A Study on Multi-word Extraction from Chinese Documents

Wen Zhang[1], Taketoshi Yoshida[1], and Xijin Tang[2]

[1] School of Knowledge Science, Japan Advanced Institute of Science and Technology, 1-1 Ashahidai, Tatsunokuchi, Ishikawa 923-1292, Japan
{zhangwen, yoshida}@jaist.ac.jp
[2] Institute of Systems Science, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100080, P.R. China
xjtang@amss.ac.cn

Abstract. As a sequence of two or more consecutive individual words inherent with contextual semantics of individual words, multi-word attracts much attention from statistical linguistics and of extensive applications in text mining. In this paper, we carried out a series studies on multi-word extraction from Chinese documents. Firstly, we proposed a new statistical method, augmented mutual information (AMI), for words' dependency. Experiment results demonstrate that AMI method can produce a recall on average as 80% and its precision is about 20%-30%. Secondly, we attempt to utilize the variance of occurrence frequencies of individual words in a multi-word candidate to deal with the rare occurrence problem. But experimental results cannot validate the effectiveness of variance. Thirdly, we developed a syntactic method based on lexical regularities of Chinese multi-word to extract the multi-words from Chinese documents. Experimental results demonstrate that this syntactical method can produce a higher precision on average as 0.5521 than AMI method but it cannot produce a comparable recall. Finally, the possible breakthrough on combining statistical methods and syntactical methods is shed light on.

Keywords: multi-word extraction, word dependency, mutual information, augmented mutual information, syntactical method.

1 Introduction

A word is characterized by the company it keeps [1]. That means not only an individual word but also the context of this word should be laid on great emphasis for further textual processing. This simple and direct motivation drives the researches on multi-word which is anticipated to capture the context information from documents. Although multi-word has no satisfactory formal definition, it can be defined as a sequence of two or more consecutive individual words, which is a semantic unit, including steady collocations (proper nouns, terminologies, etc.) and compound words. Usually, it is made up of a group of individual words and its meaning is either changed to be entirely different from (e.g. collocations) or derived by the straightforward composition of the meanings of its parts (e.g. compound words).

Y. Ishikawa et al. (Eds.): APWeb 2008 Workshops, LNCS 4977, pp. 42–53, 2008.
© Springer-Verlag Berlin Heidelberg 2008

In fact, there are some overlapping between multi-word, collocation, terminology and similar concepts to describe the unique lexical unit in natural language. For this reason, the definition of multi-word is varied according to different purposes [3, 6, 10, 12], while the fundamental idea behind these concepts is the same, that is, to find lexical term that is more meaningful and descriptive than individual word.

Generally speaking, there are mainly three types of methods developed for multi-word extraction. The first one is the linguistic methods which utilized the structural properties of phrases and sentences to extract the multi-words from document. The second one is the statistical methods based on corpus learning for word pattern discovery from documents. The third one is to combine both the linguistic methods and statistical methods.

As for the linguistic methods, Justeson and Katz analyze the grammatical structure of terminology and propose an algorithm for terminological multi-word identification from English texts [2]. Their method regulates the terminology using a regular expression. It is reported that the method can obtain coverage as 97% and at least 77% precision in noun multi-word identification, and 67% noun multi-words are conformed to the regulation given by them. Similar work can also be found in [3, 4].

Statistical methods mostly employed a series of statistical variables on words' frequency and words' position are proposed to measure the possibility of a word pair to be a multi-word[1]. For instance, Smadja used the relative offset of two words' positions occurring in a corpus to determine whether or not they constitute a multi-word [5]. His basic idea of multi-word is that the offset of two words' positions should be a uniform distribution if these two words can not constitute a multi-word. And he reported that this method is quite successful in terminological extraction with an estimated accuracy as 80%. Similar work can also be referred to [6, 9, 14 and 21].

In the aspect of combining the linguistic knowledge and statistical computation, Chen et al use the co-related text segments existing in a group of documents to identify the multi-word terms from traditional Chinese documents [7]. Chinese stop-list is utilized to split the whole sentence into text segments and the statistical measure derived from term frequency and document frequency is used to weight the text segments to determine whether or not longer segments should be further split into short segments as multi-word terms. Their method is surprisingly successful in multi-word extraction from traditional Chinese documents as they declare that their method can obtain a minimum recall as 76.39% and a minimum precision as 91.05%. However, the performance of their method is determined by the quality of the stop-lists specific for the target texts. Park et al combine the linguistic method and statistical method for domain specific glossary extraction in [8]. The linguistic method is used to produce the candidate items and the statistical method is used for multi-word ranking and selection. Similar work can also be found in [11, 15 and 20].

Our contribution in this paper includes three aspects. Firstly, we proposed AMI to measure the words' dependency with goal to cope with the two problems in MI as unilateral co-occurrence and rare occurrence. Secondly, we investigate the effect of

[1] Here we just talk about word pair, i.e. bi-gram, because if a method is validated for word pair, it can be accordingly extended to the case of more than two words.

variance on multi-word extraction. Previous work is invested to use predefined threshold for word frequency to cope with the rare occurrence problem but ours is to make use of the variance of words frequencies in a candidate as an alternative solution. Thirdly, we developed a linguistic method for multi-word extraction using syntactic patterns of multi-words.

2 Mutual Information, AMI, Variance and the Syntactic Method

In this section, MI is reviewed in order to present its two deficiencies for dependency measure of word pair. Then AMI is proposed to deal with the deficiencies of MI. Especially, variance is attempted to attack the problem of rare occurrence. And the proposed syntactic method is specified.

2.1 Mutual Information

Church and Hanks propose the association ratio for measuring word association based on the information theoretic concept of mutual information [9]. In their method, the MI between word x and y was defined as Eq.1.

$$I(x, y) = \log_2 \frac{P(x, y)}{P(x)P(y)} \tag{1}$$

$P(x)$ is the occurrence probability of term x and $P(y)$ is the occurrence probability of word y in a corpus.

Sproat and Shih develop a purely statistical method using MI to determine the word boundary in Chinese characters [10]. Their algorithm is very successful for word extraction from Chinese text but the limitation of their method is that it can merely deal with words of length with two characters. Yamamoto and Church combine residual inverse document frequency (RIDF) and MI to conduct the word extraction from Japanese text collection and they report that the substrings with higher RIDF and higher MI are more possible to be a Japanese word [11]. Kita et al compare the competence of MI and cost criteria in multi-word extraction from Japanese and English corpus [12]. Their study demonstrates that mutual information tends to extract task-dependent compound noun phrases, while the method of cost criteria tends to extract predicate phrase patterns. Boxing Chen et al use MI to compute the association score of single word pair to automatically align the bilingual multi-word units from parallel corpora of Chinese and English [13]. Their experimental results demonstrate that the performance of MI was varying with different lexicons because not all the source words have their corresponding target phrase in another language but it provided a basis for constructing a translation lexicon in which the source language and the target language are both multi-word phrases. Jian and Gao propose a method based on MI and context dependency for compound words extraction from very large Chinese Corpus and they report that their method is efficient and robust for Chinese compounds extraction [14]. However, there are too many heuristics involved to determine the context dependency and the parameters are difficult to control to obtain a robust performance.

The primary reason of applying MI for multi-word extraction is that it has the support from both information theory and mathematic proof. If word x and word y are independent form each other, i.e. x and y co-occurred by chance, $P(x,y) = P(x)P(y)$, so $I(x, y) = 0$. By analogy, $I(x, y) > 0$ if x and y are dependent of each other. The higher MI of a word pair, the more genuine is the association between two words.

MI has some inherent deficiencies in measuring association. One is the unilateral co-occurrence problem, that is, it only considers the co-occurrence of two words while ignoring the cases that when one word occur without the occurrence of another. In this aspect, Church and Gale provide an example of using MI to align the corresponding words between French word "chambre", "communes" and English word "house" [15]. The MI between "communes" and "house" is higher than "chambre" and "house" because "communes" co-occurred with "house" with more proportion than "chambre" with "house". But the MI does not consider that more absence is with "communes" than "chambre" when "house" occurred. So it determined the incorrect "communes" as the French correspondence of English word "house". The other is concerned with the rare occurrence problem [16]. As is shown in Eq.1, when we assume that $P(x)$ and $P(y)$ are very small value but $I(x, y)$ can be very large despite of the small value of $P(x, y)$ in this situation. That means the dependency between X and Y is very large despite that X and Y co-occur very small times.

In order to compare with AMI method, the traditional MI method is employed to extract multi-words from a Chinese text collection. Usually, the length of the multi-word candidate is more than two, so we need to determine at which point the multi-word candidate can be split into two parts in order to use the traditional MI formula to score the multi-word candidate. To solve this problem, all the possible partitions are generated to separate a multi-word candidate into two parts, and the one which has the maximum MI score is regarded as the most appropriate partition for this multi-word candidate. Although some practical methods are suggested to extract the multi-word in [12], the method of maximum MI score employed here is different from them, because they are bottom up methods from individual words to multi-words, and our method is a top-down method from multi-word candidate to individual words or other smaller multi-word candidates. However, essentially, they have same back principle, i.e., to split the multi-word candidate into two components, and use MI to rank the possibility of its being a multi-word. For a multi-word candidate as a string sequence $\{x_1, x_2, ..., x_n\}$, the formula for computing its MI score is as follows.

$$MI(x_1, x_2, ..., x_n) = \underset{1 \le m \le n}{Max} \{\log_2 \frac{P(x_1, x_2, ...x_n)}{P(x_1, ..., x_m)P(x_{m+1}, ..., x_n)}\} \quad (2)$$

where m is the breakpoint of multi-word which separates $\{x_1, x_2, ..., x_n\}$ into two meaningful parts, $(x_1, ..., x_m)$ and $(x_{m+1}, ..., x_n)$. Moreover, we can determine whether or not $(x_1, ..., x_m)$ and $(x_{m+1}, ..., x_n)$ are two meaningful words or word combinations by looking up the single word set and multi-word candidate set we established in the previous

step. With the maximum likelihood estimation, $P(x_1, x_2, ..., x_n) = F(x_1, x_2, ..., x_n) / N$ (N is the total word count in the corpus), so the MI method can be rewritten as follows.

$$MI = \log_2 N + \max_{1 \le m \le n} \{\log_2 F(x_1, x_2, ..., x_n) - \log_2 F(x_1, ..., x_m) - \log_2 F(x_{m+1}, ..., x_n)\} \quad (3)$$

The traditional MI score method for the multi-word candidate ranking in this paper is based on Eq.3.

2.2 Augmented Mutual Information

To attack the unilateral co-occurrence problem, AMI is proposed and defined as the ratio of the probability of word pair co-occurrence over the product of the probabilities of occurrence of the two individual words except co-occurrence, i.e., the possibility of being a multi-word over the possibility of not being a multi-word. It has the mathematic formula as described in Eq.4.

$$AMI(x, y) = \log_2 \frac{P(x, y)}{(P(x) - P(x, y))(P(y) - P(x, y))} \quad (4)$$

AMI has an approximate capability in characterizing the word pair's independence using MI but in the case of word pair's dependence with positive correlation, which means that the word pair is highly possible to be a multi-word, it overcome the unilateral co-occurrence problem and distinguish the dependency from independency more significantly. To attack the problem of rare occurrence problem, we defined the AMI for multi-word candidate more than two words as Eq.5.

$$AMI(x, y, z) = \log_2 \frac{P(x, y, z)}{(P(x) - P(x, y, z)(P(y) - P(x, y, z))(P(z) - P(x, y, z))} \quad (5)$$

In practical application for a sequence $(x_1, x_2, \ldots, x_n)$, $P(x_1, x_2, \ldots, x_n) = p$, $P(x_1) = p_1$, $P(x_2) = p_2$, …, $P(x_n) = p_n$, we have

$$AMI(x_1, x_2, ..., x_n) = \log_2 \frac{p}{(p_1 - p)(p - p_2)...(p - p_n)} \quad (6)$$

By maximum likelihood estimation,

$$\begin{aligned} AMI(x_1, x_2, ..., x_n) &= \log_2 \frac{p}{(p_1 - p)(p_2 - p)...(p_n - p)} = \log_2 \frac{F / N}{(F_1 - F)(F_2 - F)...(F_n - F) / N^n} \\ &= \log_2 \frac{N^{n-1} F}{(F_1 - F)(F_2 - F)...(F_n - F)} = (n-1)\log_2 N + \log_2 \frac{F}{(F_1 - F)(F_2 - F)...(F_n - F)} \\ &= (n-1)\log_2 N + \log_2 F - \sum_{i=1}^{n} \log_2 (F_i - F) \end{aligned} \quad (7)$$

F is the frequency of $(x_1, x_2, .., x_n)$ and F_n is the frequency of x_n. N is the number of words contained in the corpus, it is usually a large value more than 10^6. In Eq.7, $\log_2 N$ actually can be regarded as how much the AMI value will be increased by when one more word is added to the sequence. It is unreasonable that $\log_2 N$ is a large value and it makes the AMI is primarily dominated by the length of sequence. In our method,

$\log_2 N$ is replaced by α which is the weight of length in a sequence. Another problem with Eq.7 is that in some special case we have $F_i = F$ and $F_i - F = 0$, these special cases would make the Eq.8 meaningless. For this reason, Eq.7 is revised to Eq.8.

$$AMI(x_1, x_2, ..., x_n) = (n-1)\alpha + \log_2 F - \sum_{i=1}^{m} \log_2 (F_i - F) + (\text{n - m})\,\beta \tag{8}$$

m is the number of single words whose frequency are not equal to the frequency of the sequence in the corpus. β is the weight of the single word whose frequency is equal to the frequency of the sequence. This kind of single word is of great importance for a multi-word because it only occurs in this sequence such as "Lean" to "Prof. J. M. Lean".

2.3 Variance

In our method, AMI is a primary measure for ranking the multi-word candidate. Besides AMI, the variance among the occurrence frequencies of the individual words in a sequence is also used to rank the multi-word candidate as the secondary measure. The motivation of adopting variance is that multi-word often occurs as a fixed aggregate of individual words. And the result of this phenomenon is that the variance of frequencies of individual words affiliated to a multi-word would be less than that of a random aggregate of individual words. This makes the variance could be used to attack the rare occurrence problem because the sequence with rare words relative to other frequent words in the same sequence will have a greater variance than those do not have. The variance of a sequence is defined as

$$\mathrm{V}(x_1, x_2, ..., x_n) = \frac{1}{n}\sum_{i=1}^{n}(F_i - \overline{F})^2 \quad \overline{F} = \frac{1}{n}\sum_{i=1}^{n} F_i \tag{9}$$

2.4 The Syntactic Method

The syntactic method we employed here is made from Justeson and Katz's repetition and noun ending. But it is not the same as their method because Chinese syntactic structure is different from that of English. Feng et.al proposed a method based on the accessor variety of a string to extract the unknown Chinese words from texts [17]. Actually, their method is frequency based because the left accessor variety and the right accessor variety are determined by the number of different characters at the position of head and tail of the string. They used the number of the characters of these two positions to conduct the word segmentation, i.e. delimit word from a string, on Chinese characters so that the consecutive meaningful segments are extracted as words. Their idea is to great extent similar with ours proposed here. But our method does not need an outsourcing dictionary to support adhesive character elimination because adhesive characters have limited influences on multi-word extraction as their short lengths. As the counterpart of repetition in Chinese, any two sentences in a Chinese text were fetched out to match their individual words to extract the same consecutive patterns of them. And we extracted the multi-words from the extracted repetitive patterns by regulating their end words as nouns. The algorithm developed to extract Chinese multi-words is as follows.

Algorithm 1. A syntactic method to extract the multi-word from Chinese text

```
Input:
  s1, the first sentence;
  s2, the second sentence;
Output:
  Multi-words extracted from s1 and s2;
Procedure:
  s1 = {w1,w2,…,wn} s2 = {w1',w2',…,wm'} k=0
  for each word wi in s1
    for each word wj' in s2
      while(wi = wj')
        k++
      end while
      if k>1
        combine the words from wi to wi+k' as the same con-
        secutive pattern of s1 and s2 as s3 = {w1'',w2'',…,
        wk+1''}
      End if
    End for
  End for
  p = |s3|;
  for word wp'' in s3
          if wp'' is a noun
            return {w1'',…,wp''} as the output of this
procedure;
          else p = p-1;
          end if
          if p is equal to 1
            return null as the output of this procedure;
          end if
  end for
```

3 Multi-word Extraction from Chinese Documents

In this section, a series of experiments were conducted with the task to extract the multi-words from Chinese documents to evaluate the proposed methods in Section 2. Basically, we divided the experiments as two groups: one is to evaluate the statistical methods as MI, AMI and Variance; the other is to evaluate the syntactic method.

3.1 System Overview of Multi-word Extraction Using MI, AMI and Variance

The multi-word extraction using statistical methods includes primarily three steps. The first step is to generate the multi-word candidate from text using N-gram method. The second step is to rank the multi-word candidates by statistical method, respectively.

The third step is to conduct multi-word selection at different candidate retaining level (clarified in Section 3.4). Figure 1 is the implementation flow chart for multi-word extraction from Chinese documents using MI, AMI and Variance, respectively.

3.2 Chinese Text Collection

Based on our previous research on text mining [18, 19], 184 Chinese documents from Xiangshan Science Conference Website (http://www.xssc.ac.cn) are downloaded and used to conduct multi-word extraction. The topics of these documents mainly focus on the basic research in academic filed such as nanoscale science, life science, etc so there are plenty of noun multi-words (terminologies, noun phrases, etc) in them. For all these documents, they have totally 16,281 Chinese sentences in sum. After the morphological analysis[2] (Chinese is character based, not word based), 453,833 individual words are obtained and there are 180,066 noun words.

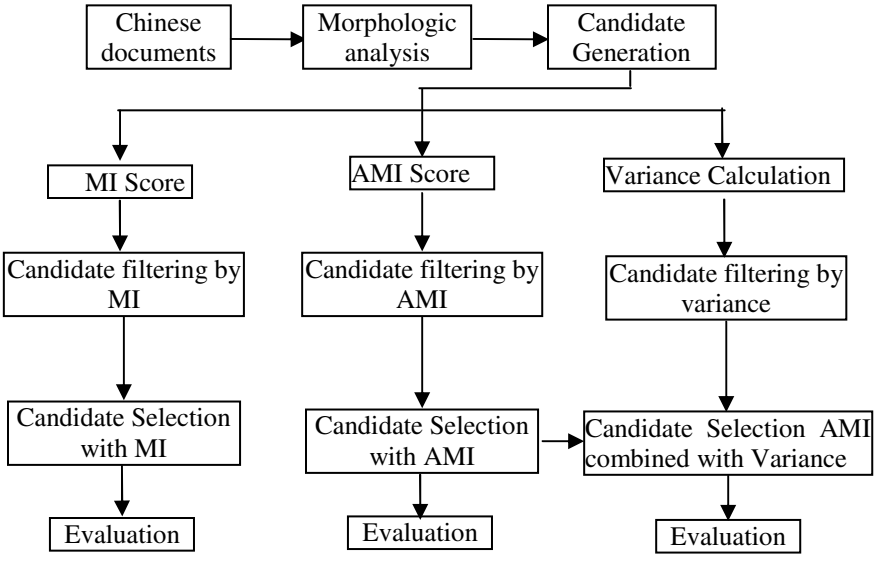

Fig. 1. Multi-word extraction from Chinese documents using the statistical methods MI, AMI and Variance, respectively

3.3 Candidate Generation

The multi-word candidates are produced by the traditional N-gram method. Assuming we have a sentence after morphological analysis as "A B C DE F G H." and H is found as a noun in this sentence, the candidates will be generated as "G H", "F G H", "E F G H", "D E F G H" and "C D E F G H" because multi-word usually has a length of 2-6 individual words.

[2] We conducted the morphological analysis using the ICTCLAS tool. It is a Chinese Lexical Analysis System. Online: http://nlp.org.cn/~zhp/ ICTCLAS/codes.html

Definition 1. Candidate Set is a word sequence set whose elements are generated from the same root noun in a sentence using n-gram method.

For example, "G H", "F G H", "E F G H", "D E F G H" and "C D E F G H" construct a candidate set generated from the root noun "H". At most only one candidate from a candidate set can be regarded as the exact multi-word for a root noun.

3.4 Requisites for Evaluation

The AMI formula in Eq.8 is used to rank the multi-word candidates. Here, α and β was predefined as 3.0 and 0. α is a heuristic value derived from our experiments on computing the AMI of all candidates. β is set to 0 as it contributes an unit in length to the candidate so that the AMI value of the sequence with this individual word will be greater than that of the sequence without this individual word. Also the variance of each candidate is calculated out for the secondary measure.

Definition 2. Candidate Retaining Level (CRL) regulates at what proportion the multi-word candidates with highest AMI are retained for further selection.

In order to match the multi-word given by our methods and the multi-word given by human experts, approximate matching is utilized.

Definition 3. Approximate matching. Assumed that a multi-word is retrieved from a candidate set as $m_1 = \{x_1, x_2, ..., x_p\}$ and another multi-word as $m_2 = \{x_1', x_2', ..., x_p'\}$ was given by human identification, we regard them as the same one if $\frac{|m_1 \cap m_2|}{|m_1 \cup m_2|} \geq \frac{1}{2}$.

The reason for adopting approximate matching is that there are certainly some trivial differences between the multi-word given by computer and human identification because human has more "knowledge" about the multi-word than computer such as common sense, background context, etc.

3.5 Multi-word Extraction Using MI, AMI and Variance

Multi-word extraction using MI employed Eq.3 to rank the candidates and AMI method employed Eq.8 to rank the candidates. It should be noticed that in the Variance method as shown in Fig 1, we used AMI as the first filtering criterion and variance as secondary measure. That is, if a candidate has the greatest AMI and the least variance in its candidate set concurrently, this candidate will be regarded as a multi-word. Otherwise, it will not be regarded as a multi-word and no multi-word comes out from this candidate set.

We varied the CRL for each method at different ratio as 70%, 50% and 30% so that the performances of the above three methods can be observed at dynamic settings. A standard multi-word base for all documents is established. 30 of 184 papers are fetched out randomly from text collection as test samples and the performances of examined methods are observed from them.

Table 1 shows the experimental results from MI, AMI and Variance, respectively. It can be seen that recall is decreasing while precision is increasing when CRL

declines from 0.7 to 0.3. The decrease of recall can be convincingly explained because fewer candidates are retained. And the increase on precision clarifies that multi-words actually have higher AMI than the candidates which are not the multi-words. On average, the greatest recall is obtained as 0.8231 at CRL 0.7 with AMI method and the greatest precision is obtained as 0.2930 at CRL 0.3 also with AMI method. This illustrates that AMI outperforms MI and Variance convincingly on all the parameter settings. The performance of MI and Variance are comparable on the whole: the precision of Variance is significantly higher than MI and the Recall of MI is better than MI method.

The motivation of variance is that the individual words of a fixed phrase which is a multi-word usually have a less variance in their occurrence frequencies. We reasonably speculate that the variance would improve the precision in multi-word extraction although the improvement of recall is not ensured. However, the experiment results did not validate our assumption of variance and the fact is that variance would reduce both the recall and the precision in multi-word extraction. We conjecture that multi-words may not have the same properties as that of the fixed phrases, that is, the individual words of a multi-word do not usually have the least variance among its candidate set although they have a low variance. For instance, assuming all individual words of a candidate are rare occurrence words, the variance of that candidate certainly is the least one in its candidate set. This candidate cannot be regarded as a multi-word because of its low occurrence.

Table 1. Performances of strategy one and strategy two on multi-word extraction from Chinese text collection at different CRLs. Av is the abbreviation of "average"; R is the abbreviation of "Recall"; P is the abbreviation of "Precision"; F is the abbreviation of "F-measure".

	MI			AMI			Variance		
CRL	Av-R	Av-P	Av-F	Av-R	Av-P	Av-F	Av-R	Av-P	Av-F
0.7	0.7871	0.2094	0.3272	0.8231	0.2193	0.3425	0.5621	0.2019	0.2913
0.5	0.5790	0.2174	0.3105	0.6356	0.2497	0.3515	0.3951	0.2317	0.2832
0.3	0.2652	0.2375	0.2419	0.3878	0.2930	0.3229	0.2040	0.2553	0.2160

3.6 Evaluation for the Syntactic Method

Table 2 is the evaluation results of the proposed syntactic method. It can be seen that the syntactic method can produce a higher precision than any one of the above statistic methods. However, the recall of this syntactical method cannot compete with the statistical methods. Furthermore, we should notice here that the F-measure of the syntactic method is also greater that any one of our statistical methods. That means the syntactic method can produce a more balancing recall and precision than those from statistical methods. In other words, the advantage of statistic method is that it can cover most of the multi-words in Chinese texts despite of its low precision but the syntactic method can produce a highly qualified multi-word extraction although its coverage is limited.

Table 2. Performance of syntactic method on multi-word extraction from Chinese text collection. Av is the abbreviation of "average".

Av-Precision	Av-Recall	Av-F-measure
0.3516	0.5520	0.4152

4 Concluding Remarks and Future Work

In this paper, we proposed three methods for multi-word extraction as AMI, Variance and a syntactic method. AMI is proposed to attack the two deficiencies inherent in MI. We pointed out that AMI has an approximate capability to characterize the independent word pairs but can amplify the significance of dependent word pairs which are possible to be multi-words. Variance was attempted to attack the problem of rare occurrence because individual words belonging to a multi-word may usually co-occur together and this phenomenon will make the variance of their occurrence frequencies very small. A syntactic method based on the simple idea as repetition and noun ending is also proposed to extract the multi-words from Chinese documents.

Experimental results showed that AMI outperforms both MI and Variance in statistical method. Variance cannot solve the problem of rare occurrence effectively because it can improve nether precision nor recall in multi-word extraction from Chinese documents. The syntactic method can produce a higher precision than the statistic methods proposed in this paper but its recall is lower than the latter. Based on this, we suggest that the performance of extraction could be improved if statistical methods are used for candidate generation and the linguistic method for further multi-word selection.

As far as the future work was concerned, the performance of multi-word extraction is still of our interest, that is, statistical and linguistic methods will be combined according their advantages in multi-word extraction. More experiments will be conducted to validate our hypotheses, especially on the solution of rare occurrence problem. Moreover, we will use the multi-words for text categorization and information retrieval, so that the context knowledge could be integrated into practical intelligent information processing applications.

Acknowledgments

This work is supported by Ministry of Education, Culture, Sports, Science and Technology of Japan under the "Kanazawa Region, Ishikawa High-Tech Sensing Cluster of Knowledge-Based Cluster Creation Project" and partially supported by the National Natural Science Foundation of China under Grant No.70571078 and 70221001.

References

1. Firth, J.R.: A Synopsis of Linguistic Theory 1930-1955. Studies in Linguistic Analysis. Philological Society. Blackwell, Oxford (1957)
2. Justeson, J.S., Katz, S.M.: Technical terminology: some linguistic properties and an algorithm for identification in text. Natural Language Engineering 1(1), 9–27 (1995)

3. Bourigault, D.: Surface Grammatical Analysis for the Extraction of Terminological Noun Phrases. In: Proceedings of the 14th International Conference on Computational Linguistics, Nantes, France, pp. 977–981 (1992)
4. Kupiec, J.: MURAX: A robust linguistic approach for question answering using an on-line encyclopedia. In: Proceedings of the Sixteenth Annual International ACM Conference on Research and Development in Information Retrieval, Pittsburgh, PA, USA, June 27 - July 1, 1993, pp. 181–190 (1993)
5. Smadja, F.: Retrieving collocations from text: Xtract. Computational Linguistics 19(1), 143–177 (1993)
6. Church, K.W., Robert, L.M.: Introduction to special issue on computational linguistics using large corpora. Computational Linguistics 19(1), 1–24 (1993)
7. Chen, J.S., et al.: Identifying multi-word terms by text-segments. In: Proceedings of the seventh international conference on Web-Age information Management Workshops (WAIMV 2006), HongKong, pp. 10–19 (2006)
8. Park, Y.J., et al.: Automatic Glossary Extraction: Beyond Terminology Identification. In: Proceedings of the 19th international conference on Computational linguistics, Taiwan, pp. 1–17 (2002)
9. Church, K.W., Hanks, P.: Word association norms, mutual information, and lexicography. Computational Linguistics 16(1), 22–29 (1990)
10. Sproat, R., Shinh, C.: A statistic method for finding word boundaries in Chinese text. Computer Processing of Chinese and Oriental Language 4(4), 336–351 (1990)
11. Yamamoto, M., Church, K.W.: Using suffix arrays to compute term frequency and document frequency for all substrings in a corpus. Computational Linguistics 27(1), 1–30 (2001)
12. Kita, K., et al.: A comparative study of automatic extraction of collocations from Corpora: mutual information vs. cost criteria. Journal of Natural Language Processing 1(1), 21–29 (1992)
13. Chen, B.X., Du, L.M.: Preparatory work on automatic extraction of bilingual multi-word units from parallel corpora. Computational Linguistics and Chinese Language Processing 8(2), 77–92 (2003)
14. Zhang, J., et al.: Extraction of Chinese Compound words: An experiment study on a very large corpus. In: Proceedings of the second Chinese Language Processing Workshop, HongKong, pp. 132–139 (2000)
15. Church, K.W., William, A.G.: Concordances for parallel text. In: Proceedings of the seventh Annual Conference of the UW Center for the New OED and Text research, Oxford, pp. 40–62 (1991)
16. Christopher, D.M., Hinrich, S.: Foundations of Statistical natural language processing, pp. 178–183. MIT Press, Cambridge (2001)
17. Feng, H.D., et al.: Accessor Variety Criteria for Chinese Word Extraction. Computational Linguistics 30(1), 75–93 (2004)
18. Zhang, W., Tang, X.J., Yoshida, T.: Web text mining on A Scientific Forum. International Journal of Knowledge and System Sciences 3(4), 51–59 (2006)
19. Zhang, W., Tang, X.J., Yoshida, T.: Text classification toward a Scientific Forum. Journal of Systems Science and Systems Engineering 16(3), 356–369 (2007)
20. Daille, B., et al.: Towards Automatic Extraction of Monolingual and Bilingual Terminology. In: Proceedings of the International Conference on Computational Linguistics, Kyoto, Japan, August 1994, pp. 93–98 (1994)
21. Fahmi, I.: C Value Method for Multi-word Term Extraction. Seminar in Statistics and Methodology. Alfa-informatica, RuG, May 23 (2005), http://odur.let.rug.nl/fahmi/talks/statistics-c-value.pdf