ICASSP 2008 Survey

Special Topics in Spoken Language Processing

Guan-Yu Menphis Chen

Department of Computer Science & Information Engineering,
National Taiwan Normal University, Taipei, Taiwan
696470203@ntnu.edu.tw

Main reference：

- Yang Liu, et al. “UNSUPERVISED LANGUAGE MODEL ADAPTATION VIA TOPIC MODELING BASED ON NAMED ENTITY HYPOTHESES,” ICASSP 2008.
- Zhengyu Zhou, et al. “RECASTING THE DISCRIMINATIVE N-GRAM MODEL AS A PSEUDO-CONVENTIONAL N-GRAM MODEL FOR LVCSR,” ICASSP 2008.

Outline

1. Unsupervised Language Model Adaptation Via Topic Modeling Based On Named Entity Hypotheses
   - Introduction
   - Training & Testing
   - Experiment
   - Conclusion

2. Recasting The Discriminative n-gram Model As A Pseudo-Conventional n-gram Model For LVCSR
   - Introduction
   - Discriminative N-gram Modeling
   - Recasting Discriminative Model
   - Pseudo-Conventional N-gram Model
   - Other Post-Processing Technique
   - Experiment
   - Conclusion

1-Introduction

- To identify implicit topics from an unlabeled corpus, one simple technique is to group the documents into topic clusters by assigning only one topic label to a document.
- Recently, several other methods in the line of latent semantic analysis have been proposed and used in LM adaptation, such as latent semantic analysis (LSA), probabilistic latent semantic analysis (PLSA), and LDA.
- Most of these existing approaches are based on the “bag of words” model to represent documents, where all the words are treated equally and no relation or association between words is considered.
- Named entities are very common in domains such as broadcast news, and carry valuable information, which we hypothesize is topic indicative and useful for latent topic analysis.

1-Introduction

- For unsupervised LM adaptation, an early attempt is a cache-based model, developed based on assumption that words appearing earlier in a document are likely to appear again.

- The focus of our work is to investigate the role of named entity information for topic modeling and LM adaptation.

- This is different from using all the words or selecting terms for topic analysis as those used in text categorization or information retrieval.

1- Training & Testing

- We evaluate two different topic modeling approaches for LM adaptation, LDA and clustering, both using NE hypotheses. Each document in the training set is labeled with NE hypotheses.

- The purpose of LDA analysis in training is to find the latent topic information for the given document collection. There are two matrix DP and WP, where DP is the document-topic matrix and WP is the word-topic matrix.

- Note that here “word” correspond to the elements used to represent the document (i.e., NEs in our experiments).

1- Training & Testing

- In the DP matrix, an entry c_{ik} represents the counts of words in a document d_i that are from a topic $z_k\ ,k=1,2,...,K$.

- In the WP matrix, an entry f_{jk} represents the frequency of a word w_j generated from a topic $z_k\ ,k=1,2,...,K$ over the training set.

- After LDA analysis, we use a hard decision to create topic clusters by assigning a topic z_i^* to a document d_i such that

$$z_i^* = arg \max_{1 \le k \le K} c_{ik} \ .$$

- Based on the documents belonging to each topic cluster, K topic N-gram LMs are trained.

1- Training & Testing

- It finds a predefined number of clusters based on a specific criterion, for which we chose the following function (maximize the within-class similarity):

$$(S_1 S_2 \ldots S_K)^* = arg\ max \sum_{i=1}^{K} \sqrt{\sum_{v,u \in S_i} sim(v,u)}$$

where K is the desired number of clusters, S_i is the set of documents belonging to the i^{th} cluster, u and v represent two documents, and $sim(v,u)$ is the similarity between two documents:

$$sim(u,v) = \frac{\vec{u} \times \vec{v}}{\|\vec{u}\| \times \|\vec{v}\|}$$

where $\vec{u}$ and $\vec{v}$ are the feature vectors representing the two documents respectively, again based on the NE hypotheses.

1- Training & Testing

- For a test document $d = w_1, w_2, \ldots, w_n$ that is generated by multiple topics under the LDA assumption, we formulate a dynamically adapted topic model using the mixture of LMs from different topics:

$$p_{LDA\text{-}adapt(w_k | h_k)} = \sum_{i=1}^{k} \gamma_i p_{z_i}(w_k | h_k)$$

where $p_{z_i}(w_k | h_k)$ stands for the i^{th} topic LM, and γ_i is the mixture weight. We propose a new weighting scheme to calculate γ_i that directly uses the two resulting matrices from LDA analysis during training:

$$\gamma_i = \sum_{j=1}^{n} p(z_i | w_j) p(w_j | d)$$

$$p(z_i | w_j) = \frac{f_{ji}}{\sum_{p=1}^{K} f_{jp}}$$

$$p(w_j | d) = \frac{freq(w_j | d)}{\sum_{q=1}^{n} freq(w_q | d)}$$

where $freq(w_j | d)$ is the frequency of a word w_j in the document d .

1- Training & Testing

- For a document $d = w_1, w_2, ..., w_n$, with a word distribution $p_d(w)$ and a cluster S with the associated topic specific LM $p_s(w)$, the cross entropy $CE(d,S)$ can be computed as the following using the unigram LM:

$$CE(d,S) = -\sum_{i=1}^{n} p_d(w_i) log_2(p_s(w_i)).$$

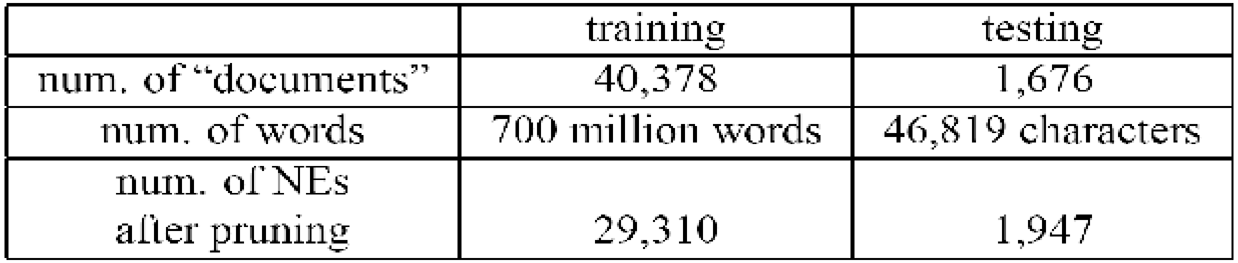
- In other word, it is the perplexity of the test document d based on the LM corresponding to topic S . For the test document, we select the cluster S^* that yields the lowest perplexity:

$$S^* = arg \min_{1 \le i \le K} CE(d, S_i)$$

1- Experiment

- The data set we used for N-best list rescoring is the GALE Mandarin 2007 Dev set. It contains about one hour of broadcast news, and 1.5 hours of broadcast conversation speech. The transcript has 2000 utterance segments in this data set.

- One goal is to infer the “topic” information based on the hypotheses, and combine the acoustic score and LM score based on the new adapted LMs to rerank the hypotheses.



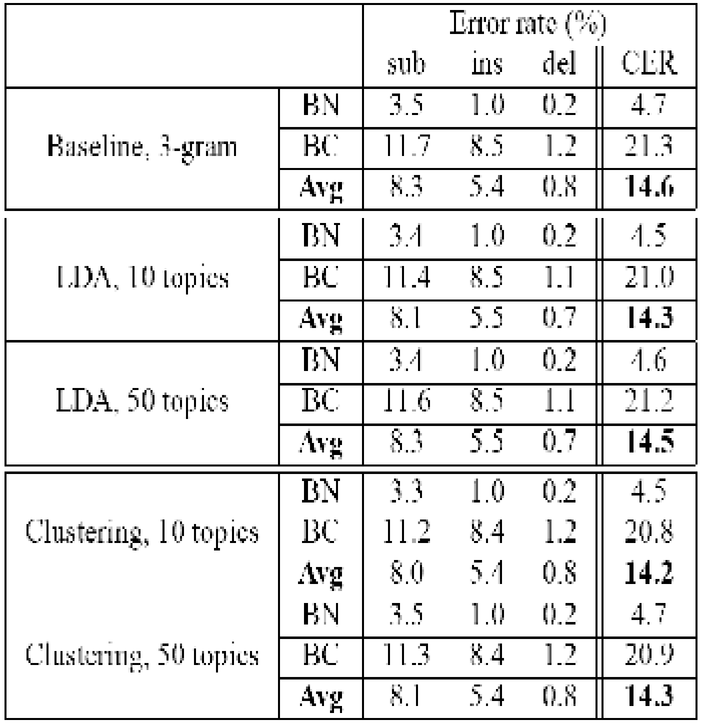
	training	testing
num. of “documents”	40,378	1,676
num. of words	700 million words	46,819 characters
num. of NEs after pruning	29,310	1,947

Table 1. Summary of data information in the large vocabulary Mandarin ASR task.

1- Experiment

- We can see that both clustering and LDA outperform the baseline trigram LM, yielding slightly lower error rate. In addition, clustering based topic modeling performs slightly better than LDA, unlike the perplexity results.

- For the two different number of topics, 10 and 50, we notice that a bigger number of topics degrades the rescoring performance in this experiment. This is also different from the perplexity results we obtained previously.



		Error rate (%)			
		sub	ins	del	CER
Baseline, 3-gram	BN	3.5	1.0	0.2	4.7
	BC	11.7	8.5	1.2	21.3
	Avg	8.3	5.4	0.8	**14.6**
LDA, 10 topics	BN	3.4	1.0	0.2	4.5
	BC	11.4	8.5	1.1	21.0
	Avg	8.1	5.5	0.7	**14.3**
LDA, 50 topics	BN	3.4	1.0	0.2	4.6
	BC	11.6	8.5	1.1	21.2
	Avg	8.3	5.5	0.7	**14.5**
Clustering, 10 topics	BN	3.3	1.0	0.2	4.5
	BC	11.2	8.4	1.2	20.8
	Avg	8.0	5.4	0.8	**14.2**
Clustering, 50 topics	BN	3.5	1.0	0.2	4.7
	BC	11.3	8.4	1.2	20.9
	Avg	8.1	5.4	0.8	**14.3**

- For LDA, this might be because of the number of mixture models we used in the current set up. For clustering-based approach with a single topic for LM adaptation, this might be explained by the smaller data size used to train the single topic adapted LM when the topic number increases.

1-Conclusion

- The experiments have shown that using the topic adapted LM, the character error rate is improved slightly compared to the baseline trigram LM using a state-of-the-art recognition system.

- Between the two topic modeling approaches, we found the difference is rather small, with clustering achieving slightly better performance than LDA, an observation different from the perplexity results.

2-Introduction

- Discriminative n-gram language modeling has been used to re-rank candidate recognition hypotheses for performance improvements in large vocabulary continuous speech recognition (LVCSR).

- Compare to the discriminative N-best re-ranking, this process of discriminative lattice rescoring has two positive advantages:
  1. Those discriminatively top-ranked efficiently identified by the A* algorithm.
  2. The rescored lattices can be further enhanced with other post-processing techniques to achieve cumulative improvement conveniently.

2-Introduction

- While maximum likelihood estimation aims to find the most likely model given the data, discriminative training attempts to minimize recognition error rate.

- The previous effort showed that discriminative n-gram modeling can effectively reduce the error rate especially when the training and testing conditions are similar.

- However, we noticed two bottlenecks: First, the discriminative n-gram model cannot be easily integrated into a single pass decoding procedure. Second, it is not straightforward to extend the discriminative n-gram modeling with other techniques to achieve cumulative improvement.

2- Discriminative N-gram Modeling

- The discriminative n-gram modeling technique defines a linear framework to re-rank the N-best recognition hypotheses.
  - We need a training data set with n speech utterances and n_i utterance hypotheses for each utterance. Define $x_{i,j}$ as the j^{th} hypothesis of the i^{th} utterance. Define $x_{i,R}$ as the utterance with lowest CER among $\{x_{i,j}\}$.
  - We need a separate test set of $y_{i,j}$ with similar definitions as the training set.
  - Define $D + 1$ features $f_d(h),\ d = 0,\ldots,D$, where is a recognition hypothesis. The features could be arbitrary functions that map h to real values.
  - Define a discriminative function as:

$$g(h,\vec{a}) = \sum_{i=0}^{D} a_i f_i(h) = \vec{a} \cdot \vec{f}(h)$$

The task of discriminative training thus involves a search for a weight vector $\vec{a}$ that satisfies the following conditions on the test set:

$$g(y_{i,R},\vec{a}) > g(y_{i,j},\vec{a}) \quad \forall i \forall j \neq R$$

2- Discriminative N-gram Modeling

- For each utterance hypothesis h , the base feature $f_0(h)$ is the recognition score which is the weighted summation of acoustic and linguistic likelihoods of h . The remaining features are the counts of each n-gram (i.e., an n-word sequence) in h . We first assign each selected n-gram with a unique id i $(1 \le i \le D)$. $f_i(h)$ is then defined as the count of the i^{th} n-gram in h .

- For instance, the unigram *"new"* and the bigram *"new solutions"* are assigned with ids j and k respectively. Given that h is *"There are new ideas and new solutions"*, $f_j(h)$ is 2 and $f_k(h)$ is 1.

- Normally, a discriminative N-gram model considers all n-grams with order $n \le N$. For example, a discriminative bigram model usually utilizes both unigrams and bigrams.

2-Recasting Discriminative Model

- Unchanging the ranking of hypotheses, we can modify the scoring method as :

$$g'(h,\vec{a}) = f_0(h) + \sum_{i=1}^{D} \frac{a_i}{a_0} f_i(h)$$

the second part can be expanded into :

$$\sum_{i=1}^{D} \frac{a_i}{a_0} f_i(h) = \frac{1}{a_0}(a_{w_1} + a_{w_2} + \ldots + a_{w_m} +$$
$$a_{w_1 w_2} + a_{w_2 w_3} + \ldots + a_{w_{m-1} w_m} +$$
$$a_{w_1 w_2 \ldots w_N} + a_{w_2 w_3 \ldots w_{N+1}} + \ldots + a_{w_{m-N+1} w_{m-N+2} \ldots w_m})$$

where $w_1 w_2 \ldots w_m$ is the corresponding word sequence of the utterance hypothesis h . $a_{w_i w_{i+1} \ldots w_{i+k}}$ is the weight of the n-gram $w_i w_{i+1} \ldots w_{i+k}$.

2-Recasting Discriminative Model

- The first part $f_0(h)$ is the score that the recognizer assigned to h , shown as follows :

$$
\begin{aligned}
f_0(h) &= \sum_{i=1}^{m} (\alpha \cdot AcScore(w_i) + \beta \cdot LmScore(w_i)) - n \cdot InsertPenalty \\
&= \alpha \sum_{i=1}^{m} AcScore(w_i) + \beta \sum_{i=1}^{m} P(w_i \mid w_1, w_2, \ldots, w_{i-1}) - n \cdot InsertPenalty
\end{aligned}
$$

where $P(w_i \mid w_1, w_2, \ldots, w_{i-1})$ is the log-domain LM likelihood provided by language model. α and β are the acoustic and language model weights.

2-Recasting Discriminative Model

- We can combine the equations as :

$$
\begin{aligned}
g'(h,\vec{a}) &= f_0(h) + \sum_{i=1}^{D} \frac{a_i}{a_0} f_i(h) \\
&= \alpha \sum_{i=1}^{m} AcScore(w_i) + \beta \sum_{i=1}^{m} P'(w_i \mid w_1, w_2, \ldots, w_{i-1}) - n \cdot InsertPenalty
\end{aligned}
$$

$$
where\ P'(w_i \mid w_1, w_2, \ldots, w_{i-1}) = P(w_i \mid w_1, w_2, \ldots, w_{i-1}) + \frac{1}{a_0 \cdot \beta}\left(a_{w_i} + a_{w_{i-1}w_i} + \ldots + a_{w_{i-N+1}w_{i-N+2}\ldots w_i}\right)
$$

- We can find that scoring an utterance hypothesis by the discriminative n-gram model is equivalent to scoring the hypothesis by the recognizer with a modified language model.

2-Recasting Discriminative Model

- We can represent the discriminative N-gram model with a pseudo-conventional L-gram model as :

$$P'(w_i \mid w_{i-L+1}, w_{i-L+2}, \ldots, w_{i-1}) = P(w_i \mid w_{i-L+1}, w_{i-L+2}, \ldots, w_{i-1}) + \frac{1}{a_0 \cdot \beta}(a_{w_i} + a_{w_{i-1}w_i} + \ldots + a_{w_{i-N+1}w_{i-N+2}\ldots w_i})$$

2- Pseudo-Conventional N-gram Model

- The pseudo-conventional n-gram can be computed using two possible methods :
  1. Compute the pseudo-conventional n-gram model offline.
     - The pseudo-conventional n-gram can be build by modifying the n-gram entries in the original n-gram model incorporated in the baseline recognizer.
     - The difficulty lies in the fact that the n-gram model in the recognizer normally does not contain all possible n-grams. This is due to the usage of the back-off strategy for n-gram modeling.
     - For example, if a bigram does not include in the model we would calculate as $P(w_2 | w_1) = b(w_1)P(w_2)$, where $b(w_1)$ is the back-off weight of w_1 .

  2. Compute the pseudo-conventional n-gram model online.

2- Pseudo-Conventional N-gram Model

- The basic idea of DLR (Discriminative lattice rescoring) is to replace the original LM score with the pseudo-conventional n-gram probability for each word node/link in a lattice based on the word history.

- As the equation below, the calculation of a pseudo-conventional n-gram probability is composed of two parts: (1) the score from the original n-gram model, and (2) the score from the discriminative n-gram model.

$$P'(w_i \mid w_{i\text{-}L+1}, w_{i\text{-}L+2}, \ldots, w_{i\text{-}1}) = \underbrace{P(w_i \mid w_{i\text{-}L+1}, w_{i\text{-}L+2}, \ldots, w_{i\text{-}1})}_{\text{original n-gram model}} + \frac{1}{a_0 \cdot \beta}\underbrace{(a_{w_i} + a_{w_{i\text{-}1}w_i} + \ldots + a_{w_{i\text{-}N+1}w_{i\text{-}N+2}\ldots w_i})}_{\text{discriminative n-gram model}}$$

2- Other Post-Processing Technique

- We extend DLR (Discriminative lattice rescoring) with N-best re-ranking procedure is applied to each utterance as follows:
  1. Select the N-best hypotheses from the corresponding lattice.
  2. Assign each hypothesis with a word mutual information score

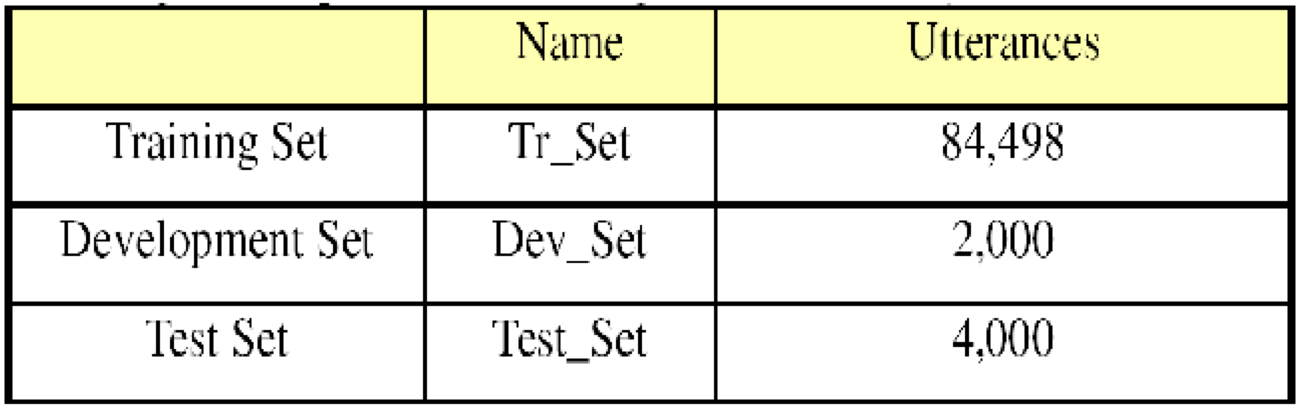
$$MI(w_1, w_2, \ldots, w_m) = \frac{\sum_{i \neq k} MI(w_i, w_k)}{C_m^2}$$

     where the mutual information $MI(w_i, w_j)$ is the co-occurrence rate of the two words within an utterance.
  3. Score each hypothesis *hypo by linear interpolation*

$$Score(hypo) = \mu \cdot MI(hypo) + (1 - \mu) \cdot LatScore(hypo)$$

     where $LatScore(hypo)$ *is the weighted summation of the acoustic and language model scores extracted from the lattice.*
  4. *The top-scoring hypothesis is the outcome of the re-ranking process.*

2-Experiment



	Name	Utterances
Training Set	Tr_Set	84,498
Development Set	Dev_Set	2,000
Test Set	Test_Set	4,000

Table 1. Data sets

- The baseline LVCSR is a state-of-the-art decoder. The cross-word triphone acoustic models were trained on a separate Mandarin dictation speech corpus of about 700 hours.
- A trigram model was trained on about 28G (disk size) domain-balanced text corpora, using a 60606-word lexicon. This baseline decoder provide a 19.86% CER (character error rate) on Test_Set.

2-Experiment

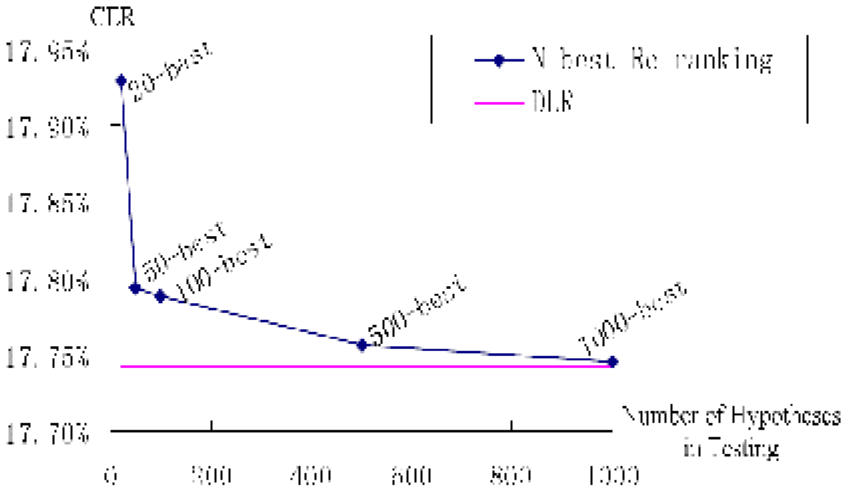


Figure 1. Evaluation using Model_N20

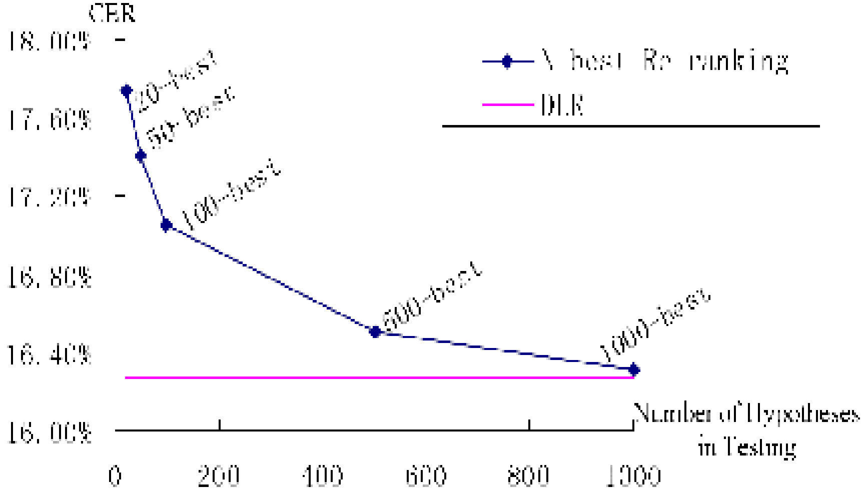
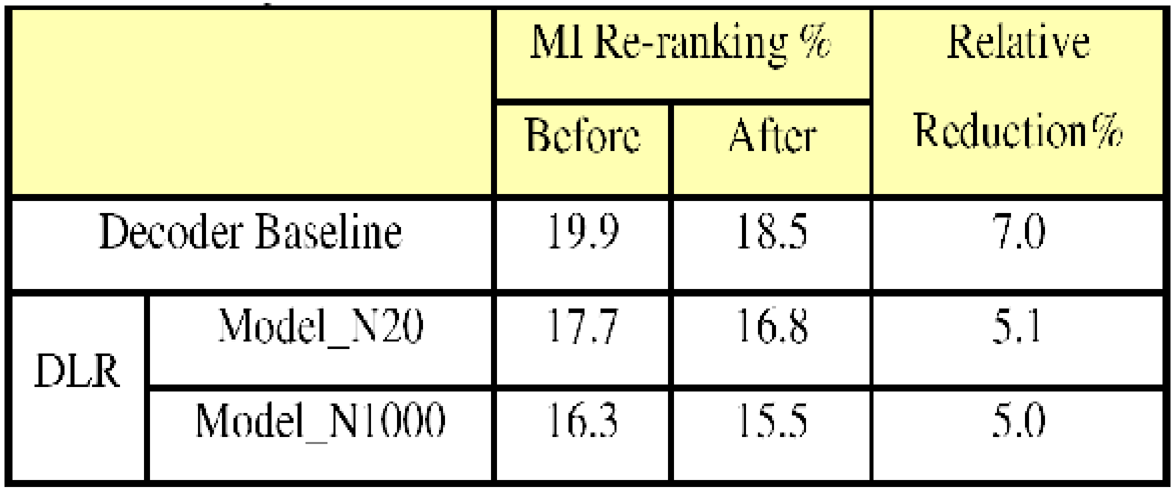


Figure 2. Evaluation using Model_N1000

- We selected the model trained on 20-best hypotheses, named Model_N20, as well as the model trained on 1000-best hypotheses, named Model_N1000.

- For discriminative N-best re-ranking, re-ranking more hypotheses brings better performance for either Model_N20 or Model_N1000, partially because the training and testing conditions are similar.

2-Experiment



		MI Re-ranking %		Relative Reduction%
		Before	After	
Decoder Baseline		19.9	18.5	7.0
DLR	Model_N20	17.7	16.8	5.1
	Model_N1000	16.3	15.5	5.0

Table 2. CERs on various baselines

- We applied the MI based 100-best re-ranking on the baseline from the original decoder.
- We observed that several percentage points of relative improvement across various performance baselines, indicating that technique combination based on discriminative lattice rescoring is feasible.

2-Conclusion

- Experiments with Mandarin LVCSR show that DLR (Discriminative lattice rescoring) can identify efficiently the best hypothesis in lattice, when compared to discriminative N-best re-ranking.
- We extended the DLR processing further with re-ranking by word mutual information and achieved cumulative improvements in recognition performance.