COMBINATION OF WORDS AND WORD CATEGORIES IN VARIGRAM HISTORIES

Reinhard Blasig

Philips Research Laboratories
Weisshausstr. 2, D-52066 Aachen, Germany
email: blasig@pfa.research.philips.com

ABSTRACT

This paper presents a new kind of language models: *category/word varigrams*. This special model type permits a tight integration of word-based and category-based modeling of word sequences. Any succession of words and word categories may be employed to describe a given word history. This provides a much greater flexibility than previous combinations of word-based and category-based language models.

Experiments on the WSJ0 corpus and the 1994 ARPA evaluation data indicate that the category/word varigram yields a perplexity reduction of up to 10 percent as compared to a word varigram of the same size, and improves the word error rate (WER) by 7 percent. Compared to a linear interpolation of a word-based and a category-based n-gram, the WER improvement is about 4 percent.

1. INTRODUCTION

Variable length n-grams (varigrams) have proven to be efficient tools for language modeling. So far, word-based [3, 5] and category-based [1,4,7] varigrams have been applied, as well as combinations thereof using linear interpolation or a backing-off scheme [8] The appealing characteristic of the combined models is their capability of utilizing the good generalization qualities of category models with respect to unseen word sequences, while retaining the accuracy of word models to effect specific word predictions

In this paper, we present a tighter coupling of the word-based and the category-based approach. The new model provides the expressiveness to describe a word history by any sequence of words and categories. For example, p(ENGINE | THREE POINT FOUR LITER) could be calculated using p(ENGINE | C30 POINT C30 LITER), where C30 is a class containing numeral words.

The paper first reviews our word-based varigram, which incorporates a backing-off structure with absolute discounting. A generalization of the technique of marginal constraints [2] is presented that permits a direct multilevel backoff, i.e. for example a direct backoff from a fourgram to a bigram. We introduce the notion of *effective counts*, which presents the marginal constraint approach from a new perspective and is helpful to generalize it to the case of combined category/word varigrams.

The varigram models presented in this paper have the special characteristic that a given word history may have different descriptions of identical length, as illustrated in the example above. Note that different history descriptions of different length are standard in n-gram language models. Their individual word predictions are usually combined using backing-off techniques.

We investigate two variants of combining these descriptions to produce the probability estimation for the predicted word. The first variant selects from all equally long history descriptions a single one according to a perplexity criterion. The second variant combines the predictions of the history descriptions by calculating an average. Section 3.1 deals with the mathematical formulation of the two variants. They differ not only in their way of calculating probabilities, but also in their approach of counting word sequence events in the training corpus, as well as in their pruning criterion. Section 4 contains the experimental results.

2. THE WORD-BASED VARIGRAM

We start from a word-based varigram $\mathcal{L}$ with absolute discounting and backing-off. Given $h = (w_{i-k}, \ldots, w_{i-1})$ as a word history of length k,[1] the probability of the successor word $w = w_i$ is calculated according to

$$p(w|h_k) = \begin{cases} \alpha(w|h_k) + \gamma(h_k)p(w|h_{k-1}) & \text{if } (h_k, w) \in \mathcal{L} \\ \gamma(h_k)p(w|h_{k-1}) & \text{if } (h_k, w) \notin \mathcal{L}, \\ & \exists w' : (h_k, w') \in \mathcal{L} \\ p(w|h_{k-1}) & \text{else.} \end{cases} \quad (1)$$

with

$$\alpha(w|h_k) = \frac{d(N(h_k, w))}{N(h_k)} \quad (2)$$

$$\gamma(h_k) = \frac{d_k N_+(h_k) + N(h_k) - \sum_{(h_k, w) \in \mathcal{L}} N(h_k, w)}{N(h_k)} \quad (3)$$

Here, the language model $\mathcal{L}$ is perceived as a set of word sequences (h, w) with associated counts $N(h, w)$ The term $N_+(h)$ denotes the number of distinct words stored as successors of h in the language model, i.e.:

$$N_+(h) = \sum_{w : (h, w) \in \mathcal{L}} 1 \quad (4)$$

It has to be mentioned that the backing-off scheme is applied down to the unigram distribution $p(w|h_0) = p(w)$ and to the zerogram distribution. The function $d()$ implements absolute discounting In our system, the discouting value depends both on the value $N(h_k, w)$ to be discounted and on the length k of the respective history h_k:[2]

$$d(N(h_k, w)) = \begin{cases} 0 & N(h_k, w) < d_k \\ N(h_k, w) - d_k & \text{else} \end{cases} \quad (5)$$

This work was partially funded by the German Federal Ministry for Education, Science, Research and Technology (BMBF) in the framework of the Verbmobil Project under Grant 01 IV 701 C9 The responsibility for the contents of this study lies with the authors.

[1] We use notation h_k, if we want to stress that history h has length k

[2] Due to the limited space, the following derivations have been done for $d_k = 1$ In the actual implementation $d_k \in [0, 1]$ have been used

The term $\gamma(h)$ is used as a normalization factor and makes sure that $\sum_{w \in V} p(w|h_k) = 1$ given that the backoff distribution $p(w|h_{k-1})$ is normalized. The normalization condition is trivially true for the zerogram: $\sum_{w \in V} \frac{1}{|V|} = 1$.

2.1. Marginal Constraints

Kneser and Ney [2] describe in depth the ideas behind marginal constraint language models. According to this concept, the terms $\alpha(w|h)$ and $\gamma(h)$ in (1) have to be substituted by

$$\alpha_M(w|h_k) = \frac{d\left(N(h_k, w) - \sum_{h' >_{\mathcal{L}} h_k} d(N(h', w))\right)}{N(h_k) - \sum_{w \in V, h' >_{\mathcal{L}} h_k} d(N(h', w))} \quad (6)$$

$$\gamma_M(h_k) = \frac{d_k N_+(h_k) + N(h_k) - \sum_{(h_k, w) \in \mathcal{L}} N(h_k, w)}{N(h_k) - \sum_{w \in V, h' >_{\mathcal{L}} h_k} d(N(h', w))} \quad (7)$$

The term $\alpha_M(w|h)$ permits a direct multilevel backoff. This is accomplished by introducing the relation $>_{\mathcal{L}}$. For a word-based language model $\mathcal{L}$ we have $h' >_{\mathcal{L}} h$, iff both h and h' are histories stored in $\mathcal{L}$ (i.e. $\exists w, w' : (h, w) \in \mathcal{L}$ and $(h', w') \in \mathcal{L}$), and h' is a history extension of h, and there is no h'' in $\mathcal{L}$, such that h'' is an extension of h and h' is an extension of h''.

2.2. The Law of "Effective Counts"

The comparison of $\alpha(w|h)$ and $\alpha_M(w|h)$ suggests the definition of an *effective count* $N_{\mathit{eff}}(h, w)$, which we define as:

$$N_{\mathit{eff}}(h, w) = d\left(N(h, w) - \sum_{h' >_{\mathcal{L}} h} d(N(h', w))\right). \quad (8)$$

The effective count is a measure for the probability mass attributed to a word sequence $(h, w) \in \mathcal{L}$ among the set of all word sequences included in the language model $\mathcal{L}$. Summing up the effective counts for all $(h, w) \in \mathcal{L}$ yields:

$$\begin{aligned}
&\sum_{(h,w) \in \mathcal{L}} N_{\mathit{eff}}(h, w) \\
&= \sum_{(h,w) \in \mathcal{L}} d\left(N(h, w) - \sum_{h' >_{\mathcal{L}} h} d(N(h', w))\right) \Bigg| d_k = 1 \\
&= \sum_{(h,w) \in \mathcal{L}} \left(N(h, w) - \sum_{h' >_{\mathcal{L}} h} N(h', w)\right) + \sum_{(h,w) \in \mathcal{L}, h' >_{\mathcal{L}} h} 1 \; - |\mathcal{L}|1 \\
&= \left(\sum_{(h',w) \in \mathcal{L} \text{ with } \forall (h,w) \in \mathcal{L} \setminus (h',w):\ h' >_{\mathcal{L}} h} N(h', w)\right) + |\mathcal{L}| - |\mathcal{L}| \\
&= N_{\text{Corpus}} \quad (9)
\end{aligned}$$

Thus, for each processed word w in the training corpus, the sum of effective counts increases by 1. Generally, the effective count will increase for the single (h', w) in $\mathcal{L}$, where h' is the most specific word sequence (according to the order $<_{\mathcal{L}}$) that matches the history of w. Only if discounting effects cause a smaller increase δ of $N(h', w)$, the remainder $1 - \delta$ will be distributed on other (h, w) with less specific h.

3. THE CATEGORY / WORD VARIGRAM

In n-gram language models, be they word-based or class-based, word histories ending on the same $n - 1$ words (or categories) are grouped into one equivalence class and thus provide identical predictions. The motivation for the language model presented in this paper is to provide higher flexibility in the definition of history classes and thus permit a better optimization of the following two conflicting criteria:

- History classes should be as large as possible. The more often members of the class have been seen in the training corpus, the more reliable can the predictions be.
- The members of a history class should be homogeneous in terms of their word predicitions.

In the framework of the category/word-based language models presented here, histories can be described by arbitrary sequences of words and word categories ("history descriptions"). The categories are optimized using bigram statistics of the given training corpus [1]. Each word is member of exactly one category.[3]

The following table illustrates history descriptions of varying length, as they were created based on the WSJ0 training corpus. The quantity $\Delta_{LL}(b)$ is the difference in training corpus log likelihood of a language model containing history description b as compared to a language model not including b. This quantity is calculated on a leaving-one-out basis [6].

The categories $Cxxx$ appearing in the table are illustrated by enumerating some of their members in the category list following the table. ⟨UNK⟩ denotes a word not contained in the language model vocabulary, and ⟨/s⟩ symbolizes the end of a sentence.

$\frac{\Delta_{LL}(b)}{N(b)}$	h	w
13.390	CHESEBROUGH	POND'S
13.310	AVANT	GARDE
10.029	COCA	COLA
9.637	SHUTTLE'S	BOOSTER
14.457	COMPANY DU	MIDI
14.414	ZIA ⟨UNK⟩	HAQ
13.544	DIVIDE AND	CONQUER
13.095	SPOKESMAN TERRY	EASTLAND
14.938	CONSTRUCTION RAIL AND	TUBULAR
14.846	M T. U	MOTOREN
13.423	LAIDLAW ADAMS AND	PECK
12.241	⟨UNK⟩ DAYS C031	NUMBERED
13.597	FRED LANGE PRESIDENT OF	LANGE
13.587	C001 HOUSTON INVESTOR CHARLES	HURWITZ
13.556	C247 CULTURE HE HAD	NURTURED
13.245	C377 THAT C108 THE	MICROCHIPS
17.330	EQUIPPED WITH C030 POINT C030	LITER
15.227	KRAMER LEVIN ⟨UNK⟩ ⟨UNK⟩ AND	FRANKEL
14.197	C009 THE MANAGEMENT BID WILL	FALTER
12.887	⟨/s⟩ A NINETEEN C030 C030	GRADUATE

$C001 = \{\text{ABOUT, ACROSS, AFTER, AGAINST, ..}\}$
$C009 = \{\text{ALTHOUGH, BECAUSE, BUT, HOW, IF, ...}\}$
$C030 = \{\text{NINETEEN, TEN, SIXTY, TWO, ZERO, ...}\}$
$C031 = \{\text{AIN'T, ARE, HAD, HADN'T, HAS, HASN'T, ...}\}$
$C108 = \{\text{ACCOMPANY, ADOPT, ADVISE, AFFECT, ...}\}$
$C247 = \{\text{AERIAL, ARCHITECTURAL, ATHLETIC, ...}\}$
$C377 = \{\text{FATHERS, FIRMS, HOUSES, MAKERS, ...}\}$

[3] A similar approach has been presented in [9]. In that work, however, the word categories are context dependent.

3.1. Calculating Probabilities

As already mentioned, category/word varigrams have the special characteristic that a given word history may have several different descriptions of equal length. Concerning the calculation of word probabilities, there are two issues arising from this fact. The first one involves the combination of the respective differing predictions. The second issue concerns the technique of backing-off, which is a little more involved, if for a given history description there may be several different backoff descriptions. We will come back to that. The rest of this section provides the background for two variants of category/word varigrams, which differ primarily in the way in which they calculate a word probability given several history descriptions of the same length.

Variant 1

Consider a history $h_k = (w_1, \ldots, w_k)$. In the framework of category/word varigrams, each of the $w_i, 1 \leq i \leq k$ may be described by either w_i itself or by its category $c(w_i)$. Consequently, there are 2^k possible descriptions for h_k. Let us denote the set of history description contained in $\mathcal{L}$ by $B_{\mathcal{L}}(h_k) \subseteq \mathcal{L}$.

The prediction quality of a history descriptions $b \in B_{\mathcal{L}}(h_k)$ can be quantified using the measure

$$\Delta_{LL}(b) = \sum_{i=1}^{N_{Corpus}} log \frac{p_{LOO}(w_i|(w_1 \ldots w_{i-1}), \mathcal{L})}{p_{LOO}(w_i|(w_1 \ldots w_{i-1}), \mathcal{L} \backslash (b, w_i))}, \quad (10)$$

where $p_{LOO}()$ denotes a probability calculated on a leaving-one-out basis. Variant 1 of the category/word varigrams now selects among all the history descriptions for h_k in $\mathcal{L}$ the one with the highest average gain in corpus log likelihood:

$$b^*_{\mathcal{L}}(h_k) = \text{argmax}_{b \in B_{\mathcal{L}}(h_k)} \frac{\Delta_{LL}(b)}{N(b)}. \quad (11)$$

$N(b)$ is the training corpus event count for description b. To actually calculate the probability $p(w|h)$ within this varigram variant, the α- and γ-expressions in formula (1) have to be changed into:[4]

$$\alpha_*(w|h_k) = \frac{d \left(N(b^*,w) - \sum_{(b',w) \in \mathcal{L}} p^*(b'|b^*) d(N(b',w)) \right)}{N(b^*) - \sum_{(b',w) \in \mathcal{L}} p^*(b'|b^*)(N(b',w) - d)} \quad (12)$$

$$\gamma_*(h_k) = \frac{d_k N_+(b^*) + N(b^*) - \sum_{(b^*,w) \in \mathcal{L}} N(b^*,w)}{N(b^*) - \sum_{(b',w) \in \mathcal{L}} p^*(b'|b^*) d(N(b',w))} \quad (13)$$

The term $p^*(b'|b^*)$ is related to the backing-off issue mentioned above. In word-based or category-based n-gram models the relation of direct backoffs is a mapping, meaning that for a given history description b' there is a unique $b <_{\mathcal{L}} b'$. In the context of category/word varigrams this is not the case. As an example using the categories of the above example category list, $b' =$ (THE TWO C377) may backoff to (C030 FIRMS) or to (C030 C377) or to a number of other history descriptions, depending on what word of class C377 actually occurs in the word history, which histories are stored in the language model and what their respective values for $\Delta_{LL}(.)/N(.)$ are. The factor $p^*(b'|b^*)$ is therefore used to express the probability that there is a direct backoff from b' to b^* in the language model. Setting

$$N_{eff}(b,w) = d \left(N(b^*,w) - \sum_{(b',w) \in \mathcal{L}} p^*(b'|b^*) d(N(b',w)) \right), \quad (14)$$

the law of effective counts for history descriptions can be proved for category/word varigrams as in (9).

[4] We use b^* as an abbreviation for $b^*_{\mathcal{L}}(h_k)$.

Variant 2

The second variant calculates its α-term as the average of the predictions $\alpha_M(w|b), b \in B_{\mathcal{L}}(h_k)$. In contrast to variant 1, where during the training of the language model an effective count will only increase for a history description qualifying as a b^*, in the second variant the effective counts for all $b \in B_{\mathcal{L}}(h_k)$ are increased, if there is no history description longer than k in $\mathcal{L}$ (disregarding discounting effects). Accordingly, the law formulated in section 2.2 is not valid for variant 2 language models. Here, the α and γ have the value:

$$\alpha_\oslash(w|h_k) = \frac{\sum_{b \in B_{\mathcal{L}}(h_k)} \alpha_M(w|b)}{|B_{\mathcal{L}}(h_k)|} \quad (15)$$

$$\gamma_\oslash(w|h_k) = \frac{\sum_{b \in B_{\mathcal{L}}(h_k)} \gamma_M(w|b)}{|B_{\mathcal{L}}(h_k)|}, \quad (16)$$

where α_M and γ_M are as defined in (6) and (7), however this time applied to history descriptions b instead of histories h

3.2. LM Generation

For category/word varigrams, the search space of all possible history descriptions for a corpus of considerable size might get huge. Thus, pruning plays an important role in generating the language model. As a pruning criterion for history descriptions b, both variants apply the measure $\Delta_{LL}(b)$. There is a separate pruning criterion for the (b,w), which is related to $\Delta_{LL}(b)$ via:

$$\text{Variant1}: \quad \Delta_{LL}(b,w) = \alpha_*(w|b)\Delta_{LL}(b) \quad (17)$$

$$\text{Variant2}: \quad \Delta_{LL}(b,w) = \alpha_M(w|b)\Delta_{LL}(b) \quad (18)$$

In our experiments, these $\Delta_{LL}(b,w)$ had slight performance advantages in comparison to a direct leaving-one-out Δ_{LL} for the (b,w) or to the pruning criterion in [3]. This is probably because the relative scaling of the $\Delta_{LL}(b)$ and the $\Delta_{LL}(b,w)$ is particularly easy with the measures presented in (17) and (18).

4. EXPERIMENTS

The experiments compare various language models in terms of perplexity as well as recognition performance. For training, the 39 million word Wallstreet Journal corpus was used. Perplexities and recognition performance were evaluated using the male portion of the DARPA NAB'94 development and evaluation sets (about 4300 spoken words). The vocabulary contains 64k words.

The following language models have been investigated:

1. a word based varigram, as described in [3]
2. a word based varigram, linearly interpolated with a category-based varigram. The WERs were calculated by optimizing both the interpolation weights and the relative size of both models on an independent 325k word WSJ corpus.
3. a category/word varigram (variant 1)
4. model 3, linearly interpolated with a category varigram
5. a category/word varigram (variant 2)
6. model 5, linearly interpolated with a category varigram

All varigrams are based on fourgram models. The category model as well as the category/word varigram both apply a categorization of the vocabulary into 750 classes. The categorization was optimized as described in [1]. The category-based components of models 2, 4 and 6 were pruned using the techniques of [3].

Figure 1 shows the language model perplexities depending on the size of the language models, measured as the number of stored word sequences.

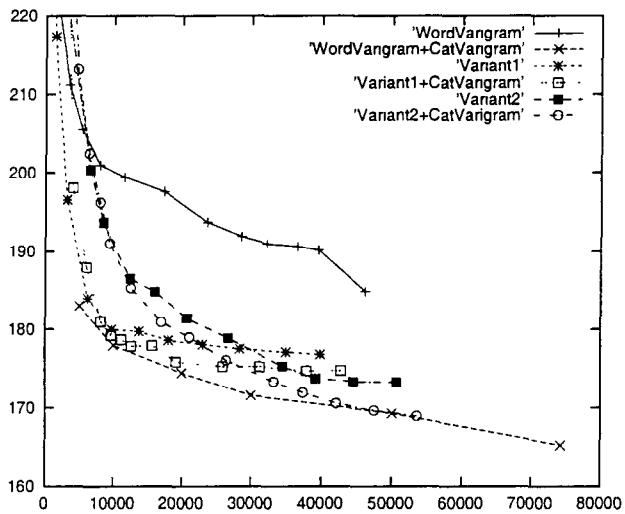

Figure 1: Perplexity (LM size in thousand word sequences)

Although the category/word varigram obviously yields a perplexity improvement as compared to the word varigram, the best overall perplexity values are provided by the linear interpolation of word-based and category-based varigrams. This seems to be mainly a result of the fact that during the generation of category/word varigrams (i.e. the search in the space of possible history descriptions), the leaving-one-out pruning criterion discards many of the (h, w) with $N(h, w) = 1$, whereas these singleton events remain in the word-based and category-based varigrams. This is also the reason why the unpruned model 2 extends to larger sizes than the category/word varigrams.

Figure 2 presents the results of word error rate experiments on the above mentioned test data. For each of the compared language models, decoding has been performed using an N-best rescoring technique with $N = 50$. Variant 1 without interpolated category model performs about as good as the linear interpolation of a word-based and a category-based varigram, whereas especially variant 2 yields a considerable reduction of WER.

5. CONCLUSIONS AND FUTURE WORK

As a well known fact, combining word-based and category-based language models can improve recognition rates. For the WSJ0

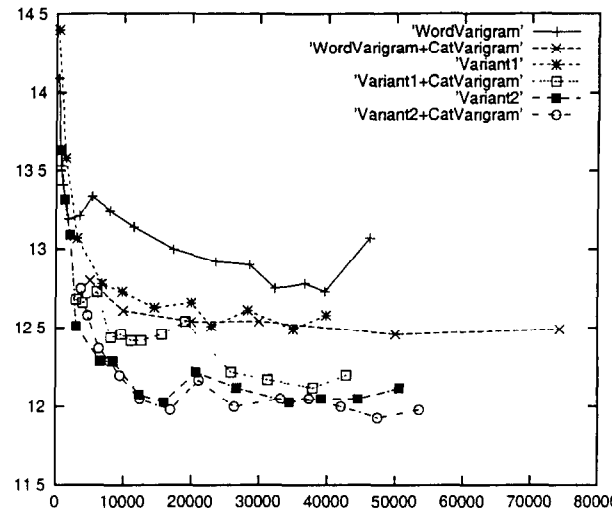

Figure 2: Word error rate (LM size in thousand word sequences)

setting, the linear interpolation of a category-based varigram to a word-based varigram yields a WER improvement of about 2–3%. The category/word-based varigrams presented in this paper allow a more efficient use of the category information, increasing the WER improvement to about 7%. Future work will concentrate on the optimization of the search process involved in language model training, and on alternative methods of combining the predictions of different history descriptions.

6. REFERENCES

[1] R. Kneser and H. Ney, "Improved Clustering Techniques for Class-Based Statistical Language Modelling". In Proc. EUROSPEECH, pages 973–976, Berlin, Germany, Sep. 1993

[2] R. Kneser and H. Ney, "Improved Backing-off for M-gram Language Modeling". In Proc. ICASSP, volume 1, pages 181–184, Detroit, MI, May 1995.

[3] R. Kneser, "Statistical Language Modeling using a Variable Context Length". In Proc. ICSLP, volume 1, pages 494–497, Philadelphia, PA, Oct. 1996.

[4] S. Martin, J. Liermann and H. Ney, "Algorithms for Bigram and Trigram Word Clustering". In Proc. EUROSPEECH, pages 1253–1256, Madrid, September 1995.

[5] H. Ney, U. Essen, and R. Kneser, "On Structuring Probabilistic Dependencies in Stochastic Language Modelling". Computer Speech and Language, 8:1–38, 1994.

[6] H. Ney, U. Essen, and R. Kneser, "On the Estimation of 'small' Probabilities by Leaving-One-Out". Pattern Analysis and Machine Intelligence, 17(12):1202–1212, 1995.

[7] T.R. Niesler and P.C. Woodland, "A variable-length category-based n-gram language model". In Proc. ICASSP, volume 1, pages 164–167, Atlanta, May 1996.

[8] T.R. Niesler and P.C. Woodland, "Combination of word-based and category-based language models". In Proc. ICSLP, volume 1, pages 220-223, Philadelphia, PA, Oct. 1996.

[9] M. Sui and M. Ostendorf, "Variable n-gram language modeling and extensions for conversational speech". In Proc. EUROSPEECH, pages 2739-2742, Rhodes, September 1997.