A VARIABLE-LENGTH CATEGORY-BASED N-GRAM LANGUAGE MODEL

T.R. Niesler and P.C. Woodland

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

ABSTRACT

A language model based on word-category n-grams and ambiguous category membership with n increased selectively to trade compactness for performance is presented. The use of categories leads intrinsically to a compact model with the ability to generalise to unseen word sequences, and diminishes the spareseness of the training data, thereby making larger n feasible. The language model implicitly involves a statistical tagging operation, which may be used explicitly to assign category assigments to untagged text. Experiments on the LOB corpus show the optimal model-building strategy to yield improved results with respect to conventional n-gram methods, and when used as a tagger, the model is seen to perform well in relation to a standard benchmark.

1. INTRODUCTION

Word-category based n-grams are a generalisation of their word-based counterparts, being based on category as opposed to word n-tuples. This allows intrinsic generalisation to unseen word sequences, and the smaller number of parameters reduces training-set sparseness. Furthermore, the latter makes larger n feasible both from a statistical as well as a storage viewpoint, a factor which has been seen to have a marked impact on language model quality [1], [8].

A language model based on category n-grams and ambiguous category membership with n increased selectively to trade compactness for performance has been developed. A consequence of the stochastic category-membership is that it allows the model to be employed as a statistical tagger, which is valuable when processing untagged corpora.

2. LANGUAGE MODEL STRUCTURE

Denote a sequence of N temporally consecutive events by $\mathbf{z}(0, N-1) \equiv \{z(0), z(1), \ldots, z(N-1)\}$, and let subscripts identify individual members thereof, so that an alphabet of size K implies $z(i) \in \{z_0, z_1, \ldots, z_{K-1}\}$. Denote sequences of words by the symbol $\mathbf{w}$ and word categories by $\mathbf{v}$. Now let the word-category relationship be described by

$$v_j = G(w_i) \qquad j \in \{0, 1, \ldots, N_{wc}-1\} \tag{1}$$

where N_{wc} is the number of different word categories. Let each word history $\mathbf{w}(0,b)$ be classified into particular equivalence class s_i defined to be an n-gram of categories:

$$s_i = S(\mathbf{w}(0,b)) = \{v(a), v(a+1), \cdots, v(b)\} \tag{2}$$

where $i \in \{0, 1, \cdots, N_{hc}-1\}$, $v(i) \in G(w(i))$, $0 \le a \le b$, and N_{hc} is the number of history equivalence classes. Since a word may belong to several categories, both G and S are in general one-to-many, and $\mathbf{w}(a,b)$ may map to multiple history equivalence classes. Assuming $P(w(i))$ to be wholly determined by $v(i)$:

$$P(w(i)|\mathbf{w}(0,i-1)) = \sum_{\forall v: v \in G(w(i))} P(w(i)|v) \cdot P(v|\mathbf{w}(0,i-1)) \tag{3}$$

Assuming furthermore that the probability of witnessing $v(i)$ depends only on the category n-gram context, the right-hand side of (3) may be decomposed further:

$$P(v|\mathbf{w}(0,i-1)) = \sum_{\forall s: s \in S(\mathbf{w}(0,i-1))} P(v|s) \cdot P(s|\mathbf{w}(0,i-1)) \tag{4}$$

The subsequent three sections treat the estimation of each component probability in (3) and (4) individually.

2.1. Estimating $P(v_j|s_m)$

For compact storage of category n-grams , we employ a tree data-structure associating each node with a particular word category, so that paths originating at the root correspond to category n-grams . From definition (2) this implies that each node represents a distinct history equivalence class[1] s_m, and therefore has associated with it a conditional probability density function $P(v|s_m)$. By not restricting the length of the individual paths through the tree, contexts of arbitrary depth are catered for. For example, in the following figure s_5 corresponds to the trigram context $v(i-2, i-1) = \{v_2, v_8\}$.

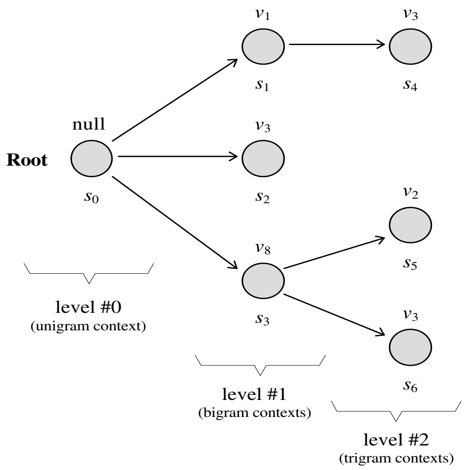

Figure 1: Illustration of a language model tree

[1] The set of all nodes therefore constitutes the set of all possible equivalence classes.

The probabilities $P(v|s_m)$ are estimated by application of Katz's back-off [5] in conjunction with nonlinear discounting [6]. Model building proceeds via the following level-by-level tree growing strategy which retains only contexts that improve a language model quality criterion, and thus allows model compactness to be maintained while employing longer n-grams where they benefit performance [7].

1. **Initialisation** : $L = -1$
2. $L = L + 1$
3. **Grow** : Add level $\#L$ to level $\#(L-1)$ by adding all the $(L+1)$-grams occurring in the training set for which the L-grams already exist in the tree.
4. **Prune** : For every (newly created) leaf in level $\#L$, apply a quality criterion, and discard the leaf if it fails.
5. **Termination** : If there are a nonzero number of leaves remaining in level $\#L$, goto step 2.

The quality criterion checks for an improvement in the leaving-one-out [2] cross-validation[2] training set likelihood by the addition of each leaf. In particular, the training set likelihood may be written as a sum of contributions of all the nodes in the tree:

$$LL_{cum}\left(\Omega^{tot}\right) = \sum_{n=0}^{N_n-1} LL_{cum}^{s_n} \tag{5}$$

where

$$LL_{cum}^{s_n} = \sum_{k=0}^{N_{vv}} N_{s_n}(v_k) \cdot \log\left(P\left(v_k|s_n, \Omega_k^{RT}\right)\right) \tag{6}$$

and where N_n is the number of nodes in the tree, LL_{cum} the log probability of the training corpus Ω^{tot}, $LL_{cum}^{s_n}$ the log probability associated with all events occurring in s_n, $N_{s_n}(v_k)$ the number of times v_k was seen in s_n in Ω^{tot}, N_{vv} the number of different categories, and $P\left(v_k|s_n, \Omega_k^{RT}\right)$ the probability of v_k occurring in s_n based on the retained part of the training set Ω_k^{RT} formed when v_k is the heldout part. Taking $\Delta LL_{cum}^{s_n}$ to denote the change in $LL_{cum}^{s_n}$ caused by the addition of the leaf, the pruning criterion is:

$$\Delta LL_{cum}^{s_n} > -\lambda \cdot LL_{cum}\left(\Omega^{tot}\right) \tag{7}$$

This requires the addition of the new node to lead to a likelihood increase of at least a threshold defined as a fraction of the total likelihood, so as to make the choice of the threshold fairly problem independent.

$P(v_j|s_m)$ may be used to calculate a perplexity indicating the confidence with which the tree predicts the following category. This is later used in language model evaluation, and will be referred to as the **category perplexity**.

2.2. Estimating $P(w_i|v_j)$

Assume each category has a sufficiently large membership to allow application of the relative frequency estimate:

$$P(w_i|v_j) = \frac{N(w_i|v_j)}{N(v_j)} \tag{8}$$

[2]Employed to avoid overfitting of the training data.

Since the language model must hypothesise categories for out-of-vocabulary (OOV) words, the probability with which these occur within each category must be estimated. Accordingly an entry "UW" is added to each category, and its count N_{uw} estimated by leaving-one-out [7] :

$$P(UW|v_j) = \frac{N_1(v_j)}{N(v_j) + \eta} \tag{9}$$

and

$$N_{uw}(v_j) = \frac{P(UW|v_j) \cdot N(v_j)}{1 - P(UW|v_j)} \tag{10}$$

where $N_1(v_j)$ is the number of words seen exactly once in both v_j and the training set, $N(v_j)$ the total number in v_j, $N_{uw}(v_j)$ is the corresponding estimated count for UW in v_j, and $\eta > 0$ a small constant introduced heuristically to ensure that the denominator of (9) is always less than one. Its effect is significant only for categories with small $N(v_j)$ (sparsely trained). The effect of η on performance was seen empirically to be weak, and $\eta \approx 5 \ldots 10$ yields satisfactory results for the LOB corpus [7].

2.3. Estimating $P(s_m|\mathbf{w}(0, i-1))$

The set of contexts to which a particular word history $\mathbf{w}(0, i-1)$ may belong as well as the probabilities associated with each may be calculated by means of a recursive approach. First define:

$\mathbf{v}_j^{hyp}(0, b)$: A possible classification of $\mathbf{w}(0, b)$ into word categories (termed a **hypothesis** hereafter).

$s_m = F_{tree}(\mathbf{v}(0, b))$: The history equivalence class corresponding to the deepest match of $\mathbf{v}(0, b)$ within the n-gram tree.

$N_{H(0,b)}$: The number of hypotheses for $\mathbf{w}(0, b)$.

Given expressions for $P\left(\mathbf{v}_j^{hyp}(0,i)\,|\,\mathbf{w}(0,i)\right)$, the desired probability of the history equivalence class may be found:

$$P(s_m|\mathbf{w}(0,i)) = \sum_{\forall j: F_{tree}\left(\mathbf{v}_j^{hyp}(0,i)\right) = s_m} P\left(\mathbf{v}_j^{hyp}(0,i) \mid \mathbf{w}(0,i)\right) \tag{11}$$

Explicit maintenance of hypotheses (as opposed to history equivalence classes) is necessary due to the varying lengths of the n-grams.

Given a set of existing hypotheses $\{\mathbf{v}_j^{hyp}(0, i-1)\}$, the set of new hypotheses is $\{\mathbf{v}_j^{hyp}(0, i-1), v_k\}$ for all (j, k) such that $j = \{0, 1, \ldots, N_{H(0,i-1)}-1\}$ and $k = \{0, 1, \ldots, N_{vv}-1\}$ where N_{vv} is the number of different POS categories. Consider now the particular postulate $\mathbf{v}_{j'}^{hyp}(0, i) = \{\mathbf{v}_j^{hyp}(0, i-1), v_k\}$, the prime over the index indicating that there is in general no fixed relation between the ordering of the two sets of hypotheses. Recalling the assumptions for equation (3) it follows that

$$\begin{aligned} P\left(\mathbf{w}(0,i) \mid \mathbf{v}_{j'}^{hyp}(0,i)\right) &= \prod_{k=0}^{i} P\left(w(k) \mid v_{j'}^{hyp}(k)\right) \\ &= P\left(w(i) \,|\, v_{j'}^{hyp}(i)\right) \cdot P\left(\mathbf{w}(0,i-1) \mid \mathbf{v}_{j'}^{hyp}(0,i-1)\right) \end{aligned} \tag{12}$$

and, from the n-gram model structure,

$$P\left(\mathbf{v}_{j'}^{hyp}(0,i)\right) = \prod_{k=0}^{i} P\left(v_{j'}^{hyp}(k) \,|\, F_{tree}\left(\mathbf{v}_{j'}^{hyp}(0,k-1)\right)\right)$$

$$= P\left(v_{j'}^{hyp}(i) \mid F_{tree}\left(\mathbf{v}_{j'}^{hyp}(0,i-1)\right)\right) \cdot P\left(\mathbf{v}_{j'}^{hyp}(0,i-1)\right) \quad (13)$$

where $\mathbf{v}_{j'}^{hyp}(0,-1)$ is the single initial null hypothesis and $F_{tree}\left(\mathbf{v}_{j'}^{hyp}(0,-1)\right)$ the associated unigram context, so that $P\left(\mathbf{v}_{j'}^{hyp}(0,-1)\right) = 1$. From (12) and (13) it follows using Bayes rule that

$$P\left(\mathbf{w}(0,i), \mathbf{v}_{j'}^{hyp}(0,i)\right) = P\left(w(i) \,|\, v_{j'}^{hyp}(i)\right) \cdot P\left(v_{j'}^{hyp}(i) \,|\, F_{tree}\left(\mathbf{v}_{j'}^{hyp}(0,i-1)\right)\right) \cdot P\left(\mathbf{w}(0,i-1), \mathbf{v}_{j'}^{hyp}(0,i-1)\right) \quad (14)$$

At instant i, the most likely postulate is that for which $P\left(\mathbf{v}_j^{hyp}(0,i) \,|\, \mathbf{w}(0,i)\right)$ is a maximum. Since $N_{H(0,i)}$ becomes extremely large as i increases, it is in practice necessary to restrict storage to the N_H^{max} most likely candidates. Letting $\mathbf{v}_q^{hyp}(0,i)$ refer to the q_{th} most likely hypothesis,

$$P\left(\mathbf{v}_j^{hyp}(0,i) \,|\, \mathbf{w}(0,i)\right) = \frac{P\left(\mathbf{w}(0,i), \mathbf{v}_j^{hyp}(0,i)\right)}{\sum_{q=0}^{N_H^{max}} P\left(\mathbf{w}(0,i), \mathbf{v}_q^{hyp}(0,i)\right)} \quad (15)$$

The summation in the denominator over only N_H^{max} hypotheses effectively proportionally distributes the probability mass associated with the discarded hypotheses over those retained, thereby ensuring that

$$\sum_{q=0}^{N_H^{max}} P\left(\mathbf{v}_q^{hyp}(0,i) \,|\, \mathbf{w}(0,i)\right) = 1 \quad (16)$$

as demanded by the language model (4). Since the denominator of (15) is common to all new hypotheses, the choice of the N_H^{max} best candidates may be made by considering the joint probabilities given by (14) rather than the conditional probabilities given by (15).

This procedure maintains a fixed maximum number of hypotheses, a significant number of which often have very low associated probabilities. When computational efficiency is an important issue[3], these may be discarded by means of a beam pruning meachanism [7].

2.4. Statistical tagging

Since the language model maintains a set of hypotheses for the input text string for which $P(\mathbf{v}(0,N-1) \,|\, \mathbf{w}(0,N-1))$ is highest, it implicitly maintains a set of the most probable category assignments for each word in $\mathbf{w}(0,N-1)$. When the categories are part-of-speech (POS) classifications, this allows statistical tagging of unlabelled text.

[3] When tagging large quantities of text, for example.

3. EXPERIMENTAL RESULTS

3.1. Building n-gram trees (LOB corpus)

Employing the method of section 2.1, n-gram language model trees with categories corresponding to POS word classes were constructed for various pruning thresholds, λ, using 95% of the LOB corpus [4], the remaining 5% forming the test-set. Tree complexities[4] and category perplexities are shown in figure 2, each point being labelled with the corresponding threshold value. In addition, perplexities obtained when pruning by simply thresholding the total number of occurrences of an event in the training-set are shown for various choices of this threshold (termed a count threshold, "CT"). This technique is commonly employed in making n-gram models more compact.

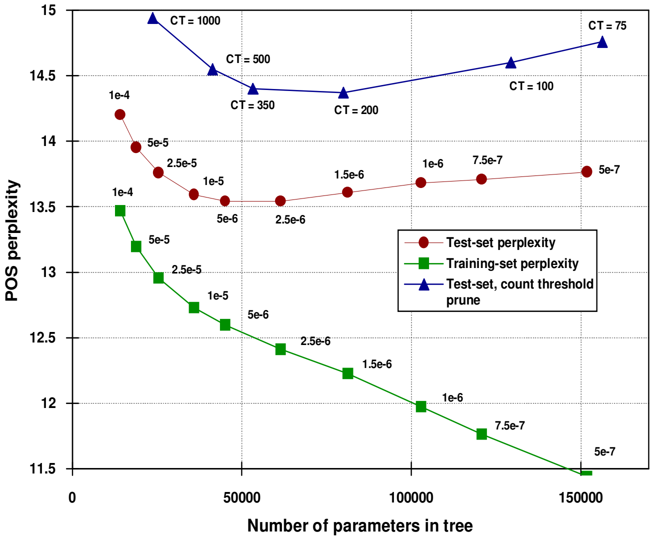

Figure 2: Language models for the LOB corpus

Figure 2 shows that, as the model complexity increases, the test-set perplexity moves through a minimum. The initial decrease is due to underfitting, and the subsequent increase to overfitting of the data. Overfitting does occur, since leaving-one-out cross-validation models the test-set only approximately, but has been reduced significantly in comparison with the use of the count-thresholds. The optimal model ($\lambda = 5e-6$) has a significantly lower perplexity than a tree of comparable size obtained by count-threshold pruning.

3.2. Word-perplexities (LOB corpus)

Three trees were constructed using $\lambda = 5e-6$, and used in the language model of section 2. The first two are bigram and trigram structures, obtained by stopping growth beyond levels 1 and 2 respectively. The third, obtained by allowing the tree-growing algorithm to execute to completion, is referred to here as a **varigram**. Table 1 shows word perplexities obtained for each model for various N_H^{max}.

The word perplexities decrease monotonically as N_H^{max} increases, demonstrating that the history equivalence class

[4] The total number of parameters in the tree has been taken as a measure of its complexity.

N_H^{max}	1	2	4	10
Bigram	671.3	610.2	604.1	603.2
Trigram	634.7	555.2	545.2	544.1
Varigram	629.3	548.9	536.7	534.1

Table 1: Word perplexities for the LOB corpus

ambiguity has a significant effect on language model performance. The results indicate that $N_H^{max} \approx 5 \ldots 10$ yields near-optimal results. Furthermore, the longer contexts of the varigram lead to a drop in perplexity with respect to the bigram and trigram structures.

A word-based trigram language model for the same corpus achieves a perplexity of 474 but contains 986,892 parameters. Thus a 11.3% decrease in perplexity is accompanied by an almost 22-fold increase in complexity.

3.3. Tagging accuracies

The varigram language model trained on the LOB training-set was used to tag the test-set. Since the tagging accuracy for OOV words is significantly lower than for words within the vocabulary, the effect of augmenting the lexicon with words from additional sources was investigated. These sources consisted of (1) the Oxford Advanced Learner's Dictionary (available electronically), (2) 5000 frequent names and surnames, and (3) OOV genitive forms of words whose baseforms were already in the vocabulary. In effect, the existing category n-grams were called on to generalise to the new words. This is not possible with word-based n-grams, for which new statistics must be gathered for each new entry. Table 2 shows tagging accuracies (TA) for the varigram (VG) model with augmented (A) and unaugmented (UA) lexica. Corresponding figures (i.e. employing the same training- and test-set) for the ACQUILEX (AQLX) tagger [3] are provided as a benchmark.

The performance of both taggers is similar, but the varigram exhibits an overall improvement as well as a considerable reduction in tagging errors for OOV words. These differences are attributed to both the longer n-gram contexts, as well as the method used to calculate unknown word probabilities. Lexicon augmentation more than halves the OOV rate, and improves the overall tagging accuracy.

	AQLX	VG (UA)	VG (A)
% OOV words	2.51	2.51	1.05
TA (overall)	94.03	95.13	95.82
TA (non-OOV)	95.77	96.31	96.26
TA (OOV words)	31.17	49.30	54.55

Table 2: LOB corpus % tagging accuracies

3.4. Word-perplexities (Switchboard corpus)

The Switchboard corpus consists of 1,860,178 words of recorded telephone speech, and has been the focus of some recent research into conversational speech recognition. A varigram language model was constructed for this corpus using a pruning threshold of 5e-6 and a 22,643 word vocabulary closed with respect to the test-set. Since the corpus is not annotated with part-of-speech information, it was tagged using the varigram built on the LOB corpus with the augmented lexicon as described in the previous section. Table 3 shows the performance of the resulting varigram as well as baseline word-based bigram and trigram models for the Switchboard dev-test set (10,179 words and 1192 sentences). When compared with the trigram, the varigram achieves a 49% higher perplexity but contains 4.5% as many parameters. The lower perplexities and larger difference in performance between the word and category models when compared with section 3.2 may be ascribed to the greater homogeneity (in style and topic) of the Switchboard corpus, as well a the larger amount of training data.

	Word bigram	Word trigram	Varigram
Parameters	305,605	1,201,176	54,547
Perplexity	108.57	92.94	138.53

Table 3: Word perplexities for the Switchboard corpus

4. CONCLUSION

A category-based language model capable of doubling as a statistical tagger and employing n-grams of varying lengths has been described. A model-building procedure optimising compactness with respect to performance has been presented, and experiments using the LOB corpus show language models constructed in this way to outperform conventional n-gram approaches. The model is most effective when dealing with corpora that are sparse due to small size or heterogeneous composition, since then the intrinsic ability to generalise to unseen word sequences is of maximum benefit.

5. REFERENCES

[1] Bahl, L; Brown, P; de Souza, P; Mercer, R. *A tree-based statistical language model for natural language speech recognition*, IEEE Trans. ASSP, vol. 37, no. 7, July 1989.

[2] Duda, R., Hart, P. ; *Pattern classification and scene analysis*; Wiley, New York, 1973.

[3] Elworthy, D. *Tagger suite user's manual*, May 1993.

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

[5] 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.

[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. *Variable-length category-based n-grams for language modelling*, Tech. report CUED/F-INFENG/TR.215, Dept. Engineering, University of Cambridge, U.K., April 1995.

[8] Shannon, C.E. *Communication theory : exposition of fundamentals*, IRE Trans. Inf. Th., no. 1, Feb. 1950.