# IMPROVED BACKING-OFF FOR M-GRAM LANGUAGE MODELING

*Reinhard Kneser*

Philips GmbH
Research Laboratories
D-52066 Aachen, Germany
kneser@pfa.philips.de

*Hermann Ney*

Lehrstuhl für Informatik VI
RWTH Aachen, University of Technology
D-52056 Aachen, Germany
ney@informatik.rwth-aachen.de

## ABSTRACT

In stochastic language modeling, backing-off is a widely used method to cope with the sparse data problem. In case of unseen events this method backs off to a less specific distribution. In this paper we propose to use distributions which are especially optimized for the task of backing-off. Two different theoretical derivations lead to distributions which are quite different from the probability distributions that are usually used for backing-off. Experiments show an improvement of about 10% in terms of perplexity and 5% in terms of word error rate.

## 1. INTRODUCTION

The task of a stochastic language model is to provide the probability of a given word sequence. Typically this is achieved by supplying conditional probabilities $p(w|h)$, where $h$ is an equivalence class of the history of word $w$, i.e. of the words preceding $w$. In the case of $M$-gram models, for example, two histories are considered to be equivalent, if they agree in the last $(M-1)$ words. Even with this simplification of equivalence classes we are faced with the problem of sparse data. The number of possible events is huge and much larger than the amount of available training data. We therefore have to estimate probabilities for events which were never observed. Many different smoothing techniques such as interpolation and backing-off strategies are in use to overcome this problem [4][5][6]. Common to most of these approaches is the use of less specific equivalence classes of the histories. Probabilities conditioned on these coarser classes can be more reliably estimated and are then used to back off the model in the case of unseen events. In the case of $M$-gram models, for example, a $(M-1)$-gram distribution can be used for backing-off.

Usually the normal probability distribution of the coarser model is taken for backing-off. In that case the information that the event is not covered by the detailed model is lost. All events are taken into consideration for the estimation of the backing-off distribution, also those already covered by the $M$-gram models. This results in a bias towards words heavily conditioned on the immediate predecessor words. Consider for example the word ***dollars*** which is a very frequent word in the Wall-Street-Journal corpus but occurs almost exclusively after numbers and some country names. The latter fact makes it very unlikely that *dollars* will occur after some word $x$ if the bigram ($x$, *dollars*) has not been observed. On the other hand, the smoothed probability estimate will be relatively high, if the unigram probability $p$(*dollars*) is taken for backing-off. This suggests to use some backing-off distribution different from the normal probability distribution.

## 2. BACKING-OFF

For the rest of the paper we assume that histories which are equivalent according to the specific equivalence relation are also equivalent according to the more general relation. For a detailed class $h$ then there is just one less specific class for which we write $\hat{h}$. From a unifying point of view most smoothing techniques can now be formulated as backing-off models which have the following form:

$$p(w|h) = \begin{cases} \alpha(w|h) & \text{if } N(h,w) > 0 \\ \gamma(h)\,\beta(w|\hat{h}) & \text{if } N(h,w) = 0 \end{cases} . \qquad (1)$$

For those events which have been observed in the training data (i.e. the occurrence count $N$ is larger than 0) we assume some reliable estimate $\alpha$ of the probability. For the remaining unseen events the estimation is done according to some less specific distribution $\beta$. The normalization term $\gamma$ is required in order to guarantee that

This work was partially funded by the German Federal Ministry for Research and Technology (BMFT) in the framework of the Verbmobil Project under Grant 01 IV 102 B. The responsibility for the contents of this study lies with the authors.

0-7803-2431-5/95 $4.00 © 1995 IEEE

$p(w|h)$ sums to unity, and is determined completely by $\alpha$ and $\beta$:

$$\gamma(h) = \frac{1 - \sum\limits_{w:N(h,w)>0} \alpha(w|h)}{\sum\limits_{w:N(h,w)=0} \beta(w|\hat{h})}. \qquad (2)$$

The various smoothing techniques differ largely in the probability estimate $\alpha$ for seen events. Among those are the Turing-Good estimates [3][5] and the linear and absolute discounting methods [6]. On the other hand the smoothing distribution is usually kept fixed to be $\beta(w|\hat{h}) = p(w|\hat{h})$. The novel idea of our approach is to leave also the parameters of this distribution $\beta$ free and to optimize them together with the other parameters.

In the following we pursue two different approaches which lead to similar solutions. Both solutions are independent of the special kind of modeling and add virtually no additional computational overhead to the model computation.

## 3. MARGINAL DISTRIBUTION AS CONSTRAINT

In the first approach we assume that we know $p(w|h)$ for $M$-grams with $N(h,w) > 0$. In other words $\alpha(w|h)$ in Eq. (1) is given and fixed. Further we assume that the less specific distributions $p(w|\hat{h})$ and $p(h|\hat{h})$ can be reliably estimated and thus are also fixed. The basic idea now is to determine $\beta(w|\hat{h})$ such that the marginal distribution of the resulting joint distribution $p(h, w|\hat{h})$ is identical to the given distribution $p(w|\hat{h})$:

$$p(w|\hat{h}) = \sum_g p(g, w|\hat{h}). \qquad (3)$$

We may write this as

$$\begin{aligned} \sum_g p(g, w|\hat{h}) &= \sum_g p(w|g, \hat{h}) p(g|\hat{h}) \\ &= \sum_{g:\hat{g}=\hat{h}} p(w|g) p(g|\hat{h}) \qquad (4) \end{aligned}$$

since we have $p(g|\hat{h}) = 0$ for histories $g$ with $\hat{g} \neq \hat{h}$. Using the special form of our model we obtain

$$\begin{aligned} \sum_g p(g, w|\hat{h}) &= \sum_{g:\hat{g}=\hat{h},N(g,w)>0} \alpha(w|g) p(g|\hat{h}) + \\ & \sum_{g:\hat{g}=\hat{h},N(g,w)=0} \gamma(g) \beta(w|\hat{h}) p(g|\hat{h}). \qquad (5) \end{aligned}$$

We move $\beta$ out of the second sum and apply the constraint Eq. (3) and get

$$\beta(w|\hat{h}) = \frac{p(w|\hat{h}) - \sum\limits_{g:\hat{g}=\hat{h},N(g,w)>0} \alpha(w|g) p(g|\hat{h})}{\sum\limits_{g:\hat{g}=\hat{h},N(g,w)=0} \gamma(g) p(g|\hat{h})}. \qquad (6)$$

In a first approximation, the sum in the denominator can be considered constant with respect to $w$. $\beta(w|\hat{h})$ is thus proportional to the numerator. A solution where $\beta(w|\hat{h})$ sums up to unity is hence obtained by normalization:

$$\beta(w|\hat{h}) = \frac{p(w|\hat{h}) - \sum\limits_{g:\hat{g}=\hat{h},N(g,w)>0} \alpha(w|g) p(g|\hat{h})}{\sum\limits_{v} [p(v|\hat{h}) - \sum\limits_{g:\hat{g}=\hat{h},N(g,v)>0} \alpha(v|g) p(g|\hat{h})]}. \qquad (7)$$

The solution Eq. (7) becomes particularly simple in the case of absolute discounting [6], i.e. when

$$\alpha(w|h) = \frac{N(h,w) - d}{N(h)} \quad \text{with} \quad 0 < d < 1 \qquad (8)$$

We assume the maximum-likelihood estimates for the marginal distributions

$$p(w|\hat{h}) = \frac{N(\hat{h}, w)}{N(\hat{h})} \quad \text{and} \quad p(g|\hat{h}) = \frac{N(\hat{h}, g)}{N(\hat{h})}, \qquad (9)$$

where the combined count $N(\hat{h}, g)$ is equal to $N(g)$ if $\hat{g} = \hat{h}$ and 0 otherwise. We then obtain

$$\beta(w|\hat{h}) = \frac{N(\hat{h}, w) - \sum\limits_{g:\hat{g}=\hat{h},N(g,w)>0} [N(g,w) - d]}{\sum\limits_{v} [N(\hat{h}, v) - \sum\limits_{g:\hat{g}=\hat{h},N(g,v)>0} [N(g,v) - d]]}. \qquad (10)$$

With the definitions

$$N_+(\cdot, \hat{h}, w) := \sum_{g:\hat{g}=\hat{h},N(g,w)>0} 1 \qquad (11)$$

and

$$N_+(\cdot, \hat{h}, \cdot) := \sum_w N_+(\cdot, \hat{h}, w) \qquad (12)$$

Eq. (10) gives

$$\beta(w|\hat{h}) = \frac{N_+(\cdot, \hat{h}, w)}{N_+(\cdot, \hat{h}, \cdot)}. \qquad (13)$$

We thus obtain a distribution which is quite different from the probability distribution $p(w|\hat{h})$. Only the information that a word has been observed in some coarse context is taken into account and the frequency of this event is ignored.

## 4. LEAVING-ONE-OUT

It is well known that maximum-likelihood estimation can not be used to estimate the parameters $\alpha$, $\beta$ and $\gamma$ directly since it leads to zero probabilities for unseen events. In order to overcome this shortcoming cross-validation techniques, such as the leaving-one-out technique [2, pp.75] can be successfully applied.

The basic idea of cross validation is to get a measure of the generalization capability of a model by testing it on data not seen during training. The leaving-one-out technique achieves this very efficiently in the following way: One single event is removed from the training data and a model is trained on the remaining data. This model is then used to estimate a leaving-one-out probability of the removed event. The sum of the logarithms of all those probabilities gives the leaving-one-out log-likelihood which serves as optimization criterion.

When applying the leaving-one-out technique to the standard backing-off model, events that were observed just once get removed from the training data and thus fall into the 'unseen' branch of Eq. (1). Events that have been observed twice or more, on the other hand, still stay in the 'seen' branch. Being only interested in terms containing $\beta$, we get the following leaving-one-out log-likelihood function:

$$F = \sum_{(g,v):N(g,v)=1} \ln[\gamma(g)\beta(v|\hat{g})] + const(\{\beta(v|\hat{g})\}). \tag{14}$$

Replacing $\gamma$ according to Eq. (2) gives

$$F = \sum_{(g,v):N(g,v)=1} \ln \frac{\beta(v|\hat{g})}{\sum\limits_{u:N(g,u)=0} \beta(u|\hat{g})} + const(\{\beta(v|\hat{g})\}). \tag{15}$$

When taking the partial derivatives of $F$ with respect to $\beta(w|\hat{h})$, we have to take into account that there are only contributions from terms containing $\beta(w|\hat{h})$:

$$\frac{\partial F}{\partial \beta(w|\hat{h})} = \sum_{g:\hat{g}=\hat{h},N(g,w)=1} \frac{1}{\beta(w|\hat{h})} - \sum_{\substack{(g,v):\hat{g}=\hat{h},N(g,v)=1 \\ N(g,w)=0}} \frac{1}{\sum\limits_{u:N(g,u)=0} \beta(u|\hat{g})} \tag{16}$$

In a first approximation, the second sum can be considered constant with respect to $w$ and by setting the derivative to zero we obtain

$$\beta(w|\hat{h}) = const(w) N_1(\cdot, \hat{h}, w), \tag{17}$$

where we use the definition

$$N_1(\cdot, \hat{h}, w) := \sum_{g:\hat{g}=\hat{h},N(g,w)=1} 1. \tag{18}$$

We may choose the proportional factor in such a way that $\beta(w|\hat{h})$ sums up to unity. With the definition of

$$N_1(\cdot, \hat{h}, \cdot) := \sum_{w} N_1(\cdot, \hat{h}, w) \tag{19}$$

we thus get the final result

$$\beta(w|\hat{h}) = \frac{N_1(\cdot, \hat{h}, w)}{N_1(\cdot, \hat{h}, \cdot)}. \tag{20}$$

This solution can be interpreted as relative counts where only singletons, i.e. events observed just once, are taken into consideration. This seems reasonable since it is well known that events seen once give a good estimate for unseen events. Note also that the solutions of the two approaches (Eqs. 13, 20) are very similar. This is even more evident, when realizing that most of the summands in Eq. (11) are really singletons.

## 5. EXPERIMENTAL RESULTS

The performance of the different backing-off distributions was evaluated in several experiments on two different tasks. The German Verbmobil corpus comprises about 30,000 words and consists of transliterations of a few hundred dialogues. Tests were carried out with a closed 2,000 word vocabulary. The much larger Wall-Street-Journal corpus (WSJ) comprises about 40 million words of preprocessed newspaper material as used in the ARPA evaluation tests [7]. An open 45k word vocabulary was used, which gave a coverage of 99.7%. For both tasks separate test sets were defined for evaluation. In the case of the Wall-Street-Journal task this test set consists of the official 1992 and 1993 evaluation material.

All experiments were carried out with trigram language models. Non-linear interpolation [6], a slight variant of absolute discounting (Eq. 8) was used for smoothing. Our standard language model, with $\beta(w|\hat{h}) = p(w|\hat{h})$ as backing-off distribution, served as baseline. Models using the 'singleton' distribution (Eq. 20) and the 'marginal constraint' distribution (Eq. 13) for backing-off were trained in addition. All backing-off distributions were themselves smoothed in order to avoid zero probabilities.

Test-set perplexities[4] were calculated for all models and tasks. In addition, recognition results were produced for the Wall-Street-Journal task by applying the different models in the trigram rescoring step of our recognizer [1]. The results are shown in Table 1.

When a huge amount of training material is available such as in the Wall-Street-Journal task, the memory space plays an important role. It is well known that the memory space needed for the storage of a language model can be drastically reduced without loss of performance by ignoring all trigrams which have been observed only once. In an additional experiment such compact trigram models were built for all three kinds of backing-off distributions. The results are also shown in Table 1.

Table 1: Perplexities and error rates for different trigram models

| Model | Perplexity | Error Rate |
|---|---|---|
| WSJ 45k, Full: | | |
| Standard | 161.0 | 11.9% |
| Singleton | 145.7 | 11.2% |
| Marginal Constraint | 144.3 | 11.1% |
| WSJ 45k, Compact: | | |
| Standard | 169.4 | 11.9% |
| Singleton | 152.6 | 11.4% |
| Marginal Constraint | 150.6 | 11.5% |
| Verbmobil: | | |
| Standard | 105.4 | - |
| Singleton | 97.9 | - |
| Marginal Constraint | 96.2 | - |

The experiments show a consistent improvement of the language models by using special backing-off distributions. For the new models we obtain a perplexity which is up to 10% lower than for the baseline model. This leads to a 5% lower word error rate in recognition. The perplexities for the 'marginal constraint' model are always slightly smaller than those of the 'singleton' models but the recognition results are more or less the same. In the 'compact' case almost all models perform a little bit worse, both in terms of perplexity and error rate. Only the standard model achieves the same recognition result as the full model.

A final experiment was performed in order to compare our best model with the official trigram language model supplied by ARPA for the 1993 evaluation. For this purpose we trained an additional model using the same training data and the same 20k vocabulary as was used for the training of the ARPA model. The perplexities are shown in Table 2. We observe that also in comparison with the official model we were able to achieve an improvement of about 9% for the trigram model.

Table 2: WSJ 20k – Perplexities for different language models

| Model | Bigram | Trigram |
|---|---|---|
| ARPA 93 | 213.3 | 147.0 |
| Marginal Constraint | 207.5 | 134.0 |

## 6. CONCLUSIONS

In this paper we have improved the standard backing-off language model. This was achieved by using backing-off distributions which were especially optimized for the backing-off model. Two related solutions were derived from different theoretical approaches. Both solutions do not depend on the specific model and do not add extra computational costs. In experiments with trigram language models we obtained an improvement of up to 10% in terms of perplexity and of 5% in terms of word error rate.

## 7. REFERENCES

[1] X. Aubert, C. Dugast, H. Ney, V. Steinbiss: "Large Vocabulary Continuous Speech Recognition of Wall Street Journal Data", *Proc. ICASSP*, Adelaide, Australia, April 1994.

[2] R. O. Duda, P. E. Hart: *Pattern Classification and Scene Analysis*, Wiley, New York, 1973.

[3] I.J. Good: "The population frequencies of species and the estimation of population parameters", *Biometrika 40*, pp. 237-264, Dec. 1953.

[4] F. Jelinek: "Self-organized language modeling for speech recognition", pp. 450-506, in A. Waibel, K.-F. Lee (eds.): *Readings in Speech Recognition*, Morgan Kaufman Publishers, 1991.

[5] S.M. Katz: "Estimation of probabilities from sparse data for the language model component of a speech recognizer", *IEEE Trans. on Acoustics, Speech and Signal Proc.*, Vol. ASSP-35, pp. 400-401, March 1987.

[6] H. Ney, U. Essen, R. Kneser: "On Structuring Probabilistic Dependences in Stochastic Language Modelling", *Computer Speech and Language*, Vol. 8, pp. 1-38, 1994.

[7] D. B. Paul, J. M. Baker: "The Design for the Wall Street Journal-based CSR Corpus", *Proceedings DARPA Speech and Natural Language Workshop*, Harriman, New York, pp. 357-362, Feb. 1992.