Stability Measure of Entropy Estimate and Its Application to Language Model Evaluation

Jahwan Kim, Sungho Ryu, and Jin H. Kim
Div. of CS, Dept. of EECS, KAIST
373-1 Yuseong Guseong
Daejeon, South Korea
{jahwan,shryu,jkim}@ai.kaist.ac.kr

Abstract

We propose in this paper a stability measure of entropy estimate based on the principle of Bayesian statistics. Stability, or how the estimates vary as training set does, is a critical issue especially for the problems where parameter-to-data ratio is extremely high as in language modeling and text compression. There are two natural estimates of entropy, one being the classical estimate and the other the Bayesian estimate. We show that the difference of them is in strong positive correlation with the variance of the classical estimate when it is not so small, and propose this difference as stability measure of entropy estimate. In order to evaluate it for language models where estimates are available but posterior distribution is not in general, we suggest to use a Dirichlet distribution so that its expectation agrees with the estimated parameters and that the total count is preserved at the same time. Experiments on two benchmark corpora show that the proposed measure indeed reflects the stability of classical entropy estimates.

1. Introduction

There are problems which are bound to suffer from insurmountable lack of data in comparison with the number of parameters. Examples include language models [6] and text compression [2]. These problems resist purely statistical analysis due to this lack of data. For instance, the number of all trigrams for 10,000-word vocabulary is 1 trillion, while the training corpus consists rarely of more than 10 million words.

Our interest mainly lies in language models. Dedicated techniques such as discounting and backoff have been devised to overcome lack of data in language models, and such techniques provide rather satisfactory solutions. However, there has been little attention paid to stability of the estimates thus obtained, i.e., to how they vary on different training sets. Especially because of this extremely high parameter-to-data ratio, the stability of these estimates is as important as the estimates themselves, in the light of bias-variance decomposition [9]. The more parameter we have, the more likely it is that the variance over the choice of the training set becomes large, i.e., the more unstable the estimated parameters are.

We investigate in this paper the stability of entropy estimates, in the framework of Bayesian statistics [3]. Two different estimates can be obtained by taking expectation and entropy of posterior distribution in different orders. The difference of these two estimates is proposed as stability measure.

The first of these estimates, which will be henceforth referred to as classical estimate, is computed by first taking expectations of parameters and then computing entropy of them, It is simply the entropy of frequency counts or its slight modification. This classical entropy estimate gives the minimal expected code length in the Bayesian sense. Thus this estimate serves the purpose better for many applications. On the other hand in Bayesian statistics, posterior distribution is computed with an assumption on the prior distribution of parameters. Then the quantity of interest is estimated by expectation with respect to the posterior distribution. Entropy can be estimated in this way, and we refer this estimate as Bayesian estimate. For n-gram language models [6], multinomial distribution fits, whose parameters follow the natural conjugate Dirichlet distribution [4]. The Bayesian estimate in this case is computed in [14].

Although the classical estimate gives the minimal expected code length, it can be highly unstable. Therefore a stability measure will complement and support the classical entropy estimate, just as error bar does any statistical estimate.

We propose the difference of these two estimates as a stability measure of the classical entropy estimate. The ra-

Proceedings of the 9th Int'l Workshop on Frontiers in Handwriting Recognition (IWFHR-9 2004)
0-7695-2187-8/04 $20.00 © 2004 IEEE

tionales are (i) the Bayesian estimate is always smaller than the classical estimate; (ii) the Bayesian estimate is known to be more stable than the classical estimate; (iii) from global point of view, both estimates decrease as the size of training set increases; (iv) asymptotically these two estimates converge to the same limit. Thus when the propose measure is much greater than the variance of the classical estimate over the choice of training set, it is then bound to be in strong positive correlation with the variance.

To compute Bayesian estimates and hence the proposed measure, posterior distribution is required. Common n-gram model estimation methods such as discounts and backoff are not Bayesian approaches and do not result in posterior distribution. We suggest in such cases to adopt a Dirichlet distribution for which the compensated sample counts provided by the estimation methods provide are used as parameters. We can think of this as modified posterior distribution obtained by shifting slightly the Bayesian posterior distribution according to external linguistic knowledge. Note that the uncertainty of parameters are preserved in the modified distribution since the total count is also unchanged. Thus the proposed stability measure can be computed for the majority of language modeling methods.

Experiments on two benchmark corpora show that the proposed measure is indeed in strong positive correlation with the variance of estimates as the size of training data changes.

This paper is consists of six sections. The first section is this introduction. We review the Bayesian and classical estimates of entropy in section 2. The definition and the analysis of the stability measure is presented in section 3, followed by its application to language models in section 4. Experimental results are shown in section 5. We discuss the use and the limit of our measure in the final section.

2. The Bayesian and classical estimates of entropy

2.1 Multinomial and Dirichlet distributions

We recall the definition of multinomial distribution and its natural conjugate, Dirichlet distribution [4]. A discrete random vector $\mathbf{X} = (X_1, \ldots, X_m)$ with the constraint $\sum_{j=1}^{m} X_j = N$, has a multinomial distribution of dimension $m-1$ with parameters N an integer and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$ where $\boldsymbol{\theta}$ is also constrained by $\sum_{j=1}^{m} \theta_j = 1$ by $\theta_j > 0$ for all $j = 1, \ldots, m$, when its probability mass function is given by $p(\mathbf{x}|\boldsymbol{\theta}, N) = N! \prod_{j=1}^{m} \theta_j^{x_j} / \prod_{j=1}^{m} x_j!$, for an instance $\mathbf{x} = (x_1, \ldots, x_m)$ of $\mathbf{X}$. Here the parameter θ_j is the probability that the event j happens, and X_j is the number of the event j in N trials.

A Dirichlet distribution is the natural conjugate distribution for the multinomial distribution, and its probability density function is of the same form as the above equation, except that θ_j's are considered as random variables and x_j's as parameters. x_j must be strictly positive for all $j = 1, \ldots, m$, but they need not be integers. The probability distribution function is given as follows.

$$p(\boldsymbol{\theta}|x_1, \ldots, x_m) = \frac{\Gamma(\sum_{j=1}^{m} n_j)}{\prod_{j=1}^{m} \Gamma(x_j)} \prod_{j=1}^{m} \theta_j^{x_j - 1},$$

where $\Gamma(x)$ is the Gamma function [1]. As shown in the above equation, it is customary to subtract one from the exponents, to facilitate computation in the presence of prior probability. We will refer to $\sum_j x_j = N$ as total count, and each x_j as sample count.

The multinomial and Dirichlet distributions are one of the basic tools of Bayesian statistics. n-gram language models [6] can be analyzed using these distributions.

2.2 Bayesian estimate of entropy

The entropy H($\boldsymbol{\theta}$) of the probability mass function given by $\boldsymbol{\theta}$ can be computed as $H(\boldsymbol{\theta}) = -\sum_{j=1}^{m} \theta_j \log_2 \theta_j$. When $\boldsymbol{\theta}$ is unknown, $H(\boldsymbol{\theta})$ itself is a random variable. The Bayesian estimate of any random variable is the expectation of it with respect to the posterior distribution. In our case with Dirichlet posterior distribution, this expectation of entropy is calculated in [14]:

$$E(H(\boldsymbol{\theta})|D) = -(\log 2)^{-1} \cdot \sum_{j=1}^{m} \frac{x_j + 1}{N + m} \big(\psi(x_j + 2) - \psi(N + m + 1)\big), \quad (1)$$

where $E(\cdot|D)$ denotes expectation with respect to the posterior, and $\psi(x)$ denotes the digamma function, defined as the logarithmic derivative of the Gamma function [1]. At integers, the digamma function has a simple formula: $\psi(x) = -\gamma + \sum_{i=1}^{x-1} i^{-1}$, where x is a positive integer and $\gamma = \lim_{n \to \infty} (\sum_{i=1}^{n} 1/i - \log n)$ is the Euler constant. In purely Bayesian case with uninformative prior and without any modification on the posterior distribution, the parameters are all integers, and eqn. (1) reduces to

$$E(H(\boldsymbol{\theta})|D) = (\log 2)^{-1} \sum_{j=1}^{m} \frac{x_j + 1}{N + m} \sum_{i=x_j+1}^{N+m} \frac{1}{i}.$$

According to [14], the Bayesian estimate is more stable than the classical estimate, i.e., the variance is less for the Bayesian estimate as the training set varies, than for the classical estimate.

2.3 Classical estimate of entropy

It is customary to estimate entropy by first estimating the parameters using expectation of the posterior and then computing entropy, i.e, by $H(E(\boldsymbol{\theta}|D))$. Comparing this to the Bayesian estimate $E(H(\boldsymbol{\theta})|D)$ of the previous section, we see that the order of expectation and entropy is interchanged.

In the presence of these two natural choice of estimate, the question arises: Which one is *the better estimate?* The answer depends on where estimates are to be used.

As well known from information theory [7], the code with $\log_2 p_j$ as the length of each codeword has the minimal expected code length $H(\mathbf{p})$. Note however, that this result holds only when $\mathbf{p}$ is explicitly known. A simple but often-unmentioned fact is that the expected code length is minimized by using $-\log_2 E(\theta_j|D)$ as the length of each codeword, in the Bayesian sense. Indeed, let l_j be the length of each codeword. The average code length is then $\sum_{j=1}^{m} p_j l_j$. But p_j is never known, and we must replace it by the random variable θ_j, which incorporates our uncertainty about it. Thus the average code length is a random variable, while we have to fix l_j's. The Bayesian way is to minimize its *expectation* $E\left(\sum_{j=1}^{m} \theta_j l_j \middle| D\right) = \sum_{j=1}^{m} E(\theta_j|D) l_j$. Then following the same path as in the case where p_j's are known, we find that this is minimized when $l_j = -\log_2 E(\theta_j|D)$. Thus the classical estimate is the minimal expected code length.

3. Stability measure of entropy estimate

There are problems which suffers from insurmountable lack of data. That of estimating language model is one such instance, where the parameter-to-data ratio ranges from 10^1 to 10^5. Techniques have been developed to obtain estimates even for non-observed events, based on linguistic characteristics. However, there has been little research efforts regarding the stability of the estimates thus obtained. Especially because of this high parameter-to-data ratio, the stability of these estimates, i.e, how they vary as the training set does, is as important as the estimates themselves, when we take the bias-variance decomposition [9] into account.

We define $\Delta(D) = H(E(\boldsymbol{\theta}|D)) - E(H(\boldsymbol{\theta})|D)$, and propose $\Delta(D)$ as stability measure of entropy estimate. We will investigate properties of $\Delta(D)$ in the following sections.

3.1 Classical estimate is no less than Bayesian estimate

It is well-known that entropy is a concave function of its arguments [7]. Therefore we may apply Jensen's inequality to the two estimates of entropy in the previous section. We obtain $H(E(\boldsymbol{\theta}|D)) \geq E(H(\boldsymbol{\theta})|D)$, i.e., the Bayesian estimate is a lower bound of the classical estimate. See fig. 1. Thus $\Delta(D) \geq 0$.

The magnitude of $\Delta(D)$, the difference between two sides in the above eqn., depends on how much the posterior distribution $p(\boldsymbol{\theta}|D)$ is peaked. When it is severely peaked, the difference is small, and when it is flat, the difference is large. This is in the same vein with why bagging works [5].

3.2 Asymptotic behavior of $\Delta(D)$

We will show that $\Delta(D) \to 0$ as the size of data $|D|$ approaches ∞. We need to assume that as $|D| \to \infty$, the posterior distribution $p(\boldsymbol{\theta}|D) \to \delta_{\boldsymbol{\eta}}$ for some fixed $\boldsymbol{\eta}$, where $\delta_{\boldsymbol{\eta}}$ denotes the point mass at $\boldsymbol{\eta}$. This assumption is not restrictive at all. It just states that our belief about the parameter must be perfectly certain with the infinitude of data, and that parameters do converge.

Under this assumption, both of the two estimates approach $H(\boldsymbol{\eta})$ as $|D| \to \infty$. Hence, $\lim_{|D|\to\infty} \Delta(D) = 0$.

3.3 $\Delta(D)$ and stability of entropy estimates

We explain in this section why $\Delta(D)$ and the variance of the classical estimate of entropy over the choice of training set are strongly correlated, and when. Let Var_D denote the latter. Throughout this section, variance always refers to the variance over the choice of the training set. Let $|D|$ denote the size of the training set D.

We first observe two tendencies about these estimates: (i) the Bayesian estimate is more stable, i.e., has less variance than the classical estimate as we vary the training set; (ii) the two estimates both decrease from the global point of view, as $|D|$ increases. (i) is reported in [14], and is also verified in our experiments. (ii) is due to the property of entropy; entropy of estimates are always greater than entropy of true parameters [7].

When $\Delta(D)$ is much greater than Var_D, e.g., by three or more orders, the magnitude of $\Delta(D)$ will not change much by different choice of D. In other words, the magnitude of $\Delta(D)$ is almost determined by $|D|$. Thus $\Delta(D)$ is in strong negative correlation with $|D|$, when it is much greater than Var_D. On the other hand, Var_D is also in strong negative correlation with $|D|$. This is nothing but one of the fundamental principles of machine learning; the more data we have, the more certain we are about the estimated parameters.

Therefore, $\Delta(D)$ and the variance Var_D of the classical estimate over the choice of training set are in strong positive correlation, when the former is much greater than the latter. This justifies our proposal to use $\Delta(D)$ as the stability measure of the classical estimate of entropy.

Care must be taken, however, when using this measure. When Var_D gets smaller and becomes similar in order to $\Delta(D)$, the above argument does not hold any more. In such cases, typically $\Delta(D)$ itself is very small.

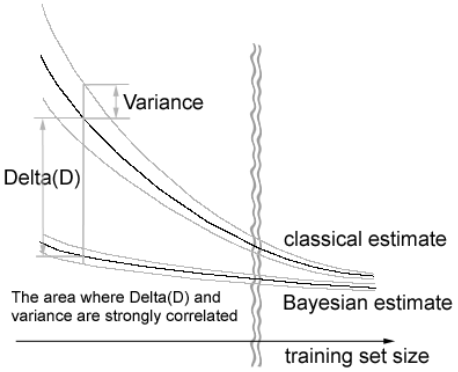

Figure 1. Bayesian and classical estimates. The classical estimate is always greater than the Bayesian estimate, and has greater variance. When the size of training set is relatively small, the proposed measure and the variance of the classical estimates are in strong positive correlation.

Note that the stability of estimates is hard to measure in any case. Cross validation with enough number of folds is the only possible way to obtain information about stability, i.e., the variance of the estimate. However, such method is computationally expensive, often making it impractical. Therefore our proposed measure provides valuable information at low computational cost.

4. Application to language models

n-gram language models [6] are widely in use as an indispensable tool in speech and character recognition problems, due to its simplicity and effectiveness. The probability of each word in a lexicon to occur following the history $n-1$ words is taken as parameter in n-gram models, under $(n-1)$-th order Markov assumption.

The number of each n-gram in a fixed context thus follows a multinomial distribution with n-gram probability as parameters. Those parameters follow Dirichlet distribution after observing data, in the Bayesian setup. Therefore we may apply the stability measure $\Delta(D)$ of the previous section to this problem of estimating parameters of language model without any change. Purely Bayesian methods with uninformative prior, sometimes called adding-one, is not commonly used in language modeling, due to its mediocre performance. The main problem of this method is that it assigns too high probability estimates for unobserved events, i.e., $1/(N+m)$, where N is the size of the training corpus and m the size of the lexicon [8].

On the other hand, since the celebrated Good-Turing formula [10], a number of estimation methods have been devised to better estimate parameters of n-gram language models by adaptation of external knowledge. See [6] for review of available methods. There are two common techniques in almost all of such methods, i.e., discounting and backoff. Discounting refers to the technique where the n-grams observed k times or more with sufficiently large k, e.g., $k \geq 5$, are assigned parameters slightly less than k/N, and the portion *discounted* in such a way is distributed to the less frequently observed ones. Backoff is the technique where parameters for infrequently observed n-grams are estimated on the basis of parameters for $(n-1)$-grams. As these methods are not developed in the Bayesian setup, only estimates of parameters are given, not the posterior distribution of them, in all such methods. To apply our stability measure, we need not only the estimate but also the posterior distribution, of parameters.

We proceed simply by assigning each parameter a Dirichlet distribution with the total count unchanged and the compensated sample counts as parameters, so that its expectation matches the estimates. In effect, the posterior is slightly shifted according as estimation method recalculate *compensated* sample counts. The uncertainty of parameters are also preserved since the total count is unchanged. Although this distribution is not derived logically with prior assumption, it incorporates both external knowledge and the observed data faithfully. This shift is similar to the prior assumption of Bayesian statistics in spirit that external knowledge is incorporated.

We will refer to thus obtained distribution as modified posterior distribution, and denote it by $p(\boldsymbol{\theta}|D')$. The stability measure computed with the modified posterior distribution will be denoted by $\Delta(D')$, or just by Δ.

In the remainder of this section, we will write down explicit formulas for several specific estimates of language models. For other smoothing methods, the modified posterior distribution can be found by fixing the total count N and adjusting each sample count accordingly.

4.1 The case of Good-Turing formula

Good-Turing estimate [10, 6] is an estimate that is central in many other smoothing techniques. It is based on the symmetry principle that all events with the same sample count must be assigned the same probability. Basically, to an event with sample count r, we compute the compensated sample count r^* by $r^* = (r+1) \cdot n_{r+1}/n_r$, where n_r denotes the number of the events with sample count r, or the count-count. Then we assign the probability r^*/N to the event. There are variants of this formula to deal with the case where $n_r = 0$ for some r. In any case, we can find the the compensated sample count r^*, computed so that $\sum_r n_r r^* = N$. Thus the modified posterior distribution

$p(\boldsymbol{\theta}|D')$ is the Dirichlet distribution with N trials and compensated sample counts as parameters.

We may now compute the Bayesian estimate of entropy based on this. Unlike the Bayesian smoothing method, there is no extra count of one due to the prior. Taking this into account with eqn. (1), we obtain

$$E(H(\boldsymbol{\theta}|D')) = -\sum_{r=0}^{\infty} \frac{n_r r^*}{N} \frac{\psi(r^*+1) - \psi(N+1)}{\log 2}$$

Therefore,

$$\Delta(D') = \sum_{r=0}^{\infty} \frac{n_r r^*}{N} \left(\frac{\psi(r^*+1) - \psi(N+1)}{\log 2} - \log_2 \frac{r^*}{N} \right)$$

r^* is in general not an integer and there is no simple formula for the digamma function $\psi(r^*+1)$ at the non-integer value r^*+1.

For purely Bayesian analysis, the sample count together with prior is always no less than 1. However, there are cases where we may have to consider the case with fractional sample counts, e.g., the language model with discount and backoff. In such cases, the estimate $-(\log 2)^{-1}\big(\psi(r^*+1) - \psi(N+1)\big)$ can be quite different from the classical estimate $-\log_2 r^*/N$, since the digamma function $\psi(x)$ is an increasing function on $x > 0$, $\psi(x) \to \infty$ as $x \to 0^+$, and $\psi(x)$ is negative for x less than approximately 1.46. This reflects the fact that our knowledge about the true parameter is extremely uncertain.

4.2 The case of backoff

Another indispensable method in language models is that of backoff. We will describe briefly about Katz backoff method [11], one of the simplest and most-widely used such methods. In this method, for unobserved or infrequently-observed n-grams, the estimates for the lower-orders are exploited. More precisely, let w_{i-n+1}^{i} be n-gram with the sample count r. Then the Katz backoff method assigns to it the compensated count $c_{\text{Katz}}(w_{i-n+1}^{i}) = d_r r$, if $r > 0$, and $c_{\text{Katz}}(w_{i-n+1}^{i}) = \alpha\left(w_{i-n+1}\right) c_{\text{Katz}}\left(w_{i-n+1}^{i-1}\right)$ if $r = 0$, where $c_{\text{Katz}}\left(w_{i-n+1}^{i-1}\right)$ denotes the estimate of the $(n-1)$-gram w_{i-n+1}^{i-1}, d_r a discount ratio, and $\alpha\left(w_{i-n+1}\right)$ a constant chosen in such a way that the total count N is unchanged. This formula defines c_{Katz} recursively in the size of n-grams, with unigram estimates given by the usual frequency counts. It is customary to use $d_r = 1$ for $r \geq 5$. The exact formula for d_r and $\alpha\left(w_{i-n+1}\right)$ can be found in [11, 6].

The key fact is that the total count N is unchanged in any case and we have the compensated sample count $c_{\text{Katz}}(\cdot)$. Therefore we may assign the Dirichlet distribution with parameters N and $c_{\text{Katz}}(\cdot)$. The stability measure then can be computed similarly:

Corpus	Language Model	Var(Tr)	test set cross entropy	Var(Test)
KAIST	Unigram	0.999	0.157	0.934
	Bigram	0.982	0.987	-0.319
	Trigram	0.965	0.999	-0.766
WSJ	Unigram	0.965	0.996	0.996
	Bigram	0.982	0.999	0.957
	Trigram	0.994	0.989	0.923

Table 1. Correlation with the proposed measure Δ

$$\Delta(D') = \sum_{\text{all } n\text{-grams } w} \frac{c_{\text{Katz}}(w)}{N} \left(-\log_2 \frac{c_{\text{Katz}}(w)}{N} + \frac{\psi\left(c_{\text{Katz}}(w)+1\right) - \psi(N+1)}{\log 2} \right).$$

5. Experiments

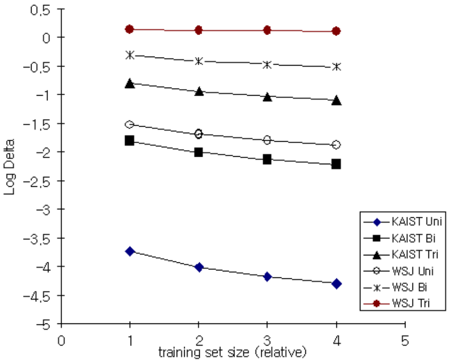

Figure 2. Δ in log scale, for the KAIST and the WSJ corpus

The proposed measure was tested on two corpora. One is the Wall Street Journal text corpus [12] of 1989, consisting of approximately one million English words. The other is the 1997 KAIST raw text corpus [13], consisting of thirty million Korean characters. The total number of unique Korean characters in this corpus is 2350, all in KS X 1001 character code. The KAIST corpus was collected from various domains such as literature, law, economics, politics and science. No linguistic preprocessing was applied to the corpora. A single Korean character plays the role of a single English word, as far as language model is concerned. Thus for the WSJ corpus the lexicon size is relatively large compared to the corpus size, while it is not so for the KAIST corpus.

Unigram, bigram, and trigram models were trained with Katz backoff method for each corpus. To measure the variance of parameters as the train set varies, each corpus are first randomly divided into the five folds of equal sizes, with domain ratio maintained as closely as possible. The size of the training set was varied from one fold to four folds. For each fixed training set size, the experiment was repeated five times with the test fold changed. The proposed measure Δ and the cross entropy are measured for each of the prepared training sets. Thus there are five different cross entropies for each fixed training set size, and the variance of them are also measured. Finally, the test set cross entropy and its variance for the fixed training set size are measured. For the training set, cross entropy is measured due to computational complexity, as an approximation of entropy. The results are shown in fig. 2.

First, as shown in table 1, there is strong positive correlation between the proposed measure Δ and the variance of the training set cross entropy. This is as explained in section 3.3. It is more subtle to interpret the result for generalization performance on the test set. What Δ can measure is the variance the entropy estimate, only when it is not too small. For the KAIST corpus, there seems to be enough data; the models are well-trained and thus the variance of entropy estimate is small. However for the WSJ corpus, the variance gets rapidly smaller as more data is used.

In general, there is a strong correlation between the proposed measure Δ and the test set cross entropy, whenever Δ is not too small. When Δ is small enough, the behavior of the entropy estimate is hard to predict based upon Δ alone.

6. Discussion

The proposed measure Δ is experimentally shown to indicate stability of a trained language model. When its scale is small enough, the model is trained with enough examples, i.e., stable and Δ alone cannot provide information on the generalization performance of the model. When its scale is big enough, the generalization performance is in strong positive correlation with Δ.

It is not clear, however, when Δ is small enough. The exact threshold will depends on rather intractable amount of other information. As there is no known criterion for overfitting, there can be no measure which exactly measures the variance [9] of generalization error, in principle. The proposed measure can only help by providing information on variance part of generalization error.

A threshold of 10^{-2} for Δ is recommended based on our experiment. When Δ is greater in order than this, there is a part of error present due to variance. When Δ is smaller in order than this, almost all error seem to be due to inherent bias of the model. An extensive experiment is desirable, to gain more insight on bias and variance of language models.

Although our interest mainly lies in the problem of language models, the proposed measure can be used wherever the posterior distribution or a similar distribution can be computed. It should be effective especially in cases where the parameter-to-data ratio is high and hence parameters are prone to be unstable.

References

[1] M. Abramowitz and I. A. Stegun, editors. *Handbook of Mathematical Functions*. U.S. Government Printing Office, Washington, D.C., 1970.

[2] T. C. Bell, J. G. Cleary, and I. H. Witten. *Text Compression*. Prentice Hall, Englewood Cliffs, N.J., 1990.

[3] J. O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York, 2nd edition, 1985.

[4] J. M. Bernardo and A. F. M. Smith. *Bayesian Theory*. John Wiley & Sons, New York, 1993.

[5] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[6] S. F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 310–318, 1996.

[7] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, New York, 1991.

[8] W. A. Gale and K. W. Church. What's wrong with adding one? In N. Oostdijk and P. de Haan, editors, *Corpus-Based Research into Language: In honour of Jan Aarts*, pages 189–200, Amsterdam, 1994. Rodopi.

[9] S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4:1–58, 1992.

[10] I. J. Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3–4):237–264, 1953.

[11] S. M. Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Trans. Acoust., Speech, Signal Processing*, 35(3):400–401, 1987.

[12] M. Marcus and et al. *The Penn Treebank Project*. `http://www.cis.upenn.edu/~treebank/home.html`.

[13] The National Academy of the Korean Language. *The 21st Century Sejong Project*. `http://www.sejong.or.kr`, 1990–.

[14] D. H. Wolpert and D. R. Wolf. Estimating functions of probability distributions from a finite set of samples. *Physical Review E*, 52(6):6841–6854, 1995.