A Hybrid Extraction Model for Chinese Noun/Verb Synonym bi-gram Collocations *

Wanyin Li and Qin Lu
Department of Computing, The Hong Kong Polytechnic University,
Hung Hom, Kowloon, Hong Kong
csclaireli@gmail.com, csluqin@comp.polyu.edu.hk

Abstract. Statistical-based collocation extraction approaches suffer from (1) low precision rate because high co-occurrence bi-grams may be syntactically unrelated and are thus not true collocations; (2) low recall rate because some true collocations with low occurrences cannot be identified successfully by statistical-based models. To integrate both syntactic rules as well as semantic knowledge into a statistical model for collocation extraction is one way to achieve a high precision while keeping a reasonable recall. This paper designs a cascade system which employs a hybrid model by integrating both syntactic and semantic knowledge into a statistical model for Chinese synonymous noun/verb collocations extraction. The grammatically bounded noun/verb collocations are extracted first from a syntactic-rule based module, which is then inputted to a semantic-based module for further retrieval of low frequent bi-gram collocations.

Keywords: Collocation extraction, statistical model, syntactic rules, semantic relationship, similarity calculation, HowNet.

1. Introduction

According to (Benson, 1990), "collocation is an arbitrary and recurrent word combination", and (Manning, 1999), "A collocation is an expression consisting of two or more words that corresponds to some conventional way of saying things". The definitions imply the feasibility of statistics calculation on collocation identification, which has been widely employed by traditional collocation extraction systems (Dunning, 1993; Smadja, 1996; Sun, 1997; Manning, 1999). These statistical models which depended on word frequencies and association strength of co-occurrence (bi-grams) made them difficult to detect bi-gram collocations with lower frequencies. Moreover, bi-grams which occur with high frequencies may not be syntactically related. For example, "这/r 就/d 要/v 从/p 各自/r 的/u 实际/n 情况/n 出发/v" (It should be thought about according to the real condition). The bi-gram "实际(…)出发" bears a higher co-occurrence frequency from the corpus based on statistical models. However, it is syntactically ill-formed as a noun/verb phrases. (Choueka, 1993) defined collocations as "a sequence of two or more consecutive words, that has characteristics of a syntactic and semantic unit", and (Cowie, 1978) defined them as "co-occurrence of two or more lexical items as realizations of structural elements within a given syntactic pattern". This paper investigates how to extract the so called bi-gram synonym collocations which satisfy the synonym substitution rule in which a co-word or head-word of a given bigram can be replaced by another synonym word (Liu, 2002) and the substituted bigram is also existed. For example "重要/位置" and "主要/位置" are one synonym collocation pair because "重要" and "主要" are synonyms and "重要/位置" and "主要/位置" co-occurs in Chinese corpus. "重要/位置" and "重要/地位" is another synonym collocation pair because "位置" and "地位" are synonyms. Based on the

* Acknowledgments "The work is partially supported by a grant from The Chiang Ching-kuo Foundation for International Scholarly Exchange under the project RG013-D-09" .

previous work (Li et. al., 2005; Li and Lu, 2006), this work proposes a hybrid approach to employ the syntactic, semantic and statistical information for the extraction of syntactically bound synony bi-gram collocations**.** This paper focuses on collocation extraction in noun/verb phrases. A sub-model of HowNet based similarity calculation is proposed to identify bi-gram synonymous collocations derived from a base noun/verb phrase structure, say in the distance of [-5, +5], especially the ones in low frequency from a monolingual corpora. Our pattern generation process from the actual training data is similar to the works in (Seretan, 2005). The named syntactic patterns in this paper are automatically learned from the actual training data according to the *F*-score of the final extracted collocations by re-applying each pattern on the training corpus. Hence the syntactic patterns are corpus independent while most of the previous works arbitrarily choose the patterns in advance.

The rest of the paper is organized as follows. Section 2 studies related works. Section 3 describes the design of the system. Section 4 presents performance analysis. Section 5 is the conclusion.

2. Related Works

Researches on collocation extraction which make use of syntactic knowledge usually employ chunk-based approach to detect the syntactic patterns such as the constructions of <PP + Verb> (Krenn, 2001; Villada Moirón, 2004), <Verb + Noun> (Wu, 2003; McCarthy *et al*., 2003; Jian, 2004), <Noun + Noun> (Wang, 2003; Seretan, 2005), and <Adjective + Noun> (Seretan *et al.*, 2004). Jian (Jian, 2004) performed a *logarithmic Likelihoood Ratio* statistics with the integration of chunks, PoS tagging and clause knowledge achieved an average precision of 89.3% in the types of <Verb+Noun> collocation extraction. Drawbacks exist in the self-confliction of rules themselves as well as the parser precision. To deal with this, *t*-test is employed to measure the co-occurrence strength of the lexical pairs which satisfy the syntactic templates learned from the actual training data.

Researches on synonymous collocation extraction using semantic knowledge are mainly based on the similarity calculation. Lin (Lin, 1997) proposed a distributional hypothesis that if two words have similar sets of collocations, they are considered similar. According to (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. 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. Researchers (Lin, 1997; Pearce, 2001; Wu and Zhou, 2003) have applied similarity-based calculation for collocation identifications. Pearce identified collocations by relying on a mapping from one word to its synonyms for each of its senses. (Wu and Zhou, 2003) are the first researches to extract synonymous collocations by mapping synonyms relationships between two different languages to automatically acquire English synonymous collocations. However, this method needs a parallel corpus which is difficult to be obtained in real case.

3. System Design

Figure 1 shows the system framework consisted of two cascaded modules. Module I, labeled as BNP/BVP bi-gram Candidates Extractor, presents a syntax model from the chunked training corpus to generate collocation patterns of Base Noun Phrase (BNP) and Base Verb Phrase (BVP). To achieve this task, the model requires certain training data with base phrase chunking information although though it does not require the targeted collocations to be annotated. After successfully extraction of the patterns, the extracted patterns are then applied to extract candidate bi-grams which are further evaluated by *t*-test. The candidate collocations from Module I will be inputted into the Module II, Synonym BNP/BVP bi-grams Extractor, in which

a HowNet based similarity model is employed to extract the synonym noun/verb bi-gram collocation candidates.

Two text corpuses are utilized in the paper. One million training corpus, named corpusTrain (Xu, 2005), tokenized by linguists with chunking information as well as PoS tagging and another testing corpus with half a year People's Daily newspaper prepared by Peking University (PKU corpus), named corpusTest which contains 11 million words with PoS tags information only, which is also the training corpus used in Module II .

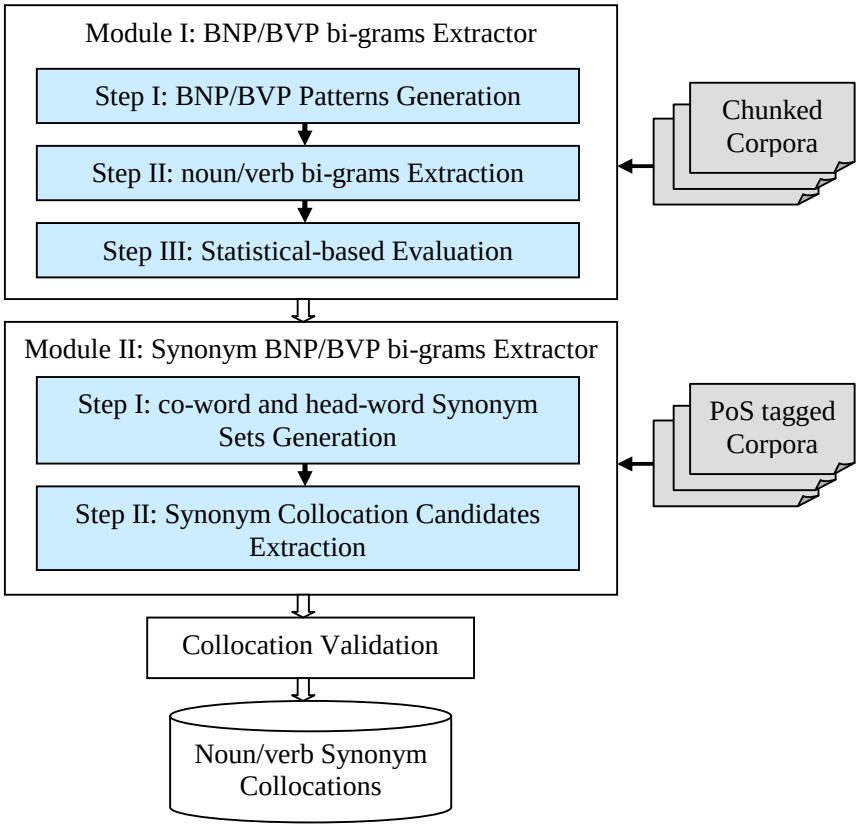

Figure 1: System Framework.

.

3.1 Module I – BNP/BVP bi-gram Collocations Extraction Model

Three steps are included in this Module:

- Step One: The syntactic pattern sets of noun/verb phrases are generated from the POS chunked corpusTrain with a raw pattern precision attached.
- Step Two: Taking a randomly selected Noun/Verb as an input word, named head-word $w_{h,}$ apply the syntactic patterns on corpusTest to extract candidate noun/verb phrases with respect to w_h. For example, given the head-word "新闻/n" and the syntactic pattern of [/n /n], the noun phrase" 新闻/n 媒介/n" will be such a candidate. The same candidate extracted when taking the head-word "媒介/n" as input will be eliminated.
- Step Three: Apply *t*-test to candidate noun/verb bi-grams of w_h to obtain syntactically bound bi-gram collocation candidates (w_h, w_c), where w_c identifies the collocated word of w_h.

3.2 BNP/BVP bi-gram Patterns Generation

BNP/BVP is a Noun/Verb Phrase that does not contain a Noun/Verb Phrase nor any Noun/Verb Phrase post-modifier. The syntactic bi-gram patterns of noun and verb phrases are automatically learned from the base phrase chunked training corpus corpusTrain. Then they are re-tested on corpusTrain in which the chunking information is removed. After which each pattern is attached with a pattern precision score such as the ones showed in the second column of Table 1. The precision threshold of the final syntactic patterns is determined from F-score of final extracted collocations, which is 30% for noun collocation patterns and 20% for verb collocation patterns. Table 1 contains the noun/verb bi-gram patterns which will be applied on corpusTest to extract the candidate noun/verb collocations.

Table 1: Bi-gram noun/verb phrases patterns.

Instances	Precision tested on corpusTrain	BNP Patterns
27,484	0.41	[/n /n]
10,856	0.53	[/vn /n]
8,421	0.38	[/n /vn]
7,198	0.62	[/a /n]
3710	0.61	[/b /n]
Instances	Precision tested on corpusTrain	BVP Patterns
22,267	0.22	[/v /n]
20,164	0.66	[/d /v]
19,548	0.43	[/v /v]
16,001	0.26	[/v /u]
3,681	0.69	[/ad /v]
2323	0.37	[/d /v /u]

3.3 Noun/Verb bi-gram Collocation Candidates Extraction

t-score is used to measure the co-occurrence strength of the syntactic bound bi-gram candidates because *t*-score achieves better performance than *z-score*, *MI*, χ^2 and *log-likelihood* for noun/verb phrase collocation extractions on Chinese corpus (Li and Lu, 2006). The first N-best from the output will be the syntactically bound noun/verb bi-gram collocations (w_h, w_c).

$$t-score = \frac{p(w_h, r, w_c) - p(w_h) \times p(w_c)}{\sqrt{\frac{p(w_h, r, w_c)}{N}}} = \frac{\frac{f(w_h, r, w_c)}{N} - \frac{f(w_h)}{N} \times \frac{f(w_c)}{N}}{\sqrt{\frac{f(w_h, r, w_c)}{N^2}}} \quad (1)$$

$f(w_h,r,w_c)$: frequency of collocations with head-word as w_h, co-word as w_c and relationship as r;
$f(w_h)$: frequency of head-word w_h; $f(w_c)$: frequency of co-word w_c
N : of the total instances of BNP/BVP;

3.4 Module II- Synonym BNP/BVP bi-gram Candidates Extraction Model

Two steps are included in this Module:

Step 1: For each Noun/Verb bi-gram collocation candidates (w_h, w_c) from Section 3.3, the synonyms against w_h and w_c are acquired respectively using the word similarity calculation based on HowNet (See Section 3.5 for details). Any word in HowNet having a similarity value above the threshold is considered a synonym head-word $w_{sh,}$ or a synonym collocated word w_{sc} for further extractions in Step 2.

- **Step 2**: For each synonym head-word w_{sh} of w_h and the collocated word $w_{c,}$, the bi-gram (w_{sh}, w_c) is taken as a synonym collocation if the pair appear at least once in the corpus. The same processing is applied to each of the synonym collocated word w_{sc} of w_c, for the bi-gram (w_h, w_{sc}).

3.5 Similarity Model Based on HowNet

The definition of synonyms in this Module is similar with the word similarity given by (Liu, 2002). A word in HowNet is defined as a set of concepts, and each concept is represented by its up to four different primitives classified as: *basic independent primitive* (weighted by β_1 in formula (4)), *other independent primitive* (weighted by β_2), *relation primitive* (weighted by β_3), and *symbol primitive* (weighted by β_4), where basic independent primitive and other independent primitive are used to calculate the semantic relationship between two concepts and the another two primitives are used to measure the syntactic relationships between two concepts. The definition of HowNet is described as a collection W of n words as below:

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

Each word w_i is described by a set of concepts S_{ij},

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

Each concept S_i is described by a set of primitives p_{ij}:

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

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

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

S_{1i} is the concept lists associated with w_1 and S_{2j} is the concept lists associated with w_2. Considering the semantic tree structure of HowNet, the primitive similarity for any two nodes p_1 and p_2 of the same primitive type can be expressed by the following formula:

$$Sim(p_1, p_2) = \frac{\min(d(p_1), d(p_2))}{Dis(p_1, p_2) + \min(d(p_1), d(p_2))} \tag{3}$$

where $d(p_i)$ is the depth of node p_i in the tree, $Dis(p_1, p_2)$ is the path length between p_1 and p_2 based on the semantic tree structure.

To integrate both semantic and syntactic information, the similarity between two concepts S_1 and S_2 is taken into consideration of all the four primitive types in weighted as:

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

$\beta_{i,i=1..4}$ is a weighting factor (Liu, 2002), where $\beta_1 + \beta_2 + \beta_3 + \beta_4 = 1$ and $\beta_1 \geq \beta_2 \geq \beta_3 \geq \beta_4$. The similarity model given here is the basis for building the synonym set where β_1 *and* β_2 represent the semantic information, and β_3 *and* β_4 represent the syntactic relationship.

3.6 Synonym Set

The synonyms set W_{syn_h} against the head-word w_h based on the similarity formula (4) is defined as below:

$$W_{syn\_h} = \{w_s : Sim(w_h, w_s) > \theta\} \quad (5^a)$$

The same definition against the co-word $w_{c,}$ of w_h to build up the synonyms set W_{syn_c} is:

$$W_{syn\_c} = \{w_s : Sim(w_c, w_s) > \theta\} \quad (5^b)$$

where $0 < \theta < 1$ is tuned from experiments (see Figure 3).

3.7 Synonym Collocations

We follow the idea from (Wu and Zhou, 2003) to define the synonym collocation pair as two collocations that are similar in meaning, but may not identical in wording. For a given collocation *(w_{sh}, $w_{c,}$ d)*, if $w_{sh} \in W_{syn\_h}$, then we deem the triple (w_{sh}, $w_{c,}$ *d)* as a synonym collocation with respect to the collocation *(w_h, $w_{c,}$ d)* if *(w_{sh}, $w_{c,}$ d)* appears at least once in the corpus, *d* identifies the position distance of [-5,+5] between w_h and w_c with respect to w_h within the running text line. Hence, the set of synonym collocations C_{syn_h} is defined as:

$$C_{syn\_h} = \{(w_{sh}, w_c, d) : Freq(w_{sh}, w_c, d) >= 1\} \quad (6^a)$$

Similarly, for $w_{sc} \in W_{syn\_c}$, the set of synonym collocations C_{syn_c} is:

$$C_{syn\_c} = \{(w_h, w_{sc}, d) : Freq(w_h, w_{sc}, d) >= 1\} \quad (6^b)$$

4. Experiments

To evaluate the proposed methodology, i.e., the effectiveness of true named collocations extraction. The strategies of a collocation dictionary and human judgment are applied. Firstly, the N-best bigrams are evaluated against the collocation dictionary built up from the colleagues in our NLP laboratory. Secondly, the remainder bi-gram candidates are then judged manually to evaluate how many among the N-best scored bi-grams are true collocations. The performance of the hybrid approach is evaluated by N-best strategies supplemented with precision and a so called local recall defined as below:

$$Local\ \mathrm{Re}call = \frac{\text{The number of Correctly Identified Collocations}}{\text{Toal number of True Collocations}} \quad (7)$$

Where total number of true collocations is defined by adding the extracted collocations from both the hybrid and statistical-based approaches and then sorted in N-best without duplication. The performance of the proposed approach is compared with the pure statistical-based approach, the syntax-based approach and the semantic-based approach.

4.1 Evaluation of Module I

30 nouns and 30 verbs are randomly selected as the head-words. Table 2 shows the performance by comparing with the statistical-based model (Xu, 2003) in which the returned word list have been further processed against the 30 nouns and verbs with the syntax-based model.

Table 2: Comparison of syntactic & statistical models.

		Extracted bigrams	Prec. Rate	Local Recall	F-Score
30 Noun Head-words					
Rule Prec. >30%	Rule-Based	3,497	81.01%	59.63%	68.69%
	Refined by *t*-test	3,114	83.26%	58.08%	68.43%
Statistical Model Only		1,484	78.84%	26.15%	39.27%
30 Verb Head-words					
Rule Prec. >20%	Rule-Based	2,615	71.74%	62.62%	66.87%
	Refined by *t*-test	2,398	75.43%	64.25%	69.39%
Statistical Model Only		818	73.15%	21.24%	32.92%

Figure 2 shows the precision variation against the local recall by the *t*-test which achieves the precision rate up to 88.39% in noun collocations and up to 83.21% in verb collocations when taking the first 70% of each respectively.

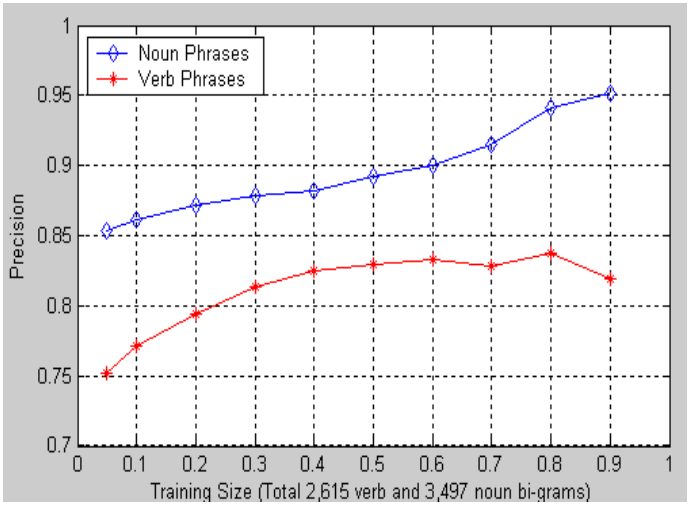

Figure 2: Variation of precision and local recall.

4.2 Evaluation of Hybrid Model

Module II aims to extract the collocations in lower co-occurrence frequency especially the ones appear less than three times in the corpus (Li et. al., 2005).

Taking total 4,278 (3,497*0.7+2,615*0.7) bi-gram noun and verb collocations from Module I as the input to Module II, total 2,573 synonym head-words and co-words are acquired with the tuned value of θ = 0.9. After Step 2 of Module II, additional 6,051 bi-gram synonym collocation candidates are extracted. 5,417 of them are true collocations to a evaluate set of manually checked "true positive" (Table 3).

Table 3: Precision of synonym collocations extraction.

Head-words in Noun	30
Synonym head-words	1,129
Synonym bi-grams extracted in Step 2 of Module II	3,078
True synonym collocations obtained in Step 2 of Module II	2,802
Precision Rate	91.03%
Head-words in Verb	30
Synonym head-words	1,444
Synonym bi-grams extracted in Step 2 of Module II	2,973
True synonym collocations obtained in Step 2 of Module II	2,615
Precision Rate	87.95%
Overall Precision Rate	89.49%

Figure 3 shows the variation of the value of θ in equation 5 with F-value, which has its best value of 0.9.

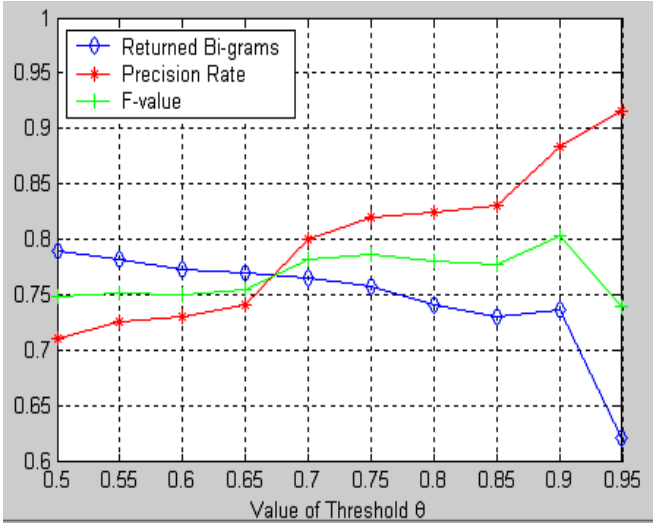

Figure 3: The choice of θ.

Table 4 presents the performance comparison for statistical-based approach, syntactic integrated approach (Module I), semantic integrated approach (Module II), and finally the hybrid approach.

Table 4: Comparison of statistical-based, Module I, II and cascade hybrid approaches.

	Models	Extracted bi-grams	Precision Rate	Local Recall Rate	F-Score
	Statistical-based	1,484	78.84%	15.52%	25.93%
	Module I Only	2,948	86.39%	33.78%	48.57%
	Module II Only	3,078	91.03%	37.16%	52.78%

Noun	Hybrid (I+II)	6,026	88.77%	70.94%	78.86%
Verb	Statistical-based	818	73.15%	10.34%	18.12%
	Module I Only	2,058	78.21%	31.24%	44.65%
	Module II Only	2,973	87.95%	45.19%	59.70%
	Hybrid (I+II)	5,031	87.91%	76.43%	81.77%
Noun + Verb		11,057	88.34%	73.69%	80.35%

5. Conclusion

The paper proposes a hybrid model to identify Chinese Noun and Verb synonym bi-gram collocations. The system achieves an average precision of 88.34% and local recall of 73.69%. The approach, aimed at low frequency word pairs which lead up to poor performance in pure statistical-based associate measure approaches, proposes distinct models reflecting diverse linguistic knowledge to improve both precision rate and recall rate.

The paper is focused on the extraction of bi-gram noun/verb synonym collocations, especially the ones may have a lower co-occurrence of less than 3 times in the running texts. Fu (Fu et al., 2010) made use of semantics of n-grams as the set of all the texts containing it to measure the information distance of n-grams which makes the n-grams of any length to their semantics applicable. To extend the current work in this paper to n-grams of free length is one valuable inspiration in the future.

References

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

Choueka Y., 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.

Cowie A. P., 1978. The place of illustrative material and collocations in the design of a learner's dictionary. *In P. Strevens, editor, In Honour of A. S. Hornby*, pages 127—139. Oxford University Press.

Dunning Ted, 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Fan Bu, Xiaoyan Zhu and Ming Li, Measuring the Non-compositionality of Multiword Expressions (2010).*COLING '10 Proceedings of the 23rd International Conference on Computational Linguistics*.

Jian, J. Y., Chang, Y. C., & Chang, J. S., 2004. Collocational Translation Memory Extraction Based on Statistical Linguistic Information. *ROCLING 2004, Conference on Computational Linguistics and Speech Processing*, Taipei.

Krenn B. and Evert S., 2001. Can we do better than frequency? a case study on extracting pp-verb collocations. *In Proceedings of the ACL Workshop on Collocations*, Toulouse, France.

Li W.Y, Lu Q. and Xu R. F., 2005. Similarity based Chinese Synonyms Collocation Extraction. *International Journal of Computational Linguistics and Chinese Language Processing*, vol.10, no.1, pp.123-144.

Li W.Y., Lu Q., 2006. Collocation Extraction for Noun Phrases using Shallow Parsing Rules and Statistical Models. *In Proceeding of 20th Paclic Asia Conference on Language, Information and Computation*, pp. 99-106, Wuhan, China.

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

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

Manning C. and Sch¨utze H., 1999. Foundations of Statistical Natural Language Processing. *MIT Press*, Cambridge.

McCarthy, D., B. Keller and J. Carroll (2003) `Detecting a continuum of compositionality in phrasal verbs'. *In Proceedings of the ACL-SIGLEX Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, Sapporo, Japan.

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

Villada Moirón, M. B., 2004. Acquisition of Dutch support verb collocations: a model comparison .ms. Groningen University. URL: http://www.let.rug.nl/˜begona/papers/svcmodels.ps.

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

Smadja Frank, Kathleen R., Mckeown, Vasileios H., 1996. Translation collocations for bilingual lexicons: a statistical approach. *Computational Linguistics*, 22:1-38.

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

Violeta Sereran, Luka Nerima, Eric Wehrli, 2004. Using the Web as a Corpus for the Syntactic-Based Collocation Identification. *In Proceedings of International Conference on Language Resources and Evaluation* , Lisbon, Portugal, pp. 1871-1874

Violeta Seretan, 2005. Induction of Syntactic Collocation Patterns from Generic Syntactic Relations. *In Proceedings of Nineteenth International Joint Conference on Artificial Intelligence (IJCAI 2005)*, Edinburgh, Scotland, pp. 1698-1699.

Wang Xinglong, Carroll J., 2003. Acquisition Of Collocations. *Technical Report for the MEANING Project (IST-2001-34460).* University of Sussex, UK.

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

Wu Andi., 2003. Learning Verb-Noun Relations to Improve Parsing. *In Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*, pp. 119-124.

Xu R. F., Lu Q., and Li Y., 2003. An automatic Chinese Collocation Extraction Algorithm based on Lexical Statistics. *In Proceedings of IEEE International Conference on Natural Language Processing and Knowledge Engineering*, Beijing, china, 321-326.

Xu R. F., Lu Q., Li Y., Li W. Y., 2005. The design and construction of the PolyU Shallow Treebank. *International Journal of Computational Linguistics and Chinese Language Processing*, Vol.10, no.3.