Using Synonym Relations In Chinese Collocation Extraction

Wanyin Li
Department of Computing,
The Hong Kong Polytechnic University,
Hung Hom, Kowloon, Hong Kong
cswyli@comp.polyu.edu.hk

Qin Lu
Department of Computing,
The Hong Kong Polytechnic University,
Hung Hom, Kowloon, Hong Kong
csluqin @comp.polyu.edu.hk

Ruifeng Xu
Department of Computing, The Hong Kong Polytechnic University,
Hung Hom, Kowloon, Hong Kong
csrfxu@comp.polyu.edu.hk

Abstract

A challenging task in Chinese collocation extraction is to improve both the precision and recall rate. Most lexical statistical methods including Xtract face the problem of unable to extract collocations with lower frequencies than a given threshold. This paper presents a method where HowNet is used to find synonyms using a similarity function. Based on such synonym information, we have successfully extracted synonymous collocations which normally cannot be extracted using the lexical statistical approach. We applied synonyms mapping to each headword to extract more synonymous word bi-grams. Our evaluation over 60MB tagged corpus shows that we can extract synonymous collocations that occur with very low frequency, sometimes even for collocations that occur only once in the training set. Comparing to a collocation extraction system based on Xtract, we have reached the precision rate of 43% on word bi-grams for a set of 9 headwords, almost 50% improvement from precision rate of 30% in the Xtract system. Furthermore, it improves the recall rate of word bi-gram collocation extraction by 30%.

1 Introduction

A Chinese collocation is a recurrent and conventional expression of words which holds syntactic and semantic relations. A widely adopted definition given by Benson (Benson 1990) stated that "*a collocation is an arbitrary and recurrent word combination*." For example, we say "warm greetings" rather than "hot greetings", "broad daylight" rather than "bright daylight". Similarly, in Chinese "行李" "包裹" "包袱" are three nouns with similar meanings, however, we say "思想包袱" rather than "思想行李", "托运行李"rather than "托运包袱".

Study in collocation extraction using lexical statistics has gained some insights to the issues faced in collocation extraction (Church and Hanks 1990, Smadja 1993, Choueka 1993, Lin 1998). As the lexical statistical approach is developed based on the "recurrence" property of collocations, only collocations with reasonably good recurrence can be extracted. Collocations with low occurrence frequency cannot be extracted, thus affecting the recall rate. The precision rate using the lexical statistics approach can reach around 60% if both word bi-gram extraction and n-gram extractions are taking into account (Smadja 1993, Lin 1997 and Lu et al. 2003). The low precision rate is mainly due to the low precision rate of word bi-gram extractions as only about 30% - 40% precision rate can be achieved for word bi-grams.

In this paper, we propose a different approach to find collocations with low recurrences. The main idea is to make use of synonym relations to extract synonymous collocations. Lin (Lin 1997) described a distributional hypothesis that if two words have similar set of collocations, they are probably similar. In HowNet, Liu Qun (Liu et al. 2002) defined the word similarity as two words that can substitute each other in the context and keep the sentence consistent in syntax and semantic structure. That means, naturally, two similar words are very close to each other and they can be used in place of the other in certain context. For example, we may either say "买书"or "订书" as "卖" and "订" are semantically close to each other. We apply this lexical phenomenal after the lexical statistics based extractor to find the low frequency synonymous collocations, thus increasing recall rate.

The rest of this paper is organized as follows. Section 2 describes related existing collocation extraction techniques based on both lexical statistics and synonymous collocation. Section 3 describes our approach on collocation extraction. Section 4 evaluates the proposed method. Section 5 draws our conclusion and presents possible future work.

2 Related Work

Methods have proposed to extract collocations based on lexical statistics. Choueka (Choueka 1993) applied quantitative selection criteria based on frequency threshold to extract adjacent n-grams (including bi-grams). Church and Hanks (Church and Hanks 1990) employed mutual information to extract both adjacent and distant bi-grams that tend to co-occur within a fixed-size window. But the method did not extend to extract n-grams. Smadja (Smadja 1993) proposed a statistical model by measuring the spread of the distribution of co-occurring pairs of words with higher strength. This method successfully extracted both adjacent and distant bi-grams and n-grams. However, the method failed to extract bi-grams with lower frequency. The precision rate on bi-grams collocation is very low, only around high 20% and low 30%. Even though, it is difficult to measure recall rate in collocation extraction (almost no report on recall estimation), It is understood that low occurrence collocations cannot be extracted. Our research group has further applied the Xtract system to Chinese (Lu et al. 2003) by adjusting the parameters to optimize the algorithm for Chinese and a new weighted algorithm was developed based on mutual information to acquire word bi-grams with one higher frequency word and one lower frequency word. The result has achieved an estimated 5% improvement in recall rate and a 15% improvement in precision comparing to the Xtract system.

All of the above techniques do not take advantage of the wide range of lexical resources available including synonym information. Pearce (Pearce 2001) presented a collocation extraction technique that relies on a mapping from a word to its synonyms for each of its senses. The underlying intuitions is that if the difference between the occurrence counts of one synonyms pair with respect to a particular word was at least two, then this was deemed sufficient to consider them as a collocation. To apply this approach, knowledge in word (concept) semantics and relations to other words must be available such as the use of WordNet. Dagan (Dagan 1997) applied similarity-based smoothing method to solve the problem of data sparseness in statistical natural language processing. The experiments conducted in his later works showed that this method achieved much better results than back-off smoothing methods in word sense disambiguation. Similarly, Hua Wu (Wu and Zhou 2003) applied synonyms relationship between two different languages to automatically acquire English synonymous collocation. This is the first time that the concept synonymous collocation is proposed. A side intuition raised here is that nature language is full of synonymous collocations. As many of them have low occurrences, they are failed to be retrieved by lexical statistical methods. Even though there are Chinese synonym dictionaries, such as 《同义辞林》 (Tong Yi Ci Lin), the dictionaries lack structured knowledge and synonyms are too loosely defined to be used for collocation extraction.

HowNet developed by Dong et al (Dong and Dong 1999) is the best publicly available resource on Chinese semantics. By making use of semantic similarities of words, synonyms can be defined by the closeness of their related concepts and the closeness can be calculated. In Section 3, we present our method to extract synonyms from HowNet and using synonym relations to further extract collocations.

Sun (Sun 1997) did a preliminary *Quantitative* analysis on Chinese collocations based on their arbitrariness, recurrence and the syntax structure. The purpose of this study is to help differentiate if a collocation is true or not according to the quantitative factors. By observing the existence of synonyms information in natural language use, we consider it possible to identify different types of collocations using more semantic and syntactic information available. We discuss the basic ideas in section 5..

3 Our Approach

Our method of extracting Chinese collocations consists of three steps.

Step 1: Take the output of any lexical statistical algorithm which extracts word bi-gram collocations. The data is then sorted according to each headword , W_h, with its co-word, W_c, listed.

Step 2: For each headword W_h used to extract bi-grams, we acquire its synonyms based on a similarity function using HowNet. Any word in HowNet having similarity value over a threshold value is chosen as a synonym headword W_s for additional extractions.

Step 3: For each synonym headword, W_s, and the co-word W_c of W_h, as its synonym, if the bi-gram (W_s , W_c) is not in the output of the

lexical statistical algorithm in Step one, take this bi-gram (W_s , W_c) as a collocation if the pair co-occurs in the corpus by additional search to the corpus.

3.1 Structure of HowNet

Different from WordNet or other synonyms dictionary, HowNet describes words as a set of concepts(义项) and each concept is described by a set of primitives(义元). The following lists for the word "打", one of its corresponding concepts

```
NO.=017144
W_C=打
G_C=V
E_C=~网球，~牌，~秋千，~太极，球~得很棒
W_E=play
G_E=V
E_E=
DEF=exercise|锻炼,sport|体育
```

In the above record, DEF is where the primitives are specified. DEF contains up to four types of primitives: the *basic independent primitive* (基本独立义元), the *other independent primitive*(其他独立义元), the *relation primitive* (关系义元), and the *symbol primitive*(符号义元), where the basic independent primitive and the other independent primitive are used to indicate the semantics of a concept and the others are used to indicate syntactical relationships. The similarity model described in the next subsection will consider both of these relationships.

The primitives are linked by a hierarchical tree to indicate the parent-child relationships of the primitives as shown in the following example:

This hierarchical structure provides a way to link one concept with any other concept in HowNet, and the closeness of concepts can be simulated by the distance between two concepts.

3.2 Similarity Model Based on HowNet

Liu Qun (Liu 2002) defined word similarity as two words which can substitute each other in the same context and still maintain the sentence consistent syntactically and semantically. This is very close to our definition of synonyms. Thus we directly used their similarity function, which is stated as follows.

A word in HowNet is defined as a set of concepts and each concept is represented by primitives. Thus, HowNet can be described by W, a collection of n words, as:

$W = \{ w_1, w_2, \ldots w_n\}$Each word w_i is, in turn, described by a set of concepts S as:

$W_i = \{ S_{i1}, S_{i2,\ldots}S_{ix}\}$,

And, each concept S_i is, in turn, described by a set of primitives:

$S_i = \{ p_{i1}, p_{i2} \ldots p_{iy}\}$

For each word pair, w_1 and w_2, the similarity function is defined by

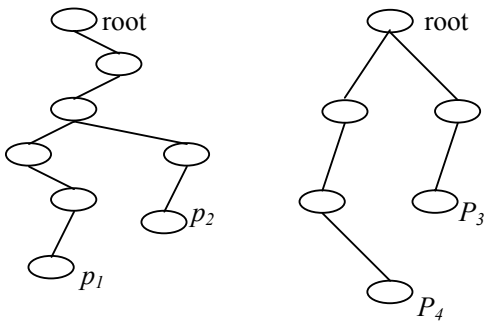
$$Sim(w_1, w_2) = \max_{i=1..n, j=1...m} Sim(S_{1i}, S_{2j}) \quad (1)$$

where S_{1i} is the list of concepts associated with W_1 and S_{2j} is the list of concepts associated with W_2.

As any concept S_i is presented by its primitives, the similarity of primitives for any p_1, and p_2 of the same type, can be expressed by the following formula:

$$Sim(p_1, p_2) = \frac{\alpha}{Dis(p_1, p_2) + \alpha} \quad (2)$$

where α is an adjustable parameter set to 1.6, and $Dis(p_1, p_2)$ is the path length between p_1 and p_2 based on the semantic tree structure. The above formula where α is a constant does not indicate explicitly the fact that the depth of a pair of nodes in the tree affects their similarity. For two pairs of nodes (p_1, p_2) and (p_3, p_4) with the same distance, the deeper the depth is, the more commonly shared ancestros they would have which should be semantically closer to each other. In following two tree structures, the pair of nodes (p_1, p_2) in the left tree should be more similar than (p_3 , p_4) in the right tree.



To indicate this observation, α is modified as a function of tree depths of the nodes using the formula $\alpha = min(d(p_1), d(p_2))$. Consequently, the formula (2) is rewritten as formular (2^a) during the experiment.

$$Sim(p_1, p_2) = \frac{\min(d(p_1), d(p_2))}{Dis(p_1, p_2) + \min(d(p_1), d(p_2))} \quad (2^a)$$

where $d(p_i)$ is the depth of node p_i in the tree . The comparison of calculating the word similarity by applying the formula (2) and (2^a) is shown in Section 4.4.

Based on the DEF description in HowNet, different primitive types play different roles only some are directly related to semantics. To make use of both the semantic and syntactic information included in HowNet to describe a word, the similarity of two concepts should take into consideration of all primitive types with weighted considerations and thus the formula is defined as

$$Sim(S_1, S_2) = \sum_{i=1}^{4} \beta_i \prod_{j=1}^{i} Sim_j(p_{1j}, p_{2j}) \quad (3)$$

where β_i is a weighting factor given in (Liu 2002) with the sum of $\beta_1 + \beta_2 + \beta_3 + \beta_4$ being 1, and $\beta_1 \geq \beta_2 \geq \beta_3 \geq \beta_4$. The distribution of the weighting factors is given for each concept a priori in HowNet to indicate the importance of primitive p_i in defining the corresponding concept S.

3.3 Collocation Extraction

In order to extract collocations from a corpus, and to obtain result for **Step 1** of our algorithm, we used the collocation extraction algorithm developed by the research group at the Hong Kong Polytechnic University(Lu et al. 2003). The extraction of bi-gram collocation is based on the English Xtract(Smaja 1993) with improvements. Based on the three Steps mentioned earlier, we will present the extractions in each step in the subsections.

3.3.1 Bi-gram Extraction

Based on the lexical statistical model proposed by Smadja in Xtract on extracting English collocations, an improved algorithm was developed for Chinese collocation by our research group and the system is called CXtract. For easy of understanding, we will explain the algorithm briefly here. According to Xtract, word cooccurence is denoted by a tripplet (w_h, w_i, d) where w_h is a given headword, w_i is a co-word appeared in the corpus in a distance d within the window of [-5, 5]. The frequency f_i of the co-word w_i in the window of [-5, 5] is defined as:

$$f_i = \sum_{j=-5}^{5} f_{i,j} \quad (4)$$

where $f_{i,j}$ is the frequency of the co-word at distance j in the corpus within the window. The average frequency of f_i, denoted by $\overline{f}_i$, is given by

$$\overline{f}_i = \sum_{j=-5}^{5} f_{i,j} / 10 \quad (5)$$

Then, the average frequency $\overline{f}$, and the standard deviation σ are defined by

$$\overline{f} = \frac{1}{n}\sum_{i=1}^{n} f_i \; ; \quad \sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_i - \overline{f})^2} \quad (6)$$

The *Strength* of the co-occurrence for the pair (w_h, w_i), denoted by k_i, is defined by

$$k_i = \frac{f_i - \overline{f}}{\sigma}, \quad (7)$$

Furthermore, the *Spread* of (w_h, w_i), denoted as U_i, which characterizes the distribution of w_i around w_h is define as:

$$U_i = \frac{\sum (f_{i,j} - \overline{f}_i)^2}{10}; \quad (8)$$

To eliminate the bi-grams with unlikely co-occurrence, the following sets of threshold values is defined:

$$C1: k_i = \frac{f_i - \overline{f}}{\sigma} \geq K_0 \quad (9)$$

$$C2: U_i \geq U_0 \quad (10)$$

$$C3: f_{i,j} \geq \overline{f}_i + (K_1 \cdot \sqrt{U_i}) \quad (11)$$

However, the above statistical model given by Smadja fails to extract the bi-grams with a much higher frequency of w_h but a relatively low frequency word of w_i, For example, in the bi-gram棘手问题, freq (棘手) is much lower than the freq (问题). Therefore, we further defined a weighted mutual information to extract this kind of bi-grams:

$$R_i = \frac{f(w_h, w_i)}{f(w_i)} \geq R_0, \quad (12)$$

As a result, the system should return a list of triplets (w_h, w_i, d), where (w_h, w_i) is considered collocations.

3.3.2 Synonyms Set

For each given headword w_h, before taking it as an input to extract its bi-grams directly, we fist apply the similarity formula described in Equation (1) to generate a set of synonyms headwords W_{syn}:

$$W_{syn} = \{w_s : Sim(w_h, w_s) > \theta\} \quad (13)$$

Where $0 < \theta < 1$ is an algorithm parameter which is adjusted based on experience. We set it as 0.85 from the experiment because we would like to balance the strength of the synonyms relationship and the coverage of the synonyms set. The setting of the parameter $\theta < 0.85$ weaks the similarity strength of the extracted synonyms. For example, for a given collocation "改善关系", that is unlikely to include the candidates "改善护照", "改善契据", "改善通连". On the other hand, by setting the parameter $\theta > 0.85$ will limit the coverage of the synonyms set and hence lose valuable synonyms. For example, for a given bi-gram "重大贡献", we hope to include the candidate synonymous collocations such as "重大成果", "重大成绩", "重大成就". We will show the test of θ in the section 4.2.

This synonyms headwords set provides the possibility to extract the synonymous collocation with the lower frequency that failed to be extracted by lexical statistic.

3.3.3 Synonymous Collocations

A phenomenal among the collocations in natural language is that there are many synonymous collocations exist. For example, 'switch on light' and 'turn on light', "财务问题" and "财政问题". Due to the domain specification of the corpus, some of the synonymous collocations may fail to be extracted by the lexical statistic model because of their lower frequency. Based on this observation, this paper takes a further step. The basic idea is for a bi-gram collocation (w_h, w_c, d) we select the synonyms w_s of w_h with the maximum similarity respect to all the concepts contained by w_h, we deem (w_s, w_c, d) as a collocation if its occurrence is greater than 1 in the corpus. There are similar works discussed by Pearce (Pearce 2001). .

For a given collocation (w_s, w_c, d), if $w_s \in W_{syn}$, then we deem the triple (w_s, w_c, d) as a synonymous collocation with respect to the collocation (w_h, w_c, d) if the co-occurrence of (w_s, w_c, d) in the corpus is greater than one. Therefore, we define the collection of synonymous collocations C_{syn} as:

$$C_{syn} = \{(w_s, w_c, d) : Freq(w_s, w_c, d) > 1\} \quad (14)$$

where $w_s \in W_{syn}$.

4 Evaluation

The performance of collocation is normally evaluated by precision and recall as defined below.

$$precision = \frac{number\ of\ correct\ Extracted\ Collocations}{total\ number\ of\ extracted\ Collocations} \quad (15)$$

$$recall = \frac{number\ of\ correct\ Extracted\ Collocations}{total\ number\ of\ actual\ Collocations} \quad (16)$$

To evaluate the performance of our approach, we conducted a set of experiments based on 9 selected headwords. A baseline system using only lexical statistics given in 3.3.1 is used to get a set of baseline data called Set A. The output using our algorithm is called Set B. Results are checked by hand for validation on what is true collocation and what is not a true collocation.

F_5	F_4	F_3	F_2	F_1	Headwo	F1	F2
*	*	*	*	*	重大	突破	*
*	*	*	*	*	重大	问题	*
*	*	*	*	*	重大	贡献	*
*	*	*	*	*	重要	成果	*
*	*	*	*	*	重要	问题	*
*	*	*	*	*	重要	指示	*
*	*	*	*	*	重要	论述	*
*	*	*	*	*	重要	批示	*
*	*	*	作出	了	重要	贡献	*
*	*	*	*	占有	重要	地位	*
*	*	*	*	*	根本	改观	*
*	*	*	*	*	根本	好转	*
*	两	国	人民	的	根本	利益	*
*	*	*	*	*	基本	原则	*
*	*	*	*	*	基本	国策	*
*	*	*	*	*	基本	生活费	*
*	*	*	党	的	基本	路线	*
*	*	*	党	的	基本	理论	*

Table 1. Sample table for the true collocation with headword "重要"

F_5	F_4	F_3	F_2	F_1	Head	F1	F2	F3	F4	F5
*	*	*	*	个	重要	*	*	*	*	*
*	*	工作	*	*	重要	*	*	*	*	*
*	*	关系	*	*	重要	*	*	*	*	*
国	*	*	*	*	重要	*	*	*	*	*
*	*	工作	*	*	重大	*	*	*	*	*
*	*	关系	*	*	重大	*	*	*	*	*
*	*	*	*	国家	重大	*	*	*	*	*
*	*	*	*	进行	重大	*	*	*	*	*
我国	*	*	*	*	重大	*	*	*	*	*
我们	*	*	*	*	重大	*	*	*	*	*
*	*	*	*	有	重大	*	*	*	*	*
*	*	*	*	*	根本	*	是	*	*	*
*	两	国	人民	的	根本	利益	*	*	*	*
*	*	*	*	最	根本	的	*	*	*	*
*	*	交通	*	的	根本	出路	在于	科学化	*	*
*	关系	*	*	*	基础	*	*	*	*	*
*	社会	*	*	*	基础	*	*	*	*	*
*	*	*	*	原则	基础	上	*	*	*	*
*	*	*	*	*	基础	打	牢	*	*	*

Table 2. Sample table for the bi-grams that are not true collocations

Table 1 shows samples of extracted word bi-grams using our algorithm that are considered synonymous collocations for the headword "重要". **Table 2** shows extracted bi-grams by our algorithm that are not considered true collocations.

4.1 Test Set

Our experiment is based on a corpus of six months tagged People Daily with 11 millions number of words. For word bi-gram extractions, we consider only content words, thus headwords are selected from noun, verb and adjective only. For evaluation purpose, we selected randomly 3 nouns, 3 verbs and 3 adjectives with frequency of low, medium and high. Thus, in **Step 1** of the algorithm, 9 headwords were used to extract bi-gram collocations from the corpus, and 253 pairs of collocations were extracted. Evaluation by hand has identified 77 true collocations in Set A test set. The overall precision rate is 30% (see Table 3).

	Noun+Verb +Adjective
Headword	9
Extracted Bi-grams	253
True collocations using lexical statistics only	77
Precision rate	30%

Table 3: Statistics in test set for set A

Using **Step 2** of our algorithm, where $\theta=0.85$ is used, we have obtained 55 synonym headwords (include the 9 headwords). Out of these 55 synonyms, 614 bi-gram pairs were then extracted from the lexical statistics based algorithm, in which 179 are consider true collocations. Then, by applying **Step 3** of our algorithm, we extracted an additional 201 bi-gram pairs, among them, 178 are considered true collocations. Therefore, using our algorithm, the overall precision rate has achieved 43%, an improvement of almost 50%. The data is summarized in **Table 4.**

	n., v, and adj.
Synonyms headword	55
Bi-grams (lexical statistics)	614
Non-synonym collocations (lexical statistics only)	179
Extracted synonym collocations Step 2	201
True synonym collocations using Step 2	178
Overall precision rate	43%

Table 4: Statistics in test set for mode B

4.2 The choice of θ

We also conducted a set of experiments to choose the best value for the similarity function's threshold θ. We tested the best value of θ with both the precision rate and the estimated recall rate using the so called remainder bi-grams. The remainder bi-grams is the total number of bi-grams extracted by the algorithm. When precision goes up, the size of the result is smaller, which in a way is an indicator of less recalled collocations. **Figure 1** shows the precision rate and the estimated recall rate in testing the value of θ.

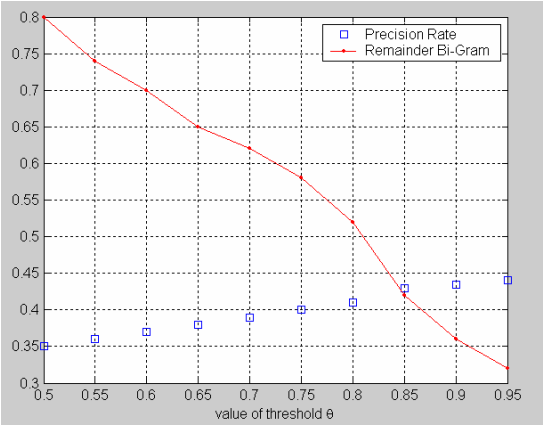


Figure 1. Precision Rate vs. value of θ

From Figure 1, it is obvious that at $\theta=0.85$ the recall rate starts to drop more drastically without much incentive for precision.

	Extracted Bi-grams using lexical statistics	Extracted Synonyms Collocations using Step 2
(1.2,1.4,12)	465	328
(1.4,1.4,12)	457	304
(1.4,1.6,12)	394	288
(1.2,1.2,12)	513	382
(1.2,1.2,14)	503	407
(1.2,1.2,16)	481	413

Table 5: Value of (K_0, K_1, U_0).

4.3 The test of (K_0, K_1, U_0)

The original threshold for CXtract is *(1.2, 1.2, 12)* for the parameters *(K_0, K_1, U_0)*. However, with synonyms collocations, we have also conducted some experiments to see whether the parameters should be adjusted. **Table 5** shows the statistics to test the value of *(K_0, K_1, U_0)*. The similarity threshold θ was fixed at 0.85 throughout the experiments.

The experimental shows that varying the value of (k_0, k_1) does not bring any benefit to our algorithm. However, increasing the value of u_0 did improve the extraction of synonymous collocations. **Figure 2** shows that U_0 =14 is a good trade-off for the precision rate and the remainder Bi-grams. The basic meaning behind the result is reasonable. According to Smadja, U_0 defined in the formula (8) represents the co-occurrence distribution of the candidate collocation (w_h, w_c) in the position of d (-5 ⩽ d ⩽ 5). For a true collocation (w_h, w_c, d), its co-occurrence in the position d is much higher than in other positions which leads to a peak in the co-occurrence distribution. Therefore, it is selected by the statistical algorithm based on the formula (10). Based on the physical meaning behind, one way to improve the precision rate is to increase the value of the threshold U_0. A side effect to an increased value of U_0 is that the recall is decreased because some true collocations do not meet the condition of co-occurrence greater than U_0. Step 2 of the new algorithm regains some true collocations lost because of a higher U_0 in Step 1.

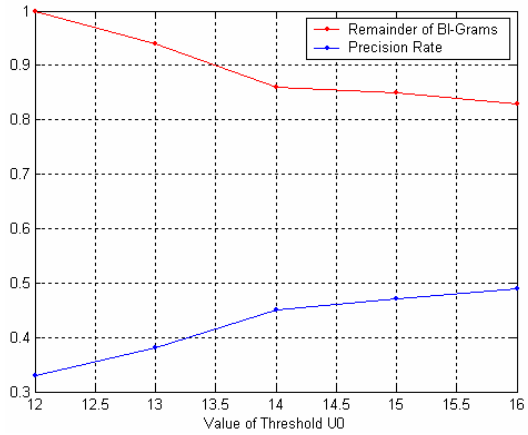


Figure 2. Precision Rate vs. Value of U_0

4.4 The comparison of similarity calculation based on formula (2) and (2^{α})

Table 6 shows the similarity value given by formula (2) where α is a constant given the value 1.6 and by formula (2^{α}) where α is replaced by a function of the depths of the nodes. Results show that (2^{α}) is more fine tuned and reflects the nature of the data better. For example, 工人 and 农民 are more similar than 工人 and 运动员. 粉红 and 红 are much similar but not the same.

Word 1	Word 2	Formula (2)	Formula (2^{α})
男人	女人	0.86	0.95
男人	父亲	1.00	1.00
男人	和尚	0.86	0.95
男人	高兴	0.05	0.10
工人	农民	0.72	0.88
工人	运动员	0.72	0.88
中国	美国	0.94	0.92
粉红	红	1.00	0.92
粉红	红色	1.00	0.92
十分	非常	1.00	1.00
十分	特别	0.62	0.95
考虑	思想	0.70	1.00
思考	考虑	1.00	1.00

Table 6: comparison of similarity calculation

5 Conclusion and Further Work

In this paper, we have presented a method to extract bi-gram collocations using lexical statistics model with synonyms information. Our method reaches the precision rate of 43% for the tested data. Comparing to the precision of 30% using lexical statistics only, our improvement is close to 50%. In additional, the recall improved 30%. The contribution is that we have made use of synonym information which is plentiful in the natural language use and it works well to supplement the shortcomings of lexical statistical method.

Manning claimed that the lack of valid substitution for a synonym is a characteristics of collocations in general (Manning and Schutze 1999). To extend our work, we consider the use of synonym information can be further applied to help identify collocations of different types.

Our preliminary study has suggested that collocation can be classified into 4 types:

Type 0 Collocation: Fully fixed collocation which include some idioms, proverbs and sayings such as "缘木求鱼" "釜底抽薪" and so on.

Type 1 Collocation: Fixed collocation in which the appearance of one word implies the co-occurrence of another one such as "裁减 员额".

Type 2 Collocation: Strong collocation which allows very limited substitution of the components, for example, "裁减 职位", "减少职位", "缩减职位" and so on.

Type 3 Collocation: Normal collocation which allows more substitution of the components, however a limitation is still required. For example, "减少 开支" "缩减 开支" "压缩开支" "消减开支".

By using synonym information and define substitutability, we can validate whether collocations are fixed collocations, strong collocations with very limited substitutions, or general collocations that can be substituted more freely.

6 Acknowledgements

Our great thanks to Dr. Liu Qun of the Chinese Language Research Center of Peking University for letting us share their data structure in the Synonyms Similarity Calculation. This work is partially supported by the Hong Kong Polytechnic University (Project Code A-P203) and CERG Grant (Project code 5087/01E)

References

M. Benson, 1990. *Collocations and General Purpose Dictionaries.* International Journal of Lexicography, 3(1): 23-35

Y. Choueka, 1993. *Looking for Needles in a Haystack or Locating Interesting Collocation Expressions in Large Textual Database.* Proceedings of RIAO Conference on User-oriented Content-based Text and Image Handling: 21-24, Cambridge.

K. Church, and P. Hanks, 1990. *Word Association Norms, Mutual Information,and Lexicography.* Computational Linguistics, 6(1): 22-29.

I. Dagan, L. Lee, and F. Pereira. 1997. *Similarity-based method for word sense disambiguation.* Proceedings of the 35th Annual Meeting of ACL: 56-63, Madrid, Spain.

Z. D. Dong and Q. Dong. 1999. *Hownet*, http://www.keenage.com

D. K. Lin, 1997. *Using Syntactic Dependency as Local Context to Resolve Word Sense Ambiguity.* Proceedings of ACL/EACL-97: 64-71, Madrid, Spain

Q. Liu, 2002. *The Word Similarity Calculation on <<HowNet>>*. Proceedings of 3rd Conference on Chinese lexicography, TaiBei

Q. Lu, Y. Li, and R. F. Xu, 2003. *Improving Xtract for Chinese Collocation Extraction.* Proceedings of IEEE International Conference on Natural Language Processing and Knowledge Engineering, Beijing

C. D. Manning and H. Schutze, 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts

D. Pearce, 2001. *Synonymy in Collocation Extraction.* Proceedings of NAACL'01 Workshop on Wordnet and Other Lexical Resources: Applications, Extensions and Customizations

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

H. Wu, and M. Zhou, 2003. *Synonymous Collocation Extraction Using Translation Information.* Proceeding of the 41st Annual Meeting of ACL

D. K. Lin, 1998. *Extracting collocations from text corpora.* In Proc. First Workshop on Computational Terminology, Montreal, Canada.

M. S. Sun, C. N. Huang and J. Fang, 1997. *Preliminary Study on Quantitative Study on Chinese Collocations.* ZhongGuoYuWen, No.1, 29-38, (in Chinese).