# A Comparative Study of Collocation Extraction Methods from the Perspectives of Vocabulary and Grammar: A Case Study in the Field of Journalism

Gu Lulu

Department of Information Science, Beijing Language and Culture University 15<sup>th</sup> Xueyuan Road, Haidian District, Beijing 10083, China Pan Yue Department of Information Science, Beijing Language and Culture University 15<sup>th</sup> Xueyuan Road, Haidian District, Beijing 10083,

China

#### Liu Pengyuan\*

Department of Information Science, Beijing Language and Culture University Center of National Language Recourse and Monitoring Research Print Media Branch 15<sup>th</sup> Xueyuan Road, Haidian District, Beijing 10083, China liupengyuan@pku.edu.cn

gululu\_ghrise@126.com

viki.pan@qq.com

### Abstract

In all the natural languages that exist in the world, word collocation plays an important role. The methods of collocation extraction are diverse. On this basis, we compare a variety of commonly used collocation methods from the perspective of vocabulary and grammar, so that researchers can choose specific research methods upon specific occasions. This paper selects 16 extraction methods as a comparative analysis strategy. Based on the People's Daily corpus, an in-depth analysis is carried out on the basis of manual labeling accuracy. It is found that in the extraction of Chinese collocations, LMI, PSM, LR, T, F prefer recognition high frequency matching, OR, DICE, J, MS,  $\chi^2$ , Z, RR, PMI,  $\Delta P(y|x)$  and  $\Delta P(x|y)$  prefer low frequency matching, and the extraction method of high frequency group is better than low frequency group. MI<sup>3</sup> can extract both high frequency and low frequency collocations at the same time, and has the characteristics of two groups of methods at the same time, and its accuracy is the highest. Different collocation extraction methods have their own characteristics

## 1 Introduction

In all natural languages that exist in the world, word collocations appear frequently and play a pivotal role. In Chinese, because the words lacks and words is not restricted by the language form, so the study of collocation is particularly important.

With the rapid development of computer technology, the study of word collocation extraction plays an important role in the field of natural language processing. Collocation extraction refers to the automatic extraction of collocations from a corpus through the computing power of a computer and programming language, Such aspects as machine translation, word sense disambiguation, language generation and information retrieval are closely related to it. Therefore, the research on the automatic extraction method of word collocation is also very important.

There are many kinds of collocation extraction methods. Researchers have adopted different collocation extraction

when extracting different syntactic structures.

<sup>\*</sup> Corresponding Author

methods in many related studies, and the collocations obtained by different extraction methods are not the same, which will inevitably affect the research conclusions to a certain extent.

Some scholars(Xue Jing, Du Youfu, Huang Lan, 2016) conduct comparative studies based on word collocation extraction systems. They mainly compare and analyze commonly used extraction systems from three aspects in order to find a system suitable for specific situations but there is no systematic analysis of the comparison between specific collocation extraction methods. There is also a comparison of extraction methods based on English corpus (Zhu Xin, 2011), but there is no comparison based on Chinese corpus, and there are only comparisons of five extraction methods. There is also a five-tuple-based word collocation extraction method (Sun Tingting, 2015), but it is only a method based on mutual information and does not involve so many statistical methods.

Therefore, the current comparative study of Chinese collocation does not involve so many collocation methods, and the corpus based on is not large-scale either.

The existing traditional word colloca tion extraction methods are mainly divid ed into three parts: rule-based methods, statistics-based methods and methods co mbining statistics and rules. This paper i s mainly based on statistical methods. O n this basis, 16 collocation extraction me thods are selected. Based on common ne ws corpus, a collocation database for co mparison of collocation methods is establ ished, and the Chinese collocations extra cted by different collocation methods are investigated and analyzed. The main res earch questions in this paper are as follo -ws:

(1) What are the vocabulary and syntax

characteristics of Chinese collocations obtained by different collocation methods?

(2) In the extraction of Chinese collocations, what are the advantages and disadvantages of different collocation methods and what situations are applicable to each?

# 2 Related Work

In order to find a system suitable for a specific situation, domestic scholars (Xue Jing, Du Youfu, Huang Lan, 2016) compare and analyze commonly used collocation extraction systems from the three aspects of corpus source, extraction method and extraction results, but this only compared extraction systems from a macro perspective.

There are also attempts to compare the effects of four word pairing automati c extraction methods based on mutual in formation, chi-square, T-test and likeliho od\_ratio on different corpus types and c orpus lengths (Zhu Xin, 2011). The rese arch results show that the length of the corpus has different extraction efficiency for different extraction methods, so that suitable collocation extraction methods c an be selected for corpora of different si zes, but this is only the comparison bet

ween the four extraction methods. Compared with relevant domestic re search, the comparative research of forei gn collocation methods is more statistical ly based on the method of word associat ion measure(AM). First seen in Choueka (1988), extracting two or more adjective words from a corpus of 11 million wor ds (New York Times). Church & Hanks (1990) used mutual information in the s etting of window 5 to extract binary sho rt-distance(and long-distance) collocations, it was not until 1993 that the Xtract sys tem proposed by Smadja could start to extract two or more word collocations (c lose or long distance collocations).

Some scholars have proposed three stages of extraction (Agrawal et al., 2018). Two statistical extraction methods (PMI and DICE) are used to calculate the relevance score. Finally, a latent semantic analysis is used to calculate the similarity of the context to study the recognizability of the paper.

# 3 Extraction Method of Word Collocation

The 16 extraction methods are Loca 1 Mutual Information(LMI), MI<sup>3</sup>, Point Mutual Information(PMI), z\_score(Z), T-t est(T), fisher(F), chi\_square( $\chi^2$ ), likelihoo d\_ratio(LLR), Poisson-Stirling Approxima tion(PSM), Minimum Sensitivity(MS), Od ds\_Ratio(OR), Relative\_Risk(RR), Dice C oefficient(DICE), Jaccard Coefficient(J),  $\Delta P(y|x)$ ,  $\Delta P(y|x)$  etc. (The abbreviations i n parentheses are written below).

The above extraction methods are developed on the basis of co-occurring word frequencies and can be further divided into four categories: methods based on information theory, methods based on hypothesis testing and correlation strength coefficients, and two methods of directional correlation measurement included in  $\Delta P$ .

### 3.1 Information theory method

What the PMI value calculates is the probability information of another word that can be provided by the frequency of a word in the corpus. This indicator measures the correlation between two things, and judges the emotional tendency of a word by judging the mutual information of a word and x and y respectively, but the point of mutual information is only a correlation judgment on two of the points.

Mutual information has a greater

impact on low-frequency words. In order to reduce this effect, Daille (2008) proposed an improved mutual information algorithm to expand the gap between high-frequency words and low-frequency words, which is called MI<sup>3</sup>. The method is simple and efficient, but still cannot extract multi-word collocations.

LMI is also a kind of mutual information, which is used to evaluate the strength of collocations between collocations.

# 3.2 Methods of hypothesis testing

Hypothesis testing is a method in mathematical statistics to infer the population from a sample based on certain assumptions. At present, Z, T, F, $\chi^2$ , etc. can all be used for collocation extraction.

Berry-Roghe (1972) proposed that Z helps to understand the positional relationship between individual values and other values in the distribution. After the standardization process, researchers can also better compare individual values in two different variable distributions.

Church (1991) was introduced to the compilation of computational lexicons and was widely used. When using the T to determine the significance of a collocation word, first he made a null hypothesis: there is no connection between two co-occurring words and cannot constitute a collocation, and then use the standard deviation to measure whether the difference between the observed frequency and the expected frequency reaches a significant level.

The British statistician Fisher propos ed that the F is very sensitive to the no rmality of the data. If the two populatio ns have the same variance (homogeneity of variance), then the F test can be use d, but the test will show extreme non-ro bustness and abnormality.  $\chi^2$  judges whether words constitute a collocation by comparing the difference between the co-occurrence frequency and the expected frequency. It uses the co-occurrence words for statistical analysis, and evaluates the correlation between the target word and the collocation word by comparing the difference between the observation frequency and the expected frequency(Agresti, 1990).  $\chi^2$  test does not require the assumption of a normal distribution on the corpus, but it needs to ensure that the sample size is large enough.

Normal distribution and  $\chi^2$  distribution are not good for judging small probability events. Dunning (1993) proposed a log-likelihood ratio test method, which mainly compares the probability of the two hypotheses, and has a better extraction effect for sparse data.

PSM is another extraction method based on the Poisson distribution, and more accurate results can be obtained based on the Poisson distribution.

### **3.3** Correlation strength coefficient

The correlation strength coefficient is a method to measure the strength of collocation by evaluating the coefficient of the correlation strength from the observed data.

In the task of word collocation extraction, MS is more commonly used (Pedersen and Bruce, 1996), which uses the concept of conditional probability to compare the magnitude of two conditional probabilities and choose a smaller value.

OR is the ratio of the probability of an event occurring to the probability of the event not occurring (Blaheta & Johnson, 2001).

RR is a concept in medicine (Olga Kolesnikova, 2016). In medicine, it is the ratio of the risk of being exposed to the disease to the risk of not being exposed to the disease. This probability can also be used to measure the relationship between the probability that the two-tuple x y is a collocation and the probability that the two-tuple is a free combination

The dice coefficient is proposed by Daudaravicius (2010). Its performance is generally higher than that of other and the extracted correlation metrics, collocation set is better than other correlation metrics, but its measured correlation score is often smaller. In addition, the collocations detected by the DICE are often related to the collocation dictionary and do not depend on the size of the corpus.

Jaccard is a statistic used to compare the similarity and diversity of sample sets. The characteristic attributes of individuals are all identified by symbolic measurements or Boolean values. Therefore, it is not possible to measure the magnitude of the specific value of the difference, but only to obtain the result of "whether it is the same".

# 3.4 Orientation correlation measure

Gries (2013) proposed a new association detection method,  $\Delta P$  is the probability that the result word appears given a clue word minus the probability that the result word appears without the clue word. The reverse conditional probability  $\Delta P(y|x)\& \Delta P(x|y)$  increase targeted association measures, which can show this directional function.

# 4 Collocation Method Comparison Database Construction

The corpus randomly selected the data of the People's Daily from February, April, and June 2005 to 2015. The original data included 862,580 lines, 84,169,476 words, and 43,909,556 words after jieba<sup>1</sup> segmentation and part-of-speech tagging.

We call the NLTK module<sup>2</sup> in the python programming software, and use 16 extraction methods to extract collocations from the data file. The extraction window is set to [-1, 1], that is, the binary collocation pairs adjacent to the left and right are extracted, because most of the extraction methods are all suitable for nearest neighbors, but some extraction methods are not suitable for multivariate. Among them, the minimum co-occurrence frequency of two-tuples is 2, because Zipf's law shows that most of the data are low-frequency words.

Due to the large number of extracted candidate collocation results, it is impossible to screen all the candidate collocations. Therefore, in the extracted results, We select the first k results of each method for evaluation, and calculate the

accuracy at the k-th position Precision@k (P@k):

$$precision@k = \frac{Correct Collocation}{k}$$

Then in the results extracted in this paper, the first 2000 results of each method are selected for evaluation, and the top2000 of these 16 methods are processed, and 11497 candidate collocations are obtained after deduplication.

# 5 Comparative Analysis

# 5.1 Overall comparison of various collocation methods

We refer the definition to of collocation by Benson (1985), divide collocations into vocabulary collocations and grammatical collocations, and а evaluation is conducted manually:

(1) Lexical collocation: After syntactic analysis, node words and collocation words can be located under a parent node, and only these two component words are included under the node.

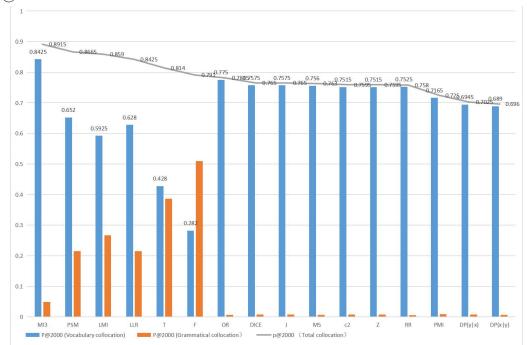


Figure1: p@2000 of each method

<sup>1</sup> https://github.com/fxsjy/jieba

<sup>2</sup> https://github.com/nltk

(2) Grammatical collocation: After syntactic analysis, node words and collocation words cannot be under the same parent node or only contain these two component words under the node. They are often a collocation composed of a function word and a non-function word.

From the perspective of the total accuracy of collocations, MI<sup>3</sup> performs the best and can reach an accuracy of 0.8915. Although the accuracy of  $\Delta P(x|y)$  is the lowest, it also has an accuracy of 0.696.

However, individually, except for F, which is better at identifying grammatical collocations and has the highest correct rate of 0.509 for grammatical collocations, various methods are better at identifying lexical collocations: OR, DICE, J, MS,  $\chi^2$ , Z, RR, PMI,  $\Delta P(y|x)$  and  $\Delta P(x|y)$  are basically composed with lexical collocations, and the recognition rate of grammatical collocations is almost close to zero.

# 5.2 The relevance of various collocation method

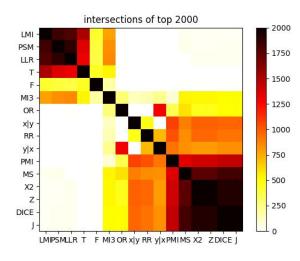


Figure2: Coincidence heat map of top2000 of various methods

From Figure 2 we can clearly see that the coincidence rate of these 16 extraction methods is divided into two major blocks: LMI, PSM, LLR, T, F have high coincidence rates, and they are clustered together. OR, DICE, J, MS,  $\chi^2$ , Z, RR, PMI,  $\Delta P(y|x)$  and  $\Delta P(x|y)$  converge on the other side, and there is almost no overlap between the two sides. But MI<sup>3</sup> is a more special one, which overlaps with both sides.

By observing the corpus, we find that OR, DICE, J, MS,  $\chi^2$ , Z, RR, PMI,  $\Delta P(y|x)$ and  $\Delta P(x|y)$  tend to extract collocations with lower co-occurrence frequency. In the corpus, more collocations of some proprietary words have been extracted, such as "Mopi Township//Xiaolugou", so more functional words are extracted, which can identify more lexical collocations. The overall recognition rate of grammar collocation is not high.

But LMI, PSM, LLR, T, F are the opposite. They prefer to recognize high-f requency collocations, so many candidate collocations containing non-function wor ds will be extracted. The use of non-fun ction words is very complicated and not as easy to distinguish as low-frequency terms. The recognition rate of its vocab ulary collocation is lower than the forme r, but the recognition rate of grammatica l collocation is higher.

Generally speaking, although there is not much difference, the methods that prefer high-frequency matching (LMI, PSM, LLR, T and F) are better than the low-frequency matching extraction methods (OR, DICE, J, MS,  $\chi^2$ , Z, RR, PMI,  $\Delta P(y|x)$  and  $\Delta P(x|y)$ ), among which the best high-frequency extraction methods are PSM and LMI, and the best extraction method in low-frequency is OR.

MI<sup>3</sup> can extract high-frequency and low-frequency collocations at the same time, and has the characteristics of two sets of methods at the same time, and its accuracy rate is the highest.

# 5.3 Comparative analysis based on syntactic structure

Combining the content word collocation framework to determine the seven categories (Wu Shuwen, 2019), this paper discusses, analyze and compare the nine most common and important syntactic structures n/n, v/n, n/v, v/v, d/v, a/n, d/a, m/n n/a. Extract the top2000 of each syntactic structure of each collocation extraction result to form 9 collocation sets, and filter these sets manually and calculate each extraction method accuracy rat.

Syntactic structure of collocation	Number of unions	Correct number	Correct rate
n/n	1062	974	0.917
a/n	1003	892	0.889
m/n	1040	888	0.854
d/a	824	674	0.818
v/n	1039	786	0.756
n/v	1054	792	0.751
n/a	992	672	0.677
v/v	1035	666	0.643
d/v	1008	454	0.450

Table1: The correct rate of various syntactic structure collocations after deduplication

As shown in the table, among these structures, the highest accuracy rate of n/n is 0.917. Except for the d/v structure, the accuracy rate is 0.450. The accuracy of other structures are all higher than 0.6. The correct rate indicates that these extraction methods are effective in terms of extracting word collocation.

#### 5.3.1 Noun structure

As shown in Figure 3, 16 extraction methods can effectively identify the noun structure, and the recognition is above 0.5 each time. The following is an analysis of each specific collocation.

Both the high-frequency extraction

method and the low-frequency extraction method can effectively identify the combination of n/n structure which has the highest correct rate of structure collocation.

For the combination of v/n and n/v, MI<sup>3</sup> is the most effective way to identify. In the low frequency group method, except for the OR extraction of v/n exceeding 0.8, the rest are all lower than 0.8. The lowest value is  $\Delta P(x|y)$ , which is close to 0.6. From this, it can be explained that compared to the n/n collocation, the recognition of the v/n collocation increases the difficulty.

The recognition accuracy of a/n is relatively highest. It can also be seen from the corpus that compared with the previous three structures, the number of a/n is relatively small, which can also be confirmed from Table 1. It is worth noting that the high-frequency group recognition is all correct, while the previous MI<sup>3</sup> with both types of extraction method features is not as accurate as the high-frequency group, indicating that it is a bit difficult for MI<sup>3</sup> to recognize a/n structure, although its recognition accuracy rate is still higher than the low-frequency group, but from Figure 3, it can be seen that all methods except  $\Delta P(x|y)$ are lower than 0.8. The accuracy rate of all methods is higher than 0.8. The a/n structure is still well recognized by various extraction methods.

In terms of identifying m/n, all methods are higher than 0.8 except for 0.685 of  $\Delta P(x|y)$ . Among them, PSM and LLR have the highest accuracy.

The accuracy of n/a is the lowest, the highest PSM is only 0.84, and the lowest  $\Delta P(y|x)$  is 0.69.

In the d/n combination, OR, DICE, J, MS,  $\chi^2$ , Z, RR, PMI,  $\Delta P(x|y)$  of the low frequency group are at least 0.87, which is better than the T and F of the high frequency group as a whole, which is a set

of syntactic structure collocations with low frequency better than high frequency.

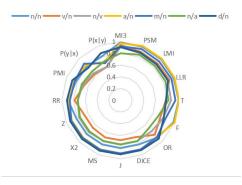


Figure3: The correct rate of each method to identify collocations containing nouns

### 5.3.2 Verb structure

The v/n and n/v structures have been discussed in 5.3.1, so they will not be repeated here.

From Figure 4, we can see that compared to the noun structure, the recognition accuracy of the verb structure fluctuates more. It can be seen that some extraction methods are difficult to recognize in this structure.

In v/v, there is also a group of syntactic structure collocations that low frequency is better than high frequency. DICE, J and MS are the highest, up to 0.925. The lowest in the low frequency group is  $\Delta P(y|x)$ , with a correct rate of 0.62, but this correct rate is also higher than 0.485 and

0.545 of T/F in the high frequency group.

The accuracy of d/v is generally low, indicating that this type of collocation is difficult to be correctly identified. The highest value is just over 0.5, which is 0.695 of DICE, J,  $\chi^2$  and Z. In comparison,  $\Delta P(y|x)$  and  $\Delta P(x|y)$  have the lowest correct rates, which are 0.275 and 0.365 respectively.

#### 5.3.3 Adjective structure

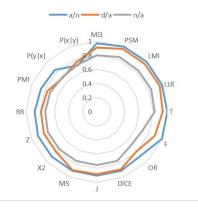


Figure 5: The correct rate of each method to identify collocations containing adjectives

The structures of a/n and n/a have been explained in the noun structure, so they won't be repeated here.

The d/a accuracy rate is slightly lower, and the highest combination is PSM and LLR, which is close to 1. The high frequency group is better than the low frequency group.



Figure4: The correct rate of each method to identify collocations containing verbs

#### 6 Conclusion

Based on the corpus of People's Daily, this paper evaluates 16 common classical statistical extraction methods, which can prove the advantages and limitations of different extraction methods.

The 16 extraction methods can be divided into three categories: high frequency group, low frequency group and MI<sup>3</sup>. PSM, LMI, LLR, T, F are biased towards extracting collocations with higher co-occurrence frequency, OR, DICE, J, MS,  $\chi^2$ , Z, RR, PMI,  $\Delta P(y|x)$  and  $\Delta P(x|y)$  are biased In order to extract collocations with higher co-occurrence frequency, MI<sup>3</sup> can extract collocations of high frequency and low frequency at the same time, and has the characteristics of two methods at the same time.

Generally speaking, the effect of the extraction method of the high-frequency group is better than that of the low-frequency group. Among them, the best high-frequency extraction methods are PSM and LMI, and the best low-frequency extraction method is OR.

For different syntactic structures, each extraction method has its own preference.

The recognition of the high-frequency group is biased towards the collocation of nouns and adjectives. PSM can recognize the structure of a/n, m/n, n/a, d/a and n/a, and it can be seen that it is sensitive to the collocation of adjectives. LMI and F can identify a/n collocation, the LLR method can effectively identify a/n, m/n and d/a collocations, and the T test is effective for identifying n/n and a/n.

The recognition of the low-frequency group is more biased towards verb collocations. From observations, OR has the highest recognition for verbs. DICE and J can effectively identify v/v and d/v MS collocations. for d/n and v/v collocations is better.  $\chi^2$  and Z can basically recognize d/v collocations.  $\Delta P(y|x)$  and  $\Delta P(x|y)$  have a somewhat low accuracy in recognizing all collocations, indicating that the recognition capabilities of these two methods are relatively weak.

 $MI^3$  can basically recognize all collocations effectively, and the correct rate of recognizing v/n and n/v collocations is high.

In the Chinese collocation study, after comparing the 16 extraction methods, the researcher can choose a specific collocation method upon a specific occasion, thereby improving the efficiency of recognition.

The problem of word collocation is very important, and related research will continue in the future. Future work will expand the size of the corpus and improve the accuracy of manual annotations to ensure the reliability of the research results.

### 7 Acknowledgments

This work was funded by the Humanities and Social Science Research Planning Fund of the Ministry of Education(18YJA740030) and Beijing Language and Culture University Chinese and Foreign Postgraduate Innovation Fund Project (21YCX171).

### References

- Xue Jing, Du Youfu, Huang Lan. A comparative study of word collocation extraction systems[J]. Modern Computer (Professional Edition), 2016(09):17-21.
- Zhu Xin. Comparative study on automatic extraction methods of word collocations [D]. Dalian Maritime University, 2011.

- Sun Tingting. Automatic extraction of word collocations based on five-tuples[J]. Electronic Design Engineering, 2015, 23(19): 75-78.
- Choueka, Y. (1988). Looking for Needles in a Haystack or Locating Interesting Collocational Expressions in Large Textual Databases. Computer-Assisted Information Retrieval (Recherche d'Information et ses Applications) - RIAO 1988, 2nd International Conference, Massachusetts Institute of Technology, Cambridge, MA, USA, March 21-25, 1988. Proceedings. DBLP.
- Kenneth Church and Patrick Hanks. Word association norms, mutual information and lexicography. Computational Linguistics, pages 22 - 29, 1990.
- Shaishav Agrawal. &Sudip Sanyal.(2018).
  Hybrid method for automatic extraction of multiword expressions. International Journal of Engineering & Technology, 7 (2.6) (2018) 33-38
- Boulaknadel S, Daille B & Aboutajdine D. (2008). A multiword term extraction program for Arabic language. In Proceedings of the 6th International Conference on Language Resources and Evalua-tion (LREC 2008), pp. 1485 1488
- Berry-Roghe, G. (1972). The Computation of Collocations and their Relevance in Lexical Studies [en linea].
- Kenneth Church and William Gale. Concordances for parallel text. In Proceedings of the 7th Annual Conference of the UW Center for the New OED and Text Research, Oxford, UK, 1991.
- Agresti, A. Categorical data analysis. John Wiley and Sons, New York. 1990

- Ted E. Dunning. Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, 19(1):61 - 74, 1993.
- Ted Pedersen. Fishing for exactness. In Proceedings of the South Central SAS User's Group Conference, pages 188 - 200, Austin, Texas, USA, 1996.
- Don Blaheta and Mark Johnson. Unsupervised learning of multi-word verbs. In Proceedings of the ACL Workshop on Collocations, pages 54 - 60, 2001.
- Olga Kolesnikova, P. (2016). Survey of Word Co-occurrence Measures for Collocation Detection. Computation y Sistemas 20(3):327-344
- Daudaravicius, V. . (2010). The Influence of Collocation Segmentation and Top 10 Items to Keyword Assignment Performance. International Conference on Computational Linguistics & Intelligent Text Processing. Springer-Verlag.
- Gries, S.T. (2013). 50-something years of work on collocations: what is or should be next. International Journal of Corpus Linguistics, Vol.18, No.1, pp.137-166. DOI: 10.1075/ijcl.18.1.09gri.
- Benson, M. . (1985). Collocations and idioms.In
  Roberr Ilson, editor,Dictionaries,
  Lexicography and Language Learning, pages 61–68. Pergamon, Oxford.
- Wu Shuwen. (2019). Research on information technology word collocation extraction method based on the cascade of multiple statistical methods (Master's thesis, Harbin Institute of Technology).