Enhanced Good-Turing and Cat-Cal: Two New Methods for Estimating Probabilities of English Bigrams

(abbreviated version)

Kenneth W. Church
William A. Gale

AT&T Bell Laboratories

Abstract

For many pattern recognition applications including speech recognition and optical character recognition, prior models of language are used to disambiguate otherwise equally probable outputs. It is common practice to use tables of probabilities of single words, pairs of words, and triples of words (*n-grams*) as a prior model. Our research is directed to 'backing-off' methods, that is, methods that build an (n+1)-gram model from an n-gram model.

In principle, n-gram probabilities can be estimated from a large sample of text, by counting the number of occurrences of each n-gram of interest and dividing by the size of the training sample. Unfortunately, this simple method, known as the "maximum likelihood estimator" (MLE), is unsuitable because n-grams which do not occur in the training text are assigned zero probability. In addition, the MLE does not distinguish among bigrams with the same frequency.

We study two alternative methods for estimating the frequency of a given bigram in a test corpus, given a training corpus. The first method is an enhanced version of the method due to Good and Turing (Good, 1953). Under the modest assumption that the distribution of each bigram is binomial, Good provided a theoretical result that increases estimation accuracy. The second method assumes even less, merely that training and test corpora are generated by the same process. We refer to this purely empirical method as the Categorize-Calibrate (or Cat-Cal) method.

We emphasize three points about these methods. First, by using a second predictor of the probability in addition to the observed frequency, it is possible to estimate different probabilities for bigrams with the same frequency. We refer to this use of a second predictor as "enhancement." With enhancement, we find 1200 significantly different probabilities (with a range of five orders of magnitude) for the group of bigrams not observed in the training text; the MLE method would not be able to distinguish any one of these bigrams from any other. Second, both methods provide (estimated) variances for the errors in estimating the n-gram probabilities. Third, the variances are used in a refined testing method that enables us to study small differences between methods. We find that Cat-Cal should be used when counts are very small, and otherwise, GT is the method of choice.

1. Materials

Our corpus was selected from articles distributed by the Associated Press (AP) during 1988. Some portions of the year were lost. The remainder was processed automatically by Riley and Liberman to remove nearly identical articles. There remained $N = 4.4 \times 10^7$ words in the corpus, with a vocabulary of $V = 400{,}653$. When we speak of "words," we use a common term to hide a number of processing decisions. Roughly, a word is a string of characters delimited by white space. For instance, *The* and *the* are different words, and so are *need* and *needs*. In addition, punctuation such as period and comma are treated as "words". Additional tokens are inserted automatically to delimit sentences, paragraphs and discourses. In the future we hope to use a more balanced sample of general English. However, for the purpose of testing methods, a large sample is desirable; the the AP corpus is considerably larger than alternatives such as the Brown Corpus. The vocabulary size is also considerably larger than the 5000 word vocabulary reported in (Nàdas, 1984).

We split the 1988 AP wire into two halves by assigning bigrams beginning with even numbered words to one sample, those beginning with odd numbered words to the other. It is important that we made this split by taking every other bigram. We have found that spliting the corpus into two half-year periods, for example, generates two quite different samples, which complicates matters considerably. Since our aim is to study methods, we have adopted this extreme measure in order to construct two very similar samples.

Our goal is to develop a methodology for extending an n-gram model to an (n+1)-gram model. We regard the model for unigrams as completely fixed before beginning to study bigrams. This includes specifying V, the vocabulary, and $e(p(x))$, an estimate of the probability of each word. We also suppose that variances of the estimates are known. Likewise, we would regard a bigram model as fixed before studying a trigram model.

2. Estimation Methods

Let r^* be the adjusted frequency for a type observed r times. Then p, the probability of the type, is estimated by r^*/N. In order to satisfy the constraint $\sum p = 1$, the adjusted frequencies must satisfy $\sum r^* = N$. Two such methods will be considered at length: the Good-Turing Method (GT) and the Categorize-Calibrate Method (CC).

These methods are considerably better than the Maximum Likelihood Estimator (MLE): $r^* = r$. The main problem with MLE is that bigrams will be assigned zero probability if they didn't happen to occur in the training sample. Moreover, there are large errors when the counts are small (e.g., less than 20). In addition, the MLE fails to distinguish among bigrams with the same count. In our application there are billions of bigrams with a count of zero, some of which are much more likely than others. Their probability is neither zero nor identical.

2.1 The Basic Good-Turing and Cat-Cal Methods

We use the adjective *basic* to distinguish these methods from the *enhanced* methods that will be discussed in the next section. The main difference is that basic methods treat bigrams at atomic objects with no internal structure; enhanced methods will "back-off" and use the unigram model when appropriate.

The Good-Turing method has been used very successfully by IBM speech recognition group (Nàdas, 1984; Nàdas, 1985; Katz, 1987). The key insight suggested by Turing and developed by Good (1953), is the use of N_r, the number of bigrams which occur r times. We may refer to N_r as the frequency of frequency r. The GT estimate is $r^* = (r+1)N_{r+1}/N_r$ and it has a variance of $r^*(1+(r+1)^*-r^*)$. In practice it is necessary to use smoothed estimates of N_r instead of raw observations, especially when N_r is small. (Smoothing will not be discussed in this paper in order to save space.)

The following table illustrates a use of the basic GT estimate (BGT). (This example was selected so that the N_r's are large enough that smoothing is not too important.)

r (=MLE)	N_r	BGT r^*	BGT σ^2
0	1.605×10^{11}	1.28×10^{-5}	1.85×10^{-5}
1	2,053,146	0.446	0.808
2	458,136	1.26	2.49
3	191,809	2.24	4.50
4	107,522	3.25	6.31
5	69,883	4.19	8.47
6	48,809	5.21	10.4
7	36,345	6.21	12.9
8	28,201	7.28	
9	22,821		

The adjusted frequencies, r^*, can be compared to the raw frequencies, r; they have the same order, and do not differ greatly. The GT method assigns some probability to bigrams which have not been seen, suggesting that we should act as if we had seen each of them 0.0000128 times instead of zero times. In order to compensate for moving 160 billion bigrams from 0 to 0.0000128, some other bigrams must be adjusted downwards. In this case, all bigrams with $r > 0$ will be adjusted downwards.

Notice that the calculation of r^* for $r = 0$ depends on N_0, the number of bigrams that we have not seen. We can calculate N_0 because V is provided by the the unigram model. (This marks a great difference in our application of the Good-Turing formula from many applications in population biology, where inferences about the population size are the desideratum.) The total universe of bigrams that we wish to know about has size $V^2 \approx 1.6\times10^{11}$. N_0 is the difference between V^2 and the number of distinct bigrams seen, $\sum_{r>0} N_r$. Note that $N_0 \approx V^2$ since $V^2 > N_0 > V^2 - N$ and $N << V^2$. In other words, most bigrams have not been seen. In our experience, the problem only gets worse as we look at larger corpora because V^2 tends to grow faster than N.

GT improves on MLE by making use of more information, namely $\{N_r\}$. CC gathers even more information. The training text is divided into two halves. *Categorize* each bigram, b, by its observed frequency $r_1(b)$ in the first part of the text. Denote the number of distinct bigrams in the category by $N_r = \sum_{b|r_1(b)=r} 1$. *Calibrate* the category by counting all occurrences of all the bigrams in the category in the second part of the text, $C_r \equiv \sum_{b|r_1(b)=r} r_2(b)$, where the $r_2(b)$ is the observed frequency of the bigram, b, in the second half. The adjusted frequency is then: $r^* = C_r/N_r$. The only assumption behind this method is that both samples are generated by the same process. This assumption is weaker than the binomial assumption of GT. We refer to this method as the *basic* Cat-Cal method (BCC); the next section will consider an *enhanced* version that makes use of the bigrams' internal structure.

Basic Cat-Cal Method

r	N_r	C_r	BCC r^*	*repeat*	BGT r^*
0	1.605×10^{11}	2,046,125	1.27×10^{-5}	1.27×10^{-5}	1.28×10^{-5}
1	2,053,146	919,645	0.448	0.448	0.446
2	458,136	577,518	1.26	1.26	1.26
3	191,809	431,839	2.25	2.25	2.24
4	107,522	347,424	3.23	3.22	3.25
5	69,883	293,953	4.22	4.23	4.19
6	48,809	257,141	5.20	5.22	5.21
7	36,345	223,574	6.20	6.19	6.21
8	28,201	205,171	7.22	7.25	7.28

The adjusted frequencies for the BCC can be compared to the adjusted frequencies for the BGT as well as to the MLE. The differences between the BCC and the BGT are limited to the third significant

figure, while the differences of either from the MLE are in the first significant figure.

The fifth column of the table, labeled *repeat*, contains the results of repeating the basic Cat-Cal method after exchanging the texts used for categorization and calibration. The differences are again limited to the third significant figure, showing that BCC agrees well with our standard. We originally established the Cat-Cal method as a standard against which to compare other methods. However, we came to realize that it could itself be used as a practical method. Thus Cat-Cal plays two roles: standard and potential method.

The CC method can be extended to compute variances as illustrated below. Note that the variances computed by the CC method agree closely with those computed by GT.

		Variances		
r	N_r	${C_r}^2$	$BCC\sigma^2$	$GT\sigma^2$
0	1.605×10^{11}	2,980,905	1.85×10^{-5}	1.85×10^{-5}
1	2,053,146	2,069,343	0.808	0.808
2	458,136	1,865,654	2.48	2.49
3	191,809	1,831,325	4.48	4.50
4	107,522	1,805,150	6.36	6.31
5	69,883	1,827,811	8.41	8.47
6	48,809	1,858,543	10.5	10.4
7	36,345	1,832,738	12.4	12.9
8	28,201	1,898,443	14.5	

2.2 The Enhanced Methods

A key suggestion of this work is the introduction of a second predictor of frequency of observation in addition to an observed frequency; accounting for the second predictor constitutes what we call an *enhanced* method. We study an enhanced Good-Turing method and an enhanced Cat-Cal method. Both enhanced methods allow us to *differentiate* among the many bigrams which have not been seen. We will show that about 1200 significantly different probabilities can be estimated for bigrams not seen in the training text.

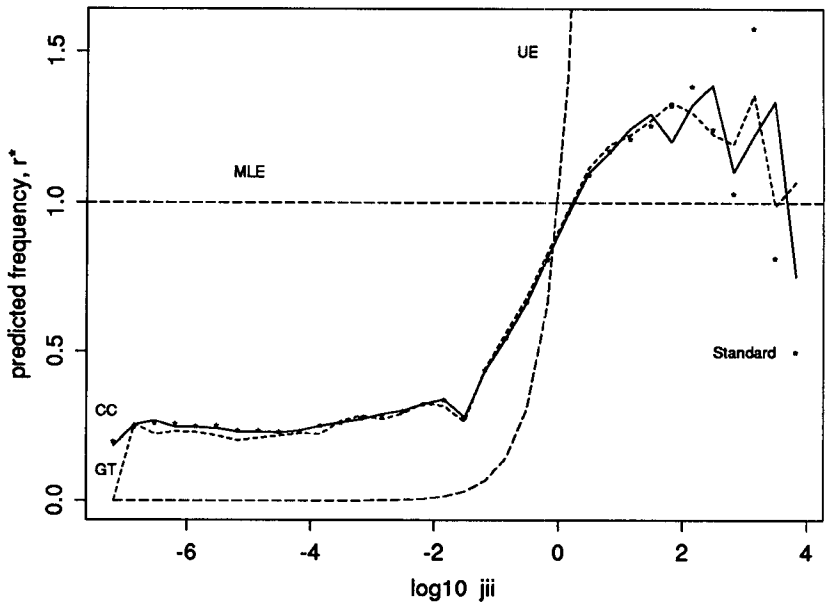
A possible second predictor for bigrams is the following: $jii = N\ e(p(x))\ e(p(y))$, where $e(p(x))$ and $e(p(y))$ are the unigram model's estimates of the probability of the first and second word in the bigram. *jii* is an acronym for "joint if independent". We refer to values of *jii* as "Unigram Estimates (UE)" when we compare them to other estimates such as MLE or GT. In many of the following plots, we group bigrams into approximately 35 bins using the binning rule: $j = \left\lfloor 3\log_{10} jii \right\rfloor$.

Other second predictors are possible. We do not know what makes one variable better than another for grouping. A necessary property of the grouping variable is that it be possible to count the number of types included in each group, because we need to know N_0. We hypothesize that if one variable predicts r better than another, then it will make a better grouping variable. It is useful for smoothing that *jii* is a continuous variable.

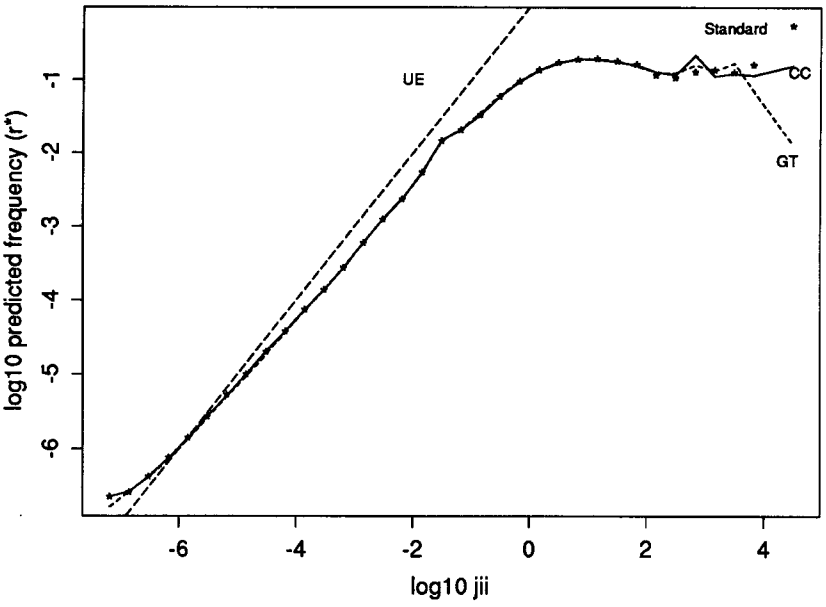
3. Qualitative Evaluation

We find that the both GT and CC estimates agree very well with the standard estimates over the entire range of data that we can test. The smallest frequency observations are the most critical. The following figure shows the results for $r = 1$. Five predicted frequencies are shown in this and following figures: (1) the standard, S, shown by points, (2) the maximum likelihood estimate, MLE, shown by long dashes, (3) the unigram estimate, UE, shown by long dashes, (4) the enhanced Cat-Cal estimate, CC, shown by a solid line, and (5) the enhanced Good-Turing estimate, GT, shown by short dashes. These estimates are plotted against the logarithm of the unigram estimator, *jii*. Note that CC and GT agree closely with the standard. They are quite distinct from either the MLE or UE but lie approximately between these two primary estimators.

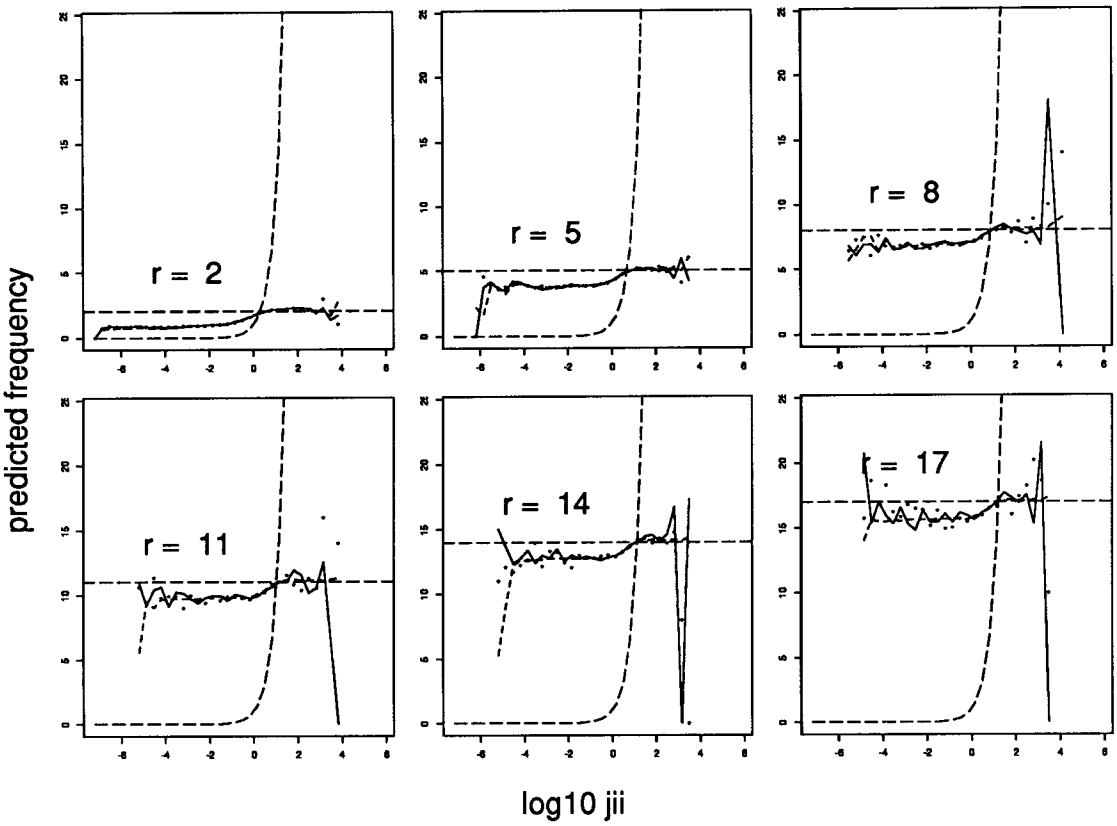
Enhanced Good-Turing and Cat-Cal Agree with the Standard for r=1



Enhanced Good-Turing and Cat-Cal Agree with the Standard for r=0



For frequency zero, the range of CC and GT is about five orders of magnitude, four orders of magnitude larger than for any other frequency. Over this range, both GT and CC agree well with the standard estimates. At the resolution shown, there is no visible difference between the three estimates for most of the range.



Note that r^* depends more on *jii* when r is small; the slope of r^* is very steep for $r = 0$, and pretty flat for $r = 17$. This means that UE is more important when r is small. We will return to this when we consider the number of significantly different probabilities.

4. Quantitative Evaluation

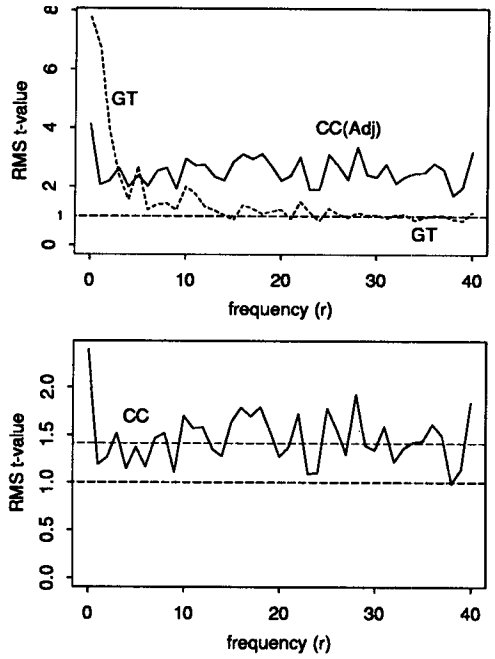
It is natural to evaluate methods with a t-score $t_{jr} = (r^*_{jr} - r^S_{jr})/\sigma_{jr}$, where r^*_{jr} is an estimate produced by one of the proposed methods for bin j and frequency r, r^S_{jr} is the standard for the same jr cell, and σ_{jr} is the standard deviation for the same jr cell. We use the GT method to estimate the standard deviation because it appears to match the CC variance while being less noisy and defined in more cells.

We have some expectations about these t-scores. A perfect predictor would give an RMS t-score of about one, because the variance of one standard observation is used as the denominator. We find that GT is nearly perfect with RMS t-scores very close to one except for small r. In contrast, CC is not perfect anywhere because both the categorization and the calibration samples have the assumed variance. However, when r is very small, it appears that the binomial assumption is inappropriate, and consequently, the the more empirical, though imperfect, CC method is preferable.

The two plots below show the RMS t-value averaged within each jii bin. The solid lines compare CC with the standard; the short dashed lines compare GT with the standard. The best performance theoretically possible is an RMS error of one, shown by a long dashed line in each panel. GT approaches this ideal quickly, though CC is preferable at very small frequencies. The CC values in the

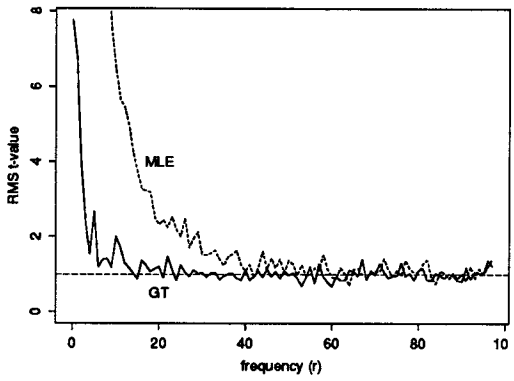
upper panel are adjusted for sample size to be comparable to GT values.

Comparison of the Enhanced Good-Turing and Cat-Cal Methods
Cat-Cal is better for small r and worse for large r



The following plot shows that MLE does not reach ideal performance within the range shown. Moreover, for frequencies less than about 40, MLE is substantially worse than GT. Over the smallest ten frequencies the MLE has RMS t-values ranging from five to thirty times those of enhanced Good-Turing estimates.

Comparison of the Enhanced Good-Turing and MLE Methods
Good-Turing is better, especially when r is small



5. How Many Significantly Different Probabilities?

In this section we show that estimates in adjacent *jii* bins differ quite significantly. This implies that interpolation is justified, and leads to an estimate of the equivalent number of significantly different estimates.

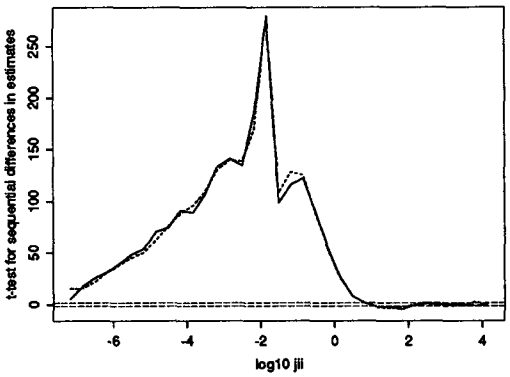
For each *jii*, let $\hat{f}_{jr}$ denote a frequency estimated for bigrams in the j^{th} bin and frequency r. Let $\hat{v}_{jr}$ be

variance of $\hat{f}_{jr}$. The following figure investigates the t-score

$$t = (\hat{f}_{jr} - \hat{f}_{(j-1)r}) \; / \; \sqrt{\hat{v}_{jr}/N_{jr} + \hat{v}_{(j-1)r}/N_{(j-1)r}}$$

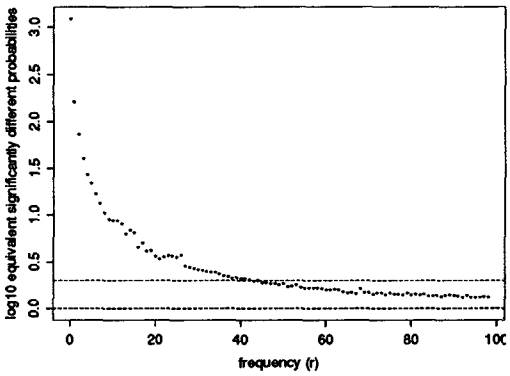
for the particularly important case of $r = 0$. The solid line shows the t-statistics for CC; the short dashed line shows the GT differences. Long dashed lines are drawn at conventional significance levels of ± 1.65. These differences are highly significant, indicating that interpolation between the observed values is justified. We estimate the equivalent number of significantly different values by taking the sum of all the t-statistics and dividing by 1.65. For $r = 0$, the equivalent number of significantly different values is 1245.

About 1200 Significantly Different Probabilities for $r = 0$



The following figure shows the equivalent number of significant differences as a function of frequency. The dashed lines are drawn at $\log_{10} 1$ and $\log_{10} 2$. While the number of significantly different values falls rapidly with increasing r, it remains above two through $r \approx 40$, and continues to be greater than one even through frequency 100. This range encompasses the majority of bigram tokens and indicates the value of a second predictor for practical applications, indicating that enhancement is of considerable value for practical applications.

Equivalent Number of Significantly Different Probabilities
we can distinguish bigrams with the same frequency very well for small frequencies



6. Conclusions

This paper has proposed two specific methods for backing-off bigram probability estimates to unigram probabilities: the enhanced Good-Turing method, and the Cat-Cal method. Three important points in this paper have extended the strength of these methods over previous methods:

- the use of a second predictor (e.g., *jii*) to exploit the structure of n-grams, the distinguishing feature between the *enhanced* Good-Turing method and the basic Good-Turing method.
- the estimation of variances for the bigram probabilities, which allows building significance tests for various practical applications, and in particular allows
- the use of refined testing methods that can show important qualitative differences even though quantitative differences may be small.

The use of a second predictor is the basis on which we distinguish the enhanced Good-Turing method (GT) proposed here from the basic Good-Turing method and the enhanced Cat-Cal (CC) from a basic Cat-Cal. If we had not introduced a second predictor, all bigrams that were observed once would be considered equally likely, and all bigrams that were observed twice would also be considered equally likely, and so on. This is extremely undesirable. Note that there are a large number of bigrams that have been seen just once (2,053,146 in a training corpus of 22 million words); we do not want to model all of them as equally probable. Much worse, there are a very large number of bigrams that have not been seen (160 billion bigrams in the same training corpus of 22 million words); we really do not want to model all of them as equally probable. By introducing the second predictor *jii* as we did, we were able to make much finer distinctions within groups of bigrams with the same number of observations r. In particular, for bigrams not seen in the training corpus, we have about 1200 significantly different estimates.

It would be interesting to consider other variables besides *jii*. One might consider, for example, the number of letters in the bigram. Katz (1987) proposes an alternative variable: the first word of the n-gram. Any variable that is not completely correlated with r would be of some use. *jii* has some advantages; it makes it possible to summarize the data so concisely that the relevant structure can be observed in a simple plot. Moreover, *jii* has a natural order and is continuous, so the number of bins can be adjusted for accuracy. In contrast, selecting the first word of the n-gram prescribes the number of bins.

The second point, the calculation of variances, is often not discussed in the literature on using the Good-Turing model for language modeling. Variances are necessary to make statements about the statistical significance of differences between observed and predicted frequencies. In other work (Church, Gale, Hanks, and Hindle, 1989), we have used variances to distinguish unusual n-grams from chance.

The third point we want to emphasize, the use of refined tests for differences in methods, is discussed in section 4. Four methods, MLE, UE, CC, and GT, were compared to the standard. t-scores were calculated for the differences between the standard and a proposed method and aggregate results across *jii*. We find that the GT method rapidly approaches ideal performance, though it is outperformed by CC when r is very small, presumably because the binomial assumption is apparently not quite satisfied for small frequencies.

There are many ways that the language model presented could be improved. We have said very little about the unigram model; in fact, the unigram model was estimated with the MLE method. One could apply the methodology developed here to improve greatly on this. One could also obtain much improved estimates by starting with a better sample; the 1988 AP corpus is not a balanced sample of general English. This paper is primarily concerned with developing methods and evaluation procedures; in future work, we hope to use these results to construct better language models.

References

Church, K, Gale, W., Hanks, P., Hindle, D., (1989) "Parsing, Word Associations and Typical Predicate-Argument Relations," International Workshop on Parsing Technologies, CMU, August.

Good, I.J., (1953), "The population frequencies of species and the estimation of population parameters," *Biometrika*, v. 40, pp. 237-264.

Kahan, S., Pavlidis, T., and Baird, H., (1987), "On the Recognition of Printed Characters of any Font or Size," *IEEE Transactions PAMI*, pp. 274-287, March, 1987.

Katz, S. M., (1987), "Estimation of probabilities from sparse data for the language model component of a speech recognizer," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, v. ASSP-35, pp. 400-401.

Nàdas, A., (1984), "Estimation of probabilities in the language model of the IBM speech recognition system," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, v. ASSP-32 pp. 859-861.

Nàdas, A., (1985), "On Turing's formula for word probabilities," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, v. ASSP-32 pp. 859-861.