Word Frequency Approximation for Chinese Using Raw, MM-Segmented and Manually Segmented Corpora*

Wei Qiao, Maosong Sun

National Lab. of Intelligent Technology & Systems,
Department of Computer Sci. & Tech.,
Tsinghua University, Beijing 100084, China

qiaow04@mails.tsinghua.edu.cn, sms@mail.tsinghua.edu.cn

Abstract. Word frequencies play important roles in many NLP-related applications. Word frequency estimation for Chinese remains a big challenge due to the characteristics of Chinese. An underlying fact is that a perfect word-segmented Chinese corpus never exists, and currently we only have raw corpora, which can be of arbitrarily large size, automatically word-segmented corpora derived from raw corpora, and a number of manually word-segmented corpora, with relatively smaller size, which are developed under various word segmentation standards by different researchers. In this paper we propose a new scheme to do word frequency approximation by combining the factors above. Experiments indicate that in most cases this scheme can benefit the word frequency estimation, though in other cases its performance is still not very satisfactory.

Keywords: word frequency estimation, raw corpus, automatically word-segmented corpus, manually word-segmented corpus.

1 Introduction

Word frequencies play important roles in many NLP-related applications, for example, TF in information retrieval. The estimation of word frequencies is easy for English whereas difficult for Chinese, because unlike English, there isn't spacing to explicitly delimit words in Chinese text. We therefore can not obtain word frequencies by simply counting word token occurrences in raw corpora.

Generally speaking, we need a 'perfect' (or, correct) manually word-segmented Chinese corpus to estimate word frequencies [5]. However, we face two fundamental difficulties. The first is that there exists serious inconsistency within/among manually segmented corpora, even when the same segmentation standard is

* The research is supported by the National Natural Science Foundation of China under grant number 60573187 and 60321002, and the Tsinghua-ALVIS Project co-sponsored by the National Natural Science Foundation of China under grant number 60520130299 and EU FP6.

adopted for annotation. Due to the characteristics of Chinese word-formation [2, 1], it is very tough to construct a 'fully' correct manually segmented corpus, although the definition of 'word' [14, 13, 11] seems very clear from the linguistic perspective. For example, a constituent, '猪肉', we can either consider it as a compounding word, *pork*, or consider it as a phrase consisting of two single-character words '猪'(pig) and '肉'(meat). Thus, the word frequency of '猪肉' could be pretty high if it is treated in the corpus in the former way, and could also be zero if treated in the latter way. The second difficulty is, according to Zipf's law, in order to obtain a statistically reliable word frequency estimation, even for a medium-sized Chinese wordlist, a balanced corpus with several hundred million characters, rather than several million characters, is required. But constructing such a huge manually segmented corpus is almost impossible, – it is both labor-intensive and time-consuming.

Since a 'perfect' manually segmented corpus is not feasible, (although the 'imperfect' manually segmented corpus is obviously useful for word frequency estimation), we have to in addition consider the possibility of making use of the following three types of corpora for the task here:

The first type is *'perfect' automatically segmented corpus*: Use a 'perfect' word segmenter to segment the corpus automatically, leading to a 'perfect' automatically segmented corpus. Then word frequencies can be easily estimated based on the corpus. Clearly, it would be ideal if a very powerful word segmenter is available [7]. Unfortunately, the state-of-the-art Chinese word segmenters are not satisfactory in performance. In the First International Chinese Word Segmentation Bakeoff in 2003 [8] organized by SIGHAN, the highest F-scores for word segmentation in the open test on four small-scale corpora were 95.9%, 95.6%, 90.4% and 91.2%, respectively. In the Second SIGHAN International Chinese Word Segmentation Bakeoff [3], the situation remains unchanged in nature, despite the minor increase in performance of word segmentation. A side-effect of such systems is that they try to solve segmentation ambiguities and recognize unknown words in context, producing a lot of unexpected inconsistencies in segmentation, which are obviously not favored by the task here.

The second type is *MM-segmented corpus*: Use 'Maximal matching'(MM), the most basic method for Chinese word segmentation, to segment the corpus automatically, then obtain the approximated word frequencies from the resulting corpus. [7] first used MM to handle large-scale texts. According to the direction of sentence scanning, MM can be further sub-categorized as forward MM (FMM) and backward MM (BMM). Experiments in [4] showed that MM is both effective and efficient (fast and easy to implement). [10] distinguished four cases in which FMM and BMM were both considered, and it provides a very strong evidence for supporting MM-based schemes to be reasonable estimations of word frequencies. Another advantage of MM-based schemes is their high consistency in word segmentation. The weak point of MM is that segmentation errors inevitably exist and when out-of-vocabulary words exist, the performance of MM will drop severely.

The third type is *raw corpus*: Use the frequency of a string of characters as an approximation (notice that we use the term 'approximation' here) of the word frequency of a constituent [9], which can be derived directly from any raw corpus. Obviously, its value is always larger than the value of word frequency for any word given a corpus. This scheme may over-estimate word frequencies seriously for some words (in particular for mono-syllabic words), but it has two good properties: the first one is that it is free from any kind of word segmentation errors; the second one is that this kind of corpus can be easily obtained and the size can be arbitrarily large.

According to the analysis above, for the task of word frequency estimation, a 'perfect' word-segmented corpus is ideal but, it doesn't exist, either manually or automatically - what we have are a variety of imperfect ones as well as raw corpora. Each type of corpora has its own advantages and drawbacks, so neither of them alone can fit the task of word frequency estimation. We have to consider a trade-off strategy which tries to utilize all the imperfect word-segmented corpora available so far, ranging from manually segmented corpora, MM-segmented corpora to raw corpora, and combine them to do sort of word frequency approximation, instead of word frequency estimation.

The remainder of this paper is organized as follows: Section 2 introduces the data set we used throughout the paper; Section 3 proposes the construct process of our trade-off scheme; Section 4 presents experiments to show the performance of the proposed scheme. And Section 5 concludes our work.

2 Data Set

In this section we introduce the corpora we used in our experiments throughout the paper.

First, two manually word-segmented corpora: The first one is the HUAYU corpus consisting of 1,763,762 characters, constructed by Tsinghua University and Beijing Language and Culture University. The second one is the BEIDA corpus consisting of 15,839,323 characters, constructed by Peking University. So the manually word-segmented corpora have totally 17,603,085 characters.

Second, the golden-standard corpus: We use a manually word-segmented corpus constructed by the National Institute of Applied Linguistics, denoted YUWEI, which contains 25,000,309 words with 51,311,659 characters. As the YUWEI corpus is sort of a noted authority and relatively large in size, we take it as golden-standard for our tests. An original wordlist is obtained from this corpus and the corresponding word frequencies can be obtained. We delete the words with frequency less than 4 from the original wordlist to form our final wordlist, which is denoted YWL and contains 99,660 entries.

Third, a raw corpus: We use a very large raw corpus, denoted RC, which contains 447,079,112 characters. Taking YWL as the wordlist, we obtain the frequency of a string of characters for each word from RC.

Fourth, MM-segmented corpora: In terms of YWL, we segment the raw corpus RC with FMM-segmenter and BMM-segmenter separately, resulting in two MM-segmented corpora.We denote them RC_FMM and RC_BMM respectively.

Thus in total, we have two moderate size manually-segmented corpora (HUAYU and BEIDA), one very large raw corpus (RC), two MM-segmented corpora (RC_FMM and RC_BMM), and a golden-standard corpus (YUWEI).

3 The Approximation Scheme

In this section we propose our trade-off scheme. In order to properly combine the five corpora which are of different size and different types, the combining process is organized in three steps. Firstly we combine the raw corpus and the two MM-segmented corpora. Secondly we combine the two manually segmented corpora. At last we combine the above two results and obtain the final approximation scheme. In the following, we introduce this step by step.

3.1 Combining Raw and MM-segmented Corpora

From each of the three corpora: the raw corpus and the two MM-segmented corpora, we can obtain word frequency for each word $w_i (i = 1, 2, \ldots, 99660)$, respectively. We use the following symbols to clarify further descriptions:

$f_{FMM}(w_i)$: Word frequency of w_i obtained from RC_FMM.
$f_{BMM}(w_i)$: Word frequency of w_i obtained from RC_BMM.
$f_{RAW}(w_i)$: Frequency of a string of characters for w_i obtained from RC.

The work of [12] indicates that in the framework of MM, the average of $f_{FMM}(w_i)$ and $f_{BMM}(w_i)$ gives the best approximation of word frequencies for 1 to 4 character words. $f_{BMM}(w_i)$ is the best for 5 characters words, and $f_{RAW}(w_i)$ the best for words with word length 6 or above. We simply follow this conclusion here.

Using $F_{RFB}(w_i)$ to represent the result of word frequency approximation by jointly considering RC, RC_FMM and RC_BMM, we have:

For words with 1-4 characters:

$$F_{RFB}(w_i) = \frac{1}{2}\left[f_{FMM}(w_i) + f_{BMM}(w_i)\right] \tag{1}$$

For words with 5 characters:

$$F_{RFB}(w_i) = f_{BMM}(w_i) \tag{2}$$

For words with 6 or more than 6 characters:

$$F_{RFB}(w_i) = f_{RAW}(w_i) \tag{3}$$

This word frequency approximation scheme is called RFB.

3.2 Combining Manually Segmented Corpora

Having two manually segmented corpora HUAYU and BEIDA, we can obtain the word frequency for each word w_i in YWL. We denote the word frequency of (w_i) derived from these two corpora $f_{HUA}(w_i)$ and $f_{BEI}(w_i)$ respectively. We simply take the sum of the two values as the result of word frequency approximation in terms of these two manually segmented corpora, denoted $F_{HB}(w_i)$:

$$F_{HB}(w_i) = f_{HUA}(w_i) + f_{BEI}(w_i) \tag{4}$$

This word frequency approximation scheme is called HB.

3.3 Combining $F_{RFB}(w_i)$ and $F_{HB}(w_i)$

With two parts of combining results $F_{RFB}(w_i)$ and $F_{HB}(w_i)$, we come up with two problems:

The first one is, these two results come from the corpora with different sizes which are extremely unbalanced: one (HUAYU+BEIDA) is 17,603,085 characters while the other (RC) is 447,079,112 characters. So we can not directly combine the results from these two corpora. We thus need to introduce a parameter α to balance the corpus size.

It is naïve that we just take the ratio value of the two corpora size as the value of α. Later on, we will adjust α through experiments to receive the most appropriate value. At this stage, we just take the size ratio as the α value, so α=25.4.

We use C_0 to denote the total number of characters of the manually segmented corpora (HUAYU+BEIDA), and let C_1 denote the total number of characters of raw corpus (RC).

We expect to integrate the manually segmented corpora (HUAYU+BEIDA) and the raw corpus (RC) into a 'new' corpus of size $2C_0$. Thus the size of RC will be reduced to C_1/α . Accordingly the word frequency $F_{RFB}(w_i)$ should be changed to $F'_{RFB}(w_i)$:

$$F'_{RFB}(w_i) = F_{RFB}(w_i)/\alpha \tag{5}$$

In order to keep the whole corpus size to be $2C_0$ after integration, the final manually segmented corpus size, denoted C'_0, should be:

$$C'_0 = 2C_0 - C_1/\alpha \tag{6}$$

Thus the word frequency $F_{HB}(w_i)$ will in turn become $F'_{HB}(w_i)$:

$$F'_{HB}(w_i) = F_{HB}(w_i) \times \frac{2C_0 - C_1/\alpha}{C_0} \tag{7}$$

The second problem concerns an observation in Chinese, i.e., the smaller the word length is, the less reliable the approximation obtained from the raw corpus, and thus the larger the weight of the approximation result from $F_{HB}(w_i)$ should

be. Here, we use a factor β as the weighting parameter. Experimentally, we set β as follows:

$$\beta = \begin{cases} 7 & \text{for one-character words} \\ 6 & \text{for two-character words} \\ 3 & \text{for three-character words} \\ 0 & \text{otherwise} \end{cases}$$

Taking the above two problems into consideration, and based on Equation 5 and Equation 7, the final word frequency approximation in terms of RC, RC_FMM and RC_BMM can be represented as:

$$F''_{RFB}(w_i) = F'_{RFB}(w_i) \times \frac{1}{1+\beta} = F_{RFB}(w_i) \times \frac{1}{\alpha(1+\beta)} \tag{8}$$

Correspondingly the final word frequency estimated by HUAYU and BEIDA should be:

$$F''_{HB}(w_i) = F_{HB}(w_i) \times \frac{2C_0 - \frac{C_1}{\alpha(1+\beta)}}{C_0} \tag{9}$$

Thus we get our final trade-off strategy, denoted $F_{RFB+HB}(w_i)$:

$$\begin{aligned} F_{RFB+HB}(w_i) &= F''_{HB}(w_i) + F''_{RFB}(w_i) \\ &= F_{HB}(w_i) \times \frac{2C_0 - \frac{C_1}{\alpha(1+\beta)}}{C_0} + F_{RFB}(w_i) \times \frac{1}{\alpha(1+\beta)} \\ &= F_{HB}(w_i) \times (1 + \frac{\beta}{1+\beta}) + F_{RFB}(w_i) \times \frac{1}{\alpha(1+\beta)} \end{aligned} \tag{10}$$

Note that for 4 or more than 4 characters words, β=0, thus in these cases Equation 10 reduces to Equation 11:

$$F_{RFB+HB}(w_i) = F_{HB}(w_i) + F_{RFB}(w_i) \times \frac{1}{\alpha} \tag{11}$$

This word frequency approximation scheme is called RFB+HB.

4 Experiments

In order to evaluate the performance of our trade-off scheme (RFB+HB), we conducted experiments from different perspectives. We compare this scheme with the other two schemes: the first one is the scheme using raw corpus and MM-segmented corpora (RFB), the second one is the scheme using only manually-segmented corpora (HB). Following experiments focus on these three schemes.

4.1 Perspective 1: The Spearman Coefficient of Rank Correlation

In terms of word frequencies derived from YUWEI, we can obtain a rank sequence for the 99,660 entries of YWL, denoted R_{YW}, which is in descending order of word frequencies. Similarly, we can also obtain a rank sequence for all these entries in terms of each of $F_{HB}(w_i)$,$F_{RFB}(w_i)$ and $F_{RFB+HB}(w_i)$, denoted R_{HB}, R_{RFB} and R_{RFB+HB} respectively. Every word w_i in YWL has its own rank numbers in R_{HB}, R_{RFB} and R_{RFB+HB}. We assign these rank numbers to w_i, with R_{YW} as a fixed index, resulting in three new rank sequences, denoted $R'_{HB}(w_i)$, $R'_{RFB}(w_i)$ and $R'_{RFB+HB}(w_i)$, accordingly.

Then we calculate the closeness between R_{YW} and each of $R'_{HB}(w_i)$, $R'_{RFB}(w_i)$ and $R'_{RFB+HB}(w_i)$, with R_{YW} as the standard rank sequence. We use the Spearman coefficient of rank correlation (SCRC) to measure the closeness between a pair of rank sequences over YWL, as given by:

$$SCRC \equiv 1 - 6 \sum_{i=1}^{99660} \frac{d_i^2}{N(N^2-1)} \ ,$$

where d_i is the difference between two rank numbers of w_i with respect to R_{YW} and R', N is the length of YWL, and R' is R'_{HB}, R'_{RFB} or R'_{RFB+HB}. Table 1 shows the values of SCRC(R_{YW}, R'), under $\alpha = 25.4$.

Table 1. SCRC values over YWL, under α=25.4.

	(R_{YW}, R'_{HB})	(R_{YW}, R'_{RFB})	(R_{YW}, R'_{RFB+HB})
SCRC	0.675	0.704	0.732

The SCRC value of the proposed scheme is the biggest among the three, indicating that the rank sequence R'_{RFB+HB} is the closest to R_{YW} compared to the rank sequences R'_{HB} and R'_{RFB}.

We also conduct an experiment to determine the most adequate value for α regarding SCRC(R_{YW}, R'_{RFB+HB}).

Fig.1 shows that $\alpha = 6.5$ receives the highest SCRC value. So in the later experiments, we fix $\alpha = 6.5$.

To further observe the performance of the proposed scheme, we continue to carry out some experiments on subsets of YWL. Table 2 and Table 3 give the SCRC values for the top part of YWL with word frequencies ≥ 10 and ≥ 200 respectively.

In these two cases, the proposed scheme also outperforms the other two schemes.

The improvements of the proposed scheme compared to the other schemes over YWL under word frequencies ≥ 4, ≥ 10, and ≥ 200, are summarized in Table 4.

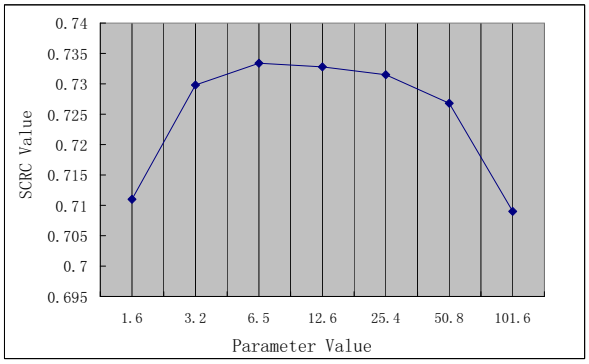

Fig. 1. SCRC value curve with respect to different α' values.

Table 2. SCRC values over YWL for word frequency ≥ 10, under $\alpha = 6.5$.

	(R_{YW}, R'_{HB})	(R_{YW}, R'_{RFB})	(R_{YW}, R'_{RFB+HB})
SCRC(word frequency≥10)	0.663	0.682	0.736

4.2 Perspective 2: Rank Sequence Deviation

Now we look at the performance of the proposed scheme in more detail, particularly its relationship with word length. We therefore define the rank sequence deviation $\sigma(R_{YW}, R')$ with respect to two rank sequences R_{YW} and R', $\sigma_{R'}$ for short, as $\sum_i |R'(w_i) - R_{YW}(w_i)|$(i over a subset of YWL), then calculate σ_{HB}, σ_{RFB} and σ_{HB+RFB}. The values of $(\sigma_{HB+RFB} - \sigma_{HB})/\sigma_{HB}$ and $(\sigma_{HB+RFB} - \sigma_{RFB})/\sigma_{RFB}$ present the varying rate of the σ value using our scheme compared to the other two schemes respectively, as listed in Table 5.

From Table 5 we can see, the proposed scheme receives the best results for 1 to 3 character words in YWL but for 4+ character words, it turns to be worse.

In order to further investigate the performance of our scheme, we divide the YWL words into three parts, i.e., high, medium and low frequency words. Fig.2 shows the coverage rate of top N frequent words to YUWEI.

Based on the coverage rate curve shown in Fig.2, we get the point HM to divide high and medium frequency words and the point ML to divide medium and low frequency words. Then we have:

High frequency words: Top 8,076 frequent words ($1 \sim$ HM), with word frequency > 281; Medium frequency words: the words from $8,077^{\text{th}}$ to $60,224^{\text{th}}$(HM

Table 3. SCRC values over YWL for word frequency ≥ 200, under $\alpha = 6.5$.

	(R_{YW}, R'_{HB})	(R_{YW}, R'_{RFB})	(R_{YW}, R'_{RFB+HB})
SCRC(word frequency≥200)	0.680	0.708	0.771

Table 4. Improvement of SCRC values over different parts of YWL, under $\alpha = 6.5$.

The part of YWL	No. of words	SCRC: $R'_{RFB+HB} - R'_{HB}$	SCRC: $R'_{RFB+HB} - R'_{RFB}$
Words with frequency≥4	99,660	0.057	0.028
Words with frequency≥10	68,100	0.073	0.054
Words with frequency≥200	10,528	0.091	0.063

Table 5. The comparison of rank sequence deviations with respect to word length.

Word length	$\frac{\sigma_{HB+RFB}-\sigma_{HB}}{\sigma_{HB}}$	$\frac{\sigma_{HB+RFB}-\sigma_{RFB}}{\sigma_{RFB}}$	Is HB+RFB the best among three schemes?
1	-22.7%	-16.2%	$\surd$
2	-19.0%	-13.4%	$\surd$
3	-13.7%	-7.3%	$\surd$
4+	15.2%	17.5%	$\times$

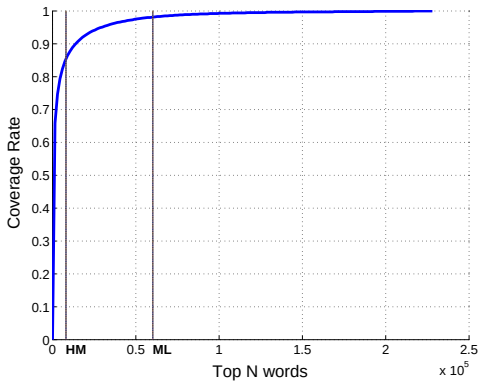

Fig. 2. The coverage rate of the top N frequent words to YUWEI.

$\sim$ ML) with word frequency > 12; Low frequency words: the remained words (ML $\sim$ 99,660), with word frequency > 3.

Then we do experiments on them respectively. The results are given in Table 6, Table 7 and Table 8. We can see that in most cases, our scheme received the best results. But for the low frequency words, especially one character words and 4+ character words, the results turn to be worse.

4.3 Perspective 3: The Coverage Rate

We select the top 50,000 frequent words from R_{HB}, R_{RFB} and R_{HB+RFB}, then calculate the coverage rates of them over YUWEI. Table 9 gives the results.

In Table 9 we can see that the coverage rate of the proposed scheme increases 3.0% and 1.9% compared to HB and RFB respectively.

Table 6. The comparison of rank sequence deviations for high frequency words.

	1 character words	2 character words	3 character words	4+ character words
$\frac{\sigma_{HB+RFB}-\sigma_{HB}}{\sigma_{HB}}$	-44.5%	-38.0%	-68.9%	-88.1%
$\frac{\sigma_{HB+RFB}-\sigma_{RFB}}{\sigma_{RFB}}$	-35.9%	-31.7%	-59.0%	-81.2%
Is HB+RFB the best among three schemes?	√	√	√	√

Table 7. The comparison of rank sequence deviations for medium frequency words.

	1 character words	2 character words	3 character words	4+ character words
$\frac{\sigma_{HB+RFB}-\sigma_{HB}}{\sigma_{HB}}$	-33.0%	-14.5%	-7.5%	-13.4%
$\frac{\sigma_{HB+RFB}-\sigma_{RFB}}{\sigma_{RFB}}$	-18.1%	-7.3%	-9.1%	-10.0%
Is HB+RFB the best among three schemes?	√	√	√	√

Table 8. The comparison of rank sequence deviations for low frequency words.

	1 character words	2 character words	3 character words	4+ character words
$\frac{\sigma_{HB+RFB}-\sigma_{HB}}{\sigma_{HB}}$	27.5%	-24.0%	-17.6%	49.1%
$\frac{\sigma_{HB+RFB}-\sigma_{RFB}}{\sigma_{RFB}}$	-3.1%	-10.9%	-6.2%	22.9%
Is HB+RFB the best among three schemes?	×	√	√	×

Table 9. Top 50,000 words coverage rate on YUWEI.

Scheme	HB	RFB	$HB+RFB$
Coverage rate	94.1%	95.2%	97.1%

4.4 Sample Analysis

Now we choose R'_{HB} and R'_{HB+RFB} to make further comparison. Comparing against R'_{HB}, there are totally 57,024 words in R'_{HB+RFB} whose ranks are better adjusted (i.e., these ranks are closer to the standard sequence $R_{YW}(w_i)$ than their ranks in R'_{HB}), which we call positive samples; 42,619 words whose ranks are worse adjusted (i.e. these ranks are farther apart from the standard sequence $R_{YW}(w_i)$ than their ranks in R'_{HB}), which we call negative samples; 17 words have the same ranks in R'_{HB} and R'_{RFB+HB}. Table 10 and Table 11 show the distribution of positive samples and negative samples over different frequency regions (high, medium, and low), respectively.

Table 10. The distribution of positive samples at different word frequency levels.

Word frequency region	Total words	# of being better adjusted	Proportion
High frequency words	8,076	5,383	66.7%
Medium frequency words	52,148	28,399	54.5%
Low frequency words	39,436	23,242	58.9%

Table 11. The distribution of negative samples at different word frequency levels.

Word frequency region	Total words	# of being worse adjusted	Proportion
High frequency words	8,076	2,679	33.2%
Medium frequency words	52,148	23,746	45.5%
Low frequency words	39,436	16,194	41.1%

Here we give some positive examples which are reasonably adjusted, such as '生物技术(biologic technology)', '知识经济(knowledge economy)', '信息高速公路(information thruway)' and '温室效应(greenhouse effect)'. These words have high frequency nowadays. When using our scheme, the ranks of this kind of words are properly adjusted ahead. We also give some negative examples, such as '周总理(Premier Zhou)', '中央红军(central red army)' and '西单商场(Xidan Market)'. These words are in rare use today, but our scheme made wrong decisions by adjusting them to higher rank positions, due to the fact that these words were frequently used historically, as reflected in RC, a very large raw corpus covering the linguistic phenomena of that time span more intensive than HUAYU, BEIDA, as well as YUWEI.

5 Conclusion and Future Work

In this paper we propose a trade-off scheme which jointly uses the raw corpus, MM-segmented corpora and manually segmented corpora to make approxima-

tion for word frequencies in Chinese. The experiments indicate that this new scheme can benefit the word frequency estimation, though in some cases it seems not very satisfactory, as indicated by '×' in Table 8. Besides, the experiments presented here are also very preliminary mainly due to the limited resources available. How to obtain a more accurate word frequency estimation for Chinese is still a big challenge.

References

1. Chen G.L.: On Chinese Morphology. In: Xuelin Publisher, Shanghai, (1994)
2. Dai X.L.: Chinese Morphology and its Interface with the Syntax. In: Ph.D Dissertation, Ohio State University, USA, (1992)
3. Emerson T.: The Second International Chinese Word Segmentation Bakeoff. In: Proceedings of the Third SIHAN Workshop on Chinese Language Processing. Jeju, Korea, (2005)
4. Liang N.Y.: CDWS: A Word Segmentation System for Written Chinese Texts. Journal of Chinese Information Processing. Vol. 1, No. 2, 44-52, (1987)
5. Liu E.S.: Frequency Dictionary of Chinese Words. Mouton and Co N.V. Publishers, (1973)
6. Liu K.Y.: Study on the Evaluation Technique for Word Segmentation of Contemporary Chinese. Applied Linguistics (Beijing). No. 1, 101-106,(1997)
7. Liu Y., Liang N.Y.: Counting Word Frequencies of Contemporary Chinese - An Engineering of Chinese Processing. Journal of Chinese Information Processing. Vol. 0, No. 1, 17-25, (1986)
8. Sproat R., Emerson T.: The First International Chinese Word Segmentation Bakeoff. Proceedings of the Second SIHAN Workshop on Chinese Language Processing. Sapporo, Japan, 133-143, (2003)
9. Sun M.S., Shen D.Y., T'sou B.K.Y.: Chinese Word Segmentation without Using Lexicon and Hand-crafted Training Data. Proceedings of 36th ACL and 17th COLING, 1265-1271, Montreal, Canada, (1998)
10. Sun M.S., T'sou B.K.Y.: Ambiguity Resolution in Chinese Word Segmentation. Proceedings of the 10th Pacific Asia Conference on Language, Information and Computation. Hong Kong, 121-126, (1995)
11. Sun M.S., Wang H. J. et al.: Wordlist of Contemporary Chinese for Information Processing. Applied Linguistics (Beijing), No. 4, (2001), 84-89
12. Sun M.S., Zhang Z.C., Benjamin KYT'sou., Lu Huaming.: Word Frequency Approximation for Chinese without Using Manually Annotated Corpus. In: Proceeding of 7th International Conference, CICLing 2006, Mexico, 105-116, (2006)
13. Tang T.C.: Chinese Morphology and Syntax. Vol. 3. Taiwan Student Publisher, Taipei, (1992)
14. Zhu D.X.: Lectures on Grammar. The Commercial Press, Beijing, (1982)