International Journal of Hybrid Information Technology
Vol.7, No.2 (2014), pp.293-308
http://dx.doi.org/10.14257/ijhit.2014.7.2.26

N-Best Re-scoring Approaches for Mandarin Speech Recognition

Xinxin Li, Xuan Wang and Jian Guan

*Computer Application Research Center,
Harbin Institute of Technology Shenzhen Graduate School
Shenzhen, China, 518055*

{lixxin2, guanjian2000}@gmail.com, wangxuan@insun.hit.edu.cn

Abstract

The predominant language model for speech recognition is n-gram language model, which is locally learned and usually lacks global linguistic information such as long-distance syntactic constraints. We first explore two n-best re-scoring approaches for Mandarin speech recognition to overcome this problem. The first approach is linear re-scoring that can combine several language models from various perspectives. The weights of these models are optimized using minimum error rate learning method. Discriminative approach can also be used for re-scoring with rich syntactic features. To overcome the speech text insufficiency problem for discriminative model, we propose a domain adaptation method that trains the model using Chinese pinyin-to-character conversion dataset. Then we present a cascaded approach to combine the two re-scoring models in pipeline that takes the probability output of linear re-scoring model as the initial weight of the discriminative model. Experimental results show that both re-scoring approaches outperform the baseline system, and the cascaded approach achieves the best performance.

Keywords: *Mandarin speech recognition, re-scoring approaches, cascaded approach, domain adaptation method*

1. Introduction

Automatic speech recognition systems usually require two basic models to obtain the optimal sentence for an audio input: acoustic model and language model [1]. Acoustic model is used to recognize audio fragments into syllables, combining Hidden Markov Models (HMMs) which process the temporal change of speech and Gaussian Mixture Models (GMMs) that predict the probability of each state of HMM using the coefficients of a speech frame, usually Mel-frequency cepstral coefficients (MFCCs) [2]. To overcome the inefficient representation of GMMs, deep neural networks adopt multiple layers of nonlinear hidden units and a output layer to fit the HMM states [3, 4].

N-gram language model (LM) is the most common language model for speech recognition and machine translation [1, 5]. It can be easily trained and achieves moderate performance for most tasks. Nevertheless, it can't capture complex linguistic information such as nest structures of noun phrases and long-distance syntactic constraints due to its simple sibling limitation. Plenty of approaches have been exploited attempting to overcome the problem in decades. N-gram language model can be improved by directly integrated with higher level models, such as dependency relations and probabilistic top-down parsing [6, 7]. Discriminative methods such as maximum entropy models can also be used to utilize global features from word sequences and syntactic structures [8]. Discriminative re-scoring methods

Copyright ⓒ 2014 SERSC

utilize various information from different sources, including word sequences, part-of-speech tags, syntactic structures [9].

In this paper, we first explore two n-best re-scoring approaches for Mandarin speech recognition. Both re-scoring methods are used to choose the optimal word sequence from n-best lists. Linear re-scoring approach can combine multiple language models from different perspectives through a linear function. These sub-models include character models, pinyin-related models, part-of-speech models, dependency model. The discriminative re-scoring approach utilizes rich global features from dependency structures instead of context-free features in previous work [9]. However, training text for acoustic model might be insufficient and inappropriate for the discriminative model. We introduce a domain adaptation method that trains the discriminative model from Chinese pinyin-to-character conversion (PTC) dataset. The PTC corpus is adequate because we can generate the data automatically from raw text. Both re-scoring approaches are evaluated on Chinese 863 speech recognition corpus. Then, a cascaded approach is proposed to combine both models, which takes the probability output of linear re-scoring model as the initial weight of discriminative model. Experimental results show that both re-scoring approaches outperform the baseline system, and the cascaded approach can further improve the performance.

2. Background

Language model (LM) is one of the most important components for natural language processing tasks, such speech recognition, machine translation, *etc.* [10]. For speech recognition problem, language model is combined with acoustic model to determine the optimal word/character sequence for an audio input. N-gram language model is the most common LM, in which the determination of one word/character is only dependent on its probability on previous n words/characters. However there are some linguistic phenomena can't be represented in word n-gram LM, such as long-distance constraints and nested structures. Techniques for improving n-gram language model have been exploited more than a decade.

2.1. Enhanced language models

Siu proposed a variable n-gram LM by representing word n-grams with trees [11]. In their tree model, node merging and combination operations were used to increase the length of n-gram LM. Therefore, the variable n-gram LM could represent high-order word relations than traditional one. Class-based model can also bring long-distance constraints and high-order word relations, because it clusters words/phrases into different classes, and the number of classes is much smaller than words [12].

Word n-gram model can also be improved by directly integrating high-level models into it. Ney proposed a model utilizing the information extracted from dependency relations between long-distance words [6]. His model is capable to integrate pair-wise word associations into word n-gram LM. The work of Chelba and Jelinek introduced a shift-reduce parser into n-gram LM [13]. Their model predicted each word using corresponding parse candidate, where the probability of each word is conditioned on previous words provided by the parse. Roark introduced a model that incorporated probabilistic top-down parsing with word n-gram LM, which can efficiently utilize syntactic information from context-free structures [7]. It calculated the probability of word sequence from the parse probability by summing the probabilities of all derivations in the beam. Wang proposed a almost-parsing language model that integrated lexical features and syntactic constraints based on constraint dependency grammar [14]. Rastrow proposed a hierarchical interpolation method for statistical language

Copyright ⓒ 2014 SERSC

models based on a shift-reduce incremental dependency parser [15]. Part-of-speech tagging model and character language model can be combined as a joint language model for speech recognition [16, 17].

Discriminative language model is another model that can extract arbitrary features from sentences and their syntactic structures. Different form generative models requiring the joint probability of the observations and the labels, it calculates the conditional probability of labels over the observations directly. Rosenfeld proposed a maximum entropy based language model that employed shallow syntactic features [8]. Magdin used Maximum Mutual Information Estimation (MMIE) method to optimize an objective function that involved a metric between correct transcriptions and their competing hypotheses, which are encoded as word graphs generated by Viterbi decoding process [18].

2.2. N-best Re-scoring approaches

Enhanced language model can directly incorporate syntactic information into traditional LMs. Nevertheless, there are still rich linguistic features difficult to be integrated into a single model. Re-scoring approaches provide an alternative and convenient way to use both typical n-gram features and arbitrary features from word sequences and corresponding syntactic structures. Instead of finding the viterbi path in the first phase, it produces n-best candidate lists, a lattice or a confusion network for further improvement.

Collins described a discriminative language model for speech recognition [9]. Averaged perceptron model was used for reranking the n-best lists. Their model can efficiently make use of words, part-of-speech tags and syntactic structures as features. Roark contrasted two parameter estimation methods for discriminative syntactic language modeling: the perceptron algorithm and a method based on maximizing the regularized conditional log-likelihood [19]. They used deterministic weighted finite state automata to encode the word lattice generated in the first phase. Confusion network dynamic programming method can be used for re-scoring on speech recognition by integrating multiple and complex knowledge sources into the model [20]. A Neural probabilistic language model can also be used for N-best re-scoring which includes syntactic features such as head words and their non-terminal labels [21]. Rastrow presented a discriminative hill climbing approach that performed efficient discriminative training procedure for long-span LMs [22]. Arisoy presented a discriminative language model that integrated linguistically motivated and statistically derived information such as semantic information and evaluated their system on Turkish [23].

3. Baseline Speech Recognition System

The baseline speech recognition consists of two basic components: acoustic model $P_{AM}(S\,|\,W)$ that represents the probability of words W on given audio trivial A, and language model $P_{LM}(W)$ that denotes the probability of word sequence W. For a given audio input A, the optimal word sequence is determined by

$$\begin{aligned} W^* &= argmax_W P_{baseline}(A,W) \\ &= argmax_W (\alpha \log P_{AM}(A\,|\,W) + \beta \log P_{LM}(W)) \end{aligned} \tag{1}$$

where α and β denote the weights of acoustic model and language model separately, and satisfy the constraint $\alpha+\beta=1$. The weights can be tuned on development data. For the acoustic model, we use a traditional GMM-HMM model as the baseline model in this paper. The optimal word sequence for an audio input is obtained by using a viterbi search algorithm in decoding phase.

3.1. Word n-gram Language Model

A language model is used to calculate a probability distribution $P(W)$ over a string W that represents how frequent the string W occurs as a sentence. The most common language model is word n-gram language model, which introduces a Markov assumption, meaning that each word depends only on its previous $(n-1)$ words. It can be denoted as

$$P(W) \quad = p(w_1)p(w_2 \mid w_1)p(w_3 \mid w_1 w_2)...p(w_l \mid w_1...w_{l-1}) = \prod_{i=1}^{l} p(w_i \mid w_1,...,w_{i-1}). \quad (2)$$

For speech recognition task, the most widely used n-gram LM is trigram, where each word is conditioned on its previous two words, denoted as

$$P(W) = p(w_1)p(w_2 \mid w_1)\prod_{i=3}^{m} p(w_i \mid w_{i-2}w_{i-1}). \qquad (3)$$

To estimate $p(w_i \mid w_{i-2}, w_{i-1})$, the frequencies with $w_{i-2}w_{i-1}$ and $w_{i-2}w_{i-1}w_i$ in training data are need to calculated. To deal with the insufficient training data problem, various smoothing methods can be adopted [10].

4. Re-scoring Approaches

In this section, we first explore two n-best re-scoring approaches for Mandarin speech recognition to overcome the problem of n-gram language model: linear re-scoring and discriminative re-scoring approaches. N-best word sequences for re-scoring are generated by baseline speech recognizer using a beam search algorithm. Then we present a cascaded approach to combine these two approaches in pipeline that takes the output probability of linear re-scoring approach as the initial weight of discriminative re-scoring approach.

4.1. Linear re-scoring approach

Previous work reveal that single word n-gram LM is insufficient for speech recognition task. A linear re-scoring approach can be used to combine different linguistic information, and choose the best one from candidate lists. The n-best word sequence lists for each audio input are first generated by baseline speech recognizer. In this paper, several sub models are introduced into our approach, including pinyin-word co-occurrence models, character-based model, part-of-speech tagging model and dependency model (detailed description in next section). To better utilize these various sub models, we use a linear function to combine their probabilities. The weights of sub models are optimized using minimum error training (MERT) method. MERT method is first proposed for machine translation problem [24], and has been successfully applied to other natural language processing tasks, such as joint word segmentation and POS tagging [25]. The probability of a candidate word sequence W for an audio input A is calculated as

$$\begin{aligned}
W^* &= argmax_{W\in GEN(A)} P_{linear}(W \mid A) \\
&= argmax_{W\in GEN(A)} \sum_{i=0}^{k} w_i * P_{sub}(S,W,C,T,D \mid A)) \\
&= argmax_{W\in GEN(A)} (w_0 * P_{baseline}(A,W) \\
&+ w_1 * P(W) + w_2 * P(C) + w_3 * P(S \mid C) \\
&+ w_4 * P_{occur}(W \mid S) + w_5 * P_{occur}(S \mid W) \\
&+ w_6 * P(T \mid W) + w_7 * P(T) + w_8 * P_{occur}(W \mid T) \\
&+ w_9 * P(D \mid W,T))
\end{aligned} \tag{4}$$

where A,S,W,C,T,D denote audio input, pinyin sequence, word sequence, character sequence, POS tag sequence, and dependency tree separately. In the re-scoring formula, $GEN(S)$ denotes all the word sequence candidates generated from audio input A, $P_{baseline}(A,W)$ represents the output probability of word sequence W on audio input A generated by baseline speech recognizer. The weights of sub models satisfy $\sum_{i=0}^{9} w_i = 1$. The probability outputs of these sub models are represented as $P_{sub}(S,W,C,T,D \mid A)$, including another word n-gram language model $P(W)$, a character n-gram language model $P(C)$, a character-based model $P(S \mid C)$, the pinyin-word co-occurrence models $P_{occur}(W \mid S)$, $P_{occur}(S \mid W)$, a part-of-speech tagging model $P(T \mid W)$, a POS N-gram language model $P(T)$, a POS-word occurrence model $P_{occur}(W \mid T)$, and a dependency model $P(D \mid W,T)$. The fact that the function $P_{linear}(W \mid A)$ equals $\sum_{i=0}^{k} w_i * P_{sub}(S,W,C,T,D \mid A)$ is tenable because a word sequence W corresponds to only one syllable sequence S, one character sequence C, one POS sequence T and one dependency tree D.

The weights $w_j (1 \le j \le k)$ for these sub models can be obtained by iteratively calculating each weight w_j with other weights fixed. Thus, the probability of each candidate is defined as

$$P_{mert}(W \mid S) = w_j \times P_j(W \mid S) + \sum_{i \ne j} w_i \times P_i(W \mid S). \tag{5}$$

The left probability $w_j \times P_j(W \mid S)$ is a variable where w_j is the weight need to be optimized, and the right probability $\sum_{i \ne j} w_i \times P_i(W \mid S)$ is fixed. It's expensive to use a direct grid search algorithm for determining each weight because re-calculating the probabilities of all candidates to find the best one is time-consuming. MERT method uses a piece-wise linear search algorithm for each j^{th} dimension. Since the optimal value of each weight w_j must be in the intersections of all lines drawn by the above formula. The optimal parameters are easy to obtain because only a few candidates need to be re-calculated when the critical value changes. The principle of MERT method has been described in Och [24, 26]. The sub models we employed and the method of calculating the probabilities of the candidates for each sub model are described in next section.

4.2. Domain adaptation discriminative re-scoring approach

Since the dependency structures of word sequence candidates are constructed, we can utilize various features extracted, including words, POS tags and syntactic information. Our discriminative re-scoring approach is built on the work of Collins's discriminative language model. Different from their context-free rule features, we extract linguistic features from dependency tree D. For an audio input A, the optimal word sequence W is chosen as

$$\begin{aligned} W^* &= argmax_{W \in GEN(A)} P_{baseline}(A,W) + P_{discriminative}(W) \\ &= argmax_{W \in GEN(A)} P_{baseline}(A,W) + \Phi(W,A) \times \bar{\alpha} \end{aligned} \tag{6}$$

The second probability $P_{discriminative}(W)$ denotes the probability produced by the discriminative re-scoring approach. Figure 1 shows an example dependency tree generated by our POS tagger and dependency parser. There are two feature sets extracted for our discriminative re-scoring model. One set contains flat features of the dependency tree, mainly the word and POS information. The second feature set we employed contains dependency relations extracted from the dependency tree, including the relations of father-son, father-son-sibling, and grandfather-father-son dependency. In this paper, we only consider the unlabeled dependency tree, and label information between dependency nodes is omitted. These detailed features are described in next section.

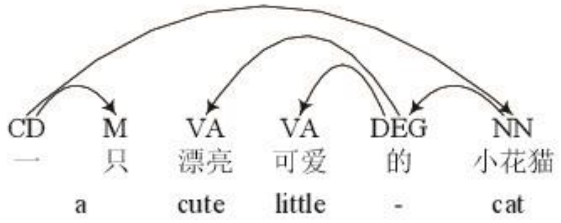

Figure 1. An example of a dependency tree

The discriminative re-scoring approach is trained using averaged perceptron algorithm. For n best dependency trees $D_k(i=1,...,n)$ generated from audio input A, the parameter $\bar{\alpha}$ is updated as $\bar{\alpha} = \bar{\alpha} + \Phi(\bar{\alpha}, D_m) - \Phi(\bar{\alpha}, D_p)$, where D_m is the dependency tree with minimum character error rate (CER), denoted as the golden tree. D_p is dependency tree D_k with greatest probability generated by the re-scoring discriminative model, as the predicted tree. $\Phi()$ represents the feature templates.

However, it is expensive to obtain large amount of annotated texts for training the discriminative model. To solve the data insufficiency problem, we introduce a domain adaptation method, which trains the discriminative model on Chinese pinyin-to-character (PTC) conversion data. The training data for Chinese PTC conversion is easy to obtain because the method that convert text sequences to pinyin sequences is very accurate, achieving 98.5%. One problem for domain adaptation method is that the

Copyright ⓒ 2014 SERSC

initial weights for both tasks are different, as $P_{baseline}(A,W)$ for speech recognition, and P_{ptc} we denoted for PTC conversion problem. We use a ratio balanced method to adapt the PTC model to speech recognition. The initial weight for Chinese PTC conversion is adjusted as

$$P_{ptc'} = \frac{P_{ptc} - P_{ptc}^{min}}{P_{ptc}^{max} - P_{ptc}^{min}} * (P_{baseline}^{max} - P_{baseline}^{min}) + P_{ptc}^{min} \tag{7}$$

where $P_{ptc}, P_{ptc'}$ represent the original and adjusted initial probabilities of Chinese PTC problem. $P_{ptc}^{max}, P_{ptc}^{min}, P_{baseline}^{max}, P_{baseline}^{min}$ denote the maximal, minimal initial weights of discriminative models for Chinese PTC conversion and speech recognition problems separately. After adjustment, the scopes of initial weights for both models become the same. Then the discriminative models trained on Chinese PTC data are used for speech recognition. The discriminative model for Chinese PTC conversion is trained similar as speech recognition, and with same feature sets.

4.3. Cascaded re-scoring approach

Both linear and discriminative re-scoring approaches utilize linguistic information from syntactic structures, but from different views. They can be cascaded in a pipeline. In previous discriminative approach, the initial probability is set to the probability output $P_{baseline}(A,W)$ produced by baseline recognizer. To utilize the linear re-scoring approach, our cascaded approach sets the probability output of the linear approach $P_{linear}(A,W)$ as the initial probability of the discriminative approach. Domain adaptation method is also used in cascaded re-scoring. Thus the word sequence W for new re-scoring approach is calculated as

$$\begin{aligned} W* &= argmax_{W \in GEN(A)} P_{cascaded}(A,W) \\ &= argmax_{W \in GEN(A)} (P_{linear}(A,W) + P_{discriminative}(W)) \end{aligned} \tag{8}$$

5. Sub Models

Our re-scoring approaches can efficiently incorporate long-distance syntactic constraints. In this section, we will describe part-of-speech tagging model and dependency parsing model used for both re-scoring approaches. Also, all sub models for the linear re-scoring approach, including the pinyin-related models, POS-related models are presented.

5.1. Character N-gram Language Model

Character n-gram language model can also be integrated with word n-gram language model to improve its performance [17]. Therefore, our linear re-scoring method can include the character n-gram LM, denoted as $P(C) = p(c_1)p(c_2 | c_1)p(c_3 | c_2c_1)\prod_{i=4}^{n} p(c_i | c_{i-3}c_{i-2}c_{i-1})$.

We use four-gram character LM, where the statistic and smoothing method for $p(c_i | c_{i-3}c_{i-2}c_{i-1})$ is calculated similar as $p(w_i | w_{i-2}w_{i-1})$.

Copyright ⓒ 2014 SERSC

5.2. Character-Based Discriminative Model

The character-based model is a discriminative model trained by averaged perceptron algorithm, which recognizes characters/words in a sentence only using the pinyin information. The pinyin sequence is generated by the speech recognizer. The parameter estimation method for averaged perceptron and the decoding algorithm for best sequence can be found in work [27, 28]. Pinyin feature templates contain $c_n(n=-2..2), c_nc_{n+1}(n=-1..0), c_{-1}c_1$. The pinyin sequences are generated by baseline speech recognizer, corresponding to the word sequences.

5.3. Pinyin-Word Co-occurrence Model

For an input audio, its generated pinyin sequence S and word sequence W, we define the pinyin-word co-occurrence as

$$P_{occur}(W \mid S)=\prod_{i=1}^{m} p(w_i \mid s_i), P_{occur}(S \mid W)=\prod_{i=1}^{m} p(s_i \mid w_i). \quad (9)$$

where $p(w_i \mid s_i)=N(w_i,s_i)/N(s_i)$, $p(s_i \mid w_i)=N(s_i,w_i)/N(w_i)$. The number of pinyins $N(s_i)$ and the number of pinyin-word pairs $N(w_i,s_i)$ can be easily obtained.

5.4. POS Models

POS information can be used to improve the performance for many natural language processing tasks, such as word segmentation and named entity recognition [29]. In this paper, we introduce POS tagging model for speech recognition. The features for POS tagging model are choose similar as [27, 30], shown in Table 1. The dataset for POS tagging model is taken from Chinese Treebank 5.0, and the distribution of training, development and test dataset is the same as [30]. Our POS tagging model achieves 95.3% F-1 value on development dataset.

Table 1. Feature templates for Chinese POS tagging model

1	$w_{-2}t_0$	$end(w_{-1})w_0start(w_1)t_0$
2	$w_{-1}t_0$	when $len(w_0)=1$
3	w_0t_0	$start(w_0)t_0$
4	w_1t_0	$end(w_0)t_0$
5	w_2t_0	$c_nt_0,(n=1,len(w_0)-2)$
6	$t_{-1}t_0$	$start(w_0)c_nt_0,(n=above)$
7	$t_{-2}t_{-1}t_0$	$end(w_0)c_nt_0,(n=above)$
8	$t_{-1}w_0$	$c_nc_{n+1}t_0,(c_n=c_{n+1})$
9	$w_0t_0end(w_{-1})$	$class(start(w_0))t_0$
10	$w_0t_0start(w_1)$	$class(end(w_0))t_0$

Besides, we also use the POS n-gram language model and POS-word co-occurrence models for our re-scoring models. The probability of a POS sequence T is defined as $P(T)=\prod_{i=1}^{n} p(t_i \mid t_1,...,t_{i-1})$. For a word sequence w and its corresponding POS sequence T, the POS-word co-occurrence is defined as

$$P_{occur}(W \mid T)=\prod_{i=1}^{m} p(w_i \mid t_i), P_{occur}(T \mid W)=\prod_{i=1}^{m} p(t_i \mid w_i), \tag{10}$$

where $p(w_i \mid t_i)$ and $p(t_i \mid w_i)$ are calculated using maximum likelihood estimation method.

5.5. Dependency model

The determination of a word sequence for an audio input might not be determined only using the flat information, such as nearby words and POS tags. The grammatical and syntactic information can be beneficial. Dependency model can bring long-distance word relations for Mandarin speech recognition. Figure 1 shows an example of a dependency tree. A deterministic transition-based algorithm can be used to determine the best dependency parse [31]. For a Chinese word and POS sentence (W,T), the probability of dependency tree D is calculated as $P(D \mid W,T)=\prod_{i} w_i \times f_i(D \mid W,T)$.

6. Experiments

Our experimental data are selected from Mandarin standard corpus of Chinese National 863 Project, recorded by 83 males and 83 females. There are totally 1560 text sentences chosen from 1993 and 1994 People's Diary for recording. These sentences are split into 3 parts, named A, B and C, of which each person only record one part. The text and person distribution of Mandarin 863 speech corpus are listed in Table 2, where parts A, B, C contain 1-521, 522-1040, and 1041-1560 sentences separately. F00 denotes the speeches recorded by female speaker 00. Thus F01, F06, F00 contain totally entire 1560 sentences. All the data for our experiments are chosen from female speakers. We will omit F in the following description.

Table 2. The dataset setting of 863 Mandarin speech corpus

parts	#sentences	speeches
A	521	F01,F02,F03,F04,F05,...
B	519	F06,F07,F08,F09,F10,...
C	520	F00,F11,F12,F13,F14,...

In this section, we perform two sets of experiments for Mandarin speech recognition. The distribution of training, development and test data is shown in Table 3. The first experiment set is trained on 4680 speech sentences, tuned and tested on 1560 speech sentences. In this setting, the training, development and test data contain same texts. The second experiment is different, where the training data contains two text parts, and the development data contains the third part.

Table 3. The distribution of training, development and test data

Group	Training data	Dev data	Test data
1	02,07,11,03,08,12,04,09,13	01,06,00	
2	01,06, 02,07, 03,08	00	05,10,14
3	01,00, 02,11, 03,12	06	
4	06,00, 07,11, 08,12	01	

The performances of speech recognition systems, including baseline system and re-scoring approaches, are evaluated using character error rate (CER) measure. The CER measure is defined as

$$CER = \frac{\#gold - \#insertion - \#deletion - \#substitution}{\#gold}. \qquad (11)$$

In the formula, $\#gold$ denotes the number of characters in golden sentences, $\#insertion$, $\#deletion$ and $\#substitution$ are separately the numbers of operations of insertion, deletion, and substitution that convert the golden sentences to predicted sentences.

6.1. Baseline System

We first describe our baseline speech recognition system. The acoustic model is trained on Mandarin standard corpus of Chinese National 863 Project, and RASC863 (annotated 10 regional accent speech corpus). We constructed a tri-phone acoustic model using HTK. The language model is trained on newswire of People's Daily and GuangMing Daily using SRILM Toolkit [32]. Finally, a forward word bigram language model and backward trigram language model are used for a two-pass decoding using Julius [33]. Table 4 shows the experimental results of our baseline system on development data and test data. From the results, we can observe that for the performance for different speeches are different, and even the corresponding texts are the same, the performance is various for different speakers.

Table 4. Results of development and test data using baselines system

Dataset	01,06,00	05,10,14	01	06	00
CER(%)	21.18	31.11	25.76	22.35	15.50

The n-best lists for each audio input are generated by the baseline system using a beam search algorithm. Figure 2 reveals the oracle of different n-best lists. It's obvious that with the increase of beam number n, the oracle of CER become smaller. The oracle decreases faster at the beginning when n increase from 1 to 5, and become slower when n become larger. Finally, we choose 500 n-best candidate lists for our re-scoring approaches.

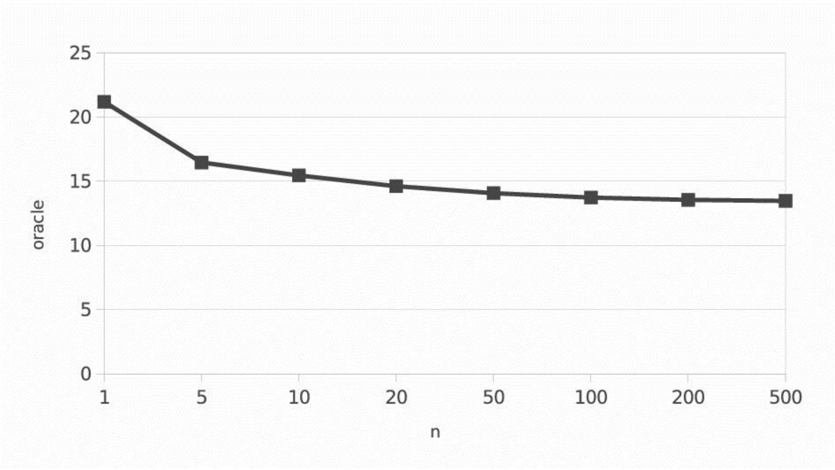

Figure 2. Oracle of n-best lists on development data 01,06,00

6.2. Linear re-scoring approach

For our linear re-scoring approach, we use a backward greedy search algorithm to determine the suitable set of sub models, and train the weights for these models. It first begins with a set containing all sub models, and iteratively attempts to remove each sub model and evaluate its performance on development data. Then the linear model will remove the sub model which achieves the best performance. The iterative procedure will stop if there is no performance improvement.

We first perform experiments on group 1, trained on 02,07,11,03,08,12,04,09,13, and evaluated the linear re-scoring approach on 01,06,00. After model selection strategy, we remove the word-pinyin occurrence model $P_{occur}(P|W)$ because it decreases the performance. The results shown in Table 5 compared the linear re-scoring model of all other sub models and the models with removing one sub model each time. It shows that linear re-scoring approach outperforms the baseline system about 1.11% decrease on CER, and all sub models are beneficial for the task, in which the character n-gram LM improves the performance most, and the word N-gram LM ranks second. Specially, the dependency model also makes an improvement for the linear re-scoring approach, outperforming character-based model, pinyin-word occurrence model and POS-word occurrence model.

Table 5. Results of linear re-scoring approach on development data

Seq	Sub Models	CER(%)
A	All	20.07,...
A/1	All - Word N-gram LM	20.62
A/2	All - Character N-gram LM	20.73
A/3	All - Character-based Model	20.25
A/4	All - Pinyin-Word Occurrence	20.22
A/5	All - POS Model	20.31
A/6	All - Word-POS Occurrence	20.38
A/7	All - POS-Word Occurrence	20.25
A/8	All - POS N-gram LM	20.29
A/9	All - Dependency Model	20.32

We then compare linear re-scoring approach with baseline system on experimental group 2,3,4, where the training and development data are totally different. Sub models are chosen the same as the first experiment. The results are shown in Table 6. The improvement for experimental group 2,3,4 is consistent with group 1. The three experiments show that the linear re-scoring approach also outperforms the baseline system when the training data and development data are different. The linear-scoring approach requires smaller training data and more generalized than discriminative re-scoring approach, which will be exhibited on next subsection.

Table 6. Results of linear re-scoring model on experimental group 2,3,4

Group	Dev data	CER(%)	CER decrease(%)
2	01	24.37	1.39
3	06	21.45	0.9
4	00	14.84	0.66

6.3. Domain adaptation discriminative re-scoring approach

Given n-best candidate dependency trees for each audio input, we can employ discriminative re-scoring approach to select the optimal one. Two type of feature sets are used in our reranking model, listed in Table 7. The first set contains only flat feature sets (W, T1, T2, T3), composed by word and POS information. w_0 and t_0 denote the word form and POS tag of current token separately, $-i, +i$ in $w_{\pm i}$ represents the i_{th} token before or after current token.The second type of feature sets are dependency relations extracted from dependency tree. In table 7, feature set D1 contains dependency features of father-son relation, D2 grandfather-father-son relation, and D3 father-son-sibling relation. In these features, G, F, S, L, R represent grandfather, father, son, left son, right son in a dependency relation separately. $F_t S_w$ denotes the combination of POS tag of father node and word of son node in a dependency relation. D_{FS} represents the direction between father node and son node in a dependency relation, which is either 1 or 0, meaning left and right direction separately. Whether the father node and son node are adjacent is denoted as A_{FS} .

Table 7. Features for discriminative re-scoring approach

W	$w_{-2}w_{-1}w_0, w_{-1}w_0, w_0$
T1	$t_{-2}t_{-1}t_0, t_{-1}t_0, t_0, t_0 w_0$
T2	$t_{-1}t_0 w_0, w_{-1}t_{-1}t_0 w_0$, $t_{-2}w_{-i}t_{-1}t_0 w_0, w_{-2}t_{-2}w_{-1}t_{-1}t_0 w_0$
D1	$F_t D_{FS} A_{FS} S_t, F_w D_{FS} A_{FS} S_w S_t$, $F_w F_t D_{FS} A_{FS} S_t, F_w F_t D_{FS} A_{FS} S_w S_t$
D2	$G_t D_{GF} A_{GF} F_t D_{FS} A_{FS} S_t$, $G_t D_{GF} A_{GF} F_t D_{FS} A_{FS} S_w S_t$ $G_t D_{GF} A_{GF} F_w F_t D_{FS} A_{FS} S_t$, $G_w G_t D_{GF} A_{GF} F_t D_{FS} A_{FS} S_t$ $G_t D_{GF} A_{GF} F_w F_t D_{FS} A_{FS} S_w S_t$, $G_w G_t D_{GF} A_{GF} F_w F_t D_{FS} A_{FS} S_t$ $G_w G_t D_{GF} A_{GF} F_w F_t D_{FS} A_{FS} S_w S_t$
D3	$L_t D_{FL} A_{FL} R_t D_{FR} A_{FR} F_t$, $L_w L_t D_{FL} A_{FL} R_w R_t D_{FR} A_{FR} F_t$ $L_t D_{FL} A_{FL} R_t D_{FR} A_{FR} F_w F_t$, $L_w L_t D_{FL} A_{FL} R_w R_t D_{FR} A_{FR} F_w F_t$

Figure 3 shows that performance of discriminative re-scoring approach on experimental group 1 and 2. It's shown that with more additional features, group 1 and 2 behave different, where CER of group 1 decreases, but group 2 increases. That might be that group 1 are trained and evaluated using speeches with same 1560 text sentences, and group 2 are different. The reason why our results on group 2 are contradicted with Collins's experiments is that we use smaller training data [9], only 1140 sentences, but Collins use 297580 transcribed utterances. However, it's expensive to obtain the transcribed texts. We proposed a domain adaptation method to train the discriminative model using Chinese pinyin-to-character dataset. The n-best lists for Chinese PTC conversion is generated using an n-gram LM.

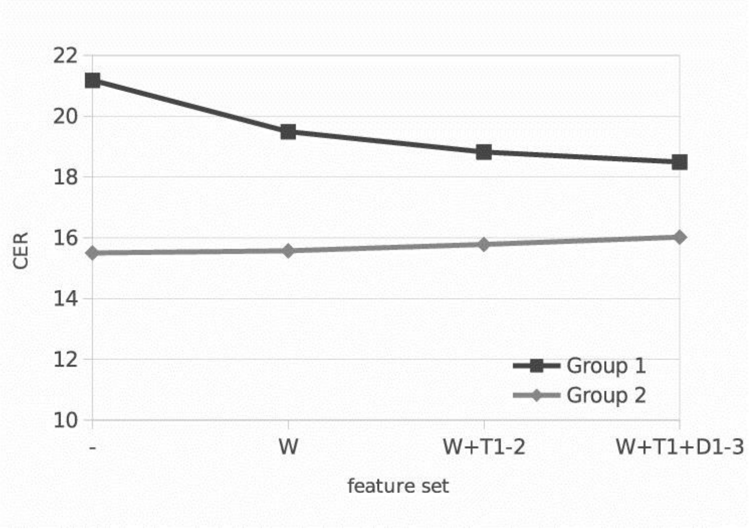

Figure 3. Results of discriminative re-scoring approach

Mandarin speech recognition and Chinese pinyin-to-character conversion perform different, and their output probabilities of baseline systems are quite different. We use the domain adaptation method described in last section to train the model on Chinese pinyin-to-character data. The training data is chosen from The Lancaster Corpus of Mandarin Chinese, containing 100000 sentences. A greedy forward feature selection strategy is used for discriminative model on Chinese PTC conversion problem. Then it's used on speech recognition, and achieves 20.38% CER on development data with feature set $W, D1$, outperforming the baseline system.

6.4. Cascaded re-scoring approach

In discriminative re-scoring approach, the initial weight is set as the output probability of the baseline system. Actually, the linear re-scoring approach can be cascaded with discriminative approach, which takes the output probability of linear re-scoring approach as the initial weight of the discriminative approach. We then perform experiments on development set, and the results are shown in the second row of Table 8. The results show that the cascaded approach can effectively utilize these two re-scoring approaches, and outperforms them both.

Copyright ⓒ 2014 SERSC

Table 8. Results of development and test data with different initial weights

Group	Data	Initial Weights	Feature Sets	CER (%)
1	01,06,00	$P_{baseline}(A,W)$	-	21.18
		$P_{baseline}(A,W)$	W,D1	20.38
		$P_{linear}(A,W)$	W,T1,D1,D3	**19.76**
2	05,10,14	$P_{baseline}(A,W)$	-	31.11
		$P_{linear}(A,W)$	-	29.73
		$P_{baseline}(A,W)$	W,D1	30.39
		$P_{linear}(A,W)$	W,T1,D1,D3	**29.3**

We then evaluated all re-scoring approaches on test data 05,10,14. The experimental results exhibit similar performance as development data. Single linear re-scoring approach outperforms discriminative re-scoring approach, and the cascade approach which uses the probability output of linear model as initial weight achieves the greatest performance, 1.8% absolute CER decrease.

7. Conclusion

In this paper, we explored two re-scoring approach for Mandarin speech recognition to overcome the problem of n-gram language model by integrating rich long-distance syntactic information. Linear re-scoring approach can combine multiple models with different perspectives through a linear function. To overcome the data insufficiency for training discriminative model, we proposed a domain adaptation method to train a model on Chinese pinyin-to-character conversion data. Experimental results show that both linear and discriminative re-scoring approach can effectively improve the performance of baseline system, and a cascaded approach achieves the best. In further research, the re-scoring approaches can be applied to other languages.

References

[1] D. Jurafsky, J. H. Martin, A. Kehler, L. K. Vander and N. Ward, “Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition”, MIT Press, **(2000).**

[2] L. R. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition”, Proceedings of the IEEE, vol. 77, no. 2, **(1989)**, pp. 257-286.

[3] M. Abdel-rahman, G. Dahl and G Hinton, “Deep Belief Networks for phone recognition”, in NIPS Workshop on Deep Learning for Speech Recognition and Related Applications, **(2009)**, November; Whistler, BC, Canada, pp. 1-9.

[4] G. Hinton, L. Deng, D, Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath and B. Kingsbury, “Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups”, IEEE Signal Processing Magazine. vol. 29, no. 6, **(2012)**, pp. 82-97.

[5] P. F. Brown, V. J. Della Pietra, S. A. Della Pietra and R. L. Mercer, “The mathematics of statistical machine translation: Parameter estimation”, Computational linguistics, vol. 19, no. 2, **(1993)**, pp. 263-311.

[6] N. Hermann, E. Ute and K. Reinhard, “On structuring probabilistic dependences in stochastic language modelling”, Computer Speech & Language, vol. 8, no. 1, **(1994)**, pp. 1-38.

[7] B. Roark, “Probabilistic top-down parsing and language modeling”, Computational Linguistics, vol. 27, no. 2, **(2001)**, pp. 249-276.

[8] R. Rosenfeld, S. F. Chen and X. J. Zhu, "Whole-sentence exponential language models: a vehicle for linguistic-statistical integration", Computer Speech & Language, vol. 15, no. 1, **(2001)**, pp. 55-73.

[9] M. Collins, B. Roark and M. Saraclar, "Discriminative Syntactic Language Modeling for Speech Recognition", Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05), **(2005)** June; Ann Arbor, Michigan, pp. 507-514.

[10] S. F. Chen and J. Goodman, "An Empirical Study of Smoothing Techniques for Language Modeling", Computer Science Group, Harvard University, **(1998)**.

[11] S. Manhung and M. Ostendorf, "Variable n-grams and extensions for conversational speech language modeling", IEEE Transactions on Speech and Audio Processing, vol. 8, no. 1, **(2001),** pp. 63-75.

[12] P. F. Brown, P. V. deSouza, R. L. Mercer, V. J. Della Pietra and J. C. Lain, "Class-based n-gram models of natural language", Computational Linguistics. vol. 18, no. 4, **(1992)**, pp. 467-479.

[13] C. Ciprian and J. Frederick, "Structured language modeling", Computer Speech & Language, vol. 14, no. 4, **(2000)**, pp. 283-332.

[14] W. Wang and M. P. Harper, "The SuperARV Language Model: Investigating the Effectiveness of Tightly Integrating Multiple Knowledge Sources", Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, **(2001)** July, pp. 238-247.

[15] A. Rastrow, M. Dredze and S. Khudanpur, "Efficient Structured Language Modeling for Speech Recognition", INTERSPEECH, **(2012)** September 9-13; Portland, Oregon, USA.

[16] D. Filimonov and M. Harper, "A Joint Language Model With Fine-grain Syntactic Tags", Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, **(2009)** August; Singapore, pp. 1114-1123.

[17] X. Y. Liu, J. L. Hieronymus, M. J. Gales and P. C. Woodland, "Syllable language models for Mandarin speech recognition: Exploiting character language models", The Journal of the Acoustical Society of America, vol. 133, no. 1, **(2013)**, pp. 519-528.

[18] M. Vladimir, "Discriminative Training of Language Models for Speech Recognition", York University, **(2009)**.

[19] B. Roark, M. Saraclar and M. Collins, "Discriminative n-gram language modeling", Computer Speech & Language, vol. 21, no. 2, **(2007)**, pp. 373-392.

[20] A. Deoras and F. Jelinek, "Iterative decoding: A novel re-scoring framework for confusion networks", IEEE Automatic Speech Recognition Understanding, **(2009)**, December 13-17; Bolzano, Italy, pp. 282-286.

[21] H. K. Kuo, L. Mangu, A. Emami and I. Zitouni, "Morphological and syntactic features for Arabic speech recognition", IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), **(2010)**, March 14-19; Dallas, Texas, USA, pp. 5190-5193.

[22] A. Rastrow, M. Dredze and S. Khudanpur, "Efficient discriminative training of long-span language models", 2011 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), **(2011)** June, pp. 214-219.

[23] E. Arisoy, M. Saraclar, B. Roark and I. Shafran, "Discriminative Language Modeling With Linguistic and Statistically Derived Features", IEEE Transactions on Audio, Speech, and Language Processing. vol. 20, no. 2, **(2012)**, pp. 540-550.

[24] J. O. Franz, "Minimum Error Rate Training in Statistical Machine Translation", Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, **(2003)** July; Sapporo, Japan, pp. 160-167.

[25] W. B. Jiang, L, Huang, Q. Liu and Y. J. Lv, "A Cascaded Linear Model for Joint Chinese Word Segmentation and Part-of-Speech Tagging", Proceedings of ACL-08: HLT, **(2008)** June; Columbus, Ohio, pp. 897-904.

[26] O. Zaidan, "Z-MERT: A Fully Configurable Open Source Tool for Minimum Error Rate Training of Machine Translation Systems", The Prague Bulletin of Mathematical Linguistics, vol. 91, no. 1, **(2001)**, pp. 79-88.

[27] M. Collins, "Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms", Proceedings of the ACL-02 conference on Empirical methods in natural language processing, **(2002)** July 6-12; Philadelphia, PA, USA, pp. 1-8.

[28] X. X. Li, X. Wang and L. Yao, "Joint decoding for Chinese word segmentation and POS tagging using character-based and word-based discriminative models", 2011 International Conference on Asian Language Processing (IALP), **(2011)**, Penang, Malaysia, pp. 11-14.

[29] H. T. Ng and J. K. Low, "Chinese Part-of-Speech Tagging: One-at-a-Time or All-at-Once? Word-Based or Character-Based?", Proceedings of EMNLP 2004, **(2004)** July; Barcelona, Spain, pp. 277-284.

[30] Y. Zhang and S. Clark, "Joint Word Segmentation and POS Tagging Using a Single Perceptron", Proceedings of ACL-08: HLT, **(2008)** June; Columbus, Ohio, USA, pp. 888-896.

[31] Y. Zhang and J. Nivre, "Transition-based Dependency Parsing with Rich Non-local Features", Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, **(2011)** June; Portland, Oregon, USA, pp. 188-193.

[32] A. Stolcke, "SRILM - An extensible language modeling toolkit", Proceedings of the International Conference on Spoken Language Processing, **(2002)**, Denver, Colorado, pp. 901-904.

[33] L. Akinobu, K. Tatsuya and S. Kiyohiro, "Julius --- An Open Source Real-Time Large Vocabulary Recognition Engine", EUROSPEECH2001: the 7th European Conference on Speech Communication and Technology, **(2010)**, September; Aalborg, Denmark, pp. 1691-1694.

Authors

Xinxin Li

He' born in China, on December 1983. He obtained his Bachelor Degree and Master Degree in 2005 and 2008 separately. Currently, He is a Ph.D. candidate at Harbin Institute of Technology Shenzhen Graduate School. His research interest covers natural language processing, and network information processing.

Xuan Wang

He received M.E.Sc. and D.E. in Harbin Institute of Technology in 1994 and 1997. Currently he is Professor , PhD Supervisor, Dean of Computer Science and Technology Department, Harbin Institute of Technology Shenzhen Graduate School. His research interests covers artificial intelligence, network multimedia, and information processing.

Jian Guan

He is a Ph.D. candidate at Harbin Institute of Technology Shenzhen Graduate School. His research interests covers artificial intelligence and speech recognition.

Copyright ⓒ 2014 SERSC