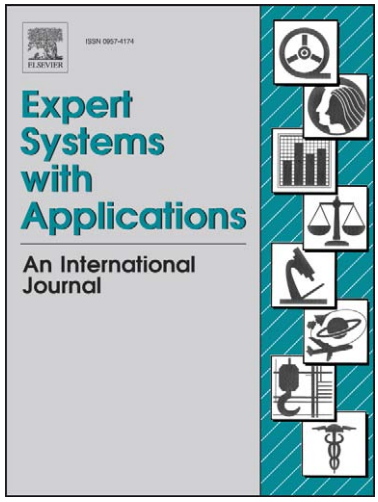

Improving effectiveness of mutual information for substantival multiword expression extraction

Wen Zhang, Taketoshi Yoshida, Xijin Tang, Tu Bao Ho

PII: S0957-4174(09)00153-5
DOI: 10.1016/j.eswa.2009.02.026
Reference: ESWA 3513

To appear in: *Expert Systems with Applications*

Received Date: 17 July 2008
Accepted Date: 6 February 2009

Please cite this article as: Zhang, W., Yoshida, T., Tang, X., Ho, T.B., Improving effectiveness of mutual information for substantival multiword expression extraction, *Expert Systems with Applications* (2009), doi: 10.1016/j.eswa.2009.02.026

Improving effectiveness of mutual information for substantival multiword expression extraction

Wen Zhang[1], Taketoshi Yoshida[1], Xijin Tang[2], Tu Bao Ho

[1] School of Knowledge Science, Japan Advanced Institute of Science and Technology,
1-1 Asahidai, 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. One of the deficiencies of mutual information is its poor capacity to measure association of words with unsymmetrical co-occurrence, which has large amounts for multi-word expression in texts. Moreover, threshold setting, which is decisive for success of practical implementation of mutual information for multi-word extraction, brings about many parameters to be predefined manually in the process of extracting multiword expressions with different number of individual words. In this paper, we propose a new method as EMICO (Enhanced Mutual Information and Collocation Optimization) to extract substantival multiword expression from text. Specifically, enhanced mutual information is proposed to measure the association of words and collocation optimization is proposed to automatically determine the number of individual words contained in a multiword expression when the multiword expression occurs in a candidate set. Our experiments showed that EMICO significantly improves the performance of substantival multiword expression extraction in comparison with a classic extraction method based on mutual information.

Keywords: substantival multiword expression, mutual information, enhanced mutual information, collocation optimization, EMICO.

1 Introduction

A word is characterized by the company it keeps [1] and the closer a set of terms, the more likely they are to indicate relevance [2]. That means not only the individual word but also the contextual information of the individual word is useful for further information processing. This simple and direct idea motivates researches on multiword expression (MWE), which expects to capture semantic concepts expressed by multi-words in text. In state of art, there is no satisfactory formal definition of MWE but some generally grammatical, syntactical or lexical characteristics to describe multiword expression [3]. In this paper, the substantival multiword expression we refer to includes merely terminology and named entity. Although it is the simplest and most frequently used MWE in text, unfortunately, we also cannot give a precise definition for the substantival multiword expression but to use some explicit properties to characterize it.

- It has been used as noun phrase in text to describe a concrete concept in context such as “federal reserve board”, “Shearson Lehman Brothers Inc”, etc.
- From grammatical parsing view, it can be parsed as an entity in a sentence and usually, it has a more stable syntactical pattern than other MWEs in text.
- In lexical composition, it often uses contiguous words composed as a word block in sentence, i.e., there is no other word inserted into a substantival multiword expression. This is not the case in most prepositional and conjunctional collocation such as “too…to…” and “so…that…”
- Like terminology, substantival multiword expression also has a length as 2-6 individual words.

The motivation for us to carry out the research on MWE extraction is that we intend to use MWE for text mining purpose and examine its performance in comparison with traditional indexing method as individual words combined with vector space model [4, 5]. We conjecture that for text representation, MWE may have superiority in both statistical and semantical quality over individual word. With this intention, we started out our research on MWE extraction [6, 7]. Especially, the focus of this paper is on using statistical method to extract MWE from text. We also follow the regulation in this area to propose an association measure to score candidates firstly and then propose a method to differentiate the substantival MWEs from all candidates automatically.

In statistical method for MWE extraction, the most frequently used association measure is mutual information (MI). Although there is also some other measures such as z-score, mutual expectation, etc, their basic ideas are very similar with MI: joint probability inverse products of independent probabilities and the assumption concerning the two words in a word pair is the same: the two words may have as many occurrences as each other, that is, their occurrence possibilities in text are almost equal. Hence, these methods can be regarded as variants of mutual information. However, we will show later that MI is not appropriate for association measure when it goes to unsymmetrical co-occurrence of these two words. Moreover, how to select candidates after association measure is another problem. Usually, a predefined parameter was set to retain a proportion of candidates with top association scores (values) as final extracted MWEs. Although LocalMaxs [8] was proposed to determine the number of individual words included in a MWE automatically, it is not appropriate for extracting substantival MWE because it often has a fixed composition and sometimes the word sequences at the association maxima is not an exact substantival MWE.

In this paper, EMICO was proposed to extract multi-words from documents. Specifically, we proposed the enhanced mutual information (EMI) to cope with the problem as unsymmetrical co-occurrence. And we developed collocation optimization (CO) to determine the number of individual words contained in a substantival MWE automatically.

The key idea of EMI is to measure word pair’s dependency as the ratio of its probability of being a multi-word to its probability of not being a multi-word. By revising the individual words’ occurrences as their occurrences subtracting their co-occurrence, respectively, EMI has considered individual word’s occurrences and its proportion contributed in its co-occurrence with other words synthetically so that association score will vary dramatically with the proportion of one word’s occurrence contributed to its co-occurrences with other words. In addition, by separating the association contributed by each word in a word sequence with more than 2 single words, rather than requiring MWE candidates to be formatted into two components as in practical implementation of mutual information, EMI can reduce the negative effect of rare occurrences to some extent.

Collocation optimization (CO) was proposed to determine the exact number of individual words in a substantival MWE automatically. We use the traditional N-gram method to produce word sequences with a same head noun and pack these sequences into a candidate set (clarified in Section 5.1). In each candidate set, we only retain one of its candidates as a substantival MWE because we conjecture that there must be at most only one correct substantival MWE for the head noun to compose a most appropriate MWE with other words. The key idea of CO is similar with LocalMaxs, that is, when an individual word is added to a MWE candidate (old MWE), the cohesiveness of the new MWE will increase if this individual word is exactly a part of this MWE candidate, otherwise, the association score of the new MWE will decline compared with the old MWE.

The remainder of this paper is organized as follows. Section 2 provides a literature review of the MWE extraction. Section 3 introduces mutual information, particularly with its practical application for MWE extraction. Section 4 proposes EMI. We will give its definition, its theoretical analysis and numerical simulation in comparison with mutual information. Section 5 proposes CO. Its mechanism will be specified together with a comparison with LocalMaxs. Section 6 specifies the details of EMICO for substantival MWE extraction together with practical performance evaluation on real corpus. Section 7 concludes this paper and indicates our future research.

2 Literature review

Generally speaking, there are four types of methods developed for MWE extraction: statistical method, linguistic method, hybrid method and machine learning. These methods are introduced as follows.

2.1 Statistical methods

In statistical methods for MWE extraction, Church and Hanks presented the concept of word association firstly, and then proposed MI as an objective measure for estimating word association norms [9]. Pecina compared 84 kinds of association measures for bigram collocation extraction and concluded that in Czech data, MI has the best performance [10]. In recent development on MWE extraction, mutual expectation (ME) is the most popular measure for words' association estimation. It combines candidate's frequency and its possibility to be a fixed phrase as the inputs of the method [8, 11]. We conjecture that ME is very suitable for extracting variable phrases but not for substantival MWE because the elements of the latter as individual words do not often change their positions and orders when they construct a substantival MWE. Silva et al proposed LocalMaxs algorithm to extract both contiguous and non-contiguous multi-word lexical units from corpora [8]. The basic idea behind LocalMaxs is that, the association score of an N-gram should be a local maximum in three sequences as N-1 gram, N-gram and N+1 gram which have same head noun. This idea is very similar with the collocation optimization we will present in this paper (clarified in Section 5.2). However, LocalMaxs and collocation optimization have some differences in essence as we will discuss them later. Smadja used relative positions of the elements of a word pair in sentences to extract fixed patterns from corpora [12]. In his method, word pair as w and w_i could be considered as a collocation if and only if the two words are repeatedly used together within a single syntactic construct, i.e., they have a marked pattern of co-appearance. Specifically, strength is used to measure co-occurrence frequency of two words, and spread is used to measure peak magnitude of their relative positions. The final decision of

their relative position if they construct a collocation is determined by a filter which only selects peaks of their relative positions. Smadja claimed that, the proposed method can extract collocations with structural consistency, and ignore the word pairs with same context, such as doctor and nurse, which is usually extracted by MI. The experiments on Brown corpus showed that precision of the proposed method for collocation retrieval was raised from 40% to 80%. Kita et al compared two statistical methods for automatic collocation extraction as MI and cost criteria in English and Japanese corpora, respectively [13]. Their studies showed that MI tends to extract task-dependent compound noun phrases, while a cost criterion tends to extract predicate phrase patterns. Chen and Du conducted a work on automatic extraction of bilingual multi-word units from parallel corpora [14]. The goal of their study is to find corresponding multi-words from one language to another language using parallel corpora learning. In their method, t-score and LocalMaxs algorithm were utilized to rank candidates and determine their lengths.

In summary, the statistical methods for multi-word extraction include two directions: one is to develop new association measures to rank candidates and the other is to develop new strategies to align the best candidate as a MWE when candidates' scores were produced.

2.2 Linguistic methods

Bourigault proposed surface grammatical analysis for the extraction of terminological noun phrases [15]. His method includes two stages: analysis and parsing. In stage of analysis, a base of rules is set up to identify frontier markers to extract maximal-length noun phrases from texts. In stage of parsing, grammatical category of lexical units was assigned to the maximal-length noun phrases to divide them into probable terminological units. Using this method, LEXTER is developed as a software package for extracting French terminologies. Justeson and Katz extracted technical terminologies from documents using a regular expression on part-of-speeches of a word sequence, together with the condition that the sequence's frequency must be more than two [16]. The experiments show that their algorithm can cover multi-word technical term with a high proportion as more than 90%. Argamon et al proposed a memory-based approach (MBSL algorithm) to learning shallow natural language patterns from corpora [17]. Their method relies on local part-of-speech information of a word sequence instead of full parsing a sentence. They firstly separated the POS sequence of a multi-word into small POS tiles and then they counted each POS tile's frequency when it occurs within a candidate's POS sequence (positive count) and when it does not occur within a multi-word's POS sequence (negative count), respectively. Hence, a candidate is ranked by positive count, negative count and context information of the tiles included in it. Their evaluation on noun phrase sequence (NP), verb-object (VO) and subject-verb (SV) showed that the recall of MBSL algorithm is around 90% and precision is between 77% and 92%.

In summary, the mostly used linguistic information for MWE extraction is words' POS tags, which is from both grammatical and syntactical requirement for a word sequence to be a MWE.

2.3 Hybrid Methods

Dias proposed an original hybrid system called HELAS to extract MWE from POS tagged corpora [11]. The key idea of his system is that ME is employed not only to score the association of words but also to score the association of POS patterns in the tagged corpora. Then a combination of both words' and POS's ME is used to evaluate the global degree of cohesiveness of a word sequence and its POS tag sequence. Finally, LocalMaxs is used to

retrieve multiword candidates by evidencing local maxima of association measure values. Chen and Chen proposed a hybrid approach to extracting noun phrases from large scale texts [18]. The input of their method is POS tagged sentences. A probabilistic partial parser was used to partition the tagged sentences into chunks. Hence, semantic head was determined for each chunk based on word's semantic usage and syntactic head was determined for each chunk based on grammatical relations. Finally, a finite state mechanism was designed to connect the chunks as many noun phrases as possible according to the chunk's semantic and syntactic heads.

In summary, the focus of hybrid methods for MWE extraction is on using both statistical and linguistic information of a word sequence to measure its possibility to be a multiword expression.

2.4 Machine Learning Methods

In machine leaning methods, Pinca used machine learning approach for MWE extraction [19]. In his method, each collocation candidate is described by a feature vector consisting of scores of 55 kinds of association measures for the candidate such as joint probability, MI, t-score, etc. Machine leaning methods as linear logistic regression, linear discriminant analysis and neural net were employed to train a combined classifier using training vectors. Hence, new coming collocation candidates were ranked by this classifier to determine whether or not they are MWEs. The experiments showed that the best association measure for ranking collocation candidates depends fully on specific data. However, machine learning methods significantly improved ranking of collocation candidates on all of their data sets than the best association measure. Duan et al developed a bio-inspired approach for multi-word expression extraction [20]. Their motivation is based on the similarity of textual sequence and gene sequence alignment. In their method, longest common sequence alignment, which originated from RNA sequence alignments [21], and heuristic knowledge, which were linguistic rules of part-of-speech information on MWE [16], was proposed to extract repetitive patterns from textual sequence and convert patterns into multi-word expressions. Zhang et al proposed a Chinese named entity recognition method using role model [22]. In their method, many kinds of roles were defined for each type of Chinese named entity such as location, person, organization, etc. They tagged a standard Chinese corpus using these roles manually and used this corpus as training data for Viterbi algorithm [23]. Hence, new Chinese named entities can be identified automatically by the trained classifier.

In summary, machine learning methods for MWE extraction employed artificial intelligence methods to discover new knowledge from either word information (frequency, association score, etc) or part-of-speech information of words (order, POS sequence, etc). Then the new knowledge was used to determine whether or not a new coming candidate is a MWE.

3 Mutual information

Mutual information (MI) is defined as the reduction in uncertainty of one random variable due to knowing about another, or in other words, the amount of information one random variable contains about another. In multi-word detection, MI can be defined as the amount of information provided by the occurrence of the word represented by Y about the occurrence of the word represented by X. Church and Hanks proposed the association ratio for measuring

word association based on the information theoretic concept of MI [9]. In their method, the MI between word x and y was defined as Eq.1.

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

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

The primary reason for applying MI to multi-word extraction is that MI has the support from both information theory and mathematical proof. If word x and word y are independent from each other, i.e. X and Y co-occur 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 on each other. The higher MI of a word pair, the more genuine is the association between the two words.

However, there are mainly two deficiencies inherent in MI for measuring the words' association. The first one is the unsymmetrical co-occurrence problem, that is, it only considers the co-occurrence of two words while ignoring those cases that one word (for a word pair) occurs without the occurrence of the other word. For instance, assuming the occurrence frequency of X is 50 and Y is 200, and their co-occurrence is 50, that is, the X only occurs in co-occurrence with Y but Y has more co- occurrence with other words than X. Certainly, the information of Y contained in X is much more than the information of X contained in Y. In this case, mutual information cannot play a reasonable role as association measure to make out that X and Y is a fixed MWE. Church and Gale give an example of using mutual information to align the corresponding words between French word "chambre", "communes" for English word "house" [24, 25]. We will discuss this example in details in next section (clarified in Section 4.2). The second deficiency of MI concerns the rare occurrence problem [25]. As is shown in Eq.1, when we assume that $P(x)$ and $P(y)$ are very small, but $I(x, y)$ can be very large despite the small value of $P(x, y)$, in this situation, the dependency between X and Y is very large, despite the fact that X and Y co-occur very few times. Actually, rare occurrence is a hard problem for linguistic data, and there is no effective remedy for it. Due to the deficiencies of MI, the proportion of "good" candidates per range of score values is quite uniformly distributed, and it is very difficult to distinguish the "good" ones from the "bad" ones. In practice, MI method is employed to extract multi-words in this paper as follows [13, 26].

1. Start out from the basic vocabulary V_0. Set $n = 0$;

2. Augment the vocabulary V_n by all word sequences by all word sequences "x y" for which $MI(x, y) > Thr$, where Thr is a predefined threshold for word sequence association score;

3. From Step 2, a new vocabulary V_{n+1} is established.

4. Adjust the vocabulary size N to reflect the new vocabulary V_{n+1}

5. Resume from Step 1 with V_{n+1} as its basis.

4 Enhanced mutual information

4.1 Motivation

The reason for unsymmetrical co-occurrence is from the unequal proportions of the words' occurrences contributing to their common co-occurrence in a word pair as is shown in Section 3. We would like to make it clear using the following example as shown in Figure 1. Given the two cases of words' co-occurrence, our problem is which one should be regarded as having greater word association than the other?

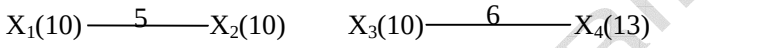

Figure 1 Two different kinds of co-occurrences of two components usually exist in documents. The number in the bracket is the frequency of occurrence of the individual component, and the number on the line is the frequency of two component's co-occurrence.

Obviously, the left co-occurrence is more balanced than the right one because both X_1 and X_2 contribute half of their occurrence to co-occurrence. However, for the right one, X_4 contributes less than half of its occurrence to co-occurrence with X_3. If word association is measured by mutual information, the left one will be determined to have a larger association score than the right one. However, the right one may be more preferred than the left one because the right pair has more co-occurrences than the left one and X_3 contributes more than half of its occurrence to co-occurrence with X_4. Nevertheless, we can consider the problem in a simple way: the sum of the proportion of occurrences from X_1 and X_2 is 1.0, but the sum of occurrence proportion from X_3 and X_4 contributing to their co-occurrence is more than 1.0. Hence, in fact, the dependent relationship (association) is intensified between X_3 and X_4. This is the very motivation for us to propose EMI for association measure.

4.2 Definition

To attack the unsymmetrical co-occurrence problem, not only the co-occurrence of the individual words in a word pair, but also their respective occurrences excluding their co-occurrences, which are the number of times when one occurs while the other one is absent, should be considered respectively. Hence, EMI is proposed and defined as the ratio of the probability of word pair occurrence over the product of the probabilities of the individual words' presences excluding the presences of the word pair, i.e., the likelihood of being a multi-word over the possibility of not being a multi-word. It has the mathematic formula described in Eq.2.

$$EMI(x,y) = \log_2 \frac{P(x,y)}{(P(x) - P(x,y))(P(y) - P(x,y))} \qquad (2)$$

We would like to take the example from Church and Gale mentioned above to illustrate the effectiveness of EMI. They use mutual information to align the corresponding words between French words "chambre", "communes" and English word "house" [27, 28] with the words occurrence shown in Table .1. We can see that if we follow the rule of mutual information, the French counterpart of "house" will be "communes", not the right counterpart as "chambre". By contrast, EMI can produce the right alignment.

Table.1 The alignment of English and French words using mutual information (MI) and EMI. "¬" means not present and the number is the frequency in each case.

	chambre	¬chambre	MI	EMI
house	31,950	12,004		
¬house	4,793	848,330	4.1495	8.9605
	communes	¬communes	MI	EMI
house	4,974	38,980		
¬house	441	852,682	4.2286	8.0200

4.3 Theoretical analysis

For the independent case between two words x and y in a bi-gram, i.e., $P(x, y) = P(x)P(y)$, we can conjecture that $P(x) >> P(x, y)$ and $P(y) >> P(x, y)$ because x (y) will co-occur with other words at the same likelihood as y (x) in corpus. Thus,

$$EMI(x, y) \approx \log_2 \frac{P(x, y)}{P(x)P(y)} = I(x, y) = 0 \qquad (3)$$

Eq.3 means that $EMI(x, y)$ has as approximately same capacity as mutual information when X and Y are independent of each other.

For dependent case between two words X and Y in a word pair, actually, the association relationship can be divided into two situations: negative correlation and positive correlation. The negative correlation between X and Y is meaningless as they will co-occur quite a few times in this case. The positive correlation between X and Y is an omen that they could constitute a MWE. In this case, $P(x, y) > P(x)P(y)$ and

$$\begin{aligned} EMI\ (x, y) &> \log_2 \frac{P(x, y)}{(P(x) - P(x)p(y))(P(y) - P(x)P(y))} \\ &= \log_2 \frac{P(x, y)}{P(x)P(y)} + \log_2 \frac{1}{(1 - P(x))(1 - P(y))} > I(x, y) \end{aligned} \qquad (4)$$

Eq.4 means that EMI will amplify the association of likely MWEs. This kind of amplification in association is beneficial for MWE extraction because it will distinguish the likely MWEs from those candidates which are not real MWEs more significantly.

Furthermore, Eq.2 can also be rewritten as Eq.5. Here, $\frac{P(x, y)}{P(x)}$ and $\frac{P(x, y)}{P(y)}$ are the proportions of x's occurrence and y's occurrence contributing to their co-occurrence, respectively, so EMI will increase when the proportions increase. This means that the association of a word pair will increase dramatically if individual words contribute more and more occurrences to their co-occurrence.

$$EMI(x, y) = \log_2 \frac{P(x, y)}{P(x)P(y)(1 - \frac{P(x, y)}{P(x)})(1 - \frac{P(x, y)}{P(y)})} \qquad (5)$$

4.4 Numerical Simulation

Figure 2 is the comparative curves of MI and EMI in characterizing associations of a word pair. We can see that the association characterized by EMI is increasing more sharply than MI when word X and Y's co-occurrence is varying from 0 to 90. This point illustrates that EMI can augment association difference between candidates which are multi-words and candidates which are not multi-words.

Figure 3 is the comparative contrast of MI and EMI's trend in describing the association with the occurrence variation of X and Y when their co-occurrence is fixed. We can see that EMI is more sensitive than MI in association value of X and Y at the edge part of the bottom plane constructed by X and Y axes. Hence, their association value will jump if X or Y's occurrence is almost the same as their co-occurrence. For instance, if we have two word pairs as (x, y) and (x', y'), where (x', y') is at the central part of XY plane and (x, y) is at the edge part of XY plane, we can see that (x', y') will have larger association than (x, y) in MI but smaller association in EMI. In real situation of substantival MWE extraction, EMI is more reasonable for characterizing the association of MWE's individual words if we consider the situation that many MWE have the same head noun such as "information processing", "waste processing", "water processing", etc and the example of X and Y we given in Section 3.

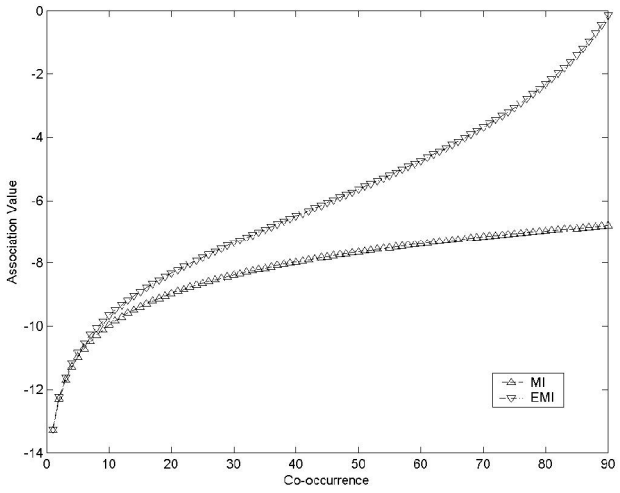

Figure 2 Plots of dependency value of MI and AMI. The frequencies of word X and word Y are fixed as 100

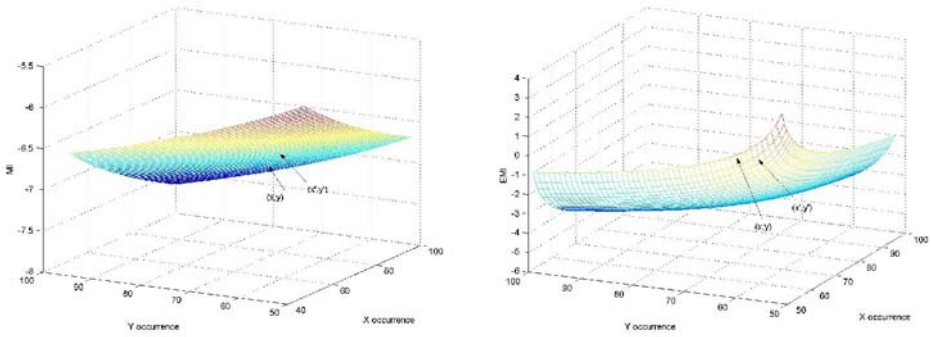

Figure 3 Plots of MI (left) and EMI (right) for association measure of (x, y) and (x', y'), respectively. Their co-occurrence is fixed as 50 and occurrences of two words are varying in their axes.

4.4 Practical Implementation

In order to employ the EMI for practical use in MWE candidate ranking, some small adaptations must be made. The first one is to extend its scope to rank MWE candidates of more than two words, i.e. longer than a bi-gram. Take a three word sequence (x,y,z) for example, the question is how to rank the possibility of its being a multi-word. Generally, if we follow the rules of MI, one solution can be used as follows.

$$EMI(x,y,z) = \log_2 \frac{P(x,y,z)}{(P(x,y) - P(x,y,z))(P(z) - P(x,y,z))} \qquad (6)$$

However, there is an intrinsic problem with formula (6): the longer the sequence, the larger its EMI is, because the EMI value of (x,y,z) is dominated by the smallest occurrence among x, y and z. For instance, if x occurred rarely, $P(x,y)$ should be very small. Even if word y occurs frequently in a corpus, $P(x,y) - P(x,y,z)$ would still be very small. For this reason, we can infer that the longer length of the sequence, the more likely it will contain a rare occurrence. Thus, sequences with the rare occurrences have usually higher EMI values than those sequences without rare occurrences. That is, if one word rarely occurs, it will reduce extremely the occurrences of the word sequence containing it. Although the rare occurrence problem is a hard nut to crack unless heuristics is involved, we can reduce its negative influence to some extent. Eq.7 is our solution for ranking the multi-word candidate of more than two words

$$EMI(x,y,z) = \log_2 \frac{P(x,y,z)}{(P(x) - P(x,y,z))(P(y) - P(x,y,z))(P(z) - P(x,y,z))} \qquad (7)$$

We can deduce that if there is a rare occurrence of x, but if y and z have many occurrences, the EMI from Eq.7 will be less influenced by x than that in Eq.6.

Hence, in practice of a sequence $(x_1, x_2, ..., x_n)$, $P(x_1, .., x_n) = p$, $P(x_1) = p_1$, $P(x_2) = p_2$, …, $P(x_n) = p_n$, we have

$$EMI(x_1, x_2, .., x_n) = \log_2 \frac{p}{(p - p_1)(p - p_2)...(p - p_n)} \quad (8)$$

By maximum likelihood estimation,

$$\begin{aligned} EMI(x_1, x_2, .., x_n) &= \log_2 \frac{F / N}{(F_1 - F)(F_2 - F)...(F_n - F) / N^n} \\ &= (n-1)\log_2 N + \log_2 F - \sum_{i=1}^{n} \log_2 (F_i - F) \end{aligned} \quad (9)$$

N is the number of words contained in the corpus, it is usually a large value, more than 10^6. Here, $\log_2 N$ can be regarded as the increased amount of EMI when one individual word is added into a multi-word candidate. However, $\log_2 N$ is usually a large value and it will make AMI value of a word sequence dominated by the length of the sequence. This result is not expected for dependency measure so $\log_2 N$ is replace by α which represents the importance of the length of a word sequence to its dependency value, i.e., EMI value. Another problem with Eq.9 is that in some special cases we have $F_i = F$ and $F_i - F = 0$, and these special cases will make Eq.9 meaningless. For this reason, Eq.9 is rewritten as Eq.10.

$$EMI(x_1, x_2, ..., x_n) = (n-1)\alpha + \log_2 F - \sum_{i=1}^{m} \log_2 (F_i - F) + (\text{n - m})\,\beta \quad (10)$$

where m is the number of single words whose frequency is not equal to the frequency of the sequence in the corpus, β is the weight of the single words whose frequency are 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”. To simplify, we set $\alpha = \beta = 0.5$ in this paper.

5 Collocation optimization

5.1 Related concepts

In order to proceed, some related concepts with substantival MWE should be clarified. As we have pointed out in Section 1, the topic of this paper is to extract terminologies and named entities using statistical methods and usually, they are noun phrases. For this reason, substantival MWE candidates are produced by traditional N-gram method. For instance, if we have a sentence after morphological analysis as “A B C DE F G H.” and H is a noun, then 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 a substantival MWE usually has a length as 2-6 words and H is the head noun of these 5 candidates. Hence, we have the following definition.

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.

The second concept we want to clarify is the method we will used for multi-word extraction. Actually, we will rank all candidates using EMI and retaining a proportion of

candidates for further selection. What is more, to simplify the process of the mutual information method specified in Section 3, we would like to define a uniform threshold here to fetch candidates with top association scores.

Definition 2: Candidate Retaining Level (CRL) is a predefined ratio used to retain a proportion of candidates with top association scores.

5.2 Association variation and mechanism of collocation optimization

Although association value can provide us some hints to select the correct candidates as a multi-word, it cannot locate its length precisely. Moreover, the substantival MWE discussed in this paper can be regarded as an extension of head noun, such as "medical information processing" and "information processing" to the head noun "processing". Although "medical information processing" and "information processing" can all be substantival MWEs in our method, in one candidate set, we only allow one candidate to be the final substantival MWE in order to better capture the context of "processing" in that candidate set.

Collocation optimization is proposed to determine the optimal length of a multi-word based on association variation. The association should be intensified when word extension from head noun is moving within the span of a MWE. Otherwise, when word extension goes beyond the span of MWE, its association will decrease. Thus, the basic assumption for collocation optimization is that the association value will increase if a correct individual word is included in MWE candidate but the association value will decrease when an incorrect individual word is included in. Hence, collocation optimization (CO) is developed based on this idea: a candidate can be regarded as most appropriate to be a substantival MWE if and only if its dependency score is maximized among all the candidates of the same head noun in its candidate set and association values of this candidate's sub-components must monotonically increase when their lengths increase.

Assuming we have a word sequence as $(x_n, x_{n-1},..., x_2, x_1)$ and x_1 is head noun, our problem is that we should find M $(2 < M \le n)$ for which $(x_M,..., x_1)$ is the most appropriate substantival MWE in the candidate set with head noun x_1. Our solution in this paper is to find M $(2 < M \le n)$ such that $EMI(x_M,..., x_1) \ge EMI(x_{M-1},..., x_1) \ge ... \ge EMI(x_2, x_1)$. The mechanism of CO is explained as follows.

Firstly, CO will measure the association value of candidates in a candidate set by the formula as Eq.11, where av_0 is a default value which is a small constant that will make $(x_n,..., x_1)$ be rejected as a multi-word. That is, if a word sequence $(x_n,..., x_1)$ does not meet the requirement of CO, it will be excluded from being selected as a multi-word.

$$av[x_n, (x_{n-1},..., x_1)] = \begin{cases} EMI(x_n,..., x_1), \text{ if } EMI(x_n,..., x_1) \ge \\ EMI(x_{n-1},..., x_1) \text{ (n > 2) or } n = 2 \\ av_0, \text{ otherwise} \end{cases} \quad (11)$$

Secondly, the optimal length M of a multi-word, which is extended from the root noun x_1 with the likely maximum length n, is determined by the rule as Eq.12. Finally, $(x_M,..., x_1)$ is extracted as substantival MWE by our method from its candidate set.

$$M = \arg\max_{1 \le m < n} av[x_m, (x_{m-1},..., x_1)] \quad (12)$$

Although CO and LocalMaxs have similar idea as using association variation to decide the length of MWE, their essences are different from each other. Firstly, the background of CO is to use candidate set to express the contextual information of head noun and then decide the length of substantival MWE while LocalMaxs does not take into account any contextual information and use a device as candidate set. Secondly, CO assumes that association values of substantival MWE candidates in a candidate set should be monotonically increasing to their lengths while LocalMaxs is to find a local a maxima and then length is decided in the local maxima as illustrated in Figure. 4.

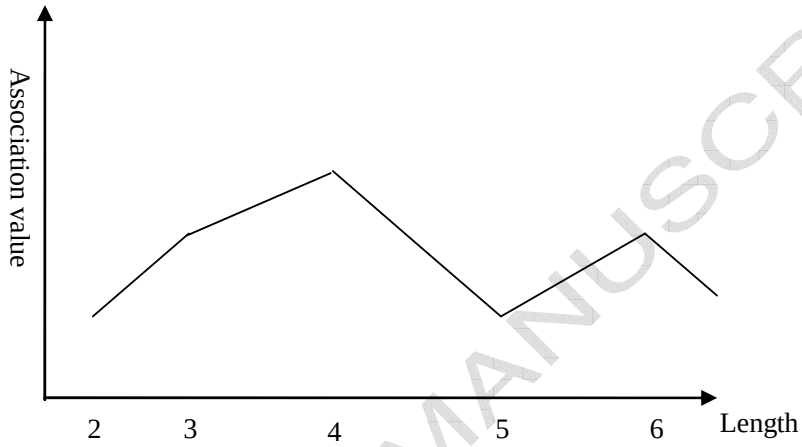

Figure.4. CO assumes association value of a substantival MWE is monotonically increasing to its length while LocalMaxs use the local maxima as correct length. If LocalMaxs method is employed, then both 4-length and 6-length candidates will be assigned as MWEs. However, for CO, only 4-length will be assigned as a substantival MWE.

5.3 The proposed method — EMICO

We combine EMI and CO to propose a new approach, EMICO, for substantival MWE extraction from text as follows.

1. Generate candidates using N-gram method with each noun word as a head noun.
2. Compute association values of candidates using EMI method.
3. Eliminate candidates whose EMI are below the predefined CRL (This point will be made clear in Section 6).
4. Dispatch each retained candidate to its original candidate set where it has been produced.
5. Use CO to select substantival MWE from each candidate set.

We can see from the above procedures of EMICO that the effectiveness of EMI is mainly on Step 3, that is, to augment the differences of candidates' dependencies so that candidates with low association values could be eliminated easily. The effectiveness of CO is mainly on Step 5 to select only one multi-word from a candidate set because in a certain context, there must be merely one candidate which is most appropriate as a substantival MWE.

6 Evaluation by experiments

6.1 Corpora and standard set

Based on our previous work on text mining [27, 28], 184 documents from Xiangshan Science Conference Website (http://www.xssc.ac.cn) are downloaded and used for the Chinese text collection to conduct multi-word extraction. The topics of these documents mainly focus on basic research in academic fields such as nano science, life science, etc., so there are plenty of noun multi-words (terms, noun phrases, etc.) in these documents. These documents contain totally 16,281 Chinese sentences in sum. After the morphological analysis[1] (Chinese is character based, not word based), 453,833 words are segmented individually and of them there are 180,066 noun words.

For the English corpus, Reuters-21578 distribution 1.0 which is available online (http://www.daviddlewis.com/resources/testcollections/ reuters21578/) is used in this paper. It contains 21,578 news articles from Reuters newswire in 1987. It was assembled and indexed with 135 categories by the personnel from Reuters Ltd in 1996. In this research, the documents from 4 categories as "crude" (520 documents), "agriculture" (574 documents), "trade" (514 documents) and "interest" (424 documents) are assigned as the target English document collection. That is, we select 2,042 documents which contain 50,837 sentences and 281,111 individual words, in sum there are 102,338 noun words after stop-word elimination[2]. Figure.3 is the framework of our experiments to evaluate the performance of EMICO for multi-word extraction compared with the traditional method based on MI.

Because of the lack of standard MWE set for texts in our text collection, from Chinese and English respectively, we fetched out 30 texts randomly and built up standard set manually to estimate the performances of EMICO method and MI method. In 30 Chinese texts, there are nearly 3,000 Chinese substantival MWEs and in 30 English texts, there are nearly 1,000 substantival MWEs (Reuters' text is relatively short so we merely use the texts whose sizes are larger than 3K for corpus learning). Table 2 and Table 3 showed some examples of substantival MWEs in our standard set.

Table 2. Chinese standard substantival MWE set. Only some examples are given due to space limitation

Doc No.	# of MW	Examples
1	47	知识产权纠纷, 遗传性疾病, 顾健人院士, 细胞生物学
2	45	生物医学, 胚胎发育, 显微成像技术, DNA分子
3	28	凝聚态物理学, 量子尺寸效应, 高温超导机理
4	40	美国乔治亚理工学院, 纳米结构, 单电子存储
5	100	生物多样性, 林业生态工程, 气象观测
6	45	白春礼院士, MEMS技术, 微系统加工技术

[1] We carried out the Chinese parts of speech using the ICTCLAS tool. It is a Chinese Lexical Analysis System. Online: http://nlp.org.cn/~zhp/ ICTCLAS/codes.html

[2] We obtain the stop-words from USPTO (United States Patent and Trademark Office) patent full-text and image database at http://ftp.uspto.gov/patft/help/stopword.htm. It includes about 100 usual words as stop-words. The part of speech of an English word is determined by WordNet2.0 which is available online: http://wordnet.princeton.edu/obtain and Java WordNet library which is online: http://sourceforge.net/projects/jwordnet.

7	37	DNA芯片，蛋白质结构，基因组序列分析
8	38	DNA序列对称性，遗传密码，北京大学物理化学研究所
9	96	羟基磷酸钙，胶原蛋白，寡链高分子
10	67	中国地质科学院，二叠纪，泥盆纪，可持续发展的蓝图
11	134	复杂性科学，算法复杂性,朱照宣教授，复杂的巨系统
12	64	开放系统，可燃性材料，火灾探测，阻燃剂分子
13	170	老年性疾病，骨质疏松，中日友好医院，传统医学
14	175	计算机硬件，核反应堆材料，虚拟实验，数值仿真
15	86	香山科学会议，人类基因组计划，药物芯片
16	101	物种起源，群体遗传学，自然选择理论，生命科学
17	130	次临界反应堆，中能强流加速器，质子加速器
18	147	生命系统，基因治疗，蛋白质合成，应用冷冻医疗
19	141	环境变迁，青藏高原，地球演化，地质信息，自然生态
20	117	陶瓷复合材料，高性能金属，薄膜孔隙，纳米材料
21	58	生物多样性，物种资料库，基因工程，转基因生物
22	70	巨磁电阻效应，霍尔效应，高温超导体，巴克豪森噪声
23	54	核磁共振现象，脑功能成像，超导MRI仪器
24	59	发育生物学，细胞生物学，分子遗传学，植物激素
25	126	光合作用研究，自然生成系统，叶绿素蛋白
26	64	岩浆活动，地磁变化，陨击效应，行星撞击
27	44	启蒙教育，科学精神，周光召院士，客观世界
28	59	冠脉形成术，帕金森病，早期乳腺癌，老年痴呆病
29	150	分子细胞膜，多肽链合成，组合化学，酶抑制剂
30	155	环境污染，化学定时炸弹问题，硫酸根离子,重金属离子

Table 3. English standard substantival MWE set. Only some examples are given due to space limitation. For convenience, all MWEs are converted into lower cases

Doc No.	# of MW	Examples
1	23	group exports, world surplus, export quota, national stock
2	28	national amusements inc，management proposal，merger plan，broadcast licenses，cable television
3	42	united states，commercial banks，cash loans，third world debt
4	34	industrial output, inflation pressures, federal reserve, monetary policy, steve slifer lehman
5	32	wall street, berger cyrus, lawrence inc, brazil citicorp
6	29	leading industrial nations, currency stability, us officials, financial markets, economic growth
7	36	federal regulators, boyd Jefferies, los angeles, new york

8	30	american express, shearson lehman brothers, nippon life insurance co
9	25	mickey levy, reagan administration, trade deficit, domestic demand
10	33	us banks, mexican committee bankers, finance minister
11	44	uk government,united states,trade industry, japanese communications companies
12	27	exchange rate, currency stability, national sovereignty
13	29	capital investment, trade deficit, treasury bonds, federal funds
14	33	trade minister, foreign ministry, sebastiao rego barros, developing countries
15	25	industrial nations, united states, developing world, debt crisis, international monetary fund
16	53	interest rate, finance ministers, central bankers, global debt, economic situation
17	31	wall street, budget plan, domestic programs, spending levels, senate budget committee
18	54	manhattan bank, money market, interest rate fluctuations, stock markets, corporate bond market
19	34	credit markets, policy shift, economic growth, us banks, developing countries, composite index, economic indicator, federal funds
20	36	monetary sources, european markets, policy coordination, finance minister, trade balances, reagan administration
21	39	us economy, washington house, trade bill, president reagan trade war, lanston co inc, lending rates,
22	24	interest rates, us inflation, central banks, us currency, overseas rates, monetary policy, funds rate
23	26	global trade imbalances, foreign investors, mellon bank
24	*30*	treasury secretary, james baker, reagan administration, government bonds, monetary sources
25	*42*	economic imbalances, world economy, industrial nations, annual meeting, finance ministers
26	*42*	prime minister, yasuhiro nakasone, president reagan, trade dispute, short-term rates, us treasury, us rates
27	*29*	stock market, white house, william schneider, dow jones
28	*29*	personal computer, ibm software standard, international business machines corp, operating system, death sentence
29	*23*	uk reserves, lending rates, general election, foreign currency, gold reserves, market tendency, chancellor exchequer
30	*57*	venice summit, third world, debt crisis, military situation, gulf co-operation, exchange rates, trade surpluses, government officials

6.2 Experimental design

Figure 5 is the complete process of using statistical methods mentioned in this paper to extract substantival MWEs comparatively. After candidate generation using N-gram method specified in Section 5.1, all the candidates and individual words, and their frequencies in corpus are stored in the candidate collection and the individual word collection, respectively. For XSSC text collection, 662, 470 unique candidates and 16,995 unique individual words are produced. And 396,764 unique candidates and 14, 838 unique individual words are obtained from the candidate generation from Reuters text collection.

It should be pointed out that, in fact, rare occurrence and extreme occurrence may cause negative effects on association measure. Although recently smoothing technique is proposed for this purpose, there is no effective remedy to attack these two problems. Hence, we set smallest number of occurrence of individual word in substantival MWE as 3 for both Chinese and English. The largest number of occurrence of individual word in Chinese substantival

MWE is set as 3000 and for English individual word; it is 1000 because above these two thresholds, the words are usually extremely frequently used words such as “dlrs” and “mln” in Reuters text.

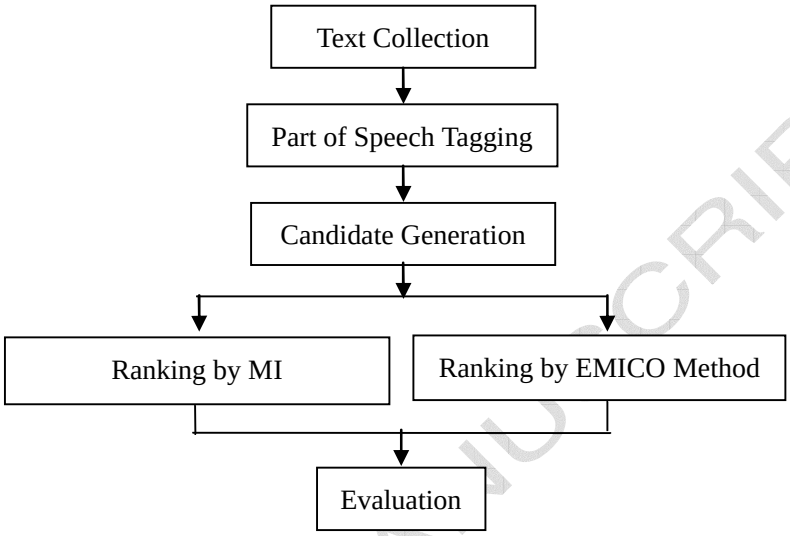

Figure 5 The framework for comparison between MI method and EMICO in substantival MWE extraction from Chinese and English text, respectively.

6.3 Evaluation

Table 4 is the Chinese substantival MWE candidates extracted by mutual information and EMICO, respectively. As for 20 instances with top association values, MI has extracted the MWE candidates with the following two characteristics: 1) the extracted MWE candidates have longer length, i.e. they contained more words than EMICO; 2) the frequency of a MWE candidate in text is almost equal to its two subordinates’ frequency such as “量变/和/质变” (3 times) to “量变” (3 times) and “和/质变” (3 times). For EMICO, the MWE it extracted has the following two characteristics: 1) it has shorter length compared with the MWE candidate from MI; 2) the frequencies of its subordinates have great difference and one of its subordinates has frequency almost as many as frequency of this MWE candidate. For instance, the frequency of “遗传性/疾病” is 5, the frequency of “遗传性” is 6 and the frequency of “疾病” is 283. We can explain the difference of MWE candidate length between MI and EMICO is made by two factors: the first is from the mechanism of MI used for extracted MWE candidates more than two words, that is, the longer is the candidate, the larger association value it will have in MI method; the second is from the mechanism of collocation optimization in EMICO, that is, it decided the “optimal” length of a MWE candidate based on association variation of candidates in a candidate set that makes the longer candidates having less possibility to be a MWE candidate than short candidates. The different characteristics in MWE candidate frequency between MI and EMICO can be explained as the result that EMICO has laid more emphasis on unsymmetrical co-occurrence when measuring word

pair's association than MI. As for the 20 MWE candidates with smallest association value, the extracted MWE candidates from both MI method and EMICO have very similar characteristics: the frequency of extracted MWE candidate is very small whereas the frequencies of its subordinates are relatively very large. For instance, in MI method, "探测/土星/卫星" has the frequency as 3 but "探测" has the frequency as 153 and "土星/卫星" has the frequency as 27. In EMICO method, "发电/燃料" has the frequency as 7 but "发电" has the frequency as 25 and "燃料" has the frequency as 79. Thus, we can regard the MWE candidates with smallest association values in both MI and EMICO as with independent subordinates.

Table 4. Chinese substantival MWE candidates extracted by mutual information and EMICO

	substantival MWEs extracted by mutual information	substantival MWEs extracted by EMICO
20 MWE candidates with largest association measures	探险/和/考察/的/一个/热点	国立/卫生/研究院
	圣/巴巴/拉/分校/alan/heeger/教授	遗传性/疾病
	例/prp/基因/129/密码	心血管/疾病
	聘请/第一/军医/大学/钟/世镇/院士	香山/科学/会议
	质变/和/量变	吴/旻/院士
	美国/Syracuse/大学/张/建舜/副教授	侯/云德/院士
	固态/液态/和/气态	顾/健人/院士
	和/首都/医科/大学/罗/述谦/教授	知识/产权
	廿一/世纪/世界/空间	锂离子/电池
	铵/硫酸/水泥	电光/调制器
	探究/海洋/生命/的/奥秘	憨/笨/院士
	在/具有/脊椎骨/这/一/动物/类型	红细胞/生成素
	有/斑岩/型/和/斑岩	水下/航行/器
	院士/首都/医科/大学/罗/述谦	解放军/总/医院
	绶/琯/院士	圣/巴巴/拉/分校
	量变/和/质变	理事会/理事长/程
	农林/废弃物/为主/以/低/质地	澜沧江/怒江
	具有/脊椎骨/这/一/动物/类型	大成/智慧/工程
	建立/神经/信息学/工作/平台/和/电子	昆明/动物/所
	廿一/世纪/世界/空间/天文台	前列腺/癌/发病率
20 MWE candiates with smallest association measures	以/形成/生物/多样性/研究	学科/交叉/发展
	来/纳米/生物/效应	序列/分析/蛋白质
	应力/下/深部	西部/开发/战略

微观/和/宏观/	如下/具体/建议
对/海洋/环境/的/需求	西线/调
能量/转移/荧光	它/可以/影响/天
能量/传输/过程/的/研究	有/中介/机构
原初/的/反应	它/处于/化学/生物
面临/着/许多/挑战	酶/或/受体
基/转移/酶	它/具有/灵敏度
优势/互补/资源	发电/燃料
探测/土星/卫星	活/细胞/单/分子
入/颗粒物/在/环境/影响	使/分散/资源
氧/生理学/与/医学/的/研究	单/分子/间/相互作用/方法
和/卫星/通信/产业	使/空气/净化
是/香山/科学/会议/国际	和/退化/危机
古/地理/与/沉积	与/超常/凝固/过程
通信卫星/是/通信	研究员/中科院/计算/所
中/的/浓度/分布	奋斗/目标
和/神经/递/质/受体	意大利/日本

Table 5. English substantival MWE candidates extracted by MI and EMICO

	substantival MWEs extracted by mutual information	substantival MWEs extracted by EMICO
20 MWE candidates with largest association measures	action/ec/industrial/union/leader/lord ray talking/us/regulators/trying/drive matches/orders/electronically/allows/anonymous/negotiation move/broaden/arbitraging/opportunities/sfe/traders capital/adequacy/ratio/goes/effect forced/reagan/automatically/impose/quotas/tariffs brokerage/firms/office/switzerland/clearing/members several/lawmakers/argued/new trade regan/protégé/sprinkles/chances roger/hemminghaus/chairman/refining/marketing/company packages/may/effect/essential/future/bank long/co/investors/more/interested/stock kangyos/increase/stake/chekiang/first/bank year tax/rates/greatly/increase/private/production result/consultations/countries/whether/corrective foreign/investors/brokerage/houses/us/oil/companies own/politicking/talking/us/regulators/trying/drive allies/reagan/administration/vice-president/george/bush/treasury intervention/so/many/nations/unprecedented/recent/years	burnham/lambert donaldson/Lufkin kidder/peabody harris/upham/co dai-ichi/kangyo/bank kidder/peabody/co marketing/years goldman/sachs/co cubic/feet jardine/fleming/securities bourses/zurich/geneva burnham/lambert/inc chairman/council ford/motor bourses/zurich/basle jardine/fleming bourgeois/liberalism drexel/burnham feedgrains/sorghum/barley/oats integrated/circuit

20 MWE candidates with smallest association measures	administrations/opposition	limitation/japanese/banks
	applications/supercomputers	referring/both/brazil
	yen/paris/accord	expand/international/prominence
	interest/rate/currency swap	tokyo/traders/holidays
	japan/us/west/germany	established/three-way
	us/japan/west/germany	half-point/cut
	treasury/secretary/baker	advisory/asset/management
	market/stop/selling	m0/measure
	greater/flexibility	paris/club/western
	opening/closing	across-the/board
	cojuangco/shares	paris/club/first
	speculative/selling	exposure/brazil
	higher/government/borrowing	james/capel/london
	Banking/supervisory/office	makers/boost/chip/imports
	nakasone/prime	james/baker/considered/tokyo
	producer/quota/shares	conduct/offshore/funding
	debt/forgiveness	currently/valued
	opposition/administrations	electric/credit/notes
	billion/yen/economic /package	guaranty/trust/co
	negotiations/international/coffee	james/baker/october

Table 5 is the English substantival MWE candidates extracted by MI and EMICO, respectively. For the top 20 MWE candiates, similarly with the situation in Chinese, MWEs extracted by MI has longer length than those from EMICO and their frequencies are at close ranges of their subordinates' frequencies. We also can use the same reasons as in Chinese to explain this outcome. For the last 20 MWE candidates, most candidates have similar characteristics as in Chinese. Moreover, we found an interesting case for MI method: it gave the MWE candidate "treasury/secretary/baker" a low score despite it is a right substantival MWE. We checked that the frequency as "treasury/secretary/baker" is 34, "treasury" is 138 and "secretary/baker" is 40. This is a typical case as unsymmetrical co-occurrence in word pair and a high association measure would be given to "treasury/secretary/baker" if EMICO is employed. Unfortunately, MI cannot deal with this kind of case.

Figure 6 is the performances of MI and EMICO on Chinese and English substantival MWE extraction, respectively. Recall and precision were used to measure their performances by comparing the extracted substantival MWE and standard substantival MWE in the randomly selected 30 documents for both Chinese and English. Here, for MI, CRL is the threshold of percentage at which point the candidates with larger association values were regarded as substantival MWEs. And for EMICO, CRL is used to eliminate candidates with smaller association values than the value at this point for further selection.

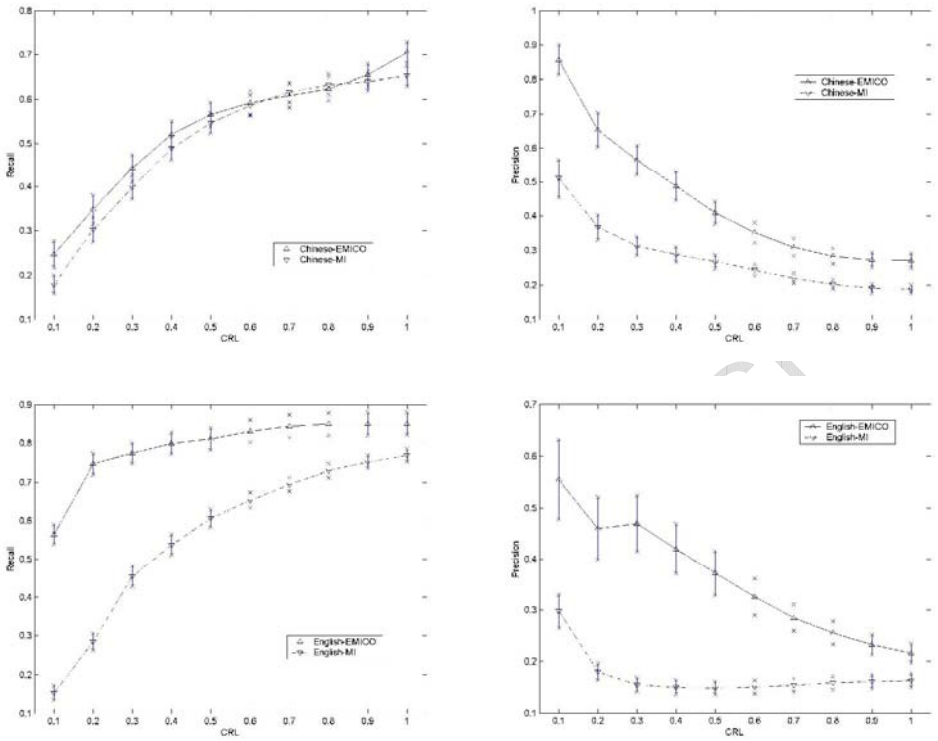

Figure 6 The performances of mutual information and EMICO for Chinese and English substantival MWE extraction at different CRLs, respectively.

We can see that the recall increases and precision declines when CRL is tuned increasingly from 0.1 to 1.0 for both Chinese and English. The smaller CRL with higher precision shows the effectiveness of both EMICO and MI method for MWE extraction. The precision produced by EMICO is convincingly better than that from mutual information method in both Chinese and English. In recall, EMICO also shows it superiority over mutual information method except two points on Chinese corpus. Furthermore, we can see that in precision, the difference between EMICO and mutual information method at small CRLs is larger than that in large CRLs. This outcome exactly illustrates that when CRL declines, EMICO removes the candidates which are not substantival MWEs while mutual information method removes candidates at equal probabilities among those candidates which are substantival MWEs and those candidates which are not substantival MWEs. It is worth noticing that the difference in both precision and recall between mutual information and EMICO is narrowed when CRL increase. We conjecture that this phenomenon happens because at large CRL level, all candidates have same possibility to be selected as substantival MWEs so that real substantival MWEs can not distinguish themselves from the false substantival MWEs. When CRL declines, the difference between the real and false substantival MWEs is become more and more significant.

7 Concluding remarks and future work

In this paper, a new approach, EMICO (Enhanced Mutual Information and Collocation Optimization) is proposed for substantival MWE extraction from texts. Specifically, EMI is proposed to measure association of word pair and collocation optimization is proposed to determine the optimal length of a MWE. With EMI, association of a word pair is measured by the ratio of the probability of the individual words' being a MWE to the probability of them not being a MWE. The benefits of EMI include the following two aspects. Firstly, it amplifies the significance of the intensively dependent word pairs, and distinguishes their association value from less dependent word pairs, not as uniformly distributed association value as in MI. Secondly, EMI solves unsymmetrical co-occurrence problem by synthetically considering the proportion of individual words' occurrence contributed to their common co-occurrence, not like the situation in mutual information. Collocation optimization, which is based on association variation of the candidates in the same candidate set, is proposed to determine the optimal length of a substantival MWE because we conceive that individual words in a MWE are prone to cluster together and thus the association of words in the MWE is intensified when an individual word which should be included in this MWE is included in.

To evaluate the performance of EMICO, we carried out a series experiments on the task of substantival MWE extraction on both Chinese and English documents. The experimental results demonstrate that, compared with MI, EMICO can improve the substantival MWE extraction performance singificantly. The better precision of EMICO illustrates that the association of individual words in a MWE is better characterized by EMICO rather than by mutual information. And the better recall of EMICO illustrates that EMICO has a greater potential than mutual information method to capture MWEs from texts. These two points indicate that EMICO is a promising statistical method for substantival MWE extraction. Although we only use EMICO for substantival MWE extraction for our research purpose, we argue that it also can be extended to extract MWEs in other types such as verbal phrases, preposition phrases, etc.

In future, we will use EMICO for MWE extraction other than substantival MWEs. To do this, we will combine EMICO with linguistic methods to improve performance of MWE extraction. Furthermore, we will use the MWEs extracted by our method for text categorization and information retrieval, so that the contextual 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 Grants No.70571078 and 70221001.

References

1. Firth, J. R.: A Synopsis of Linguistic Theory 1930-1955. Studies in Linguistic Analysis. Philological Society. Oxford: Blackwell. 1957.
2. Hawking, D., Thistlewaite, P.: Proximity Operators – So Near And Yet So Far. Proceedings of TREC-4, (1996) 131-144
3. Multiword Expression Project: http://mwe.stanford.edu/index.html

4. Zhang, W., Yoshida, T., Tang, X. J.: Text classification based on multi-word with support vector machine. Knowledge-based Systems. 2008, In Press.
5. Salton, G., Wang, A., & Yang, C. S.: A vector space model for information retrieval. Journal of the Society for Information Science, 18, 613-620.
6. Zhang, W., Yoshida, T., Ho, T. B., Tang, X. J.: Augmented mutual information for multi-word extraction. International Journal of Innovative Computing, Information and Control. 2008, In Press.
7. Zhang, W., Yoshida, T., Tang, X. J.: Multi-word extraction from Chinese text collection. Proceedings of APWeb'2008 Workshops. LCNS 4977, 2008. In Press.
8. Silva, J. F., Dias, G., Guillore, S., Lpoes, J. G. P.: Using LocalMaxs Algorithm for the extraction of contiguous and non-contiguous multiword lexical units. Progress in Artificial Intelligence, LNAI 1695, (1999) 113-132.
9. Church, K. W., Hanks, P.: Word association norms, mutual information, and lexicography. Computational Linguistics, 16(1), (1990) 22-29.
10. Pecina, P.: An extensive empirical study of collocation extraction methods. Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL, Sydney, Australia, (2006) 953 – 960.
11. Dias, G.: Multiword Unit Hybrid Extraction. In Proceedings of the Workshop on Multiword Expressions of the 41st Annual Meeting of the Association of Computational Linguistics. Sappora, Japan, (2003) 41-19.
12. Smadja, F.: Retrieving collocations from text: Xtract. Computational Linguistics. 19(1), (1993) 143-177.
13. Kita, K., Kato, Y., Omoto, Yano, Y.: A comparative study of automatic extraction of collocations from corpora: mutual information vs. cost criteria. Journal of Natural Language Processing, 1(1), (1992) 21-33.
14. Chen, B. X. and Du, L. M.: Preparatory work on automatic extraction of collocations from Corpora. Computational Linguistics and Chinese Language Processing. 1(1), (1992) 21-29.
15. Bourigault, D.: Surface Grammatical Analysis for the Extraction of Terminological Noun Phrases. Proceedings of the 14th International Conference on Computational Linguistics. Nantes, France, (1992) 977-981.
16. Justeson, J. S., Katz, S. M.: Technical terminology: some linguistic properties and an algorithm for indentification in text. Natural Language Engineering, 1(1), (1995) 9-27.
17. Argamon, S., I. Dagan, I., Krymolowski, Y.: A memory-based approach to learning shallow natural language patterns. Proceedings of the 17th international conference on Computational Linguistics, Montreal, Canada, (1998) 67-73.
18. Chen, K. H. and Chen, H. H.: Extracting Noun Phrases from Large-Scale Texts: A Hybrid Approach and Its Automatic Evaluation. In Proceedings of the 32nd Annual Meeting of Association of Computational Linguistics, New York, (1994) 234-241.
19. Pecina, P.: A machine learning approach to multiword expression extraction. Proceedings of Towards a Shared Task for Multiword Expressions Workshop (MWE 2008) at the sixth International conference on Language Resources and Evaluation (LREC 2008), Marrakech, Morocco, May 2008.
20. Duan, J . Y., Lu, R. Z., Wu, W. L., Hu, Y., Tian, Y.: A bio-inspired approach for multi-word expression extraction. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics table of contents, Michigan, USA, (2005) 605-613.
21. Smith, T. F. and Waterman, M. S.: Identification of common molecular subsequences. Journal of Molecular Biology, 147, (1981) 195-197.
22. Zhang, H. P., Liu, Q., Yu, H. K., Cheng, X. Q., Bai. S.: Chinese Named Entity Recognition Using Role Model. Computational Linguistics and Chinese Language Processing, 8 (2), (2003) 29-60.
23. Rabiner, L. R. and Juang. B. H.: An introduction to Hidden Markov Models. IEEE Signal Processing Magazine, (1986) 4-166.
24. 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 (1991) 40-62.

25. Maning, C. D. and Schutze, S.: Foundations of Statistical Natural Language Processing. MIT Press. Cambridge, Massachusetts. (2001). 178-183.
26. Jelinek, F.: Self-organized language modeling for speech recognition. Readings in Speech Recognition. Morgan Kaufmann Publishers (1990) 450-506.
27. Zhang, W., Tang, X. J., Yoshida, T.: Web Text Mining on A Scientific Forum. International Journal of Knowledge and System Sciences. 3(4) (2006) 51-59.
28. Zhang, W., Tang, X. J., Yoshida, T.: Text Classification Toward a Scientific Forum. Journal of Systems Science and Systems Engineering. 16(3) (2007) 356-369.