COOCCURRENCE SMOOTHING FOR STOCHASTIC LANGUAGE MODELING

Ute Essen, Volker Steinbiss

Philips GmbH Forschungslaboratorien, Aachen
P. O. Box 1980, D-5100 Aachen, Germany
email: essen@pfa.philips.de *and* steinbiss@pfa.philips.de

ABSTRACT

Training corpora for stochastic language models are virtually always too small for maximum-likelihood estimation, so smoothing the models is of great importance. This paper derives the cooccurrence smoothing technique for stochastic language modeling and gives experimental evidence for its validity. Using word-bigram language models, cooccurrence smoothing improved the test-set perplexity by 14% on a German 100,000-word text corpus and by 10% on an English 1-million word corpus.

1. INTRODUCTION

Good stochastic language models are important for achieving high performance in large-vocabulary speech recognition. Although large text corpora are used to train these language models, the number of training observations is typically small as compared to the number of free model parameters. Many events are not observed in training and thus would be assigned zero probabilities by maximum-likelihood estimation. As on principle no word sequence should be excluded from recognition, zero probabilities must be avoided. This is achieved by smoothing the language model parameters.

We propose a novel smoothing technique for language modeling which is motivated by the *cooccurrence smoothing* method used for acoustic modeling [2], which was introduced by Sugawara [1]. A central point is the estimation of confusion probabilities of word pairs. The resulting confusion matrix - the *cooccurrence matrix* - is used for smoothing the conditional word probabilities of the language model.

We start with a general derivation of the cooccurrence smoothing technique for stochastic language modeling. Then, explicit formulas for the bigram model case are given. Experiments on a 100,000-word German and a 1-million-word English corpus show the validity of our approach.

2. THEORY

2.1 Estimation of the Cooccurrence Matrix

In estimating the parameters of a stochastic language model on a training corpus, smoothing becomes an essential technique as training corpora are virtually always too small for maximum-likelihood estimation (if they are not, better refine the model). One possibility to obtain a more reliable estimate of the conditional probabilities of a word given some context is to take advantage of observations of other words that behave 'similarly' to this word.

Our measure of similarity is the probability of word w' being substituted by word w, so we estimate how likely w is to be observed in the same contexts in which w' has been seen. Here, a *context* k is an equivalence class [w] of the complementary word sequence

$$w = \ldots, w_{n-2}, w_{n-1}, w_{n+1}, w_{n+2}, \ldots$$

preceding and following the word w_n under consideration. Two simple examples are the contexts defined by

- the predecessor word: $[w] = w_{n-1}$, or (A)
- the successor word: $[w] = w_{n+1}$. (B)

Let W_n be a random variable denoting *word at position n* and $W = \ldots, W_{n-1}, W_{n+1}, W_{n+2}, \ldots$ the random variable for the *complementary word sequence*. Let us assume for the moment that not one but two experiments are performed, one being marked with a prime. Thus, $W_n{=}w_n$ *and* $W_n'{=}w_n'$ means that word w_n' was observed in experiment 1 and word w_n in experiment 2. Assuming the context to be unknown but fixed, the probability

$$P\,(W_n{=}w_n,\ W_n'{=}w_n')$$

of this pair of observations is modelled by

$$= P\,(W_n{=}w_n,\ W_n'{=}w_n' \mid [W]{=}[W'])$$

$$= \sum_{\text{contexts } k} P(W_n{=}w_n, W_n'{=}w_n' \mid [W]{=}[W']{=}k)\, P([W]{=}k)$$

$$= \sum_{k} P(W_n{=}w_n \mid [W]{=}k)\, P(W_n'{=}w_n' \mid [W']{=}k)\, P([W]{=}k)$$

The confusion probabilities, forming the *cooccurrence matrix*, then are

$$P_C\,(W_n{=}w_n \mid W_n'{=}w_n')$$

$$= \sum_{k} \frac{P(W_n{=}w_n \mid [W]{=}k)\, P(W_n'{=}w_n' \mid [W']{=}k)\, P([W]{=}k)}{P(W_n'{=}w_n')}$$

We feel that it is not obvious how to choose the type of context and how to estimate the conditional probabilities $P(W_n{=}w_n|[W]{=}k)$ in order to *optimally* estimate the cooccurrence matrix. A straightforward approach is to specify the context via the m-1 words directly preceding w_n , i. e. $[w] = (w_{n-m+1}, \ldots, w_{n-1})$. This amounts to using a *stochastic m-gram language model* for calculating the cooccurrence matrix.

2.2 Smoothing of the Language Model

We start from a basic stochastic m-gram language model P_B . In order to obtain a more robust estimate for the conditional probability of word w_n following the word sequence $w_{n-m+1}, \ldots, w_{n-1}$,

$$P_B\,(w_n \mid w_{n-m+1}, \ldots, w_{n-1})\,,$$

we take account of conditional probabilities of words w_n' that behave similarly to w_n . Using the confusion probabilities

$$P_C\,(W_n{=}w_n \mid W_n'{=}w_n') =: P_C\,(w_n \mid w_n')$$

derived above, the cooccurrence-smoothed probabilities P_S are defined as

$$P_S\,(w_n \mid w_{n-m+1}, \ldots, w_{n-1}) =$$

$$\sum_{w_n'} P_C\,(w_n \mid w_n')\, P_B\,(w_n' \mid w_{n-m+1}, \ldots, w_{n-1})$$

Note that, as the m-gram model can be used to estimate the cooccurrence probabilities, in the end the model is used for smoothing itself!

2.3 Special Case: Bigram Language Model

For the bigram language model case, the smoothing formula of section 2.2 turns into

$$P_S(w_n \mid w_{n-1}) = \sum_{w_n'} P_C(w_n \mid w_n')\, P_B(w_n' \mid w_{n-1}) \quad (1)$$

Instead of smoothing over the observations w_n, an interesting variant in the bigram case is to smooth over the conditioning events, namely the predecessor words w_{n-1}:

$$P_S(w_n \mid w_{n-1}) = \sum_{w'_{n-1}} P_B(w_n \mid w'_{n-1})\, P_C(w'_{n-1} \mid w_{n-1}) \qquad (2)$$

or over both:

$$P_S(w_n \mid w_{n-1}) = \sum_{w'_n,\, w'_{n-1}} P_C(w_n \mid w'_n)\, P_B(w'_n \mid w'_{n-1})\, P_C(w'_{n-1} \mid w_{n-1}) \qquad (3)$$

3. EXPERIMENTAL RESULTS

3.1 Approach

In our experiments we combined the smoothing formulas (1) and (2) with two cooccurrence matrices derived from the contexts (A) and (B) of section 2.1 (a standard and a 'reversed' word bigram context). As 1-A and 2-B can be shown to be identical, three language models had to be evaluated.

We used non-smoothed maximum-likelihood estimates for the bigram and unigram probabilities (i. e. relative frequencies) in order to separate the effects of different smoothing methods. In the experiments we combined the cooccurrence-smoothed bigram with the bigram, unigram and zerogram probabilities using linear interpolation.

3.2 Corpora

Experiments were run on two text corpora, which both were separated into a training (3/4) and an evaluation part (1/4).

Corpus I is a German text corpus of newspaper articles comprising 100,000 words. *Corpus II* is an English corpus (the LOB corpus) comprising 1 million words from a more heterogeneous text collection.

3.3 Results

We compared three variants of cooccurrence smoothing with each other and with a standard model. The standard model is the linear interpolation of a bigram, a unigram and a zerogram (or floor) model; the three variants to be compared were obtained by additionally interpolating with a cooccurrence-smoothed bigram component. All interpolation parameters were exclusively estimated on the training section of the corpus using the leaving-one-out method [3].

The main result is summarized in *Table 1.* Method 1-A is smoothing over the words w_n (formula (1)) with a cooccurrence matrix which has been estimated based on the predecessor words (A), i. e. using a standard bigram model. In comparison with our standard model, method 1-A resulted in a 10.3% reduction of test-set perplexity on corpus II. The improvement is larger on the smaller corpus I (14.4%) which is due to the fact that the word bigram probabilities were less reliably estimated: The fraction of word bigrams observed in the test partition that were not in the training partition is 50% for corpus I and 12% for corpus II.

Table 1. *Evaluation of test-set perplexities for three variants of cooccurrence smoothing on two text corpora, which are described in section 3.2.*

Method	Standard	1-A	1-B	2-A
Corpus I	696	**596**	696	674
Corpus II	562	**504**	541	538

With the two other methods (1-B and 2-A) we only achieved minor improvements (0% - 5%). This might be explained by the fact that only in case 1-A (or 2-B, resp.) exactly the relation between w_n and w_{n-1} is modelled. In

Table 2. *Interpolation parameters as estimated with the leaving-one-out method from the respective training portions (method 1-A).*

Corpus / Language Model		Zero-gram	Uni-gram	Cooc.-smoothed Bigram	Bi-gram
I	Standard	.16	.38	-	.46
	1-A	.15	.01	.71	.12
II	Standard	.06	.31	-	.63
	1-A	.04	.001	.49	.47

consequence of these results we did not try method (3), which in addition leads to time / memory problems.

Going back to method 1-A, *Table 2* shows the relative contributions of the partial language models. The cooccurrence-smoothed bigram almost replaces the unigram part, but the zerogram part - the constant value $\frac{1}{V}$ (V = vocabulary size) - remains important: both $P_C(w_n | .)$ and $P_B(w_n | .)$ are zero for words w_n not observed in the training section. *Table 3* indicates that the gain of the interpolated cooccurrence-smoothed model 1-A lies in the bigrams that have not been observed in the training data while both words have been observed in other contexts.

4. CONCLUSIONS AND FUTURE WORK

Cooccurrence smoothing has so far successfully been applied in acoustic modeling for the smoothing of discrete probabilities of hidden Markov models. We have derived the cooccurrence smoothing technique for stochastic language modeling and have shown its validity by experiments on two corpora: The 1-million word English LOB corpus and a 100,000-word German newspaper text corpus. Test-set perplexities were improved by 10.3% and 14.4%. We plan to further evaluate the new method by speech recognition experiments.

Table 3. *The differences in test-set perplexities between the standard method and method 1-A, evaluated on the following (overlapping) subsets: Bigrams where*

a) the predecessor w_{n-1}

b) the word w_n itself

c) the bigram (w_{n-1}, w_n)

were not observed in training. (Negative figures indicate improvements.)

Subset	(a)	(b)	(c)
Corpus I	+0.2%	+7.5%	-23.0%
Corpus II	+1.3%	+30.7%	-26.7%

REFERENCES

[1] K. Sugawara, M. Nishimura, K. Toshioka, M. Okochi, T. Kaneko: *Isolated Word Recognition Using Hidden Markov Models*, Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, Tampa, FL, pp. 1-4, March 1985.

[2] R. Schwartz, O. Kimball, F. Kubala, M.-W. Feng, Y.-L. Chow, C. Barry, J. Makhoul: *Robust Smoothing Methods for Discrete Hidden Markov Models*, Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, Glasgow, pp. 548-551, April 1989.

[3] H. Ney, U. Essen: *On Smoothing Techniques for Bigram-Based Natural Language Modelling*, Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, Toronto, pp. 825-828, May 1991.