Prompt-based Re-ranking Language Model for ASR

Mengxi Nie, Ming Yan, Caixia Gong
DiDi Chuxing, Beijing, China
niemengxi@didiglobal.com, lucyyan@didiglobal.com, gongcaixia@didiglobal.com

Abstract

In Automatic Speech Recognition (ASR), the language model re-ranking based on unlabeled text can improve the performance and realize flexibly scene adaptation. The scheme of ASR re-ranking is usually to build a language model and then use it to reorder the speech recognition N-best hypotheses. Recently, BERT-based re-ranking has achieved impressive results, benefiting from the powerful modeling capability of contextual semantic. In the view of that BERT’s non-autoregressive structure limits the calculation speed of the language model scores (perplexity, ppl), we use a classification method in prompt paradigm instead of the re-ranking method based on ppl. The prompt-based re-ranking scheme simplifies the pipeline of re-ranking as well as ensures the performance. Experiments on AISHELL-1 dataset show the effective of our proposed method. On the test set, the inference speed is accelerated by 49 times and compared to baseline the Character Error Rate (CER) is relatively decreased by 13.51% ~ 14.43%.

Index Terms: automatic speech recognition, language model, re-ranking, prompt, attention

1. Introduction

With the development of neural network, end-to-end ASR has attracted extensive attention[1, 2, 3]. As we all know, large amount of labeled data is required to achieve an available recognition performance. Manually annotating audios is time-consuming and expensive. While the language model is obtained much more easily due to its self-supervised training and only relying on text corpus. Introducing extra language models into the end-to-end ASR is a popular and effective means to improve the recognition performance[4, 5, 6].

There are two kinds of methods, language model fusion and language model re-ranking. The main difference is that the language model is introduced at different decoding stage of speech recognition. Language model fusion is involved in the first-pass decoding of speech recognition, including shallow fusion[7, 8], deep fusion[4, 9] and cold fusion[10]. Considering the real-time requirement, the structure and size of the language model for fusion are limited[11]. Language model re-ranking is used in the second-pass decoding to select the best transcript from the N-best first-pass hypotheses[5, 6, 12]. Since the hypotheses set is usually small and text sequences are processed faster than speech sequences, the language model for re-ranking has more diverse and flexible structure. Previous studies on language model re-ranking is usually divided into two steps. Firstly, an extra text-based language model is built to score each hypothesis, and then the score combined with the decoding score optionally is used for ranking. Both N-gram and neural network are common methods for building language model and the latter has stronger long-distance modeling capability[13, 14].

Recently, Bidirectional Encoder Representation from Transformers (BERT)[15] has shown great advantages in natural language processing because of its powerful language representation and feature extraction ability. Simultaneously, BERT-based re-ranking has also achieved impressive results [6, 16]. BERT language model is used to calculate the perplexity (ppl) score of every hypothesis, which requires masking each token of the sequence in turn and is time-consuming due to the non-autoregressive structure. Besides, the commonly re-ranking scheme scores and then ranks each hypothesis, without using the information between hypotheses. While the essence of re-ranking task is to select the recognition result with the lowest character error rate from hypotheses list, and does not pay attention to the order of hypotheses. Based on the above motivation, we treat the re-ranking task as a classification task to directly obtain the best result, like[6]. The classification method can not only reduce the inference time, but also capture richer and more suitable semantic information. In addition, most of language models for re-ranking are only trained with the gold text data and cannot obtain the exclusive information from the specify speech recognition system[17]. For example, some kind of transcription error is always present in the system. The proposed method uses the recognition results as the training data, which make the model gain the exclusive information.

However, the related research[18] and our prior experiments demonstrate that model collapse is occurred when the BERT is directly used to train the classification model, due to imbalanced or insufficient categorical data. In order to solve this problem and ensure performance, we introduce prompt into the language model re-ranking task. The reason is that the prompt-based learning can reduce the gap between the pre-trained model and downstream tasks[18, 19]. As far as we know, we are the first to apply the prompt to the re-ranking task. Furthermore, we integrate the first-pass decoding score into the model based on attention, which provides good initial information and makes the training easier to converge.

The paper is organized as follows: Section 2 outlines the related work. Section 3 details the proposed prompt-based re-ranking method. In Section 4, the performance of our method is evaluated. And the conclusion is given in Section 5.

2. Related work

In the section, we introduce the language model re-ranking for ASR and some related approaches. Given a text sequence $S = \{\omega_1, \omega_2, \ldots, \omega_N\}$, the language model can calculate the probability $P(S)$ of the sentence. Perplexity(ppl) is often used to evaluate the quality of the sentence. The smaller the ppl score, the better the sentence. The commonly used re-ranking scheme is that building an extra language model to give the corresponding ppl scores for N-best hypotheses, then the best transcript is selected according to the scores. Optionnally, the score for re-ranking can be obtained by combining with other information using some strategies, for example, taking the weighted sum of the decoding score and the language model score.

In the above re-ranking scheme, there are two main re-
search points. One is how to build the language model, and the other is the design of the re-ranking strategy. The N-gram LM and RNN LM for re-ranking have been discussed in many papers[5, 12] and developed in ASR systems. Recently, BERT-based re-ranking is also explored in [6, 16]. Instead of a standard language model objective, BERT achieves a bidirectional language model by two tasks, masked language model and next sentence prediction. Although the character accuracy outperforms that of LSTM, it suffers from slow inference speed because of the special mask decoding method. To this end, we propose the prompt-based re-ranking method. There are also some researches that focus on the loss function of language model. A representative approach is to involve candidate hypotheses and the decoding scores in training through MWER[17, 20]. The authors point that the language model should be adapted to the specific ASR system, because the hypotheses with low CER may not be semantically fluent and the transcription errors are relatively fixed in the system. This idea has also been extended to our proposed method. For the design of the re-ranking strategy, some work uses lattice [21] or interpolation algorithm [22] to calculate the final score. There is also some work to improve performance by adding several different language models according to the specific topic, style, or domain [23]. In the paper, we use the other information through the mechanism of attention, which accelerates the convergence.

3. Methods

3.1. BERT-based LM re-ranking

As the above mentioned, the LM score refers to the sentence perplexity. According to the chain rule, the perplexity is decomposed into the form of multiplying the probability of token given its entire history. And the negative log-likelihood is generally used, as follows:

\[
PPL(S) = P(\omega_1, \omega_2, ..., \omega_N)^{-\frac{1}{N}} = \prod_{i} P(\omega_i|h_i)^{-\frac{1}{N}} (1)
\]

\[
LM\_score = -log(PPL(S)) = \frac{1}{N} \sum_{\omega_i} log(P(\omega_i|h_i)) (2)
\]

where \( N \) is the length of the text sequence and \( h_i \) represents the context of \( \omega_i \). The neural language model is generally built by maximizing the conditional probability of each token in corpus. BERT achieves a bidirectional language model based on the self-attention through masked-LM task. Thus it is necessary to mask a token in the sentence each time to calculate the corresponding conditional probability, as shown in Figure 1. This inference way is relatively slow, limiting BERT-based LM re-ranking developed in ASR systems requiring real-time rate. Besides, all hypotheses are concatenated together as model input, which make the model grasp more useful information. To this end, we convert re-ranking based on ppl into a prediction task [6], which directly predicts the best hypothesis. The difference is that we do not classify the hypotheses, but make the prediction in the scheme of language model with prompt.

3.2. Prompt-based learning paradigm

Recently, benefiting from the pre-training and fine-tuning paradigm, many natural language processing (NLP) researches have made great progress. In the paradigm, a language model is firstly pre-trained on large text datasets. Then the above model will be introduced additional parameters and fine-tuned using the task-specific objective functions for different downstream tasks [24]. Traditional supervised learning trains a model to take \( X \) as input and output \( Y \) in the form of \( P(X|Y) \), which requires a lot of labeled data. However, prompt-based learning is only based on the language model, which directly outputs the probability of text that can be used to predict \( Y \). It solves some NLP tasks by providing the pre-trained language model with "task descriptions", namely prompt which guides the learning direction. The method reduces the gap between the pre-trained model and downstream tasks [19], which cat not only speed up the training but also achieve better performance. In addition, a small amount of labeled data is required because of that factual and commonsense knowledge can be retrieved from a pre-trained language model [18].

In the prompt-based learning paradigm, the original input \( X \) is modified into a textual string \( X' \) with prompt according to a template. The template has two slots: an input slot \( X \) and an answer slot \( Z \) for an intermediate generated answer text \( Z \) that will later be mapped into original label \( Y \) [24]. After the template is determined, filling \( X \) into slot \( [X] \) to get \( X' \) as the input, the language model predicts the content of slot \( [Z] \). Here is an example of sentiment analysis mentioned in the [24]. If the task is to judge whether the emotion of a sentence is positive or negative, we can map these two categories to the corresponding natural language words, such as "good", "fantasy" and "boring". The template and prompt of the task is as follows:

- Template: [X] It’s a [Z] movie.
- Input: \( X = I \) love this movie.
- Prompt: \( X' = I \) love this movie. It’s a [Z] movie.

Actually, the template can be not only the natural language, but also the continuous values of text embedding. And the answer slot can also be multiple. More methods and templates of prompt are not described in details. In the paper, the most basic prompt is used.
3.3. Prompt-based re-ranking

In the section, how to apply prompt-based learning to re-ranking task is described. As mentioned, BERT-based LM re-ranking is limited by the speed of inference in the commonly re-ranking scheme. Besides, the common method to select the hypothesis with highest LM score obtains a transcript with relatively fluent semantics. The purpose of re-ranking is to get the oracle transcript with the lowest CER. As shown in Table 1, among the 4-best hypotheses, there are two transcript(L1, L2) with similar semantic scores. If the re-ranking model has some prior knowledge, for example, these sentences are transcribed by a specified speech recognition system, the task objective is select the sentence with high character accuracy and the scene is in the taxi, the L2 is better than L1.

Table 1: The example of hypotheses.

<table>
<thead>
<tr>
<th>index</th>
<th>hypotheses</th>
<th>LM score</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td>实话 楼面价三千元</td>
<td>-18.23</td>
</tr>
<tr>
<td>L2</td>
<td>起始楼面价三千元</td>
<td>-19.66</td>
</tr>
<tr>
<td>L3</td>
<td>七十楼面价三千元</td>
<td>-24.75</td>
</tr>
<tr>
<td>L4</td>
<td>起始楼面价三千亿元</td>
<td>-25.13</td>
</tr>
</tbody>
</table>

Table 2: The mapping between Z and Y.

<table>
<thead>
<tr>
<th>Z</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1, 2</td>
<td>1</td>
</tr>
<tr>
<td>2, 3</td>
<td>2</td>
</tr>
<tr>
<td>3, 4</td>
<td>3</td>
</tr>
<tr>
<td>4, 5</td>
<td>4</td>
</tr>
</tbody>
</table>

Therefore, the re-ranking task is treated as a classification task to predict the best index in hypotheses list. And prompt is introduced into BERT-based re-ranking to provide the prior knowledge. As shown in Figure 2, the left is the original classification task, and the right is the prompt-based re-ranking. The prompt template is "[X]这个识别结果中字错误率最低的是第[Z]个" and is expressed in English as "[X] Among the K hypotheses, [Z] is the one with the lowest CER." The template is selected from several manual designed templates. The evaluation criterion is the CER of the validation set on the original BERT model. The slot [X] will be filled with the N-best hypotheses connected by the symbol [SEP]. The value space of [Z] is {1, 2, ..., K, 1, 2, ...,} and is expressed in English as {1st, 2nd, ..., Kth, first, second, ...} determined by the number of hypotheses. The mapping between Y and Z is shown in the Table 2. During training and inference, the BERT-based LM model is used to predict the token at this location of [Z]. We follow the training method of BERT-based MLM. The slot [Z] is replaced with the token [MASK]. Like the seminal paper of BERT, the position vector of the token involved in each hypothesis and the segment vector to indicate the change of hypotheses are also taken as input. It is optimized based on the cross-entropy loss function as Equation 3, where \( \omega_{\text{mask}} \) represents the masked token in the slot [Z].

\[
f_{\text{loss}} = - \sum \log(P({\omega_{\text{mask}}} | S_{\omega_{\text{mask}}, i})) \quad (3)
\]

3.4. Addition of other information

It’s mentioned above that other information can be combined for re-ranking in the commonly scheme. The information includes decoding score, ppl obtained from other LMs and so on. The decoding score represent acoustically related features. And it’s considered that the decoding score can provide guidance for training through giving the model a good initialization. In the view of that this score is not entirely reliable and only for reference, we add a attention block to introduce the information into the model as shown in Figure 2. The confidence is determined by the model through soft attention mechanism(Equation 4). The context embedding of [MASK] is assigned \( Q \), and the dimension is \( h \). While the decoding score is assigned to \( V \) and mapped to \( h \)-dim as \( K \). Then the context embedding of [MASK] and the output of the attention block are concatenated together and input to a feed-forward layer. Another information can also be added into the model in this way.

\[
I_{\text{other}} = \text{softmax}(QK) * V \quad (4)
\]

4. Experiments

In the section, the proposed method is evaluated by some experiments to prove that the inference speed is improved while
the character accuracy is not lost. It’s also considered whether performance is improved after adding the decoding score.

4.1. Setups

All experiments are conducted on the AISHELL-1 dataset[25] that is an open source Mandarin speech corpus. The dataset consists of training set, verification set and test set with a total of 178 hours, involving 400 speakers, and each speaker records about 300 sentences. The text covers 11 fields such as smart home, automatic driving and industrial production, etc. We directly use the division of the dataset. In addition to model training, the verification set is also used for prompt template selection and adjustment of interpolation parameters. The test set is for the final performance evaluation of models. The 4-best hypotheses are obtained by Kaldi[26], an open source speech recognition tool. In the commonly re-ranking scheme, BiLSTM-based and BERT-based LM are trained and the train data is the reference text about 120,000 sentences. We also do the pre-training and finetuning experiments on BiLSTM. A large text data set called THUNews (more than 20 million sentences) is used for pre-training. The prompt-based re-ranking model uses the 4-best hypotheses as its train data. We use the pre-trained BERT-Base model\(^1\) for Chinese as our base model. The number of layers is 12 layers and the hidden size is 768. The size of BERT dictionary is about twenty thousand, while the BiLSTM model is based on more than 5,000 commonly used Chinese characters. The network structure of BiLSTM includes a hidden layer and a projection layer. The size of the hidden layer is 800 and the projection layer is 200. The learning rates are 5e-5 and 0.1, respectively. The optimization algorithm is Adam[27].

4.2. Results

Table 3: The character error rates of all models.

<table>
<thead>
<tr>
<th>Method</th>
<th>dev</th>
<th>test</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>11.39%</td>
<td>12.95%</td>
</tr>
<tr>
<td>oracle</td>
<td>7.89%</td>
<td>9.40%</td>
</tr>
<tr>
<td>without the decoding score</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BiLSTM</td>
<td>11.58%</td>
<td>12.96%</td>
</tr>
<tr>
<td>BiLSTM THU</td>
<td>10.47%</td>
<td>11.97%</td>
</tr>
<tr>
<td>BiLSTM THU(finetune)</td>
<td>10.42%</td>
<td>11.91%</td>
</tr>
<tr>
<td>BERT-based(ppl)</td>
<td>9.54%</td>
<td>10.90%</td>
</tr>
<tr>
<td>BERT-based(class)</td>
<td>10.31%</td>
<td>16.22%</td>
</tr>
<tr>
<td>prompt-based</td>
<td>9.51%</td>
<td>11.20%</td>
</tr>
<tr>
<td>with the decoding score</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BiLSTM</td>
<td>11.39%</td>
<td>12.95%</td>
</tr>
<tr>
<td>BiLSTM THU</td>
<td>10.40%</td>
<td>11.91%</td>
</tr>
<tr>
<td>BiLSTM THU(finetune)</td>
<td>10.36%</td>
<td>11.87%</td>
</tr>
<tr>
<td>BERT-based(ppl)</td>
<td>9.50%</td>
<td>10.88%</td>
</tr>
<tr>
<td>prompt-based(attention)</td>
<td>9.40%</td>
<td>11.08%</td>
</tr>
</tbody>
</table>

In Table 3, we list the re-ranking CER of each model. The first row is the baseline without re-ranking that is the CER of the top-1 transcript. The oracle represents the best result that can be achieved with this hypotheses set. The models of BiLSTM and BiLSTM THU are trained using AISHELL-1 and THUNews dataset, respectively. BiLSTM THU(finetune) is obtained by fine-tuning the model of BiLSTM THU on the AISHELL-1. BiLSTM-based methods and BERT-based(ppl) all use ppl for re-ranking and the decoding score is introduced by interpolation. BERT-based(class) represents the method of directly classifying hypotheses. As can be seen from Table 3, the performance of BERT-based methods is significantly better than that of BiLSTM-based methods. The method of directly classifying hypotheses suffers from overfitting. While the CER of the proposed methods are comparable to that of BERT-based LM re-ranking and is relatively reduced by 13.51% ∼ 14.43% compared with the baseline. The Table 4 shows the average inference time of each method. The prompt-based re-ranking is much faster than BERT-based(ppl), about 49 times, even faster than the BiLSTM-based method.

Furthermore, the character accuracy with the introduction of decoding score is higher than that without decoding score from Table 3. Additionally, the information of decoding score can accelerate model training convergence as we expected. Figure 3 shows the training curves with or without decoding score attention, where the right curve converges before the left one and is smoother.

5. Conclusions

End-to-end ASR models have been increasingly used in practical system and face challenges in various scenarios. Language model re-ranking is an effective method to improve the performance, which can support the rapid domain adaptation. BERT is a powerful model used for language model re-ranking, but the inference speed of BERT-based re-ranking needs to be optimized. To this end, we turn the regression task into classification task, and introduce the prompt to ensure model performance. Moreover, the decoding score is added into model by a soft-attention block, which makes the model converge faster. The experiment results demonstrate the effectiveness of our proposed method as we expected. In the future, we may explore three points. One is one-shot learning to reduce the need for labeled data. One is template design such as automatic learning template. The other is model compression which can reduce the amount of parameters to facilitate deployment, and can further improve inference speed.

Table 4: The inference time of each method.

<table>
<thead>
<tr>
<th>Method</th>
<th>Time-consuming(ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BiLSTM</td>
<td>34.77</td>
</tr>
<tr>
<td>BERT-based(ppl)</td>
<td>588.47</td>
</tr>
<tr>
<td>BERT-based(class)</td>
<td>10.46</td>
</tr>
<tr>
<td>prompt-based</td>
<td>11.32</td>
</tr>
<tr>
<td>prompt-based(attention)</td>
<td>12.00</td>
</tr>
</tbody>
</table>

Figure 3: The training curves of the proposed methods.
6. References


[25] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, “Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline,” *Beijing Shell Shell Technology Co. Ltd, Beijing, China;Beijing, China;Beijing, China;Beijing, China;Beijing, China;*.
