Properties Analysis of Smoothing Method for Mandarin Language Models

Feng-Long Huang[+] and **Yih-Jeng Lin**

Text-To-Speech System Laboratory
Department of Applied Mathematics, National Chung-Hsing University,
Taichung 40227, Taiwan, R. O. C.
flhuang@amath.nchu.edu.tw, yclin@amath.nchu.edu.tw

ABSTRACT

In this paper, a set of statistical properties is used to evaluate several well-known smoothing methods. We first propose a set of properties to analyze the statistical behaviors of these methods. Furthermore, we present a new smoothing method which complies with all the proposed properties. Finally, we implement three Mandarin language models and then evaluate the cross entropies on various size N.

Keywords:*Statistical properties, Smoothing Method, Cross entropy, Language models.*

1. Introduction

Language models (LM) have been used in various tasks of natural language processing (NLP). An event can be regarded as a possible type of *n*-gram in LM, n>=1. We can calculate the probability for the each occurred event according to its count in corpora.

It is important in NLP to compute the probability of a sequence of words $W = w_1 w_2 w_3 ... w_m = w_1^m$. This probability will be denoted by $P(W) = P(w_1^m)$. We can use the chain rule of probability to decompose the probability:

$$P(w_1^m) = P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1^2)...P(w_m \mid w_1^{m-1})$$
$$= \prod_{k=1}^{m} P(w_k \mid w_1^{k-1}). \quad (1)$$

In practice, the probability should be estimated on the assumption that each word depends only on a limited number of preceding words. In the *n*-gram model, the conditional probability in Eq. (1) can be written as:

$$P(w_i \mid w_{i-n+1}^{i-1}) = \frac{C(w_{i-n+1}^{i-1} w_i)}{\sum_w C(w_{i-n+1}^{i-1} w)} \quad (2)$$

where C(·) denotes the count of an event in the training corpus.

The probability P of Eq. (2) is the relative frequency and such a method of parameter estimation is called *maximum likelihood estimation* (MLE). As shown in Eq (2), we can estimate the probability of a word sequence W with MLE.

Because of zero count of an event, such a method will lead to the degradation of performance. For a given word w_{i-1} in bigram model, if a bigram $w_{i-1}w_i$ never occur in the training corpus, then $C(w_{i-1}w_i)$ is equal to 0. It is apparent that Eq. (2) evaluates to zero. Yet an unseen event in a word sequence W does not mean that $P(W)$ should be zero. The schemes used to resolve this problem are called *smoothing*. Smoothing methods are usually used to re-estimate the probability for each possible event. There are some well-known methods: *Additive discount*, *Good-Turing*, *Witten-Bell*, *Absolute discount*, and so on.

2. Previous Smoothing Methods

In this section, we review some famous smoothing methods and only consider *n*-gram Mandarin language models in our paper. Let a type t_i be a possible event in an *n*-gram model and $C(t_i)$ be the count (the number of times) that type t_i occurred in the training corpus. We use N to denote the number of occurrences of all the type, That is,

$$N = \sum_i C(t_i). \quad (3)$$

Also we use B (bins) to denote the number of possible types. Then we have $B=V$, $B=V^2$, and $B=V^3$ for the unigram, bigram, and trigram models,

[+] Correspondence author.

correspondingly, where V is the vocabulary size (the number of Mandarin characters in our discussion).

Additive smoothing method is intuitively simple. A small δ is added into all types (including all seen and unseen types). Typically, $0< \delta <=1$. The case δ =1 is called *add-1* smoothing. The adjusted count c^* is defined as:

$$c^* = (c + d)\frac{N}{N + Bd}, \quad c \geq 0. \tag{4}$$

According to the previous experiments [1], the performance was usually degraded by using *add-1* smoothing.

Good-Turing was first described by Good in 1953 [2]. Some related works are [3] and [4]. Let n_c denote the number of types with count c in the corpus. For example, n_0 represent that the number of types with zero count and n_1 means the number of types which exactly occur once. Therefore, n_c will be described as:

$$n_c = \sum_{i:C(t_i)=c} 1. \tag{5}$$

Based on *Good-Turing* smoothing, the redistributed count c^* will be presented in term of n_c, n_{c+1} and c as:

$$c^* = (c+1)\frac{n_{c+1}}{n_c} \tag{6}$$

We discuss two of five smoothing schemes introduced by *Wetten-Bell*[1][6]; called *W-B A* and *C*. In method *W-B A*, just one count is allocated to the probability that an unseen bigram will occur next. The probability mass P_{mass} assigned to all unseen bigrams can be summed up to $1/(N+1)$. Let $P^*_{i,N}$ be the smoothed probability of type t_i in a training data od size N. Then,

$$P^*_{i,N} = \begin{cases} \dfrac{1}{U(N+1)} & \text{for } C(t_i) = 0, \\ \dfrac{C(t_i)}{N+1} & \text{for } C(t_i) \geq 1, \end{cases} \tag{7}$$

where U is the number of unseen types, i.e.,

$$U = \sum_{i:C(t_i)=0} 1. \tag{8}$$

Each smoothed count c^* in *W-B C* is described as:

$$c^* = \begin{cases} \dfrac{S}{U}\dfrac{N}{N+S}, & \text{if } c = 0 \\ c\,\dfrac{N}{N+S}, & \text{if } c > 0 \end{cases} \tag{9}$$

[1] There are 5 methods in [6]; method *A*, *B*, *C*, *P* and *X*. We just discuss two of them (*W-B A* and *C*) in this paper.

where S is the number of kinds of seen types, i.e.,

$$S = \sum_{i:C(t_i)\geq 1} 1. \tag{10}$$

The discounted probability will be expressed for seen bigrams as:

$$P^*_{i,N} = \frac{c_i}{N+S}, \quad \text{if } c_i > 0 \tag{11}$$

Absolute discount, introduced by authors of Ney and Essen [5], is an interpolating scheme and looks like method of Jelinek and Mercer [3]. The method interpolates the higher and lower order models; the higher order distribution will be calculated just subtracting a static discount D from each n-gram with non-zero count.

3. Proposed Properties

In this section, we propose five properties which can be regarded as statistical features. These properties will be further used to analyze the statistical behaviors of the smoothing methods.

Property 1: The smoothed probability for any one type t_i should fall between 0 and 1 (0,1), which is described as:

$$0 < P^*_{i,N} < 1 \tag{12}$$

Property 2: The summation of smoothed probability P^* for all the types is necessarily equal to 1 on any training size N. Total probability P is summed as:

$$\sum P^*_{i,N} = \sum_{t_i \in seen\ types} P^*_{i,N} + \sum_{t_j \in unseen\ types} P^*_{j,N} = 1 \tag{13}$$

Property 3: Let $Q^*_{c,N}$ be the smoothed probability of a type with count c on a training corpus of size N. That is,

$$Q^*_{c,N} = P^*_{i,N}, \quad C(t_i) \geq 1. \tag{14}$$

The smoothed probability assigned to the type t_i with different count should satisfy all the following inequality equations:

$$Q^*_{c,N} < Q^*_{c+1,N}, \qquad \text{for } c=0,1,2,\ldots, \tag{15}$$

Inequality Eq. (15) describes the concept that smoothed probability for any type with same count should be the same. Instead, the probability for type t with count c+1 should be larger than that of type with count c.

Property 4: Comparing to the probability P prior to smoothing process, the smoothed probability P^* for all types will be changed. Since we assign some probability to the unseen types, the smoothed

probability of he unseen types should be higher than zero obtained from MLE. Property 4 can be expressed as follows:

$$Q^*_{0,N} > Q_{0,N}, \quad \text{for } c = 0 \tag{16}$$

$$Q^*_{c,N} < Q_{c,N}, \quad \text{for } c \geq 1 \tag{17}$$

Property 5: We have $B=S+U$ for the language models. When the number of training size is Increased, all the smoothed probability Q^* for type with same count on training size N+1 should be smaller than Q^* on training size N. For instance, when an incoming type (say t_{next}) occurs, the training size is increase by one (now $N=N+1$). The smoothed probability Q^* on N+1 training set should be less than the probability Q^* on N for $c \geq 0$, except the P^* for the incoming bigram t_{next}.

In other words, in addition to the P^* of t_{next} at training size N+1, all other smoothed probability Q^* at training size N+1 will be decreased than those at training size N. In summary, property 5 can be expressed as:

$$Q^*_{c,N} > Q^*_{c,N+1}, \tag{18}$$

$$Q^*_{c,N} < Q^*_{c+1,N+1}. \tag{19}$$

4. Properties Analysis

From the statistical points, smoothed probability for bigrams computed from various smoothing methods should still comply with these properties. Based on the statistical properties, we will analyze the rationalization of each smoothing models. For simplicity, only *Good-Turing* will be analyzed using five properties in this section.

4.1 The Analysis of Good-Turing Smoothing

Referring to Eq. (7), total number of smoothed count can be computed as:

$$\sum_i c_i n_i = c_0 n_0 + c_1 n_1 + c_2 n_2 + \ldots = N, \text{ for } i >= 0.$$

Properties 1, 2 and 3 do not hold. For instance, the following case: n_m is equal 0, $(m-1)^* = (m-1+1)\frac{n_m}{n_{m-1}} = 0$, and $m^* = (m+1)\frac{n_{m+1}}{n_m} = \infty$ (violate property 1 and 2). In such a case, it is obvious that:

$$Q^*_{m-2,N} > Q^*_{m-1,N} \text{ and } Q^*_{m,N} > Q^*_{m+1,N}$$

Hence, the results also violate the property 3.

It is possible that one of n_m for certain amount of training data set will be zero. The smoothed probability for unseen and seen bigrams with c counts, property 4 does not hold.

When a new type t_{next} is read in, then training size is increased by one ($N=N+1$). As shown in Eq. (6), the smoothed count $c^* = (c+1)\frac{n_{c+1}}{n_c}$. Supposed that the type t_{next} is ever seen with count c on training size N, upon the t_{next} appears, $N=N+1$, $n_c=n_c-1$ and $n_{c+1}=n_{c+1}+1$, the smoothed probability for types with c on training size N and N+1 can be computed as:

$$Q^*_{c,N} = (c+1)\frac{n_{c+1}}{n_c}/N \text{ and } Q^*_{c,N+1} = (c+2)\frac{n_{c+1}+1}{n_c-1}/(N+1)$$

Therefore, the ratio of Q^* is:

$$\frac{Q^*_{c,N}}{Q^*_{c,N+1}} = \frac{(N+1)\frac{n_{c+1}}{n_c}}{N\frac{n_{c+1}+1}{n_c-1}} = \frac{(N+1)(n_c-1)n_{c+1}}{Nn_c(n_{c+1}+1)} \tag{20}$$

According to Eq. (18), $Q^*_{c,N} > Q^*_{c,N+1}$. Therefore, Eq. (19) should be greater than 1. In fact, $N>>n_c$ and $N>>n_{c+1}$ while Eq. (20) may be < 1 on certain situation. Hence, property 5 does not hold.

For the type t_{next}, what is the relation between the smoothed probabilities P^* on training size N and N+1?

As we know:

$$P^*_{c,N} = (c+1)\frac{n_{c+1}}{n_c}/N, \quad P^*_{c+1,N+1} = (c+2)\frac{n_{c+2}}{n_{c+1}+1}/(N+1).$$

then:

$$\frac{P^*_{c,N}}{P^*_{c+1,N+1}} = \frac{(c+1)(N+1)n_{c+1}(n_{c+1}+1)}{(c+2)Nn_c n_{c+2}} \tag{21}$$

According to Eq. (19) of property 5, Eq. (21) should be less than 1. It is obvious that Eq. (21) may be greater than 1 in certain situations, while it is possibly less than 1. Therefore, property 5 does not hold.

4.2 Our Smoothing Method and Its Properties

We will propose a new smoothing method; *Yu-Huang* (called *Y-H* hereafter) and then analyze the statistical behaviors of these methods.

We describe another smoothing scheme; in which the probability mass for unseen bigrams is assigned $Ud/(N+1)$. Consequently, it varied with N and U; the number of training data and types of unseen types.

The smoothed probabilities will be calculated as:

$$P^*_{i,N} = \begin{cases} \dfrac{d}{(N+1)} & \text{for } C(t_i) = 0, \\ \dfrac{C(t_i)}{N}\dfrac{N+1-Ud}{N+1} & \text{for } C(t_i) \geq 1, \end{cases} \tag{22}$$

and

$$d < \min\{\frac{N}{N+2U}, \frac{N+2}{2U}\} \tag{23}$$

When computing the smoothed probability P^*, our proposed method don't employ interpolating scheme to combine the high order models with lower order

models. As shown of Eq. (22), $(N+1-Ud)/(N+1)$ is the normalization factor of Q^* for seen types. The probabilities Q will be discounted by the normalization factor and then the remained Q^* are redistributed to unseen types; which share uniformly the distributed probability mass $Ud/(N+1)$

We will analyze further the statistical behaviors of *Y-H* smoothing. As shown in Eq. (22), the smoothed probabilities for seen and unseen types will be (0,1). Therefore, property 1 does hold. The total probabilities for seen and seen types can be summed as:

$$\sum_{i:C(t_i)\geq 1} Q^*_{c,N} + \sum_{i:C(t_i)=0} Q^*_{c,N} = \sum_{i:C(t_i)\geq 1} \frac{C(t_i)}{N}\frac{N+1-d}{N+1} + \sum_{i:C(t_i)=0} \frac{d}{N+1}$$
$$= \frac{N}{N}\frac{N+1-Ud}{N+1} + \frac{Ud}{N+1} = 1$$

So, property 2 does hold. The smoothed probability Q^* for bigrams with c and $c+1$ counts on training size n is calculated as follows. For $c=0$ and 1,

$$Q^*_{1,N} - Q^*_{0,N} = \frac{1}{N}\frac{N+1-Ud}{N+1} - \frac{d}{N+1}$$
$$= \frac{N+1-Ud-Nd}{N(N+1)} = \frac{N+1-d(N+U)}{N(N+1)} \qquad (24)$$

Due to the condition of d (see Eq. (22)), Eq. (24) is larger than 0. For $c>1$,

$$Q^*_{c,N} - Q^*_{c+1,N} = \frac{c(N+1-Ud)}{N(N+1)} - \frac{(c+1)(N+1-Ud)}{N(N+1)}$$
$$= \frac{-(N+1-Ud)}{N(N+1)} \quad \text{for } c>=1. \qquad (25)$$

Eq. (25) will be less than 0. Referring to the results of Eqs. (24) and (25), we can conclude that property 3 does hold. Original and smoothed probability for types is as follows:

$$Q^*_{c,N} - Q_{c,N} = \frac{c(N+1-Ud}{N(N+1)} - \frac{c}{N}$$
$$= \frac{-cUd}{N(N+1)} < 0 \qquad \text{for } c>=1. \qquad (26)$$

$$Q^*_{c,N} - Q_{c,N} = \frac{d}{N+1} - 0 > 0 \qquad \text{for } c=0. \qquad (27)$$

As shown of Eq. (26) and (27), we can conclude that property 4 does hold. Finally, we analyze property 5. Smoothed probabilities for types with count c on N and $N+1$ training data are calculated as:

$$Q^*_{0,N} - Q^*_{0,N+1} = \frac{d}{(N+1)} - \frac{d}{(N+2)}$$
$$= \frac{d}{(N+1)(N+2)} > 0 \text{ for } c=0.$$

$$Q^*_{c,N} - Q^*_{c,N+1} = \frac{c(N+1-Ud)}{N(N+1)} - \frac{c(N+2-Ud)}{(N+1)(N+2)}$$
$$= c\frac{N+2-2Ud}{N(N+1)(N+2)} \qquad \text{for } c>=1. \qquad (28)$$

Numerator $(N+2-2Ud)$ should be larger than 0 because $d<(N+2)/2U$. We can prove that Eq. (28) is greater than 0. From the results above, property 5 does hold.

4.3 Summary of the Properties

In Table 1, the relationship between 6 smoothing methods and five proposed properties are shown. Among these smoothing methods, there isn't any method which completely comply with 5 proposed properties. *Additive discounting* and *W-B C* don't comply only one of five properties. All other methods do not comply with more than two properties. However, Our smoothing *Y-H* does satisfy all properties. Notations O and X denote the method does and does not comply with the property, respectively.

Table 1: The relationship of 6 methods and proposed statistical properties.

Method \ property	1	2	3	4	5
Add-one discount	O	O	O	X	O
Good-Guring	X	X	X	X	X
Witten-Bell (A)	O	O	O	O	X
Witten-Bell (C)	O	O	X	O	O
Absolute discount	O	O	X	O	X
Yu-Huang	O	O	O	O	O

5. Evaluation of Cross Entropy

5.1 Data Sets and Empirical Language Models

In the following experiments, a text sources is used as data sets; the Mandarin news texts collected from Internet and divided into two parts; training set and test set. The html tags and all unnecessary symbols are extracted and there are about 30M(10^6) Mandarin characters of news texts.

We construct three language models: Mandarin character unigrams, character bigrams and trigram model, to evaluate the cross entropy ***CH*** of smoothing methods discussed. The cross entropies of first two models are evaluated on various data size on ratio 4:1 of training and test sets, from 4M to 12M Mandarin characters. The third model employs 30M characters as training set (trigrams).

In our experiments, δ =1 for *additive* smoothing and discount constant D =1 for *absolute discount* smoothing. A set of cut-off value K on various training size N for Good-Turing is used to avoid the

failure (such as violating the properties in Section 3) of probability estimation.

5.2 Results

Figure 1~3 display three empirical results of cross entropies ***CH*** for six smoothing methods discussed in this paper. Basically, the method with lower cross entropy may perform better in NLP.

Figure 1 gives the cross entropies for Mandarin characters unigram model. Although the difference of ***CH*** between various methods is not obvious, *Good-Turing* and *Y-H* methods have lower *CH*. The average of cross entropy is near 9.81. Figure 2 shows Mandarin character model. For all methods, the cross entropies will decrease gradually on increasing training data set *N*. Among these methods, *Good-Truing* always obtains lower *CH* through different *N* by using a cut-off *K*. Our *Yu-Huang* obtains 17.8 in average and a little higher *CH* than that of *Good-Truing*, while bwer than that of all other methods. Figure 3 shows trigram model on training size N=30M characters. Three methods; *Good-Truing*, *W-B C* and *Yu-Huang*, generate closer *CH*. It is apparent that *Add-1* always obtains highest *CH* For both bigram and trigram models.

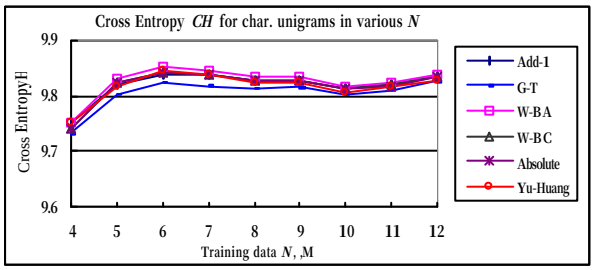

Figure 1: Cross Entropies of 6 smoothing methods for Mandarin character *unigram*.

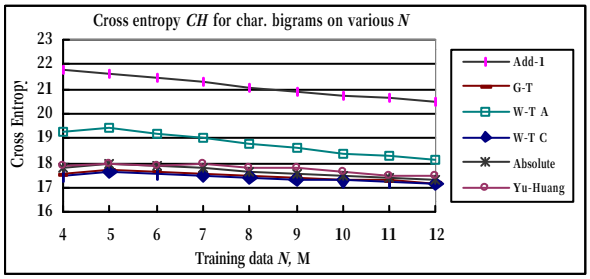

Figure 2: Cross Entropies of 6 smoothing methods for Mandarin character *bigram*.

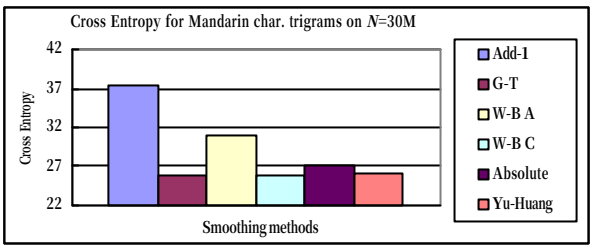

Figure 3: Cross Entropies of 6 smoothing methods for trigram on 30M Mandarin characters.

6. Conclusion

In the paper we propose 5 statistical properties to evaluate 5 well-known smoothing methods employed to solve the zero-count problem for language model. An effective smoothing method is also proposed and evaluated by these properties. For each method, every property is proven and 5 previous methods can't satisfy these properties while our method satisfies all the properties; which presents the method will fit the application of NLP and holds better statistical behaviors. Based on the experiment results, our smoothing method always gets lower ***CH*** than three previous methods and almost equal to that of *Good-Truing* smoothing method.

References

[1] Chen Standy F. and Goodman Joshua, 1999, An Empirical study of smoothing Techniques for Language Modeling, Computer Speech and Language, Vol. 13, pp. 359-394.

[2] Good I. J., 1953, The Population Frequencies of Species and the Estimation of Population Parameters, Biometrika, Vol. 40, pp. 237-264.

[3] Jelinek F. and Mercer R. L., 1980, Interpolated Estimation of Markov Source Parameters from Spars Data, Proceedings of the Workshop on Pattern Recognition in Practice, North-Holland, Amsterdam, The Northlands, pp. 381-397.

[4] Nádas A., 1985, On Turing's Formula for Word Probabilities, IEEE Trans. On Acoustic, Speech and Signal Processing, Vol. ASSP-33, pp. 1414-1416.

[5] Ney H. and Essen U., 1991, On Smoothing Techniques for Bigram-Based Natural Language Modeling, IEEE International conference on Acoustic, Speech and Signal Processing, pp. 825-828.

[6] Witten L. H. and Bell T. C., 1991, The Zero-Frequency Problem: Estimating the Probabilities of Novel Events in Adaptive Text Compression, IEEE Transaction on Information theory, Vol. 37, No. 4, pp. 1085-1094.