Extracting Chinese Multi-Word Units from Large-Scale Balanced Corpus

LIU Jianzhou, HE Tingting
Department of Computer Science
Central China Normal University
430079, Wuhan, China
xxzzsoft@sina.com hett@163.net

LIU Xiaohua
Center of Modern Educational Technology
Huazhong University of Science & Technology
430074, Wuhan, China
xx_liuxh@sohu.com

Abstract

Automatic Multi-word Units Extraction is an important issue in Natural Language Processing. This paper has proposed a new statistical method based on a large-scale balanced corpus to extract multi-word units. We have used two improved traditional parameters: mutual information and log-likelihood ratio, and have increased the precision for the top 10,000 words extracted through the method to 80.13%. The results of the research indicate that this method is more efficient and robust than previous multi-word units extraction methods.

1 Introduction

Natural language processing is a project based on knowledge, thus human's linguistic knowledge must be stored in the computer and the process of human's comprehending and producing languages be formalized before the computer commands human's linguistic potency. Since multi-word units (words formed with at least two characters) are the primary embodiment of semantics (Pinchuck 1977; Sager 1990), research on these words is the starting point for different natural language processing applications. Automatic multi-word units (MWUs) extraction has great theoretical and practical significance to such language information processing research as information indexing, machine translation, voice recognition, document classification as well as thesaurus compiling.

Presently, the rapid developments in different professional fields (e.g. computer science, medicine) mean continuous creation of new MWUs, and it is impossible to list them exhaustively in a lexicon. Therefore, automatic extraction of MWUs is a very important issue. Compared with western languages, as for Chinese there is no space between characters and words are hard to define, thus automatic Chinese MWUs extraction will surely confront even more difficulties.

In this paper, we have proposed a statistical method based on a large-scale balance corpus to realize the automatic extraction of MWUs. The goal is to extract sets of words with exact meaning from the corpus. Our method mainly consists of three phases (the first two phases include 3 steps respectively and the third phase includes 4 steps). First, select "seeds" (two character word) ready for extension; then extend these seeds at the front or back by K characters; finally, by comparing these parameters, determine which are MWUs. We have assessed the experiment data by measuring precision rates, and the result indicates that our method is more efficient and robust compared with other approaches.

The rest of this paper is structured as follows. In section 2, we describe in detail the method and all statistical parameters used. In section 3, we make a comprehensive analysis and just evaluation of the experimental data. In section 4, we outline the related works as well as their results. Finally, we give out conclusions and introduce part of our later research work in section 5.

2 Multi-word Units Extraction Algorithm

Our automatic MWUs extraction algorithm takes three phases. Two improved parameters are applied to measure the association ratio of adjacent characters.

2.1 Technique description

2.1.1 Introduction to mi_f and logL_f parameters

In this method, we use the parameters mi_f and logL_f to measure the association ratio between characters. These two parameters are improved from the mutual information and log-likelihood by Silva & Lopes (for detailed illustration please see Silva & Lopes (1999)). At present, the relatively common formula to calculate mutual information is:

$$mi(x,y) = \log(\frac{p(x,y)}{p(x)\cdot p(y)}) \qquad (3.1)$$

We think this formula only fits bigrams (two character word). It is hard to use this formula to deal with n-grams (n>2), because it is a knotty problem to divide n-grams effectively into x, y parts. Therefore, we will use the parameter mi_f improved by Silva & Lopes, and the calculation formula is defined as follows:

$$mi\_f(w_1...w_n) = \log(\frac{p(w_1...w_n)}{Avp}) \qquad (3.2)$$

Where

$$Avp = \frac{1}{n-1}\cdot\sum_{i=1}^{i=n-1} p(w_1...w_i)\cdot p(w_{i+1}...w_n) \qquad (3.3)$$

$w_1...w_n$ is an n-gram, $p(w_1...w_n)$ is the probability that occurs in the given corpus. Since we cannot directly calculate the probability $p(w_1...w_n)$, we are able to estimate it by applying MLE (Maximum Likelihood Estimation) method. The estimation formula is:

$$p(w_1...w_n) = \frac{f(w_1...w_n)}{N} \qquad (3.4)$$

$f(w_1...w_n)$ stands for the occurrence frequency of $w_1...w_n$ in the corpus. N stands for the number of words in the corpus.

Ted Dunning originally proposed the parameter log-likelihood ratio, and the formula was defined as follows: (for detailed illustration please see Dunning (1993))

$$-2\log\lambda = 2[\log L(p_1,k_1,n_1) + \log L(p_2,k_2,n_2) - \log L(p,k_1,n_1) - \log L(p,k_2,n_2)]$$

Where

$$p_1 = \frac{k_1}{n_1}, p_2 = \frac{k_2}{n_2}, p = \frac{k_1+k_2}{n_1+n_2}, k_1 = f(x,y), k_2 = f(-x,y), n_1 = f(x,*), n_2 = f(-x,*)$$

And

$$\log L(p,n,k) = k\log p + (n-k)\log(1-p)$$

In this method, we have obtained the parameter logL_f according to Silva & Lopes processing method, and the calculation formula is:

$$\log L\_f(w_1...w_n) = 2\cdot(\log L(\frac{kf1}{nf1},kf1,nf1) + \log l(\frac{kf2}{nf2},kf2,nf2) -$$

$$\log L(\frac{kf1+kf2}{nf1+nf2},kf1,nf1) - \log L(\frac{kf1+kf2}{nf1+nf2},kf2,nf2)) \qquad (3.5)$$

Where

$$kf1 = f(w_1...w_n),\ kf2 = Avy - kf1,\ nf1 = Avx,\ nf2 = N - nf1$$

Avx and Avy are respectively defined as：

$$Avx = \frac{1}{n-1} \cdot \sum_{i=1}^{i=n-1} f(w_1 ... w_i), \quad Avy = \frac{1}{n-1} \cdot \sum_{i=2}^{i=n} f(w_i ... w_n) \qquad (3.6)$$

2.2 Illustrations of MWUs Extraction Algorithm

Our algorithm is done in 3 phases. First, select the seeds for extension; then extend the seeds; finally, determine which extensions are MWUs. Here we are to introduce these phases in a detailed way.

2.2.1 Seeds ready for extension

The algorithm of selecting seeds is:

```
Input:   A Corpus L
Output:  Seeds list db_two
Step 1:  Collect all unigram frequencies and possible bigrams frequencies from L in
         DB
Step 2:  For all 4-grams w x y z in L, remove one count for x y in DB if
             - mi_f(x, y)<mi_f(w, x)-k or mi_f(x, y)<mi_f(y, z)-k
Step 3:  Set the lower limit of the parameters and filter DB.
```

In step 1, we build a database DB to store the frequency information of each adjacent pair of words and unigram in the corpus for the sake of calculating the values of mi_f and logL_f.

The purpose of Step 2 is for the consideration of phrase boundary. If these adjacent characters are divided by a phrase boundary, we will subtract 1 from its frequency rate. Give a 4-grams (w，x，y，z), if it satisfies one of the following two conditions:

mi_f(x, y)<mi_f(w, x)-k or mi_f(x, y)<mi_f(y, z)-k

we will suppose these exists a phrase boundary between x and y, and we should subtract 1 from the frequency of xy.

Step 3 is to select seeds. If the frequency of a bigrams and its parameter surpass the fixed threshold, we will select it as a seed.

2.2.2 Extension of Seeds

The main algorithm of the extension of seeds is:

```
Input:   Corpus L, seeds list db_two
Output:  n-grams list db_n (n = three, four, five or six)
Step 1:  For each seed in list db_two, extend the seed at the front or back by K
         characters
             First, extend seeds to 3-grams, stored the data in list db_three
Step 2:  Then extend other n-grams, until n = 2×(K＋1), store the data in list db_n
         respectively
Step 3:  Set the fixed threshold of the parameters respectively according to the
         number n, and filter db_n
```

In step 1 of the algorithm, we are extending the collected seeds at the front or back by K words (let K = 2). When extending 3-grams, the information we need includes the frequencies of these 3-grams, the values of mi_f, logL_f, sc and the id of the extended seeds.

For example, suppose we have a seed “经济”(economy), which occurred in a corpus in the following contexts:

➢ …面对现代的有计划的商品经济市场，除了应有点现代知识外…

then all possible extensions are: (经济，市), (经济市，场), (品，经济), (品经济,市), (品经济市，场), (商，品经济), (商品经济，市) and (商品经济市，场).

When we extend "经济" into 4-grams, (经济市，场) are possible to be extended, so we collect the frequency of the four characters and calculate the values of mi_f, logL_f of them and the id of seed ("经济") together with the value of sc (sc = logL_f(经济市场) －logL_f(经济市)).

2.2.3 Definition of MWUs

The algorithm to determine MWUs is:

```
Input:   n-grams list db_n，seeds list db_two
Output:  MWUs list M
Step 1:  Unite all list db_n to a list M
Step 2:  As for list M，order by id asce, logL_f desc
Step 3:  For each n-grams in M:
             Determine if it is MWUs.
             If it is MWUs,
                 isMWus = 1
             else
                 isMWUs = 0
Step 4:  Filter the list by field "isMWUs"
```

This part is used to select and output the MWUs found through this method. Step 3 is to deal with nested words. If an n-gram whose logL_f value is higher than other n-grams and not contained by others, we consider this n-gram is MWUs. For example, there are some records in our experiment data as follows:

Char	Cofreq	Mi_f	logL_f	Id_bigram	IsMWUs
苛捐杂税	17	14.507073547	167.59326860	668	1
苛捐杂	17	12.884431613	156.50126415	668	0
苛捐	17	13.015326921	144.10458893	668	0

(Table 1)

According to the above judgment method, we will consider "苛捐杂税" as a multi-word unit, while "苛捐杂" and "苛捐" are not.

3 Results, Evaluation and Discussion

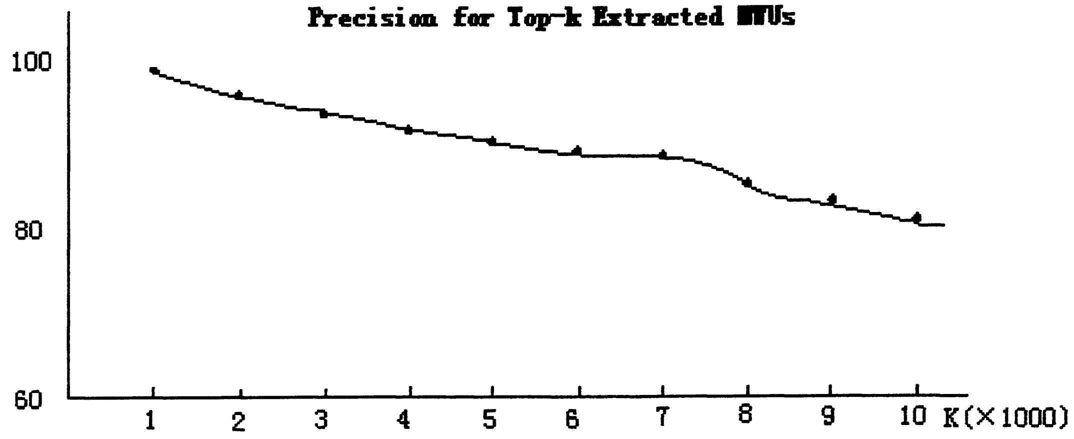
3.1 The corpus

At present, we mainly test our method with closed data. The corpus we used is the large-scale balanced corpus of Chinese Language Committee. The test data is the core part of the corpus, containing about 20 millions Chinese words.

3.2 The results

We have tested our method with the above corpus. For top 10,000 extracted MWUs, we achieved 80.13% precision and increased above 6% compared with the 74.4% precision achieved by Partick &

Dekang (2001). For the top 1,000 words extracted by our method, we achieved 97.60% precision. The detailed results are as follows:



(Fig. 1)

At the same time, we make a research on seeds extension and n-grams precision. We selected 2,000 seeds from the results, and selected the extracted MWUs extended by these seeds. We extracted 3,180 MWUs. The n-grams precision is as follows:

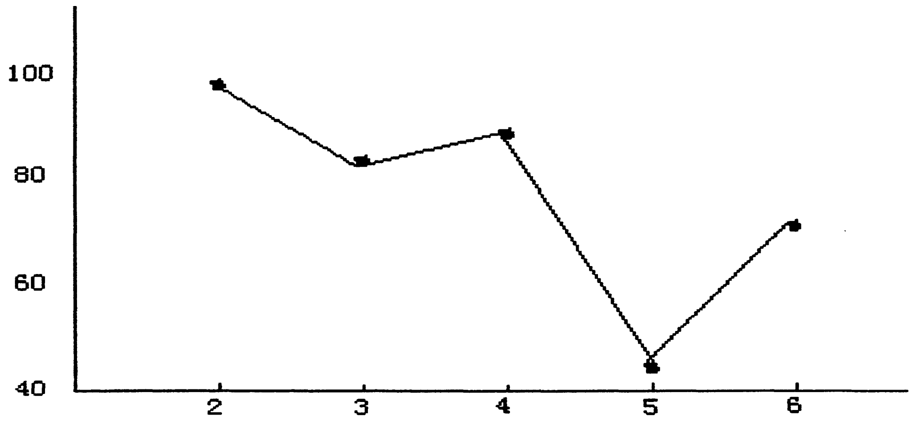
n-grams	2-grams	3-grams	4-grams	5-grams	6-grams	Average
Precision (%)	98.20	84.24	88.04	40.70	73.79	87.77

(Table 2)

From the table 2, we can see the average precision rate is much higher than Fung's n-grams average precision rate 54.09% (Fung 1998)

3.3 Analysis

From the results in 3.2, it is not hard to see that he 2-grams extraction proves to have the most ideal result, while 5-grams the worst. The precision change tendency is shown as follows:



(Fig. 2)

It is not difficult to see from the Fig.2 that the general tendency of the precision rate is descendence, and that of the precision rate of odd-grams is even sharper to form two valleys in the curve. This on one hand indicates that Chinese MWUs are primarily m-grams (m is even number), and on the other hand reveals that in this method certain parameters discriminating odd-grams are not the ideal.

During the process of extraction, we have discovered that many words begin with “的” (of), “和” (and), “是” (is) and “了” (int.), especially those in 5-grams. Customarily, these words are not MWUs, thus we can further filter them by means of lexical knowledge in order to raise the precision rate.

From the above experiment data, we can see that the proposed extraction method is quite successful. This is mainly because we have synthesized advantages of the two parameters: mutual information and log-likelihood ratio, and avoided their disadvantages. Generally, mutual information can well reflect the association degree of characters, but it also has its shortage in that it overestimates the function of low frequency words. The log-likelihood ratio is an efficient parameter to solve the problem. The disadvantage of log-likelihood ratio is that for those high frequency words that are rarely adjacent, its value turns out to be pretty high. This problem is solved by mutual information in some sense.

Also, selecting the two improved parameters is another main factor that contributes to the improvement of efficiency, which is shown in 3.1 and 3.2. We apply mi_f and logL_f because we have revealed certain shortages during the process of extraction with two traditional parameters: mutual information and log-likelihood ratio. The problem lies in that these two traditional parameters are hardly utilized in a just way in the process of seed extension. For example, when we calculate the value of mutual information of the 4-gram $w_1w_2w_3w_4$ by using the formula 3.1, it is a big problem how to divide $w_1w_2w_3w_4$ into two parts x and y. Theoretically, when divided, x and y should be words, but how to define x, y as two words itself is a problem to be solved. In Patrick & Dekang's term extraction algorithm, $w_1w_2w_3w_4$ is generally supposed to be divided into two words or terms (Patrick & Dekang 2001), yet this is very hard to realize, and with the increase of the length of n-grams, the division method of xy can be more various. So we think this is not the best solution. The improved parameters in our method fully solve the problem, and in fact the final results fully prove that this method is better than others.

4 Related works

Traditional approaches of MWUs extraction mainly used rules. However, not all MWUs can be created by rules, there are many words which are not created by rules (SUN Honglin 1998). Our method is mainly based on statistics. Several methods have been proposed for extracting MWUs from corpus by statistical approaches. In this section, we will briefly describe some of them.

Patrick & Dekang (2001) proposed a method based on statistics to automatically extracting domain specific terms from a segmented Chinese corpus. It contains about 10MB of Chinese news text. 10,268 terms extracted from that corpus, with the precision of 74.4%.

Ming-Wen Wu etc (1993) presented a method using mutual information and relative frequency. 9,124 multi-word units are extracted from the corpus, which consists of 74,404 words, with the precision of 47.43%. In this method, the MWUs extraction problem is formulated as classification problem. It also needs a training corpus to estimate parameters for classification model. In our method, we didn't make use of any training corpus. Another difference is that they use the method for English MWUs extraction while we extract Chinese MWUs in our experiments.

Fung (1998) presented a simple system for Chinese MWUs extraction-CXtract. CXtract uses predominantly statistical lexical information to find term boundaries in large text. Evaluations on the corpus consisting of 2 million characters show that the average precision is 54.09%.

5 Conclusion and future works

In our MWUs automatically extraction method, we use two parameters, mi_f and logL_f, deriving from mutual information and log-likelihood ratio. The results of our experiment show that our extraction method is successful. The precision of the top-10,000 MWUs extracted by our method reaches 80.13%, and the precision of the top-1,000 extracted MWUs reaches 97.06%. It is impossible to calculate the overall recall, so we only give the precision.

In future works, we will prepare a test on open data. Furthermore, we will do research on automatic term extraction. Considering the characteristics of term, we may extract MWUs from the corpus of professional fields compared with the MWUs from this balance corpus so as to ensure that the extracted MWUs are terms, instead of common words.

References

Dunning. T. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. Association for Computational Linguistics, 19(1)61-76 1993.

K. Church & K. Hanks. 1990. in Word Association Norms, Mutual Information and Lexicography. Computational Linguistics, 16(1):22-29.

Silva. J. & Lopes. G. 1999. A local Maxima Method and a Fair Dispersion Normalization for Extracting Multiword Units. In Proceedings of the 6th Meeting on the Mathematics of Language, p.369-381.

Patrick Pantel & Dekang Lin. 2001. A Statistical Corpus-Based Term Extractor. Canadian Conference on AI 2001. p.36-46

Fung. P. 1998. Extracting key term from Chinese and Japanese texts. The International Journal on Computer Processing of Oriental Language. Special Issue on Information Retrieval on Oriental Language. p.99-121.

Diana Maynard & Sophia Ananiadou. 1999. Identifying Contextual Information for Multi-Word Term Extraction.

K. T. Frantzi and S. Ananiadou. 1999. The C-Value/NC-Value domain independent method for multi-word term extraction. Journal of Natural Language Processing, 6(3):145-179.

Joana Paulo, Margarita Correia, Nuno J. Mamede & Caroline. 2002. Using Morphological, Syntactical and Statistical Information for Automatic Term Acquisition.

Kyo Kageura, Bin Umino. 1996. Methods of Automatic Term Recognition. Terminology, 3(2):259-289 1996.

A. Verdejo and Gonzalo J. 2001. Corpus-based Terminology Extraction applied to Information Access. Corpus Linguistics 2001; Lancaster, UK.

Smadja. Frank. 1993. Retrieving collocations from text: Xtract. Computational Linguistics, 19(1): 143-177.

Damerau. F. J. 1990. Evaluating Domain-Oriented Multi-Word Terms from Texts. Information Processing and Management 29(4), 433-447.

Luhn, H. P. 1957. A Statistical Approach to Mechanized Encoding and Searching of Literary Information. IBM Journal of Research and Development 2(2), 159-165.

Lin. D. 1998. Extracting collocations from text corpora. In Proceedings of COLING/ACL-98 Workshop on Computational Terminology. Montreal, Canada.

Salton. G, Yang. C. S and Yu. C. T. 1975. A Theory of Term Importance in Automatic Text Analysis. Journal of the American Society for Information Science 26(1), 33-44.

Sun. M., Shen. D. and Tsou B. K. 1998. Chinese Word Segmentation without Using Lexicon and Handcrafted Training Data. In Proceedings of COLING-ACL/98, P.1265-1271.

Cohen. J. D. 1995. Highlights: Language- and Domain-Independent Automatic Indexing Terms for Abstracting. Journal of the American Society for Information Science 46(3), 162-174.

Lee-Feng Chien. 1997. PAT-tree-based Keyword Extraction for Chinese Information retrieval. ACMSIGIR'97, Philadelphia, USA， p.50-58.

WU, Dekai and Xuanyin XIA. 1995. Large-scale Automatic Extraction of an English-Chinese Lexicon. Machine Translation 9(3-4), p.285-313.

Damerau, F. J. 1993. Evaluating Domain-Oriented Multi-Word Terms from Texts. Information Processing and Management 29(4). p.433-447.

Ming-Wen Wu and Keh-Yih 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. p.207-216.

Pinchuck, Isadore. 1997. Scientific and Technical Translation. Andre Deutsch.

Sager, Juan C. 1990. A Practical Course in Terminology Processing. John Benjamins B.V.

SUN Honglin and DUAN Huiming. 1998. Chinese Phrase Information Database about Natural Language Processing. Term Standardization and Information Technology (2).