An Effective Combination of Different Order N-grams

Sen Zhang, Na Dong
Speech Group, INRIA-LORIA B.P.101 54602 Villers les Nancy, France
zhangsen@yahoo.com

Abstract

In this paper an approach is proposed to combine different order N-grams based on the discriminative estimation criterion, on which the parameters of n-gram can be optimized. To raise the power of modeling language information, we propose several schemes to combine conventional different order n-gram language model. We employ Newton Gradient method to estimate the assumption probabilities and then test the optimally selected language model. We conduct experiments on the platform of conversion from Chinese pinyin to Chinese character. The experimental results show that the memory capacity of language model can be remarkably lowered with little loss of accuracy.

1. Introduction

In Chinese natural language processing domain, the parameters of n-gram can be estimated by calculating the frequency of word pair in text corpus and then normalizing the frequency. This conventional language model cannot satisfy our requirements since it is dependent of discriminative capability. We propose discriminative estimation approach, which can directly relate the estimation of n-gram parameters to its discriminative capability. We optimize the parameters of n-gram on the criterion of discriminative estimation by using Newton Gradient method.

When we establish N-gram we artificially introduce an assumption over the relationship among adjacent words. Uni-gram is based upon the assumption that all words appear in the corpus independently. Bi-gram assumes that only contiguous words correlate with each other and tri-gram puts a constraint on the language information that one word can be predicted only by its two predecessor words. Some words are free of context and some depend on short or long history information under some circumstances. In this sense single N-gram could model the language phenomena with some compromise. This paper addresses the impact of the different assumption from different order n-gram on the performance of language model and proposed the combination and optimal selection of different n-gram to battle with artificial assumption and possible data sparsity problem.

In the following sections, we first bring up with the discriminative estimation criterion. Next we describe the assumption from N-gram. In the following section we introduce the scheme for combination of different order N-gram. In the next section, an approach to optimal selection of different order language model is proposed. At last we report the experimental results on the platform of conversion from Chinese pinyin to Chinese character.

2 Discriminative Estimation

In natural language processing, statistical language model has been proved to be a successful method and great efforts have been taken in building n-gram language model. (R. Isotani, 1994) reports the result of the research on stochastic language model with local and global language information.

N-gram can be looked as Markov chain model, so maximum likelihood estimation can be used in n-gram similar to the estimation of HMM parameters (L. Bahl, 1983). At the same time, the estimation of n-gram parameters is not surely relative to discriminative capability.

Discriminative capability means the power that n-gram gets rise to correct results with higher score or probability contrast to wrong results. In speech recognition, discriminative training of HMM has been proposed from the viewpoint of pattern recognition (P. Chang, 1993; W. Chou, 1995). In this training approach, high recognition rate for training set is the motivation. The complicated objective function results in the complex formula and insupportable computational cost for parameter estimation.

To simplify the above estimation criterion, we introduce the following objective function.

$$MAX\left\{\sum \frac{p\left(w_1^N|\theta\right)}{p\left(w_1^N|\lambda\right)}\right\}$$

where è and ë denote the correct word string and all possible word string, respectively.
We can estimate the parameters of n-gram using Newton gradient method.
For the parameters in the numerator,

$$p(\omega_i) = p(\omega_i) + \alpha \cdot \frac{1}{\sum_j P\left(w_1^N|\theta_j\right)} \frac{\partial p\left(w_1^N|\theta\right)}{\partial p(\omega_i)} \qquad (1)$$

For the parameters in the denominator,

$$p(\omega_i) = p(\omega_i) - \alpha \cdot \sum_k \frac{p\left(w_1^N|\theta\right)}{\left(\sum_j p\left(w_1^N|\theta_j\right)\right)^2} \frac{\partial p\left(w_1^N|\theta_k\right)}{\partial p(\omega_i)} \qquad (2)$$

where á denotes the length of step.

The above formula describe that discriminative estimation increases the value of the parameters in the numerator, that is, the correct word string, and decreases the value of the parameters in the denominator.

3 Assumption of N-Gram Model

As we describe above, we introduce to N-gram language model an assumption about the mutual information indicated in contiguous words. We assume over uni-gram that adjacent words are independent of each other and a word string is made up of words without any mutual information. On the confidence of this assumption and the probability of words conditioned over this assumption, the

conventional n-gram language model can be evaluated by the following conditional probability:

$$p(w_1^N|\omega_1)=\prod p(w_i|\omega_1)$$

where ω_1 denotes the assumption that words are independent.

For bi-gram it is assumed that two contiguous words can imply some language information, which can be modeled by the conditional probability that one word is followed by a specific word as follows:

$$p(w_1^N|\omega_2)=\prod p(w_{i+1}|w_i,\omega_2)$$

where ω_2 denotes the assumption that only two adjacent words are dependent.

The assumption over tri-gram is similar to that over bi-gram, but three adjacent words are taken into consideration for the conditional probability.

$$p(w_1^N|\omega_3)=\prod p(w_{i+1}|w_i,w_{i-1},\omega_3)$$

where ω_3 means the assumption that one word is relative to its two predecessor words.

In speech dictation, we can generally obtain N-best candidates by using bi-gram and then tri-gram. We notice that tri-gram cannot handle beyond N-best candidates, in which correct results may be included. To make full use of the power of different order N-grams, it is important to combine different order N-grams together instead of two-pass or multi-pass.

4 Combination of Different N-Gram Model

We can obtain the probabilities of a word string conditioned on different assumptions using traditional n-gram language model. In order to calculate the probability that a word string is generated regardless of any assumption, we can introduce the probability that one assumption is true and merge it together with the probability of the word string conditioned on different assumptions.

We employ different n-gram to analyze one sentence and then combine the analysis result together. We apply single an assumption to describe the relationship among words for each n-gram. From this viewpoint we can address it as sentence-level analysis.

$$p(w_1^n)=\sum p(w_1^n|\omega_i)p(\omega_i) \qquad (3)$$

where

$$\omega_i, p(w_1^n|\omega_i) and \quad p(\omega_i)$$

denote assumption i, conditional probability of the sentence over assumption i and probability of assumption i, respectively. Here $p(\omega_i)$ means that the assumption ω_i is true.

In practice, we can apply the assumption from different n-gram to word level analysis. When we process the next word following a sub-string, we can view this word as the production of different n-gram and merge the result of different n-gram together. To be more detailed,

$$p(w_1^n, w_{n+1}) = \sum p(w_1^n) p(w_{n+1} | w_n \cdots w_{n-m}, \omega_i) p(\omega_i) \quad (4)$$

where

$$w_1^n, w_{n+1}, \omega_i, p(w_1^n), p(w_{n+1} | w_n \cdots w_{n-m}), p(\omega_i)$$

mean the word string containing n words, (n+1)-th word, assumption i, probability of word string, conventional n-gram and the probability of assumption i, respectively.

In formula 4 we take the probability of assumption into account independent of specific words. Actually whether an assumption is true or not strongly depends on context information. Hereby we introduce word-specific assumption probability to formula 4 and calculate the probability of word sub-string using the following formula 5.

$$p(w_1^n, w_{n+1}) = \sum p(w_1^n) p(w_{n+1} | w_n \cdots w_{n-m}, \omega_i) p(\omega_i | w_n \cdots w_{n-m}) \quad (5)$$

where

$$p(\omega_i | w_n \cdots w_{n-m})$$

is the probability of assumption ω_i.

In order to reduce the computational complexity over the probability of a sentence, we propose that the probability of sub-string can be merged regardless of the history information and used assumptions as indicated in the following expression.

$$p(W_1^n, w_{n+1}) = p(W_1^{n+1}) = \sum p(W_1^n) p(w_{n+1} | w_n \cdots w_{n-m}, \omega_i) p(\omega_i \mid w_{n+1}, w_n \cdots w_{n-m}) \quad (6)$$

5 Optimal Selection of N-grams

To reduce the memory occupancy, we design some schemes to optimally select elements from different order N-grams.

1. If the probability of one assumption over some word pair is close to 1.0, then select this element.
2. If the probabilities of several assumptions over some word pair are comparable, then choose the element with greatest conditional probability. If the conditional probabilities are very close, then choose the simplest elements.
3. If the difference on the assumption probabilities is remarkable, but none of them is close to 1.0, we choose the elements with bigger conditional probability and assumption probability.

6 Experiments

We conduct several experiments using the tagged text corpus by Peking University which contains one million characters and covers political materials, novels, technical papers, grammatical papers and so on. N-gram is built up on the basis of this corpus and sparse data problem is very serious due to the limitation of corpus. We select 200 sentences for evaluation, which can be successfully processed by N-gram. The total number of characters for test is 3,335. Chinese pinyin of a sentence is character flow without segmentation and tone information. We use dynamic programming (DP) to achieve the

conversion between Chinese pinyin and Chinese character. We use uni-gram and bi-gram to test the availability of the proposed approaches.

The following experimental results show that the combination of different N-gram can raise the transform rate remarkably.

Table 1. Conversion result

N_gram	D	S	I
Uni-gram	2	388	0
Bi-gram	0	29	0
Formula 3	0	27	0
Formula 4	0	24	0
Formula 5	0	13	0
Hybrid	0	15	0

D-Deletion, S-Substitution, I-Insertion

Fig. 1. Correctness rate and error reduction rate

From Fig. 1 and table 1, we can obtain that the error rate is reduced by 6.9% when we adopt probability of assumptions in sentence level. The word-specific assumption probability gives an error rate reduction of 55.2% while the word-independent assumption probability decreases the error rate by 17.2%. If we build the hybrid language model by optimally selecting some elements from different N-grams, the accuracy is lowered from 99.6% to 99.5%, but the memory capacity for new language model is decreased remarkably by 30%.

7 Conclusion

This paper reports the result of research on the combination of different order N-gram for conversion from Chinese pinyin to Chinese character. We bring up with three schemes to achieve the combination by introducing the assumption probability, which can be in sentence level, word level and word-specific, respectively. The experimental results show that the error rate of conversion could be decreased remarkably.

8 Reference

A maximum likelihood approach to continuous speech recognition. 1983 *IEEE trans. on PAMI, PAMI-5(20, pp.179,).* L. Bahl, R. Jelinek.

A stochastic language model for speech recognition integrating local and global constraint. 1994, *ICASSP'94, pp5-8.* R. Isotani, S. Matsunaga.

Discriminative training of dynamic programming based speech recognizers. 1993, *IEEE trans. on Audio and Speech Processing, Vol. 1(2), pp.135.* P. Chang, B. Juang.

Signal conditioned minimum error rate training, 1995, *Eurospeech'95, pp.495.* W. Chou, B. Juang.