COMBINATION OF WORD-BASED AND CATEGORY-BASED LANGUAGE MODELS

T.R. Niesler and P.C. Woodland

Cambridge University Engineering Department
Trumpington Street, Cambridge CB2 1PZ, U.K.

ABSTRACT

A language model combining word-based and category-based *n*-grams within a backoff framework is presented. Word *n*-grams conveniently capture sequential relations between particular words, while the category-model, which is based on part-of-speech classific ations and allows ambiguous category membership, is able to generalise to unseen word sequences and therefore appropriate in backoff situations. Experiments on the LOB, Switchboard and WSJ0 corpora demonstrate that the technique greatly improves language model perplexities for sparse training sets, and offers significantly improved complexity versus performance tradeoffs when compared with standard trigram models.

1. INTRODUCTION

Language models using word-categories are intrinsically more compact and better at generalising to unseen word sequences than their word-based counterparts. Despite this, word-based models continue to deliver superior performance by capturing sequential r elationships between particular words, and remain the mainstay of state-of-the-art large-vocabulary speech recognition systems. This paper presents a technique that attempts to retain the advantages of each of these approaches by allowing backoffs to take place from word- to category-based *n*-gram probability estimates.

The category-based language model component of the combined model is based on variable-length word-category *n*-grams[1] [3], and in this work the categories correspond to p art-of-speech classifications as defined in the LOB corpus [1]. Words are permitted to belong to multiple categories, and consequently the model bases its probability estimates on a set of possible classifications of the word history into c ategory sequences. Each such classification has an associated probability, and is updated recursively for each successive word in a sentence during operation of the model. The word based *n*-gram language model component employs the Katz back-off in conjunc tion with Good-Turing discounting [2].

[1] referred to as "**varigram**" models hereafter

2. EXACT MODEL

Consider the following language model[2], which backs off from a word- to a category-based probability estimate :

$$P_{wc}(w \mid \Phi_c) = \begin{cases} P_w(w \mid \Phi_w) & if w \in W_T(\Phi_w) \\ \beta(\Phi_c) \cdot P_c(w \mid \Phi_c) & otherwise. \end{cases} \quad (1)$$

where :

- w is the word for which we would like to estimate the probability of occurrence.
- Φ_w is the word-history upon which the probability estimate of the word-based *n*-gram language model is based, referred to the **word-level context** hereafter. For a trigram it consists of the preceding two words.
- Φ_c is the word history and associated set of category history postulates upon which the probability estimate of the category-based model is based, termed the **category-level context** hereafter. Due to the recursive way in which these cate gory history postulates are maintained [3], the category-level context is in general only completely defined by the entire word history (i.e. to the beginning of the current sentence). Thus the number of category-level contexts for a given w ord-level context is potentially huge, and the mapping from Φ_w to Φ_c is one-to-many.
- $P_w(w \mid \Phi_w)$ is the probability estimate for w obtained from the word-based language model.
- $P_c(w \mid \Phi_c)$ is the corresponding probability obtained from the category-based language model.
- $W_T(\Phi_w)$ is the set of words in word-context Φ_w for which the word-model estimates will be used, backoffs occurring in other cases.
- $\beta(\cdot)$ is the backoff weight, $\beta(\cdot) > 0$.

The estimate (1) has been designed to employ word *n*-grams in capturing significant sequential dependencies between particular

[2] to be denoted by the abbreviation "**WTCBO**" hereafter

words, while the category-based component models less frequent word combinations. From the requirement

$$\sum_{\forall w} P_{wc}(w \mid \Phi_c) = 1.0 \qquad \forall \ \Phi_c \tag{2}$$

it follows from (1) that

$$\beta(\Phi_c) = \frac{1.0 - \sum_{\forall w \in W_T} P_w(w \mid \Phi_w)}{1.0 - \sum_{\forall w \in W_T} P_c(w \mid \Phi_c)} \tag{3}$$

3. APPROXIMATE MODEL

Due to the large number of different possible Φ_c used by the category model, precalculation of $\beta(\cdot)$ according to equation (3) is not feasible. It would be convenient to obtain backoff constants for every Φ_w i nstead, but the dependence of the denominator of (3) upon Φ_c does not permit this. Run-time calculation of $\beta(\cdot)$ using equation (3) increases the computational complexity of a probability calc ulation by a factor or approximately W_T in comparison with a model for which these parameters are precalculated. To circumvent this, we note that the Φ_c is most strongly influenced by the most recent words, and hence make the approximation :

$$P_c(w \mid \Phi_c) \simeq P_c(w \mid \hat{\Phi}_c) \tag{4}$$

where $\hat{\Phi}_c$ is the the category-level context corresponding to Φ_w when assuming no prior knowledge of the words preceding Φ_w. Since there is a unique $\hat{\Phi}_c$ for each Φ_w, we define:

$$\hat{P}_c(w \mid \Phi_w) \stackrel{def}{=} P_c\left(w \mid \hat{\Phi}_c\right) \qquad for \quad \Phi_w \rightarrow \hat{\Phi}_c \tag{5}$$

and therefore we approximate the backoff weights by

$$\beta(\Phi_c) \simeq \beta(\hat{\Phi}_c) = \beta(\Phi_w) \qquad for \quad \Phi_w \rightarrow \hat{\Phi}_c \tag{6}$$

$$\beta(\Phi_w) = \frac{1.0 - \sum_{\forall w \in W_T} P_w(w \mid \Phi_w)}{1.0 - \sum_{\forall w \in W_T} \hat{P}_c(w \mid \Phi_w)} \tag{7}$$

This choice of $\beta(\cdot)$ in general no longer satisfies (2) however, and so we adjust the backoff model (1) as follows :

$$P_{wc}(w \mid \Phi_c) = \begin{cases} P'_w(w \mid \Phi_c) & if \ w \in W_T(\Phi_w) \\ \beta(\Phi_w) \cdot P_c(w \mid \Phi_c) & otherwise. \end{cases} \tag{8}$$

where $P'_w(w \mid \Phi_c)$ is some approximation of $P_w(w \mid \Phi_w)$. Now for (2) to be satisfied, we may choose to define:

$$P'_w(w \mid \Phi_c) \stackrel{def}{=} k(w \mid \Phi_w) \cdot (1 - \beta(\Phi_w)) + \beta(\Phi_w) \cdot P_c(w \mid \Phi_c) \tag{9}$$

while requiring that

$$\sum_{\forall w \in W_T} k(w \mid \Phi_w) = 1.0 \tag{10}$$

The quantity $(1 - \beta(\Phi_w))$ may be interpreted as a probability mass which must be distributed among the elements of W_T by a suitable choice of $k(w \mid \Phi_w)$. The adopted approach is to distribute this mass using the ratio

$$\frac{P_w(w \mid \Phi_w)}{\sum_{\forall w \in W_T} P_w(w \mid \Phi_w)} \tag{11}$$

and hence assign probability mass to *n*-grams in approximately the same proportion as the word-based model. Proceeding from (9), (11), and employing (5) this leads to:

$$\alpha(w \mid \Phi_w) = k(w \mid \Phi_w) \cdot (1 - \beta(\Phi_w)) \tag{12}$$

where, omitting the arguments of $\hat{P}_c(w \mid \Phi_w)$, $P_w(w \mid \Phi_w)$ and $\beta(\Phi_w)$ for brevity,

$$\alpha(w \mid \Phi_w) = \left(1 - \beta + \beta \cdot \sum_{\forall w \in W_T} \hat{P}_c\right) \cdot \frac{P_w}{\sum_{\forall w \in W_T} P_w} - \beta \cdot \hat{P}_c \tag{13}$$

It is easy to show that, for this choice of $k(w \mid \Phi_w)$, equation (10) holds. Furthermore, the approximate backoff (8) converges to the exact backoff (1) as the estimates $\hat{P}_c(w \mid \Phi_w)$ approach the exact values $P_c(w \mid \Phi_c)$ [4]. Finally, although equation (13) guarantees $P'_w(w \mid \Phi_c) > 0$ when the approximation (4) is perfect , it does not do so in general. In order to achieve this, it is sufficient to require that $\alpha(w \mid \Phi_w) \leq 0$ so that from equation (13) it follows that:

$$\beta \geq \frac{P_w}{\hat{P}_c \cdot \left(\sum_{\forall w \in W_T} P_w\right) + P_w \cdot \left(1 - \sum_{\forall w \in W_T} \hat{P}_c\right)} \tag{14}$$

where the arguments of $\beta(\Phi_w)$, $\hat{P}_c(w \mid \Phi_w)$ and $P_w(w \mid \Phi_w)$ have once again been omitted. While calculating $\beta(\Phi_w)$, the equality in equation (14) should be enforced whenever the inequality is v iolated. Referring back to equation (12), this is equivalent to demanding that the probability mass distributed to each word be positive. In practice, this adjustment is required infrequently.

The use of the estimates $\hat{P}_c(\cdot)$ allows the backoff constants $\alpha(w \mid \Phi_w)$ and $\beta(\Phi_w)$ to be precalculated, making the model (8) significantly more computationally efficient than the exact model (1) while continuing to employ the probabilities delivered by the category model in backoff situations.

4. MODEL COMPLEXITY: FINDING W_T

Thus far it has been assumed that, for each word-level context Φ_w, a set of words W_T has been established for which probabilities will be calculated according to the word-based language model. An obvious choice for W_T would be the set of all word s seen within the context Φ_w in the training set. Denote this choice by $\mathbf{W}_T$, and note that when $W_T = \mathbf{W}_T$, backing-off occurs only for truly unseen events.

The approach taken here however has been to reduce the size of W_T by eliminating words which do not afford the word-based model much predictive power in relation to the category-based model. Since this process eliminates *n*-grams from the word-based mod el component, it allows the complexity of the WTCBO language model to be reduced. Note that since the complexity of the category-based component does not change, it sets the minimum overall complexity[3].

To reduce the number of words in W_T we discard those *n*-grams with the smallest effect on the training set likelihood. In particular, an *n*-gram is retained when

$$\Delta\overline{LP} > \delta \tag{15}$$

where $\Delta\overline{LP}$ is the change in mean per-word log probability when using the word- instead of the category-model:

$$\Delta\overline{LP} = \frac{N(w\,|\,\Phi_w)\cdot\left(log\left(P_w(w\,|\,\Phi_w)\right) - log\left(\hat{P}_c(w\,|\,\Phi_w)\right)\right)}{N_{tot}} \tag{16}$$

and N_{tot} is the total number of words in the training set[4].

5. RESULTS

In order to gauge its performance, the described backoff technique has been applied to the LOB, Switchboard and WSJ0 text corpora. In each case language models of various complexities were generated by varying the size of W_T as described in the previou s section, and the resulting perplexities compared with those of a word trigram trained on the same data. The complexity of the latter was controlled by the standard technique of discarding *n*-grams occurring fewer than a threshold number of times in the t raining text (i.e. varying the *n*-gram cutoffs). Identical thresholds were employed for both bigrams and trigrams in all cases.

5.1. LOB corpus

The LOB corpus [2] consists of approximately 1 million words of text drawn from a variety of sources, including for example fiction, news reportage and religious writing. Training- and test-sets were created by splitting the material evenly acro ss these topics in the ratio 95:5, resulting in a vocabulary size of 41097 words. Category language models of differing complexities were built using pruning thresholds of 1e-4 and 5e-6 as described in [3]. Table 1 shows details of the varigram (abbreviated VG) and word-trigram (abbreviated WTG) language models, and figure 1 the performance of the resulting two WTCBO models[5]. In bot h cases these achieve significant perplexity reductions relative to the trigram.

	VG 1 (1e-4)	VG 2 (5e-6)	WTG
Parameters	13,585	44,380	1,142,457
Perplexity	482.04	458.34	413.14

Table 1: Language models for the LOB corpus

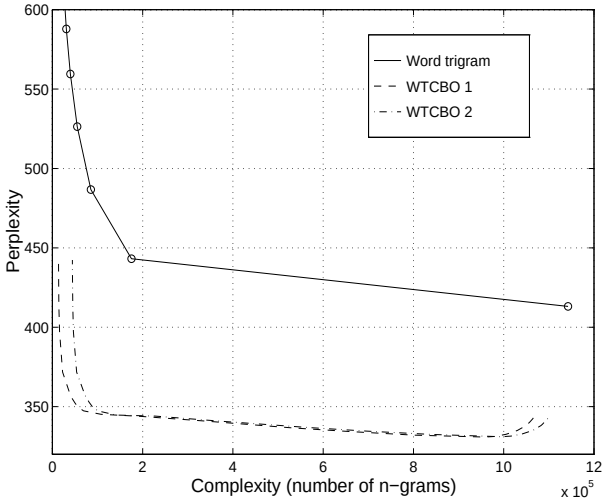

Figure 1: WTCBO models for the LOB corpus

5.2. Switchboard corpus

The Switchboard corpus consists of approximately 1.9 million words of spontaneous telephone conversations concerning a predefined set of topics, and has been the focus of some recent research into conversational speech recognition. A 22,643 word vocabular y closed with respect to the test-set was used, the test-set being the Switchboard dev-test set containing 10,179 words and 1192 sentences. Varigram models were constructed again using pruning thresholds of 1e-4 and 5e-6. Table 2 shows individual model details and figure 2 the performance of the two resulting WTCBO models.

	VG 1 (1e-4)	VG 2 (5e-6)	WTG
Parameters	13,627	54,547	1,183,880
Perplexity	155.40	145.28	96.57

Table 2: Language models for the Switchboard corpus

[3]The overall complexity of the WTCBO model is taken as the sum of the total number of *n*-grams in the word- and category-based components.

[4]Normalisation by the quantity N_{tot} makes $\Delta\overline{LP}$ and thus also δ fairly corpus-independent.

[5]WTCBO 1 and 2 are built using VG 1 and 2 respectively.

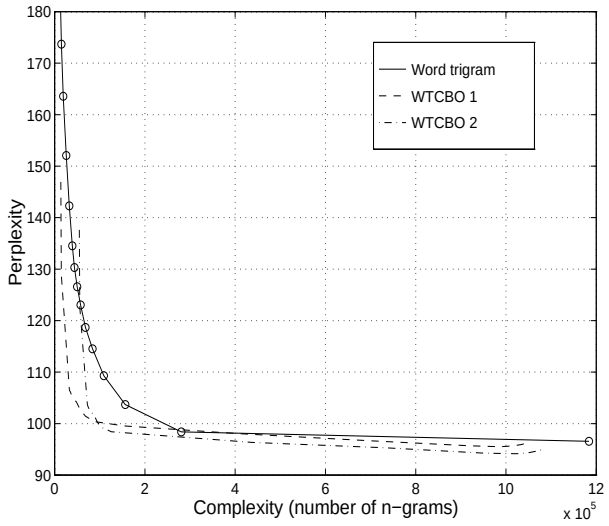

Figure 2: WTCBO models for the Switchboard corpus

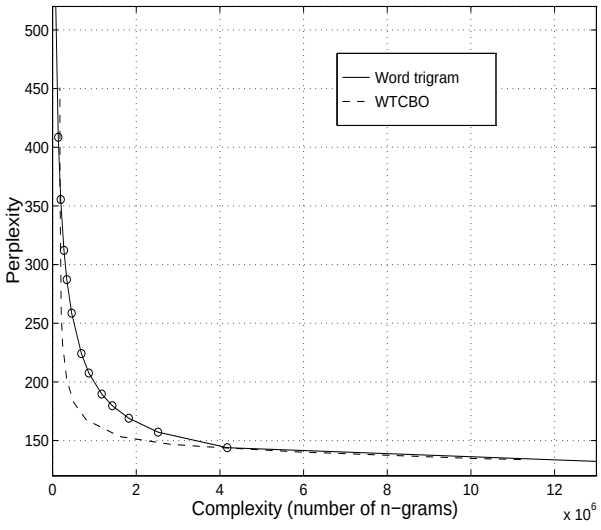

Figure 3: WTCBO models for the WSJ0 corpus

The larger varigram leads to a WTCBO model with lower minimum perplexity and improved performance for model complexities exceeding approximately 60,000 *n*-grams, while the smaller leads to performance that is slightly diminished in this region but better for smaller numbers of *n*-grams. The WTCBO model offers a slight improvement in perplexity with respect to the trigram (approximately 2.7% in figure 2) and, depending on the choice of varigram complexity, a signif icantly improved complexity vs. performance tradeoff characteristic.

The limited number of conversational topics in the Switchboard corpus leads to a reduction in the training-set sparseness and better coverage by the word trigram than for LOB (the backoff rate drops by 41% from the latter to the former). Thus there is le ss need for the generalising ability of the category-model, and consequently a smaller perplexity improvement.

5.3. WSJ0 corpus

This corpus consists of approximately 37 million words of text drawn from the Wall Street Journal over the period 1987-89 inclusive. A 65K vocabulary was used to build language models. The standard 2.1 million word set-aside dev-test text for WSJ0 was us ed as a test-set. Table 3 shows individual language model details, and figure 3 the performance of the WTCBO and trigram models.

	Varigram	Word trigram
Parameters	174,261	13,047,678
Perplexity	481.73	132.21

Table 3: Language models for the WSJ0 corpus

From figure 3 we see that, while in this case the WTCBO model does not offer substantial perplexity improvements over the trigram, it still allows a significantly better complexity vs. performance tradeoff. As was found for the Switchboard corpus, perplexity improvements are small when the word-model is well-trained, which it is for WSJ0 due to the large amount of training data.

6. CONCLUSION

A language model which backs-off from a word-based to a category-based *n*-gram estimate has been introduced. This technique greatly improves perplexities for sparse corpora, and offers significantly better complexity vs. performance tradeoffs when compared with standard trigram models.

7. ACKNOWLEDGEMENT

Thomas Niesler is supported by a scholarship from St. John's College, Cambridge.

8. REFERENCES

1. Johansson, S; Atwell, R; Garside, R; Leech, G. *The tagged LOB corpus user's manual*; Norwegian Computing Centre for the Humanities, Bergen, 1986.
2. 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.
3. Niesler, T.R; Woodland, P.C. *A variable-length category-based n-gram language model*, ICASSP 96, vol. 1, pp. 164-7.
4. Niesler, T.R; Woodland, P.C. *Word-to-category backoff language models*, Tech. report CUED/F-INFENG/TR.258, Dept. Engineering, University of Cambridge, U.K., May 1996.