IN PROC. ICASSP ’99, PHOENIX, VOL. 1, PAGES 525 - 528

DISCRIMINATIVE ESTIMATION OF INTERPOLATION PARAMETERS FOR LANGUAGE MODEL CLASSIFIERS

V. Warnke, S. Harbeck, E. Nöth, H. Niemann and M. Levit

Chair for Pattern Recognition
University of Erlangen - Nuremberg
Martensstr. 3, 91058 Erlangen , Germany
{warnke,snharbec,noeth,niemann,mllevit}@informatik.uni-erlangen.de

ABSTRACT

In this paper we present a new approach for estimating the interpolation parameters of language models (LM) which are used as classifiers. With the classical maximum likelihood (ML) estimation theoretically one needs to have a huge amount of data and the fundamental density assumption has to be correct. Usually one of these conditions is violated, so different optimization techniques like maximum mutual information (MMI) and minimum classification error (MCE) can be used instead, where the interpolation parameters are not optimized on its own but in consideration of all models together. In this paper we present how MCE and MMI techniques can be applied to two different kind of interpolation strategies: the linear interpolation, which is the standard interpolation method and the rational interpolation. We compare ML, MCE and MMI on the German part of the Verbmobil corpus, where we get a reduction of 3% of classification error when discriminating between 18 dialog act classes.

1. INTRODUCTION

Language models (LM) are very important for automatic speech recognition systems; they are widely used in word recognizers to estimate the probability of a word chain in order to reduce the number of possible paths in forward decoding or to find the best word chain in a word hypotheses graph or lattice. If LM are trained class dependent and run in parallel, they can serve as classifiers for tasks like topic spotting, language identification and dialog act (DA) classification. LM work on every kind of symbol sequence with a finite vocabulary, e.g. word sequences, phoneme sequences, or codebook class sequences; they are thus applicable to many domains, even if there is no word information available.

We use LM classifiers in the task of topic spotting on the Switchboard-corpus with codebook classes or phonemes as basic symbol [9]. Furthermore, we perform language identification using codebook class sequences produced by one or more vector quantizer and decide for one language on the scores computed by our LM classifiers [4]. Within the speech-to-speech translation project Verbmobil [2] we use language models to classify and segment incoming turns in units of DA for the shallow processing module which uses a template based translation as fall strategy if the deep linguistic analysis fails. DA are, e.g., “greeting”, “confirmation of a date”, “suggestion of a place”. For our experiments we use the 18 DA from the first phase of Verbmobil which were defined based on their illocutionary force [5].

To address the problem of sparse data, often language models use interpolation strategies to get reliable performance even when there is not enough data available. The classical approach is to have a certain kind of interpolation strategy and to optimize the free interpolation parameters using maximum likelihood (ML) estimation. According to [7] this method is optimal when the fundamental density assumption is valid and enough data is available. At least one of these conditions is violated, so we will not get the optimal classifier. In [1] maximum mutual information (MMI) estimation was proposed as an alternative to ML estimation. MMIE training tries to find the parameter set maximizing the a posteriori probability of training data, which tends to be more reasonable since classification is usually performed by finding the model with the highest a posteriori probability. Another non-parametric approach is minimum classification error training, which tries to minimize a representation of error rate directly [6].

2. STOCHASTIC LANGUAGE MODELS

In most cases language models are used to calculate the probability of a word sequence $w = w^1 \ldots w^T$ in a given language or context. We use our *polygram language models*[11] which are a special kind of *stochastic n*-gram model to estimate the probability of every kind of *symbol sequence* where a symbol could be a word, phoneme or a codebook class.

2.1. Maximum Likelihood Estimation

Using polygrams the probability of the symbol sequence $w^1 \ldots w^T$ can be approximated with a N symbol history:

$$P(w) = \prod_{j=1}^{T} P(w^j \mid \underbrace{w^{\max(1,j-N+1)} \ldots w^{j-1}}_{:=v}).$$

This work was funded by the German Federal Ministry of Education, Science, Research and Technology (BMBF) in the framework of the Verbmobil Project under Grant 01 IV 102 H/0. The responsibility for the contents lies with the authors.

With this history v we can estimate the conditional probabilities $P(w \mid v)$ from a given training corpus simply by using the maximum likelihood (ML) estimation:

$$\hat{P}^k(w^j|v) = \frac{\#(vw^j)}{\#(v)}, \text{ with } k = \mid w^j v \mid,$$

where $\#(\cdot)$ denotes the frequency of its argument in the training data. Of course, one would like to choose a large number of N for the history length – the approximation made by a LM of higher order gets closer to the real probability. Unfortunately, the number of parameters to estimate increases exponentially with the size of N, and thus the ML estimates become far from being reliable because of the limited training data.

A compromise with respect to this conflict between the model context size N and the training data volume can be made by introducing a weighted interpolation scheme.

2.2. Interpolation

The basic idea of applying interpolation methods is to fall back on the probability estimation of subsequences shorter than N. An example is the *linear interpolation* which uses all subsequences up to the length N [8] :

$$\widetilde{P}(w^j \mid v) = \lambda_0 \cdot \frac{1}{L} + \lambda_1 \cdot P(w^j) + \lambda_2 \cdot P(w^j \mid w^{j-1}) + \sum_{n=3}^{N} \lambda_n \cdot P(w^j \mid w^{j-n+1} \dots w^{j-1}).$$

The fraction $1/L$ accounts for unseen sequences, where L is the number of words known to the LM, and ensures that no probabilities are set to zero. The interpolation coefficients λ_n can be estimated using the *Expectation Maximization (EM)* algorithm on a given validation set if we perform ML optimization.

Another interpolation method is the *rational interpolation* [11]; it gives a higher weight to those n-grams which have been seen more frequently in the training set using a weighting function $g_k(v)$:

$$\widetilde{P}(w^j|v) = \sum_k \lambda^k g_k(v) \hat{P}^k(w^j|v).$$

With the weighting function $g_k(v)$ defined as a hyperbolistic function

$$g_k(v) = \frac{\#(w^{j-k,j-1})}{\#(w^{j-k,j-1}) + C},$$

with the *bias* C we obtain the formula

$$\widetilde{P}(w^j|v) = \frac{\sum_k \lambda^k \frac{\#(w^{j-k,j})}{\#(w^{j-k,j-1})+C}}{\sum_k \lambda^k \frac{\#(w^{j-k,j-1})}{\#(w^{j-k,j-1})+C}} = \frac{\sum_k \lambda^k \phi^{j,k}(w)}{\sum_k \lambda^k \psi^{j,k}(w)}.$$

The classification of an utterance is done by choosing the LM which has the best a posteriori probability.

3. OPTIMIZATION METHODS

Before using the discriminative optimization techniques for estimation of LM interpolation parameters we describe MMI and the MCE approaches in more detail.

3.1. Maximum Mutual Information Estimation

The MMI approach is a discriminative extension of the maximum a posteriori estimation (MAP)[1]. In contrast to ML the a posteriori probability of one model is maximized under the assumption that one pattern of this model was observed. The objective function to maximize is the following

$$R(\Lambda) = \prod_{i=1}^{I} P(M_{q(i)}|w_i) = \prod_{i=1}^{I} \frac{P_{q(i)}(w_i) P_{q(i)}}{P(w_i)}$$

where I is the number of sentences in the validation set, Q is the number of considered language models, P_q is the a priori probability of model M_q and $q(i)$ refers to the correct model for the sentence i. For MMIE the denominator $P(w_i)$ is written as

$$R(\Lambda) = \prod_{i=1}^{I} \frac{P_{q(i)}(w_i) P_{q(i)}}{\sum\limits_{q=1}^{Q} P_q(w_i) P_q},$$

3.2. Minimum Classification Error

Another discriminative approach proposed in [3] is the minimum classification error approach. It has been successfully used in the domain of estimation of HMM parameters e.g. in [10]. The basic idea is the functional representation of the error function of the classifier. It is based on the *Sigmoid* function which is 1 for every correctly classified phrase and 0 otherwise. Instead of the real Sigmoid function an exponential approximation is used

$$\hat{\sigma}(k, x) = \frac{1}{1 + exp(-kx)}$$

One slightly difference from the classical version approach for the MCE leads to the objective function which is to be constructed the following way:

- Choose criterion $g_q(\Lambda, w_i)$ which is the basic score for the underlying classifier; in our researches we apply the Bayes classifier which leads to

  $$g_q(\Lambda, w_i) = -\log P(w_i|M_q) + \log P(M_q)$$

- For every phrase w_i find model $M_r(i)$ such that

  $$r(i) = \underset{q \neq q(i)}{\text{argmin}}\, g_q(\Lambda, w_i)$$

  The model $M_r(i)$ is hence the model with the highest probability of observation w_i but not the correct one.

- Build difference function δ with

  $$\delta(\Lambda, w_i) = g_{r(i)}(\Lambda, w_i) - g_{q(i)}(\Lambda, w_i)$$

- The probability to perform a correct classification of phrase w_i is written as

  $$R_i(\Lambda) = \frac{1}{1 + e^{-\delta(\Lambda, w_i)}}$$

  which tends to zero if $P(w_i|M_{q(i)}) \ll P(w_i|M_{r(i)})$ and to one if $P(w_i|M_{q(i)}) \gg P(w_i|M_{r(i)})$.

The overall objective function for MCE can be written as

$$R(\Lambda) = \prod_{i=1}^{I} R_i(\Lambda) = \prod_{i=1}^{I} \frac{1}{1 + e^{-\delta(\Lambda, w_i)}}$$

which can be interpreted as the probability for no classification errors for the whole validation set. Supposed we use uniform a priori distribution P_q, this can be written as

$$\begin{aligned} R(\Lambda) &= \prod_{i=1}^{I} \frac{1}{1 + e^{\log P_{r(i)}(w_i) - \log P_{q(i)}(w_i)}} \\ &= \prod_{i=1}^{I} \frac{P_{q(i)}(w_i)}{P_{q(i)}(w_i) + P_{r(i)}(w_i)}. \end{aligned}$$

4. USING MCE AND MMI FOR LANGUAGE MODEL INTERPOLATION

Estimation of the interpolation parameters as described in section 2 is done using the *General Probabilistic Descend* (GPD)[6] algorithm which implies the estimation of gradient vector for the objective function

$$x^{(l)} = x^{(l-1)} + \alpha^{(l)} \nabla F(x^{(l-1)}),$$

The value of α is estimated using the standard Monte-Carlo algorithm. Instead of optimizing of the interpolation parameters λ_t^k we substitute them by $\lambda_t^k = (\mu_t^k)^2$ for the linear interpolation and by

$$\lambda_t^k = \frac{(\mu_t^k)^2}{\sum_i (\mu_t^i)^2}$$

for the rational interpolation. This allows us to exclude the two stochastic conditions imposed on the weights λ_t^k.
The element of the gradient vector for the MMI algorithm can be transformed to

$$\begin{aligned} \frac{\partial \log R(\Lambda)}{\partial \mu_t^s} &= \sum_{\substack{i=1 \\ q(i)=t}}^{I} \sum_{j=1}^{|w_i|} \frac{\left(\widetilde{P}_t(w_i^j|v)\right)'_{\mu_t^s}}{\widetilde{P}_t(w_i^j|v)} \\ &\quad - \sum_{i=1}^{I} \frac{(P_t(w_i))'_{\mu_i^s} P_t}{\sum\limits_{q=1}^{Q} P_q(w_i) P_q}. \end{aligned}$$

and for the MCE algorithm to

$$\begin{aligned} \frac{\partial \log R(\Lambda)}{\partial \mu_t^s} &= \sum_{\substack{i=1 \\ q(i)=t}}^{I} \frac{(P_t(w_i))'_{\mu_i^s}}{P_t(w_i) + P_{r(i)}(w_i)} \frac{P_{r(i)}(w_i)}{P_t(w_i)} \\ &\quad - \sum_{\substack{i=1 \\ r(i)=t}}^{I} \frac{(P_t(w_i))'_{\mu_i^s}}{P_t(w_i) + P_{q(i)}(w_i)}. \end{aligned}$$

For both objective functions we need the derivations of the model related probabilities $P_q(w)$ and $\widetilde{P}_q(w^j|v)$ for the interpolation parameters μ_t^s. For the linear interpolation these are:

$$\left(\widetilde{P}_t(w^j|v)\right)'_{\mu_t^s} = \frac{2\mu_t^s}{\sum\limits_m (\mu_t^m)^2} \left(\hat{P}_t^s(w^j|v) - \widetilde{P}_t(w^j|v)\right)$$

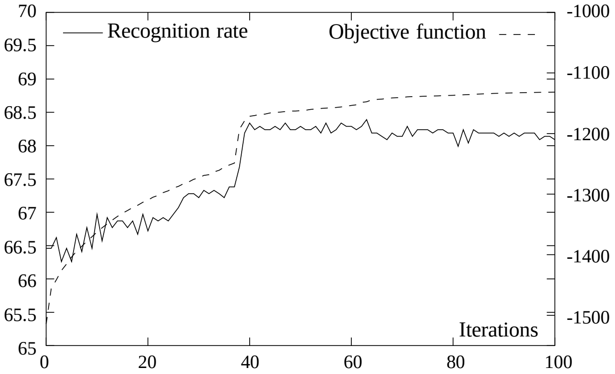

Figure 1: High correlation of the objective function and recognition rate on the validation set.

and

$$(P_t(w))'_{\mu_t^s} = \frac{2\mu_t^s}{\sum\limits_m (\mu_t^m)^2} P_t(w) \sum_{j=1}^{|w|} \left(\frac{\hat{P}_t^s(w^j|v)}{\widetilde{P}_t(w^j|v)} - 1 \right).$$

and for the rational interpolation:

$$\left(\widetilde{P}_t(w^j|v)\right)'_{\mu_t^s} = 2\mu_q^s \frac{\varphi_t^{j,s}(w) - \psi_t^{j,s}(w)\widetilde{P}_t(w^j|v)}{\sum_k (\mu_t^k)^2 \psi_t^{j,k}(w)},$$

$$\begin{aligned} (P_t(w))'_{\mu_t^s} &= 2\mu_t^s P_q(w) \\ &\quad \sum_{j=1}^{|w|} \left(\frac{\varphi_t^{j,s}(w)}{\sum\limits_k \left(\mu_t^k\right)^2 \varphi_t^{j,k}(w)} - \frac{\psi_t^{j,s}(w)}{\sum\limits_k \left(\mu_t^k\right)^2 \psi_t^{j,k}(w)} \right) \end{aligned}$$

5. EXPERIMENTS AND RESULTS

Until now we tested the different optimization techniques on data from the Verbmobil corpus for the task DA classification. We use a training set with 19795 phrases and a test set with 2540 phrases for the experiments with 18 DA classes with lexicon size of 4500 words. The validation set used for interpolation parameter optimization contains 1980 phrases which we excluded from the training set.

To get a feeling of the methods efficiency it is necessary to know how the applied objective function fits the recognition rate on the validation set during the iterations. In our experiments we reached a high correlation of these two values for both optimization methods (see Figure 1) which justifies the choice of the discriminative techniques we have made.

Furthermore it is remarkable that the monotone growth of the objective function does not implicate the permanent improvement of the recognition rate not even on the validation set. Indeed the resulting gradient vector composes of gradient vectors for every phrase of the validation set. This means that the general improvement of the recognition quality can be accompanied by the partial loss of the recognition rate owing those very phrases whose gradient direction has been suppressed by the majority of the set.

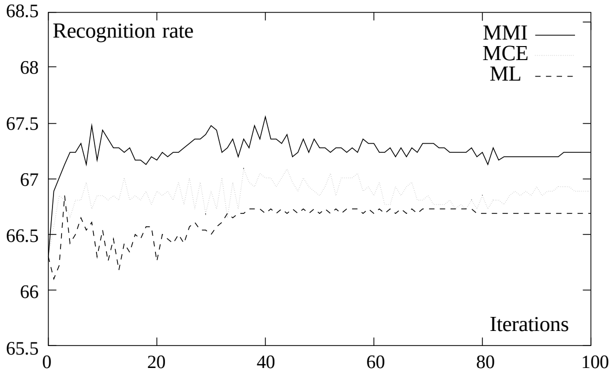

Figure 2: Comparison of recognition rates for ML, MCE and MMI estimations during interpolation process when using trigrams and rational interpolation.

As it can be seen in Figure 2 the recognition rate on the test set of MMIE is after a small number of interpolation iterations already much better than for ML whereas recognition rates for MCE and for ML are nearly the same. After 100 iterations we got a reduction of error rate of more than 3 percent when comparing MMIE and ML. Even MCE proved to be slightly better than ML. This makes sense because MMI seems to be “more discriminative” than MCE. In fact: on each iteration step of MMI optimization every phrase from the validation set causes alteration of all models coefficients whereas with MCE only parameter of two models ($r(i)$ and $q(i)$) are to be modified.

Comparing our different interpolation strategies for both optimization techniques the rational interpolation outperforms the linear interpolation even if we use different n-gram sizes (see Table 1).

n	*linear*			*rational*		
	ML	*MMI*	*MCE*	*ML*	*MMI*	*MCE*
2	63.3 %	66.7 %	66.5 %	66.3 %	67.6 %	66.9 %
3	63.6 %	66.9 %	66.4 %	66.6 %	67.2 %	66.8 %

Table 1: Recognition rates for MMI, MCE and ML using linear and rational interpolation for bi- and trigrams.

6. CONCLUSION AND FUTURE WORK

In [1] it was shown, that using discriminative optimization techniques for estimation of HMMs parameters improved recognition rate. We applied MMI and the MCE techniques in order to estimate interpolation parameters of LM. We could show that discriminative optimization techniques of interpolation coefficients improve recognition results for the 18 class problem in the task of dialog act classification. The best results we achieved using the rational interpolation and MMI estimation which cuts our error by 3 percent in comparison to ML estimation. In the future we are going to test the discriminative techniques in other tasks like topic spotting on the Switchboard-corpus and language identification. We would like to extend the estimation techniques to different interpolation strategies and to estimate the n-gram probabilities themself using discriminative methods.

7. REFERENCES

[1] L. Bahl, P.Brown, P. de Souza, and R. Mercer. Maximum Mutual Information Estimation of Hidden Markov Model Parameters for Speech Recognition. In *Int. Conf. on Acoustics, Speech and Signal Processing*, pages 49–52, Tokyo, 1986.

[2] T. Bub and J. Schwinn. Verbmobil: The Evolution of a Complex Large Speech-to-Speech Translation System. In *Int. Conf. on Spoken Language Processing*, volume 4, pages 1026–1029, Philadelphia, 1996.

[3] W. Chou, B.H. Juang, and C.H. Lee. Segmental GPD Training of HMM Based Speech Recognizer. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 473–476, San Francisco, CA, 1992.

[4] S. Harbeck, E. Nöth, and H. Niemann. Multilingual Speech Recognition in the Context of Multilingual Information Retrieval Dialogues. In *Proc. of the Workshop on TEXT, SPEECH and DIALOG (TSD'98)*, pages 375–380, Brno, 1998. Masaryk University.

[5] S. Jekat, A. Klein, E. Maier, I. Maleck, M. Mast, and J. Quantz. Dialogue Acts in VERBMOBIL verbmobil–report–65–95, April 1995.

[6] B. Juang, P. Chang, W. Chou, and C. Lee. Minimum error rate training for dynamic time warping and hidden markov model recognizers. In *IEEE Speech Recogition Workshop*, pages 14–15, 1991.

[7] A. Nadas. A decision theoretic formulation of a training problem in speech recognition and a comparison of training by unconditional versus conditional maximum likelihood. pages 814–817, 1983.

[8] H. Ney, U. Essen, and R. Kneser. On Structuring Probabilistic Dependences on Stochastic Language Modelling. *Computer Speech & Language*, 8(1):1–38, 1994.

[9] E. Nöth, S. Harbeck, H. Niemann, and V. Warnke. A Frame and Segment Based Approach for Topic Spotting. In *Proc. European Conf. on Speech Communication and Technology*, pages 275–278, Rhodes, Greece, 1997.

[10] K.K. Paliwal, M. Bacchiani, and Y. Sagisaka. Minimum classification error training algorithm for feature extractor and pattern classifier in speech recognition. pages 541–544, 1995.

[11] E.G. Schukat-Talamazzini, F. Gallwitz, S. Harbeck, and V. Warnke. Rational Interpolation of Maximum Likelihood Predictors in Stochastic Language Modeling. In *Proc. European Conf. on Speech Communication and Technology*, pages 2731–2734, Rhodes, Greece, 1997.