Discriminative Language Model With Part-of-speech for Mandarin Large Vocabulary Continuous Speech Recognition System

Yujing Si, Zhen Zhang, Qingqing Zhang, Jielin Pan, Yonghong Yan
The Key Laboratory of Speech Acoustics and Content Understanding,
Chinese Academy of Sciences, Beijing 100190, P.R.China
{siyujing, zhangzhen, zhangqingqing, panjielin, yanyonghong}@hccl.ioa.ac.cn

***Abstract*—Statistical language model, trained by a large number of text corpus, is an integral component in many speech and natural language model processing systems, such as speech recognition and machine translation. It is a probabilistic model which describes the distribution pattern of natural language. Over the last few decades, N-gram language model (LM) is the most popular technique since it is simple and effective. However, the training of the N-gram language model is based on the maximum likelihood rule resulting in suboptimal output in speech recognition systems. In this paper, a discriminative training based language model (DLM) which directly focused on minimizing speech recognition word error rate (WER) was employed to improve the performance of speech recognition system. In particular, the part-of-speech (POS) feature was used to train DLM as well as the n-gram features. Experimental results showed that DLM with n-gram features gave 1% absolute reduction in word error rate (WER). Combining n-gram features with POS feature, DLM could obtain another 0.4% absolute reduction in WER.**

Keywords- speech recognition; language model; DLM; POS

I. INTRODUCTION

With the development of internet, more and more embedded devices are getting into people`s life, such as IPad, IPhone, Google Nexus, Galaxy Note and so on. These devices have many interesting functions which make people`s life more convenient. However, people sometimes feel laborious when communicating with these devices because the interface of devices is not friendly enough.

Recently, speech recognition technique [1] has attracted more and more attentions because speech is the most nature way for human to communicate with each other and speech recognition technique could drastically improve the User Experience (UE) of intelligent devices, such as Siri developed by Apple Corporation.

Statistical language model is an integral part of speech recognition system. The purpose of language model is to assign a non-zero probability to any possible words sequence. Over the last few decades, N-gram language model (LM) [2] is the most popular technique since it is simple and effective. However, the training of the N-gram language model is based on the maximum likelihood rule resulting in suboptimal output in speech recognition systems. The result obtained by N-gram LM usually has the highest probability but it may be not the optimal. In other words, the probability of another hypothesis in the n-best list may be lower but it is more similar to the reference. Discriminative training of language models [3] has been recently introduced to obtain improved parameter estimates for language models (see Section II). The advantage of discriminative parameter estimation to MLE is that discriminative training takes negative examples into account as well as the positive examples and therefore results in a better discrimination between alternative classes. Positive examples are the correct transcriptions and negative examples are erroneous candidate transcriptions. DLM parameters are trained by optimizing an objective function that is directly related to the system performance, word error rate (WER) in ASR systems. Estimation of the DLM parameters is relatively straightforward for linear models. In addition to the improved parameter estimates with discriminative training, another advantage of this DLM approach to the conventional n-gram language model is that it is a featured-based approach. Consequently, it allows for easy integration of many relevant knowledge sources into language Model [4].

In this paper, we explore DLM in the context of Mandarin Large Vocabulary Continuous Speech Recognition (LVCSR). DLM is used to re-rank the n-best list obtained by the first pass of LVCSR and output a new hypothesis which has minimal WER. In particular, DLM was trained with POS feature as well as the conventional n-gram features and system performance could be further improved.

The rest sections are organized as follows: the precious work is described in Section II, the framework of our ASR system is described in Section III and the principle of discriminative language modeling is provided in Section IV. In section V, we detail the features used to train DLMs. Finally, experimental results are given in section VI followed by conclusion in section VII.

II. PRECIOUS WORK

Discriminative estimation of hidden Markov model (HMM) parameters in the form of maximum mutual information estimation (MMIE) or minimum phone error (MPE) has made a great achievement in the large vocabulary continuous speech recognition system (LVCSR) [5]. Recently, discriminative estimation of language model has also been proposed to improve the performance of language model [6]. In [7], a generalized probabilistic descent (GPD) algorithm was used to train relatively small language models which attempt to minimize string error rate. GPD based discriminative language model is actually an instance of the widely known minimum classification error (MCE) training

Published by Atlantis Press, Paris, France.

[8]. And then, linear models have been successfully applied to discriminative language model for speech recognition [3]. In linear models based DLM, model parameters are used to define a score, S(W,A), on the word sequence which also includes the likelihood P(W,A) from the baseline recognizer:

$$S(W,A)=\alpha_0 \log P(W,A)+\sum_i \alpha_i \Phi_i(W) \quad (1)$$

Here $\Phi_i(W)$ represents sentence level features and α_i is the parameter of feature, which can be estimated discriminatively using many methods such as the perceptron algorithm [9], a method based on maximizing the regularized conditional log-likelihood and so on. A comparison of various training methods for DLMs is given in [10]. Recently, semi-supervised discriminative language modeling has also been explored [11] [12] [13]. The work in [4] is the most related to our approach. However, we focus on improving the performance of Mandarin speech recognition on a more difficult task. And, n-gram features as well as POS tags were used to train DLM.

III. THE FRAMEWORK OF ASR SYSTEM

Our Mandarin speech recognition system [14] is a two-pass system containing three main modules: a front-end, Large Vocabulary Continuous Speech Recognition (LVCSR) and n-best re-scoring module as shown in Figure 1. The front-end module consists of two components: Voice Activity Detection (VAD), Feature Extract and Processing (FEP). The function of VAD is to detect the start point and end point of the submitted input voice. Once the start point has been detected, voice data are sent to the following FEP to online extract feature and further transformation. The final feature is recognized and the output of LVCSR module is a *word lattice* from which the n-best list is extracted by n-best extractor module using Viterbi algorithm [15]. In the second pass, discriminative language model is employed to reorder the hypotheses in the n-best list. Thus, we finally obtain a new best hypothesis which is considered as final recognized result.

IV. DISCRIMINATIVE LANGUAGE MODELING

A. Problem definition

Generally speaking, discriminative language model, minimizing the word error rate, is used to re-rank the n-best list or word graph output by the first pass of LVCSR, and put a new hypothesis whose WER is minimal as the final recognized result. In practice, DLM uses the original score of speech recognizer as well as more advanced language features to re-score the hypotheses of n-best list. The discriminative language model is defined as followings [16]:

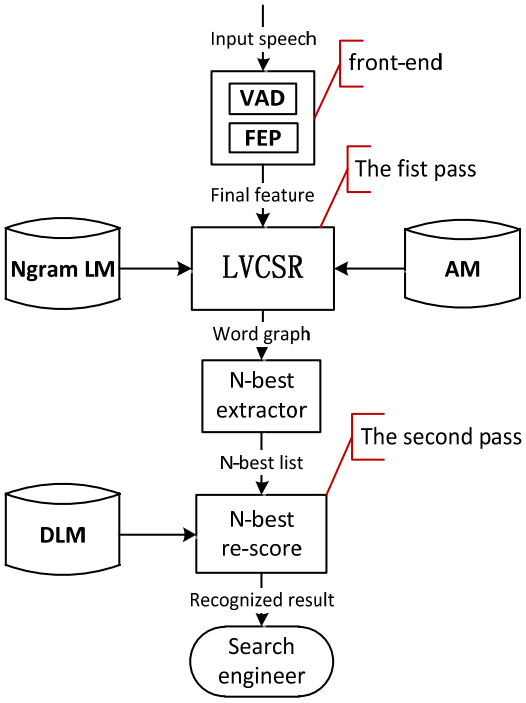

Figure 1. The framework of LVCSR

1) The top N recognized hypotheses of a speech signal x_i can be represented as $Gen(x_i)=\{W_{i,j}\}\ \ 1\le i\le M$.

2) The training data of DLM can be defined as $(x_i, W_i^r)\ \ 1\le i\le L$, in which L is the number of the training data. W_i^r is the hypothesis whose WER is minimal in the n-best list.

3) Define a (D+1) dimension feature vector $f_d(w_{i,j})$ for every hypothesis. $f_0(w_{i,j})$ represents the base score and $f_{d\ge 1}(w_{i,j})$ means the advanced language features such as n-gram features.

4) Define the parameters of features as $(\lambda_0,\lambda_1,\cdots,\lambda_D)$ which need to be estimated.

Based on the above definitions, the new score of hypothesis given by DLM is:

$$Score(W_{i,j},\lambda)=\lambda\times f(W_{i,j})=\sum_{d=0}^{d=D}\lambda_d f_d(W_{i,j}) \quad (2)$$

And the new hypothesis is:

$$W_i^* = \arg\max_{W_i^j\in GEN(x_i)} score(W_{i,j},\lambda) \quad (3)$$

The purpose of discriminative language modeling training is to get the optimal parameters λ which minimize the word error rate.

B. The perceptron algorithm

The perceptron algorithm [9] takes the N-best re-ranking task as a binary classification problem, attempting to award

the positive examples and punish the negative examples by adjusting the parameters. Instead of minimizing the training error directly, perceptron optimizes a minimum square error (MSE) loss function.

The perceptron algorithm based discriminative language model is described in Figure 2. It can be seen that the perceptron algorithm is incremental, meaning that the language model D is built one training example at a time, during several passes over the training set. The framework of DLM makes it easy to integrate the advanced language features.

Inputs: Training examples (x_i, y_i)

Initialization: Set $\overline{\alpha} = 0$

Algorithm:

For t=1…T

For i=1…N

Calculate $z_i = \arg\max_{z \in GEN(x_i)} \Phi(x_i, z) \cdot \overline{\alpha}$

If($z_i \neq y_i$)then $\overline{\alpha} = \overline{\alpha} + \Phi(x_i, y_i) - \Phi(x_i, z_i)$

Output: Parameters $\overline{\alpha}$

Figure 2. The perceptron algorithm

V. DLM FEATURES

This section describes the features which are used to train Mandarin DLMs. Features are extracted from the N-best list output by the first pass of LVCSR. In this work, we used the n-gram features as well as the POS features obtained by tools ICTCLAS [17]. Besides, sentence start symbol "<s>" and sentence end symbol "</s>" are added to each hypothesis before feature extraction.

A. N-gram features

N-grams which are defined as the count of n-grams in the candidate hypothesis were widely-used features in DLM training [18]. An example word trigram feature is as follows:

$\Phi_i(x, y) =$ Number of times "wish word peace" is seen in a hypothesis of n-best list.

In the n-best list re-scoring phase, we select the recognizer score (an interpolation of the acoustic and language model score) of each hypothesis as the base feature f_0 and define the remaining features, $f_i(h), i = 1 \cdots D$ as unigram/bigram/trigram features.

B. Part-of-speech features

In grammar, a part of speech (also a word class, a lexical class, or a lexical category) [19] is a linguistic category of words (or more precisely lexical items), which is generally defined by the syntactic or morphological behavior of the lexical item in question. Common linguistic categories include noun and verb, among others.

In DLMs, part-of-speech tag sequences provide the model with a limited amount of syntactic information, e.g. that noun usually follows determiner. Moreover, adding POS tags can be seen a form of smoothing. POS features have been successfully used to train RNNLM [20] and better results could be obtained.

VI. EXPERIMENTS

A. Experimental setup

The ASR system used in this work is based on the conversational telephone speech recognition system [21]. It employs a two-pass search strategy. In the first pass, decoder only uses a trigram LM to generate multiple recognition hypotheses which can be compactly represented in a data structure called word lattice and then n-best list is extracted from *word lattice* [22]. In the second pass, a discriminative language model is used to re-rank the n-best list and then output a new hypothesis whose WER is minimal as the final recognized result. The SRILM tools [23] were used to obtain the statistical back-off N-gram language models. In addition to the transcriptions of the acoustic training data, generic web data collected by ourselves was also used. The transcriptions data contains about 13M words and the generic web data contains about 2 billion words in total. All All N-gram LMs are translated into the binary files and loaded using KenLM [24].

The training set of DLM contained 34,018 utterances and a Mandarin Chinese spontaneous conversational speech corpus containing 3200 utterances was used as a development set to optimize DLM. Moreover, a Mandarin Chinese spontaneous containing 2476 utterances was used to evaluate the discriminative language modeling.

B. Experimental results

We used the DLM to re-score the 1000-best list. Firstly DLM trained by only n-gram features was evaluated and then POS tags were added. The performance of DLM was shown in TABLE I. It can be seen that DLM with only n-gram features was helpful and gave 1% absolute reduction in word error rate (WER) (from 49.6% to 48.6%). Combining n-gram features with POS tags, DLM could obtain another 0.4% absolute reduction in WER (from 48.6% to 48.2%). Finally, the DLM trained by both n-gram features and POS tags gave 1.4% absolute reduction in WER.

TABLE I. THE PERFORMANCE OF DLM IN N-BEST LIST RESCORING

	WER(%)
baseline	49.6
Perceptron	48.6
Perceptron+POS	48.2

VII. CONCLUSION

In this study, we present our initial efforts in applying Discriminative Language Model (DLM) in the n-best list re-scoring phase to improve the performance of our Mandarin speech recognition system. Experimental results showed

that DLM was very effective and compensated to the n-gram LM. In particular, the application of part-of-speech features as well as the conventional n-gram features was helpful to reduce WER.

In the future work, we attempted to explore other more advanced language features such as relevance Information. Moreover, we planned to compare various methods of training discriminative language model in order to further improve the system performance.

ACKNOWLEDGMENT

This work is partially supported by the National Natural Science Foundation of China (Nos. 10925419, 90920302, 61072124,11074275,11161140319) and the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant Nos. X-DA06030100, XDA06030500).

REFERENCE

[1] S. J. Young, N. Russell, and J. Thornton, "Token passing: a simple conceptual model for connected speech recognition systems," *Cambridge University Engineering Department*, pp. 1-23, 1989.

[2] S. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," 1996, pp. 310-318.

[3] B. Roark, M. Saraclar, and M. Collins, "Discriminative< i> n</i>-gram language modeling," *Computer Speech & Language*, vol. 21, pp. 373-392, 2007.

[4] E. Arısoy, M. Saraçlar, B. Roark, and I. Shafran, "Discriminative language modeling with linguistic and statistically derived features," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, pp. 540-550, 2012.

[5] D. Povey, "Discriminative training for large vocabulary speech recognition," Ph. D. thesis, Cambridge University, 2003.

[6] J. T. Huang, X. Li, and A. Acero, "Discriminative training methods for language models using conditional entropy criteria," 2010, pp. 5182-5185.

[7] H. K. J. Kuo, E. Fosler-Lussier, H. Jiang, and C. H. Lee, "Discriminative training of language models for speech recognition," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, 2002, pp. I-325-I-328.

[8] B. H. Juang, W. Hou, and C. H. Lee, "Minimum classification error rate methods for speech recognition," *Speech and Audio Processing, IEEE Transactions on*, vol. 5, pp. 257-265, 1997.

[9] M. Collins, "Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms," in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, 2002, pp. 1-8.

[10] Z. Zhou, J. Gao, F. K. Soong, and H. Meng, "A comparative study of discriminative methods for reranking LVCSR n-best hypotheses in domain adaptation and generalization," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, 2006, pp. I-I.

[11] A. Celebi, H. Sak, E. Dikici, M. Saraclar, M. Lehr, E. Prud'hommeaux, *et al.*, "Semi-supervised discriminative language modeling for Turkish ASR," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 2012, pp. 5025-5028.

[12] K. Sagae, M. Lehr, E. Prud'hommeaux, P. Xu, N. Glenn, D. Karakos, *et al.*, "Hallucinated n-best lists for discriminative language modeling," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 2012, pp. 5001-5004.

[13] Z. Li, Z. Wang, S. Khudanpur, and J. Eisner, "Unsupervised discriminative language model training for machine translation using simulated confusion sets," in *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, 2010, pp. 656-664.

[14] S. Cai, Z. Zhang, T. Li, J. Pan, and Y. Yan, "Development of a Chinese song name recognition system," in *Natural Computation (ICNC), 2011 Seventh International Conference on*, 2011, pp. 941-945.

[15] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, pp. 257-286, 1989.

[16] B. Roark, M. Saraclar, and M. Collins, "Corrective language modeling for large vocabulary ASR with the perceptron algorithm," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, 2004, pp. I-749-52 vol. 1.

[17] H. P. Zhang, H. K. Yu, D. Y. Xiong, and Q. Liu, "HHMM-based Chinese lexical analyzer ICTCLAS," in *Proceedings of the second SIGHAN workshop on Chinese language processing-Volume 17*, 2003, pp. 184-187.

[18] B. Roark, M. Saraclar, M. Collins, and M. Johnson, "Discriminative language modeling with conditional random fields and the perceptron algorithm," in *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, 2004, p. 47.

[19] E. Brill, "A simple rule-based part of speech tagger," in *Proceedings of the workshop on Speech and Natural Language*, 1992, pp. 112-116.

[20] P. W. Yangyang Shi, Catholijn, M. Jonker, "Towards Recurrent Neural Networks Language Models with Liguistic and Contextual Features," *interspeech*, 2012.

[21] J. Shao, T. Li, Q. Zhang, Q. Zhao, and Y. Yan, "A one-Pass real-time decoder using memory-efficient state network," *IEICE TRANSACTIONS on Information and Systems*, vol. 91, pp. 529-537, 2008.

[22] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus in speech recognition: word error minimization and other applications of confusion networks," *Computer Speech & Language*, vol. 14, pp. 373-400, 2000.

[23] A. Stolcke, "SRILM-an extensible language modeling toolkit," 2002.

[24] K. Heafield, "KenLM: Faster and smaller language model queries," 2011.