# Word-to-category backoff language models

T.R. Niesler and P.C. Woodland

CUED/F-INFENG/TR.258

12 May 1996

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

trn@eng.cam.ac.uk
pcw@eng.cam.ac.uk

# 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 classifications 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.

# Contents

# 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 relationships between particular words, and remain the mainstay of state-of-the-art large-vocabulary speech recognition systems. This report presents a technique attempting to retain the advantages of each approach by allowing backoffs to take place from word- to category-based *n*-gram probability estimates.

The category-based component of the combined model is based on variable-length word-category *n*-grams[1] [3], and in this work the categories correspond to part-of-speech classifications as defined in the LOB corpus [1]. Words may belong to multiple categories, and consequently the model bases its probability estimates on a set of possible classifications of the word history into category 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 conjunction with Good-Turing discounting [2].

# 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.
- $\Phi_w$ 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 bigram it is the preceding word, and for a trigram the preceding two words.
- $\Phi_c$ is the word history and associated set of category history equivalence-class postulates upon which the probability estimate of the category-based model is based, termed the **category-level context** hereafter. Due to the recursive nature in which the category history equivalence class 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 is potentially huge, and the mapping from $\Phi_w$ to $\Phi_c$ 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 $\Phi_w$ 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 to capture significant sequential dependencies between particular words, while the category-based component models less frequent word combinations. From the requirement

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

it follows from (1) that

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

[1] Referred to as "**varigram**" models hereafter.
[2] to be denoted by the abbreviation "**WTCBO**" hereafter

# 3. APPROXIMATE MODEL

Due to the large number of different possible $\Phi_c$ used by the category model, precalculation of $\beta(\cdot)$ according to equation (3) is not feasible. It would be more convenient to obtain backoff constants for every word-level context instead, but the dependence of the denominator of (3) upon $\Phi_c$ does not permit this, since it is not uniquely fixed by the word-level context. Run-time calculation of $\beta(\cdot)$ using equation (3) increases the computational complexity of a probability calculation within any particular context by a factor or approximately $W_T$ in comparison with a model for which these parameters are precalculated, and thus represents a significant computational burden. To circumvent this, we note that the category-level context 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 $\Phi_w$ when assuming no prior knowledge of the words preceding $\Phi_w$. Note that there is a unique $\hat{\Phi}_c$ for each $\Phi_w$, so that we may define:

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

and therefore we approximate the backoff weights by

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

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

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

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

where $P'_w(w \mid \Phi_c)$ is some approximation of $P_w(w \mid \Phi_w)$. Now for (2) to be satisfied, it follows from (8) that:

$$\sum_{\forall w \in W_T} P'_w\left(w \mid \Phi_c\right) = \left(1 - \beta\left(\Phi_w\right)\right) + \beta\left(\Phi_w\right) \cdot \sum_{\forall w \in W_T} P_c\left(w \mid \Phi_c\right) \tag{9}$$

so we may choose to define

$$P'_w\left(w \mid \Phi_c\right) \stackrel{def}{=} k\left(w \mid \Phi_w\right) \cdot \left(1 - \beta\left(\Phi_w\right)\right) + \beta\left(\Phi_w\right) \cdot P_c\left(w \mid \Phi_c\right) \tag{10}$$

where, in order to satisfy (9) we require that

$$\sum_{\forall w \in W_T} k(w \mid \Phi_w) = 1.0 \tag{11}$$

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

$$\frac{P_w(w \mid \Phi_w)}{\sum\limits_{\forall w \in W_T} P_w(w \mid \Phi_w)} \tag{12}$$

and therefore distribute probability mass to *n*-grams approximately in the same proportion as the word-based language model. Proceeding from (10), (12), and employing the approximation (5) we are lead to require :

$$\frac{k(w \mid \Phi_w)\cdot(1-\beta(\Phi_w))+\beta(\Phi_w)\cdot\hat{P}_c(w \mid \Phi_w)}{1-\beta(\Phi_w)+\beta(\Phi_w)\cdot\sum\limits_{\forall w \in W_T}\hat{P}_c(w \mid \Phi_w)} = \frac{P_w(w \mid \Phi_w)}{\sum\limits_{\forall w \in W_T} P_w(w \mid \Phi_w)} \tag{13}$$

from which it follows that

$$\alpha\left(w \mid \Phi_w\right) = \left(1-\beta(\Phi_w)+\beta(\Phi_w)\cdot\sum_{\forall w \in W_T}\hat{P}_c(w \mid \Phi_w)\right)\cdot\frac{P_w(w \mid \Phi_w)}{\sum\limits_{\forall w \in W_T} P_w(w \mid \Phi_w)} - \beta(\Phi_w)\cdot\hat{P}_c(w \mid \Phi_w) \tag{14}$$

where

$$\alpha\left(w \mid \Phi_w\right) = k(w \mid \Phi_w)\cdot(1-\beta(\Phi_w)) \tag{15}$$

It is easy to show that, for this choice of $k(w \mid \Phi_w)$, equation (11) holds. Furthermore, as

$$\hat{P}_c\left(w \mid \Phi_w\right) \Rightarrow P_c\left(w \mid \Phi_c\right) \tag{16}$$

we find that

$$\begin{aligned}
P'_w\left(w \mid \Phi_c\right) \quad &\Rightarrow \quad \left(1-\frac{1-\sum\limits_{\forall w \in W_T} P_w\left(w \mid \Phi_w\right)}{1-\sum\limits_{\forall w \in W_T} P_c\left(w \mid \Phi_c\right)}+\frac{1-\sum\limits_{\forall w \in W_T} P_w\left(w \mid \Phi_w\right)}{1-\sum\limits_{\forall w \in W_T} P_c\left(w \mid \Phi_c\right)}\cdot\sum_{\forall w \in W_T} P_c\left(w \mid \Phi_c\right)\right)\cdot\left(\frac{P_w\left(w \mid \Phi_w\right)}{\sum\limits_{\forall w \in W_T} P_w\left(w \mid \Phi_w\right)}\right) \\
&= \quad \left(1-\left(\frac{1-\sum\limits_{\forall w \in W_T} P_w\left(w \mid \Phi_w\right)}{1-\sum\limits_{\forall w \in W_T} P_c\left(w \mid \Phi_c\right)}\right)\cdot\left(1-\sum_{\forall w \in W_T} P_c\left(w \mid \Phi_c\right)\right)\right)\cdot\frac{P_w\left(w \mid \Phi_w\right)}{\sum\limits_{\forall w \in W_T} P_w\left(w \mid \Phi_w\right)} \\
&= \quad P_w\left(w \mid \Phi_w\right)
\end{aligned}$$

which means that the approximate backoff (8) converges to the exact backoff (1) as the estimates $\hat{P}_c\left(w \mid \Phi_w\right)$ approach the exact values $P_c\left(w \mid \Phi_c\right)$.

Finally, although equation (14) guarantees

$$P'_w\left(w \mid \Phi_c\right) > 0 \tag{17}$$

when the approximation (4) is perfect, it does not do so in general. In order to guarantee (17) it is sufficient to require:

$$\alpha(w \mid \Phi_w) \leq 0 \tag{18}$$

so that, using equation (14) it follows that:

$$\beta(\Phi_w) \geq \frac{P_w(w \mid \Phi_w)}{\hat{P}_c\left(w \mid \Phi_w\right) \cdot \left( \sum_{\forall w \in W_T} P_w\left(w \mid \Phi_w\right) \right) + P_w\left(w \mid \Phi_w\right) \cdot \left( 1 - \sum_{\forall w \in W_T} \hat{P}_c\left(w \mid \Phi_w\right) \right)} \tag{19}$$

While calculating $\beta(\Phi_w)$, the equality in equation (19) should be enforced whenever the inequality is violated. Referring back to equation (15), this is equivalent to demanding the probability mass distributed to each word to be positive. In practice, this adjustment is required infrequently.

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

## 4. MODEL COMPLEXITY : DETERMINING $W_T$

Thus far it has been assumed that, for each word-level context $\Phi_w$, 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 words seen within the context $\Phi_w$ 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 whose presence 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 model 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].

The reduction of the number of words in $W_T$ has been accomplished in two ways, both leading to similar results.

### 4.1. Testing *n*-gram counts

Consider a word $w$ which has been seen in in context $\Phi_w$ a total of $N\left(w \mid \Phi_w\right)$ times and denote the number of times the context $\Phi_w$ itself was seen by $N\left(\Phi_w\right)$. Based on the category-model probability estimate $\hat{P}_c\left(w \mid \Phi_w\right)$, the number of times we would expect to see $w$ in $\Phi_w$ is

$$\hat{N}_c\left(w \mid \Phi_w\right) = \hat{P}_c\left(w \mid \Phi_w\right) \cdot N\left(\Phi_w\right) \tag{20}$$

Assuming words to occur independently within each context, thus exhibiting a binomial distribution, the variance on the count $N\left(w \mid \Phi_w\right)$ is :

$$\sigma^2\left(w \mid \Phi_w\right) = \hat{P}_c\left(w \mid \Phi_w\right) \cdot \left(1 - \hat{P}_c\left(w \mid \Phi_w\right)\right) \cdot N\left(\Phi_w\right) \tag{21}$$

Using the normal approximation of the binomial distribution, we determine whether the actual count $N\left(w \mid \Phi_w\right)$ exceeds the expected count $\hat{N}_c\left(w \mid \Phi_w\right)$ by a certain fraction $\delta$ with a certain confidence $\eta$ by testing whether

$$N\left(w \mid \Phi_w\right) - (1+\delta) \cdot \hat{N}_c\left(w \mid \Phi_w\right) > \eta \cdot \sigma\left(w \mid \Phi_w\right) \tag{22}$$

where $w$ is retained in $W_T$ when the test succeeds. In practice the value of $\eta$ is fixed (e.g. to 2.33 for 1% or 1.645 for 5% confidence levels), and the value of $\delta$ is varied to control the size of $W_T$.

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

### 4.2. Testing the effect on overall probability

Testing the *n*-gram counts as described in the previous section allows *n*-grams to be retained in $W_T$ when their counts are seen to differ from the expected ones in a statistically significant manner. The effect of such pruning decisions upon the perplexity of the resulting language model is not clear, however, since a frequent *n*-gram occurring only slightly more often in the word- than in the category-model might be discarded, although it has a significant effect on the overall likelihood due to its high frequency. Therefore a second pruning criterion has been implemented, which discards those *n*-grams with the smallest effect on the training set likelihood first. The perplexity-complexity tradeoff of the resulting models is similar to that achieved with the count-pruning method.

In particular, an *n*-gram is retained when

$$\Delta\overline{LP} > \delta \tag{23}$$

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

$$\Delta\overline{LP} = \frac{N\left(w \mid \Phi_w\right) \cdot \left(log\left(P_w\left(w \mid \Phi_w\right)\right) - log\left(\hat{P}_c\left(w \mid \Phi_w\right)\right)\right)}{N_{tot}} \tag{24}$$

where $N_{tot}$ is the total number of words in the training corpus [4].

## 5. MODEL BUILDING PROCEDURE

In summary of the preceding sections, the following steps need to be taken in order to construct a word-to-category backoff language model :

- Build a category-based language model (here varigram models have been constructed as described in [3]).
- Build a word-based *n*-gram model.
- By application of equation (5), determine the probabilities $\hat{P}_c\left(w \mid \Phi_w\right)$ for each *n*-gram in the word model.
- For each context in the word-model, determine the set $W_T$ using either equation (22) or equation (23). This is essentially a process of pruning *n*-grams from the word model.
- For the remaining *n*-grams in the word-based model, calculate the $\alpha$ and $\beta$ values according to equations (14) and (7).
- Apply the language model according to equation (8).

## 6. 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 previously, and the resulting perplexities compared with those achieved using 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 training text (i.e. varying the *n*-gram cutoffs). Identical thresholds were employed for both bigrams and trigrams in all cases.

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

### 6.1. 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 across 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 the details of these individual language models, and figure 1 the performance of the resulting two WTCBO models[5]. In both cases these achieve significant reductions in perplexity relative to the trigram, and offer more favourable complexity vs. performance tradeoff.

| | Varigram 1 (1e-4) | Varigram 2 (5e-6) | Word trigram |
|---|---|---|---|
| Parameters | 13,585 | 44,380 | 1,142,457 |
| Perplexity | 482.04 | 458.34 | 413.14 |

**Table 1: Language models for the LOB corpus**

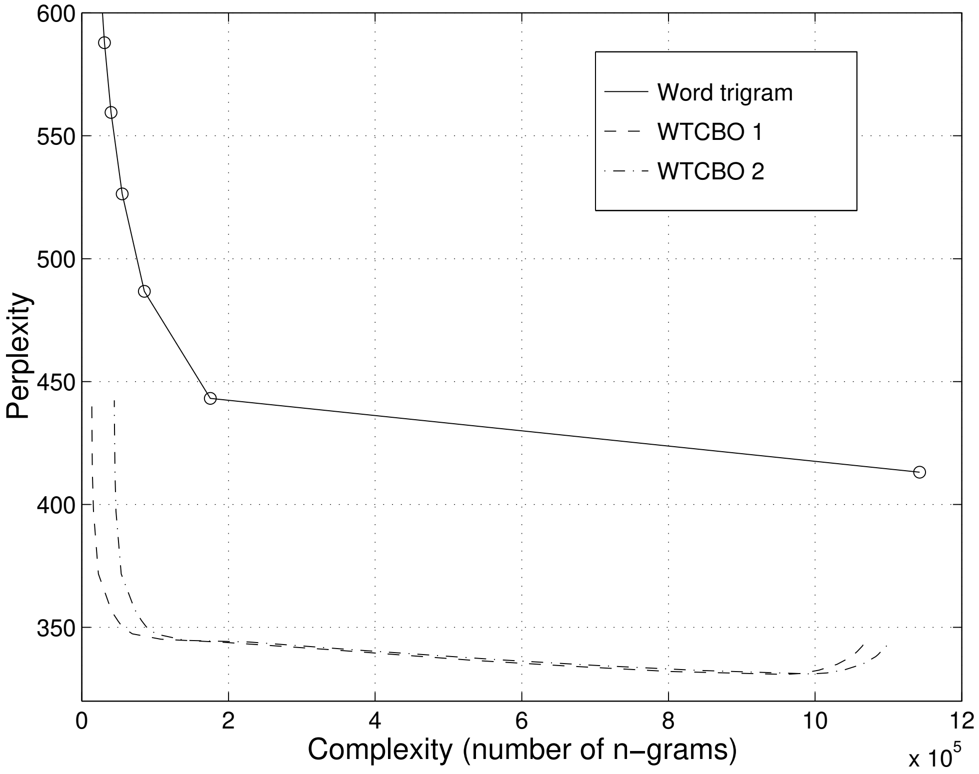


**Figure 1: Performance of word-to-category backoff and trigram language models for the LOB corpus**

### 6.2. Switchboard corpus

The Switchboard corpus consists of approximately 1.9 million words of spontaneous telephone conversations concerning a predefined number of topics, and has been the focus of some recent research into conversational speech recognition. A 22,643 word vocabulary 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 language models were constructed again using pruning thresholds of 1e-4 and 5e-6, and table 2 shows the characteristics of these individual models while figure 2 shows the performance of the resulting two WTCBO models.

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

| | Varigram 1 (1e-4) | Varigram 2 (5e-6) | Word trigram |
|---|---|---|---|
| Parameters | 13,627 | 54,547 | 1,183,880 |
| Perplexity | 155.40 | 145.28 | 96.57 |

**Table 2: Language models for the Switchboard corpus**

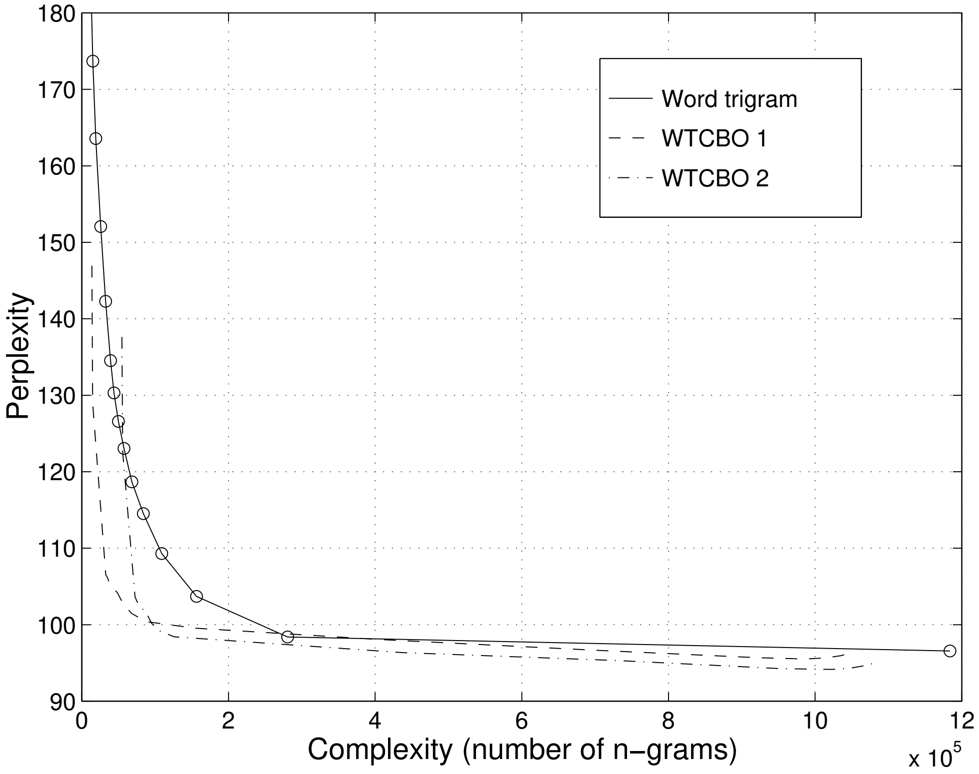


**Figure 2: Performance of the WTCBO and trigram language models for the LOB corpus**

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 significantly 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 less need for the generalising ability of the category-model, and consequently a smaller perplexity improvement.

## 6.3. 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 used 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**

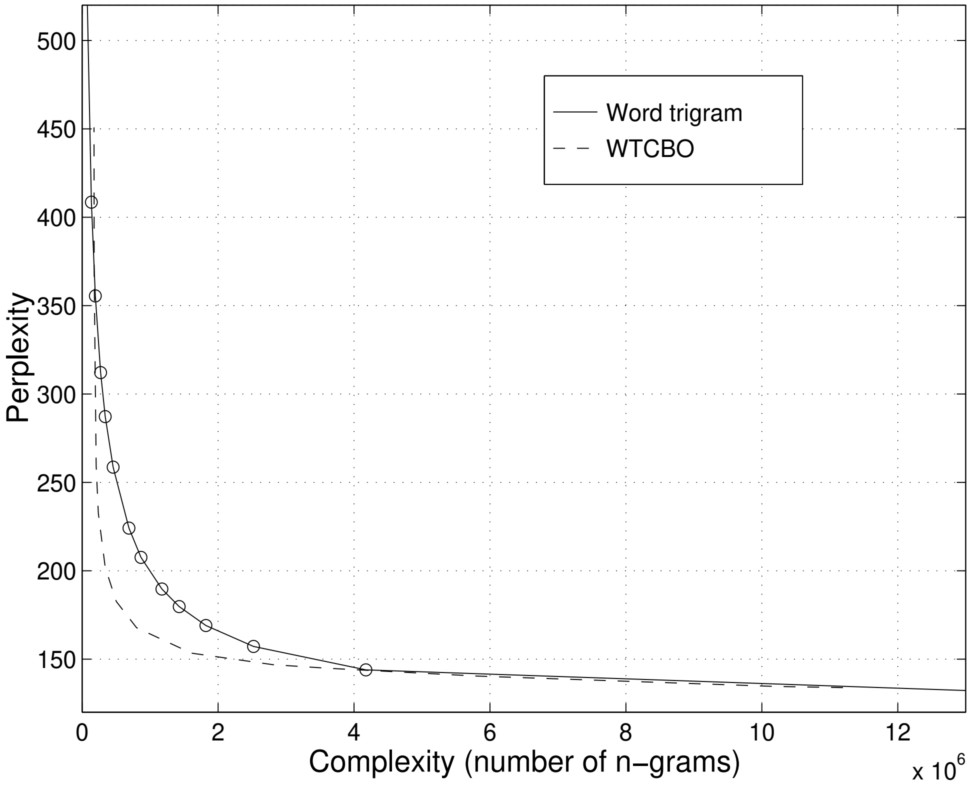


**Figure 3: Performance of word-to-category backoff and trigram language models for the WSJ0 corpus**

From figure 3 we see that, while in this case the WTCBO model does not offer substantial perplexity improvements over the word-based 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.

## 7. SUMMARY AND CONCLUSION

A language model which backs-off from a word-based to a category-based *n*-gram model has been introduced. The category-model, which is based on part-of-speech classifications and allows words to belong to multiple categories, is able to generalise to unseen word sequences and therefore appropriate in backoff situations. This technique greatly improves perplexities for sparse corpora, and offers significantly better complexity vs. performance tradeoffs when compared with standard trigram models.

# 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 1987, pp. 400 - 401.

[3] Niesler, T.R; Woodland, P.C. *A variable-length category-based n-gram language model*, ICASSP 1996, Atlanta. vol. 1, pp. 164-7.