International Journal of Knowledge
and Language Processing
Volume **2**, Number 1, January **2011**

www.ijklp.org
KLP International ©2011 ISSN 2191-2734
pp. **53-58**

Collocation Extraction Using Square Mutual Information Approaches

Huarui Zhang[1], Yongwei Zhang[2] and Jingsong Yu[3]
[1]Institute of Computational Linguistics
Peking University
Beijing, China
hrzhang@pku.edu.cn

[2,3]School of Software and Microelectronics
Peking University
Beijing, China
[2]zhangywibb@gmail.com, [3]yjs@ss.pku.edu.cn

Received December 2010; revised January 2011

ABSTRACT. *MI (Mutual Information) has been proposed for measure of collocation long before, although still widely applied today in various fields, it has the disadvantage of heavily favoring rarely occurring items.*

A new improved Square Mutual Information approach is proposed to solve this problem. Supported by experimental results, the precision of this new method is better than that of MI and other modified approach such as combination of external and internal measures. Another advantage of this new approach is that it remains language independent.

Keywords: Collocation, association measure, square mutual information, improved square mutual information

1. **Introduction.** Statistical approach of collocation extraction has been a dominant trend for years, from [4, 9, 6] to [5, 7, 1]. Mutual Information (MI) is one of most early and widely used measures, referred the by the majority of research papers on collocation extraction.

In [8], a total of 82 association measures are empirically tested, 6 among which are mutual information and derived measures. However, the new approach proposed in this paper is not found in the full list.

Our main interest lies on the improvement of mutual information related measures. One intuitional motivation is that mutual information is originated from information theory, while many information-theoretic approaches have been quite successful in NLP. Another motivation from the opposite direction is that mutual information is sometimes considered as a poor measure for collocation extraction. Despite the disadvantage of heavily favoring rarely occurring items, we think that MI can be improved to get better performance.

We will first review one of such attempt to modify MI [2, 3].

2. **Unithood: Chen's approach.** Chen [2, 3] calculates unithood measure by combining the external measure and the internal measure.

The external measure is based on two rates: the left dependent rate (LD) and the right dependent rate (RD).

$$LD(w_1 \dots w_n) = \frac{\max\limits_{a \in A} f(aw_1 \dots w_n)}{f(w_1 \dots w_n)}$$

$$RD(w_1 \dots w_n) = \frac{\max\limits_{b \in B} f(w_1 \dots w_n b)}{f(w_1 \dots w_n)}$$

where $w = w_1 w_2 \dots w_\mathrm{n}$
$f(w)$ is the frequency of a string w,
A is the full set of all the left neighbor elements of w,
a is any element of set A,
B is the full set of all right neighbor elements of w,
b is any element of set B.

The external measure, denoted as IDR (independent rate), is given by.

$$IDR(w_1..w_n) = (1 - 1/\ f(w_1..w_n)) \times \sqrt{(1 - LD(w_1..w_n)) \times (1 - RD(w_1..w_n))} \quad (3)$$

The internal measure is based on C*onnectRate*($w_i w_{i+1}$), which is given by

$$ConnectRate(w_i w_{i+1}) = \frac{p(w_i w_{i+1}) - p(w_i)p(w_{i+1})}{p(w_i w_{i+1})}$$

The minimum of *ConnectRate*($w_i w_{i+1}$), denoted as *MinConnectRate*($w_1..w_n$), is the internal measure.

$$MinConnectRate(w_1..w_n) = \min_{1 \le i \le n-1} ConnectRate(w_i w_{i+1})$$

The final formula of unithood measure, denoted as *UnitRate*($w_1..w_n$), is the product of external measure *IDR*($w_1..w_n$) and internal measure *MinConnectRate*($w_1..w_n$).

$$UnitRate(w_1..w_n) = IDR(w_1..w_n) \times MinConnectRate(w_1..w_n)$$

It can be seen that *ConnectRate*($w_i w_{i+1}$) is a transformation of MI, which can be derived from MI directly. This suggests that Chen's approach also belongs to the family of MI, with which we will compare the results of our new method.

3. **Improved square mutual information: New approach.** We add a new term to square MI, which increases the influence of high frequency combinations by logarithmic scale.

The bigram version is given by

$$SquareMI(x,y) = \log \left(\frac{f(xy)^2 \times \log (1+f(xy))}{f(x)\times f(y)}\right)$$

where *x*, *y* is the adjacent part of combination *xy*,
f(*x*), *f*(*y*) is the frequency of part *x*, *y*,
f(*xy*) is the frequency of combination *xy*.

While the n-gram version is

$$= \log \left(\frac{f(w_1 ... w_n)^n \times \log (1+f(w_1 ... w_n))}{\prod_{i=1}^{n} f(w_i)}\right)$$

where $w = w_1 w_2 \ldots w_n$,
$f(w_i)$ is the frequency of part w_i,
$f(w_1 \ldots w_n)$ is the frequency of combination w.

4. **Results and Discussion.** The evaluations and results are as below:

The first part of the evaluation data is the People’s Daily Corpus (January 1998) segmented and annotated by Institute of Computational Linguistics, Peking University.

The second part of the evaluation data is Financial Times (http://www.ftchinese.com/), mainly Chinese text translated from original English text.

The evaluation is based on the following assumption: The connection between collocations and words is similar to that between words and Chinese characters. If a method is suitable for extracting words from Chinese character combinations, then it is suitable for extracting collocations from word combinations.

TABLE 1. Comparison of precisions

Number of collocations	**Mutual Information(%)**	**Unit Rate(%)**	**Square MI(%)**
Top 100	68.00	86.00	95.00
Top 500	69.60	87.58	88.18
Top 1000	66.70	81.60	87.20
Top 5000	63.02	67.34	76.10
Top 10000	58.46	58.75	64.75
Top 15000	53.29	53.55	57.32
Top 21296	47.92	49.15	50.26

The top 21296 terms are selected for evaluation, in parallel with Chen’s approach (denoted as UnitRate hereafter) for better comparability, as shown in Table 1.

The precision changes with the number of collocations selected. As shown in Figure 1, 2, and 3, the horizontal axis is number of collocations (100 as a unit), while the y-axis is precision.

From Figure 1 we can see that our improved square mutual information approach is better than Chen's method and pointwise mutual information method.

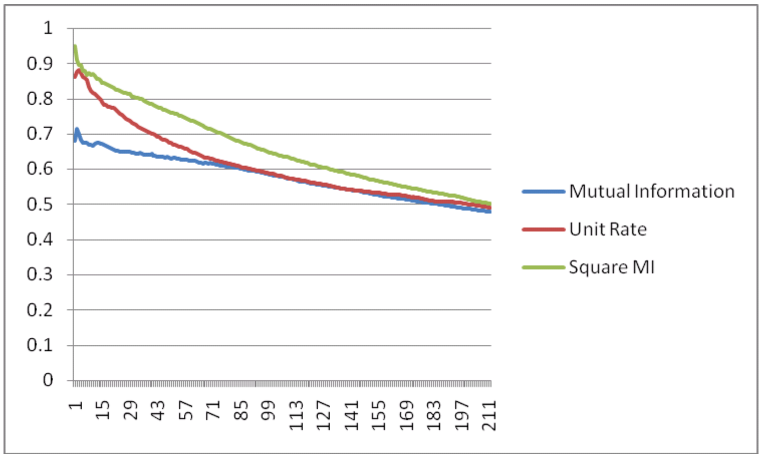

FIGURE 1. Comparison with MI and UnitRate.

In [2], Chen's methods achieved higher precision than that by repeating his method. One conjecture is that preprocessing and/or postprocessing are done before/after the extraction. After we remove the word extraction result containing Chinese characters in stop list, the precision curve becomes Figure 2.

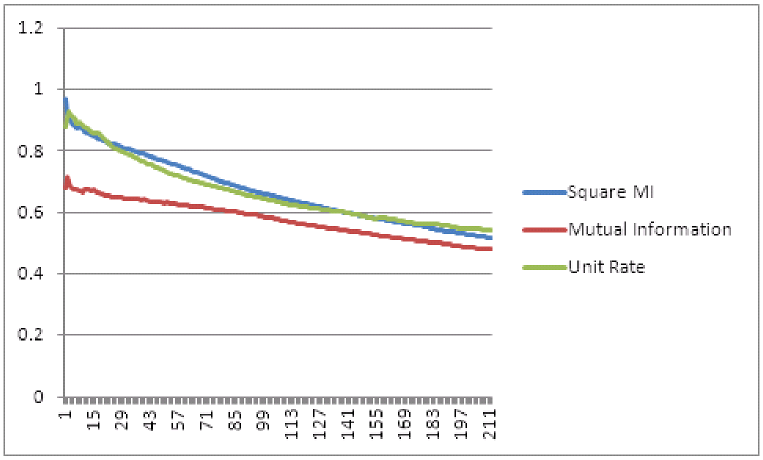

FIGURE 2. Comparison with UnitRate after filtering.

From Figure 2 we can see that after the removal of words containing Chinese characters in stop list, Chen's method get much closer result to our improved square mutual information method.

Figure 3 shows the change in precision curve of our improved square mutual information method before and after the removal of words containing stopping Chinese characters.

The minor change in precision curve of our method suggests that our method can do better even before the use of filtering, which means our method is more effective and can be language independent.

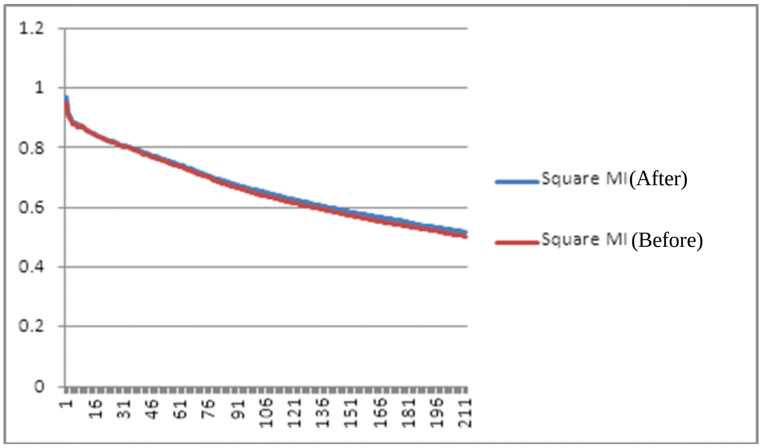

FIGURE 3. Improved Square MI (before and after filtering).

Expert Evaluation: A randomly-chosen sample of the result is manually checked by human experts, and the approved percentage is shown in Table 2.

TABLE 2. Comparison of expert evaluation

Number of collocations	**Unit Rate(%)**	**Square MI(%)**
Top 100	82	84
Top 500	72	78
Top 1000	58	63
Top 3000	53	56
Top 5000	40	43
Top 10000	38	38

From these comparisons, we find that our improved square mutual information approach obtains a better precision in collocation extraction.

5. **Conclusions.** The new improved square mutual information approach over performs pointwise mutual information method completely. Although simpler than Chen’s approach, our approach is still more effective than Chen’s when no filter is applied. Human evaluation on chosen sample also confirms the advantage of this new approach.

Acknowledgment. This work is partially based on the segmented and annotated Chinese corpus developed by Institute of Computational Linguistics at Peking University under the leadership of Professor Shiwen YU.

REFERENCES

[1] I. A. Bolshakov, E. I. Bolshakova, A. P. Kotlyarov and A. Gelbukh, Various Criteria of Collocation Cohesion in Internet: Comparison of Resolving Power, *Computational Linguistics and Intelligent Text Processing*, Lecture Notes in Computer Science, vol.4919, pp.64-72, 2010.

[2] Chen Yirong, *The Research on Automatic Chinese Term Extraction Integrated with Unithood and Domain Feature*, Master Thesis in Peking University, Beijing, 2005.

[3] Yirong Chen, Qin Lu, Wenjie Li, Zhifang Sui and Luning Ji, A Study on Terminology Extraction Based on Classified Corpora, *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, pp.2383-2386, 2006.

[4] K. Church and P. Hanks, Word association norms, mutual information and lexicography, *Computational Linguistics*, vol.16, no.1, pp.22–29, 1990.

[5] S. Evert, *The Statistics of Word Cooccurrences: Word Pairs and Collocations*, PhD dissertation, IMS, University of Stuttgart, 2004.

[6] C. Manning and H. Schutze, *Foundations of statistical natural language processing*, MIT Press, Cambridge, MA, 1999.

[7] B. T. McInnes, *Extending the Log Likelihood Measure to Improve Collocation Identification*, M.S. Thesis, Department of Computer Science, University of Minnesota, Duluth, 2004.

[8] P. Pecina, Lexical association measures and collocation extraction, *Lang Resources & Evaluation*, vol.44, pp.137–158, 2010.

[9] J. Pustejovsky, P. Anick, and S. Bergler, Lexical semantic techniques for corpus analysis, *Computational Linguistics*, vol.19, no.2, pp.331-358, 1993.