HYBRID LANGUAGE MODELS AND SPONTANEOUS LEGAL DISCOURSE

P.E.Kenne and Mary O'Kane

The University of Adelaide
South Australia 5005
Australia
{pek,mok}@dvcr.adelaide.edu.au

ABSTRACT

The idea of using multiple speech recognizers and alternate language models in a spoken language task is a familiar one. The problems which arise are deciding which language model(s) and/or recognizer(s) to use, or when to change language models. We use the idea of local perplexity of a corpus to determine when to change to an alternate language model and compare this strategy to a number of other strategies in the context of court interactions.

1. INTRODUCTION

The idea of using multiple speech recognizers and alternate language models in a spoken language task is a familiar one [1]. The problems which arise are deciding which language model(s) and/or recognizer(s) to use and when to change language models. For example, when working with recordings of court proceedings to automatically produce transcript, if a speaker change from a witness to a lawyer is detected (for exmaple, if lawyers and witnesses are recorded on separate tracks) it is simple to change from a language model for witnesses' speech to a language model for lawyers' speech.

We investigate this approach and several others.

2. DATA

The transcripts from two 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.* Table 1 gives details of training and test set sizes.

	c1	c2
Training size (words)	155K	290K
Test size (words)	20K	40K
% Coverage	88	92

Table 1: Training/test set details

In case c1, the training set is the first ten days of a trial, and the test set is the next two days. In case c2 the training set is the first twenty days of a trial and the test set is the next three days. Tables 2 and 3 give further details about each of the cases. In these tables we have combined training and test set details. Witnesses are assumed not to ask questions (not always true but the total number of questions asked by witnesses is very small). See Kenne *et al* [4] for further characterization of Australian court cases.

	Total Words	Total Questions	Total Statements	Average utterance length
Judge1	17122	88	1027	15.3
Lawyers				
L1	14699	614	122	19.9
L2	10217	200	173	27.4
L3	36805	1112	594	21.6
L4	47921	1497	646	22.4
L5	859	0	11	78.0
L6	16	0	1	16.0
Witnesses				
W1	215	0	11	19.5
W2	76	0	4	19.0
W3	1729	0	169	10.2
W4	3798	0	219	17.3
W5	5896	0	540	10.8
W6	5197	0	580	9.0
W7	550	0	51	10.8
W8	1861	0	168	11.0
W9	24333	0	1821	13.4

Table 2: Details for case c1

3. LANGUAGE MODELS AND LOCAL PERPLEXITY

Several language model types were tested: word bigram, word trigram and word phrase bigram [5], all using a linear backing-off strategy [2]. For each model type and each case we used 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 wit-

	Total Words	Total Questions	Total Statements	Average utterance length
Judge2	8369	48	606	12.8
Lawyers				
L7	12924	362	142	25.6
L8	18186	507	257	23.8
L9	116920	4421	264	25.0
L10	52544	2214	132	22.4
Witnesses				
W10	383	0	83	4.6
W11	9366	0	555	16.9
W12	1457	0	124	11.8
W13	2290	0	245	9.4
W14	1152	0	95	12.1
W15	3352	0	130	25.8
W16	9395	0	682	13.8
W17	3906	0	266	14.7
W18	2563	0	109	14.7
W19	55851	0	3002	18.6
W20	887	0	71	12.5
W21	380	0	71	5.4
W22	549	0	44	12.5
W23	1901	0	153	12.4
W24	4733	0	339	14.0
W25	467	0	45	10.4
W26	6502	0	540	12.0
W27	15447	0	792	19.5

Table 3: Details for case c2

nesses are recorded on separate tracks, so determining that a speaker has changed is straightforward.) In both cases, the judge's speech was included with the lawyers' for training and testing.

For a given language model and corpus

$$w_1 w_2 \ldots w_N$$

the perplexity PP of this corpus with respect to the language model is defined as

$$PP = \exp(-\frac{1}{N} \Pr(w_1 w_2 \ldots w_N)).$$

The perplexity may thought of as approximately the average branching factor in the language model. Most of the literature reports a single value of PP for a language model/corpus combination. However, there is no reason why a number of values of perplexity (corresponding to different parts of the corpus) may not be calculated.

Figure 1 shows the local perplexity (for a word bigram model) for the start of case c2 test set, using fixed windows of 600 and 1200 words, with 90% overlap in both cases. The perplexity of this data is 42.

Local perplexity was calculated for the remainder of these experiments by using fixed size of 600 words and 90% overlap. For the court data sets, changing language models does not substantially change the shape of the local perplexity curve (see Kenne and O'Kane [3]). A different language model does however, translate the curve up or down.

4. RESULTS

Results from applying different language models to the test sets are given in tables 4-9.

Here, "Both" is the language model trained on all speakers, L+W changes models from the lawyer model to the witness model when a change from lawyer to witness is detected and Hybrid uses local perplexity to change models. In the hybrid case the default action is to start with the model trained on all speakers, and if the local perplexity becomes sufficiently large (if the local perplexity exceeds the test set perplexity plus 12%) a switch is made to the lawyer or witness language model as appropriate.

Observe that for case c2 the perplexity and word error rate increase when using separate language models for lawyers and witnesses in some cases. This seems to be due to the fact that case c2 has a number of expert witnesses giving evidence (W19 and W27) and their language is atypical of the other witnesses'. A possible solution for this difficulty is to use to lawyers' language model for such speakers.

There are several areas of further work: there is a (almost) closed set of lawyers who appear in cases in some areas of law and it would be of interest to develop language models for individuals. Similarly, although there is insufficient data to adequately model a judge's similar techniques may be applied. This would also allow techniques such as those described by Rudnikiy [6] to be used.

Model	Word Error Rate	Effective Perplexity
Both	14.9	23
L+W	12.2	20
Hybrid	11.7	19

Table 4: Results for case c1: word bigram

Model	Word Error Rate	Effective Perplexity
Both	10.8	42
L+W	11.2	36
Hybrid	9.8	34

Table 5: Results for case c2:word bigram

5. REFERENCES

1. Chevalier, H., Ingold, C., Kunz, C., Moore, C., Roven, R., Yamron, J., Baker, B., Bamberg, P., Bridle, S., Bruce, T., and Weader, A. Large-vocabulary speech recognition in specialized domains. In *Proceedings ICASSP 95*, pages 217–220, May 1995.

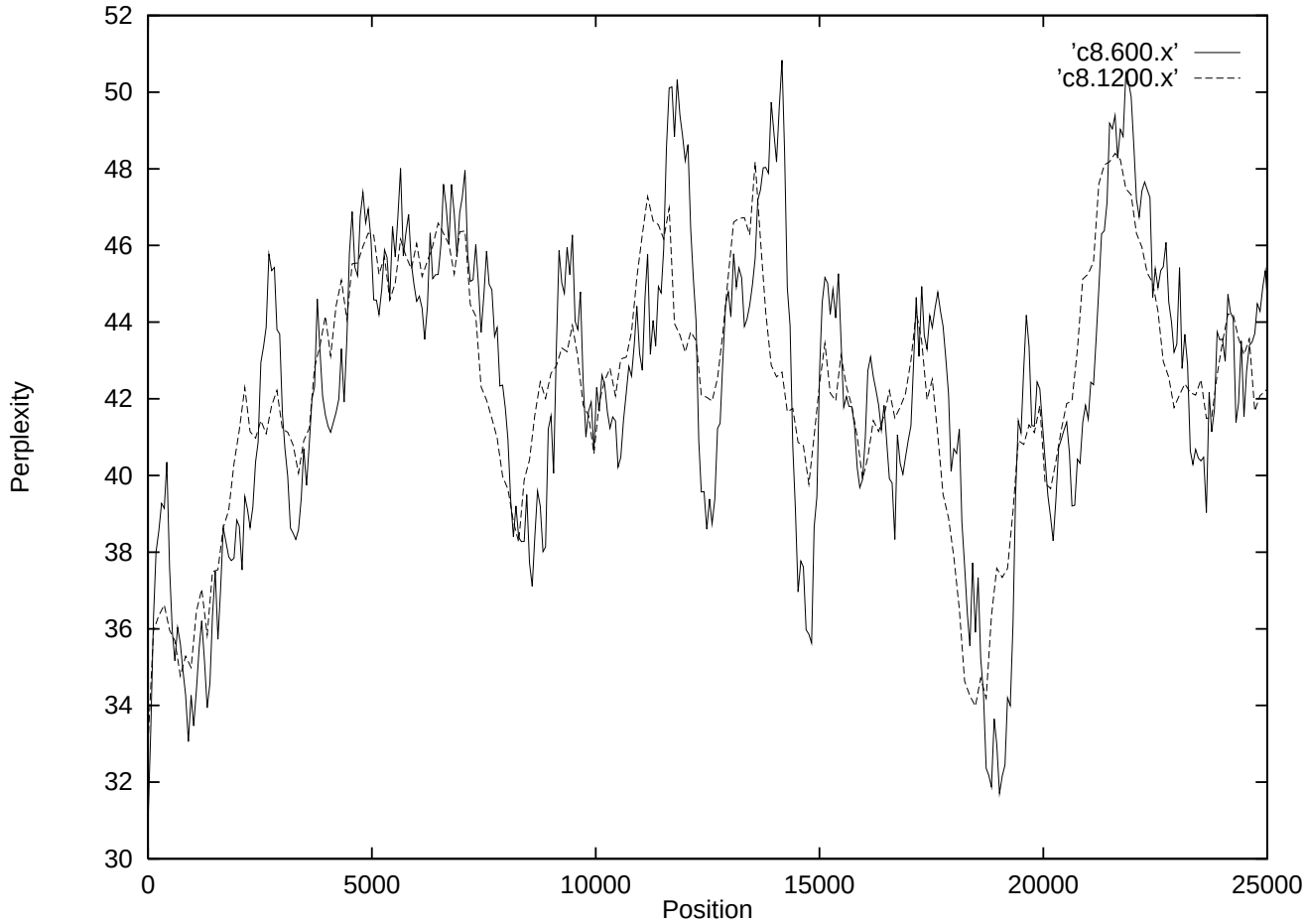

Figure 1: Local perplexity for case c2 for windows of size 600 and 1200

Model	Word Error Rate	Effective Perplexity
Both	9.7	13
L+W	9.2	11
Hybrid	9.2	10

Table 6: Results for case c1: word trigram

Model	Word Error Rate	Effective Perplexity
Both	11.9	28
L+W	12.4	29
Hybrid	9.8	26

Table 7: Results for case c2:word trigram

Model	Word Error Rate	Effective Perplexity
Both	11.8	16
L+W	11.8	14
Hybrid	10.9	13

Table 8: Results for case c1: word phrase bigram

Model	Word Error Rate	Effective Perplexity
Both	12.4	36
L+W	12.3	36
Hybrid	12.2	34

Table 9: Results for case c2:word phrase bigram

2. Ney H., Essen U., and R. Kneser. On structuring probabalistic dependences in stochastic language modelling. *Computer Speech and Language*, 8:1–38, 1994.

3. Kenne P.E. and O'Kane M. Topic change and local perplexity in spoken legal dialogue. In *Proceedings of the International Conference on Spoken Language Processing ICSLP96*, October 1996.

4. Kenne P.E., O'Kane M., and Pearcy H.G. An australian speech database derived from court recordings. In *Proceedings of the Fifth Australian International Conference on Speech Science and Technology*, pages 808–813, December 1994.

5. Kenne P.E., O'Kane M., and Pearcy H.G. Language modeling of spontaneous speech in a court context. In *Proceedings Eurospeech 95*, pages 1801–1804, September 1995.

6. A.I. Rudnicky. Language modeling with limited domain data. In *Proceedings of the Spoken Language Systems Technology Workshop*, pages 66–69, January 1995.