TOWARDS IMPROVED LANGUAGE MODEL EVALUATION MEASURES

Philip Clarkson and Tony Robinson

Cambridge University Engineering Department,
Trumpington Street, Cambridge CB2 1PZ, UK.
{prc14,ajr}@eng.cam.ac.uk

ABSTRACT

Much recent research has demonstrated that the correlation between a language model's perplexity and its effect on the word error rate of a speech recognition system is not as strong as was once thought. This represents a major problem for those involved in developing language models. This paper describes the development of new measures of language model quality. These measures retain the ease of computation and task independence that are perplexity's strengths, yet are considerably better correlated with word error rate. This paper also shows that mixture-based language models are improved by applying interpolation weights which are optimised with respect to these new measures, rather than a maximum likelihood criterion.

1. INTRODUCTION

For many years, perplexity [2] has been the measure by which the quality of language models has been evaluated. There are good reasons for this; it is a simple, well-understood measure that fits into the maximum likelihood framework, and which can be computed quickly. However, recent work on language modelling has demonstrated that the correlation between a language model's perplexity and its effect on the word error rate (WER) of a speech recognition system is not as strong as was once thought. There are many examples of cases in which a language model has a much lower perplexity than the baseline model, but does not result in a reduction in WER, and often results in a degradation in recognition accuracy (see, for example, [6]). This paper discusses alternative measures to perplexity. These measures are better correlated with WER, yet retain the ease of computation and task independence that are perplexity's strengths.

The calculation of perplexity is based on the probability that the language model assigns to some test text. If the language model is successful it will assign a high probability to this test text, with the result that the language model will have a low perplexity. Thus perplexity is based solely on the probabilities of the words which actually occur in the test text. It disregards the way in which the remaining probability mass is distributed over the alternative words, which may be competing with the correct word in the decoder of a speech recognition system. We will show that this additional information is important, and that including it in our methods of evaluating language models leads to measures which are better correlated with WER. Finally, we will show how the information from the new measures can be used to select more appropriate interpolation weights for mixture-based language models. Such interpolation weights lead to a small, but statistically significant improvement in WER as compared to the original maximum likelihood weights.

2. MOTIVATION – SAME-PERPLEXITY LANGUAGE MODELS

Previous work has shown that it is possible to construct mixture-based language models which have lower perplexities than the baseline model, but result in higher error rates [6]. If one reduces the amount of training data available to these mixture-based models, their perplexities and the resulting WERs are likely to be increased. Indeed, if one selects the appropriate amounts of training data for each language model, it is possible to generate mixture-based models that have the same perplexity as the baseline trigram model, but result in a higher WER. These models will therefore differ in some important way, despite having identical perplexities. By investigating the manner in which the models differ it is to be hoped that some light might be shed on the discrepancy between WER and perplexity.

A mixture-based language model with the same perplexity as the baseline trigram model was constructed, and its effect on WER was evaluated. The results are summarised in Table 1.

Model	% Data	Perplexity	WER
Baseline	100	134.4	37.9
Mixture-based	42	134.4	39.3

Table 1: *Summary of same-perplexity language models*

Consider a function $f_{\mathcal{M}}(x)$ which indicates the frequency with which words are assigned a log probability of x by the language model $\mathcal{M}$. The value μ of the mean log probability of the words in the test text can be computed given the values of this function:

$$\mu = \int_{-\infty}^{0} x f_{\mathcal{M}}(x) dx. \quad (1)$$

Since perplexity is based on the mean log probability of the words in the test text w_1^n:

$$PP = P(w_1^n)^{-1/n} = e^{-\frac{1}{n}\sum_{i=1}^{n}\log[P(w_i|w_1^{i-1})]} = e^{-\mu}, \quad (2)$$

$f_{\mathcal{M}}(x)$ contains at least as much useful information as the value of perplexity, and possibly somewhat more.

The function $f_{\mathcal{M}}(x)$ was estimated by partitioning the probability range into 100 bins which are spaced equally in the log domain. For each language model, the number of words in the test text which have language model probabilities in each bin was computed. Figure 1 shows the resulting functions for the mixture-based model and the baseline trigram model.

The key observation to make from this graph is that the probability distribution curve for the mixture-based model is almost identical to that of the baseline trigram model. These models result in different WERs, yet there is not sufficient information in the probabilities of the words in the test text to

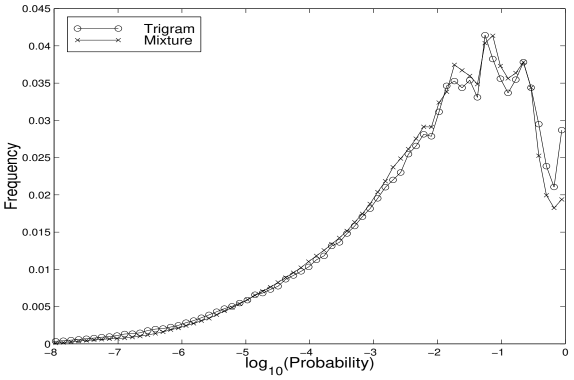

Figure 1: *Probability distribution graph. Comparison of* $f_{trigram}(x)$ *and* $f_{mixture}(x)$

distinguish between them. It seems then, that the information needed to discriminate between these models is not contained in the probabilities of the words which actually occur in the test text. Rather, it seems that one must consider the way in which the remaining probability mass is distributed over the *alternative* words, which will compete with the correct word in the decoder of a speech recogniser. It is this observation which motivates the work in the rest of this paper.

3. EXPERIMENTAL CONDITIONS

In order to investigate the correlation between WER and new language model evaluation measures, it is clearly necessary to have a large set of language models upon which to base the experiments.

A set of 50 language models was constructed. These models comprise bigram, trigram, mixture- and cache-based models, which have been trained on either the broadcast news corpus [10] or British National Corpus [4]. Different quantities of the training corpora were used to train each language model, and various cutoffs were applied. The WER which results from using each language model was computed.

This work was carried out for the broadcast news task. All WER results are based on the six shows of the 1996 Hub 4 development test. The results were generated by rescoring lattices produced by a simplified version of the 1996 Hub 4 Abbot system [7].

The strength of the correlation between the new evaluation schemes and WER was evaluated using three correlation coefficients: the Pearson product-moment correlation coefficient r, the Spearman rank-order correlation coefficient r_s, and the Kendall rank-order correlation coefficient T. Detailed information on these correlation coefficients can be found in many standard statistics books (see, for example, [3] for the Pearson product-moment correlation coefficient and [11] for the Spearman and Kendall rank-order correlation coefficients).

The language model evaluation schemes which will be described in this paper are evaluated with respect to the reference transcription of the broadcast news shows upon which the WER scores are based, rather than the much larger language model test text. This circumvents the problems caused by any potential mismatch between the language model test text and the recognition task itself.

4. MEASURES OF LANGUAGE MODEL QUALITY

4.1. Perplexity

Perplexity has been used as a method of evaluating language models for several years and in this paper it serves as the baseline measure. The correlation of WER and perplexity on the test set of fifty language models was evaluated according to the three correlation coefficients described above. The results are shown in Table 2.

	r	r_s	T
Perplexity	0.955	0.955	0.840

Table 2: *Correlation of perplexity with WER*

4.2. Rank

Perplexity measures the language model's success according to the probability it assigns to each of the words in the test text. An alternative approach is to evaluate the language model according to the proportion of words which have a higher probability than the target word at each point in the test text. By so doing, the measure encodes the quality of the target word's prediction relative to the other words with which it will be competing within the speech decoder.

The *rank* of the target word, given a particular history, is defined as the word's position in an ordered list of the word probabilities. For each language model, the rank of each word in the reference file was calculated. Hence the *mean log rank* of each language model was computed, and the strength of the correlation between this measure and WER evaluated. The results are shown in Table 3.

	r	r_s	T
Mean log rank	-0.967	-0.957	-0.846

Table 3: *Correlation of mean log rank with WER*

These results indicate that the mean log rank is at least as well correlated with WER as perplexity is.

4.3. Entropy

Given a particular word history w_1^i, the *entropy* of the distribution in bits is given by

$$H = -\sum_{w \in \mathbf{V}} \hat{P}(w \mid w_1^i) \log_2 \hat{P}(w \mid w_1^i). \tag{3}$$

Therefore, the entropy is related to the expected value of the log probability given the word history in the following way:

$$E(\log_2 \hat{P}(w \mid w_1^i)) = -H. \tag{4}$$

Since perplexity is based on the mean log probability of words in the test text, and the log probability and entropy are related in this way, the measure that was developed was based on the mean entropy over the test text.

The strength of the correlation between this measure and the WER was calculated, and the results are shown in Table 4.

These results show that the entropy is not as well correlated with WER as some of the other proposed language model evaluation measures. In particular, it is inferior to perplexity. However, there is a clear correlation displayed by these results, so

	r	r_s	T
Mean entropy	-0.799	-0.792	-0.602

Table 4: *Correlation of mean entropy with WER*

some useful information is certainly present in this measure. In Section 5 the manner in which this information can be more fruitfully used will be investigated.

4.4. Low probability estimates

The set of fifty language models makes it possible not only to investigate new language model evaluation measures, but also to evaluate previously proposed ones. In [1] language models are compared according to the number of words in the test text which receive probability estimates below a certain threshold. The premise is that recognition errors are strongly correlated with very low language model estimates. Therefore, the correlation between WER and measures of the form

$$L(x) = \# \text{ of words } w \text{ in test text such that } \hat{P}(w) \leq x \quad (5)$$

was investigated. The results are shown in Table 5.

	r	r_s	T
$L(2^{-5})$	-0.919	-0.893	-0.726
$L(2^{-10})$	-0.915	-0.917	-0.768
$L(2^{-15})$	-0.833	-0.817	-0.640
$L(2^{-20})$	-0.646	-0.544	-0.388

Table 5: *Correlation of low probability estimates with WER*

These results indicate that the number of very low probability estimates in the test set is not as well correlated with WER as some of the other measures investigated in this paper. Certainly the measure used in [1], $L(2^{-15})$, is much less well correlated with WER than perplexity is for the set of language models investigated here.

5. COMBINING INFORMATION SOURCES

5.1. Correlation between measures

The probability, rank and entropy were computed for each of the words in the test text according to the baseline broadcast news language model. The value of r_s for each pair of features was then calculated. The results are shown in Table 6.

Feature 1	**Feature 2**	r_s
Probability	Rank	-0.985
Probability	Entropy	-0.378
Rank	Entropy	0.381

Table 6: *Correlation between language model features*

These results clearly show that there is a very strong correlation between a word's probability and its rank. That is, the two features provide very similar information. Conversely, there seems to be much less correlation between a word's probability and the entropy of the distribution at that point in the test text. Thus, the information provided by these features is, in some sense, orthogonal. Given that both features provide information which is useful in predicting WER, it seems that if the information sources can be combined, a superior measure of language model quality would result.

5.2. Combination of log probability and entropy

In order to develop measures of language model quality which are better correlated with WER, the information from the probability of the target word and the entropy at each point in the test text was combined.

Since the entropy H is the negative value of the expected log probability of the next word, the values that were combined were the log probability of the target word $\log_2(\hat{P}(w_i \mid w_1^{i-1}))$ and the negative entropy:

$$-H(w_1^{i-1}) = \sum_{w \in \mathbf{V}} \hat{P}(w \mid w_1^{i-1}) \log_2(\hat{P}(w \mid w_1^{i-1})). \quad (6)$$

These values were combined using linear interpolation, both in the log domain, leading to a measure which will be referred to as $C_{\text{log}}(\lambda)$ and after converting back from the log domain, giving a measure called $C_{\text{exp}}(\lambda)$. If the test text is w_1^n, then these measures can be expressed as

$$C_{\text{log}}(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \Big[-\lambda H(w_1^{i-1}) + (1-\lambda) \log_2(\hat{P}(w_i \mid w_1^{i-1})) \Big] \quad (7)$$

and

$$C_{\text{exp}}(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \Big[\lambda 2^{-H(w_1^{i-1})} + (1-\lambda) \hat{P}(w_i \mid w_1^{i-1}) \Big]. \quad (8)$$

The values of these measures were computed for a range of values of λ. The strength of the correlation between the resulting measures and WER was computed. The results are shown in Table 7.

	r	r_s	T
$C_{\text{log}}(0)$ (Baseline)	0.966	0.955	0.840
$C_{\text{log}}(0.05)$	0.969	0.960	0.853
$C_{\text{log}}(0.1)$	0.971	0.965	0.868
$C_{\text{log}}(0.2)$	0.971	0.964	0.863
$C_{\text{log}}(0.3)$	0.964	0.957	0.837
$C_{\text{exp}}(0.001)$	0.970	0.962	0.853
$C_{\text{exp}}(0.002)$	0.970	0.963	0.856
$C_{\text{exp}}(0.01)$	0.965	0.955	0.842
$C_{\text{exp}}(0.02)$	0.959	0.952	0.835

Table 7: *Correlation of combined measures with WER*

These results show that combining the information from the two sources leads to language model evaluation measures which are better correlated with WER than either of the individual measures. In particular, $C_{\text{log}}(0.1)$ performs considerably better than perplexity in this respect. This clearly demonstrates that information concerning the manner in which the probability mass is distributed over non-target words is useful in predicting WER.

6. APPLICATION TO LANGUAGE MODEL DEVELOPMENT

In [5] and [6], mixture-based models were described. Small topic- or style-homogeneous language models were constructed, and their output was combined using interpolation weights chosen to satisfy a maximum likelihood criterion (and hence to minimise perplexity). This led to models which had considerably lower perplexities than the baseline trigram model, but no decrease in WER. Furthermore, in [6], it was shown that even choosing the interpolation weights to maximise the likelihood of the reference transcription of the speech being decoded led to no improvement in WER.

This paper has described the development of measures of language model quality which correlate better with WER than perplexity does. Since the ultimate aim of mixture-based models is to reduce WER, the interpolation weights should be chosen with this in mind. Therefore, we attempt to choose interpolation weights which are optimised with respect to our new measures.

The probability estimate from a mixture-based model is simply a linear combination of the probability estimates from a set of component models [5]:

$$\hat{P}(w_i \mid w_1^{i-1}) = \sum_j \lambda_j \hat{P}_{\mathcal{M}_j}(w_i \mid w_1^{i-1}). \tag{9}$$

The aim, therefore, is to select interpolation weights λ_j in order to maximise a more appropriate measure. In this case, we aim to maximise $C_{\log}(0.1)$. This technique was applied to generate new interpolation weights for mixture-based models based on 30 mixture components trained on both the broadcast news corpus and the British National Corpus. In both cases, supervised adaptation (in which the interpolation weights are chosen based on the reference transcription) and unsupervised adaptation (where the interpolation weights are based on the first recognition pass) were applied. Lattice rescoring experiments were carried out using the resulting interpolation weights. The results are presented in Table 8, and are compared with the results of using the conventional maximum likelihood weights.

Training text	Adaptation	Weights	
		ML	New
Broadcast news	Unsupervised	38.2%	38.1%
Broadcast news	Supervised	38.0%	37.9%
BNC	Unsupervised	42.3%	41.9%
BNC	Supervised	41.8%	41.6%

Table 8: *Comparison of WERs achieved by maximum likelihood ("ML") and WER optimised interpolation weights ("New")*

These results show that the new weights chosen to optimise $C_{\log}(0.1)$ perform consistently better than the the old maximum likelihood weights. While the difference in performance is small, the improvements are consistent, and the overall difference between the WER optimised and maximum likelihood weights is statistically significant at the 1% level according to the matched pairs sentence segments word error test [9].

It should be noted that the ability to choose more appropriate interpolation weights does not merely allow for improved adaptive language models to be created. Any language model in which multiple information sources are combined using weights chosen to minimise perplexity can be improved using this method. Such techniques are currently used in many state-of-the-art systems for the transcription for broadcast news. In the 1998 broadcast news systems from both LIMSI [8] and Cambridge University's HTK group [12], individual language models are constructed for each source of training data, and combined using linear interpolation with weights chosen to minimise perplexity. In such cases, it is to be expected that WER could be reduced by the application of WER optimised interpolation weights.

7. CONCLUSIONS

This paper has described the development of new measures of language model quality. It has been shown that some measures based on the entire probability distribution (rather than simply the probability of the target words) are better correlated with WER than perplexity is. Moreover, it has been shown that mixture-based language models are improved by applying interpolation weights which are optimised with respect to these new measures.

ACKNOWLEDGEMENTS

The authors wish to thank Roni Rosenfeld, Kristie Seymore and Stan Chen of Carnegie Mellon University for many useful and enlightening discussions.

REFERENCES

[1] L.R. Bahl, P.F. Brown, P.V. de Souza, and R.L. Mercer. A Tree-Based Statistical Language Model for Natural Language Speech Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(7), 1989.

[2] L.R. Bahl, F. Jelinek, and R.L. Mercer. A Maximum Likelihood Approach to Continuous Speech Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(2), 1983.

[3] M. Berenson, D. Levine, and D. Rindskopf. *Applied Statistics. A first course.* Prentice-Hall International, 1988.

[4] L. Burnard. *Users Reference Guide for the British National Corpus*. Oxford University Computing Services, May 1995.

[5] P.R. Clarkson and A.J. Robinson. Language Model Adaptation using Mixtures and an Exponentially Decaying Cache. In *Proceedings IEEE ICASSP*, Munich, Germany, 1997.

[6] P.R. Clarkson and A.J. Robinson. The Applicability of Adaptive Language Modelling for the Broadcast News Task. In *Proceedings ICSLP*, Sydney, Australia, 1998.

[7] G. Cook, D. Kershaw, J. Christie, and A. Robinson. The Transcription of Broadcast Television and Radio News: The 1996 Abbot System. In *ARPA Spoken Language Technology Workshop*, 1997.

[8] J-L. Gauvain, L Lamel, and M. Jardino. The LIMSI 1998 Hub-4E Transcription System. In *Proceedings of DARPA Broadcast News Workshop*, Herndon, Virginia, USA, 1999.

[9] L. Gillick and S.J. Cox. Some Statistical Issues in the Comparison of Speech Recognition Algorithms. In *Proceedings IEEE ICASSP*, Glasgow, UK, 1989.

[10] D. Graff. The 1996 Broadcast News Speech and Language-Model Corpus. In *Proceedings ARPA Workshop on Human Language Technology*, 1996.

[11] S. Siegel and N. Castellan. *Nonparametric Statistics for the Behavioural Sciences*. McGraw Hill, 2nd edition, 1988.

[12] P. Woodland, T. Hain, G. Moore, T. Niesler, D. Povey, A. Tuerk, and E. Whittaker. The 1998 HTK Broadcast News Transcription System: Development and Results. In *Proceedings of DARPA Broadcast News Workshop*, Herndon, Virginia, USA, 1999.