Computational Linguistics and Chinese Language Processing
Vol. 10, No. 1, March 2005, pp.123-144

Similarity Based Chinese Synonym Collocation Extraction

Wanyin Li*, Qin Lu* and Ruifeng Xu*

Abstract

Collocation extraction systems based on pure statistical methods suffer from two major problems. The first problem is their relatively low precision and recall rates. The second problem is their difficulty in dealing with sparse collocations. In order to improve performance, both statistical and lexicographic approaches should be considered. This paper presents a new method to extract synonymous collocations using semantic information. The semantic information is obtained by calculating similarities from HowNet. We have successfully extracted synonymous collocations which normally cannot be extracted using lexical statistics. Our evaluation conducted on a 60MB tagged corpus shows that we can extract synonymous collocations that occur with very low frequency and that the improvement in the recall rate is close to 100%. In addition, compared with a collocation extraction system based on the Xtract system for English, our algorithm can improve the precision rate by about 44%.

Keywords: Lexical Statistics, Synonymous Collocations, Similarity, Semantic Information

1. Introduction

A collocation refers to the conventional use of two or more adjacent or distant words which hold syntactic and semantic relations. For example, the conventional expressions “warm greetings”, “broad daylight”, “思想包袱”, and “托运行李” all are collocations. Collocations bear certain properties that have been used to develop feasible methods to extract them automatically from running text. Since collocations are commonly found, they must be recurrent. Therefore, their appearance in running text should be statistically significant, making it feasible to extract them using the statistical approach.

* Department of Computing, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong
Tel: +852-27667326; +852-27667247 Fax:+852-27740842
E-mail: {cswyli, csluqin, csrfxu}@comp.polyu.edu.hk

A collocation extraction system normally starts with a so-called headword (sometimes also called a keyword) and proceeds to find co-occurring words called the collocated words. For example, given the headword "基本", such bi-gram collocations as "基本理论", "基本工作", and, "基本原因" can be found using an extraction system where "理论", "工作", and "原因" are called collocated words with respect to the headword "基本." Many collocation extraction algorithms and systems are based on lexical statistics [Church and Hanks 1990; Smadja 1993; Choueka 1993; Lin 1998]. As the lexical statistical approach was 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 both the recall rate and precision rate. The precision rate achieved using the lexical statistics approach can reach around 60% if both word bi-gram extraction and n-gram extraction are employed [Smadja 1993; Lin 1997; Lu *et al*. 2003]. The low precision rate is mainly due to the low precision rate of word bi-gram extraction as only about a 30% - 40% precision rate can be achieved for word bi-grams. The semantic information is largely ignored by statistics- based collocation extraction systems even though there exist multiple resources for lexical semantic knowledge, such as WordNet [Miller 98] and HowNet [Dong and Dong 99].

In many collocations, the headword and its collocated words hold specific semantic relations, hence allowing collocate substitutability. The substitutability property provides the possibility of extracting collocations by finding synonyms of headwords and collocate words. Based on the above properties of collocations, this paper presents a new method that uses synonymous relationships to extract synonym word bi-gram collocations. The objective is to make use of synonym relations to extract synonym collocations, thus increasing the recall rate.

Lin [Lin 1997] proposed a distributional hypothesis which says that if two words have similar sets of collocations, then they are probably similar. According to one definition [Miller 1992], two expressions are synonymous in a context C if the substitution of one for the other in C does not change the truth-value of a sentence in which the substitution is made. Similarly, in HowNet, Liu Qun [Liu *et al.* 2002] defined word similarity as two words that can substitute for each other in a context and keep the sentence consistent in syntax and semantic structure. This means, naturally, that two similar words are very close to each other and they can be used in place of each other in certain contexts. For example, we may either say "买书"or "订书" since "买" and "订" are semantically close to each other when used in the context of buying books. We can apply this lexical phenomena after a lexical statistics-based extractor is applied to find low frequency synonymous collocations, thus increasing the recall rate.

The rest of this paper is organized as follows. Section 2 describes related existing collocation extraction techniques that are based on both lexical statistics and synonymous collocation. Section 3 describes our approach to collocation extraction. Section 4 describes the

data set and evaluation method. Section 5 evaluates the proposed method. Section 6 presents our conclusions and possible future work.

2. Related Works

Methods have been proposed to extract collocations based on lexical statistics. Choueka [Choueka 1993] applied quantitative selection criteria based on a 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. However, the method can not be extended to extract n-grams. Smadja [Smadja 1993] proposed a statistical model that measures the spread of the distribution of co-occurring pairs of words with higher strength. This method can successfully extract both adjacent and distant bi-grams, and n-grams. However, it can not extract bi-grams with lower frequency. The precision rate of bi-grams collocation is very low, only around 30%. Generally speaking, it is difficult to measure the recall rate in collocation extraction (there are almost no reports on recall estimation) even though it is understood that low occurrence collocations cannot be extracted. Sun [Sun 1997] performed a preliminary *Quantitative* analysis of the strength, spread and peak of Chinese collocation extraction using different statistical functions. That study suggested that the statistical model is very limited and that syntax structures can perhaps be used to help identify pseudo collocations.

Our research group has further applied the Xtract system to Chinese [Lu *et al.* 2003] by adjusting the parameters so at to optimize the algorithm for Chinese and developed a new weighted algorithm based on mutual information to acquire word bi-grams which are constructed with one higher frequency word and one lower frequency word. This method has achieved an estimated 5% improvement in the recall rate and a 15% improvement in the precision rate compared with the Xtract system.

A method proposed by Lin [Lin 1998] applies a dependency parser for information extraction to collocation extraction, where a collocation is defined as a dependency triple which specifies the type of relationship between a word and the modifiee. This method collects dependency statistics over a parsed collocation corpus to cover the syntactic patterns of bi-gram collocations. Since it is statistically based, therefore it still is unable to extract bi-gram collocations with lower frequency.

Based on the availability of collocation dictionaries and semantic relations of words combinatorial possibilities, such as those in WordNet and HowNet, some researches have made a wide range of lexical resources, especially synonym information. Pearce [Pearce 2001] presented a collocation extraction technique that relies on a mapping from one word to its synonyms for each of its senses. The underlying intuition is that if the difference between the occurrence counts of a synonym pair with respect to a particular word is at least two, then they

can be considered a collocation. To apply this approach, knowledge of word (concept) semantics and relations with other words must be available, such as that provided by WordNet. Dagan [Dagan 1997] applied a similarity-based smoothing method to solve the problem of data sparseness in statistical natural language processing. Experiments conducted in his later research showed that this method could achieve much better results than back-off smoothing methods in terms of word sense disambiguation. Similarly, Hua [Wu 2003] applied synonyms relationships between two different languages to automatically acquire English synonymous collocations. This was the first time that the concept of synonymous collocations was proposed. A side intuition raised here is that a natural language is full of synonymous collocations. As many of them have low occurrence rates, they can not be retrieved by using lexical statistical methods.

HowNet, developed by Dong et al. [Dong and Dong 1999] is the best publicly available resource for Chinese semantics. Since semantic similarities of words are employed, synonyms can be defined by the closeness of their related concepts and this closeness can be calculated. In Section 3, we will present our method for extracting synonyms from HowNet and using synonym relations to further extract collocations. While a Chinese synonym dictionary, Tong Yi Ci Lin (《同义辞林》), is available in electronic form, it lacks structured knowledge, and the synonyms listed in it are too loosely defined and are not applicable to collocation extraction.

3. Our Approach

Our method to extract Chinese collocations consists of three steps.

Step 1: We first take the output of any lexical statistical algorithm that extracts word bi-gram collocations. This data is then sorted according to each headword, w_h, along with its collocated word, w_c.

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 a similarity value exceeding a threshold is considered a synonym headword, w_s, for additional extractions.

Step 3: For each synonym headword, w_s, and the collocated word, w_c, of w_h, if the bi-gram (w_s, w_c) is not in the output of the lexical statistical algorithm applied in Step 1, then we take this bi-gram (w_s, w_c) as a collocation if the pair appears in the corpus by applying an additional search on the corpus.

3.1 Bi-gram Collocation Extraction

In order to extract Chinese collocations from a corpus and to obtain result in Step 1 of our algorithm, we use an automatic collocation extraction system named CXtract, developed by a research group at Hong Kong Polytechnic University [Lu *et al.* 2003]. This collocation

extraction system is based on English Xtract [Smaja 1993] with two improvements. First, the parameters (K_0, K_1, U_0) used in Xtract are adjusted so as to optimize them for a Chinese collocation extraction system, resulting in an 8% improvement in the precision rate. Secondly, a solution is provided to the so-called high-low problem in Xtract, where bi-grams with a high frequency the head word, w_h, but a relatively low frequency collocated word, w_i can not be extracted. We will explain the algorithm briefly here. According to Xtract, a word concurrence is denoted by a triplet (w_h, w_i, d), where w_h is a given headword and w_i is a collocated word appeared in the corpus with a distance d within the window [-5, 5]. The frequency, f_i, of the collocated word, w_i, in the window [-5, 5] is defined as

$$f_i = \sum_{j=-5}^{5} f_{i,j} \tag{1}$$

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

$$\overline{f_i} = \sum_{j=-5}^{5} f_{i,j} / 10 . \tag{2}$$

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

$$\overline{f} = \frac{1}{n}\sum_{i=1}^{n} f_i ; \quad \sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_i - \overline{f})^2} . \tag{3}$$

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

$$k_i = \frac{f_i - \overline{f}}{\sigma} . \tag{4}$$

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

$$U_i = \frac{\sum (f_{i,j} - \overline{f_i})^2}{10} . \tag{5}$$

To eliminate bi-grams which are unlikely to co-occur, the following set of threshold values is defined:

$$C1 : k_i = \frac{f_i - \overline{f}}{\sigma} \geq K_0 \tag{6}$$

$$C2 : U_i \geq U_0 \tag{7}$$

$$C3 : f_{i,j} \geq \overline{f_i} + (K_1 \cdot \sqrt{U_i}) \tag{8}$$

where the threshold value set (K_0, K_1, U_0) is obtained through experiments. A bi-gram (w_h, w_i, d) will be filtered out as a collocation if it does not satisfy one of the above conditional thresholds. Condition **C1** is used to measure the "recurrence" property of collocations when the bi-grams (w_h, w_i, d) with co-occurrences frequencies higher than K_0 times the standard deviation over the average are selected. **C2** is used to select bi-gram pairs (w_h, w_i, d) having a spread values that are larger than a given threshold, U_0. A lower U value implies that the bi-gram is evenly distributed in all 10 positions and thus is not considered a "rigid combination". **C3** is used to select bi-grams in these "certain positions". Only if certain peak positions exist, the co-occurrence bi-grams are considered collocations. The values of (K_0, K_1, U_0) are set to (1, 1, 10), which are the optimal parameters for English according to Xtract. For the CXtract, the values of (K_0, K_1, U_0) are adjusted to (1.2, 1.2, 12) which are suitable for the Chinese collocation extraction.

However, Xtract cannot extract high-low collocations when w_h has a quite high frequency and its co-word w_i has a relatively low frequency. For example, "棘手问题" is a bi-gram collocation. But because *freq (*棘手*)* is much lower than the *freq (*问题*),* this bi-gram collocation cannot be identified, resulting in a lower recall rate. In CXtract, an additional step is used to identify such high-low collocations by measuring the conditional probability as follows:

$$R_i = \frac{f(\mathrm{w_h}, w_i)}{f(w_i)} \geq R_0, \tag{9}$$

which measures the likelihood of occurrence of w_h given w_i, thus discounting the absolute frequency of w_i. CXtract outputs a list of triplets (w_h, w_i, d), where (w_h, w_i,) is considered to be a collocation.

3.2 Construct Synonyms Set

In **Step 2** of our system, for each given headword w_h, we first need to find its *synonym set* W_{syn}, which contains all the words that are said to be the synonyms of w_h. As stated earlier, we estimate the synonym relation between words based on semantic similarity calculation in HowNet. Therefore, before explaining how the synonym set can be constructed, we will introduce the semantic structure of HowNet and the similarity model built based on HowNet.

3.2.1 Semantic Structure of HowNet

Because we hope to explore the different semantics meanings that each word carries, word sense disambiguation is the main issue when we calculate the similarity of words. For example, the word "打" used with the words "酱油" as in "打酱油" and "网球" as in "打网球" has the meanings of buy("卖") and exercise("锻炼"), respectively. As a bilingual semantic and syntactic knowledge base, HowNet provides separate entries when the same word contains

more than one concept. Unlike WordNet, in which a semantic relation is a relation between synsets, HowNet adopts a constructive approach to semantic representation. It describes words as a set of concepts (义项) and describes each concept using a set of primitives（义元）,which is the smallest semantic unit in HowNet and cannot be decomposed further. The template of word concepts is organized in HowNet as shown below:

NO.= the record number of the lexical entries
W_C/E = concept of the language (Chinese or English)
E_C/E = example of W_C/E
G_C/E = Part-of-speech of the W_C/E
DEF = Definition, which is constructed by primitives and pointers

For example, in the following, for the word "打", we list the two of its corresponding concepts:

NO.=000001
W_C=打
G_C=V
E_C=~酱油，~张票，~饭，去~瓶酒，醋~来了
W_E=buy
G_E=V
E_E=
DEF=buy|买

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|锻练, sport|体育

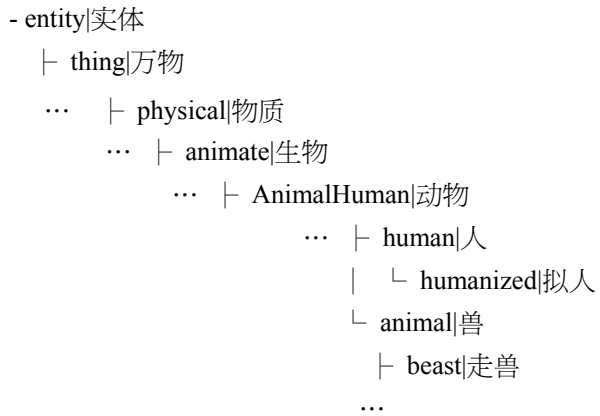
Note: Replace all the graphics above by simple text. In the above records, DEFs are where the primitives are specified. DEF contains up to four types of primitives: *basic independent primitives* (基本独立义元), *other independent primitives* (其他独立义元), *relation primitives* (关系义元), and *symbol primitives* (符号义元), where basic independent primitives and other independent primitives are used to indicate the basic concept, and the

other types are used to indicate syntactical relationships. For example, the word “生日” has all four types of primitives as shown below:

NO.=072280
W_C=生日
G_C=n
E_C=祝贺~，过~，~聚会
W_E=birthday
G_E=n
E_E=
DEF=time|时间, day|日, @ComeToWorld|问世, $congratulate|祝贺

The basic independent primitive “time|时间” defines the general classification of “birthday|生日”. The other independent primitive “day|日” indicates that “birthday|生日” is related to “day|日”. The symbol primitives “@ComeToWorld|问世” and “$congratdulate|祝贺” provide more specific, distinguishing features to indicate syntactical relationships. The pointer “@” specifies “time or space”, indicating that “birthday|生日” is the time of “ComeToWorld|问世”. Another pointer “$” specifies “object of V”, which means that “birthday|生日” is the object of “congratulate|祝贺”. In summary, we find that “birthday|生日” belongs to “time|时间” in general and is related to “day|日” which specifies the time of “ComeToWorld|问世”.

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



Note: Replace all the graphics above by simple text.

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

3.2.2 Similarity Model Based on HowNet

Liu Qun [Liu 2002] defined word similarity as two words that can substitute for each other in the same context and still keep the sentence syntactically and semantically consistent. This is very close to our definition of synonyms. Thus, in this work, we will directly use the similarity function provided by Liu Qun, which is stated below.

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

$$W = \{w_1, w_2, \ldots w_n\}.$$

Each word w_i is, in turn, described by a set of concepts S

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

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

$$Sim(w_1, w_2) = \max_{i=1\cdots n, j=1\cdots m} Sim(S_{1i}, S_{2j}) \tag{10}$$

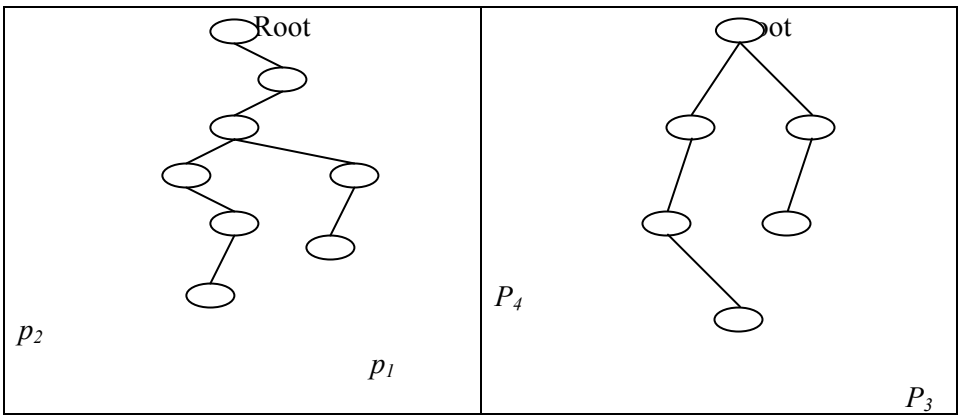
where S_{1i} is the list of concepts associated with w_1 and S_{2j} is the list of concepts associated with w_2.

As any concept, S_i is represented 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:

$$Sim(p_1, p_2) = \frac{\alpha}{Dis(p_1, p_2) + \alpha} \tag{11}$$

where α is an adjustable parameter with a value of 1.6 according to Liu [Liu 2002]. $Dis(p_1, p_2)$ is the path length between p_1 and p_2 based on the semantic tree structure. The above formula does not explicitly indicate 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, the more commonly shared ancestors they have, which means that they are semantically

closer to each other. In the 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 clarify this observation, α is modified as a function of the tree depths of the nodes using the formula $\alpha=min(d(p_1),d(p_2))$. Consequently, the formula (11) was rewritten as formula (11ª) below for our experiments.

$$Sim(p_1, p_2) = \frac{\min(d(p_1), d(p_2))}{Dis(p_1, p_2) + \min(d(p_1), d(p_2))} \tag{11ª}$$

where $d(p_i)$ is the depth of node p_i in the tree. Calculating the word similarity by applying formulas (11) and (11ª) will be discussed in Section 4.4.

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

$$Sim(S_1, S_2) = \sum_{i=1}^{4} \beta_i \prod_{j=1}^{i} Sim_j(p_{1j}, p_{2j}) \tag{12}$$

where β_i is a weighting factor given in [Liu 2002], where the sum of $\beta_1 + \beta_2 + \beta_3 + \beta_4$ is 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 for the corresponding concept S. The similarity model given here is the basis for building a synonyms set where β_1 *and* β_2 represent the semantic information, and β_3 *and* β_4 represent the syntactic relation.

3.2.3 The Set of Synonyms Headwords

For each given headword w_h, we apply the similarity formula in Equation (10) to generate its synonym set, $W_{syn,}$ which is defined as

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

where $0 < \theta < 1$ is an algorithm parameter which is adjusted based on experience. We set it to 0.85 based on our experiment because we wanted to balance the strength of the synonym relationship and the coverage of the synonym set. Setting the parameter $\theta < 0.85$ will weaken the similarity strength of the extracted synonyms. For example, a given collocation "改善关系" is unlikely to include the candidates "改善护照" and ，"改善契据". On the other hand, setting the parameter $\theta > 0.85$ will limit the coverage of the synonym set, thus valuable synonyms will be lost. For example, for a given bi-gram "重大贡献", we hope to include candidate synonymous collocations such as "重大成果", "重大成绩", and "重大成就". We will discuss the test on θ in section 5.2.

3.3 Synonyms Collocation

H. Wu [Wu 2003] defined a synonymous collocation pair as two collocations that are similar in meaning, but not identical in wording. Actually, in natural language, there exist many synonym collocations. For example, "switch on light" and "turn on light", "财务问题" and "财政问题". However, the sparse appearance of word combinations in a training corpus due to the limitation on the corpus size itself, some synonym collocations may not be extracted by the statistical method because of their lower co-occurrence frequencies. Based on this observation, we perform a further step. Our basic idea is to use a bi-gram collocation (w_h, w_c, d) to further obtain the synonym set W_{syn} of w_h, quantified by the similarity function. Then, for each w_s in W_{syn}, we consider (w_s, w_c, d) as a collocation if it indeed appears in the corpus at least a given number of times.

Our definition of a synonym collocation as follows. 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) to be a synonyms collocation with respect to the collocation (w_h, w_c, d) if (w_s, w_c, d) appears in the corpus N times, where N is a threshold value which we set to 2 in our experiment. Therefore, we define the collection of synonym collocations C_{syn} as

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

where $w_s \in W_{syn}$.

Our experimental results show that the precision rate of synonym collocation extraction is around 80% when we use the knowledge of HowNet. Some pseudo collocations can be automatically excluded because of the fact that they do not appear in the corpus. For example, for the headword "增长" in the collocation "增长见识", the synonym set extracted from our system contains {"增加", "增高", "增多"}, so the pseudo-collocations "增高见识", "增加见识", and "增多见识" will be excluded because they are not being used customarily used and,

thus, do not appear in the corpus. We checked them using Google and found that they did not appear either. On the other hand, for the collocated word "见识", our system extracts the synonyms set {"眼光", "眼界"}, and the word combination "增长眼界" appears twice in our corpus, thus according to our definition, it is a collocation. Therefore, the collocations "增长见识" and "增长眼界" are synonym collocations, and we can successfully extract "增长眼界" even though its frequency is very low (below 10 in our system).

4. Data Set and Evaluation Method

We modified Liu Qun's similarity model based on HowNet to obtain the synonyms of specified words. HowNet is a Chinese-English Bilingual Knowledge Dictionary. It includes both word entries and concept entries. There are more than 60 thousand Chinese concept entries and around 70 thousand English concept entries in HowNet. Both Chinese and English word entries are more than 50 thousand.

The corpus we used contains over 60MB of tagged sentences. Our experiment was conducted using tagged corpus of 11 million words collected six months from the People's Daily. For word bi-gram extraction, we considered only content words, thus, headwords were nouns, verbs or adjectives only.

In order to illustrate the effect of our algorithm, we used the statistically based system discussed in Section 3.1 as our baseline systems where the output data is called Set A. Using the output of the baseline system, we could further apply our algorithm to produce a data set called Set B.

The collocation performance is normally evaluated based on precision and recall as defined below:

$$precision = \frac{number\ of\ correct\ Extracted\ Collocations}{total\ number\ of\ extracted\ Collocations}, \tag{15}$$

$$recall = \frac{number\ of\ correct\ Extracted\ Collocations}{total\ number\ of\ actual\ Collocations}. \tag{16}$$

However, in collocation extraction, the absolute recall rate is rarely used because there are no benchmark "standard answers". Alternatively, we can use recall improvement to evaluate our system as defined below.

$$recall = \frac{(N_{none\_syn} + N_{syn})/X - N_{none\_syn}/X}{N_{syn}/X}, \tag{17}$$

where $N_{none\text{-}syn}$ stands for the number of non-synonyms collocations extracted by a statistical model, N_{syn} stands for the number of synonym collocations extracted based on synonym

relationships, and *X* stands for the total number of collocations in the corpus with respect to the given headwords.

Because there are no readily available "standard answers" for collocations, our results were checked manually to verify whether each candidate bi-gram was a true collocation or not. Since the output from the baseline system obtained using 60MB of tagged data consisted of over 200,000 collocations, we had to use the random sampling method to conduct an evaluation. In order to perform a fair evaluation, we tried to avoid subjective selection of words. Therefore, we randomly selected 5 words for each of the three types of words, namely, 5 nouns, 5 verbs, and 5 adjectives. Because headwords we chose were completely random and we did not target any particular words, our results should be statistically sound. Following is a list of the 15 randomly selected words used for the purpose evaluation:

nouns: 基础, 思想, 研究, 条件, 评选;
verbs: 改善, 加大, 增长, 提起, 颁发;
adjectives: 明显, 全面, 重要, 优秀, 大好

Table 1 shows samples of word bi-grams extracted using our algorithm that are considered collocations of the headwords "重大", "改善" and "加大". Table 2 shows bi-grams extracted by our algorithm that are not considered true collocations.

Table 1. Sample table for true collocations of the headwords "重大","改善","加大"

F_5	F_4	F_3	F_2	F_1	Headword	F1	F2	F3	F4
*	*	*	*	*	重大	意义	*	*	*
*	*	*	*	*	重大	影响	*	*	*
*	*	*	*	*	重大	作用	*	*	*
*	*	*	*	*	改善	关系	*	*	*
*	*	*	*	*	改善	*	环境	*	*
*	*	*	*	*	改善	*	交通	*	*
*	*	*	*	*	改善	*	结构	*	*
*	*	*	*	进一步	改善	*	*	*	*
*	*	*	*	明显	改善	*	*	*	*
*	*	*	*	*	改善	*	条件	*	*
*	*	*	*	*	改善	*	状况	*	*
*	*	*	*	进一步	加大	*	*	*	*
*	*	*	*	*	加大	*	力度	*	*
*	*	*	*	*	提起	公诉	*	*	*
*	*	*	*	*	提起	诉讼	*	*	*
*	*	*	*	*	增加	*	负担	*	*

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

F_4	F_3	F_2	F_1	Headword	F1	F2	F3	F4	F5
*	*	*	*	重大	政治	*	*	*	*
*	中	*	*	重大	*	*	*	*	*
*	*	*	着	重大	*	*	*	*	*
*	*	*	作出	重大	*	*	*	*	*
*	*	*	*	改善	*	*	關係	*	*
*	*	要	*	改善	*	*	*	*	*
*	*	将	*	改善	*	*	*	*	*
*	*	*	*	改善	金融	*	*	*	*
*	*	*	*	改善	农村	*	*	*	*
*	*	*	将	加大	*	*	*	*	*
*	*	*	*	加大	科技	*	*	*	*
*	*	*	*	加大	农业	*	*	*	*
*	*	*	*	加大	*	企業	*	*	*
*	*	*	*	加大	投入	*	*	*	*
*	*	*	要	加大	*	*	*	*	*
*	*	*		加大	*	*	企業		*

5. Evaluation and Analysis

5.1 Improvement in precision and recall rates

In Step 1 of the algorithm, 15 headwords were used to extract bi-gram collocations from the corpus, and 703 pairs of collocations were extracted. Evaluation by hand identified 232 true collocations in the set A test set. The overall precision rate was 31.7% (see Table 3).

Table 3. Statistics of the test set for set A

	n. + v. + a.
Headwords	15
Extracted Bi-grams	703
True collocations obtains using lexical statistics only	232
Precision rate	31.7 %

In Step 2 of our algorithm, where $\theta = 0.85$ was used, we obtained 94 synonym headwords (including the original 15 headwords). Out of these 94 synonym headwords, 841 bi-gram pairs were then extracted from the baseline system, and 243 were considered true collocations. Then, in Step 3 of our algorithm, we extracted an additional 311 bi-gram pairs; among them, 261 were considered true collocations. Because the synonym collocation extraction algorithm has

achieved a high precision rate of around 84% (261/311 = 83.9%) according to our experimental result as shown in Table 4.

Table 4. Statistics of the test set for mode B

	n. + v. + a.
Synonym headwords	94
Bi-grams (lexical statistics)	841
Non-synonym collocations (lexical statistics only)	243
Synonym collocations extracted in Step 3	311
True synonym collocations obtained in Step 3	261
Overall precision rate	83.9%

Since the data for Set B consisted of the additional extracted collocations. When we employed both Set A and Set B together as an overall system, the precision increased to 44 % ((243+261)/(841+311) = 43.7%), an improvement of almost 33% (43.7%-32.9%)/32.9% = 32.8%) comparing with the precision rate of the baseline system as shown in Table 5. As stated earlier, we are not able to evaluate the recall rate. However, compared with the statistical method indicated by Set A, an additional 261 collocations were recalled. Thus, we can record the recall the improvement which is ((243+261) – 243) /243= 107.4% as shown in Table 5.

Table 5. Comparison of sets A and B

Precision Rate of the Statistic Model (Set A)	Precision Rate if the Synonyms Model(Set B)	Overall Precision Rate	Overall Improvement in Recall
32%	84%	44%	107.4%

5.2 A analysis of the loss / gain in recall

To test the average recall improvement achieved with synonym collocation extraction, we experimented on three set tests with 9, 15, and 21 distinct headwords respectively. The results are shown in Table 6.

Table 6. Statistics of three test sets

Test 1			Test 2			Test 3		
Set A	Headwords	9	Set A	Headwords	15	Set A	Headwords	21
	Bi-grams	253		Bi-grams	703		Bi-grams	153
	Collocations	77		Collocations	232		Collocations	445
Set B	Synonym Headwords	55	Set B	Synonym Headwords	94	Set B	Synonym Headwords	121
	Bi-grams	614		Bi-grams	841		Bi-grams	203
	Non-synonym Collocations	179		Non-synonym Collocations	243		Non-synonym Collocations	576
	Extracted Synonym Collocations	201		Extracted Synonym Collocations	311		Extracted Synonym Collocations	554
	Synonym Collocations	178		Synonym Collocations	261		Synonym Collocations	476
Recall improvement: 99.49%			Recall improvement: 107.4%			Recall improvement: 82.6%		
Average improvement in recall: 96.5%								

The above table shows that the average recall improvement was close to 100% when using the synonyms relationships were used in the collocation extraction. With different choices of headwords, the improvement averaged about 100% with a standard deviation of 0.106, which indicates that our sampling approach to evaluation is reasonable.

5.3 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 based on both the precision rate and the estimated recall rate using so-called remainder bi-grams. The remainder bi-grams are all the bi-grams extracted by the algorithm. When the precision goes up, the size of the result is smaller, indicating a decreasing of recalled collocations. Figure 1 shows the precision rate and estimated recall rate recorded when we tested the value of θ.

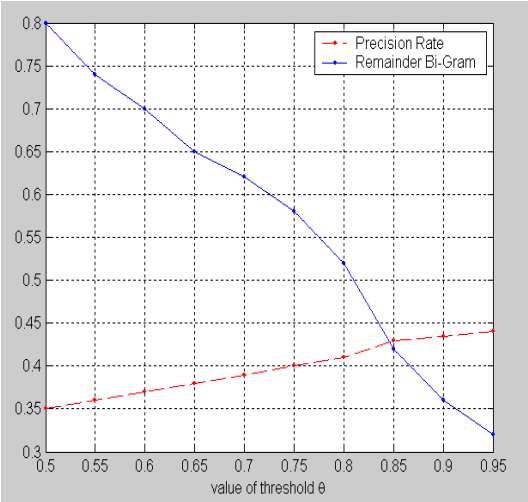


Figure 1. Precision rate vs. the value of θ

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

5.4 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 respect to synonym collocations, we also conducted some experiments to see whether the parameters should be adjusted. Table 7 shows the statistics used to test the value of (K_0, K_1, U_0). The similarity threshold θ was fixed at 0.85 throughout the experiments.

Table 7.Values of (K0, K1, U0)

	Bi-grams extracted using lexical statistics	Synonym collocations extracted in Step2
(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

The experimental results show that varying the value of (K_0, K_1) does not benefit our algorithm. However, increasing the value of U_0 does improve the extraction of synonymous collocations. Figure 2 shows that U_0=14 provides a good trade-off between the precision rate and the remainder Bi-grams. This result is reasonable. According to Smadja, U_0 as defined in equation (8) represents the co-occurrence distribution of the candidate collocation (w_h, w_c) at the position d ($-5 \le d \le 5$). For a true collocation (w_h, w_c, d), its co-occurrence frequency at the position d is much higher than those at other positions, which leads to a peak in the co-occurrence distribution. Therefore, it is selected by the statistical algorithm based on equation (10). Based on the physical meaning, one way to improve the precision rate is to increase the value of the threshold U_0. A side effect of increasing the value of U_0 is a decreased recall rate because some true collocations do not meet the condition of co-occurrence frequency in the ten positions greater than U_0. Step 2 of the new algorithm regains some true collocations that are lost because of the higher value of U_0 in Step 1.

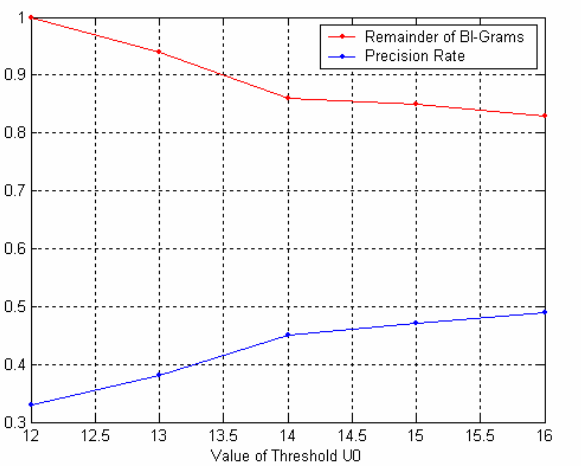


Figure 2. Precision rate vs. the value of U_0

5.5 A comparison of similarity calculation using equations (11) and (11^a)

Table 8 lists the similarity values calculated using equation (11), where α is a constant with a given value of 1.6, and equation (11^a), where α is replaced with a function of the depths of the nodes. Results show that (11^a) is finer tuned, and that it also 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 similar but not the same.

Table 8. Comparison of calculated similarity results

Word 1	Word 2	Formula(11)	Formula(11^a)
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

5.6 An Example

Table 9. Substitution of headwords and collocated words for the collocation "迅速增长"

Substitution headword	Substitution collocated word	Freq. in corpus	Freq. in Google results	Substitution collocated word	Freq. in corpus	Freq. in Google results
迅速增加		15	17,000	迅捷增长	0	7
迅速增多		2	14,900	迅速增长	20	224,000
迅速增高		0	744	飞快增长	0	2,530
	快速增长	111	1,280,000	飞速增长	4	48,100
	急遽增长	4	64,100	高速增长	60	543,000
	急促增长	0	201	火速增长	2	211
	急速增长	2	19,700	全速增长	3	607
	急骤增长	0	1,020	神速增长	0	55
	迅猛增长	4	84,600	麻利增长	0	0
	迅疾增长	0	98	湍急增长	0	0

The above example shows for the collocation "迅速增长", how each word is substituted and the statistical data for the synonym collocations. Our system extracts twenty candidate synonym collocations. Seven of them are synonym collocations with frequencies below than 10. Four of them have frequencies above 10, which means that they can be extracted by using statistical models only. Another nine of them do not appear in our corpus, which including two pseudo collocations "麻利增长"and "湍急增长".

6. Conclusions and On-Going Work

In this paper, we have presented a method to extract bi-gram collocations using a lexical statistics model with synonym information. Our method achieved a precision rate of 44% for the tested data. Comparing with the precision of 32% obtained using lexical statistics only, our method results in an improvement of close to 33%. In addition, the recall improvement achieved reached 100% on average. The main contribution of our method is that we make use of synonym information to extract collocations which otherwise cannot be extracted using a lexical statistical method alone. Our method can supplement a lexical statistical method to increase the recall quite significantly.

Our work focuses on synonym collocation extraction. However, Manning [Manning 99] claimed that the lack of valid substitutions for synonyms is a characteristic of collocations in general [Manning and Schutze 1999]. Nevertheless, our method shows that synonym

collocations do exist and that they are not a minimal collection that can be ignored in collocation extraction.

To extend our work, we will further apply synonym information to identify collocations of different types. Our preliminary study has suggested that collocations can be classified into 4 types:

Type 0 collocations: These are fully fixed collocations which including some idioms, proverbs, and sayings, such as "缘木求鱼"，"釜底抽薪" and so on.

Type 1 collocations: These are fixed collocation in which the appearance of one word implies the co-occurrence of another one as in "历史包袱".

Type 2 collocations: These are strong collocation which allow very limited substitution of components, as in, "裁减职位", "减少职位", "缩减职位" and so on. These collocations are classified with type 3 collocations when substitution can occur at only one end, not both ends.

Type 3 collocations: These are loose collocations which allow more substitutions of components; however a limitation is still required to restrict the substitution as in "减少开支", "缩减开支", "压缩开支", "消减开支".

By using synonym information and defining substitutability, we can validate whether collocations are fixed collocations, strong collocations with very limited substitutions, or general collocations that can be substituted more freely. Based on this observation, we are currently working on a synonym substitution model for classifying the collocations into different types automatically.

Acknowledgements and notes

Our great thanks go to Dr. Liu Qun of the Chinese Language Research Center of Peking University for letting us share their data structure in the Synonym Similarity Calculation. This work was partially supported by Hong Kong Polytechnic University (Project Code A-P203) and a CERG Grant (Project code 5087/01E). Ms. Wanyin Li is currently a lecturer in the department of Computer Science of Chu Hai College, Hong Kong.

References

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

Choueka, Y., "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,* 1993, pp. 21-24, Cambridge.

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

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

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

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

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

Lin, D. K., "Using Collocation Statistics in Information Extraction," *Proceedings of the Seventh Message Understanding Conference* (MUC-7), 1998.

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

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

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

Miller, G., WordNet, http://www.cogsci.princeton.edu/~wn/, 1998.

Miller, G and C. Fellbaum, "Semantic networks of English," *In Beth Levin & Steven Pinker (eds.), Lexical and conceptual semantics*, 1992, pp. 197-229.

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

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

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

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

Yang, E., G. Zhang and Y. Zhang, "The Research of Word Sense Disambiguation Method Based on Co-occurrence Frequency of HowNet," *Communication of COLIPS*, 8(2) 1999, pp. 129-136.

Yao, T., W. Ding and G. Erbach, "CHINERS: A Chinese Name Entity Recognition System for the Sports Domain," *Second SIGHAN Workshop on Chinese Language Processing*, 2003, pp. 55-62.