Modeling of Long Distance Context Dependency in Chinese

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

Abstract

Ngram modeling is simple in language modeling and has been widely used in many applications. However, it can only capture the short distance context dependency within an N-word window where the largest practical N for natural language is three. In the meantime, much of context dependency in natural language occurs beyond a three-word window. In order to incorporate this kind of long distance context dependency, this paper proposes a new MI-Ngram modeling approach. The MI-Ngram model consists of two components: an ngram model and an MI model. The ngram model captures the short distance context dependency within an N-word window while the MI model captures the long distance context dependency between the word pairs beyond the N-word window by using the concept of mutual information. It is found that MI-Ngram modeling has much better performance than ngram modeling. Evaluation on the XINHUA new corpus of 29 million words shows that inclusion of the best 1,600,000 word pairs decreases the perplexity of the MI-Trigram model by 20 percent compared with the trigram model. In the meanwhile, evaluation on Chinese word segmentation shows that about 35 percent of errors can be corrected by using the MI-Trigram model compared with the trigram model.

1 Introduction

Language modeling is the attempt to characterize, capture and exploit the regularities and constraints in natural language. Among various language modeling approaches, ngram modeling has been widely used in many applications, such as speech recognition, machine translation (Katz 1987; Jelinek 1989; Gale and Church 1990; Brown et al. 1992; Yang et al. 1996; Bai et al 1998; Zhou et al 1999; Rosenfeld 2000; Gao et al 2002). Although ngram modeling is simple in nature and easy to use, it has obvious deficiencies. For instance, ngram modeling can only capture the short distance context dependency within an N-word window where currently the largest practical N for 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 不仅/而且 ("not only/but also") and 医生/护士 ("doctor/nurse"). Psychological experiments in Meyer 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 et al. 1990; Hiddle et al. 1993; Rosenfeld 1994; Zhou et al.1998; Zhou et al 1999). Obviously, the preference relationships between words can expand from a short to long distance. While we can use traditional ngram modeling to capture the short distance context dependency, the long distance context dependency should also be exploited properly.

The purpose of this paper is to propose a new MI-Ngram modeling approach to capture the context dependency over both a short distance and a long distance. Experimentation shows that this new MI-Ngram modeling approach can significantly decrease the perplexity of the new MI-Ngram model compared with traditional ngram model. In the meantime, evaluation on Chinese word segmentation shows that this new approach can significantly reduce the error rate.

This paper is organized as follows. In section 2, we describe the traditional ngram modeling approach and discuss its main property. In section 3, we propose the new MI-Ngram modeling approach to capture context dependency over both a short distance and a long distance. In section 4, we measure the MI-Ngram modeling approach and evaluate its application in Chinese word segmentation. Finally we give a summary of this paper in section 5.

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:

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

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

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

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.

Ngram modeling has been widely used in estimating $P(w_i \mid h)$. Within an ngram model, the probability of a word occurring next is estimated based on the $n-1$ previous words. That is to say,

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

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

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

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

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

Where $C(\bullet)$ represents the number of times the sequence occurs in the training data. In practice, due to the data sparseness problem, some smoothing techniques, such as linear interpolation (Jelinek 1989; Chen and Goodman 1999) and back-off modeling (Katz 1987), are applied.

Obviously, an ngram model assumes that the probability of the next word w_i is independent of 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 are thus the probabilities:

$$P(w_n \mid w_1 ... w_{n-1}) \quad \text{For all } w_1, w_w, ..., w_n \in V .$$

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

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

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

From equation (2.3) $P(w_i \mid w_1^{i-1}) \approx P(w_i \mid w_{i-n+1}^{i-1})$, we have:

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

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

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

Obviously, we can get

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

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 between 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 between the word string pair (w_{i-n+1}^{i-1}, w_i). d is the distance of two word strings in the word string pair and is equal to 1 when the two word strings are adjacent.

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

From the view of mutual information, an ngram model assumes the mutual information independency between (w_1^{i-n}, w_i). 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.

As trigram model is most widely used in current research, we will mainly consider the trigram-based model. By re-writing equation (2.6), the trigram model estimates the log probability of the string $P(S)$ as:

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

3 MI-Ngram Modeling

Given history $H = w_1^{i-1} = w_1 w_2 ... w_{i-1}$, we can assume $X = w_2^{i-1} = w_2 w_3 ... w_{i-1}$. Then we have

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

and

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

Since

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

Here we assume

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

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.

We can re-writing equation (3.3) by using equation (3.4):

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

$$= \log P(w_i) + \log \frac{P(w_i X)}{P(w_i)P(X)} + MI(w_1, w_i, i)$$

$$= \log \frac{P(w_i X)}{P(X)} + MI(w_1, w_i, i)$$

$$= \log P(w_i \mid X) + MI(w_1, w_i, i) \quad (3.5)$$

Then we have

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

By applying equation (3.6) repeatedly, we have:

$$\log P(w_i \mid w_1^{i-1})$$
$$= \log P(w_i \mid w_2^{i-1}) + MI(w_1, w_i, i)$$
$$= \log P(w_i \mid w_3^{i-1})$$
$$+ MI(w_2, w_i, i-1) + MI(w_1, w_i, i)$$
$$\cdots\cdots$$
$$= \log P(w_i \mid w_{i-n+1}^{i-1})$$
$$+ \sum_{k=1}^{k=i-n} MI(w_k, w_i, i-k+1) \quad (3.7)$$

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

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

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

$$= \log P(w_1) + \sum_{i=2}^{i=n} \log(w_i \mid w_1^{i-1})$$
$$+ \log P(w_{n+1} \mid w_1^n) + \sum_{i=n+2}^{m} \log P(w_i \mid w_1^{i-1})$$
$$= \log P(w_1) + \sum_{i=2}^{i=n} \log(w_i \mid w_1^{i-1})$$
$$+ \log P(w_{n+1} \mid w_2^n)$$
$$+ MI(w_{n+1}, w_1, n+1) + \sum_{i=n+2}^{m} \log P(w_i \mid w_1^{i-1})$$
$$\cdots\cdots$$
$$= \log P(w_1) + \sum_{i=2}^{i=n} \log(w_i \mid w_1^{i-1})$$
$$+ \sum_{i=n+1}^{m} \log P(w_i \mid w_{i-2}^{i-1})$$
$$+ \sum_{i=n+1}^{m} \sum_{k=1}^{k=i-n} MI(w_k, w_i, i-k+1) \quad (3.8)$$

In equation (3.8), the first three items are the values computed by the 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 over the long distance outside the N-word window. That is, the new model as shown in equation (3.8) consists of two components: an ngram model and an MI model. Therefore, we call equation (3.8) as an MI-Ngram model and equation (3.8) can be re-written as:

$$\log P_{MI-Ngram}(S)$$
$$= \log P_{Ngram}(S) \quad (3.9)$$
$$+ \sum_{i=n+1}^{m} \sum_{k=1}^{k=i-n} MI(w_k, w_i, i-k+1)$$

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

$$\log P_{MI-Trigram}(S)$$
$$= \log P_{Trigram}(S) \quad (3.10)$$
$$+ \sum_{i=4}^{m} \sum_{k=1}^{k=i-3} MI(w_k, w_i, i-k+1)$$

Compared with traditional ngram modeling, MI-Ngram modeling incorporates the long distance context dependency by computing mutual information of the long distance dependent word

pairs. Since the number of possible long distance dependent word pairs may be very huge, it is impossible for MI-Ngram modeling to incorporate all of them. Therefore, for MI-Ngram modeling to be practically useful, how to select a reasonable number of word pairs becomes very important. Here two approaches are used (Zhou et al 1998 and 1999). One is to restrict the window size of possible word pairs by computing and comparing the perplexities[1] (Shannon C.E. 1951) of various long distance bigram models for different distances. It is found that the bigram perplexities for different distances outside the 10-word window become stable. Therefore, we only consider MI-Ngram modeling with a window size of 10 words. Another is to adapt average mutual information to select a reasonable number of long distance dependent word pairs. Given distance d and two words A and B, its average mutual information is computed as:

$$
\begin{aligned}
&AMI(A,B,d) \\
&= P(A,B,d)\log\frac{P(A,B,d)}{P(A)P(B)} \\
&+ P(A,\overline{B},d)\log\frac{P(A,\overline{B},d)}{P(A)P(\overline{B})} \\
&+ P(\overline{A},B,d)\log\frac{P(\overline{A},B,d)}{P(\overline{A})P(B)} \\
&+ P(\overline{A},\overline{B},d)\log\frac{P(\overline{A},\overline{B},d)}{P(\overline{A})P(\overline{B})}
\end{aligned}
\qquad (3.11)
$$

Compared with mutual information, average mutual information takes joint probabilities into consideration. In this way, average mutual information prefers frequently occurred word pairs. In our paper, different numbers of long distance dependent word pairs will be considered in MI-Ngram modeling within a window size of 10 words to evaluate the effect of different MI model size.

4 Experimentation

As trigram modeling is most widely used in current research, only MI-Trigram modeling is studied here. Furthermore, in order to demonstrate the effect of different numbers of word pairs in MI-Trigram modeling, various MI-Trigram models with different numbers of word pairs and the same window size of 10 words are trained on the XINHUA news corpus of 29 million words while the lexicon contains about 56,000 words. Finally, various MI-Trigram models are tested on the same task of Chinese word segmentation using the Chinese tag bank PFR1.0[2] of 3.69M Chinese characters (1.12M Chinese Words).

Table 1 shows the perplexities of various MI-Trigram models and their performances on Chinese word segmentation. Here, the precision (P) measures the number of correct words in the answer file over the total number of words in the answer file and the recall (R) measures the number of correct words in the answer file over the total number of

[1] Perplexity is a measure of the average number of possible choices there are for a random variable. The perplexity PP of a random variable X with entropy $H(X)$ is defined as:

$$PP(X) = 2^{H(X)}$$

Entropy is a measure of uncertainty about a random variable. If a random variable X occurs with a probability distribution $P(x)$, then the entropy $H(X)$ of that event is defined as:

$$H(X) = -\sum_{x\in X} P(x)\log_2 P(x)$$

Since $x\log_2 x \to 0$ as $x \to 0$, it is conventional to use the relation $0\log_2 0 = 0$ when computing entropy.

The units of entropy are bits of information. This is because the entropy of a random variable corresponds to the average number of bits per event needed to encode a typical sequence of event samples from that random variable' s distribution.

[2] PFR1.0 is developed by Institute of Computational Linguistics at Beijing Univ. Here, only the word segmentation annotation is used.

words in the key file. F-measure is the weighted harmonic mean of precision and recall: $F = \frac{(\beta^2 + 1)RP}{\beta^2 R + P}$ with $\beta^2 = 1$.

Table 1 shows that

- The perplexity and the F-measure rise quickly as the number of word pairs in MI-Trigram modeling increases from 0 to 1,600,000 and then rise slowly. Therefore, the best 1,600,000 word pairs should at least be included.
- Inclusion of the best 1,600,000 word pairs decreases the perplexity of MI-Trigram modeling by about 20 percent compared with the pure trigram model.
- The performance of Chinese word segmentation using the MI-Trigram model with 1,600,000 word pairs is 0.8 percent higher than using the pure trigram model (MI-Trigram with 0 word pairs). That is to say, about 35 percent of errors can be corrected by incorporating only 1,600,000 word pairs to the MI-Trigram model compared with the pure trigram model.
- For Chinese word segmentation task, recalls are about 0.7 percent higher than precisions. The main reason may be the existence of unknown words. In our experimentation, unknown words are segmented into individual Chinese characters. This makes the number of segmented words in the answer file higher than that in the key file.

It is clear that MI-Ngram modeling has much better performance than ngram modeling. One advantage of MI-Ngram modeling is that its number of parameters is just a little more than that of 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 context dependency of word pairs using the concept of mutual information.

Table 1: The effect of different numbers of word pairs in MI-Trigram modeling with the same window size of 10 words on Chinese word segmentation

Number of word pairs	Perplexity	Precision	Recall	F-measure
0	316	97.5	98.2	97.8
100,000	295	97.9	98.4	98.1
200,000	281	98.1	98.6	98.3
400,000	269	98.2	98.7	98.4
800,000	259	98.2	98.8	98.5
1,600,000	250	98.4	98.8	98.6
3,200,000	245	98.3	98.9	98.6
6,400,000	242	98.4	98.9	98.6

6 Conclusion

This paper proposes a new 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 traditional ngram model by using the concept of mutual information. It is found that MI-Ngram modeling has much better performance than ngram modeling.

Future works include the explorations of the new MI-Trigram modeling approach in other applications, such as Mandarin speech recognition and PINYIN to Chinese character conversion.

References

Bai S.H., Li H.Z., Lin Z.M. and Yuan B.S. 1989. Building class-based language models with contextual statistics. *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP'1998). pages173-176.*

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). 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. and Church K.W. 1990. Poor estimates of context are worse than none. *Proceedings of DARPA Speech and Natural Language Workshop*, Hidden Valley, Pennsylvania, pages293-295.

Gao J.F., Goodman J.T., Cao G.H. and Li H. 2002. Exploring asymmetric clustering for statistical language modelling. *Proceedings of the Fortieth Annual Meeting of the Association for Computational Linguistics (ACL'2002).* Philadelphia. pages183-190.

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

Jelinek F. 1989. 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. pages450-506.

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.* 35. 400-401.

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. pages98-116. Acdemic Press.

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

Rosenfeld R. 2000. Two decades of language modelling: where do we go from here. *Proceedings of IEEE.* 88:1270-1278. August.

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 GuoDong and Lua Kim Teng, 1998. Word Association and MI-Trigger-based Language Modeling. *Proceedings of the Thirtieth-sixth Annual Meeting of the Association for Computational Linguistics and the Seventeenth International Conference on Computational Linguistics (COLING-ACL'1998)*. Montreal, Canada. pages10-14. August.

Zhou GuoDong and Lua KimTeng. 1999. Interpolation of N-gram and MI-based Trigger Pair Language Modeling in Mandarin Speech Recognition, *Computer, Speech and Language*, 13(2), 123-135.