International Journal of Innovative
Computing, Information and Control
Volume **5**, Number **2**, February **2009**

 ISSN 1349-4198
pp. 543–554

AUGMENTED MUTUAL INFORMATION FOR MULTI-WORD EXTRACTION

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

[1]School of Knowledge Science
Japan Advanced Institute of Science and Technology
1-1 Asahidai, Tatsunokuchi, Ishikawa 923-1292, Japan
{ zhangwen; yoshida; bao }@jaist.ac.jp

Institute of Systems Science
[2]Academy of Mathematics and Systems Science
Chinese Academy of Sciences
Beijing 100190, P. R. China
xjtang@amss.ac.cn

Received January 2008; revised May 2008

Abstract. *In order to extract multi-words from documents, mutual information (MI), as a statistical method, is the most popular solution under consideration. However, there are two kinds of deficiencies inherent in MI. One is the problem of unilateral co-occurrence, and the other is rare occurrence problem. To attack these two problems, augmented mutual information (AMI) is proposed in this paper to measure word dependency for multi-word extraction. We prove theoretically that AMI has the capacity to approximate MI to capture the independency of individual words, but it will amplify the significance of dependent individual words which may be possible multi-words. And our experimental results on Chinese multi-word extraction demonstrate that AMI method has superior performance to traditional MI method.*
Keywords: Multi-word extraction, Mutual information, Augmented mutual information, Word dependency

1. **Introduction.** A word is characterized by the company it keeps [1]. That means not only the individual word but also its context should be emphasized for further processing. This simple and direct idea motivates the research on multi-words, which is expected to capture the context information of the individual words. Although multi-word has no satisfactory formal definition, it can be defined as a sequence of two or more consecutive individual words, which is a semantic unit, including steady collocations (e.g. proper nouns, terminologies, etc.) and compound words [2-4,7,10,11,16]. Usually, it is made up of a group of individual words, and its meaning is either changed to be entirely different from (e.g. collocation) or derived from the straight-forward composition of the meanings of its parts (e.g. compound phrase).

Generally speaking, there are mainly two types of methods developed for multi-word extraction. One is the linguistic method, which utilizes the structural properties of phrases and sentences to extract the multi-words from documents [2,3]. The other is the statistical method, based on corpus learning with mutual information for word occurrence pattern discovery [4,5]. There are also some other multi-word extraction methods which combine both linguistic knowledge and statistical computation [6-10]. However, as for the statistical methods for multi-word extraction, most of them employ MI directly, or an adaptation of MI without theoretical proof.

Our motivation in this paper is primarily concerned with the statistical method for multi-word extraction, i.e., to propose a more reliable method for word dependency measure to discriminate dependent word pairs from independent word pairs for multi-word extraction efficiently. With this motivation, we propose AMI, which is evolved from MI, for multi-word extraction, with the goal of overcoming the inherent deficiencies of MI regarding unilateral co-occurrence and rare occurrence.Intuitively, the key idea of AMI is that we measure the words' dependency, considering the possibility of their being a multi-word over the possibility of them not being a multi-word.

The benefit of AMI to attack unilateral co-occurrence is that it will amplify the word dependency for the word pair which might be a multi-word, especially for the word pair which has dominant co-occurrence. But for the word pair which does not have many co-occurrences, i.e., it has less possibility of being a multi-word; AMI has approximately the same ability as MI to measure its dependency as 0. The benefit of AMI to address the rare occurrence problem is in that AMI can reduce the influence of the rare occurrences when measuring the word's dependency, not the situation that the dependency is dominated by the rare occurrences as in MI.

We prove mathematically that AMI has better performance than MI under the condition of words' dependency, while it has approximately the same effectiveness as MI on the condition of words' independency. Experimental results demonstrate that AMI outperforms classical MI, gauging by precision and recall, when smaller and smaller numbers of candidates with highest AMI and MI scores respectively are retained for multi-word selection, more superiority is indicated in AMI.

The rest of this paper is organized as follows. Section 2 provides a review of MI and presents its two deficiencies as unilateral co-occurrence and rare occurrence problem. Section 3 propose the AMI method, and it is proved theoretically not only to have the capability to capture the independency of individual words but also to amplify the significance of dependent individual words which are likely to be a multi-word. Section 4 gives a further explanation of AMI's superiority over MI for word dependency measure. Section 5 shows the experiments of Chinese multi-word extraction to validate that AMI method is superior to MI method in practical application. Section 6 concludes this paper and indicates our further research.

2. 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 propose the association ratio for measuring word association based on the information theoretic concept of mutual information [11]. In their method, the MI between word x and y was defined as Eq.(1).

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

$P(x)$ is the occurrence probability of term x, $P(y)$ is the occurrence probability of word y in a corpus and $P(x,y)$ is the co-occurrence probability of words sequence (x,y).

The primary reason for applying MI to multi-word extraction is that it has 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, $P(x,y) > P(x)P(y)$ so $I(x,y) > 0$, if x and y are dependent on each other. The higher MI of a word pair, the more genuine is the association between two words.

However, there are mainly two deficiencies inherent in MI for measuring the words' dependency. The first one is the unilateral co-occurrence problem, that is, it only considers the co-occurrence of two words, while ignoring those cases that one is present while the other is absent [12,13]. The second deficiency of MI method concerns the rare occurrence problem [13]. 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. Although rare occurrence is a hard problem for linguistic data, and usually there is no effective remedy for it, we may attempt to reduce its passive influence in word dependency measure. 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 "bad" ones [10].

$$MI(x_1, x_2, ..., x_n) = Max_{1 \le m \le n} log_2 \frac{P(x_1, x_2, ..., x_n)}{P(x_1)P(x_2)...P(x_n)} \tag{2}$$

where m is the breakpoint of multi-word which separates $(x_1, x_2, ..., x_n)$ into two meaningful parts, $(x_1, x_2, ..., x_m)$ and $(x_{m+1}, x_2, ..., x_n)$. Moreover, we need to determine whether or not $(x_1, x_2, ..., x_m)$ and $(x_{m+1}, x_2, ..., x_n)$ are two meaningful words or word combinations by looking up the single word set and multi-word candidate set we established later. With the maximum likelihood estimation, $P(x_1, x_2, ..., x_n) = F(x_1, x_2, ..., x_n)/N$ (N is the total word count in the corpus), so the MI method can be rewritten as follows.

$$\begin{aligned} MI &= (n-1)log_2 N \\ &+ Max_{1 \le m \le n}\{log_2 F(x_1, ..., x_n) - log_2 F(x_1, ..., x_m) - log_2 F(x_{m+1}, ..., x_n)\} \end{aligned} \tag{3}$$

The traditional MI score method for the multi-word candidate ranking in this paper is based on Eq.(1).

3. **Augmented Mutual Information.** As mentioned in Section 2, MI has two inherent primary deficiencies. One is the unilateral co-occurrence problem, and the other is rare occurrence problem. To attack the unilateral co-occurrence problem, not only the co-occurrences, but also their individual occurrences excluding their co-occurrences, which are the number of cases when one occurs while the other one is absent, should be considered particularly. AMI is proposed and defined as the ratio of the probability of word pair occurrence over the product of the probabilities of absences of both individual words, i.e., the possibility of being a multi-word over the possibility of not being a multi-word. It has the mathematic formula described in Eq.(4).

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

AMI has approximately the same capability for characterizing the word pair's independence as MI. But in the case of word pair's dependency with positive correlation, which means that the word pair is highly likely to be a multi-word, AMI will amplify the difference between "true" dependency and "false" dependency, which is caused by unilateral co-occurrence.

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

$$AMI(x,y) \approx log_2 \frac{P(x,y)}{P(x)P(y)} = I(x,y) = 0 \tag{5}$$

Eq.(5) means that $AMI(x,y)$ has approximately the same competence with MI when x and y are independent of each other.

For the dependent case between two words x and y in a bi-gram, generally, the dependence relationship can be divided into two situations, as negative correlation and positive correlation among x and y. Negative correlation between them is meaningless for multi-word extraction, as x and y would co-occur quite a few times in this case. Positive correlation between x and y is an omen that x and y could constitute a multi-word. In the case of positive correlation between x and y, $P(x,y) > P(x)P(y)$ and

$$\begin{aligned} AMI(x,y) &> log_2 \frac{P(x,y)}{(P(x) - P(x)p(y))(P(y) - P(x)P(y))} \\ &= I(x,y) + log_2 \frac{1}{(P(\overline{x}))(P(\overline{y}))} = 0 \end{aligned} \tag{6}$$

Eq.(6) means that AMI will amplify the significance of the possible multi-word candidate. Combining Eq.(5) and Eq.(6), we can conclude that the possible multi-word candidate would be distinguished from those candidates whose parts are independent from each other. This analysis is the exact theoretical motivation of AMI.

In order to extend AMI to rank multi-word candidate 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. If we follow the idea of MI method, the solution will be as follows.

$$AMI(x,y,z) > log_2 \frac{P(x,y,z)}{(P(x,y) - P(x,y,z))(P(z) - P(x,y,z))} \tag{7}$$

However, there is another intrinsic problem with formula (7), the longer the sequence, the larger AMI is, because its value is dominated by the smallest occurrence among x, y and z. For example, if x occurred rarely, $P(x,y)$ should be very small, even if y occurs frequently, so $P(x,y) - P(x,y,z)$ is 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 higher AMI values than those without. That is, if one word is a rare occurrence, this rare occurrence will reduce the occurrences of the component extremely it is contained. Although the rare occurrence problem is a hard nut to crack unless heuristics is involved, we can alleviate it to some extent. Eq.(8) is our solution for ranking the multi-word candidate of more than two words.

$$AMI(x,y,z) > log_2 \frac{P(x,y,z)}{(P(x) - P(x,y,z))(P(y) - P(x,y,z))(P(z) - P(x,y,z))} \tag{8}$$

We can expect that if there is a rare occurrence of x, but if y and z have many occurrences, the AMI from Eq.(8) will be less influenced by x than that from Eq.(7).

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

$$AMI(x_1, x_2, ..., x_n) = log_2 \frac{p}{(p_1 - p)(p_2 - p)...(p_n - p)} \tag{9}$$

By maximum likelihood estimation,

$$AMI(x_1, x_2, ..., x_n) = (n-1)log_2 N + log_2 F - \sum_{i=1}^{n} log_2(F_i - F) \tag{10}$$

N is the number of words contained in the corpus, it is usually a large value, more than 10^6. In Eq.(10), $log_2 N$ actually can be regarded as how much the AMI value will be increased when one more word is added to the candidate. Because $log_2 N$ is a large value and it makes the AMI is primarily dominated by the length of sequence. It is not suitable in our method so $log_2 N$ is replace by α which is the weight of length in a sequence. Another problem with Eq.(10) is that in some special cases we have $F_i = F$ then $F_i - F = 0$, these special cases would make Eq.(10) meaningless. For this reason, Eq.(10) is rewritten as Eq.(11).

$$AMI(x_1, x_2, ..., x_n) = (n-1)\alpha + log_2 F - \sum_{i=1}^{m} log_2(F_i - F) + (n-m)\beta \tag{11}$$

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, as "Lean" to "Prof. J. M. Lean".

4. **The Advantages of AMI for Multi-word Extraction.** Currently, multi-word is extracted from documents using the traditional MI method. The primary intention of this paper is to propose AMI for multi-word extraction. Although the theoretical proof of AMI's superiority over MI is specified in Section 3, we would like to explain AMI's mechanism more concretely and more pellucid.

Actually, the advantage of AMI over MI can be clarified in two aspects. The first one is that AMI put more emphasis on the co-occurrence of x and y than that of MI. Figure 1 shows the variation trend of AMI and MI with co-occurrence frequency, respectively. We can see that the dependency measured by MI is linearly increased with the co-occurrence frequency and the dependency measured by AMI is acceleratedly increased with the co-occurrence frequency. Thus, it can be concluded that the differences of multi-word candidates' dependencies in AMI will be much larger than that of MI. For this reason, the candidates with high co-occurrence will be identified more easily with AMI measure than those with MI.

The second one is that the proportions of x and y's occurrences contributing to their co-occurrence are integrated to measure the dependency of them. According to the rewritten formation of Eq.(4) as Eq.(12), $\frac{P(x,y)}{P(x)}$ and $\frac{P(x,y)}{P(y)}$ are proportions of x's occurrence and Y's occurrence contributing to their co-occurrence, respectively and AMI will increase when the proportions increase. This point can be illustrated in Figure 2 as a contrast of MI and AMI. We can see from it that the profile of MI with respect to x and y is a plane and the profile of AMI is a curve surface with sharp increase at the edges of x and y. The sharp increase in AMI can produce different measure results in MI. Taking word pairs (x, y) and (x', y') in Figure 2 for example, we can see that in MI, (x, y) has larger dependency than (x', y') and the result is converse when it comes to AMI. The reason for this outcome is

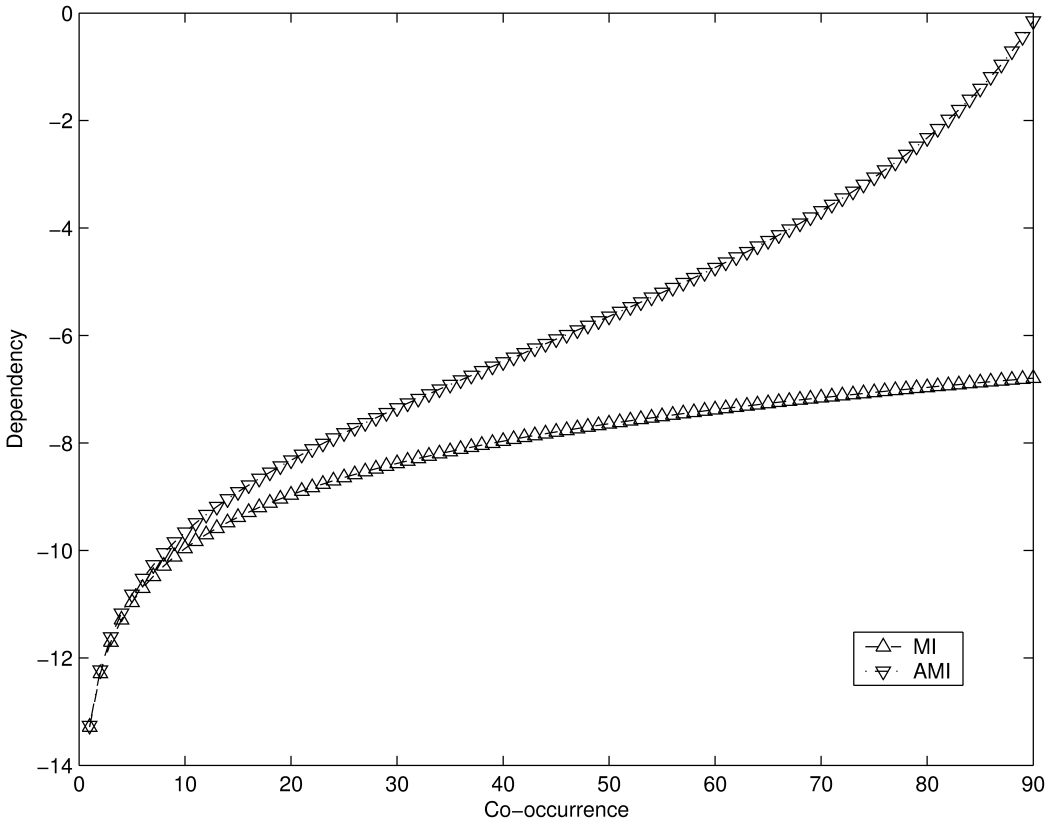

FIGURE 1. Plots of variation trend of dependency value of MI and AMI. The frequency of X and Y are fixed as 100.

that (x', y') is nearer to the edge part of the curve surface than (x, y) which is at central part of the plane.

$$AMI(x,y) = log_2 \frac{P(x,y)}{P(x)P(y)(1-\frac{P(x,y)}{P(x)})(1-\frac{P(x,y)}{P(y)})} \tag{12}$$

5. **Experiment.** In this section, a series of experiments are carried out to compare the AMI and traditional MI on multi-word extraction from a Chinese text collection. Only the noun multi-words are under consideration for the comparison, because there are plenty of them in our documents and they can be easily identified manually for constructing the standard noun multi-word base to evaluate the extraction performance of AMI and MI. The classic performance measures in information retrieval, recall, precision and F-measure, are used to evaluate the performance of multi-word extraction of AMI and MI methods at different parameter settings.

5.1. **Design of experiment.** It should be pointed out that in the experiments; the multi-words we want to extract from documents currently are noun phrases. Furthermore, the multi-words are extracted from documents with the traditional n-gram method [6] plus root noun extension strategy. That is, firstly, a noun is located in a sentence in the document; next, the words before this noun are also captured to constituent a multi-word candidates; then, MI and AMI are used to give the dependencies of these candidates, respectively; finally, the candidates with dependencies above a predefined threshold are regarded as multi-words in the documents. Please write down your subsection.

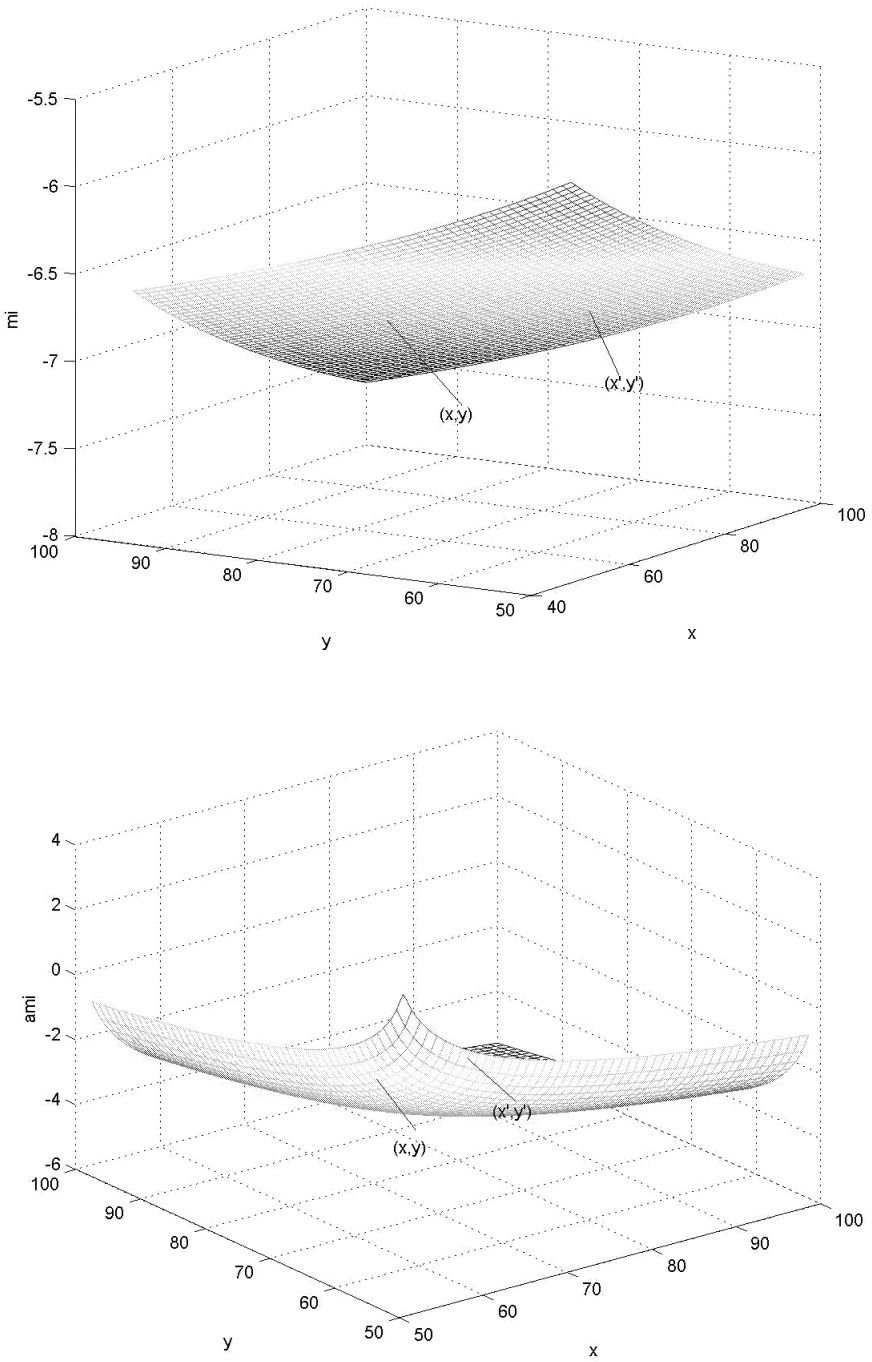

FIGURE 2. The contrast of MI (above) and AMI (below) for dependency measure of $(x, y), (x', y')$, respectively. The co-occurrence frequency is fixed as 50 and the frequencies of two components are varying in their axes.

The initial conditions for multi-word extractions include five aspects as follows: 1) a corpus containing enough number of documents; 2) for character-based language, a morphological analysis tool is needed to segment the sentence into meaningful words; 3) dependency measure for rank the dependencies of candidates; 4) a benchmark multi-word set used for performance evaluation; 5) For AMI formula in Eq.(9), α and β are predefined as 3.0 and 0.

In the experiments, the following method is used to extract the multi-words from documents using MI and AMI measure, respectively. Actually, we knew this method from [14] and [15].

1) Start out from the basic vocabulary V_0 , set $n = 0$; 2) Load the vocabulary V_n , by all word sequences "x y" for which $MI(x, y) > Thr$ for MI. (or $AMI(x, y) > Thr$ for using AMI), Thr is a predefined threshold for word sequence dependency; 3) From Step 2, a new vocabulary V_{n+1} is established. 4) Resume from Step 1 with V_{n+1} as its basis. 5) Realign the multi-words included in the longer multi-words. For instance, "医学情报 " ("medical information") will be deleted from the multi-word vocabulary if " 医学情报处理" ("medical information processing") is extracted in the multi-word vocabulary. 6) Dispatch the multi-words in the final vocabulary into documents in which they occurred.

5.2. **Chinese text collection.** Based on our previous work on text mining [16, 17, and 18], 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 (terminologies, noun phrases, etc.) in these documents. For all these documents, there are 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. All the 184 documents are used for corpus learning to extract Chinese multi-words. However, because there lacks of a standard multi-word base for all documents in our text collection, only 30 of 184 documents are fetched out randomly from the text collection and a benchmark multi-word dataset is established manually to estimate the performances of AMI method and MI method in multi-word extraction form this text collection. Table 1 is the basic information of our benchmark multi-word dataset.

5.3. **Candidate generation.** If we have a sentence after morphological analysis such as "A B C DE F G H." and H is found to be a noun in this sentence, 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 multi-word usually has a length of 2-6 words.

Definition 5.1. *Candidate Set is a word sequence set whose elements are generated from the same root noun in a sentence using n-gram method.*

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

In order to gauge the performances of AMI method and MI method at different parameter settings, we set a predefined proportion of all the candidates to retain the multi-word candidates with highest AMI or MI value for further multi-word selection, and the remaining candidates with low AMI or MI will be removed from possible candidates.

Definition 5.2. *Candidate Retaining Level (CRL) is a predefined proportion at which point the multi-word candidates with highest AMI or MI value will be retained for further selection.*

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

TABLE 1. Basic information of the benchmark multi-word set. Only some examples in the benchmark data set are given due to the space limitation.

DocNo.	*#ofMW*	*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	25	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
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

5.4. **Evaluation.** In order to observe the performance of AMI method and MI method at different parameter settings, CRL is varied at different ratio, as 70%, 50% and 30% for comparison. Moreover, in order to match the multi-words given by AMI or MI method and the multi-words given by human experts to determine whether the extracted multi-words are correct, approximate matching is utilized as follows.

Definition 5.3. *approximate matching: assuming that a multi-word is retrieved from a candidate set as , and another multi-word, is given by human identification, we regard them as the same one if* $\frac{|m_1 \bigcap m_2|}{|m_1 \bigcup m_2|} \geq \frac{2}{3}$.

The reason for adopting approximate matching is that there are certainly some trivial differences between the multi-words given by our methods and the multi-words provided by human identification, because humans have more "knowledge" about the multi-word than the "knowledge" integrated into our methods, such as common sense, background context, etc. Taking the Chinese name for example, human beings can easily identify that the family name is a part of a full name, but it is not so easy for the full name to

be extracted perfectly by neither AMI nor MI, because many persons may have the same family name, the result is that AMI or MI will be decreased if a family name is added to the extracted name as a multi-word.

Tables 2-3 show the evaluation results of multi-word extraction methods of AMI method and MI method. We can see from Table 2 that under the condition of CRL as 0.7, 0.5 and 0.3, AMI has a better average recall than MI. Moreover, at CRL as 0.7, multi-word extraction on 21 (including equivalents) of 30 documents has obtained better recall with AMI than MI. At CRL as 0.5, multi-word extraction on 24 of 30 documents has obtained better (including equivalents) recall with AMI than MI. At CRL as 0.3, multi-word extraction on 26 of 30 documents has obtained better (including equivalents) recall with AMI than MI. In a word, AMI has shown its superiority on recall of multi-word extraction in all of the indicators at all CRLs except the maximum recall at CRL 0.7 and minimum recall at CRL 0.3.

It is shown in Table 3 that in all CRLs, AMI has better performances than MI in terms of average precision, maximum precision and minimum precision. In details, at CRL as 0.7, multi-word extraction on 19 documents has obtained better performance with AMI than MI. At CRL as 0.5, AMI has produced better performance (including equivalents) in multi-word extraction on 24 documents than MI, which is much larger than the number of documents in which MI's performance is better than AMI's. At CRL as 0.3, AMI has obtained better performance (including equivalents) with on 26 documents than MI, which is also much larger the number of documents in which MI has obtained better performance than AMI. These are all the evidences for that AMI can produce better performance than MI in multi-word extraction.

TABLE 2. Recalls of AMI method and MI method on multi-word extraction from Chinese texts at different CRLs.

	AMI Method			*MI Method*		
CRL	Ave-Rec	Max-Rec	Min-Rec	Ave-Rec	Max-Rec	Min-Rec
0.7	0.8231	0.9194	0.6950	0.7871	0.9420	0.6521
0.5	0.6356	0.7741	0.4528	0.5790	0.7307	0.3773
0.3	0.3878	0.5806	0.1132	0.2652	0.5363	0.1538

TABLE 3. Precisions of AMI method and MI method on multi-word extraction from Chinese texts at different CRLs.

	AMI Method			*MI Method*		
CRL	Ave-Pre	Max-Pre	Min-Pre	Ave-Pre	Max-Pre	Min-Pre
0.7	0.2193	0.3592	0.1405	0.2094	0.3317	0.1016
0.5	0.2497	0.4304	0.1497	0.2174	0.3676	0.1317
0.3	0.2930	0.5142	0.2002	0.2375	0.4736	0.1111

The greatest average recall is obtained as 0.8231 at CRL 0.7 with AMI method, and the greatest precision is obtained as 0.2930 at CRL 0.3 also with AMI method. In contrast, the least recall, 0.2652, is obtained at CRL 0.3 with MI method, and the least precision is obtained, 0.2094, at CRL 0.7, also with MI method.

Furthermore, for both AMI and MI, it can be seen that the recall decreases and precision increases when CRL declines from 0.7 to 0.3. The decrease of recall can be explained as a result from that less and less of candidates are retained fro selection during this process. The increase in precision clarifies that multi-words actually has higher AMI or MI value

than the candidates which are not a multi-words. Moreover, the performance differences at which AMI lags behind AMI are becoming larger and larger when CRL is varied from 0.7 to 0.3. This point illustrates that the performance superiority of AMI over MI is more and more significant when CRL becomes smaller and smaller.

The greater value of recall validates that AMI method can retrieve more multi-words. The greater value of precision of AMI method reflects that it generates more accurate output as multi-words than MI method. And, the greater value in F-measure manifests that the AMI method can achieve more reliable results.

6. **Concluding Remarks and Future Work.** In this paper, a statistical method, AMI, is proposed to rank the dependency of individual words for multi-word extraction. The key idea of AMI is that we measure the words' dependency considering the possibility of their being a multi-word over the possibility of them not being a multi-word.

For the problem of unilateral co-occurrence, the probability of co-occurrence is subtracted from the probabilities of occurrences of individual words respectively, so that only the occurrences of individual words in a multi-word candidate when they do not co-occur are considered to measure their dependency. For the problem of rare occurrence, the AMI method is designed to consider the individual words separately, other than as the traditional two parts as required in MI method. Although our solution can not overcome the rare occurrence problem completely, it can alleviate the influence of rare occurrence as an inherent deficiency caused by the randomness of linguistic data.

Furthermore, we prove mathematically that AMI has the capacity approximately the same as MI for measuring the independent individual words, but AMI amplifies the significance of the dependent individual words, which may be combined a multi-word. Moreover, we conduct a series of experiments to extract multi-words from a Chinese text collection. The experimental results are consistent with our theoretical analysis that AMI method is superior to MI method.

As far as our future work is concerned, multi-word extraction is still of our interest. We will combine the statistical and linguistic methods based on their superiorities in multi-word extraction, and extend our work of multi-word extraction to English text corpus. More experiments will be conducted to validate our hypotheses, especially on the solution of rare occurrence problem. Furthermore, we will use the multi-words for the task of text classification [19] and some real applications such as [20, 21], so that the context knowledge can be integrated into practical intelligent information processing applications.

Acknowledgment. This work is supported by Ministry of Education, Culture, Sports, Science and Technology of Japan under the "Kanazawa Region, Ishikawa High-Tech Sensing Cluster of Knowledge-Based Cluster Creation Project" and partially supported by the National Natural Science Foundation of China under Grant No.70571078 and 70221001. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] J. R. Firth, *A Synopsis of Linguistic Theory 1930-1955, Studies in Linguistic Analysis*, Philological Society, Oxford, Blackwell, 1957.

[2] J. S. Justeson and S. M. Katz, Technical terminology: Some linguistic properties and an algorithm for identification in text, *Natural Language Engineering*, vol.1, no.1, pp.9-27, 1995.

[3] D. Bourigault, Surface grammatical analysis for the extraction of terminological noun phrases, *Proc. of the 14th International Conference on Computational Linguistics*, Nantes, France, pp.977-981, 1992.

[4] F. Smadja, Retrieving collocations from text: Xtract, *Computational Linguistics*, vol.19, no.1, pp.143-177, 1993.
[5] K. W. Church and L. M. Robert, Introduction to special issue on computational linguistics using large corpora, *Computational Linguistics*, vol.19 no.1, pp.1-24, 1993.
[6] J. S. Chen, C. H. Yeh and R. N. Chau, Identifying multi-word terms by text-segments, *Proc. of the Seventh International Conference on Web-Age Information Management Workshops*, Hong Kong, pp.10-19, 2006.
[7] Y. J. Park, R. J. Byrd and K. B. Boguraev, Automatic glossary extraction: Beyond terminology identification, *Proc of the 19th International Conference on Computational linguistics*, Taiwan, pp.1-17, 2002.
[8] J. S. Chang, S. D. Chen, S. J. Ker, Y. Chen, J. Liu, A multiple-corpus approach to recognition of proper names in Chinese texts, *Computer Processing of Chinese and Oriental Languages*, vol.8, no.1, pp.75-85, 1994.
[9] I. Fahmi, C value method for multi-word term extraction, *Seminar in Statistics and Methodology.* Alfa-informatica, RuG, May 23, 2005. *online: http://odur.let.rug.nl/fahmi/ talks/statistics-c-value.pdf*
[10] B. Daille, E. Gaussier and J. M. Lange, Towards automatic extraction of monolingual and bilingual terminology, *Proc. of the International Conference on Computational Linguistics*, Kyoto, Japan, pp.93-98, 1994
[11] K. W. Church and P. Hanks, Word association norms, mutual information, and lexicography, *Computational Linguistics*, vol.16, no.1, pp.22-29, 1990.
[12] K. W. Church and A. G. William, Concordances for parallel text, *Proc. of the Seventh Annual Conference of the UW Center for the New OED and Text Research*, Oxford, pp.40-62, 1991.
[13] C. D. Manning and H. Schutze. *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge, Massachusetts, 2001.
[14] K. Kita, Y. Kato, T. Omoto and Y. Yano, A comparative study of automatic extraction of collocations from Corpora: Mutual information vs. cost criteria, *Journal of Natural Language Processing*, vol.1, no.1, pp.21-29, 1992.
[15] F. Jelinek, Self-organized language modeling for speech recognition, in *Readings in Speech Recognition*, A. Waibel and K. F. Lee (eds.), Morgan Kaufmann Publishers, pp.450-506, 1990.
[16] W. Zhang, X. J. Tang and T. Yoshida, Web text mining on a scientific forum, *International Journal of Knowledge and System Sciences*, vol.3, no.4, pp.51-59, 2006.
[17] W. Zhang, X. J. Tang and T. Yoshida, Text classification toward a scientific forum, *Journal of Systems Science and Systems Engineering*, vol.16, no.3, pp.356-369, 2007.
[18] W Zhang, T. Yoshida and X. J. Tang, A study on multi-word extraction from Chinese documents, *Proc. of the 10th Asia-Pacific Web Conference Workshops*, Shenyang, China, 2008.
[19] W. Zhang, T. Yoshida and X. J. Tang, Text classification based on multi-word with support vector machine, *Knowledge-based Systems*, in press, 2008.
[20] G. T. Raju, P. S. Satyanarayana and L. M. Patnaik, Knowledge discovery from web usage data: Extraction and applications of sequential and clustering patterns - A survey, *International Journal of Innovative Computing, Information and Control*, vol.4, no.2, pp.381-389, 2008.
[21] D. B. Bracewell, J. J. Yan and F. J. Ren, Single document keyword extraction for internet news articles, *International Journal of Innovative Computing, Information and Control*, vol.4, no.4, pp.905-913, 2008.