A New Word Language Model Evaluation Metric For Character Based Languages

Peilu Wang, Ruihua Sun, Hai Zhao, and Kai Yu

Institute of Intelligent Human-Machine Interaction
MOE-Microsoft Key Lab. of Intelligent Computing and Intelligent Systems
Department of Computer Science and Engineering
Shanghai Jiao Tong University, 200240, Shanghai, P. R. China
{plwang1990,sun.r.h,zhaohai,kai.yu}@sjtu.edu.cn

Abstract. Perplexity is a widely used measure to evaluate word prediction power of a word-based language model. It can be computed independently and has shown good correlation with word error rate (WER) in speech recognition. However, for character based languages, character error rate (CER) is commonly used instead of WER as the measure for speech recognition, although language model is still word based. Due to the fact that different word segmentation strategies may result in different word vocabulary for the same text corpus, in many cases, word-based perplexity is incompetent to evaluate the combined effect of word segmentation and language model training to predict final CER. In this paper, a new word-based language model evaluation measure is proposed to account for the effect of word segmentation and the goal of predicting CER. Experiments were conducted on Chinese speech recognition. Compared to the traditional word-based perplexity, the new measure is more robust to word segmentation and shows much more consistent correlation with CER in a large vocabulary continuous Chinese speech recognition task.[1]

Keywords: language model evaluation, character error rate, perplexity

1 Introduction

In speech recognition, language model plays an important role. It models the prior probabilities of all possible word sequence that a speech recogniser can deal with. It is independent of acoustic observations and defines the search space of a speech recogniser. In speech recognition, word error rate (WER) is usually used as the ultimate evaluation metric for the whole system. Although WER can also be used to evaluate language model given a fixed acoustic model, it is not convenient to do so because acoustic data is required and decoding is timeconsuming.

[1] This research was partly supported by the Program for Professor of Special Appointment(Eastern Scholar) at Shanghai Institutions of Higher Learning and the China NSFC project No.61222208

To conveniently evaluate the quality of an estimated language model, perplexity was proposed and has been the most widely used metric [5]. Perplexity is essentially the exponent of the cross entropy between the real word sequence distribution and the estimated word sequence distribution. Its calculation is independent of acoustic data and can be done quickly. More importantly, it was shown that perplexity has good correlation with WER [1, 6]. Hence, it has been used for decades to evaluate language model in speech recognition. However, there has also been a long argument about the correlation between perplexity and WER. Previous works showed that the good correlation between perplexity and WER only exists in certain cases [3] and modifications of perplexity has been proposed to improve the correlations in more general cases [3, 4, 2].

In these studies, different factors are changed to construct different language models, such as corpus size, smoothing algorithm, interpolation weight and so on. Then the correlation between perplexity and WER of all different language models is investigated. However, all the previous works, to our best knowledge, have not explicitly considered the influence of voabulary on language model training. It may be because that vocabulary is normally fixed before language model training given certain training corpus and consequently does not have remarkable influence. Although this is a common case in word based languages, in character based languages such as Chinese, the influence of vocabulary can not be neglected. Since character based languages are not naturally defined with spaces appearing between words, corpus needs to be segmented to form words before language model training. Different segmentation strategies will generate different word vocabularies with totally different size and components which lead to different probability distribution and final recognition result. We will show in the following chapter that in this situation, perplexity is incompetent to predict the recognition performance.

What's more, for character based languages, character error rate(CER) was used to evaluate the final performance instead of word error rate because character becomes the basic unit while language model is still trained based on word, since word based language model always tends to get a better performance in application. This mismatch makes it harder for perplexity to do an accurate evaluation that perplexity only considers the probability distribution of each word but ignores the information of word itself. For example, it is intuitive that the length of word have relation with the CER because word with more characters will cause more incorrectly recognised characters in CER calculation and this effect will not be recognised by perplexity.

In this paper, traditional word based perplexity is extended to take the effect of vocabulary construction into consideration. Two new evaluation functions are proposed, one is taking the vocabulary size into consideration and the other one is considering the vocabulary size as well as the length of word. Experiments are performed to investigate the correlation between different versions of perplexity and CER, where the segmentaion strategy and word vocabulary are the variable quantities. The result shows that these new measures are more robust and

present much more consistent with CER while the influence of word length is not as strong as we thought.

The rest of the paper is arranged as follows. Section 2 reviews traditional word based perplexity and proposes two modified versions for character based languages. Experiments are described in section 3, followed by conclusion.

2 Character based perplexity

2.1 Word based perplexity and its limitation

In natural language processing, it is assumed that the appearance of word in sentences satisfying some specific kind of probability distribution referred to as language model. The model that can best reflect such distribution is called the real model but limited to the calculation ability, it is impossible to achieve this real model in practice. Therefore, the quality of language model is always assessed by quantitatively measuring the difference between the estimated language model and the real model. This can be done by asking how well the estimated model can predict the words generated from the real word distribution. For a given test word sequence $\mathbf{w} = \{w_1, \cdots, w_N\}$, where N is the number of words, the perplexity (*ppl*) of the estimated language model $q(\mathbf{w})$ is defined as

$$ppl = 2^{-\frac{1}{N}\log_2 q(\mathbf{W})} = 2^{-\frac{1}{N}\sum_{i=1}^{N}\log_2 q(w_i|h_i)} \tag{1}$$

where w_i is the i^{th} word of the whole test set $\mathbf{w}$ and $h_i = \{w_1, \cdots, w_{i-1}\}$ denotes the word history of w_i . Assuming the real word sequence distribution is p($\mathbf{w}$), better estimated model q($\mathbf{w}$) of the unknown distribution p($\mathbf{w}$) will tend to assign higher probabilities to the test word sequences. Thus, they have lower perplexity, meaning that they are less surprised by the test sample.

Considering $\log_2 q(w_i|h_i)$ represents the bits needed to record the information of word w_i given history h_i, the exponent in equation (1) can be regarded as the number of bits needed per word to represent the test set if the coding scheme used is based on $q(\cdot)$. Low *ppl* means the estimated model requires few bits per word to compress the test set which means the model is more close to the real model.

In most cases, ppl calculated by equation (1) works quite well, but when the vocabulary changes, it always tends to behave poorly. Since language model is word based, even for character based language, a word vocabulary is required to determine the set of valid words. Words not appeared in the vocabulary will not be taken into consideration when calculating the *ppl*. The size and composition of the word vocabulary will severely affect *ppl*s evaluation. For example, considering two language model LM_A and LM_B , LM_A has only 50 words, and the probability of each word is equal which is 1/50, and LM_B has 100 words and the probability of each word is also equal for convenience. According to the equation (1), the *ppl* of LM_A is 50 while the *ppl* of LM_B is 100. Although the *ppl* of LM_A is much lower than LM_B , it is likely that, LM_A which contains more words will get a better performance in application due to better coverage of words.

2.2 Character based perplexity

Considering the definition of perplexity(*ppl*), it uses the average bits needed to compress the test set as the criterion to evaluate language model but ignores the vocabulary size. As the example given in last paragraph, it is obviously unfair to compare the number of bits if the two language model have different vocabulary size. Therefore, equation (2) is extended to take the size of vocabulary into consideration. Since this function is designed for conquering problem appearing in character based languages, it is denoted as character-based perplexity (*cppl*) for convenience. The extended function is defined as:

$$cppl = 2^{-\frac{1}{N|V|}\sum_{i=1}^{N}\log_2 q(w_i|h_i)} \tag{2}$$

where $|V|$ is the number of words of vocabulary V . This is an empirical function that introduces the size of vocabulary as a balancing factor. The language model which has a smaller vocabulary size tends to have larger $q()$ and therefore will get a smaller exponent and smaller *ppl*. In contrast, in equation (2), the exponent will become larger with smaller vocabulary size which will neutralize the effect of $q(\cdot)$.

What's more, as mentioned before, in character based language, the information of word itself is also an influence factors. Since character based languages are not naturally defined with spaces appearing between words, these words which are decided by segmentation and corpus contains far more possibilities than those in word based languages. For example, on a 310M Chinese text corpus, the size of vocabulary after word segmentation can be more than 1000k! Not only the vocabulary size, the words in vocabulary constructed from different segmentation strategies may also have notable difference. To make it easy to consider this difference, the effect of word length is considered, since it seems intuitive that longer word will cause more error characters if it is incorrectly recognized in speech recognition. The bits needed to transfer the word into equation (2) is further introduced and a refined character-based perplexity, referred to as $cppl_2$ is defined as below:

$$cppl = 2^{-\frac{1}{N|V|}\sum_{i=1}^{N}\frac{1}{1+\log_2(|w_i|)}\log_2 q(w_i|h_i)} \tag{3}$$

where $|w|$ denotes the number of characters of word w, i.e. word length. This is also an empirical function that considering both of the effect of the vocabulary composition as well as the vocabulary size.

3 Experiments

To investigate the correlation between the new language model evaluation measures and CER, experiments were performed on a large vocabulary Chinese speech recognition task. The acoustic model is a cross-word triphone model trained on about 200 hours of read speech using the minimum phone error (MPE) criterion. It has about 3000 clustered states and an average of 12 Gaussian components per state. The acoustic model was fixed for all experiments. The text

corpus used to train language models were extracted from Weibo[2] consisting of 42M sentences and 101M characters. A series of trigram language models were then trained during the experiments. The test data for calculating perplexity and CER consists of 2040 sentences, about 20K characters. All these sentences were preprocessed to ensure that they were composed with 6763 simplified Chinese characters and other symbols were filtered. The toolkit to train language model was SRILM[7] and HTK toolkit[8] was used to decode the lattice transcript.

In this experiment, 10 different language models were constructed. Unlike previous works which mostly focused on adjusting the smoothing algorithm or interpolation weight, different language models were generated by utilizing different segmentation strategies in this experiment. To achieve many different segmentation strategies, backward maximal matching(BMM) word segmentation algorithm was used with different vocabularies. These vocabularies was consciously constructed to let the segment result varied obviously, having apparent divergence in word length and vocabulary size to better check the performance of *cppl* and $cppl_2$. The pseudocode generating these vocabularies is shown in Algorithm 1.

These vocabularies are constructed by merging the bigram and trigram in trigram count with high frequency. In our algorithm, if the n-gram(n>1) words having high frequency which is represented by the appearance times counted in held out corpus, it is supposed to be a new word and is added to the new vocabulary generated for the next segmentation strategy. The criterion judging high frequency is determined by the input parameter *mc* which represents the number of new word will be added. When the new vocabulary is used for segmentation, many bigrams and trigrams will be recognized as a integrated word which will increase the average word length of the segmented corpus.

The basic information of the 10 segmented corpus to train language models is summarized in Table 1.

Table 1. *The average word length and vocabulary size of different language model training corpuses*

corpus no	avg word length	vocab size
1	1.0	6k
2	1.44	11k
3	1.62	16k
4	1.72	21k
6	1.79	25k
7	1.85	30k
8	1.89	35k
9	1.93	39k
10	1.97	44k

[2] Chinese version twitter

Algorithm 1 Generating segmentation dictionary

INPUT1 held out corpus *hc*
INPUT2 merge count *mc*
INPUT3 number of generated dictionaries *num*
OUTPUT generated dictionaries *vocabs*
segment *hc* by characters and get the segmented data *sc*
for i=0;$i < num$;i++ **do**
  state trigram count *tc* from *sc*
  for each element e in *tc* **do**
    if e is trigram **then**
      merge the e to unigram e_u
      remove e and add e_u to *tc*
      for each bigram b in e **do**
        if b is in *tc* **then**
          let b.count-=e.count
        end if
      end for
    end if
    if e is bigram **then**
      merge the e to unigram e_u
      remove e and add e_u to *tc*
    end if
  end for
  sort *tc* order by count
  let $c = 0$
  let *vocab* be the dictionary for new segmentation
  for each element e in *tc* **do**
    if e is merged by trigram or bigram **then**
      c+=1
    end if
    if $c > mc$ **then**
      break
    end if
    add e to *vocab*
  end for
  add *vocab* to *vocabs*
  segment *hc* using BMM algorithm with *vocab* and get the segmented data *sc*
end for

3.1 Correlation between CER and ppl

With the trigram language models trained on the 10 different text corpra, normal word-based perplexities were calculated and CERs were generated after full decoding on the acoustic data. The correlation between CER and word-based perplexity *ppl* is shown in figure 1

It can be seen that, there is no positive correlation between CER and *ppl*. To quantify the correlation between different metrics with character error rate, linear correlation coefficient(or Pearson coefficient) was calculated to measure

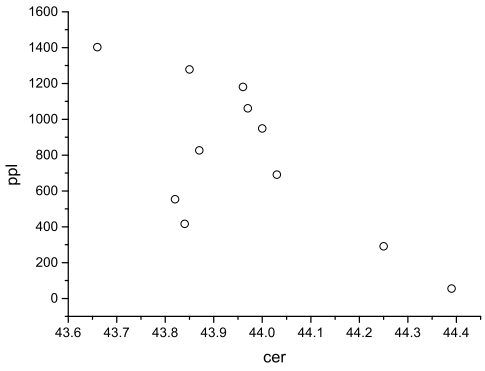

Fig. 1. *Correlation of CER and ppl when segmentation strategy varies*

the degree of linear correlation. The Linear correlation coefficient of CER and *ppl* is -0.70. The coefficient of CER and log(*ppl*) is -0.79. All of the correlation coefficients are negative in this experiment. It is inconsistent with the expectation that CER is positively correlated with *ppl*.

To further investigated the issue, another experiment has been performed. Here, the segmentation strategy is fixed and the size of corpus to train language model varied from 10M to 100M. The result is shown in figure 2

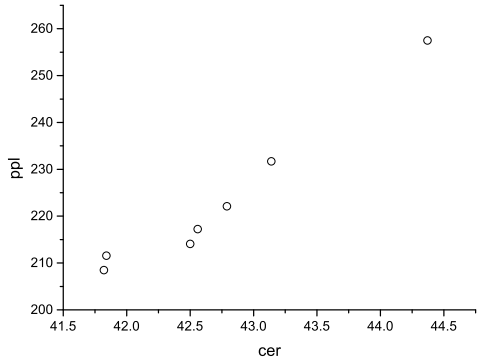

Fig. 2. *The correlation of CER and ppl when size of training corpus varies*

CER and *ppl* correlates quite well in this experiment, which is a consistent observation as the previous work on perplexity. From the above two experiments, the correlation between CER and *ppl* varies from positive to negative which is

quite inconsistent, and therefore, we conclude *ppl* is incompetent for evaluating CER.

3.2 Correlation between CER and cppl

The setup of this experiment was same as the previous experi- ment except equation (2) was used to calculate the *cppl* instead of *ppl*. The correlation between *cppl* and CER when segmentation strategies varies is shown in figure 3 and when the corpus size changes, the correlation is shown in figure 4

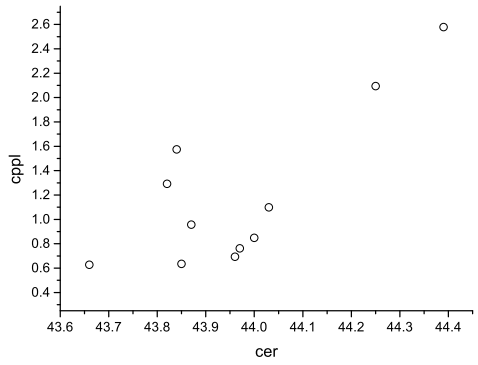

Fig. 3. *The relationship between CER and cppl when segmentation strategy varies*

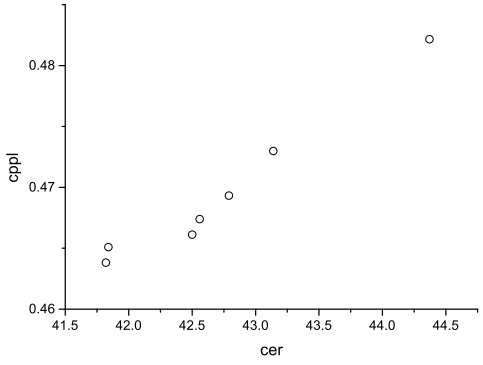

Fig. 4. *The relationship between CER and cppl when size of training corpus varies*

The Linear correlation coefficient in figure 3 is 0.78 which is higher than the absolute value of *ppl* but a little lower than absolute value of log(*ppl*) and in figure 4 is 0.97 which is equal to *ppl*. In both figures, *cppl* shows a positive correlation with CER.

Experiment testing the performance of $cppl_2$ which considers the influence of word length was also performed. Equation (3) was used to calculate the $cppl_2$. The correlation result was similar to *cppl* but with a litter increase in correlation coefficient. The improvement is shown in table 2

Table 2. *Linear correlation coefficients comparison*

measure	wrd seg vary	corpus size vary
ppl	-0.70	0.97
cppl	0.78	0.97
$cppl_2$	0.80	0.98

This comparison showed that considering word length slightly improved the correlation coefficients, but this influence was very tiny compared to the effect caused vocabulary size change.

In the above experiments, it has been shown that perplexity is incompetent predicting language models quality for character based languages. One main reason is that perplexity is not only affected by the probability distribution of language model but also by the scale of vocabulary size. Since it is only the probability distribution deciding the language models performance in speech recognition, the influence of vocabulary size will observably interfere the correlation. Therefore, the proposed metric *cppl* empirically neutralizing this effect retained inconsistence with character error rate in the two experiments.

The experiment about $cppl_2$ shows that taking the word length into consideration does not have apparent improvement to the evaluation. It infers that word length may not be as important to the correlation as we thought. This is because by our analysis, the influence of vocabulary composition varies is very complex and length only describing a simple physical attribute of a word without reaching its probability attributes or its character element is inadequate to neutralizing the effect caused by vocabulary change. Therefore, our future work will focuses on the further investigation of the influence caused by vocabulary composition, more information and more complex model about the word in vocabulary will be considered.

4 Conclusion

In this paper, perplexity is shown incompetent to predict CER for character based language, since the segmentation strategies which change the vocabulary composition will distinctly affect the evaluation of perplexity. To address this

problem, word-based perplexity has been extended. A new metric taking vocabulary size into consideration is proposed. It is shown to successfully neutralize the influence of vocabulary change and is more robust. Length of word in vocabulary is also considered while it is proved having little effect about the final correlation. The main factor about the influence of the vocabulary composition should be further investigated.

References

1. Lalit R Bahl, Frederick Jelinek, and Robert L Mercer. A maximum likelihood approach to continuous speech recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (2):179–190, 1983.
2. Stanley F Chen, Douglas Beeferman, and Roni Rosenfield. Evaluation metrics for language models. 1998.
3. Philip Clarkson, Tony Robinson, et al. Towards improved language model evaluation measures. *Procc of EUROSPEECH*, 99:1927–1930, 1999.
4. Akinori Ito, Masaki Kohda, and Mari Ostendorf. A new metric for stochastic language model evaluation. In *Proceedings of the Sixth European Conference on Speech Communication and Technology*, volume 4, pages 1591–1594, 1999.
5. Fred Jelinek, Robert L Mercer, Lalit R Bahl, and James K Baker. Perplexitya measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63, 1977.
6. Dietrich Klakow and Jochen Peters. Testing the correlation of word error rate and perplexity. *Speech Communication*, 38(1):19–28, 2002.
7. Andreas Stolcke et al. Srilm-an extensible language modeling toolkit. In *Proceedings of the international conference on spoken language processing*, volume 2, pages 901–904, 2002.
8. Steve Young, Gunnar Evermann, Dan Kershaw, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev, and Phil Woodland. The htk book (for htk version 3.2). *Cambridge University Engineering Department*, 2002.