LEXICON OPTIMIZATION FOR CHINESE LANGUAGE MODELING

ZHAO Jun[1]

University of Science & Technology of China

junzhao@ustc.edu

GAO Jianfeng, CHANG Eric, LI Mingjing

Microsoft Research China

{jfgao, echang, mjli} @microsoft.com

ABSTRACT

In this paper, we present an approach to lexicon optimization for Chinese language modeling. The method is an iterative procedure consisting of two phases, namely *lexicon generation* and *lexicon pruning*. In the first phase, we extract appropriate new words from a very large training corpus by statistical approaches. In the second phase, we prune the lexicon to a pre-set memory limitation using a perplexity minimization criterion. Experimental results show up to 6% character perplexity reduction comparing to the baseline lexicon.

1. INTRODUCTION

Almost all techniques to statistical language processing are word based. In particular, word-based statistical language modeling (LM) has been successfully applied to many domains such as speech recognition [1], information retrieval [2], and spoken language understanding [3].

Although word-based LM works very well for western languages, where the words are well defined, it is quite difficult to apply for Chinese. Chinese language is based on characters. There are no spaces between characters and word boundaries are not explicitly marked. Therefore, the "word" in Chinese is actually not well defined, and there does not exist a commonly accepted lexicon. Furthermore, the segmentation of a sentence into a string of words is not unique. These factors make language modeling very sophisticated in Chinese language, and the "out of vocabulary (OOV)" problem especially serious.

This paper presents a new approach to lexicon optimization for Chinese language modeling. The method is an iterative procedure consisting of two phases. The first phase is the mutual information based new words generation. The second phase is the lexicon pruning using a perplexity minimization criterion. Experimental results show up to 6% character perplexity reduction comparing to our baseline lexicon.

In Section 2, we give more details about Chinese processing and discuss related works briefly. In Section 3, we describe the method of lexicon optimization in detail. In Section 4, we present experimental results. Finally, we give our conclusions.

2. CHINESE PROCESSING AND RELATED WORKS

Chinese language is based on characters. There are 6763 frequently used Chinese characters. Each Chinese word is a semantic concept that is about 1.6 characters long on average. But there is no standard lexicon of words -- linguists may agree on some tens of thousands of words, but they will dispute tens of thousands of others.

Furthermore, sentences are written without spaces between words. So a sequence of characters will have many possible parses in the word segmentation stage.

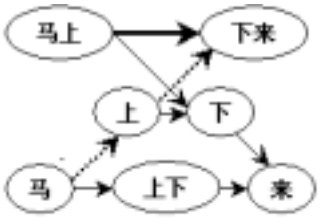

Figure 1. The word graph of Chinese sentence " 马上下来"

Figure 1 shows the segmentation of a simple sentence with only four characters. Here, these four characters can be parsed in five ways into words. For example, the dotted path represents "*dismounted a horse*", and the bolded path represents "*immediately coming down*". This figure also shows seven possible "words", some of which (e.g., 上下) might be disputable on whether they should be considered "words".

One might believe that the character-based LM can bypass the above problems. However, previous work [4] has found that a Chinese LM built on characters did not yield good results, because the words constructed with a few characters always carry different meanings from characters. Therefore, a lexicon is indispensable for Chinese LM. There are many related works on lexicon modeling in recent years. In [4][5], the elements of the

[1] The work was done while the author was visiting Microsoft Research China.

lexicon can be words or any other "segment patterns", which are extracted from the training corpus by statistical approaches. E. P. Giachin [6] has studied the effects of adding phrase *bi-grams* to the lexicon, and has found that it can greatly reduce the perplexity of a language model. Nevertheless, blindly adding new words to the lexicon will damage the quality of the lexicon while removing (or decompose) some compound words can improve the lexicon [7].

Although [4] and [5] present a new direction for Chinese lexicon modeling, their methods are hard to implement and evaluate efficiently. [6] and [7] provide some heuristics of adding and removing lexicon items, but they did not take Chinese into account. In this paper, we propose an efficient method of lexicon optimization for Chinese language modeling.

3. LEXICON OPTIMIZATION

Extending previous methods described above, we propose a new approach to lexicon optimization for Chinese language modeling. Our method is an iterative procedure consisting of two phases, namely *lexicon generation* and *lexicon pruning*. In the first phase, we extract appropriate new words (lexicon items) from a very large training corpus by statistical approaches. In the second phase, we prune the lexicon to a pre-set memory limitation using a perplexity minimization criterion. In what follows, we will describe each phase in detail.

3.1 Lexicon Generation

We hereby investigate statistical approaches to Chinese new words extraction from very large corpus. The basic idea is that a Chinese new word should appear as a stable sequence in corpus. That is, the components in the new word are strongly correlated, while the components lie at both ends should have low correlations with outer words.

Our method is similar to [8, 9, 10, 11]. It consists of two steps. At first, a list of candidate Chinese new words is extracted from a very large corpus by using *mutual information*. Then, *relative frequency* and *context dependency* are used to remove undesirable words.

3.1.1 Mutual Information

According to our study on Chinese corpora, most words are less than 5 characters long, and the average length of words in the segmented-corpus is of approximately 1.6 characters. Therefore, only word *bi-gram*, and *tri-gram* in the corpus are of interest to us in new words extraction. For simplicity, we discuss *bi-grams* only in this paper.

We use *mutual information* as a criterion to evaluate the correlation of different components in the new word.

Theoretically, the *mutual information* of two random *X* and *Y* is given by:

$$MI(X,Y) = H(Y) - H(Y/X) = H(X) + H(Y) - H(X,Y) \qquad (1)$$

where *H(.)* is the entropy. The mutual information between two symbols x and y is interpreted as:

$$\log P(x,y) / P(x)P(y) \qquad (2)$$

Similarly, in our experiments described in the next section, we estimate the *mutual information MI(x,y)* of a *bi-gram (x, y)* by the following 3 forms:

$$MI(x,y) = \log(P(x,y) / P(x) + P(y)) \qquad (3)$$

$$MI(x,y) = \log P(x,y)^{\alpha} / P(x)P(y) \qquad (4)$$

$$MI(x,y) = P(x,y)\log P(x,y) / P(x)P(y) \qquad (5)$$

where *P(.)* is the probability, and is the α is the coefficient tuned to maximize the performance.

The extracted words should be of higher value of *MI* than a pre-set threshold.

3.1.2 Relative Frequency and Context Dependency

Using *mutual information* alone results in many undesirable new words that are either of low frequency or of semantic incomplete.

An obvious solution to reduce noisy lexicon is to set a threshold on *bi-gram* frequency. All *bi-grams* with lower frequency are removed before estimating the *mutual information*.

We also argue that the extracted Chinese words should be semantic complete. That is, we should generate a whole word, not a part of it. For example, 导弹防御计划(missile defense plan) is a complete word, and 导弹防御 (missile defense) is not, although both have relatively high value of mutual information.

Therefore, we use another feature, called *context dependency*. The contexts of the word 防御(defense) are illustrated in Figure 2.

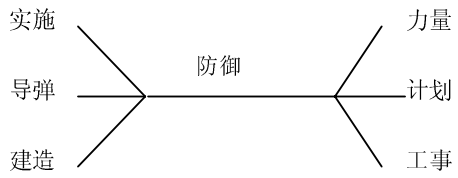

Figure 2.

A compound word *X* has left context dependency if

$$LSize = |L| < t1 \text{ or}$$

$$MaxL = MAX_{\alpha} \frac{f(\alpha X)}{f(X)} > t2$$

Where *t1*, *t2* are threshold value, *f(.)* is frequency, *L* is the set of left adjacent strings of *X*, $\alpha \in L$ and $|L|$ means the number of unique left adjacent strings.

Similarly, a compound word *X* has right context dependency if

$$RSize = |R| < t1 \text{ or}$$

$$MaxR = MAX_{\beta} \frac{f(\beta X)}{f(X)} > t2$$

Where *t1*, *t2* are threshold value, *f(.)* is frequency, *R* is the set of right adjacent strings of *X*, $\beta \in R$ and $|R|$ means the number of unique left adjacent strings.

The extracted complete new words should have both left and right *context dependency*.

3.2 Lexicon Pruning

The procedure of lexicon pruning is entirely automatic and works according to a perplexity minimization criterion. It states by identifying the word consisting more than two shorter words/characters, $w = (w_1, \ldots, w_n)$, when removing it from the lexicon (i.e. by dividing into several shorter words), produces the lowest increased perplexity. By iteratively repeating this action, the lexicon becomes smaller and smaller, while the increase of the character perplexity is minimized.

The core of the algorithm is the computation of the perplexity increase of each word $w = (w_1,\ldots,w_n)$ before and after it is removed from the lexicon (i.e. divided into shorter words $w_1,\ldots,w_n$). It is impractical to estimate the perplexity increase of each word by re-segmenting the whole training data after removing the word, especially when the corpus is large. Therefore, we present an efficient method to estimate the perplexity difference approximately based on the *simple substitution assumption* as follows.

After a word is removed, we simply substitute all occurrence of the word in the corpus by its shorter components.

For example, there is a word sequence, $S = (w_1,w_2,w_3)$, where w_2 is a new word consisting of two shorter words, say $w_2 = (w_{21}, w_{22})$. After removing w_2 from the lexicon, the word sequence is of the form $S = (w_1,w_{21}\ w_{22},,w_3)$. That is, after re-segmentation, there are no new words created, which consist of characters across original words, such as a word consisting of characters in w_1 and w_{21}.

Following this assumption, the perplexity of the LM after removing a word from the lexicon can be re-estimated more efficiently. Considering *bi-grams* only, suppose that we keep the original frequencies of all words and word pairs. When a word $w = (w_1,w_2)$ is removed, all parameters need to be re-computed include 1) the word frequency of w_1, and w_2, and 2) the frequency of word pairs containing w, w_1, and w_2. Then we can use the method suggested in [12] to estimate the relative change in LM perplexity before and after removing the word.

We therefore use a simple thresholding algorithm for lexicon pruning:

1. Select a threshold θ.
2. Compute the relative perplexity increase due to pruning each word (of length more than 2 characters) individually.
3. Remove all words that raise the perplexity by less than θ.

4. EXPERIMENTAL RESULTS

We set up two sets of experiments to test our method. We use the CMU-SLM toolkit [13] to build and evaluate the *bi-gram* LMs.

4.1 Automatic Lexicon Generation

In the first set of experiments, we examined the impact of the lexicon size on the performance of the Chinese LM.

The lexicon is generated on a large training corpus. The training corpus consists of documents from different domains of novel, news, technical report, etc., with approximately 27 million characters. The open test data we use consists of 0.5 million characters that have been proofread and balanced among domain and style.

The initial lexicon of our iterative method contains 6763 frequently used Chinese characters. Our resulting lexicons are compared with the baseline lexicon, which is carefully designed by Chinese linguists, with approximately 53,000 items.

Figure 3 and 4 compare the character perplexity (PPC) based on the lexicons with different sizes. The horizontal axis denotes the size of the lexicons. The rightmost point is the character perplexity of baseline lexicon. The vertical axis is the character perplexity.

As shown in Figure 3 and 4, as the size of the lexicon increases, the perplexity decreases constantly. At the same size of the baseline lexicon, our method achieves similar performance. It turns out that a Chinese lexicon with comparable quality to manually generated lexicon can be obtained automatically from large training corpus using our method.

4.2 Improve the Baseline Lexicon

In the second set of experiments, we use the baseline lexicon mentioned above as the initial lexicon. We expect improvements over the baseline by using our method of lexicon optimization. We also examine which forms of *bi-gram mutual information*, mentioned in Section 3.1, achieve better performance.

The training corpus consists of local news, with approximately 50 million characters. The first test data we use consists of local news, with 52 million characters. The other test data, containing 9 million characters, consists of documents from various domain including shopping, news, entertainment, etc.

The results are presented in table 1 and 2. The first column contains the lexicons generated by different forms of MI. The second and third column contains word perplexity (PPW) and character perplexity respectively. The last column contains the character perplexity reductions (PPC RED) compared with the baseline lexicon.

Row 1 shows the results of the baseline lexicon. Row 2-5 show the results of different forms of *mutual information* corresponding to equation (2)-(5).

We can see that using the *mutual information* of form (5), we got the best result, while using form (2), we obtained the worst result (even worse than the baseline lexicon). It turns out that in case of *bi-grams*, the information of the *bi-gram* relative frequency is very important in the estimation of the probability of the generation of a new word. It acts as a weighted factor of the relative entropy.

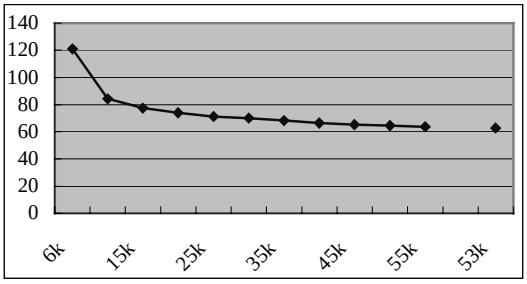

Figure 3. Lexicon size vs. PPC (on open test data)

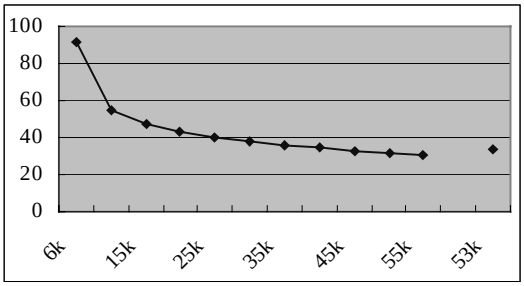

Figure 4. Lexicon size vs. PPC (on training data)

DIC	PPW	PPC	PPC RED
Baseline	376.70	48.89	
MI (2)	405.96	49.39	-1.04%
MI (3)	510.83	47.22	3.42%
MI (4)	442.81	47.72	2.38%
MI (5)	694.16	45.76	6.39%

Table 1. Perplexity results on the first test data

DIC	PPW	PPC	PPC RED
Baseline	750.34	100.32	
MI (2)	825.75	101.53	-1.20%
MI (3)	987.86	98.61	1.71%
MI (4)	845.20	99.07	1.25%
MI (5)	1387.77	98.05	2.26%

Table 2. Perplexity results on the second test data

5. CONCLUSIONS

In this paper, we present an approach to lexicon optimization for Chinese language modeling. The method is an iterative procedure consisting of two phases, namely *lexicon generation* and *lexicon pruning*. In the first phase, we extract appropriate new words from a very large training corpus by statistical approaches. In the second phase, we prune the lexicon to a pre-set memory limitation using a perplexity minimization criterion. Experimental results show that a Chinese lexicon with comparable quality to manual-make lexicon can be obtained automatically from large training corpus using our method. Furthermore, when using the baseline lexicon as the initial lexicon, our iterative method achieves up to 6% character perplexity reduction.

6. ACKNOWLEDGEMENTS

We would like to thank Kai-Fu LEE, Zheng CHEN, and other members from Microsoft Research China, for their help in developing the ideas and implementation in this paper. We would also like to thank Jiang ZHU and Honghui SUN for providing the data used in the experiments.

7. REFERENCES

1. F. Jelink (1990). “Self-organized language modeling for speech recognition”, in *Readings in Speech Recognition*, Alex Waibel and Kai-Fu Lee (eds.), Morgan-Kaufmann, San Mateo, CA, pp. 450-506.
2. D. Miller, T. Leek, R. M. Schwartz (1999). “A hidden Markov model information retrieval system”, in Proc. 22nd International Conference on Research and Development in Information Retrieval, Berkeley, CA, pp. 214-221.
3. V. W. Zue (1995). “Navigating the information superhighway using spoken language interfaces”, IEEE Expert, vol. 10, no. 5, pp. 39-43.
4. K. C. Yang, T. H. Ho, L. F. Chien, L. S. Lee (1998). “Statistics-based segment pattern lexicon – a new direction for Chinese language modeling”, in Proc. IEEE 1998 International Conference on Acoustics, Speech, Signal Processing, Seattle, WA, pp. 169-172.
5. J. Gao, H. F. Wang, M. Li, K. F. Lee (2000). “A Unified Approach to Statistical Language Modeling for Chinese”, *IEEE ICASSP2000*.
6. E. P. Giachin (1995) “Phrase bigrams for continuous speech recognition”, *IEEE ICASSP95*.
7. André Berton, Pablo Fetter, and Peter Regel-Brietzmann (1996) “Compound words in large-vocabulary German speech recognition systems”. *ICSLP96*.
8. L. F. Chien, (1997) “PAT-tree-based adaptive keyphrase extraction for intelligent Chinese Information retrieval”, *ACM SIGIR’97*, Phliadelphia, USA, 50-58
9. D. Wu and X. Xia. (1995). "Large-scale automatic extraction of an English-Chinese lexicon". *Machine Translation* 9(3-4), pp.285-313.
10. M. W. Wu and K. Y. Su. (1993). "Corpus-based Automatic Compound Extraction with Mutual Information and Relative Frequency Count," *Proceedings of R. O. C. Computational Linguistics Conference* VI. Nantou, Taiwan, R. O. C., pp.207-216.
11. J. Zhang, J. Gao, M. Zhou (2000). “Extraction of Chinese Compound Words – An Experimental Study on a Very Large Corpus” to appear in Proceedings of *IRAL*.
12. A. Stolcke (1998). “Entropy-based Pruning of Backoff Language Models” in Proc. *DRAPA* News Transcriptionand Understanding Workshop, Lansdowne, VA., pp.270-274
13. P. R. Clarkson and R. Rosenfeld (1997). “Statistical Language Modeling Using the CMU-Cambridge Toolkit” in Proc. of *ESCA Eurospeech*.