Modeling of Long Distance Context Dependency

ZHOU GuoDong
Institute for Infocomm Research
21 Heng Mui Keng Terrace
Singapore 119613
Email: zhougd@i2r.a-star.edu.sg

Abstract

Ngram models are simple in language modeling and have been successfully used in speech recognition and other tasks. However, they can only capture the short distance context dependency within an n-words window where currently the largest practical n for a natural language is three while much of the context dependency in a natural language occurs beyond a three words window. In order to incorporate this kind of long distance context dependency in the ngram model of our Mandarin speech recognition system, this paper proposes a novel MI-Ngram modeling approach. This new MI-Ngram model consists of two components: a normal ngram model and a novel MI model. The ngram model captures the short distance context dependency within an n-words window while the MI model captures the context dependency between the word pairs over a long distance by using the concept of mutual information. That is, the MI-Ngram model incorporates the word occurrences beyond the scope of the normal ngram model. It is found that MI-Ngram modeling has much better performance than the normal word ngram modeling. Experimentation shows that about 20% of errors can be corrected by using a MI-Trigram model compared with the pure word trigram model.

1 Introduction

Language modeling is the attempt to characterize, capture and exploit the regularities and constraints in a natural language and has been successfully applied to many domains. Among all the language modeling approaches, ngram models have been most widely used in speech recognition (Jelinek 1990; Gale and Church 1990; Brown et al. 1992; Yang et al. 1996) and other applications. While ngram models are simple in language modeling and have been successfully used in speech recognition and other tasks, they have obvious deficiencies. For instance, ngram models can only capture the short-distance dependency within an n-words window where currently the largest practical N for a natural language is three.

In the meantime, it is found that there always exist many preferred relationships between words. Two highly associated word pairs are “not only/but also” and “doctor/nurse”. Psychological experiments in Meyer D. et al. (1975) indicated that the human’s reaction to a highly associated word pair was stronger and faster than that to a poorly associated word pair. Such preference information is very useful for natural language processing (Church K.W. et al. 1990; Hiddle D. et al. 1993; Rosenfeld R. 1994 and Zhou G.D. et al.1998). Obviously, the preference relationships between words can expand from a short to long distance. While we can use conventional ngram models to capture the short distance dependency, the long distance dependency should also be exploited properly.

The purpose of this paper is to propose a new modeling approach to capture the context dependency over both a short distance and a long distance and apply it in Mandarin speech recognition.

This paper is organized as follows. In Section 2, we present the normal ngram modeling while a new modeling approach, named MI-ngram modeling, is proposed in Section 3. In Section 4, we will describe its use in our Mandarin speech recognition system. Finally we give a summary of this paper.

2 Ngram Modeling

Let $S = w_1^m = w_1 w_2 ... w_m$, where w_i’s are the words that make up the hypothesis, the probability of the word string, $P(S)$, can be computed by using the chain rule:

$$P(S) = P(w_1)\prod_{i=2}^{m} P(w_i \mid w_1^{i-1}) \qquad (2.1)$$

By taking a log function to both sides of Equation (2.1), we have the log probability of the word string, $\log P(S)$:

$$\log P(S) = \log P(w_1) + \sum_{i=2}^{m} \log P(w_i \mid w_1^{i-1}) \quad (2.2)$$

So, the classical task of statistical language modeling becomes how to effectively and efficiently predict the next word, given the previous words, that is to say, to estimate expressions of the form $P(w_i \mid w_1^{i-1})$. For convenience, $P(w_i \mid w_1^{i-1})$ is often written as $P(w_i \mid h)$, where $h = w_1^{i-1}$, is called history.

Traditionally, simple statistical models, known as ngram models, have been widely used in speech recognition. Within an ngram model, the probability of a word occurring next is estimated based on the $n-1$ previous words. That is to say,

$$P(w_i \mid w_1^{i-1}) \approx P(w_i \mid w_{i-n+1}^{i-1}) \qquad (2.3)$$

For example, in bigram model (n=2) the probability of a word is assumed to depend only on the previous word:

$$P(w_i \mid w_1^{i-1}) \approx P(w_i \mid w_{i-1}) \qquad (2.4)$$

And the probability $P(w_i \mid w_{i-1})$ can be estimated by using maximum likelihood estimation (MLE) principle:

$$P(w_i \mid w_{i-1}) = \frac{C(w_{i-1}w_i)}{C(w_{i-1})} \qquad (2.5)$$

Where $C(\bullet)$ represents the number of times the sequence occurs in the training text. In practice, due to the data sparseness problem, some smoothing technique (e.g. Good Turing in [Chen and Goodman 1999]) is applied to get more accurate estimation.

Obviously, an ngram model assumes that the probability of the next word w_i is independent of the word string w_1^{i-n} in the history. The difference between bigram, trigram and other ngram models is the value of n. The parameters of an ngram model are thus the probabilities: $P(w_n \mid w_1 ... w_{n-1})$ for all $w_1, w_i, ..., w_n \in V$.

Given $S = w_1 w_2 ... w_m$, an ngram model estimates the log probability of the word string, $\log P(S)$, by re-writing Equation (2.2):

$$\log P_{ngram}(S) = \log P(w_1) + \sum_{i=2}^{n-1} \log P(w_i \mid w_1^{i-1})$$
$$+ \sum_{i=n}^{m} \log P(w_i \mid w_{i-n+1}^{i-1}) \qquad (2.6)$$

where m is the string length, w_i is the i-th word in the string S.

From the ngram model as in Equation (2.3), we have:

$$\frac{P(w_1^{i-1}w_i)}{P(w_1^{i-1})} \approx \frac{P(w_{i-n+1}^{i-1}w_i)}{P(w_{i-n+1}^{i-1})}$$

$$\frac{P(w_1^{i-1}w_i)}{P(w_1^{i-1})P(w_i)} \approx \frac{P(w_{i-n+1}^{i-1}w_i)}{P(w_{i-n+1}^{i-1})P(w_i)}$$

$$\log\frac{P(w_1^{i-1}w_i)}{P(w_1^{i-1})P(w_i)} \approx \log\frac{P(w_{i-n+1}^{i-1}w_i)}{P(w_{i-n+1}^{i-1})P(w_i)} \qquad (2.7)$$

Obviously, the normal ngram model has the assumption:

$$MI(w_1^{i-1}, w_i, d=1) \approx MI(w_{i-n+1}^{i-1}, w_i, d=1) \quad (2.8)$$

where $MI(w_1^{i-1}, w_i, d=1) = \log\frac{P(w_1^{i-1}w_i)}{P(w_1^{i-1})P(w_i)}$ is the mutual information of the word string pair (w_1^{i-1}, w_i), and $MI(w_{i-n+1}^{i-1}, w_i, d=1) = \log\frac{P(w_{i-n+1}^{i-1}w_i)}{P(w_{i-n+1}^{i-1})P(w_i)}$ is the mutual information of the word string pair (w_{i-n+1}^{i-1}, w_i). d is the distance of the two word strings in the word string pair and is equal to 1 when the two word strings are adjacent.

For a pair (A, B) over a distance d where A and B are word strings, the mutual information $MI(A,B,d)$ reflects the degree of preference relationship between the two strings over a distance d. Several properties of the mutual information are apparent:

- For the same distance d, $MI(A,B,d) \neq MI(B,A,d)$.
- For different distances d_1 and d_2, $MI(A,B,d_1) \neq MI(A,B,d_2)$.

- If A and B are independent over a distance d, then $MI(A,B,d)=0$.

$MI(A,B,d)$ reflects the change of the information content when two word strings A and B are correlated. That is to say, the higher the value of $MI(A,B,d)$, the stronger affinity A and B have. Therefore, we can use the mutual information to measure the preference relationship degree of a word string pair.

Using an alternative view of equivalence, an ngram model is one that partitions the data into equivalence classes based on the last n-1 words in the history. Viewed in this way, a bigram induces a partition based on the last word in the history. A trigram model further refines this partition by considering the next-to-last word and so on.

As the word trigram model is most widely used in current research, we will mainly consider the word trigram-based model. By re-writing Equation (2.2), the word trigram model estimates the log probability of the string $\log P(S)$ as:

$$\log P_{Trigram}(S) = \log P(w_1) + \log(w_2 \mid w_1) + \sum_{i=3}^{m} \log P(w_i \mid w_{i-2}^{i-1}) \qquad (2.9)$$

3 MI-Ngram Modeling

Given $H = w_1^{i-1} = w_1 w_2 ... w_{i-1}$ and $X = w_2^{i-1} = w_2 w_3 ... w_{i-1}$, we have

$$H = w_1 X \qquad (3.1)$$

and

$$P(w_i \mid H) = P(w_i \mid w_1 X). \qquad (3.2)$$

By taking a log function to both sides of Equation (3.2), we have

$$\begin{aligned} &\log P(w_i \mid H) \\ &= \log \frac{P(Hw_i)}{P(H)} \\ &= \log P(w_i) + \log \frac{P(Hw_i)}{P(H)P(w_i)} \\ &= \log P(w_i) + MI(w_i, H, 1) \qquad (3.3) \end{aligned}$$

Now we assume

$$\begin{aligned} MI(H, w_i, d=1) &= MI(X, w_i, d=1) \\ &\quad + MI(w_1, w_i, d=i) \qquad (3.4) \end{aligned}$$

where $H = w_1^{i-1}$, $X = w_2^{i-1}$ and $i > N$. That is to say, the mutual information of the next word with the history is assumed equal to the summation of that of the next word with the first word in the history and that of the next word with the rest word string in the history. Then we can re-write Equation (3.3) by using Equation (3.4),

$$\begin{aligned} &\log P(w_i \mid H) \\ &= \log P(w_i) + MI(H, w_i, 1) \\ &= \log P(w_i) + MI(X, w_i, 1) + MI(w_1, w_i, i) \\ &= \log P(w_i) + \log \frac{P(w_i X)}{P(w_i)P(X)} + MI(w_1, w_i, i) \\ &= \log \frac{P(w_i X)}{P(X)} + MI(w_1, w_i, i) \\ &= \log P(w_i \mid X) + MI(w_1, w_i, i) \qquad (3.5) \end{aligned}$$

That is, we have

$$\log P(w_i \mid w_1^{i-1}) = \log P(w_i \mid w_2^{i-1}) + MI(w_1, w_i, i) \qquad (3.6)$$

By applying Equation (3.6) repeatedly, we have a modified estimation of the conditional probability:

$$\begin{aligned} &\log P(w_i \mid w_1^{i-1}) \\ &= \log P(w_i \mid w_2^{i-1}) + MI(w_1, w_i, i) \\ &= \log P(w_i \mid w_3^{i-1}) \\ &+ MI(w_2, w_i, i-1) + MI(w_1, w_i, i) \\ &\cdot\cdot \;\; \cdot\cdot \;\; \cdot\cdot \\ &= \log P(w_i \mid w_{i-n+1}^{i-1}) + \sum_{k=1}^{k=i-n} MI(w_k, w_i, i-k+1) \qquad (3.7) \end{aligned}$$

Obviously, the first item in equation (3.7) contributes to the log probability of the normal word ngram within an N-words window while the second item is the mutual information which contributes to the long distance context dependency of the next word w_i with the previous words $w_j (1 \le j \le i-N, i > N)$ outside the n-words window of the normal word ngram model.

By using Equation (3.7) iteratively, Equation (2.2) can be re-written as:

$$\log P(S)$$

$$= \log P(w_1) + \sum_{i=2}^{m} \log P(w_i \mid w_1^{i-1})$$

$$= \log P(w_1) + \sum_{i=2}^{i=n} \log(w_i \mid w_1^{i-1})$$

$$+ \log P(w_{n+1} \mid w_1^n) + \sum_{i=n+2}^{m} \log P(w_i \mid w_1^{i-1})$$

$$= \log P(w_1) + \sum_{i=2}^{i=n} \log(w_i \mid w_1^{i-1})$$

$$+ \log P(w_{n+1} \mid w_2^n)$$

$$+ MI(w_{n+1}, w_1, n+1) + \sum_{i=n+2}^{m} \log P(w_i \mid w_1^{i-1})$$

$$\cdots \cdots \cdots$$

$$= \log P(w_1) + \sum_{i=2}^{i=n} \log(w_i \mid w_1^{i-1})$$

$$+ \sum_{i=n+1}^{m} \log P(w_i \mid w_{i-2}^{i-1})$$

$$+ \sum_{i=n+1}^{m} \sum_{k=1}^{k=i-n} MI(w_k, w_i, i-k+1) \qquad (3.8)$$

From Equation (3.8), we can see that the first three items are the values computed by the normal word trigram model as shown in Equation (2.9) and the forth item $\sum_{i=n+1}^{m} \sum_{k=1}^{k=i-n} MI(w_k, w_i, i-k+1)$ contributes to summation of the mutual information of the next word with the words in the history w_1^{i-n}. Therefore, we call Equation (3.8) as a MI-Ngram model and rewrite it as:

$$\log P_{MI-Ngram}(S)$$

$$= \log P_{Ngram}(S)$$

$$+ \sum_{i=n+1}^{m} \sum_{k=1}^{k=i-n} MI(w_k, w_i, i-k+1) \qquad (3.9)$$

As a special case of N=3, the MI-Trigram model estimate the log probability of the string as follows:

$$\log P_{MI-Trigram}(S)$$

$$= \log P_{Trigram}(S)$$

$$+ \sum_{i=4}^{m} \sum_{k=1}^{k=i-3} MI(w_k, w_i, i-k+1) \qquad (3.10)$$

Compared with the normal word ngram model, the novel MI-Ngram model also incorporates the long distance context dependency by computing the mutual information of the distance dependent word pairs. That is, the MI-Ngram model incorporates the word occurrences beyond the scope of the normal ngram model.

Since the number of possible distance-dependent word pairs may be very huge, it is impossible for the MI-Ngram model to incorporate all the possible distance-dependent word pairs. Therefore, for the MI-Ngram model to be practically useful, how to select a reasonable number of word pairs becomes most important. Here two approaches are used (Zhou G.D., et al 1998):

One approach is to restrict the window size of possible word pairs by computing and comparing the conditional perplexities (Shannon C.E. 1951) of the long distance word bigram models for different distances. Conditional perplexity is a measure of the average number of possible choices there are for a conditional distribution. The conditional perplexity of a conditional distribution with the conditional entropy $H(Y|X)$ is defined to be $2^{H(Y|X)}$. Given two random variables X and Y, a conditional probability mass function $P_{Y|X}(y|x)$, and a marginal probability mass function $P_Y(y)$, the conditional entropy of Y given X, $H(Y|X)$, is defined as:

$$H(Y \mid X) = -\sum_{x \in X} \sum_{y \in Y} P_{X,Y}(x, y) \log_2 P_{Y|X}(y \mid x) \qquad (3.11)$$

For a large enough corpus, the conditional perplexity is usually an indication of the amount of information conveyed by the model: the lower the conditional perplexity, the more information it conveys and thus a better model. This is because the model captures as much as it can of that information, and whatever uncertainty remains shows up in the conditional perplexity. Here, the corpus is the XinHua corpus, which has about 57M(million) characters or 29M words. For all the experiments, 80% of the corpus is used for training while the remaining 20% is used for testing.

Table 1 shows that the conditional perplexity is lowest for d = 1 and increases significantly as we move through d = 2, 3, 4, 5 and 6. For d = 7, 8, 9, the conditional perplexity increases slightly while further increasing d almost does not increase the conditional perplexity. This suggests that significant information exists only in the last 6 words of the history. In this paper, we restrict the maximum window size to 10.

Table 1: Conditional perplexities of the long-distance word bigram models for different distances

Distance	Perplexity	Distance	Perplexity
1	230	7	1479
2	575	8	1531
3	966	9	1580
4	1157	10	1599
5	1307	11	1611
6	1410	20	1647

Another approach is to adapt average mutual information to select a reasonable number of distance-dependent word pairs:

$$\begin{aligned} AMI(A;B) = {} & P(A,B)\log\frac{P(AB)}{P(A)P(B)} \\ & + P(A,\overline{B})\log\frac{P(A\overline{B})}{P(A)P(\overline{B})} \\ & + P(\overline{A},B)\log\frac{P(\overline{A}B)}{P(\overline{A})P(B)} \\ & + P(\overline{A},\overline{B})\log\frac{P(\overline{A}\overline{B})}{P(\overline{A})P(\overline{B})} \end{aligned} \tag{3.12}$$

Obviously, Equation (3.12) takes the joint probability into consideration. That is, those frequently occurring word pairs are more important and have much more potential to be incorporated into the MI-Ngram model than less frequently occurring word pairs.

4 Experimentation

We have evaluated the new MI-Ngram model in an experimental speaker-dependent continuous Mandarin speech recognition system (Zhou G.D. et al 1999). For base syllable recognition, 14 cepstral and 14 delta-cepstral coefficients, energy(normalized) and delta-energy are used as feature parameters to form a feature vector with dimension 30, while for tone recognition, the pitch period and the energy together with their first order and second order delta coefficients are used to form a feature vector with dimension 6. All the acoustic units are modeled by semi-continuous HMMs (Rabiner 1993). For base syllable recognition, 138 HMMs are used to model 100 context-dependent INITIALs and 38 context-independent FINALs while 5 HMMs are used to model five different tones in Mandarin Chinese. 5,000 short sentences are used for training and another 600 sentences (6102 Chinese characters) are used for testing. All the training and testing data are recorded by one same speaker in an office-like laboratory environment with a sampling frequency of 16KHZ.

As a reference, the base syllable recognition rate and the tone recognition rate are shown in Table 2 and Table 3, respectively. As the word trigram model is most widely used in current research, all the experiments have been done using a MI-Trigram model which is trained on the XINHUA news corpus of 29 million words(automatically segmented) while the lexicon contains about 28000 words. As a result, the perplexities and Chinese character recognition rates of different MI-Trigram models with the same window size of 10 and different numbers of distance-dependent word pairs are shown in Table 4.

Table 2: The top-n recognition rates of base syllables

Top-N Base Syllables	1	5	10	15	20
Recognition Rate of Base Syllables	88.2	97.6	99.2	99.5	99.8

Table 3: The recognition rates of the tones

	tone 1	tone 2	tone 3	tone 4	tone 5
tone 1	90.4	0.8	0.6	0.8	7.4
tone 2	8.3	81.1	5.4	0.2	4.9
tone 3	5.0	20.9	43.0	29.1	2.0
tone 4	4.3	0.2	1.8	93.5	0.2
tone 5	24.1	8.6	0.9	8.2	58.2

Table 4: The effect of different numbers of word pairs in the MI-Trigram models with the same window size 10 on the Chinese character recognition rates

Number of word pairs	Perplexity	Recognition Rate
0	204	90.5
100,000	196	91.2
200,000	189	91.7
400,000	183	92.1
600,000	179	92.3
800,000	175	92.4
1,000,000	172	92.5
1,500,000	171	92.5
2,000,000	170	92.6
2,500,000	170	92.5
3,000,000	168	92.6
3,500,000	169	92.6
4,000,000	168	92.7

Table 4 shows that the perplexity and the recognition rate rise quickly as the number of the long distance-dependent word pairs in the MI-Trigram model increase from 0 to 800,000, and then rise slowly. This suggests that the best 800,000 word pairs carry most of the long distance context dependency and should be included in the MI-Ngram model. It also shows that the recognition rate of the MI-Trigram model with 800,000 word pairs is 1.9% higher than the pure word trigram model (the MI-Trigram model with 0 long distance-dependent word pairs). That is to say, about 20% of errors can be corrected by incorporating only 800,000 word pairs to the MI-Trigram model compared with the pure word trigram model.

It is clear that MI-Ngram modeling has much better performance than normal word ngram modeling. One advantage of MI-Ngram modeling is that its number of parameters is just a little more than that of word ngram modeling. Another advantage of MI-Ngram modeling is that the number of the word pairs can be reasonable in size without losing too much of its modeling power. Compared to ngram modeling, MI-Ngram modeling also captures the long distance dependency of word pairs using the concept of mutual information.

4. CONCLUSION

This paper proposes a novel MI-Ngram modeling approach to capture the context dependency over both a short distance and a long distance. This is done by incorporating long distance-dependent word pairs into normal ngram modeling by using the concept of mutual information. It is found that MI-Ngram modeling has much better performance than word ngram modeling.

REFERENCE

Brown P.F. et al. (1992). Class-based Ngram models of natural language. *Computational Linguistics* 18(4), 467-479.

Chen S.F. and Goodman J. (1999). An empirical study of smoothing technique for language modeling. *Computer, Speech and Language.* 13(5). pp.359-394.

Church K.W. et al. (1991). Enhanced good Turing and Cat-Cal: two new methods for estimating probabilities of English bigrams. *Computer, Speech and Language* 5(1), 19-54.

Gale W.A. & Church K.W. (1990). Poor estimates of context are worse than none. *Proceedings of DARPA Speech and Natural Language Workshop*, Hidden Valley, Pennsylvania, pp. 293-295.

Hindle D. et al. (1993). Structural ambiguity and lexical relations. *Computational Linguistics* 19(1), 103-120.

Jelinek F. (1990). Self-organized language modeling for speech recognition. In *Readings in Speech Recognition.* Edited by Waibel A. and Lee K.F. Morgan Kaufman. San Mateo. CA. pp.450-506.

Meyer D. et al. (1975). Loci of contextual effects on visual word recognition. In *Attention and Performance V*, edited by P.Rabbitt and S.Dornie. pp. 98-116. Acdemic Press.

Rabiner L.R. et al. (1993). *Foundamentals to Speech Recognition*. Prentice Hall.

Rosenfeld R. (1994). Adaptive statistical language modeling: A Maximum Entropy Approach. *Ph.D. Thesis*, Carneige Mellon University.

Shannon C.E. (1951). Prediction and entropy of printed English. *Bell Systems Technical Journal* 30, 50-64.

Yang Y.J. et al. (1996). Adaptive linguistic decoding system for Mandarin speech recognition applications. *Computer Processing of Chinese & Oriental Lang*uages 10(2), 211-224.

Zhou G.D. and Lua K.T. (1998). Word association and MI-Trigger-based language modeling, *COLING-ACL'98*. Montreal Canada, 8-14 August.

Zhou G.D. and Lua KimTeng (1999). Interpolation of N-gram and MI-based Trigger Pair Language Modeling in Mandarin Speech Recognition, *Computer, Speech and Language*, Vol. 13, No. 2, pp.123-135.