# A BOOTSTRAP METHOD FOR CHINESE NEW WORDS EXTRACTION

*Shan He, Jie Zhu*
Department of Electrical Engineering, Shanghai Jiaotong University
Shanghai 200030, China

## ABSTRACT

A bootstrap approach for extracting unknown words from a Chinese text corpus is proposed in this paper. Instead of using a non-iterative segmentation-detection approach, the proposed method iteratively extracts the new words and adds them into the lexicon. Then the augmented dictionary, which includes potential unknown words (in addition to known words), is used in the next iteration to re-segment the input corpus until stop conditions are reached. Experiments show that both the precision and recall rates of segmentation are improved.

## 1. INTRIDUCTION

Unlike the English and other European language, Chinese text is a sequence of ideological representation without word delimiters. Therefore, whenever Chinese text processing is concerned, an extra word segmentation process is required. This segmentation process will be used for all Chinese text processing related fields such as machine translation, natural language processing and information retrieval. Most word identification approaches share one common algorithm: matching [1][2]. The basic strategy is to match the input character string with a large set of entries stored in a pre-compiled lexicon to find all (or part of) possible segmentations. Here the completeness of lexicon is crucial for the accuracy of segmentation. If a word/phase is not included in the lexicon, it will be segmented into shorter character/word sequences. But the word set of a natural language is open-ended. So the problem of unknown words is inevitable, especially when using a general lexicon in a specific domain. The best solution is to detect new words from the special corpus and add them into the lexicon. This is why we propose a bootstrap procedure: The first version of the lexicon contains only a general vocabulary that may lack specific domains words. When implementing our method, we enhance the lexicon after each iteration and re-segment the corpus with it. The accuracy of segmentation should be better than that of the previous version. The framework of our system is shown in Fig.1.

In section 2, we briefly introduce our segmentation module. In section 3, we describe the new words extraction standard. In section 4, we describe in detail the proposed method -- *a bootstrap method for Chinese new words extraction* and present experimental results. Finally, we present our conclusion in section 5.

## 2. SEGMENTATION MODULE BASED ON HEURISTIC RULES

Word segmentation algorithms find the most plausible segmentation. In our system, we implement the heuristic match method proposed in [3]. Complex Maximal Matching Rule is the main rule to match, and the other three rules to resolve ambiguities.

- *Complex Maximal Matching Rule:* It says that the most plausible segmentation is the three-word sequence with the maximal length. In segmentation, if an ambiguity does occur, the matching algorithm looks ahead two more words. If there is more than one chunk with maximum length, we apply the word length rule.
- *Word Length Rule:*

(1) Largest average word length rule picks the word from the chunk with largest average word length. It is more likely to encounter multi-character words than one-character words. If there are chunks with the same average length, we use the next rule.

(2) Smallest variance of word lengths rule picks the chunk with smallest variance of word lengths. This rule supposes that word length is usually evenly distributed.

If these two rules fail, we apply the probability rule.

- *Probability Rule:* The rule picks the chunk with the largest sum of morphemic freedom of one-character words. The frequency of occurrence of a character can serve as an index of its degree of morphemic freedom. So the formula used to calculate the sum of morphemic freedom is to sum log (frequency) of all one-character word(s) in a chunk.

Since it is very unlikely that two characters will have exactly the same frequency value, there should be no ambiguity after this rule has been applied.

## 3. NEW WORDS EXTRACTION STANDARD

According to [4], the segmentation units should have the following feathers:

(1) A string of characters that has a high frequency or high co-occurrence frequency among the components should be treated as a segmentation unit when possible;

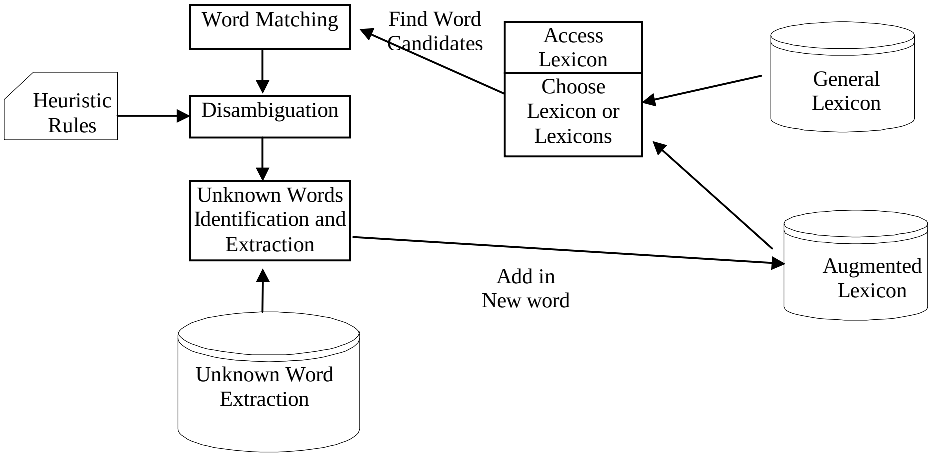

**Fig. 1**. The diagram of the bootstrap system

(2) Strings with overt segmentation markers should be segmented; and

(3) String with complex internal structure should be segmented when possible.

Based on this definition, we combine the following two features to detect new words.

- *Mutual information***:** MI is a useful criterion to evaluate the correlation of different components. The average length of Chinese words is of approximately 1.6 characters. Therefore, only word *bi-gra*m, *tri-gra*m, an*d quad-gram* in the corpus are of interest.

To *bi-gra*m: $MI(x,y) = \log\left(\frac{P(x,y)}{P(x)\times P(y)}\right)$

To *tri-gra*m: $MI(x,y,z) = \log\left(\frac{P(x,y,z)}{P_I(x,y,z)}\right)$

$$P_I(x,y,z) = P(x)P(y)P(z) + P(x)P(y,z) + P(x,y)P(z)$$

Where $P(\ )$ is the probability function. The estimation of MI of *quad- grams* is similar to that of tri-grams.

- *Entropy*: The entropy measure can be used to indicate the degree of randomness or uncertainty. It is defined by the probabilities of an event. For example, if $P(w_iW)$ and $P(Ww_i)$ represent the probability that the word $w_i$ occurs to the left and right of a word *"W*", the entropy of the set of left or right neighboring characters are

$$H_L(W) = \sum_{i=1}^{V_L} P(w_iW) \times \log P(w_iW), \sum_{i=1}^{V_L} P(w_iW) = 1$$

$$H_R(W) = \sum_{i=1}^{V_R} P(Ww_i) \times \log P(Ww_i), \sum_{i=1}^{V_R} P(Ww_i) = 1$$

Where $V_L$ and $V_R$ denote the numbers of all the possible words to the left and right of the word *"W*". The entropy measure reaches a maximum if all $P(\ )$ are equal, which implies that all words can appear to the left (or right) of *"W*" with equal probability. If most probability density are contributed by a few words, the entropy will be low, which means that only a few words can appear to the left (or right) of *"W*".

[5] proposes a criterion to take out plausible new words. It requests that only when the occurrence of a string and the entropy of its left and right neighboring characters are all high enough, can the string be considered as a potential new word.

In our system, we use both MI and Entropy indices. Entropy is used to filter out wrong extracted words. For example, in addition to extract "警戒力", "警戒", MI will take out "戒力" also. According to the entropies, the string "警戒力" is the most likely new word, "警戒" is next most likely and "戒力" is not words at all. Experimental results show that this combined measure can identify a large number of new words from a large corpus in a very short time.

# 4. EXPERIMENT

## 4.1. Procedure

Our experiment was carried out as follows:

(1) We use the general lexicon that is made up of two parts to segment the corpus. The first part of the lexiocn consists of 124,499 multi-character entries. This list is created by merging several Chinese word lists accessible on the Internet. The lengths of the lexical entries varies from two characters to eight characters. The second part consists of 13,060 characters and their frequency of usage. Character frequency was used in the last ambiguity resolution rule.

We segmented the corpus with the general lexicon. After that, new words were divided into several characters and domain-specific compound phrases were segmented into their constituent words.

(2) We calculated MI and entropy of the bi-gram, tri-gram, and quad-gram of neighboring words and sort them, which were all new word candidates.

| | | | | | |
|---|---|---|---|---|---|
| 梳化 | sofa | 罪案率 | case rate | 体育器材 | sports equipments |
| 豆奶 | soy milk | 玉兰树 | yulan tree | 不可否认 | incontestable |
| 狗粮 | dog’s food | 贵宾厅 | the hall for honored guest | 贩卖奴隶 | slave trade |
| 航机 | airplane | 环保式 | environmental protection style | 老奸巨滑 | a crafty old scoundrel |
| 熏香 | perfume | 归属感 | the feeling of ascription | 广播公司 | broadcast company |
| 恒河 | Ganges | 泰姬陵 | the tomb of the Thailand’s imperial concubine | 四大金刚 | four Guardians |
| 卡其 | khaki | 德鲁士 | Druses | 热血救国 | save the nation with passion |
| 趣致 | sentiment | 陈群娣 | Qundi Chen | 四肢百骸 | all limbs and bones |
| 浅窄 | shallow and narrow | 赵香珠 | Xiangzhu Zhao | 格子头巾 | checked muffler |
| 忿慨 | indignation | 科威特 | Kuwait | 爱克斯光片 | X-ray |
| 劫杀 | foray | 马可勃罗 | Marco Polo | 公益图书馆 | commonwealth library |
| 路易 | Louis | 人声沸腾 | noisy and confused | 格林威治村 | Greenwich village |
| 两伊 | Iran and Iraq | 蒸气熨斗 | steam iron | 连环谋杀案 | serial murder cases |
| 小周 | Zhou | 大好前途 | splendid outlook | 末期爱滋病 | terminal stages of AIDS |
| 木兰 | Ms. Mu Lan | 劳碌奔波 | work hard | 大英博物馆 | British museum |
| 琥珀色 | amber | 银行户口 | account | 十字架项链 | cross necklace |
| 氧气筒 | air tank | 异香扑鼻 | strong sweet-scent | 鞋厂负责人 | manager of a shoe factory |
| 浅褐色 | sandy beige | | | 圣地亚哥动物园 | zoo in San Diego |

**Fig. 2.** A portion of new words from the first iteration

| | | | | | |
|---|---|---|---|---|---|
| 八爪鱼 | octopus | 半途劫杀 | foray in midway | 红白格子头巾 | red and white checked muffler |
| 木兰从军 | Ms. Mu Lan enlisted | 路易十三 | Louis 13th | 贩卖奴隶行为 | the deeds of slave trade |
| 两伊战争 | the war between Iran and Iraq | 恒河三角洲 | Ganges delta | 体育器材店铺 | sports equipments store |
| 卡其军裤 | khaki military trousers | 琥珀色香料 | amber flavor | | |
| | | 英国广播公司 | BBC | | |

**Fig. 3**. A portion of new words from the second iteration

| n-gram | Iteration number | P(%) | R(%) | WPR | FM |
|---|---|---|---|---|---|
| 2 | 1 | 63.83 | 70.77 | 67.30 | 67.12 |
| | 2 | 65.88 | 76.67 | 71.28 | 70.87 |
| | 8 | 67.76 | 78.21 | 72.99 | 72.61 |
| 3 | 1 | 25.33 | 76.48 | 50.91 | 38.06 |
| | 2 | 27.75 | 77.32 | 52.54 | 40.84 |
| | 8 | 28.63 | 80.42 | 54.53 | 42.23 |
| 4 | 1 | 34.48 | 90.24 | 62.36 | 49.90 |
| | 2 | 36.77 | 91.46 | 64.12 | 52.45 |
| | 8 | 36.97 | 91.98 | 64.48 | 52.74 |

**Tab. 2**. Performance of new word identification (Convergence is reached at iteration 8)

(3) Instead of using absolute thresholds for unknown words extraction, we used a *relative mode* of filtering to extract only the most likely 10% new since the best thresholds couldn’t be reliably estimated before hand.

(4) We divided the output from (3) into lists of bi-gram, tri-gram, etc, and added them into the lexicon.

(5) Use the refined lexicon to re-segment the corpus. Then go to (2).

We iterated this procedure until a stop condition is satisfied.

## 4.2. Stop Condition

Two commonly used measures for unknown word detection are defined as follows:

$$P = \frac{number\ of\ correctly\ detected\ new\ words}{number\ of\ truly\ new\ words\ in\ the\ corpus}$$

$$R = \frac{number\ of\ correctly\ detected\ new\ words}{total\ number\ of\ detected\ new\ words}$$

The precision and recall rates are, in many cases, two contradictory performance indices. When one performance rate is raised, the other rate might be degraded. To make a fair comparison, the *weighted precision-recall* (WPR), which reflects the average of these two indexes, is proposed here to evaluate the joint performance of precision and recall:

$$WPR\left(w_p, w_r\right) = w_p * P + w_r * R\left(w_p + w_r = 1\right)$$

where $w_p$ , $w_r$ are weighting factors for precision and recall respectively. The *F-measure* (FM) [6], defined as follows, is another joint performance metric that allows lexicographers to weigh precision and recall differently:

$$FM\ (\beta) = \frac{(\beta^2 + 1) * P * R}{\beta^2 P + R}$$

where $\beta$ encodes users' preference for precision or recall. Our stop condition is based on these two indices. We used human evaluation to help the calculation of them. When the changes of them were small, we stop the iteration.

## 4.3. Experiment result

A portion of the list of the new words after the first and second iterations are shown in Fig2 and Fig3 respectively. We can find that most of the new words from the first iteration are entity names and idioms. Some of the new words detected from the second iteration are based on the words from the first iteration.

We use human evaluation to assess the performance of our method. Table 2 shows the metrics in the first two iterations and the 8th iteration.

Table 2 shows that the recalls for the embedded new words are as high as 79%, 80% and 92%, respectively, for 2-, 3-, 4-gram after 8 iterations. So we can conclude that about (79%-92%) new words are included in the extracted lists. Figure 3 shows the change in precision and recall for bi-gram words in each iteration. The performance increases almost monotonically. Therefore, the proposed method provides a way to stably improve precision and recall without offsetting the performance of each other.

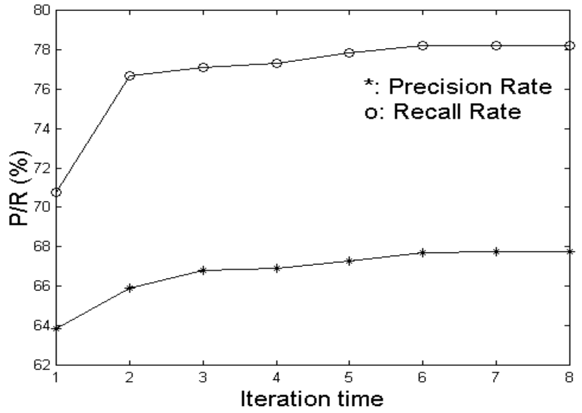

**Fig. 3**. Performance of identifying bi-gram new words in each iteration

( $w_p = w_r$ = 0.5, $\beta$ =1)

## 5. CONCLUSIONS AND FUTURE WORK

The iterative word segmentation and new word extraction approach proved to be an effective and easy method of improving accuracy of segmentation. Its advantages are as follows:

a. The patterns of new words are complicated and numerous. We don't need to hand-code each pattern, yet most high frequency words are extractable from the corpus.

b. The lexicon is self-incremental and general lexicon can be adapted for special domain use.

c. It is easy to control the balance between the precision and the recall of the detection algorithm.

Based on the performance of our method, we suppose that it can even work without the original general lexicon. Beginning from the n-gram statistics of the corpus, it can generate the primary lexicon for further processing. The issue is left for further experiments.

## 6. REFERENCES

[1] Chen, K.J. & S.H. Liu, 1992,"Word Identification for Mandarin Chinese Sentences," Proceedings of 14th Coling, pp. 101-107.

[2] Lee, H. J. & C .L .Yeh, 1991, "Rule-based Word Identification for Mandarin Chinese Sentences- A Unification Approach," Computer Processing of Chinese and Oriental Languages, Vol. 5, No. 1, 97-118.

[3] Tsai, C. H. (1996), "MMSEG: A word identification system for Mandarin Chinese text based on two variations of the maximum matching algorithm," Unpublished manuscript, University of Illinois at Urbana-Champaign.

[4] Huang, C.R., K.J. Chen, & Li-Li Chang, 1997, "Segmentation Standard for Chinese Natural Language Processing," International Journal of Computational Linguistics and Chinese Language Processing, 47-62.

[5] Tung, Cheng-Huang and Hsi-Jian Lee, "Identification of Unknown Words from Corpus," Computer Processing of Chinese & Oriental Languages, Vol. 8, pp. 131-145, Dec. 1994.

[6] Hobbs, Jerry R. "FASTUS: A Cascaded Finite-State Transducer for Extracting Information from Natural-Language Text," Proc. of ROCLING IX, pp. 199-231, Natl. Cheng-Kung Univ., Tainan, Taiwan, Aug. 1996.

[7] Chang, J.S. & Su, K.Y., "An Unsupervised Iterative Method for Chinese New Lexicon Extraction".