. The static interpolation of the prior model with the the baseline DI model results in an 8% reduction in perplexity.

Model	Description	Perplexity
Baseline	DI model	25.3
Prior	NAB model	426
Model 1	4-2-1 discounted	24.8
Model 2	4-3-2 discounted	26.9
Model 3	5-3-2 discounted	27.0
Model 4	0.1 NAB prior + 0.9 baseline	23.2

Count Thresholds	Number of Features
4-2-1	725 unigram 7200 bigram 38500 trigram
4-3-2	725 unigram 5200 bigram 15400 trigram
5-3-2	645 unigram 5200 bigram 15400 trigram

5. CONCLUSIONS

The choice of a suitable initial corpus and prior model, and the selection of features used for adaptation is important and challenging. The NAB corpus used here is somewhat more suitable for adaptation to Switchboard corpus than it is for ATIS, as reflected in the relative improvements of the adapted models. The computational complexity of the generalized iterative scaling algorithm used to construct the MDI adapted models is a limitation of this approach. The simpler method of interpolating the prior model with one built on the corpus of interest performed slightly better in terms of perplexity than the MDI method and seems preferable at this point, when sufficient data from the corpus is available to build a smoothed model. We are currently conducting recognition experiments to determine if the same can be concluded when word error rates are considered. In addition, we are further investigating various prior models and selection of features.

6. ACKNOWLEDGMENTS

The authors wish to acknowledge the many helpful discussions with their team members and other participants of the LM95 Workshop at the Center for Language and Speech Processing, Johns Hopkins University in 1995.

REFERENCES

[1] F. Jelinek and R. L. Mercer, "Interpolated estimation of Markov source parameters from sparse data," in *Pattern Recognition in Practice* (E. Gelsema and L. Kanal, eds.), Amsterdam: North-Holland, 1980.

[2] S. M. Katz, "Estimation of probabilities from sparse data for the language model component of a speech recognizer," *IEEE Trans. Acoustic,*

Speech, Signal Proc., vol. ASSP-35, pp. 400–401, March 1987.

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

[4] S. D. Pietra, V. D. Pietra, R. L. Mercer, and S. Roukos, "Adaptive language modeling using minimum discrimination estimation," in *Proc. ICASSP-92*, (San Fransisco), pp. I–633–636, March 1992.

[5] P. S. Rao, M. D. Monkowski, M. A. Picheny, and S. Roukos, "Language model adaptation via minimum discrimination information," in *Proc. ICASSP-95, Vol 1*, pp. 161–164, 1995.

[6] J. N. Darroch and D. Ratcliff, "Generalized iterative scaling for log-linear models," *Ann. Math. Stat.*, vol. 43, no. 5, pp. 1470–1480, 1972.

[7] I. Csiszar and G. Tusnady, "Information geometry and alternating minimization procedures," *Statistics & Decisions*, vol. Supplement Isuue, no. 1, pp. 205–237, 1984.