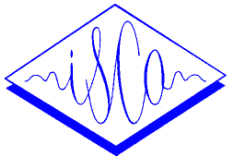

7th International Conference on Spoken Language Processing [ICSLP2002]
Denver, Colorado, USA
September 16-20, 2002

IMPROVED KATZ SMOOTHING FOR LANGUAGE MODELING IN SPEECH RECOGNITON

Genqing WU, Fang ZHENG[*]*, Wenhu WU, Mingxing XU, and Ling JIN*

Center of Speech Technology, State Key Laboratory of Intelligent Technology and Systems
Department of Computer Science and Technology, Tsinghua University, Beijing, 100084, China
[wgq,fzheng,wwh,xumx,jil]@sp.cs.tsinghua.edu.cn, http://sp.cs.tsinghua.edu.cn

ABSTRACT

In this paper, a new method is proposed to improve the canonical Katz back-off smoothing technique in language modeling. The process of Katz smoothing is detailedly analyzed and the global discounting parameters are selected for discounting. Further more, a modified version of the formula for discounting parameters is proposed, in which the discounting parameters are determined by not only the occurring counts of the n-gram units but also the low-order history frequencies. This modification makes the smoothing more reasonable for those n-gram units that have homophonic (same in pronunciation) histories. The new method is tested on a Chinese Pinyin-to-character (where Pinyin is the pronunciation string) conversion system and the results show that the improved method can achieve a surprising reduction both in perplexity and Chinese character error rate.

1. INTRODUCTION

The statistical language model (LM) is widely used in automatic speech recognition applications. Usually the language model is trained from a large text corpus using Markov model [1], in which the word sequence is treated as the observation of an $n-1$ order Markov process called n-gram. That is to say, only the most recent $n-1$ words are used to predicate the coming word. In this assumption framework, the probability of a word sequence $s = w_1 w_2 ... w_n$ can be estimated as $P(s) = P(w_1)P(w_2 \mid w_1)\prod_{i=3}^{n} P(w_i \mid w_{i-2} w_{i-1})$, and then some methods, such as maximum likelihood estimation (MLE) are used to estimate the conditional probabilities of the n-gram units. But no matter how widely the training corpus covers, the data sparseness problem is still very serious. In order to deal with this problem, some smoothing techniques are proposed, such as deleted interpolation smoothing, Katz smoothing [2], Kneser-Ney smoothing [3], and so on. Now smoothing is one of the key issues in language modeling.

The basic idea of language model smoothing is simple, and it just tries to *take out* some occurring counts from the seen units and then redistribute them to the unseen ones. Actually, the detailed measures are quite different in these methods. Deleted Interpolation method uses interpolation weights to combine uni-gram, bi-gram and tri-gram probabilities. Firstly the whole corpus is divided into two distinct parts; the larger one is used to estimate the conditional probabilities of the coming words given the corresponding histories, while the smaller used to estimate the interpolation weights among the three relative probabilities. The Back-off method is based on the simple idea that the seen units should give the unseen ones some appearance counts, and then the unseen units should partition the counts according to the low order n-gram probabilities. Both the back-off method and the deleted interpolation can achieve good performance in experiments, but they behave quite differently when the size of the training corpus varies. The Back-off method outperforms when the size of training corpus increases, while deleted interpolation does when the size of training corpus is small.

The Katz smoothing is one of the most important back-off smoothing methods, and it was widely used in speech recognition systems. But we argue that it still has the defect that it often misleads the decoding process when the n-gram units with homophonic histories are compared (this will be detailedly discussed in Section 3). The motivation of this paper is to make the smoothing more reasonable in this case.

In Section 2, a review on canonical Good Turing estimation [4] and Katz smoothing is given. In Section 3, our detailed analysis and improving on Katz smoothing is described. In Section 4, the experiments results are given and discussed. Conclusion is drawn in Section 5.

2. CANONICAL GOOD TURING AND KATZ SMOOTHING

As mentioned above, the main idea of the Good Turing method is to take out some frequencies from the seen n-gram units to the unseen ones, thus the n-gram units counts space can be shared. Detailed speaking, if a n-gram units occurs r times, it will be treated as r^* times, which is calculated as follows

$$r^* = (r+1)\frac{n_{r+1}}{n_r} \qquad (1)$$

where n_r denotes the number of n-grams occur exactly r times. After normalization, the relative frequency

[*] The author is currently with Beijing d-Ear Technologies Co., Ltd.

$$P_{GT}(w_1 \ldots w_n) = \frac{r^*}{\sum_{r=0}^{\infty} n_r r^*} \quad (2)$$

is taken as the re-estimated probability. Obviously, the probability sum of the n-grams occurring in the training data is

$$\sum_{w_1^n : c(w_1^n) > 0} P_T(w_1^n) = 1 - \frac{n_1}{N} \quad (3)$$

and then the discounted probabilities n_1 / N will be partitioned averagely by the unseen n-gram units.

Katz smoothing extends Good Turing to a new version. It redistributes the discounted probabilities to the unseen units via the recursive utilization of lower level conditional distributions. The following is the formula of standard Katz smoothing for tri-gram model

$$P_{Katz}\left(w_i \mid w_{i-2}^{i-1}\right) = \begin{cases} C\left(w_{i-2}^{i}\right) / C\left(w_{i-2}^{i-1}\right), & if \quad r > r_T \\ d_{r,3} C\left(w_{i-2}^{i}\right) / C\left(w_{i-2}^{i-1}\right), & if \quad 0 < r \le r_T \\ \alpha\left(w_{i-2}^{i-1}\right) P_{Katz}\left(w_i \mid w_{i-1}\right), & if \quad r = 0 \end{cases} \quad (4)$$

where $C(\cdot)$ stands for the occurring count of the specified event, r stands for $C(w_{i-2}^{i})$ for convenience, r_T is a count threshold for discounting purpose, $\alpha(w_{i-2}^{i-1})$ and $d_{r,3}$ (subscript 3 means it is for tri-gram) are the smoothing parameters for tri-gram, $d_{r,3}$ is calculated as follows

$$d_{r,3} = \frac{\frac{r^*}{r} - \frac{(r_T+1)n_{r_T+1}}{n_1}}{1 - \frac{(r_T+1)n_{r_T+1}}{n_1}} \quad (5)$$

If $d_{r,3}$ is determined, $\alpha(w_{i-2}^{i-1})$ can be calculated as

$$\alpha(w_{i-2}^{i-1}) = \frac{1 - \sum_{w_i : r > 0} P_{katz}\left(w_i \mid w_{i-2}^{i-1}\right)}{1 - \sum_{w_i : r > 0} P_{katz}\left(w_i \mid w_{i-1}\right)} \quad (6)$$

The Katz smoothing outperforms the Good Turing because it redistributes different probabilities to different unseen units. For example, word pair "*economic information*" and "*economic fluctuation*" will be assigned with different probabilities according to the fact that the second words (the coming word) are not the same.

Katz smoothing method has been widely used in speech recognition system since it was raised in 1987 and it yields good performance. But it still has its shortages, which will be discussed in section 3.

3. PROPOSED IMPROVING APPROACHES

It should be pointed out that $d_{r,n}$ is the key factor in Katz smoothing. Every *n*-gram units occurring exactly r times will be treated as $d_{r,n} \bullet r$ times, that is to say, $(1 - d_{r,n}) \bullet r$ times will be removed from every such units and redistributed to the unseen ones, so the back-off parameter $\alpha(\bullet)$ is determined by $d_{r,n}$.

Some ones calculate $d_{r,n}\left(w_{i-n+1}^{i-1}\right)$ instead of $d_{r,n}$ [5], that is to say, the discounting parameters are calculated according to the histories respectively, the n-gram units with the same history will share the same $d_{r,n}$, so $n_1\left(w_{i-n+1}^{i-1}\right)$ and $n_{k+1}\left(w_{i-n+1}^{i-1}\right)$, which denote the number of units occurring once and *k*+1 times with the history w_{i-n+1}^{i-1}, will take the place of n_1 and n_{k+1} in equation (5). This is very unreasonable because of the fact that the data sparseness is very serious in large vocabulary language model and $n_r\left(w_{i-n+1}^{i-1}\right)$ is very small for most of the history w_{i-n+1}^{i-1}.

In our system, a corpus with 200 million Chinese characters is used for training and about 7 million distinct bi-gram units and 29 million distinct tri-gram units are observed, Table 1 is the occurring counts distribution of the n-gram units. Parameters $n_{r,2}$ and $n_{r,3}$ denote the counts of bi-gram and tri-gram units occurring r times without considering the history, while $n_{r,3}(h_1)$ and $n_{r,3}(h_2)$ are examples of tri-gram units counts distribution considering the corresponding bi-gram histories h_1 and h_2. In our training process, r_T =5. In addition, $c(h)$ is the occurring count of the history h.

r	**1**	**2**	**3**	**4**	**5**	**6**	**C(h)**
$n_{r,2}$	3,815k	1,040k	482k	285k	190k	137k	/
$n_{r,3}$	21,529k	3,842k	1,432k	743k	449k	307k	/
$n_{r,3}(h_1)$	201	35	17	6	4	3	628
$n_{r,3}(h_2)$	3	1	0	1	0	0	9

Table 1: n-gram units occurring count distribution

Then the d_r's can be calculated respectively from the data in Table 1, for the sake of convenience, we call $d_{r,3}(h_1)$ and $d_{r,3}(h_2)$ local d_r. The result is shown in Table 2.

R	**1**	**2**	**3**	**4**	**5**
$d_{r,2}$	0.420	0.611	0.730	0.787	0.828
$d_{r,3}$	0.297	0.518	0.663	0.732	0.804
$d_{r,3}(h_1)$	0.284	0.702	0.419	0.817	0.890
$d_{r,3}(h_2)$	Overflows, incalculable				

Table 2: The global d_r and the local d_r

It is shown in Table 2 that the global discount parameters $d_{r,2}$ and $d_{r,3}$ are very reasonable, and the local discount parameters $d_{r,3}(h_1)$ are similar to $d_{r,3}$, but when $n_{r,3}(h)$ decreases (as $n_{r,3}(h_2)$ does), the local discount parameters become quite unreliable or even incalculable. Because the frequencies of all most all the units are very low, so it is more reliable to choose the global discount parameters $d_{r,2}$ and $d_{r,3}$ for all the units and our elementary experiments prove that this selection is wise.

But there are still some problems in this baseline when the homophonic (same in pronunciation) histories are considered. Let's take tri-gram as an illustration. According to equation (4), the discounted probabilities with the history $h = w_{i-2}^{i-1}$ is

$$P_{discounted}\left(w_{i-2}^{i-1}\right) = \sum_{w_i : c\left(w_{i-2}^{i}\right)>0} \left(1 - P_{Katz}\left(w_i \mid w_{i-2}^{i-1}\right)\right) = \sum_{r=1}^{r_T} \frac{\left(1-d_{r,3}\right)\cdot r \cdot n_r}{C\left(w_{i-2}^{i-1}\right)} \quad (7)$$

Applying the data in Table 1 into equation (7), we can get $P_{discounted}(h_1)$=0.3226, and $P_{discounted}(h_2)$=0.4606. This result shows that the discounted probabilities $P_{discounted}(h)$ with different histories h's do not differ too much (with the same magnitude) though the count $c(h)$ varies greatly with h (As can be seen in Table 1, $c(h_1) = 628$ and $c(h_2) = 9$). It can be explained in this way: $c(h_1)$ is much larger than $c(h_2)$, but most of the high-order n-gram units $h_1 w$'s occur no more than the discounting threshold r_T (shown in Table 1) and they will be discounted as their counterpart $h_2 w$'s do, which makes the proportion of h_1's discounted probabilities almost the same with that of h_2's. But this will lead to an unfair result that hundreds of n-grams units $h_1 w$'s with histories h_1 partition the remaining probabilities as several n-grams units $h_2 w$ do. Thus $P_{Katz}(w \mid h_1)$ may be much smaller than $P_{Katz}(w \mid h_2)$. Considering h_1 and h_2 are bi-gram units, if they are the same in the second word ($h_1 = w_{i-2}w_{i-1}, h_2 = w_{i-2}^{'}w_{i-1}$), the smoothing may becomes unreasonable. For example, if $c(w_{i-2}w_{i-1}w_i) = 10$ and $c(w_{i-2}^{'}w_{i-1}w_i) = 1$, according to the canonical Katz smoothing equations, it's easy to get

$$P_{Katz}\left(w_i \mid w_{i-2}w_{i-1}\right) = 10/628 = 0.016$$

$$P_{Katz}\left(w_i \mid w_{i-2}^{'}w_{i-1}\right) = d_1 * 1/9 = 0.033$$

Obviously,

$$P_{Katz}\left(w_i \mid w_{i-2}w_{i-1}\right) < P_{Katz}\left(w_i \mid w_{i-2}^{'}w_{i-1}\right)$$

Then if w_{i-2} and $w_{i-2}^{'}$ are homophonic words (or just similar in pronunciation), the smoothing result above will mislead the decoding in speech recognition. The whole process can be illustrated as in Figure 1.

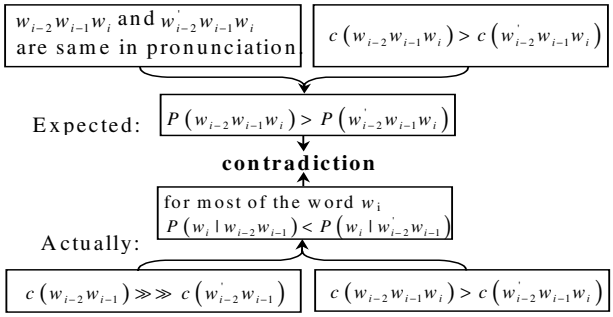

Figure 1: how the canonical Katz smoothing misleads the decoding when there are homophonic histories

In Figure 1, $P(w_{i-2}w_{i-1}w_i)$ and $P(w_{i-2}^{'}w_{i-1}w_i)$ denote the probabilities of event $e = w_{i-2}w_{i-1}w_i$ and event $e^{'} = w_{i-2}^{'}w_{i-1}w_i$ in the global event space respectively. If all the tri-gram units are considered in the global event space, it is reasonable to expect $P(e) > P(e^{'})$ since event e occurs more, that is to say, it is expected that $w_{i-2}w_{i-1}w_i$ will outgo $w_{i-2}^{'}w_{i-1}w_i$ in speech recognition decoding process. But in fact, $w_{i-2}w_{i-1}w_i$ and $w_{i-2}^{'}w_{i-1}w_i$ have the same (or just similar) pronunciation, so $P_{Katz}(w_i \mid w_{i-2}w_{i-1})$ and $P_{Katz}(w_i \mid w_{i-2}^{'}w_{i-1})$ will be compared in the speech recognition process and the unreasonable smoothing result $P_{Katz}(w_i \mid w_{i-2}w_{i-1}) < P_{Katz}(w_i \mid w_{i-2}^{'}w_{i-1})$ (this inequality is true for most of the word w_i) will mislead the decoding.

Unfortunately, the problem caused by homophonic words is very serious and the contradiction shown in Figure1 happens frequently, especially for Chinese language, so the decoding process is misled times without number.

As analyzed above, the local discounting factors $d_r(hw)$'s are unreliable and the global discounting factors d_r's are not suitable for all the n-gram units, especially for those with homophonic histories. In order to deal with these problems, the calculation of discounting factor should be improved. The simplest idea is just to discount more counts from those n-gram units with a low frequency history like h_2, that is to say, just leave a few counts for themselves, so $d_r(hw)$ should be cut down more when $c(h)$ decreases. Our proposed improving method for discounting is shown as follows

$$d_{r,n}(h)=\begin{cases}0.25d_{r,n} & 0<c(h)\le r_T\\ 0.50d_{r,n} & r_T<c(h)\le 3r_T\\ 0.75d_{r,n} & 3r_T<c(h)\le 5r_T\\ 1.00d_{r,n} & 5r_T<c(h)\le 10r_T\\ \min(1,\ 1.50d_{r,n}) & 10r_T<c(h)\end{cases} \quad (8)$$

where $d_{r,n}$ are global discounting factors of canonical Katz smoothing. The improved factors are calculated based on not only the global canonical Katz discounting factors but also the occurring counts of the histories. The back-off parameters $\alpha(\bullet)$ should be calculated based on the improved $d_{r,n}(hw)$. In addition, the subsection of equation (8) is determined by the fact that the frequencies of more 97% of bi-gram units are under $10r_T$.

4. EXPERIMENTS AND ANALYSIS

The language model used in the following experiments is a tri-gram model trained from a huge corpus containing about 200 million Chinese characters. The corpus covers the 4-year text data of *People's Daily* (from 1993 to 1994 and from 1996 to 1997) and a few texts from other newspapers. The vocabulary consists of 51,007 Chinese words.

We select three corpora of different properties. Corpus A (1,800 sentences, 23,310 characters) is from the Chinese National High-Tech Project 863, which is similar to the training corpus and with media-high perplexity; Corpus B (375 sentences, 3,466 characters) is news from the web of Hong Kong Phoenix TV (http://www.phoenixtv.com.cn), which is similar with the training corpus and with media perplexity; Corpus C (25,457 sentences, 231,606 characters) is a political book, which is quite different from the training one. Because the result of canonical Katz smoothing with local discounting factors $d_r(hw)$ is much worse than the smoothing with global discounting factors, so these results are not listed.

4.1. Perplexity comparison

	Corpus A	Corpus B	Corpus C
Canonical Smoothing with global d_r	54	253	401
Improved smoothing	46	223	360
Perplexity reduction	14.8%	11.9%	10.0%

Table 3: Tri-gram perplexity measurement

4.2. Accuracy comparison

We apply our improved model and the canonical Katz smoothing model to a Chinese Pinyin-to-Character Conversion system. The conversion error rates are listed in Table 4

	Corpus A	Corpus B	Corpus C
Canonical Smoothing with global d_r	1.53%	9.49%	15.21%
Improved smoothing	1.27%	8.11%	13.87%
CER reduction	17.0%	14.5%	8.8%

Table 4: Character error rate (CER) comparison

As can be seen in Table 3 and Table 4, the improved method achieves a surprising reduction both in perplexity and Character error rate; moreover, the method is effective for the corpora with different perplexities. The reasons have been discussed in section 3.

5. CONCLUSION

The canonical Katz smoothing in n-gram language model is detailedly analyzed in this paper. Firstly, it can be concluded that it is advisable to use the global discounting parameters instead of the local discounting parameters because the latter suffers more from data sparseness problem. Secondly, if history $h1$ and $h2$ are same or similar in pronunciation but the count of $h1$ is much larger than that of $h2$, the standard Katz smoothing is unfair to those n-gram units with history $h1$ when they are compared with other n-gram units with history $h2$. Considering the fact there are many homophonic words in Chinese vocabulary, this defect will misleads the speech recognition decoding. Our proposed method gives a modified version of the smoothing formula that makes the smoothing parameters more reasonable. Experiments show that the improving method can achieve a surprising reduction both in perplexity and character error rate.

6. REFERENCES

[1] F. Jelinek and R. L. Mercer, "Interpolated estimation of Markov source parameters from sparse data," *Pattern Recognition in Practice*, E. S. Gelsema and L. N. Kanal, Eds. Amsterdam: North-Holland, 1986

[2] S.M. Katz, "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer," *ICASSP'87,* 35(3): pp. 400~401, 1987.

[3] R. Kneser and H.Ney, "Improved Backing-off for m-gram Language Modeling," *ICASSP'95*, Vol.1, pp.49~52, 1995.

[4] I. J. Good. "The Population Frequencies of Species and the Estimation of Population Parameters," *Biometrika*, 40(3 and):pp237~264, 1953.

[5] J. Wu and F. Zheng, "On Enhancing Katz-smoothing Based on Back-off Language Model," *ICSLP'2000,* Vol. 1, pp. 198~201, 2000.