MINIMUM WORD ERROR BASED DISCRIMINATIVE TRAINING OF LANGUAGE MODELS

Jen-Wei Kuo and Berlin Chen

Graduate Institute of Computer Science & Information Engineering,
National Taiwan Normal University, Taipei, Taiwan
{rogerkuo, berlin}@csie.ntnu.edu.tw

ABSTRACT

This paper considers discriminative training of language models for large vocabulary continuous speech recognition. The minimum word error (MWE) criterion was explored to make use of the word confusion information as well as the local lexical constraints inherent in the acoustic training corpus, in conjunction with those constraints obtained from the background text corpus, for properly guiding the speech recognizer to separate the correct hypothesis from the competing ones. The underlying characteristics of the MWE-based approach were extensively investigated, and its performance was verified by comparison with the conventional maximum likelihood (ML) approaches as well. The speech recognition experiments were performed on the broadcast news collected in Taiwan.

1. INTRODUCTION

Statistical language modeling, which aims to capture the regularities in human natural language and quantify the acceptance of a given word sequence, has continuously been an important research issue in a wide variety of applications of natural language processing (NLP) over the past three decades. The *n*-gram modeling (especially the bigram and trigram modeling) approach, which determines the probability of a word given the previous *n*-1 word history, is most prominently used [1]. The *n*-gram language models are normally trained based on the maximum likelihood (ML) criterion. Nevertheless, for complicated NLP tasks such as speech recognition, in which the principal role of the language models is to help resolve the acoustic confusion and thus separate the correct hypothesis from the competing ones, the ML-trained language models are not always capable of achieving minimum recognition error rates. To circumvent this problem, in the recent past, the discriminative training, which was developed in an attempt to correctly discriminate the recognition hypotheses for the best recognition results rather than just to fit the model distributions, has been successful introduced to language model estimation [2]-[3], and the minimum classification error (MCE) approach is often considered as one of the representatives in this category. The MCE-based language models have been shown to have superior performance over the conventional ML-based ones on small vocabulary speech recognition tasks [3]; however, to our knowledge, there are still no known results reported for large vocabulary speech recognition tasks.

In this paper, we present a minimum word error (MWE) approach for discriminative training of language models for large vocabulary continuous speech recognition. MWE and its variant, MPE (minimum phone error), are viewed as two alternatives to MCE [4]. Either MWE or MPE is intended to maximize the expected word or phone accuracy and can be easily computed using different kinds of search structures such as word graphs, lattices or confusion networks. They also have been recently shown very effective for large vocabulary speech recognition, especially in training [4]-[7] or adaptation [8] of acoustic models, and in feature extraction [9][10]. Hence, we study here the use of the MWE criterion for language model training, in order to make use of the word confusion information as well as the local lexical constraints inherent in the acoustic training corpus, in combination with those constraints obtained from the background text corpus, for properly guiding the speech recognizer to separate the correct hypothesis from the competing ones. Unlike the other discriminative approaches, such as the MCE-based approach, in which only the correct hypothesis can provide a positive contribution to model estimation, in the MWE-based approach, those hypotheses that have word accuracies higher than the average also can provide positive contributions, which are weighted in proportion to both the word accuracy and the model likelihood. The underlying characteristics of the MWE-based approach were extensively investigated, and its performance was verified by comparison with the conventional ML-based approaches as well. The speech recognition experiments were carried out on the broadcast news collected in Taiwan.

The rest of this paper is organized as follows. More details of MWE-based language model estimation are explained in Section 2. In Section 3, the system configuration of our experiments on the Mandarin broadcast news transcription task is described. Then, a series of speech recognition experiments are presented in Section 4. Finally, conclusions are drawn in Section 5.

2. MWE OBJECTIVE FUNCTION AND LANGUAGE MODEL ESTIMATION

Given a training set of observation sequences $O=\{O_1,..,O_u,..,O_U\}$, the MWE criterion for acoustic and language model training aims to minimize the expected word errors of these observation sequences using the following objective function [4]:

$$F_{MWE}(O,\Lambda,\Gamma)=\sum_{u=1}^{U}\sum_{S_u}\frac{P_\Lambda(O_u\mid S_u)^\kappa P_\Gamma(S_u)}{\sum_{S'_u}P_\Lambda(O_u\mid S'_u)^\kappa P_\Gamma(S'_u)}RawAcc(S_u),\quad(1)$$

where Λ and are Γ respectively the acoustic model and language model parameter sets, S_u is one of the hypothesized word sequences for the training observation sequence O_u, $P_\Lambda(O_u\mid S_u)$ is the acoustic score for S_u, $P_\Gamma(S_u)$ is the language model probability for S_u, $RawAcc(S_u)$ is the "raw word accuracy" of S_u in comparison to the corresponding reference transcript, and κ is a scaling factor for reducing the dynamic range of acoustic scores. As κ is set equal to zero, it means that each hypothesis is merely weighted by its normalized language model likelihood. Hence, for each training sequence O_u, the MWE criterion gives a weighted average over the raw accuracy $RawAcc(S_u)$ of all hypothesized word sequences S_u, and the hypothesized word sequences of each training sequence O_u can be efficiently approximated by a word graph structure as depicted in Figure 1. The raw accuracy for each S_u can be calculated in terms of the sum of the accuracy of each word contained in S_u:

$$RawAcc(S_u)=\sum\nolimits_{w_i\in S_u} WordAcc(w_i), \tag{2}$$

where $WordAcc(w_i)$ is the "raw word accuracy" for a word w_i in S_u, which can be defined as follows:

$$WordAcc(w_i)=\max_{z_j\in Z_u}\left\{\begin{array}{ll}-1+2e(z_j,w)/l(z_j), & z_j=w_i\\ -1+e(z_j,w)/l(z_j), & z_j\neq w_i\end{array}\right\}, \tag{3}$$

where Z_u is the set of words in the corresponding reference transcript, and $e(z_j,w_i)$ is the overlap length in time for a word z_j in Z_u and a hypothesized word w_i in S_u, $l(z_j)$ is the length in time for z_j. Unlike the other discriminative approaches, such as the MCE-based approach, in which only the correct hypothesis can provide a positive contribution to model estimation, in the MWE-based approach, those hypotheses that have word accuracies higher than the average also can provide positive contributions, which are weighted in proportion to both the word accuracy and the model likelihood. It is obvious that under the MWE criterion, we can optimize the acoustic model and language model parameters jointly; however, we would like the optimization to concentrate here on finding a better language model by performing the optimization of language model parameters as a standalone process.

We thus can maximize the objective function defined in Equation (1) by applying the Extended Baum-Welch (EBW) algorithm [11] to update the language model probabilities as follows:

$$P_{MWE}(w_i\mid h)=\frac{P_\Gamma(w_i\mid h)\left[\dfrac{\partial F_{MWE}(O,\Lambda)}{\partial P_\Gamma(w_i\mid h)}+C\right]}{\sum_k P_\Gamma(w_k\mid h)\left[\dfrac{\partial F_{MWE}(O,\Lambda)}{\partial P_\Gamma(w_k\mid h)}+C\right]}, \tag{4}$$

where $P_{MWE}(w_i\mid h)$ and $P_\Gamma(w_i\mid h)$ are respectively the new estimated and old language model probabilities for the word w_i given that its previous word history is h (for example, h can be a word pair w_{i-2},w_{i-1} for trigram modeling and a single word w_{i-1} for bigram modeling), and C is a constant used to ensure positive language model probabilities. The language model probability $P_{MWE}(w_i\mid h)$ can be further estimated using the following equation:

$$P_{MWE}(w_i\mid h)=\frac{\gamma^{num}_{(h,w_i)}-\gamma^{den}_{(h,w_i)}+C\cdot P_\Gamma(w_i|h)}{\sum_{w_k}\left[\gamma^{num}_{(h,w_k)}-\gamma^{den}_{(h,w_k)}+C\cdot P_\Gamma(w_k|h)\right]}, \tag{5}$$

where

$$\gamma^{num}_{(h,w_i)}=\sum\nolimits_{u=1}^{U}\max\left(0,\gamma^{MWE}_{(h,w_i)\in u}\right), \tag{6}$$

$$\gamma^{den}_{(h,w_i)}=\sum\nolimits_{u=1}^{U}\max\left(0,-\gamma^{MWE}_{(h,w_i)\in u}\right), \tag{7}$$

and

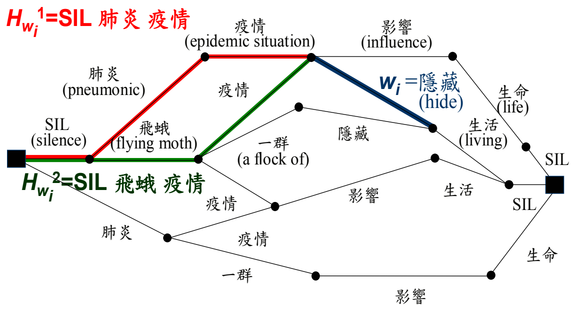

Figure 1: An illustration of the word graph, in which each arc, together with its corresponding start and end speech frames, represents a candidate word hypothesis.

$$\gamma^{MPE}_{(h,w_i)\in u}=\gamma_{(h,w_i)\in u}\left(c^u_{(h,w_i)}-c^u_{avg}\right) \tag{8}$$

For each training observation sequence O_u, $\gamma_{(h,w_i)\in u}$ is the posterior probability of word sequence (h,w_i) occurring in the associated word graph, c^u_{avg} is the expected word accuracy over all hypothesized word sequences, and $c^u_{(h,w_i)}$ is the average word accuracy over all hypothesized word sequences that contain (h,w_i). The statistics required for probability updates can be efficiently accumulated by performing the forward-backward algorithm on the word graph. More detailed derivations of the MWE training formulae also can be found in [5].

On the other hand, as it is well known that there are no explicit marks, such as the spaces or blanks, separating words in the Chinese language, the Chinese language thus often suffers from the word tokenization problems. The performance evaluation metric in Mandarin speech recognition usually is the character error rate (CER) rather than the word error rate (WER). In this study, the "raw word accuracy" in Equation (1) is replaced with the "raw character accuracy."

3. EXPERIMENTAL SETUP

The large vocabulary continuous speech recognition system as well as the experimental speech and language data used in this paper will be described in this section.

3.1. Front-End Signal Processing

Each frame of the speech data is represented by a 39 dimensional feature vector, which consists of 12 MFCCs and log energy, and their first and second differences. Utterance-based cepstral mean subtraction (CMS) is applied to all the training and testing materials.

3.2. Speech Corpus and Acoustic Model Training

The speech corpus consists of about 200 hours of MATBN Mandarin television news (Mandarin Across Taiwan Broadcast News) [12], which were collected by Acdemia Sinica and Public Television Service Foundation of Taiwan during November 2001 and April 2003. All the speech materials were manually segmented into separate stories, and each of them is pronounced by one anchor speaker, several field reporters and interviewees. Some stories contain background noise, speech and music. All the 200 hours of speech data are equipped with corresponding orthographic transcripts, in which about 25 hours of gender-balanced speech data of the field reporters collected during November 2001 to December 2002 were used to bootstrap the acoustic training. Another set of 1.5 hour speech data of the field reporters collected within 2003 were reserved for testing. On the other hand, the acoustic models chosen here for speech recognition are 112 right-context-dependent INITIALs and 38 context-independent FINAL's. Each INITIAL is represented by an HMM with 3 states while each FINAL with 4 states. Gender-independent models were used.

The acoustic models were first trained at optimum settings using the ML criterion as well as the Baum-Welch training algorithm. MPE-based and MMI-based (Maximum Mutual Information) [13] discriminative acoustic model training approaches were respectively further applied to those acoustic models previously trained by the ML criterion. Unigram language model constraints were used in accumulating the training statistics from the word graphs for discriminative training. For MPE-based discriminative training, both silence and short pause labels are also involved in the calculation of the accuracies of the hypothesized word sequences.

3.3. Lexicon and *N*-gram Language Modeling

The recognition lexicon initially consists of 67K words. A set of about 5K compound words was automatically derived using the forward and backward bigram statistics and was then added to the lexicon to form a new lexicon of 72K words. The background language models used in this paper consist of trigram and bigram models, which were estimated based on the ML criterion and using a text corpus consisting of 170 million Chinese characters collected from Central News Agency (CNA) in 2001 and 2002 (the Chinese Gigaword Corpus released by LDC). The *n*-gram language models were trained with Katz backoff smoothing using the SRI Language Modeling Toolkit (SRILM) [14].

On the other hand, the orthographic (or reference) transcripts of the 25-hour training utterances, consisting of about 500K Chinese characters, are postulated to be stylistically consistent with the broadcast news speech to be tested and regarded here as the in-domain text corpus. Thus, they are believed to contain the extra linguistic information that is helpful for the language modeling of speech recognition. For the ML-based approach, a new set of language models were trained on these orthographic transcripts and then were interpolated with the ML-trained background language models using the following equation:

$$\tilde{P}_{ML}(w_i \mid h) = \alpha \cdot P_{ML-BG}(w_i \mid h) + (1-\alpha) \cdot P_{ML-IN}(w_i \mid h), \quad (9)$$

where $P_{ML-BG}(w_i \mid h)$ and $P_{ML-IN}(w_i \mid h)$ respectively are the ML-trained background language model probability and the ML-trained in-domain one, and α is a tunable weighting parameter. While for the MWE-based approach proposed in this paper, both the orthographic transcripts of the training utterances and their associated word graphs (consisting of 23M Chinese characters) that were generated beforehand by the speech recognition system, were all involved in language model training.

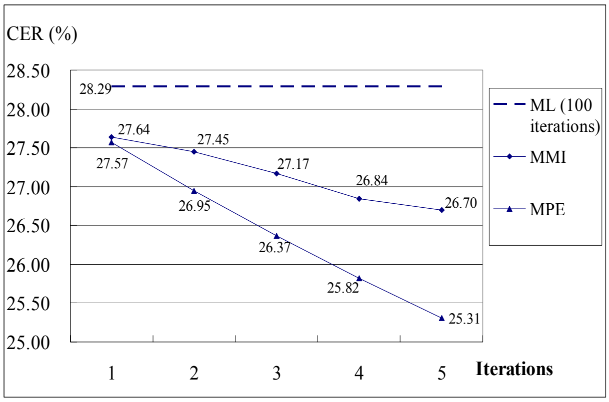

Figure 2: The speech recognition results (CERs) for three different baseline systems trained with ML, MMI and MPE criteria, respectively.

3.4. Speech Recognition

The speech recognizer was implemented with a left-to-right frame-synchronous Viterbi tree-copy search as well as a lexical prefix tree organization of the lexicon. At each speech frame, a beam pruning technique, which considered the decoding scores of path hypotheses together with their corresponding unigram language model look-ahead scores and syllable-level acoustic look-ahead scores [15], was used to select the most promising path hypotheses. Moreover, if the word hypotheses ending at each speech frame had scores higher than a predefined threshold, their associated decoding information, such as the word start and end frames, the identities of current and predecessor words, and the acoustic score, will be kept in order to build a word graph for further language model rescoring. In this study, the word bigram language model was used in the tree search procedure while the trigram language model was used in the word graph rescoring procedure.

4. EXPERIMENTAL RESULTS

4.1. Baseline Experimental Results

The baseline broadcast news systems were alternatively configured with three different sets of acoustic models, respectively trained using the ML, MMI, and MPE criteria. Figure 2 shows the character error rates (CERs) for acoustic model training on the 25 hour speech corpus, in which both the MMI and MPE training approaches started with the acoustic models obtained by 100 ML training iterations, and used the information contained in the associated word graphs of training utterances to accumulate the statistics for discriminative acoustic model training. Moreover, in all three baseline systems, the trigram and bigram language models were estimated using the background text corpus collected form CNA during 2001 and 2002. As can be seen from Figure 2, the ML-trained acoustic models (at the 100th iteration) yield a CER of 28.29%. On the other hand, either the MMI or MPE training approach works out very well. Both of them can provide a great boost to the acoustic models initially trained by ML, and the acoustic models trained by MPE further outperform those models trained by MMI, consistently at all training iterations. In summary, the MPE-trained

	ML-BG	ML-IN + ML-BG	MWE-IN + ML-BG				
Iterations	-	-	1	2	3	4	5
CERs (%)	25.31	24.14	24.12	24.09	24.09	24.09	24.09

Table 1: The speech recognition results (CERs) for the MWE-based language model training approach, in comparison to the conventional ML-based approaches.

acoustic models (at the 5th iteration) give a relative CER reduction of 10.53% over those trained by ML. Thus, the MPE-trained acoustic models are chosen as the default acoustic model set for the following experiments.

4.2. Experiments on MWE-based Language Models

We investigate here the use of the speech training utterances and its associated reference transcripts and word graphs for language model training, in conjunction with those lexical constraints obtained from the background text corpus. The experimental results are shown in Table 1. First, Column 2 (ML-BG) denotes the recognition result using the set of language models trained based on the ML criterion and with the background text corpus. The CER for such language models is 25.31%, which is just the best result shown in Figure 2. Then, another set of language models trained based on the ML criterion and with the reference transcripts of speech training utterances (500K Chinese characters) are used to combine with the above ML-trained background languages via simple linear model interpolation, as expressed in Equation (9). It can be seen form Column 3 (ML-IN+ML-BG) that, the CER can be significantly reduced from 25.31% to 24.14% (a relative improvement of 4.62%), which indicates that the stylistic (or wording) information inherent in the reference speech transcripts are indeed very helpful for language modeling, though the size of the reference speech transcripts is very small. Finally, we examine the performance level of the MWE-based language models, which were trained on the reference speech transcripts and their associated word graphs and also were linearly interpolated with the ML-trained background language models during speech recognition. Columns 4 to 8 (MWE-IN+ML-BG) are the recognition results for the MPE-based language models trained at different iterations. It can be found that, the MWE-based approach also provides significant improvements over the baseline ML-based approach (ML-BG) that only adopts the background text corpus as the training material. Comparing to the ML-based approach (ML-IN+ML-BG) that additionally includes the reference speech transcripts as the training material, the MWE-based approach are consistently better than the ML-based approach as well for all training iterations (a best result of 24.09% CER was initially achieved), though the relative performance gains are not apparent.

In the mean time, we are extensively experimenting on the ways to improve the performance of the MWE-based language models, including trying different sets of training settings, investigating the joint training of discriminative acoustic and language models and the possibility to estimate the MWE-based language models on a larger collection of broadcast news without reference transcripts in a purely unsupervised manner, etc.

5. CONCLUSIONS

In this paper, we have explored the use of the MWE criterion for discriminative training of language models for large vocabulary continuous speech recognition. The underlying characteristics of the MWE-based language model training approach have been investigated, and its performance was verified by comparison with the conventional ML-based approaches as well. More in-deep investigation of the MWE-based approach, as well as comparison and integration with other language modeling approaches are also currently undertaken [16].

6. REFERENCES

[1] R. Rosenfeld, “Two Decades of Statistical Language Modeling: Where Do We Go from Here,” *Proc. IEEE*, 88 (8), 2000.

[2] Z. Chen, K. F. Lee, and M. J. Li, “Discriminative training on language model,” in *Proc. ICSLP 2000*.

[3] H. K. Kuo, E. Fosler-Lussier, H. Jiang, and C. H. Lee, “Discriminative Training of Language Models for Speech Recognition,” in *Proc. ICASSP 2002*.

[4] D. Povey and P.C. Woodland, “Minimum Phone Error. and I-Smoothing for Improved Discriminative Training,” in *Proc. ICASSP 2002*.

[5] D. Povey “Discriminative Training for Large Vocabulary Speech Recognition,” Ph.D Dissertation, Peterhouse, Cambridge, July 2004.

[6] L. Wang and P. C. Woodland, “Discriminative Adaptive Training using The MPE Criterion,” in *Proc. ASRU 2003*.

[7] K. Yu and M. J. F. Gales, “Adaptive Training using Structured Transforms,” in *Proc. ICASSP 2004*.

[8] L. Wang and P. C. Woodland, “MPE-Based Discriminative Linear Transform for Speaker Adaptation,” in *Proc. ICASSP 2004*.

[9] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, G. Zweig, “fMPE: Discriminatively Trained Features for Speech Recognition,” in *Proc. ICASSP 2005*.

[10] B. Zhang and S. Matsoukas, “Minimum Phoneme Error based Heteroscedastic Linear Discriminant Analysis for Speech Recognition,” in *Proc. ICASSP 2005*.

[11] P. S. Gopalakrishnan, D. Kanevsky, A. Nadas, D. Nahamoo, “A Generalization of the Baum Algorithm to Rational Objective Functions," in *Proc. ICASSP 1989*.

[12] H. M. Wang, B. Chen, J. W. Kuo and S. S. Cheng, “MATBN: A Mandarin Chinese Broadcast News Corpus,” *International Journal of Computational Linguistics & Chinese Language Processing*, June 2005.

[13] D. Povey and P. C. Woodland, “Large Scale Discriminative Training of Acoustic Models for Speech Recognition,” *Computer Speech & Language*, Vol. 16, pp. 25-47, 2002.

[14] A. Stolcke, “SRI language Modeling Toolkit,” version 1.3.3, http://www.speech.sri.com/projects/srilm/.

[15] B. Chen, J. W. Kuo, W. H. Tsai, "Lightly Supervised and Data-Driven Approaches to Mandarin Broadcast News Transcription," in *Proc. ICASSP 2004*.

[16] B. Chen, W. H. Tsai, J. W. Kuo, “Statistical Language Model Adaptation for Mandarin Broadcast News Transcription,” in *Proc. ISCSLP 2004*.