COMPARISON OF PART-OF-SPEECH AND AUTOMATICALLY DERIVED CATEGORY-BASED LANGUAGE MODELS FOR SPEECH RECOGNITION

T.R. Niesler, E.W.D. Whittaker and P.C. Woodland

Cambridge University Engineering Department
Trumpington Street, Cambridge, CB2 1PZ, England

{trn,ewdw2,pcw}@eng.cam.ac.uk

ABSTRACT

This paper compares various category-based language models when used in conjunction with a word-based trigram by means of linear interpolation. Categories corresponding to parts-of-speech as well as automatically clustered groupings are considered. The category-based model employs variable-length *n*-grams and permits each word to belong to multiple categories. Relative word error rate reductions of between 2 and 7 % over the baseline are achieved in N-best rescoring experiments on the Wall Street Journal corpus. The largest improvement is obtained with a model using automatically determined categories. Perplexities continue to decrease as the number of different categories is increased, but improvements in the word error rate reach an optimum.

1. INTRODUCTION

Language models based on *n*-grams of word-categories[1] are intrinsically able to generalise to unseen word sequences, and hence offer improved robustness to novel or rare word combinations. In isolation, such models represent a competitive alternative to standard word-based approaches when the training set is small and consequently sparse [7]. For large training sets, category-based methods offer improved performance when used in combination with word *n*-gram language models [8], [9].

The work presented here studies the performance of various types of category-based language models when used in conjunction with a conventional word-based backoff trigram. Category definitions corresponding to part-of-speech classifications, as well as categories derived by means of a clustering algorithm that optimises the training set likelihood have been used. Since words may have multiple grammatical functions, the former requires the language model to allow words to belong to multiple categories. The category-based model employs variable-length *n*-grams, where n is increased selectively to optimise performance while restricting model size [7].

2. VARIABLE-LENGTH CATEGORY-BASED *N*-GRAMS

This section describes the category-based language model structure and training algorithm. Two alternative approaches to defining the particular categories that will be used by these models are treated in sections 3 and 4.

Let $w(i)$ and $v(i)$ denote the i^{th} word in the corpus and its category respectively, while w_j and v_k denote a particular word and category from the lexicon[2], where $j \in 0 \ldots N_w - 1$ and $k \in 0 \ldots N_v - 1$, and N_w and N_v are the number of different words and categories respectively. Now let the set of categories to which w_j belongs be denoted by $V(w_j)$, where $j \in \{0, 1, \ldots, N_v - 1\}$. Furthermore, let each word history $\mathbf{w}(0, b)$ be classified into an equivalence class h_r defined to be an *n*-gram of categories, i.e.:

$$h_r = H\left(\mathbf{w}\,(0, b)\right) = \left\{ v\,(a)\,, v\,(a+1)\,, \cdots, v\,(b) \right\} \quad (1)$$

where $r \in \{0, 1, \cdots, N_h - 1\}$, $0 \leq a \leq b$, and N_h is the number of history equivalence classes. Since a word may belong to several categories, V is in general many-to-many, and $\mathbf{w}(a, b)$ may map to multiple history equivalence classes. Assuming $P(w(i))$ to be wholly determined by $v(i)$:

$$P\left(w(i)|\mathbf{w}(0,i-1)\right) = \sum_{\forall\, v : v \in V(w(i))} P\left(w(i)|v\right) \cdot P\left(v|\mathbf{w}(0,i-1)\right) \quad (2)$$

Assuming furthermore that the probability of witnessing $v\,(i)$ depends only on the category *n*-gram context, the right-hand side of (2) may be decomposed further:

$$P\left(v|\mathbf{w}\,(0,i-1)\right) = \sum_{\forall\, h : h \in H(\mathbf{w}(0,i-1))} P\big(v|h\big) \cdot P\big(h|\mathbf{w}\,(0,i-1)\big) \quad (3)$$

Figure 1 illustrates the interrelation of the components of equations (1), (2) and (3).

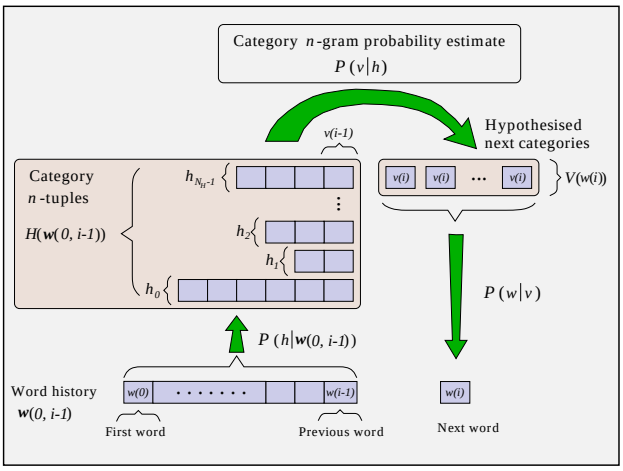

Figure 1: Operation of the category-based language model.

[1] A word-category is taken to mean some set of words.

[2] The possible category assignments for each word in the vocabulary.

When the mapping $V(\cdot)$ is many-to-one, each word belongs to exactly one category and equation (2) reduces to:

$$P\big(w(i)|\mathbf{w}(0,i-1)\big) = P\big(w(i)|V\big(w(i)\big)\big) \cdot P\big(V\big(w(i)\big)|\mathbf{w}(0,i-1)\big) \quad (4)$$

Also the mapping $H(\cdot)$ is many-to one, and therefore equation (3) simplifies to:

$$P\big(v|\mathbf{w}\,(0,i-1)\big) = P\big(v|H\big(\mathbf{w}(0,i-1)\big)\big) \quad (5)$$

Referring again to figure 1, this corresponds to having only a single category *n*-tuple for $H(\mathbf{w}\,(0,i-1))$ and a single hypothesised next category for $V(w(i))$.

The number of categories is generally much smaller than the number of words in the vocabulary, and hence the number of possible *n*-grams is much smaller than for a word-based model. This diminishes the training set sparseness, and makes larger n feasible both from a statistical as well as a storage viewpoint. The variable-length *n*-gram strategy increases the length of individual *n*-grams selectively during model construction according to the expected benefit to predictive performance [7],[9]. In particular, an *n*-gram is extended to an $(n+1)$-gram when this decreases the leaving-one-out training set likelihood by at least a fraction λ_{ct}. This optimises performance while minimising the number of model parameters, and allows λ_{ct} to control the extent to which performance is sacrificed for reduced model size.

3. PART-OF-SPEECH CATEGORY DEFINITIONS

Category definitions in terms of part-of-speech classifications are available in tagged corpora such as LOB [2], and may be used to construct language models as described in section 2. This has been done for one of the language models reported here. Since words may have multiple grammatical functions and hence part-of-speech assignments, the mapping $V(\cdot)$ is many-to-many. The language model described in section 2 was trained on the LOB corpus, and then employed as a statistical tagger in order to obtain the part-of-speech classification of each word in the Wall Street Journal training set [7].

4. AUTOMATIC DETERMINATION OF LEXICAL CLASSES BY CLUSTERING

As an alternative to using the linguistically predefined categories described in the preceding section, the category definitions may be determined automatically by means of an optimisation algorithm that clusters words into groups. In this work we have adopted the greedy algorithm described in [4] and [6], which employs a maximum likelihood optimisation criterion. Words are moved between categories so as to increase the log-likelihood (LL) of the training text with respect to the bigram category language model described by equation (6). Bigram counts for each word-pair are obtained from the same training text. This model constrains each word to belong to only one category, and hence $V(\cdot)$ is many-to-one as in equations (4) and (5).

$$P\big(w(i)\,|\,w(i-1)\big) = P\big(w(i)\,|\,V\big(w(i)\big)\big) \cdot P\big(V\big(w(i)\big)\;|\;V\big(w(i-1)\big)\big) \quad (6)$$

Using the maximum likelihood estimates of the probabilities in equation (6), the log-likelihood of the training text is found to be given by:

$$LL(V) = \sum_{w} N(w)\log N(w) + \sum_{\forall v_i, v_j} N(v_i, v_j)\log\frac{N(v_i, v_j)}{N(v_i)N(v_j)}, \quad (7)$$

where, $N(w)$ and $N(v)$ represents the unigram word and category counts respectively. $N(v_i, v_j)$ is the bigram category count (the number of times v_j follows v_i), and may be found by summing the word bigram counts over the category member words as follows:

$$N(v_i, v_j) = \sum_{\forall w_x : w_x \in v_i} \sum_{\forall w_y : w_y \in v_j} N(w_x, w_y). \quad (8)$$

The summation in the second term of equation (7) takes place over all possible category pairs, (v_i, v_j). Maximising the log-likelihood of the text is only dependent on the second term of equation (7) which is of a form similar to the mutual information between categories. The first component is unaffected by the distribution of words among the categories and hence is constant for a fixed vocabulary and training set.

The algorithm computes the change in log-likelihood after moving each word from its present category to each remaining category in turn. The word is assigned to the category for which this increase is greatest. For N_v categories, this log-likelihood difference may be computed in $\mathcal{O}\,(N_v)$ time by changing only the counts of those categories which are affected by the move.

With I iterations through the vocabulary, the complexity of the algorithm is given by $\mathcal{O}\,(I \cdot ({N_v}^2 \cdot N_w + {N_w}^2))$. This scales as log_2 of the number of unique bigrams stored and hence the complexity of the algorithm is largely unaffected by the size of the corpus that is clustered. More details on the update calculations may be found in [5].

The initial partition of the vocabulary is obtained by assigning the $N_v - 1$ most frequent words each to their own unique category, and the remaining $N_w - N_v + 1$ words to the ${N_v}^{th}$ class. This is an initialisation proposed in [6].

Various approximations can be applied to improve the execution speed of the clustering algorithm [13]. For example, one might choose not to move words which occur fewer than a certain number of times in the training text. However, no approximations were applied to the clustering algorithm used in this paper. Functions from the CMU-Cambridge Toolkit [1] were used by the clustering process to facilitate the collection and storage of word bigram data from corpora. The times taken to perform one iteration of the algorithm on a Sun Ultra 2 for the range of categories considered are given in Table 1.

No. of categories	Time (hrs:mins)
150	4:42
200	5:10
500	13:44
1000	34:39
2000	134:18

Table 1: Time taken to perform one iteration of the automatic clustering algorithm on a Sun Ultra 2.

5. EXPERIMENTAL RESULTS

5.1. Test- and training-corpora

The language model training corpus comprises approximately 37 million words of newspaper text drawn from 1987-89 issues of the Wall Street Journal (WSJ) [11]. Approximately the first 19,000 words from each of these years were taken from the language model setaside development test to yield a 59,000-word test corpus (LM-DEV) for perplexity calculation purposes. A baseline word-based backoff trigram language model (WTG) employing Turing-Good discounting [3] was derived from the training corpus. Singleton bigrams and trigrams were discarded, and a minimum unigram count of 10 was enforced.

Recognition experiments were performed for the development (R-DEV) and evaluation (R-EVAL) tests that formed part of the 1994 ARPA CSR HUB-1 evaluation.

	Sentences	**Speakers**	**Words**
R-DEV	310	20	7,388
R-EVAL	316	20	8,190

Table 2: Characteristics of the 1994 ARPA CSR HUB-1 development (R-DEV) and evaluation (R-EVAL) test sets.

Lattices were generated for these two test sets at Cambridge University using the HTK large-vocabulary speech recognition system with a 65K vocabulary and backoff bigram language model [14]. Before performing experiments with category-based models, these lattices were rebuilt using the baseline word-trigram to ensure that all language models were trained on the same text.

5.2. Category-based models

Variable-length category-based *n*-gram language models were produced for the WSJ corpus as described in section 2. The categories were based either on part-of-speech classes (POS) as described in section 3, or were determined automatically (CLUST) by means of the clustering algorithm treated in section 4. For the latter method, the number of distinct categories N_v could be varied. The pruning threshold λ_{ct} was adjusted to give optimal perplexities on the LM-DEV test set, the only exception being C-0, where this parameter was chosen to yield a language model with approximately as many *n*-grams as that based on part-of-speech categories. Table 3 summarises the language model characteristics, showing also their size in terms of the total number of *n*-grams N_{ng} as well as the perplexity (PP) measured on the LM-DEV test set.

As outlined in section 2, the length of individual *n*-grams is extended selectively during model construction. Figure 2 illustrates the proportion of the total number of *n*-grams for each n occurring in selected models of table 3. Note that the model based on part-of-speech categories employs proportionally more *n*-grams with larger n than the models based on automatically clustered categories, and that increasing the number of categories leads to a preference for shorter *n*-grams.

5.3. Recognition performance

The Entropic Lattice and Language Modelling Toolkit [10] was used to generate N-best lists from the LM-DEV and R-DEV lattices, and to rescore these lists by linear interpolation of the category-based and baseline word-trigram language models. The

Name	**Category type**	N_v	N_{ng}	PP
POS	POS	152	909,542	448.5
C-0	CLUST	150	1,038,766	301.1
C-1	CLUST	150	2,134,923	289.5
C-2	CLUST	200	2,697,015	265.8
C-3	CLUST	500	4,677,642	212.2
C-4	CLUST	1000	6,384,707	184.4
C-5	CLUST	2000	8,376,952	167.8
WTG	-	-	4,884,863	148.8

Table 3: Details of the category-based language models used in rescoring experiments, showing number of categories (N_v), number of *n*-grams (N_{ng}) and perplexity on LM-DEV.

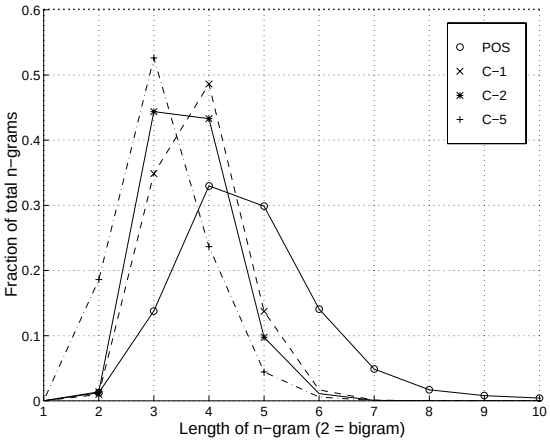

Figure 2: Fraction of total *n*-grams at each n for selected language models from table 3.

N-best lists comprised the top 100 hypotheses, and the interpolation weights were chosen to minimise the word error rate (WER) on R-DEV. Table 4 lists the performance of each of the category-based models listed in table 3.

Considering the best-performing model (C-3), table 5 shows how performance changes when limiting the length of the *n*-grams to n_{max}. The number of *n*-grams of length n_{max} is shown as $N_{ng}(n_{max})$.

6. DISCUSSION

From tables 3 and 4 we see that, even when the number of *n*-grams in the language model is approximately the same (as it is for POS and C-0, for example), a language model using the automatically determined categories exhibits better performance than one using the part-of-speech definitions. Investigation of the former has shown the clustering approach to make better use of the available number of categories by distributing the words more evenly among them. Very frequent words often appear almost as sole occupants of a category, while less frequent words are grouped together. The part-of-speech definitions, on the other hand, group words exclusively on grounds of their grammatical function, and do not take frequencies of occurrence into account. This often

	Perplexity			% WER	
	LM-DEV	R-DEV	R-EVAL	R-DEV	R-EVAL
POS	139.4	185.4	183.2	11.5	12.3
C-0	142.2	189.7	187.1	11.1	12.2
C-1	139.1	186.4	184.4	11.1	11.9
C-2	136.9	184.5	181.9	11.0	11.9
C-3	131.7	180.8	175.7	10.8	11.7
C-4	129.7	179.0	175.0	10.9	11.8
C-5	129.4	179.9	176.1	10.9	12.0
WTG	148.8	206.2	201.8	11.9	12.5

Table 4: Performance when interpolating the baseline trigram with category-based models.

n_{max}	$N_{ng}(n_{max})$	Perplexity		% WER
		LM-DEV	R-EVAL	R-EVAL
2	212,055	152.0	196.5	12.4
3	2,700,175	139.2	181.6	11.8
4	1,467,805	132.8	176.5	11.7
5	266,009	131.8	175.7	11.7
6	28,611	131.7	175.7	11.7
7	2,219	131.7	175.7	11.7
8	268	131.7	175.7	11.7

Table 5: Effect of maximum *n*-gram length on the performance of language model C-3.

results in several frequent words appearing in the same category, as well as infrequent words occupying categories by themselves. Such assignments are not optimal from a statistical point of view.

Linear interpolation of both types of category-based model with the word-based trigram language model has led to reductions in both perplexity as well as word error rate in all cases. The model C-3, which employs 500 categories, results in the lowest word error rate, leading to a relative improvement of 7 %. Further increases in the number of categories shows performance to deteriorate slightly, as the model's ability to generalise to unseen word sequences begins to be undermined. It is this ability which complements the word-based trigram, and hence the performance of the interpolated language model reaches an optimum.

Finally, the results in table 5 show that performance in terms of word error rate is affected most strongly by bigrams, trigrams and 4-grams, to a much smaller extent by 5-grams, and insignificantly by 6-, 7- and 8-grams. This is noteworthy since longer *n*-grams complicate the integration of a language model into a recognition search or lattice rescore.

7. CONCLUSION

The combination of word-based and category-based *n*-gram language models by linear interpolation has been shown to lead to improvements in both perplexity and recogniser word error rate. The number of categories employed by the category-based model reaches an optimum beyond which the performance of the interpolated model begins to deteriorate. Categories determined by an automatic clustering procedure resulted in larger performance improvements than categories based on part-of-speech classifications, notably because the former allows the number of different categories to be increased. Although *n*-grams of arbitrary length were permitted, performance was influenced most strongly by category bigrams, trigrams and 4-grams.

8. ACKNOWLEDGEMENTS

Thomas Niesler was supported by a scholarship from St. John's College, Cambridge, and Ed Whittaker is funded by a studentship from the E.P.S.R.C.

9. REFERENCES

[1] Clarkson, P.R; Rosenfeld, R; *Statistical language modelling using the CMU-Cambridge Toolkit*, Proc. Eurospeech-97, Rhodes, pp. 2707 - 2710, 1997.

[2] Johansson, S; Atwell, R; Garside, R; Leech, G. *The tagged LOB corpus user's manual*; Norwegian Computing Centre for the Humanities, Bergen, 1986.

[3] Katz, S. *Estimation of probabilities from sparse data for the language model component of a speech recogniser*; IEEE Trans. ASSP, vol. 35, no. 3, March 87, pp. 400-1.

[4] Kneser, R; Ney, H; *Improved clustering techniques for class-based statistical language modelling*, Proc. Eurospeech-93, Berlin, pp. 973-976, 1993.

[5] Martin, S; Liermann, J; Ney, H; *Algorithms for bigram and trigram clustering*, Proc. of Eurospeech-95, Madrid, pp. 1253-1256, 1995.

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

[7] Niesler, T.R; Woodland, P.C. *A variable-length category-based n-gram language model*, Proc. ICASSP-96, Atlanta, pp. 164-7, 1996.

[8] Niesler, T.R; Woodland, P.C; *Combination of word-based and category-based language models*, Proc. ICSLP-96, Philadelphia, pp. 220-3, 1996.

[9] Niesler, T.R; *Category-based statistical language models*, PhD thesis, Dept. Engineering, University of Cambridge, U.K., June 1997.

[10] Odell, J.J; Niesler, T.R; *Lattice and language modelling toolkit v2.0*, Reference manual, Entropic Cambridge Research Laboratories Inc., 1996.

[11] Paul, D.B; Baker, J.M; *The design for the Wall Street Journal-based CSR corpus*, Proc. ICSLP-92, pp. 899-902, 1992.

[12] Rosenfeld, R; *Adaptive statistical language modelling : a maximum entropy approach*, PhD thesis, School of Computer Science, CMU, April 1994.

[13] Ueberla, J.P; *More efficient clustering of n-grams for statistical language modelling*, Proc. Eurospeech-95, Madrid, pp. 1257-1260, 1995.

[14] Woodland, P.C; Leggetter, C.J; Odell, J.J; Valtchev, V; Young, S.J; *The 1994 HTK large vocabulary speech-recognition system*, Proc. ICASSP-95, Atlanta, pp. 73-76, 1995.