Proceedings of the Institute of Acoustics

A HYBRID GRAMMAR–BIGRAM LANGUAGE MODEL WITH DECODING OF MULTIPLE (N–BEST) HYPOTHESES FOR SPEECH RECOGNITION

G.J.F. Jones, J.H. Wright, H. Lloyd–Thomas & E.N. Wrigley

Centre for Communications Research, University of Bristol, Queens Building, University Walk, Bristol BS8 1TR

ABSTRACT

The most likely sentence decoded by automatic speech recognition cannot be guaranteed to be that uttered by the speaker. Errors may occur either in the words of the sentence hypothesis or grammatical derivation or both. For this reason considerable research has been undertaken into the development of algorithms to find the N-best most likely sentence hypotheses. In this paper we explore the development of an N-best hybrid language model incorporating both a bigram language model and a probabilistic context free grammar (PCFG). This hybrid successfully combines the broad coverage of the language of the bigram with the grammatical derivations of the PCFG. A simple hybrid N–best is successfully formed from the merging of the outputs from two separate N–best sentence hypotheses lists. However, in this model where the two approaches are entirely separate when a bigram derived sentence is chosen all grammatical structure is lost. To overcome this disadvantage we propose a consolidated language model designed to maintain the maximum degree of grammatical structure by linking grammar derived phrases using bigram type probabilities in the non–terminal symbols.

1 INTRODUCTION

The probabilistic relationships within and between the component stages of automatic speech recognition systems mean that the most highly scoring sentence hypothesis will be formed on the basis of previous experience of the manner of speech and the use of language. Of course, this may not accurately model the input to the system, the speaker may vary his or her speaking style or use of language or there could be variation in environmental factors outside the speaker's control. These factors mean that the highest scoring hypothesis from the recognition system may not actually correspond to the sentence uttered. In order for there to be alternative derivations available when the highest scoring hypothesis is in error the incorporation of derivation of N–best sentence hypotheses has attracted considerable interest in recent years. A number of papers have describing approaches to the computation of such lists have appeared in the literature [1] [2] [3] [4]. The details of these algorithms are reviewed in section 3.

The derivation of the N–best list increases the computational load of recognition system. In systems processing a large vocabulary and continuous speech utterance this increased overhead may render real–time processing impossible. For this reason many systems incorporate a low cost language model as the first stage after the pattern matcher to reduce the search space before the data is further processed by more computationally expensive language models. In our investigation we apply this approach to the incorporation of a Probabilistic Context–Free Grammar parsed by an enhanced LR parser as a high level language model. In this system the derived hypotheses lists for the two language models are available separately and in a combined hybrid list. It is observed that the hybrid provides the best overall performance.

We also investigate the use of an alternative topology processing the data in parallel with a bigram and the PCFG with LR parser. This second situation whilst computationally more expensive

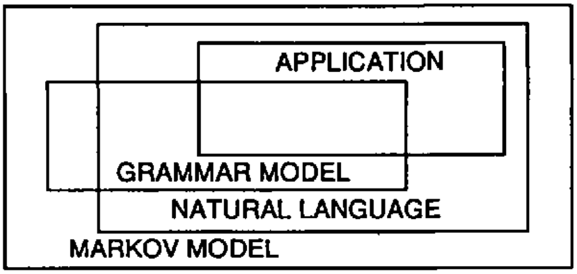

Figure 1: Undergeneration and overgeneration by language models

avoids the potential loss of high scoring grammatical derivations from the N–best list which may occur if the data is processed sequentially. While the bigram N–best may form a long N–best list its coverage of potential word order is very broad, and while this allows any utterance from the vocabulary to be processed it also allows many non grammatical derivations to appear in the list. Since we hope that the grammar model is good reflection of the general use of the language many of these bigram derivations are actually very unlikely. A smaller N–best list derived by the grammar model at the front end may actually be more useful, although even with efficient implementation it may be too slow in practice for a large task. Experimental results for both approaches are reported in this paper.

The next stage of the development of the hybrid is to attempt to combine them to achieve the advantages of both systems. This will give the combined model a coverage of the language similar to the bigram whilst generating the maximum amount of grammatical structure within the hypotheses. The principles of this consolidated language model are described in section 6.

2 THE HYBRID LANGUAGE MODEL

2.1 The scope of the language models

Two important problems with the use of language models in speech recognition are undergeneration and overgeneration. Undergeneration is the failure of the model to cover a sentence actually used by the speaker, whereas overgeneration is the coverage by the language model of sentences which are never used and do not need to recognised, many of these may in fact be complete nonsense. The relationship between language use in a particular application, the whole natural language, a grammar model and Markov model are shown in Figure 1.

Markov language models, usually as a bigram, are often implemented in such a way as to eliminate undergeneration entirely as shown in Figure 1. This implies that there is a lot of redundancy in such language models. A grammar model has very much reduced overgeneration but does suffer from undergeneration. This undergeneration occurs because a grammar model can never be complete. New linguistic constructions can be used by the speaker at any time. It is to overcome this that we designed our original hybrid model [5]. If the grammar model is a good representation of the language as typically used within a task it will generally be invoked as providing the most likely

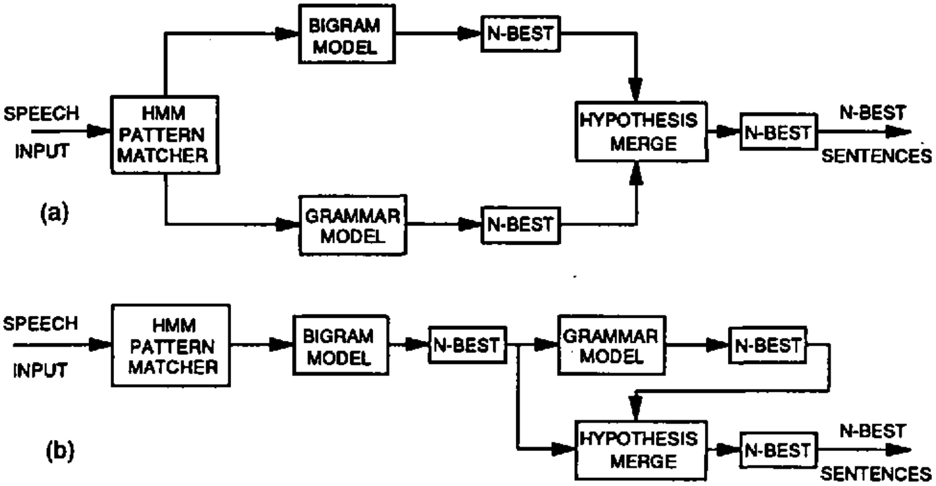

Figure 2: System topologies for generation of multiple hypotheses: (a) parallel (b) pipeline.

derivation. For utterances outside the scope of the grammar model the Markov model is invoked as a back-up.

2.2 System topologies for hybrid language models

The initial 1-best hybrid language model operates both language models in parallel. The output is chosen competitively from the outputs of the two models [5]. For speech recognition systems which incorporate more than one language model and N-best decoding there are two basic options for system topology. The language models can be used in a closely coupled parallel arrangement or they can be arranged in a pipeline arrangement. Examples of both topologies are shown in Figure 2.

This latter topology usually places the models in order of increasing computational complexity allowing the search space to be constrained progressively. The broad low cost bigram is typically placed first followed by more expensive ones for example those based on PCFGs. The entropy reduction associated with the application of a particular language model is traded against the cost of applying it. The input to the PCFG in this latter case is the N-best list output from the preceding language model. The disadvantages of this approach are that the lower order language model may not generate an N-best list sufficiently long to include the correct sentence and that many of the sentences it produces will due to the vast overgeneration described previously often be complete nonsense. Thus, the grammar is processing many word sequences which it would have discarded as ungrammatical early in the process of derivation and the correct solution, which it is quite possible it would have scored highly, may be lost. The parallel arrangement avoids these two problems with respect to the grammar, however, the increased scope of the search space may render this arrangement too computationally expensive in complex tasks. For example, in the pipeline arrangement the grammar does not have to deal with the problems associated with word boundaries in continuous speech recognition. However, it is unable to make a contribution to guiding the system to the correct choice of word hypotheses in this stage of the recognition process.

3 THE N-BEST ALGORITHM

The additional computational load associated with N-best algorithms has been mentioned several times already. This factor has been critically important to the differences between the algorithms released in this area to date. In continuous speech recognition the boundaries between word utterances and even the number of words in a phrase are not well defined. If the absolute N-best sentence hypotheses are to be derived these variations must be taken into account exactly. To do this for more than a small task vocabulary generates a very large search space. An algorithm to calculate this exact list based on a bigram language model and Viterbi type search is described in [1]. There are important additional considerations, for example the computation must be of word-sequences rather than state sequences and it must find all such sequences within a specified beam. It is found in practice that even the incorporation of this beam search to remove low scoring hypotheses does not prevent this algorithm being too expensive for real-time computation [6]. Various approximations have been proposed that reduce the computational load, for example rather than allowing multiple seqmentations for a particular word only the highest scoring is preserved [4] or the segmentation is based only on the preceding word [7]. In addition a highly efficient algorithm known as the Forward-Backward algorithm has been proposed to make an initial search to reduce the search space before applying the full N-best search [6].

4 USING A PCFG AS A LANGUAGE MODEL

The Markov language model provides no grammatical derivation for its sentence hypotheses. A language model based on a grammar model not only provides a syntactic interpretation of the sentence but uses the linguistic structures to guide the generation of derivation hypotheses. A theoretical framework and parsing algorithm for PCFGs has been described previously [8]. this approach is based on the generalised LR parsing algorithm of Tomita [9]. This algorithm is an efficient extension of the popular LR algorithm which is much used in compilers. The extension allows for the presence of ambiguity, which seems to be an unavoidable feature of non-trivial natural language grammars. The Tomita algorithm has been extended first to probabilistic grammars and second to permit the type of stochastic input which would be expected as output from a hidden Markov model pattern matcher. Thus, this grammar model can be used in either of the hybrid topologies discussed previously.

In the pipeline arrangement the parser processes sentences derived by the bigram language model. The parser attempts to generate grammatical derivations of the word sequence. Since the parser is able to handle syntactic ambiguity more than one derivation of each word sequence may be found. Each derivation will have associated with it a likelihood score from the grammar. Alternatively, the word sequence may be found not to have any grammatical derivation and, thus, it is outside the scope of the grammar. In this case the word sequence is given a likelihood score of zero. The derivations as calculated by the parser are then arranged into a separate N–best list. The length of this list will depend on the number of derivations associated with each bigram derived sentence since each derivation for a word sequence is considered to be a different sentence hypothesis.

When connected directly to the HMM pattern matcher, as in the parallel hybrid topology, the word likelihoods are delivered to the parser which assembles the plausible sentences using the grammar as a guide. A probability–weighting for each sentence then emerges, based on the grammar and the word likelihoods. This weighting is used as the criterion for pruning out low scoring derivations and prevents a combinatorial explosion. The efficient representation of syntactic structure in a parse forest [10] permits the algorithm to exhibit a time–dependence which is a polynomial function of the length of sentence. The probabilities are devolved through the parse forest in such a way that the possible sentences can be ranked by overall probability. The parser forms the overall N-best derivations into an output list. In the same way as for the pipeline N–best grammar derivation list each derivation associated with a word sequence is treated at all points as a separate derivation.

5 FORMING THE SIMPLE N–BEST HYBRID LIST

The N–best sentence hypotheses lists from the two language models need to be merged to form an overall hybrid N–best list. The following approach has been adopted to this area. Identical word sequences from the two lists are compared. The highest scoring one is selected as the chosen hypothesis, this means that the grammatical derivation may be lost if it scores lower than the bigram hypothesis with the same word sequence. Since there may be several grammar derivations associated with a particular word sequence this comparison must be carried out for each such derivation. The policy adopted in this situation is to select the bigram at most once, for example, if it scores better than the second highest scoring derivation it is selected. It will also score better than subsequent derivations of the same word sequence. These are discarded as unreliable but the bigram derivation is not added to the list again since it doesn't represent an independent hypothesis to one already in the list. Bigram word sequences (and in the case of the parallel system grammar derivations as well) which are not found in the both lists are placed in the hybrid N–best list on the basis of their score in this list alone. There are several possible approaches to scoring the hybrid list. The highest scoring probability from the hypothesis comparison may be used or some combination of the scores associated with the compared hypotheses. This aspect is considered further in the experimental work.

6 THE CONSOLIDATED LANGUAGE MODEL

The hybrid system described above produces a merged output from two separate language models which are very different: the Markov model is sequential and statistical, the grammar model is hierarchical and rule–based. Considered on their own, the Markov model would benefit from some additional structure, and the grammar model would benefit from the additional flexibility brought by the sequential nature of trained n-grams. Work is in progress on a consolidated language model which combines the advantages of the two models and produces a single N–best output list.

The essence of the approach is to re-score the initial set of N–best bigram sentences (in a pipeline) by running a substring parser over them, building grammar structure wherever possible. In addition to the grammar rules, the nonterminal symbols are also related using bigrams. These link the islands of syntactic structure, and in fact operate at all levels. All scores (including the original bigram model score) are allowed to be carried up through the layers of interpretation in a competitive way, although only a beam of the best ones is retained.

This approach overcomes the main disadvantage of a grammar, namely that each sentence is either inside or outside of the language. Even when a sentence is "close" to the language the parser will normally fail and in the hybrid system the score is then based purely on the word or preterminal bigrams, whereas in this approach what grammar structure is present is also allowed to participate in the scoring. With proper training (which is an issue which requires further work) the net result should be improved recognition.

There is a further benefit. There is such a discontinuity between an n-gram model and a full grammar that a measure of progress from the one to the other is impossible at present, and this has held back the application of enhanced language models in speech recognition. The consolidated model should enable progress to be measured (in terms of recognition performance) incrementally as a function of added structure, thus creating a much-needed continuity.

Further details and results using the consolidated model will be reported elsewhere.

7 EXPERIMENTAL RESULTS

Experimental investigation was made of the effectiveness of the basic hybrid language model. The experiments were carried out on a PC based DSP32C Telephony board. The example under consideration is a relatively simple task, we have a vocabulary of 100 words and the speech is isolated. A basic evaluation of system performance is still possible. A 7-state HMM of each word was enroled on the speech board. The training was limited to only 5 utterances of each word and the system is speaker dependent. The probabilistic grammar used for the experiments consisted of 230 rewrite rules governing the 100-word vocabulary. A bigram table was obtained from a corpus of text generated at random from the PCFG, using the rule probabilities. The bigram model was smoothed using the standard Good-Turing technique [11]. The Markov model is thus representative of the language governed by the grammar and is able to effectively complement the grammar in the hybrid.

Two sets of test sentences were generated. The first was generated at random from the PCFG, again using the rule probabilities. These were not part of the training set used to train the bigram. The second set was generated at random from the bigram model trained using the standard Held-Out method [11], but with all grammatically correct sentences filtered out. This second set is therefore intended to activate the bigram language model. Both sets contained nearly 100 sentences.

The performance of the parallel and pipeline topologies was analysed over three sets of test data. It was found that in both arrangements the appropriate component language model was selected on the majority of occasions. For the separate N-best lists the PCFG performed better than the bigram model for the grammar compatible sentences. For the ungrammatical sentences the bigram model was successfully chosen by the hybrid model without being distracted by the incorrect hypotheses generated by the PCFG.

Figure 3 shows the cumulative distribution of the rank of the correct sentence for the pipeline

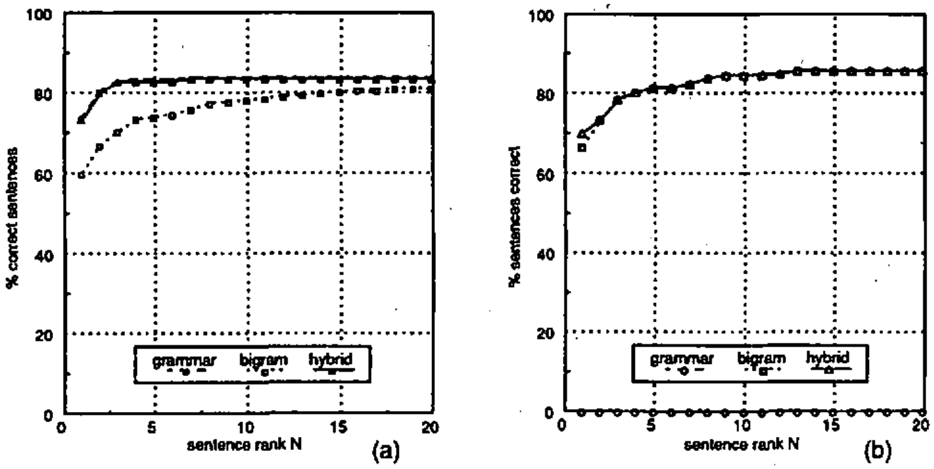

Figure 3: Cumulative distribution of rank of correct sentence for the hybrid language model : (a) grammar compatible test sentences (b) grammar incompatible sentences.

arrangement for both grammatical and ungrammatical sentences. For the PCFG only the highest scoring derivation is recorded in this result since we are only comparing the correct word sequence of the sentences. The performance of the parallel arrangement was found to be almost identical. The pattern matcher correct word performance was about 70%. The average length of sentences was 8.3 words for grammar and 6.5 words for ungrammatical sentences.

These experiments were repeated with several different approaches to the scoring of the merged hypotheses lists eg using the score of the selected hypothesis in a comparison or a linear combination of the scores of the hypotheses being compared. It was found that there was no detectable advantage in any of these methods. The results shown are for N-best lists with the scores of the selected hypothesis in each case.

8 CONCLUSIONS AND FURTHER WORK

A hybrid language model has been described which can successfully be applied to the generation of N–best lists of hypotheses. It generates grammatical derivations for utterances within the scope of grammar while successfully providing wide coverage of use of the words from a task vocabulary in an ungrammatical fashion.

The next stage of this work is examination of the effectiveness of the consolidated language model proposed in this paper for speech recognition. The computational load associated with the different language models and topologies must also be analysed. In addition, further language processing stages, for example the first–order dependence model described in [12], are being investigated.

9 ACKNOWLEDGEMENTS

This work is financially supported by the U.K. Science and Engineering Research Council, by Ensigma Limited, Chepstow and by the Defence Research Agency, RSRE, Malvern. Thanks are extended to the staff of Ensigma Limited for assistance with HMM software and DSP hardware and also to the Centre for Communications Research for provision of computing facilities.

REFERENCES

[1] R.Schwartz and Y.-L.Chow, "The N-best algorithm: an efficient and exact procedure for finding the N most likely sentence hypotheses," in *Proceedings of ICASSP-90*, pp. 81–84, IEEE, 1990.

[2] R.Schwartz and S.Austin, "A comparison of several approximate algorithms for finding multiple (n-best) sentence hypotheses," in *Proceedings of ICASSP-91*, (Toronto), pp. 701–704, IEEE, 1991.

[3] F.K.Soong and E.-F.Huang, "A tree-trellis based fast search for finding the n-best sentence hypotheses in continuous speech recognition," in *Proceedings of ICASSP-91*, (Toronto), pp. 705–708, IEEE, 1991.

[4] V.Steinbiss, "A search organization for large-vocabulary recognition based on n-best decoding," in *Proceedings of EUROSPEECH-91*, (Genoa), pp. 1217–1220, 1991.

[5] G.J.F.Jones, J.H.Wright, E.N.Wrigley, M.J.Carey, and E.S.Parris, "Isolated-word sentence recognition using probabilistic context-free grammar," in *Proceedings of EUROSPEECH-91*, (Genoa), pp. 487–489, 1991.

[6] S.Austin, R.Schwartz, and P.Placeway, "The forward-backward search algorithm," in *Proceedings of ICASSP-91*, (Toronto), pp. 697–700, IEEE, 1991.

[7] R.Schwartz, S.Austin, F.Kubala, J.Makhoul, L.Nguyen, P.Placeway, and G.Zavaliagkos, "New uses for the n-best sentence hypotheses within the byblos speech recognition system," in *Proceedings of ICASSP-92*, (San Fransico), pp. 1(1–4), IEEE, 1992.

[8] J.H.Wright, "LR parsing of probabilisitc grammars with input uncertainty for speech recognition," *Computer Speech and Language*, vol. 4, pp. 297–323, October 1990.

[9] M.Tomita, *Efficient parsing for Natural Language.* Kulwer Academic Publishers, 1986.

[10] E.N.Wrigley and J.H.Wright, "Computational requirements of probabilistic LR parsing for speech recognition using a natural language grammar," in *Proceedings of EUROSPEECH-91*, (Genoa), pp. 761–764, 1991.

[11] K.W.Church and W.A.Gale, "A comparison of the enhanced good-turing and deleted estimation methods for estimating probabilities of English bigrams," *Computer Speech and Language*, vol. 5, pp. 19–54, January 1991.

[12] J.H.Wright, "Adaptation of grammar-base language models for continuous speech recognition," in *Proceedings of EUROSPEECH-91*, (Genoa), pp. 203–206, 1991.