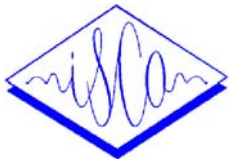

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

CHARACTERISTICS OF CHINESE LANGUAGE MODELS FOR LARGE VOCABULARY TELEPHONE SPEECH

Roger H.Y. Leung, Chi-Yan Choy, Hong C. Leung
Department of Electronic Engineering,
The Chinese University of Hong Kong
{hyleung, cychoy, hcleung}@ee.cuhk.edu.hk.

ABSTRACT

This paper is concerned with language modeling (LM) for large vocabulary speech recognition in Mandarin Chinese. As the language characteristics of Chinese are quite unique, we investigate some novel techniques in language modeling. We also borrow some of techniques that have been applied to other languages. Experiments have been conducted on the Call Home Mandarin, HUB4, and HUB5 corpora obtained from the Linguistic Data Consortium (LDC). The training set consists of 9.8 hours of spontaneous speech and 100K words in text. The test set consists of 1.6 hours of spontaneous speech and 20K words in text. We have found that our results compare favorably to the results reported in the literature.

10.21437/Eurospeech.1999-413

1. INTRODUCTION

N-gram LM simultaneously encodes syntax, semantics and pragmatics. They concentrate on local dependencies. It is especially effective for structural languages, such as Chinese, where word order is important and the contextual effects among neighbor words are strong. However, people tend to speak more freely (less constraint in syntax or grammar) in telephone conversation. Also, dis-fluency problem is very severe in Call Home corpus. Those problems make the building of LM even more difficult. This paper discusses several methods to handle these problems. We have studied the following:

1. Word LM versus Character LM: Chinese is a syllable-based language with one Chinese character corresponding to a syllable. Furthermore, the definition of a Chinese word is not well defined, in contrast to many western languages such as English. Therefore, we have investigated the use of character LM and word LM.
2. Discount Methods: Multiple discount methods have been proposed in the past. We have studied the effects of 4 well-known discount methods, and will report the properties and advantages of each.
3. Interpolation Methods: Two interpolation LMs have been explored, based on a newspaper corpus of HUB4 and spoken transcription of HUB5.
4. Integration with Acoustic Modeling: A large vocabulary speech recognizer has been developed, utilizing the LMs we have developed.

2. EXPERIMENTAL DATABASE

All our experiments were performed on Call Home Mandarin Corpus [1], which is a large vocabulary, telephone-bandwidth conversational speech corpus. The speech corpus contains dialogues between two Mandarin speaking people through long distance calls. They have no specific topics, and the speech files were transmitted through the telephone networks. There are 80 conversations in the training set and 20 in the development test set. The training set contains 19K sentences and 5774 unique words. The average word length is about 1.39 character, which is about the same as written Chinese [3]. The detailed statistics are shown in Table 1.

Transcription	**Training**	**Development Testing**
# of dialog	80	20
# of sentence	19,965	5,378
# of word	127,063	34,699
# of character	177,148	48,218
# of unique word	5,774	2,936
# of unique character	2,098	1,466
Ave. word length (character/word)	1.394	1.390

Table 1: Detailed statistics of Call Home Corpus

3. ACOUSTIC MODELING

Mandarin is a monosyllabic and tonal language. The total number of phonologically allowed syllables is about 1300. Each syllable is assigned a tone and there are a total of five lexicon tones. Tones can be separately recognized using the pitch contour information. If tone information is ignored, the 1300 tonal syllable can be reduced to only 408 base syllables. We used the 408 base syllables as speech units in all our experiments. We used the same architecture for all the recognizers in our experiments: hidden Markov

Model (HMM) technique for acoustic modeling. Each of the 408 HMMs has 8-states and 8 mixtures. The HMM states are arranged in a left-to-right, no-state-skipping topology. The segmental k-means algorithm is used for training and the Viterbi algorithm is used for decoding. 13 MFCC, 13 ΔMFCC, Energy and an ΔEnergy are used as feature vectors for the recognizer. Since our main theme is to compare the LMs, no context-dependent models have been used to improve the recognition result.

4. LANGUAGE MODELING

The simplest lexical unit for Chinese is the character. There are more than 10,000 Chinese characters. Most Chinese word is composed of one to four characters. The combinations are usually based on the lexical meaning of the characters similar to prefix or suffix. However, as the rules of word formation are not well defined, the total number of Chinese words is not well defined as well. For training purposes, we used a lexicon of the 5,774 most frequently used words.

4.1 Construction of Language Model

While the bigram and trigram probabilities can be statistically estimated [4]. We have found that the unseen data problem is very severe. Hence, we adopted backoff model. In this model, a language probability backoffs from a trigram to a bigram, and then to a unigram estimation if necessary. An intuitive impression for the quality of the LM can be conceived from Table 2. The simple backoff LM is used to predict the potential words for the sentence "<s> 他 都 不 知道 他 的 條件 有 多麼 好 </s>". Table 2 lists the ranks of correct words. For example, knowing the preceding word "有", the LM estimates that the most likely next word is "個", and the word 機會,時間… are all more likely than the actual word "多麼" which is estimated as the 219-th likeliness, given that particular past.

We can observe that the LM is quite effective at predicting most function words (e.g. 我,你,他) but that is uncertain about some content words (e.g. 條件, 多麼). Another observation is that the LM provides powerful constraint for a speech recognizer. In the above example, the correct words are always within the top 800 candidates, instead of 5774 words.

Word	<s>	他	都	不	知道	他
Rank	1	10	17	2	1	8

Word	的	條件	有	多麼	好	</s>
Rank	8	785	40	219	34	1

Table 2: Rank of correct word in the simple backoff LM

4.2 Word LM and Character LM:

Character LM has advantage that its OOV rate is low while word LM has more powerful constraint in syntax. Figure 1 compares the word perplexity of the word-level and character-level LMs, as functions of the OOV [5]. As usual, we notice that increasing the vocabulary size always reduces the OOV rate. We have also found that the word-level bigram consistently results in lower word perplexity than the character-level bigram, suggesting that the word-level LM should be more appropriate for speech recognition.

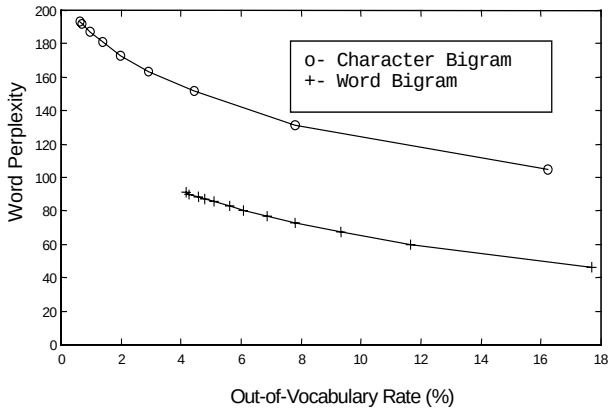

Figure 1: Perplexity vs. OOV % for Word LM & Character LM

4.3 Smoothing:

Although backoff technique helps to alleviate the zero probability problem, the backoff model is poor in estimating the actual probability of unseen events. Hence, four different smoothing techniques are employed. The first mentioned smoothing method is Witten-Bell smoothing method [7]. The main idea behind the Witten-Bell smoothing is that "Use the count of things you have seen to help estimating the count of things you haven't seen." This idea gives a simple but useful estimation for unseen events. A more complex smoothing method, called Good-Turing smoothing method [8], is also introduced. The basic idea of Good-Turing smoothing is to re-estimate the low or zero count n-gram by looking at the number of n-grams with a higher count. Another two smoothing methods are absolute and linear smoothing [9]. They are sharing the same idea that each original count is subtracted by certain value. In the case of absolute discount, the subtracting value is an 'absolute' constant. Therefore, we have 'absolute discount' as her name. Simultaneously, the subtracting value of linear discount is 'linear' to the original count. Hence, we have the name 'linear discount'.

The performance of the smoothing methods can be found in Table 3 . All the four smoothing methods give more than 39% reduction in perplexity for the Call Home task. Good-Turing smoothing is the most effective one among these

methods. Specifically, the Good-Turing smoothed word bigram and character bigram perplexities are 90.88 and 193.09 respectively. These results compare favorably to the simple backoff bigram perplexities of 175.13 and 317.93. As shown in Table 3, the simple absolute discount method gives the perplexity value 91.97, which is better than the more well-known and more sophisticated Witten-Bell algorithm. For the reason of comparison, character perplexity is translated to word perplexity using to formula $PP_C = \sqrt[L]{PP_W}$, where PP_W is the average word perplexity, PP_C is the average character perplexity, and L is the average length of a word.

Discount Method	Word-LM Perplexity	Character-LM Perplexity
Simple Backoff	175.13	$63.13^{1.39} = 317.93$
Good Turing	90.88	$44.10^{1.39} = 193.09$
Absolute	91.97	$44.25^{1.39} = 194.01$
Witten Bell	93.88	$44.47^{1.39} = 195.35$
Linear	106.00	$48.22^{1.39} = 218.62$

Table 3: Comparison of different Discounting Methods for the Call Home task

4.4 Interpolation:

In this experiment, the probability of a given sentence assigned by the interpolated model is defined as a weighted sum of the probability assigned by the original models:

$$P(w \mid ILM) = (1-\alpha)P(w \mid CHLM) + \alpha P(w \mid HBLM)$$

$$P(w \mid ILM) = \begin{cases} (1-\alpha)P_{CH\_2}(w_n \mid w_{n-1}) + \alpha P_{HB\_2}(w_n \mid w_{n-1}) \\ (1-\alpha)P_{CH\_2}(w_n \mid w_{n-1}) + \alpha\beta_{HB\_1}(w_{n-1})P_{HB\_1}(w_n) \\ (1-\alpha)\beta_{CH\_1}(w_{n-1})P_{CH\_1}(w_n) + \alpha P_{HB\_2}(w_n \mid w_{n-1}) \\ (1-\alpha)\beta_{CH\_1}(w_{n-1})P_{CH\_1}(w_n) + \alpha\beta_{HB\_1}(w_{n-1})P_{HB\_1}(w_n) \end{cases}$$

where α is the interpolation ratio, β is the backoff weight, and CH-LM stand for Call Home LM, HB-LM stand for HUB5 LM.

The weight α is found using the estimation maximization (EM) algorithm [10] which minimizes the perplexity of the interpolated model over the training data. The Interpolated LM Bigram is then generated through the formula above. There are four possible cases: (1)both CH-LM and HB-LM has the bigram, (2) only CH-LM has the bigram, (3) only HB-LM has the bigram, (4) both CH-LM and HB-LM do not have the bigram. Both perplexity and OOV rate can be improved by interpolating the Call Home LM with HUB5 conversation transcription. By using the Estimation Maximization (EM) algorithm, it is found that the perplexity is optimized when α=0.2.

Figure 2 shows the change of word perplexity at different interpolation ratios for simple backoff LMs. There is 6.3% improvement in perplexity.

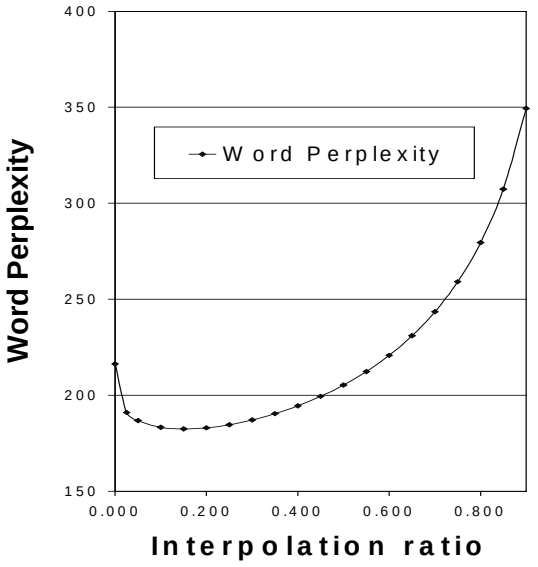

Figure 2: Word Perplexity of (Call Home & HUB5) Interpolated LM

The improvement of perplexity and OOV rate is less significant in the smoothed LM but it is still noticeable. As shown in Table 4, our interpolation methods have successfully reduced the OOV by 6.8%. In addition, the perplexity improves from 90.88 to 88.90.

Language Model	Perplexity	OOV
Smoothed FWB	90.88	1441
Interpolated and Smoothed FWB	88.90	1343

Table 4: Perplexity and OOV reduction of Interpolated LM

5. SYSTEM EVALUATUION

Perplexity and speech recognition results are summarized in the following table. We have found that our results compare favorably to the results reported in the literature. By using the simple word bigram LM, we have the syllable accuracy increased from 20.17% to 27.24%, which is a 35% improvement when compared to the system without LM. In addition, the Good Turing smoothed word bigram produces the syllable accuracy 31.46%. It is a significant improvement when compared with the simple backoff character bigram. Moreover, the interpolated LM can further improve the accuracy to 32.02%, which is 0.8% improvement over the smoothed one.

Language Model	Word Perplexity	Syllable Accuracy	Character Accuracy
No LM	N/A	20.17%	N/A
Fair Character Bigram	N/A	27.24%	23.25%
Fair Word Bigram (FWB)	175.13	29.37%	25.24%
Smoothed FWB	90.88	31.75%	27.94%
Interpolated and Smoothed FWB	88.90	32.02%	28.13%
Cheating Word Bigram	18.35	38.28%	36.09%

Table 5: Recognition results by using different LMs

6. CONCLUSION

The following summarizes our findings:

1. Word LM and Character LM: By using character level LM, the vocabulary size is reduced from 5774 to 2098, which is 63% reduction. Also, the out-of-vocabulary (OOV) rate is reduced from 4.15% to 0.62%. However, experiments show that character level LM results in lower accuracy than word LM by 7.3% due to the loss of linguistic information This result was supported by our perplexity and OOV analysis of both LMs.

2. Smoothing: There is a severe sparse data problem. For example, 26% of the word bigram (58% of word trigram) appearing in the test set never took place in the training transcription. Our investigated smoothing methods result in more than 39% improvement in perplexity.

3. Interpolation: Interpolation improves both perplexity and OOV rate. Our interpolation methods have successfully reduced the OOV by 6.8%. In addition, the CallHome LM is interpolated with HUB4 LM (newspaper text corpus) and HUB5 LM (conversation transcription) individually. By interpolating HUB5 LM, we achieves a further 2.17% improvement in perplexity, and a 0.8% improvement in accuracy.

7. ACKNOWLEDGEMENTS

This project is partially supported by Sir Edward Youde Memorial Fund.

8. REFERENCES

[1] R. Agarwal, B.Wheatley, Y. Muthusamy, and T.Staples, "Diagnostic Profiling for Speech Technology Development: Call Hone Analysis", Proceedings of Speech Research Symposium, P.P 131-137, June, 1995.

[2] Fu-Hua Liu, Micheal Picheny, Patibandla Srinivasa, Michael Monkowski and Julian Chen, "Speech Recognition on Mandarin Call Home: A large-Voculary, Conserational, and Telephone Speech Corpus", International Conference on Spoken Language Processing, 1996.

[3] Editors, "Some Characteristics of Chinese Language", Language Information Sciences Research Center Newletter 2, City University of Hong Kong, 1997.

[4] L.R. Bahl, F. Jelinek, & R.L. Mercer, "A Maximum Likelihood approach to continuous Speech Recognition", IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-5(2): 179-190, 1983.

[5] Kyuwoong Hwang, "Vocabulary Optimization Based on Perplexity", IEEE Trans ASSP, 1997.

[6] A. Nadas, "Estimation of probabilities in the language model of the IBM speech recognition system", IEEE Trans, Acoustic, speech and Signal Proc., vol.32, pp.859-861, Aug. 1984.

[7] Timothy C. Bell, J.G. Cleary & Ian H. Witten, "Text Compression", Englewood Cliffs, N.J, Prentice-Hall, 1990.

[8] Good I.J., "The population frequencies of species and the estimation of population parameters.", Biometrika,40:237-264, 1953.

[9] Hermann Hey, Ute Essen & Reinhard Kneser, "On the Estimation of 'Small' Probabilities by Leaving-One-Out", IEEE Transactions of Pattern Analysis & Machine Intelligence, vol. 17, No.12, p1202-1212, Dec. 1995.

[10] R. Rosenfeld, "Adaptive Statistical Language Modeling: A Statistical approach PhD Thesis", School of Computer Science, Carnegie Mellon university, April 1994.