THE APPLICABILITY OF ADAPTIVE LANGUAGE MODELLING FOR THE BROADCAST NEWS TASK

Philip Clarkson *Tony Robinson*

Cambridge University Engineering Department,
Trumpington Street, Cambridge CB2 1PZ, UK.

ABSTRACT

Adaptive language models have consistently been shown to lead to a significant reduction in language model perplexity compared to the equivalent static trigram model on many data sets. When these language models have been applied to speech recognition, however, they have seldom resulted in a corresponding reduction in word error rate. This paper will investigate some of the possible reasons for this apparent discrepancy, and will explore the circumstances under which adaptive language models can be useful. We will concentrate on cache-based and mixture-based models and their use on the Broadcast News task.

1. INTRODUCTION

The performance of an automatic speech recognition system can depend critically on the suitability of its language model. For example, a system trained to recognise speech read from the Wall Street Journal will be equipped with a language model trained on many millions of words from previous editions of the newspaper, and will perform very well on its specified task. However, when presented with speech of a different style, or on a topic not commonly discussed within the Wall Street Journal, such a system will often perform very badly.

The Broadcast News task is much more varied than that of recognising text read from the Wall Street Journal. There is a much greater variety in both subject matter and linguistic style, as the speech might be read text from an auto-cue one minute, and spontaneous answers to interview questions the next. This makes the language modelling task much more challenging.

A 130 million word corpus of transcribed news broadcasts exists, and this enables us to train a language model which is, in general, appropriate. However, we would prefer to have a language model which is less general, and always appropriate to the particular topic and style of speech which is currently being used. This is the motivation for adaptive language modelling – if we can tune our language model to these local fluctuations in linguistic style then we believe that we will improve recognition performance.

There has been a good deal of research into adaptive language modelling for many speech recognition tasks – including Broadcast News. Much of this work has reported that the adaptive language models result in a large reduction in the language model perplexity compared to the baseline trigram model, but do not result in an equivalent reduction in word error rate (see, for example [6, 8, 9]). The focus of this paper is to investigate possible reasons for this apparent discrepancy.

We will describe two simple adaptive language models, and show that they do indeed lead to a significant perplexity reduction, but not a reduction in the word error rate of a speech recogniser. We will then describe supervised adaptation experiments which investigate the hypothesis that the poor performance of the adaptive language models is due to the errorful nature of the initial transcription. Finally we investigate whether the lack of improvement in recognition performance is due to the baseline language model being well adapted to begin with, by replacing it with one trained on more general text.

2. TWO ADAPTIVE LANGUAGE MODELS

Two adaptive language models, a cache-based model and a mixture-based model, will be used in the experiments presented in this paper. Neither model is especially sophisticated. However, as both lead to significant perplexity reductions, with little or no improvement in word error rate, they are adequate to illustrate the points made by this paper.

2.1. Cache-Based Model

The cache-based model [7] is based on the premise that words which have occurred recently are more likely to re-occur than a static language model would predict. Therefore the most recent words are stored in a cache, and their language model probabilities are boosted relative to those words which do not appear in the cache. Typically this is achieved by linearly interpolating the baseline N-gram language model with a cache-based component:

$$P(w_i) = (1-\lambda)P_{\text{trigram}}(w_i \mid w_{i-2}w_{i-1}) + \lambda P_{\text{cache}}(w_i) \quad (1)$$

where $P_{\text{cache}}(w_i)$ will be high if the word is contained in the cache, and zero otherwise.

When we compute the perplexity of the cache-based language models, a word's cache-based probability $P_{\text{cache}}(w_i)$ is computed simply as its frequency in the previously seen portion of the current article. The value of λ was chosen to maximise the likelihood of a portion of held-out text.

The speech recognition experiments described here compute the cache-based probabilities slightly differently. Since we do not apply the cache until we have already generated a first-pass transcription (as we use the cache-based model to rescore lattices), we can use information from the future portion of the article, as well as the past. Therefore we compute the cache-based probabilities such that they are equal to the word frequencies of the whole first pass transcription of the current article (with the segment currently being decoded removed). Different values of the interpolation parameter λ were investigated.

2.2. Mixture-Based Model

The mixture-based language model used in this paper has the same structure as the model described in [3]. The training text is split up into articles, and these are clustered into a set of components using a k-means style clustering algorithm. A standard backoff trigram language model is then constructed for each component, as well as a "full" language model which is trained on the entire training text. These language models are then interpolated according to interpolation weights which are chosen on an article-by-article basis to maximise the likelihood of some held-out text (for perplexity experiments) or of the first-pass transcription (when applied to speech recognition).

Thus we have:

$$P(w_i|w_{i-2}w_{i-1}) = \sum_{j=0}^{k} \lambda_j P_{\text{model } j}(w_i|w_{i-2}w_{i-1}) \quad (2)$$

where "model 0" is the full language model, and k represents the number of components into which the training text is clustered.

3. BASELINE RESULTS

The effect on perplexity and word error rate of both adaptive language models described in the previous section was investigated. The baseline language model was a standard back-off trigram model trained on the 130 million word Broadcast News corpus, with a 65,000 word vocabulary and bigram and trigram cutoffs of 1.

The perplexity results are based on the 17 million words of held-out language model text from the Broadcast News corpus. Of this, 5 million words are used to estimate appropriate values for the interpolation weights (the global value of λ in (1) and the article-specific values of λ_j in (2)), and the remaining 12 million are used for the actual perplexity computation. The word error rate results are based on the six shows of the 1996 Hub 4 development test, and were generated by rescoring lattices produced by a simplified version of the 1996 Hub 4 Abbot system [4]. The lattice word error rate (i.e. the word error rate which would result if we chose the path through each lattice with the least errors) for these lattices was 7.0%.

The cache-based model resulted in a 12% reduction in perplexity (with the optimal choice for λ being 0.09), and the mixture-based model reduced perplexity by up to 13% as the number of mixture components was increased to 50 (see Table 1).

Note that there is little reduction in perplexity as we increase the number of mixture components above 30, but the increase in the computation time and memory required to train and use the model is significant. For this reason, when we used mixture-based models in speech recognition experiments, we used a model with 30 mixture components.

Model	Perplexity
Baseline	134.4
Cache-based	118.9
Mixture-based (10 components)	121.3
Mixture-based (20 components)	119.3
Mixture-based (30 components)	117.9
Mixture-based (40 components)	117.1
Mixture-based (50 components)	116.7

Table 1: The effect on perplexity of cache- and mixture-based language model adaptation

The results in Table 2 show that the cache-based model does not improve speech recognition performance, and that the mixture-based model actually degrades performance, despite the substantial reduction in perplexity that both models yield. The perplexity reductions indicate that we are, at some level, modelling language better, and yet we are not observing the anticipated reductions in word error rate. The next sections explore some of the possible reasons for this apparent discrepancy.

Model	Word Error Rate
Baseline	37.9%
Cache-based ($\lambda = 0.05$)	37.9%
Cache-based ($\lambda = 0.1$)	38.0%
Mixture-based (30 components)	38.2%

Table 2: The effect on word-error-rate of cache- and mixture-based language model adaptation

4. SUPERVISED ADAPTATION

Most adaptive language models – including the two described in this paper – use an initial transcription as the basis for the adaptation. This initial transcription is likely to contain errors (if it didn't then there would be no need to adapt our language model), and this could obviously affect the quality of the adaptation. This appears to represent a major problem for adaptive language modelling. A good initial transcription will benefit little from language model adaption (and would probably indicate that the baseline language model was well adapted in the first place), whereas a poor initial transcription will contain so much noise, and so little useful information, that successful adaptation will prove difficult. This problem is not insurmountable. One could, for example, use confidence measures and base the adaptation only on the areas of the initial transcription in which we have high confidence. But the first step is to investigate how much the errorful nature of the initial transcription actually affects performance. We have done this by performing supervised adaption experiments in which the adaptation is based not on the first-pass output of the decoder, but on the reference transcription.

We conducted supervised adaptation experiments using both the cache- and mixture-based models. The cache-based model used the reference transcription to estimate the cache-based component's probabilities, and the mixture-based model estimated the interpolation weights using the reference transcription. Table 3 shows the effect of using supervised adaptation as compared to unsupervised adaptation. It can be seen that while the supervised adaptation performs slightly better than unsupervised adaptation, it is not a major effect. It therefore seems that the errorful nature of the initial transcription is not the primary reason for the poor performance of these adaptive language models. This isn't to say that it isn't a factor at all; merely that there are other issues which need to be addressed first.

	Word Error Rate	
Model	**Unsupervised**	**Supervised**
Baseline	37.9%	
Cache ($\lambda = 0.05$)	37.9%	37.6%
Cache ($\lambda = 0.1$)	38.0%	37.7%
Mixture	38.2%	38.0%

Table 3: Comparison of supervised and unsupervised adaptation

5. A MORE GENERAL BASELINE LANGUAGE MODEL

So far we have shown that our attempts to adapt the baseline Broadcast News language model to the individual sub-topics and styles of discourse within the Broadcast News task have not had positive results in terms of recognition performance. However, our baseline language model is already well adapted to the target domain (it is, after all, trained on transcribed Broadcast News shows); might there be something to be gained from language model adaptation if our baseline language model was not so specific to the target domain?

The British National Corpus (BNC) [1] is a 100 million word corpus of British English taken from very diverse domains (novels, advertising pamphlets, transcribed spontaneous conversation, etc.). We trained a trigram language model on this corpus, and applied the cache- and mixture-based adaptation approaches to this model in order to investigate whether a more general language model might benefit more from language model adaptation.

Table 4 shows that the general BNC language model has a much higher perplexity than the Broadcast News language model, but that the adaptation techniques reduce the perplexity of the BNC language model by more than they do for the Broadcast News model. Cache-based adaptation reduces the perplexity of the BNC model by 22%, and mixture-based adaptation by 16%. Table 5 shows that, in this scenario, the reductions in perplexity do translate to reductions in word error rate, with both forms of adaptation leading to reductions in word error rate, even for unsupervised adaptation. This indicates that the usefulness of these language model adaptation techniques varies according to the baseline language model, and seems to suggest that they are more likely to be of use in situations where the baseline language model is less well suited to the target domain.

Training text	**Adaptation**	**Perplexity**
Broadcast News	None	134.4
BNC	None	277.5
BNC	Cache	216.5
BNC	Mixture	233.6

Table 4: The effect on perplexity of a more general language model

Training text	**Adaptation**	**Word Error Rate**
Broadcast News	None	37.9%
BNC	None	42.9%
BNC	Cache (unsupervised)	42.1%
BNC	Cache (supervised)	41.3%
BNC	Mixture (unsupervised)	42.3%
BNC	Mixture (supervised)	41.8%

Table 5: The effect on word error rate of a more general language model

6. CONCLUSIONS AND FUTURE WORK

This paper has described two simple adaptive language models, and shown that while they lead to substantial reductions in perplexity over the baseline Broadcast News language model, they do not result in improved recognition performance. The reductions in perplexity indicate that in some way these adaptive models are better at *modelling language*. However, these results, as well as those given in several other papers [6, 8, 9] show that even fairly large reductions in perplexity are no guarantee of a reduction in word error rate. We have shown that this is not primarily due to the errorful nature of the first-pass transcription. In addition, we have shown that these adaptive language models can reduce error rates in situations where the baseline language model is not initially well adapted to the target domain.

Work which has researched alternatives to perplexity [2, 5] has reported measures which are better predictors of word error rate than perplexity, although as these measures become more complicated, it seems that one might simply be better abandoning perplexity and similar measurements altogether, and simply evaluating word error rate directly. One cannot make strong claims about the likely effect on word error rate of a language model simply by considering the probabilities of "correct" word strings as perplexity and similar measures do. It is important to also consider the language model's effect on the relative probabilities of alternative word strings which the decoder might consider.

It is our belief that the potential usefulness of a modified language model can only be evaluated by considering the errors that were made by the recognition system with the initial language model. A reduction in perplexity will not, for example, improve recognition performance if it comes about simply by boosting the probabilities of words which were correctly recognised in the first pass.

The next stage in this research is to attempt to answer the following questions:

- Which are the words for which the probabilities differ most between the baseline and the adapted language model?
- Which are the words that our recognition system is most prone to make errors on?
- How much intersection is there between these sets, and what can that tell us about the likely effect on word error rate of our adaptive language model?

7. ACKNOWLEDGMENTS

Philip Clarkson is supported by an EPSRC studentship.

8. REFERENCES

1. L. Burnard (editor). *Users Reference Guide for the British National Corpus*. Oxford University Computing Services, May 1995.
2. S. Chen, D. Beeferman, and R. Rosenfeld. Evaluation Metrics for Language Models. In *Proceedings of the ARPA Workshop on Human Language Technology*, 1998.
3. P.R. Clarkson and A.J. Robinson. Language Model Adaptation using Mixtures and an Exponentially Decaying Cache. In *Proceedings IEEE ICASSP*, 1997.
4. G.D. Cook, D.J. Kershaw, J.D.M. Christie, C.W. Seymore, and S.R. Waterhouse. Transcription of Broadcast Television and Radio News: The 1996 Abbot System. In *Proceedings IEEE ICASSP*, 1997.
5. R. Iyer, M. Ostendorf, and M. Meteer. Analysing and Predicting Language Model Improvements. In *Proceedings IEEE Workshop on Speech Recognition and Understanding*, 1997.
6. R. Kneser, J. Peters, and D. Klakow. Language Model Adaptation Using Dynamic Marginals. In *Proceedings Eurospeech*, 1997.
7. R. Kuhn and R. De Mori. A Cache-Based Natural Language Model for Speech Reproduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(6):570–583, 1990.
8. S. Martin, J. Liermann, and H. Ney. Adaptive Topic-Dependent Language Modelling Using Word-Based Varigrams. In *Proceedings Eurospeech*, 1997.
9. K. Seymore and R. Rosenfeld. Using Story Topics for Language Model Adaptation. In *Proceedings Eurospeech*, 1997.