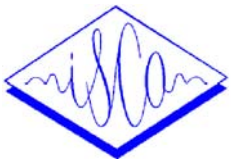

EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

HYBRID LANGUAGE MODELS: IS SIMPLER BETTER?

P.E.Kenne and Mary O'Kane
The University of Adelaide
South Australia 5005
Australia
Tel. +61 8 83033282 FAX:+61 8 83034417
E-mail: pek@dvcr.adelaide.edu.au

ABSTRACT

The use of several n-gram and hybrid language models with and without cache is examined in the context of producing court transcripts. Language models with cache (in which words which have recently been uttered are preferred) have seen considerable use. The suitability of cache models (with fixed size cache) in the production of court transcripts is not clear. A decrease in perplexity and an improvement in the word error rate is observed with some of the models when using a cache, however, performance deteriorates with increasing cache size.

1. INTRODUCTION

An increasingly common approach to developing spoken language systems is to use multiple speech recognizers and a number of alternate language models [1]. Problems which arise from this approach include how to decide which language model and recognizer to use at a given time, and how to change language model or recognizer. We work with court transcripts and examine several approaches in deciding when to change language models. This work extends that described in [2].

One form of language model adaptation which has seen considerable use is the cache language model [3], in which words which have been recently uttered are preferred. These language models capture a notion of locality of topic, and lead to an increase in performance of speech recognition systems. There are applications which seem to be ill suited to the use of caching language models. In the results presented below, the use of a language model with cache is of limited value in the automatic production of court transcripts.

2. DATA

The transcripts from five Australian court cases were used for training and testing. These transcripts do not include non-speech events such as "um" etc, but they do include repetitions, false starts etc. Details of training and test sets are given in table 1 below.

Case	c1	c2	c3	c4	c5
Training(words)	180K	290K	650K	1.2M	860K
Test (words)	20K	40K	60K	65K	64K
% Coverage	88	92	92	85	90

Table 1: Training and test set details

Each of the training sets consists of approximately the first 25% of the transcript of the case, and the test sets are each the next two or three days of transcript. Note that this summary only reports the results for case c3; the results for the other cases are similar.

3. THE EXPERIMENTS

Several language model types were used: word bigram, word trigram, word phrase bigram and word phrase trigram, with and without a cache [3], all using a linear backing-off strategy [4]. Word phrase models are ones in which the tokens may be either words or phrases. The phrases are often (but not always) commonly occurring phrases, for example, in court transcripts, the phrase "May it please the court" is likely to be treated as a token. The word phrase models were generated by using the method described in [5]. Higher order n-grams were not used due to lack of data.

For each type of model, we had a model trained only on lawyers' speech, a model trained on both lawyers' and witnesses' speech and a model trained only on witnesses' speech. (Lawyers and witnesses are recorded on separate tracks, so determining that a speaker has changed is straightforward.) The judge's speech is not modelled separately but is included with the lawyers' speech.

The methods used to determine a change of language model were change of speaker, and two hybrid models based on local perplexity. Local perplexity is calculated in the same way as the perplexity of an entire corpus is calculated: the corpus is divided into sections and the perplexity is calculated for each section. For all the models used here, local perplexity was calculated by using a fixed-size, moving window of 600 words with a 90% overlap. In the hybrid models, the default action is to start with the model trained on all speakers, and if the local

perplexity becomes too large (if it exceeds the test set perplexity plus 15%) a switch is made to the lawyer or witness language model as appropriate. The second hybrid model uses utterance length and local perplexity as criteria for changing language models. Here a change is made to the lawyer language model (if not already using it) when the local perplexity becomes too large, and the utterance length exceeds the mean witness utterance length plus 1.5 standard deviations.

Effective perplexity and recognition results obtained for the five cases using various language models with no cache are given in tables 2-6 below. Here 63(15.2) means an effective perplexity of 63 and a word error rate of 15.2%. "Both" is the language model trained on all speakers, L+W changes models from the lawyer model to the witness model when a change of speaker from lawyer to witness is detected, Hybrid-1 uses local perplexity to change models and Hybrid-2 uses local perplexity together with utterance length to change models. 2-g and 3-g are word bigram and trigram models and wp-3 is a word phrase trigram model.

Model Type	Both	L+W	Hybrid-1	Hybrid-2
2-g	32(21.2)	27(19.2)	26(17.4)	25(17.3)
3-g	24(17.9)	24(16.1)	22(15.8)	22(15.8)
wp-3	22(17.1)	20(16.2)	19(15.3)	19(15.2)

Table 2: Effective perplexity and error rate for case c1; no cache

Model Type	Both	L+W	Hybrid-1	Hybrid-2
2-g	53(18.9)	44(16.2)	41(15.1)	40(14.9)
3-g	34(16.7)	33(15.1)	30(13.8)	28(13.5)
wp-3	33(16.1)	30(14.3)	29(13.6)	28(13.4)

Table 3: Effective perplexity and error rate for case c2; no cache

Model Type	Both	L+W	Hybrid-1	Hybrid-2
2-g	63(15.2)	54(15.8)	51(13.7)	50(13.2)
3-g	42(16.9)	41(15.4)	37(13.6)	34(13.4)
wp-3	40(15.8)	38(15.2)	37(13.5)	36(13.2)

Table 4: Effective perplexity and error rate for case c3; no cache

Model Type	Both	L+W	Hybrid-1	Hybrid-2
2-g	96(22.4)	89(19.8)	86(17.2)	84(16.8)
3-g	60(19.9)	57(16.2)	52(14.7)	50(14.4)
wp-3	56(18.7)	52(17.9)	49(16.3)	50(16.8)

Table 5: Effective perplexity and error rate for case c4; no cache

Model Type	Both	L+W	Hybrid-1	Hybrid-2
2-g	66(15.9)	58(14.6)	52(12.9)	50(13.2)
3-g	47(16.2)	43(13.9)	38(12.1)	37(13.6)
wp-3	42(15.1)	39(13.3)	36(12.0)	36(12.5)

Table 6: Effective perplexity and error rate for case c5; no cache

The effects of varying cache size for the different language models and cases are shown in figures 1-5 below for the word trigram model type.

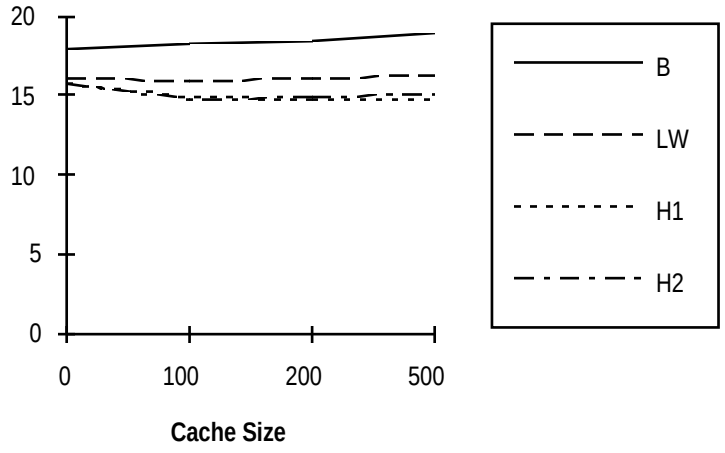

Figure 1: Error rate versus cache size, case c1

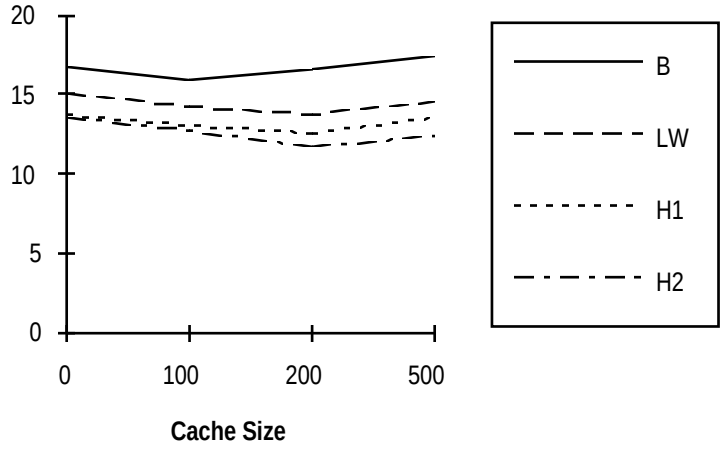

Figure 2: Error rate versus cache size for case c2

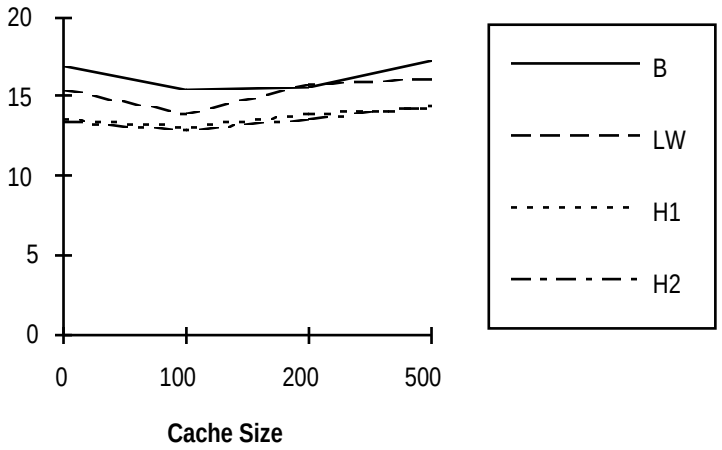

Figure 3: Error rate versus cache size for case c3

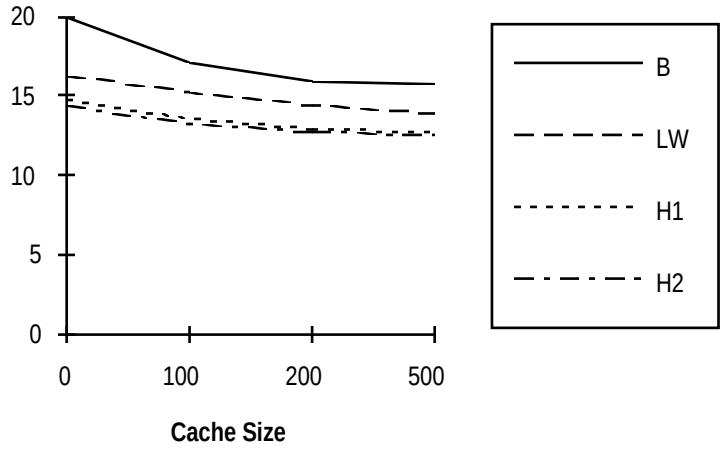

Figure 4: Error rate versus cache size for case c4

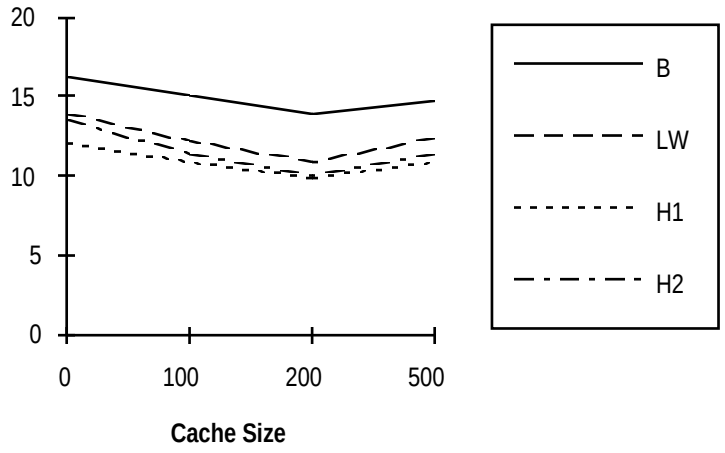

Figure 5: Error rate versus cache size for case c5

Observe that increasing cache size does not give decreasing error rates for all the cases. Similar results are observed for the other model types.

4. CONCLUSION

Improvements in performance using language models with cache are smaller than those reported elsewhere [3], and indeed performance deteriorates with increasing cache size for a number of the cases. This is probably due to the fact that in a court dialogue, the notion of locality, which caching models try to capture, is less well captured by a fixed size cache. Average utterance lengths vary widely both between cases and within cases. Intra-case variance is illustrated by long opening and closing remarks made by lawyers, whilst relatively short utterances are usually characteristic of lawyers interacting with most witnesses. Case c4 is atypical in that it has testimony given by a number of expert witnesses, whose utterances are more typical of those made by lawyers. This is illustrated by the improvements in error rate for increasing cache size. Further investigation with variable sized caching is needed for court dialogues.

5. REFERENCES

[1] H. Chevalier *et. al.*, *"Large-vocabulary speech recognition in specialised domains*", Proceedings ICASSP 95, pp. 217-220, Detroit, 1995.

[2] P. E. Kenne and M. O'Kane, "*Hybrid language models and spontaneous legal discourse,"* Proceedings ICSLP 96, pp. 717-720, Philadelphia, 1996.

[3] S. Sekine, "A New Direction for Sublanguage N.L.P.", *Journal of Natural Language Processing*, Vol.2 (2) pp. 75-87, 1995.

[4] H. Ney, U. Essen and R. Kneser, "On structuring probabilistic dependencies in stochastic language modelling", *Computer Speech and Language* Vol.8, pp. 1-38, 1994.

[5] P. E. Kenne, M. O'Kane and H. Pearcy, "*Language modeling of spontaneous speech in a court context*", Proc. EUROSPEECH'95, pp. 1801-1804, Madrid, 1995.