Statistical Analysis of Chinese Language and Language Modeling Based on Huge Text Corpora[1]

Hong Zhang, Bo Xu, Taiyi Huang

National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences
P.O Box 2728, Beijing 100080, P.R China
{ hongzh, xubo, huang }@nlpr.ia.ac.cn

Abstract. This paper presents the statistical characteristics of Chinese language based on huge text corpora. From our investigation, we find that in writing Chinese it is more likely to use long words, while in other language styles the words are shorter. In large text corpora, the number of bigram and trigram can be estimated by the size of the corpus. In the recognition experiments, we find the correlation is weak between the perplexity to either the size of the training set or the recognition character error rate. However, in order to attain good performance, the large training set above tens of million words is necessary.

1 Introduction

Statistical Language model is effective in large vocabulary continuous speech recognition. However, construction of language model needs supporting of large text corpora, which in practice is related to large amount of consumption of human and monetary resource. If there is an empirical research on huge text corpora as the reference of design or selection of the training set for language modeling, the possible waste of resource will be reduced to the minimum.

There are already many text corpora for public purpose of English and other western languages. However the case for Chinese language is not so favorable. Since no previous report of very large Chinese text corpora, this is the first extensive investigation of statistical characteristics of Chinese language on very large text corpora.

The Chinese language is very different from the western languages. The basic unit is Chinese character. There is no separation between words in the Chinese text. Theoretically, any combination of Chinese characters can be defined as a word. In our investigation, the words are defined in a 40k lexicon, in which all Chinese characters are included as the single-character words. Therefore, the out-of-vocabulary problem is circumvented. The words from text are produced by an automatic word

[1] The work described in the paper is funded by the National Key Fundamental Research Program (the 973 Project) under No.G1998030504, and the National 863 High-tech Project under No. 863-306-ZD03-01-1.

T. Tan, Y. Shi, and W. Gao (Eds.): ICMI 2000, LNCS 1948, pp. 279-286, 2000.

segmentation algorithm provided by the Chinese Language Model Toolkit, which is developed by the National Lab of Pattern Recognition.

In this paper, the investigation on text corpora of Chinese language includes two aspects: one is the statistics on words and N-grams, the other is the performance of language models on different scale of corpus.

The rest of the paper is organized as the following: Section 2 introduces the language modeling theory briefly. Section 3 presents the statistics of Chinese language. In section 4 we report the recognition experiments of language models. The correlation between the perplexity and the performance of recognizers equipped with the language model is analyzed in section 5. Section 6 is the conclusion of this paper.

2 Language Modeling Theory

The most commonly used language model in large vocabulary speech recognition system is the trigram model [4], which is to determine the probability of a word given the previous two words: $p(w_3/w_1,w_2)$. The simplest way to approximate this probability is by the maximum likelihood (ML) estimate

$$P(w_3 \mid w_1, w_2) = f(w_3 \mid w_1, w_2) = \frac{C(w_1, w_2, w_3)}{C(w_1, w_2)} \tag{1}$$

where f(|) denotes the relative frequency function, and C( ) denotes the count function.

2.1 Smoothing

Since even in very large corpora of training text, many possible trigram pairs are not encountered. To avoid the language model assign $P(W) = 0$ to strings W containing these gram pairs, it is necessary to smooth the gram pair frequencies, which is the task of smoothing technique.

In the NLPR Toolkit, a smooth method devised by Katz based on Good–Turing estimation [5], was implemented. The basic idea is from the formula (in the case of trigram language model):

$$\hat{P}(w_3 \mid w_1, w_2) = \begin{cases} f(w_3 \mid w_1, w_2) & if \quad C(w_1, w_2, w_3) \geq K \\ \alpha Q_T(w_3 \mid w_1, w_2) & if \quad 1 \leq C(w_1, w_2, w_3) < K \\ \beta(w_1, w_2)\hat{P}(w_3 \mid w_2) & otherwise \end{cases} \tag{2}$$

Where $Q_T(w_3 \mid w_1, w_2)$ is a Good-Turing type function and $\hat{P}(w_3 \mid w_2)$ is a bigram probability estimate having the same form as $\hat{P}(w_3 \mid w_1, w_2)$

$$\hat{P}(w_3 \mid w_2) = \begin{cases} f(w_3 \mid w_2) & if \quad C(w_2, w_3) \geq L \\ \alpha Q_T(w_3 \mid w_2) & if \quad 1 \leq C(w_2, w_3) < L \\ \beta(\_, w_2) f(w_3) & otherwise \end{cases} \tag{3}$$

2.2 Evaluation Measure

According to the information theory [4], perplexity (PP) is defined to measure the text source complexity as following,

$$PP = 2^{LP} \tag{4}$$

Where *LP* is the logprob, which is defined by

$$LP = \lim_{n\to\infty} -\frac{1}{n} \sum_{i=1}^{n} \log Q(w_i \mid w_1, \ldots, w_{i-1}) \tag{5}$$

Where $Q(w_i/w_1,\ldots w_{i-1})$ denotes the recognizer's estimate of the text production probabilities that is embedded in the language model.

3 Statistics on Large Chinese Text Corpora

3.1 Processing of the Chinese Text Material

As described in section 1, there are no separated signs between the words in Chinese language. In practice, the Chinese words are segmented from the text stream according to a lexicon, in which words are defined as combination of Chinese characters. Besides, a step of preprocessing is also necessary to refine the raw text before the word segmentation, as shown in figure 1.

All the works concerned by this paper of the processing of the text material are under the auxiliary of NLPR Chinese Language Model Toolkit (v 1.0) [6].

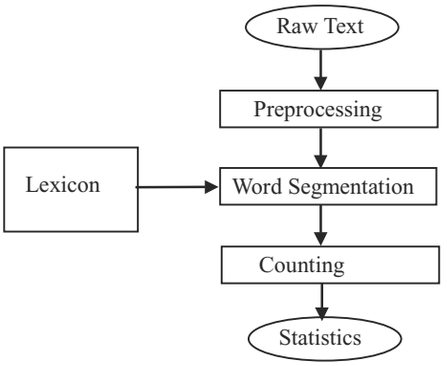

Fig. 1. Processing of Chinese Text Material.

3.2 Average Word Length

The average word length refers to the average length in numbers of Characters of words, which is defined as

$$L_w = \frac{N_c}{N_w} \tag{6}$$

Where L_w is the average word length, N_c is the number of Chinese characters in the text after preprocessing, N_w is the number of Chinese word after word segmentation.

In our investigation, all the Chinese words are defined according to a 40K lexicon, in which the average word length is 2.2266 Nc/W.

Table 1. Statistics of Average Word Length

Classes	Intent	Average Word Length
Newspapers and magazines	People's Daily (81,94,95,96)	1.6527
	Reference News(92)	1.6558
	Daily of Science and Technology (91-96)	1.6527
	The Computer(98)	1.5414
	The Chinese Computer (98)	1.6164
	Chinese Reader's Digests (98, Bound Edition of 200 volumes)	1.5297
Books	Translated Foreign literature	1.4994
	Ancient Chinese literature	1.3519
	Modern Chinese literature	1.4581
	Culture	1.5291
	Techniques	1.5152

The average word lengths of some classes of text material are shown in Table 1.

The statistics of the corpora reveals that the average word length shows prominent feature related to the kind of the text. The average word length of most texts from newspapers are higher than 1.60, while that of most novels and scientific publications are less than 1.5. The main language style of newspapers is writing Chinese, and in books, especially in novels, there are many oral and informal expressions of Chinese.

3.3 Numbers of Bigram and Trigram

The number of N-gram is investigated with increment amount of text. The correlations between the number of bigram and trigram to the size of corpus are shown in figure 2 and figure 3, respectively.

The statistics of N-gram pairs on corpora of totally 2960 million words shows that numbers of bigram pairs and trigram pairs increase with the enlarging of the size of training sets. The increasing speed of pair numbers is a little slow when the training set is larger than 1000 million words. In the log-log scale, there exists an approximate linear relation between number of bigram and trigram pairs and the size of the

training set in million words. According to the data we can conclude two coarse estimation equations:

$$\log_{10}(NB) = 0.6127 \log_{10}(SC) + 5.7301 \quad (7)$$

$$\log_{10}(NT) = 0.7928 \log_{10}(SC) + 6.1053 \quad (8)$$

Where NB is the number of bigram in millions, NT is the number of trigram in millions, and SC is the size of corpus in million words.

In large text corpora, the number of bigram and trigram can be estimated by the size of the corpus. From these relationships, the amount of raw material can be predicted according to the size of the language model, which is often limited by the computing capability of the hardware of the recognizer.

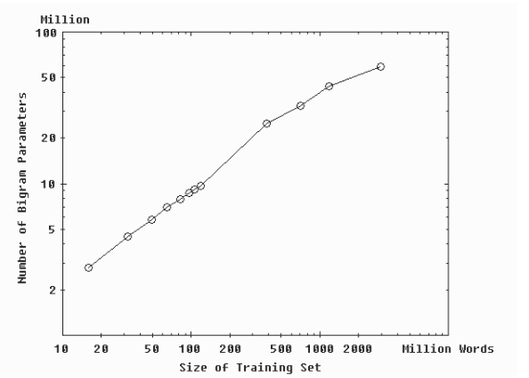

Fig. 2. The number of bigram with different size of corpus.

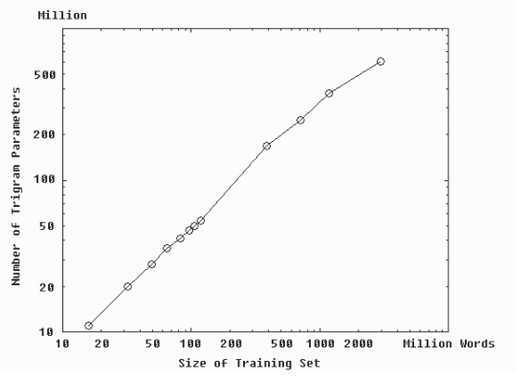

Fig. 3. The number of trigram with different size of corpus.

4 Performance of Language Models

4.1 Corpora for Language Modeling

There are 11 corpora constructed for testing of performance. All the corpora are in an extending style, as shown as the following:

$$C_{i+1} = C_i \bigcup C_{extend}, i = 1,2,...n-1 \quad (9)$$

where C denotes the corpus, n is the total number of the corpora.

The first 8 corpora are confined within economical affairs from the Economic Daily. The 9th is consisted of pure newspapers. The last two contain newspapers, magazines, books and scientific digests.

4.2 Recognition Experiments

The test corpus is reading speech of economical news from 10 female speakers, which is a sub set of Corpus99 developed by the National Lab of Pattern Recognition.

The recognizer is an embedded system of FlyingTalk, the recognition engine developed by NLPR.

The recognition results in Character Error Rate (CER) concerned with every language model are shown in table 2. The results show that, before the size of training set attains to 1000 million words, the CER decreases with the increasing of training set. When the training set is very large and extending broadly with content, the CER increases and the performance of the recognizer drops.

Table 2. Performance of language models

LM	training set (Million Words)	CER(%)	Perplexity
eco_83_85	16	25.67	389
eco_83_88	32	24.90	287
eco_83_91	49	23.35	244
eco_83_93	65	21.96	211
eco_83_95	82	21.77	195
eco_83_96	96	21.57	185
eco_83_97	106	20.93	178
eco_83_98	118	20.38	166
newspaper	707	18.29	189
balance	1180	22.19	151
general	2960	22.58	169

4.3 Language Evaluations by Perplexity

The language models are evaluated with the economics part of the text corpus from which the transcription of Corpus99 are based, rather than the much larger language model test text. This circumvents the problems caused by any potential mismatch between the language model test text and the recognition task itself.

The values of perplexity of each language model are shown in Table 2. We can find that the correlation between the perplexity and the CER is not very strong.

5 Discussion

5.1 Correlation between CER and Perplexity

In our investigation, the correlation coefficient r is computed according to [1]

$$r = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} \tag{10}$$

The correlation coefficient between CER and the perplexity of language models of the data of section 4 is 0.7626, which shows a loose relation of the two variables.

5.2 Analysis of the Perplexity

Perplexity is a widely used measure of language model quality. However, same as the results in section 4, many recent work by other researchers (for example, in [3]) has demonstrated that the correlation between a language model's perplexity and its effect on a speech recognition system is not as strong as was once thought.

According to equation (5) we know that perplexity is based solely on the probabilities of the words which actually occur in the test text. It does not consider the alternative words which may be competing with the correct word in the decoder of a speech recognizer.

Table 3. Comparison of different recognition results on perplexity computed by Language Model "balance".

Recognition results	With LM balance	With LM newspaper	With LM eco_83_98	With LM eco_83_97
Perplexity computed by LM balance	498	626	572	603

From Table 2 we know that language model "balance" has the lowest perplexity, however, the performance of recognizer with this language model is not as good as it should be according to the perplexity. Table 3 compares the perplexity values of

model “balance” on the recognition result texts of recognizer with model “balance”, “newspaper”, “eco_93_98” and “eco_83_97”, the latter 3 models have higher perplexity in Table 2 but lower CER than model “balance”. Table 3 implies that, in the space of words array in the decoder of the recognizer, model “balance” has picked the most reasonable result. Or in other words, in the view of model “balance”, the more correct results picked by model “newspaper”, “eco_93_98” and “eco_83_97”, are less probable.

Since the purpose of language model in a recognizer is to pick the correct words from competing words with acoustic similarity in the decoder, the feasible evaluation measure of language model should have the ability to stand for this point. Some researchers already started the exploration of new evaluation measures ([2],[3]). Besides innovation on structure of language model, this is another significant direction in the developing of theory of language modeling.

6 Conclusions

This paper presents the statistical characteristics of Chinese language and its statistical language models based on huge text corpora. From our investigation on the large Chinese text corpora, we find that, the statistics of the corpora reveals that the average word length shows prominent feature related to the class of the text, and the number of bigram and trigram can be estimated by the size of the corpus. In the recognition experiments, we find the correlation is weak between the perplexity and either the size of the training set or the recognition character error rate. The drawback of perplexity exists intrinsically, so new evaluation measures should be explored.

References

[1] Berenson, M., Levine, D. and Mercer, R.L., “Applied Statistics, A First Course.” Prentice-Hall International, 1988.

[2] Chen,S., Beeferman, D., and Rosenfeld, R. (1998). “Evaluation Metrics for Language Models.” In Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, 1998.

[3] Clarkson, P. and Robinson, T. “Towards Improved Language Model Evaluation Measures”, In proceedings of EUAROSPEECH’99, Sep. 5-9, 1999 Budapest, Hungary.

[4] Jelinek, F. (1998).“Statistical Methods for Speech recognition”, The MIT Press, 1998.

[5] Katz, S.M, “Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer.” IEEE Transactions on Acoustics, Speech and Signal Processing, 35(3): 400 – 401, 1987.

[6] Zhang, H., Huang, T.， and Xu, B. (2000), The NLPR Chinese Language Model Toolkit (V1.0) for Large Text corpus, 2000 International Conference on Multilingual Information Processing (2000 ICMIP), Urumqi, China, July 20-25, 2000.