TCtract-A Collocation Extraction Approach for Noun Phrases Using Shallow Parsing Rules and Statistic Models

Wan Yin Li, Qin Lu, James Liu

Department of Computing
The Hong Kong Polytechnic University
Hung Hom, Kowloon, Hong Kong
{cswyli, csluqin, csnkliu}@comp.polyu.edu.hk

Abstract. This paper presents a hybrid method for extracting Chinese noun phrase collocations that combines a statistical model with rule-based linguistic knowledge. The algorithm first extracts all the noun phrase collocations from a shallow parsed corpus by using syntactic knowledge in the form of phrase rules. It then removes pseudo collocations by using a set of statistic-based association measures (AMs) as filters. There are two main purposes for the design of this hybrid algorithm: (1) to maintain a reasonable recall while improving the precision, and (2) to investigate the proposed association measures on Chinese noun phrase collocations. The performance is compared with a pure statistical model and a pure rule-based method on a 60MB PoS tagged corpus. The experiment results show that the proposed hybrid method has a higher precision of 92.65% and recall of 47% based on 29 randomly selected noun headwords compared with the precision of 78.87% and recall of 27.19% of a statistics based extraction system. The F-score improvement is 55.7%.

Keywords: Collocation Extraction**,** Typed Collocations, Phrase Rules, Association Measures.

1 Introduction

Statistical approaches in collocation extraction still need improvement in terms of precision ([1], [10], [11]). There are two reasons for this. First, any given set of words has a very large number of possible combinations. One way to make statistical approaches more precise is simply to reduce the context to a fixed small size window of consecutive words. However, this method simply removes collocations beyond the span of the predefined window. The second cause of imprecision is that grammatical analysis itself is imprecise. Candidate collocations which are retrieved using statistical approaches may be "grammatically" related in the extracting system's terms yet in fact consist of syntactically unrelated items which do not qualify as valid collocations. Ultimately, in the absence of massive computing power or extensive word lists the problems with statistical approaches would appear to be insuperable. For this reason it would appear that it may be useful to introduce other identifying features as hybrids to improve the precision of collocation extraction. On the face of it, linguistic knowledge is one area that would seem to offer an abundance of potentially useful text-related features such as semantic knowledge, morphological knowledge and, as will be discussed in this paper, syntactic knowledge,

Studies that have made use of syntactic knowledge mainly fall into two broad categories. The first category makes use of syntactic filters ([1], [21]), which will not be considered in this paper. The second category makes use of collocation patterns ([3]), such as Adjective-Noun, Noun-Noun, Verb-Noun, Verb-Object, Subject-Verb, etc. However, the syntactic knowledge, such as phrase rules in this work, is mainly used for identifying occurrences of particular phenomena of certain given rules. It is not likely to apply the phrase rules individually because both inter-conflicts and the precision of the rules tend to introduce noise if no further strategies are used. Furthermore, there exist statistical regularities in natural languages. Thus, the hybrid syntactic-statistical approaches seem to be viable in collocation extraction, and indeed, such approaches to collocation extraction have been applied in a number of projects using European language corpora. These will be described in the next section.

In this paper, we present the *TCtract* system, a syntactic-based Chinese collocation extraction system enhanced by the use of additional statistical measures. Central to the design is that the extracted candidate collocations should be a pair having a defined grammatical relationship. A recent study on collocation [2] expressed a similar idea, saying that "collocation research is especially valuable if it aims at finding typed collocations, that is, collocations selected on the basis of some morpho-syntactic properties, as opposed to the extraction of typeless ones". At this stage, *TCtract* extracts one kind of typed collocation, noun phrase collocations. These collocations are represented in a set of phrase rules validated by using a test corpus and manual checking. It will be easy to extend the rule-based representation in the future to include other linguistic knowledge such as grammar, syntax, phrase and semantics by adding related rules to represent their defining relations.

The rest of the paper is organized as follows. Section 2 gives a brief review of the works for the automatic acquisition of typed collocations in syntactic-based method. Section 3 describes the Framework of the approach. Section 4 presents the performance comparison of the system with a statistic-based approach. Section 5 concludes the work and discusses possible extension in the future.

2 Previous Work

Early statistic based collocation extraction systems ([1], [4]) mainly use syntactic knowledge for filtering or error correction. These statistic models depended on the word frequency and association strength of co-occurrence (bi-grams) making them difficult to detect low frequency collocations. More recent works ([6], [7], [8], [21]) show a growing interest in integrating syntactic components in performing collocation extraction over (shallow) parsing trees rather than over sentences with the availability of linguistic knowledge such as collocation dictionary and shallow paring tree bank([4], [17], [18]). Furthermore, there are interesting reports on retaining only the concordance tokens of certain syntactic patterns, namely typed collocation, such as in the types of <PP+Verb> ([5], [13]), <Verb+Noun> ([6], [7]), <Noun+Noun> ([8], [9]).

The existing linguistic features used in these approaches include chunking information, PoS tagging, clause information, head identification, and sentence boundary. A recent work discussed in [7] performed a *logarithmic Likelihoood Ratio* statistics with the integration of chunks, PoS tagging and clause knowledge to achieve an average precision of 89.3% in <Verb+Noun> collocation extraction. The observation on the features selection is that it should rely on the nature of the target languages, the properties of the applied corpus, the candidate extraction strategies, and the type of collocations to be identified. For example, in English, the structure of noun phrase is head-first such as in the example "the book he liked" which differs from the head-last structure in Chinese such as "他喜欢的书". A more complicated example in Chinese can be seen in the in "会议通过并颁布了[澳门/ns 特别/a 行政区/n 基本法/n]BNP" where "基本法" is the head in the last position. Thus, for Chinese, the last noun in the noun phrases is normally considered the default head.

The associate measures (AMs) tested in recent reports include the most commonly used *MI(mutual information)*, χ^2-test, *log-likelihood* ratio, *t*-score, and some less used measures such as *log-linear* model adopted by [9], *Fisher's score* tested by [15], and *relative entropy* model evaluated by [14]. A detail evaluation involved in the <PP+Verb> collocation extraction reported in [13] showed that *t*-score achieves the best precision values over other AMs. Another report from [5] showed that the *log-likehood* and χ^2 *-test* works well for the identification of support verb constructions. The statistical measures applied in this paper include frequency, mutual-information, *z-score*, χ^2, *log-likelihood*, *t*-score. As there are rare attempts made on typed collocation extraction in Chinese, one of aims of this work is to evaluate the usefulness of different association measures for extracting typed collocations from Chinese corpus.

3 TCtract

Collocations are defined as a recurrent and conventional expression of words which holds semantic relations and fits predefined syntactic structures. This paper focuses on the bi-gram noun phrase collocations extraction.

3.1 Resources and Evaluation Methods

Two corpuses are used in the experiments: one is a one million small data corpus, namely corpusS [18], tokenized by linguists with chunking information as well as PoS tagging. Another is a larger corpus with half a year People's Daily newspaper prepared by Peking University [20], called corpusL which contains 11 million data with PoS tag information only. The BNP patterns are extracted from corpusS first. Then, a set of candidate lexeme pairs matched the types predefined are extracted from corpusL and further passed to the association measures procedure for further evaluation. As stated early, different association measures may be effective in the exaction of different types of collocation. In this work, AMs applied include *MI*, *z*-score, χ^2-test, *log-likelihood* ratio, and *t*-score.

3.2 The Rule Base

The rule base initially includes only phrase rules and later enhanced with syntactic and semantic rules. There are 11 types of phrase types in the corpus including BNP, BAP, BVP, BDP, BQP, BTP, BFP, BNT, BNS, BNZ, and SV (see details in [18]). This paper focuses none phrase (BNP) collocation extraction only.

3.3 The Hybrid Approach

The hybrid approach to extract noun phrase collocation consists of three stages: (1) preprocessing for data preparation, (2) noun phrase collocation extraction to assign each extracted candidates a weight to indicate its co-occurrence strength based on the AMs applied. (3) pseudo- collocation elimination.

Stage One: Dataset preparation which includes two steps.

- Step One: extracts the temporal rule set of noun phrases based on the BNP chunk in corpusS. The temporal rule set is divided into an Accept Rule Set (Aset) and Reject Rule Set (Rset). Then, the temporal rule set is further tested on the same closed test data without the chunking labels and verified manually to validate the rule sets. The rules were also supplemented from other sources as will be discussed later.
- Step Two: Applies Aset to corpusL to extract candidate noun phrases.

Stage Two: Noun phrase collocation extraction which consists of two steps.

- Step One: Divides the candidate noun phrases into bi-gram noun phrases and n-gram noun phrases because of the different statistic measures applied to.
- Step Two: Applies statistic association measures (AMs) discussed in Section 3.3.2 to bi-gram noun phrases to obtain candidate bi-gram noun phrase collocations.

Stage Three: Pseudo-collocation elimination by using rejection rules.

- Applies the Rset to candidate bi-gram noun phrase collocations. The result in this stage is considered noun phrase collocations.

3.3.1 Extracting Noun Phrase Patterns

The noun phrase rules are generated from the instances which are initially extracted from corpusS, then refined by re-testing on the same corpus without BNP chunks, and finally manually checked as shown in **Table 1**. An additional set of noun phrase rules are extracted from a manually verified collection of 4,300 collocations, referred to as the Golden Answer Set previously [20].

Table 1. Bi-gram Noun Phrases.

# of instances	Precision tested on corpusS	Refined BNP Pattern
27484	0.41	[/n /n]
16306	0.81	[/nr /nr]
10856	0.53	[/vn /n]
8421	0.38	[/n /vn]
7198	0.62	[/a /n]
3710	0.61	[/b /n]
+[/i /n] [/l /n]	[/p /n] [/s /n]	[/q /n] [/j /n]

+the last line contains the supplement rules based on the Golden Answer Set

As the structure of noun phrase in Chinese is head-last, head-last patterns are also added to catch the co-words appeared within a window of five words (n-gram) specified as below in **Table 2**:

Table 2. N-gram Noun Phrases.

# of instances	Precision tested on corpusS	Head-last BNP Rules
774	0.30	[/*(1, 5) /n]
16306	0.31	[/*(1, 5) /vn]

*means the PoS types included in **Table 1**

The rule set is further extended by looking for one more PoS on either the left or the right of an examining BNP chunk. For example, in the example "在/p[长期/b 艰苦/a]BAP 的/u [斗争/vn 岁月/n]BNP 里/f, /w", the extended rule would be /u [/vn /n]BNP /f. This extension was only carried out on the first five patterns shown in **Table 1** which covered over 90% BNP instances in the whole corpus. All the collected rules are then tested on corpusS and checked manually to divide them into Aset and Rset with the distribution information shown in **Table 3**.

Table 3. Extended Bi-gram Rules.

	# of instances	# of Acceptation Rules	# of Rejection Rules
l_1[/n /n]r_1	7202	34	86
l_1[/nr /nr]r_1	5802	24	47
l_1[/vn /n]r_1	4083	20	36
l_1[/n /vn]r_1	2191	18	30
l_1[/a /n]r_1	1896	22	15

3.3.2 The AMs Measure

According to collocation's recurrence assumption of correlation between statistical association, five association measures are used to measure the candidate noun phrase pairs obtained from the Stage one as listed below:

1. *Log-likelihood* ratio

The *log-likelihood* in [7] is defined as formula (1):

$$LLR(x;y)_i = -2\log_2 \frac{p_1^{k_1}(1-p_1)^{n_1-k_1}(1-p_2)^{n_2-k_2}}{p^{k1}(1-p)^{n_1-k_1}\, p^{k2}(1-p)^{n_2-k_2}} \qquad \textbf{(1)}$$

where

k_1: of pairs contain x and y simultaneously; k_2: of pairs contain x but not y

n_1: of pairs contain y; n_2: of pairs that does not contain y

$p_1 = k_1/n_1$; $p_2 = k_2/n_2$; $p = (k_1 + k_2) / (n_1 + n_2)$

2. *t*-score, *z*-score, *MI*

The statistical measures *t*-score, *z*-score, and *MI* are formulated below:

$$z-score = \frac{O-E}{\sqrt{E}} \quad (2)$$

$$t-score = \frac{O-E}{\sqrt{O}} \quad (3)$$

$$MI = \log\frac{O}{E} \quad (4)$$

$$E = \frac{f_x \cdot f_y}{N} \quad (5)$$

where

N : of the total instances of BNP; O : of the total instances of pair (x;y)

f_x: of the total instances of x; f_y: of the total instances of y

The above methods try to compare the observed frequencies of collocation candidates with the expected frequencies based on the assumption of independence in the target pairs (x:y). Krenn [13] did a thorough evaluation among t-score, z-socre and MI measures and showed that t-score over performed the other AMs for <PP+Verb> collocations in a German corpus. This work define**s** the instances of bi-gram BNP as: <n+n>; <nr+nr>; <vn+n>; <n+vn>; <a+n>; <m+n>; <r+n>; <b+n>; <i+n>; <l+n>; <p+n>; <s+n>; <q+n>; <j+n>.

3. χ^2-test

The χ^2-test is formulated below:

$$\chi^2 = \frac{N(f_x \cdot O_{\overline{x}\overline{y}} - O_{\overline{x}y} \cdot O_{x\overline{y}})^2}{f_x \cdot f_y \cdot (O_{\overline{x}\overline{y}} + O_{\overline{x}y}) \cdot (O_{x\overline{y}} + O_{\overline{x}\overline{y}})} \quad (6)$$

Where

N : of the total instances of BNP; O : of the total instances of pair (x;y)

f_x: of the total instances of x; f_y: of the total instances of y

$O_{\overline{x}\overline{y}}$: of pairs do no contain x and y simultaneously

$O_{\overline{x}y}$: of pairs contain y but not x; $O_{x\overline{y}}$: of pairs contain x but not y

The statistical measures given in formulas (2) to (4) assume that the data are normally distributed which is proven to be untrue for English [16]. It is also been proven untrue for Chinese corpus [20]. Therefore, the χ^2-test is used in this work because it does not assume normal distribution probabilities.

3.3.3 Rejection Rules

After applied the association measures in Stage two, there are still some pseudo-collocations remained because of their high co-occurrence frequency which makes the AMs identify them as collocations. For example “很多/m 机遇/n”, “亿万/m 资产/n”, “某些/r 单位/n”, “各项/r 资金/n”, “某种/r 道德/n” etc. These pseudo-collocations will be weeded out by rejection rules such as [/r /n] and [/m /n] when they are identified by these rejection rules. As another case, when there are multiple nouns appear together, they are often being identified as bi-gram collocations because of the bi-gram rule [/n /n], but are not correct. For example, in the none bi-gram examples, “国际/n 关系/n 学院/n”, “军事/n 全球/n 定位/vn 系统/n”,

"学校/n 师生/n 学术/n 水平/n", the underlined noun pairs are not true collocations and thus should be avoided to be extracted. Thus in the n-gram rules, [/n /n] becomes a rejection rule. In addition, a relative n-gram accepting rule [/*(1, 5) /n] is added as shown in **Table 2**.

4 Experimental Result

The evaluation was conducted based on the five AMs with the performance shown in **Fig. 1**. To investigate the proposed method with the statistic-based approach, a pure statistical based system [20] is used as the baseline for comparison.

4.1 Candidate Sets

After Stage One, 63,225 BNP instances are returned which fall in 9,740 different BNP patterns. The number is reduced to 463,531 instances in 2,781 types after deleting the ones with precision less than 30% and the frequency $f < 3$. The phrase rules are further refined as given in **Table 1** which has the total coverage over 90% of the whole corpus. To test the AMs, 29 noun headwords are randomly selected against the selected headwords. After applied the refined phrase rule sets on the corpusL, a total of 3,497 candidate bi-gram noun phrase collocations. The precision is 83.58% verified manually.

4.2 Evaluation with AMs

In Stage Two, different AMs are applied to investigate their contribution to noun phrase collocation extraction. Therefore, for each AMs, the improvement of precision is evaluated against the loss of recall by sorting the first-n candidate collocations obtained from Stage One. The precision curves for the five AMs are presented in **Fig. 1**.

From the experiment results, *t*-score achieved slightly better performance than the other three AMs for the noun phrase collocation while *Log-likelihood* has the worst performance. This result on log-likelihood agrees with Evert's [19] observation for adjective-noun collocations candidate data.

Fig. 1 shows that the *t*-score achieves precision of 87.87% with the recall loss of 30%, which is 4.29% improvement compared to 83.58% by applying the pure phrase rule-based stage. Under this case, we obtain 50% of the recall which is not the main target of the method. Hence, we hope to mainly compare the precision while maintaining a reasonable recall.

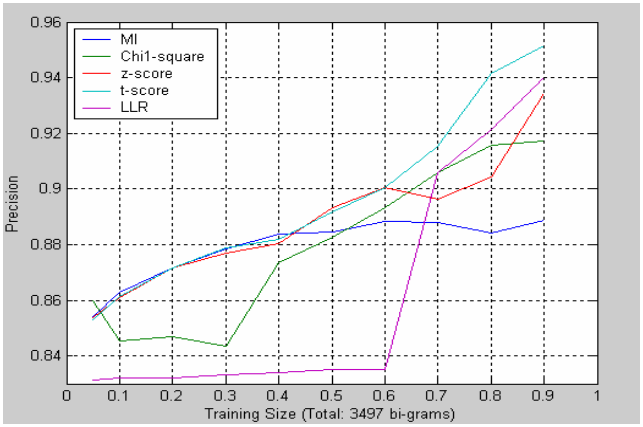

Fig. 1. The Precision Curves for the Five AMs (In colorful curve).

4.3 Comparison of the Proposed Method with Statistic Method

The rule-based hybrid method are compared with the baseline statistical model [20] using a bi-direction strength, spread and χ^2 values. The results are listed in **Table 4** and **Table 5**.

29 randomly selected noun headwords are applied in both systems (see **Table 4**). In the recall rate evaluation, the total number of actual collocation is calculated by adding up the extraction results of *TCtract* and the baseline system in a total of 4,300 verified manually. Results show that the performance of *TCtract* is much better than the baseline in terms both precision and recall (see **Table 5**) with the F-score improvement of 55.7%. The majority type of collocations missed was the verb-noun collocations which are defined as verb phrases with the phrase pattern of [v + n]BVP.

Table 4. Comparison of Precision and Recall for *TCtract* and Statistic Model.

	Headword	Extracted bi-grams	True Collocations	Precision Rate	Recall Rate	F-Score
Rule-based	29	3497	2922	83.58%	57.95%	63.92%
Refined by AMs	29	2448	2151	87.87%	50%	63.73%
Eliminate by Stage Three	29	2182	2021	92.65%	47%	62.95%
Statistic Model	29	1484	1169	78.84%	27.19%	40.43%

Table 5. Comparison of *TCtract* and Statistic Model.

Overlaped	Appeared in Statistic Model But missed in *TCtract*	Appeared in *TCtract* But missed in Statistic Model
636	848	2801

5 Possible Extensions and Future Work

This work aims to extract noun phrase Chinese collocations. An encouraging result from the experiments with the precision of 92.65% and recall of 47% is obtained. The future work will further investigate other types of collocations such as base verb phrase, base adjective phrase collocations and hope to prove or summarize that certain AMs are more suitable for identifying some classes of collocations than others.

Another direction is to employ the syntactic-rules to explore the grammatically well structured collocations such as <Verb + Object>, <Subject + Verb> etc. Furthermore, chunking the sentences into smaller syntactic structures by using the chunking information makes it easier to identify the adjacent relationship between each chunk, hence recognize n-grams collocations. For example, BVP+BNP is one of valid sequences for the most prevalent <Verb + Noun> collocation, from the clause "[会/v 铸就 /v]BVP [高尚/a 的/u 灵魂/n]BNP", a n-gram collocation of "铸就高尚的灵魂" can be extracted.

6 Acknowledgements

This work is supported by the Hong Kong Polytechnic University (Project Code Z-08K and RT31).

References

1. Frank Smadja, Kathleen R., Mckeown, Vasileios Hatzivassiloglou: Translation collocations for bilingual lexicons: a statistical approach. Computational Linguistics, (1996) 22:1-38
2. Kis, Balázs., Villada, B., Bouma, G., Biro, T., Nerbonne, J., Ugray, G. and Pohl, G.: A new approach to the corpus-based statistical investigation of hungarian multi-word lexemes. Proceedings of 4th International Conference on Language Resources and Evaluation (LREC) 2004, Lisbon, Portugal (2004) pp. 1677-80
3. Yan Zhang, Bo Xu and Chengqing Zong: Rule-based Post-processing of Pinyin to Chinese Characters Conversion System. ISCSLP'2000, Beijing, china (2000) pp.291-294
4. Dekang Lin: Extracting collocation from Text corpora. First Workshop on Computational Terminology, (1998) pp. 57-63
5. Villada Moir´on, M. B.: Acquisition of Dutch support verb collocations: a model comparison .ms. Groningen University. (2004) URL: http://www.let.rug.nl/˜begona/papers/svcmodels.ps.
6. Wu, Andi: Learning Verb-Noun Relations to Improve Parsing. Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, (2003) pp. 119-124
7. Jian, J. Y., Chang, Y. C., & Chang, J. S.: Collocational Translation Memory Extraction Based on Statistical Linguistic Information. Paper presented in ROCLING 2004, Conference on Computational Linguistics and Speech Processing, Taipei (2004)
8. Xinglong Wang, John Carroll: Acquisition Of Collocations, http://www.lsi.upc.es/~nlp/meaning/meaning.html (2003)
9. Violeta Seretan: Induction of Syntactic Collocation Patterns from Generic Syntactic Relations. In Proceedings of Nineteenth International Joint Conference on Artificial Intelligence (IJCAI 2005), Edinburgh, Scotland (2005) pages 1698-1699
10. Christopher Manning and Heinrich Sch¨utze: Foundations of Statistical Natural Language Processing. MIT Press, Cambridge (1999)
11. Ted Dunning: Accurate methods for the statistics of surprise and coincidence. Computational Linguistics (1993) 19(1):61–74
12. Frank Smadja: Retrieving collocations form text: X-tract. Computational Linguistics (1993) 19(1):143–177
13. Brigitte Krenn and Stefan Evert: Can we do better than frequency? a case study on extracting pp-verb collocations. In Proceedings of the ACL Workshop on Collocations, Toulouse, France (2001)
14. Krenn, B.: Empirical implications on lexical association measures. Proceeding of the Ninth EURALEX International Congress, Stuttgart, Germany (2000)
15. Weeber, M., Vos, R. And Baaye, H.: Extracting the lowest-frequency words: Pitfalls and possibilities. Computational Linguistics (2000) 26(3), pp 301-317.
16. K.W. Church, and R. L.Mercer: Introduction to the Special Issue on Computational Linguistics Using Large Corpora. Computational Linguistics (1993) Vol. 19, pp. 1-24
17. http://www.cis.upenn.edu/~treebank/
18. Xu R. F. et al., The design and construction of the PolyU Shallow Treebank, *International Journal of Computational Linguistics and Chinese Language Processing*, vol.10, no.3, 2005
19. Stefan Evert, Ulrich Heid, and Wolfgang Lezius, Methoden zum Vergleich von Signifikanzmaßen zur Kollokationsidentifikation. In Proceedings of KONVENS 2000, VDE-Verlag, Germany, (2000) pp. 215 – 220.
20. Ruifeng Xu, Qin Lu, and Yin Li, An automatic Chinese Collocation Extraction Algorithm based on Lexical Statistics, In Proceedings of IEEE International Conference on Natural Language Processing and Knowledge Engineering, Beijing, chian, (2003) pp.321-326
21. Seretan, Violeta, Luka Nerima, Eric Wehrli: Using the Web as a Corpus for the Syntactic-Based Collocation Identification. In Proceedings of International Conference on Language Resources and Evaluation *(LREC 2004)*, Lisbon, Portugal (2004) pages 1871-1874