The Design and Construction of A Chinese Collocation Bank

Ruifeng Xu[*], Qin Lu[*], Sujian Li[**]
* Dept. of Computing, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong
** Institute of Computational Linguistics, Peking University, Beijing, China
E-mail: csrfxu@comp.polyu.edu.hk, csluqin@comp.polyu.edu.hk, lisujian@pku.edu.cn

Abstract

This paper presents an annotated Chinese collocation bank developed at the Hong Kong Polytechnic University. The definition of collocation with good linguistic consistency and good computational operability is first discussed and the properties of collocations are then presented. Secondly, based on the combination of different properties, collocations are classified into four types. Thirdly, the annotation guideline is presented. Fourthly, the implementation issues for collocation bank construction are addressed including the annotation with categorization, dependency and contextual information. Currently, the collocation bank is completed for 3,643 headwords in a 5-million-word corpus

1. Introduction

Collocation is a lexical phenomenon in which two or more words are habitually combined and commonly used in a language to express certain semantic meaning. The combined use of words, collocation, is an important lexical knowledge. It is essential for distinguishing different senses of a word in context and identifying an appropriate word to fill in the given context. Consequently, collocation knowledge were widely employed in many natural language processing systems, such as in word sense disambiguation [Sinclair 1991], machine translation [Gitaski et al. 2000], information retrieval [Mitr et al. 1997] and natural language generation [Smadja 1993]. Although everyone understands the importance of collocation knowledge, it still cannot be compiled easily into a collocation dictionary. Even the definition of collocation is subject to debate and different interpretations. The previous lexicographic research on Chinese collocation is mainly on observation based on the knowledge of linguistic experts. The lack of a large-scale collocation corpus with true collocations annotated, which we refer to as ***collocation bank***, become a barrier to accurately collect and analyze the characteristics of collocations for building automatic collocation extraction models.

In this paper, we present our work in the design and construction of a Chinese collocation bank. It the first attempt to provide a large scale and accurate collocation knowledge resource which can be shared for Chinese collocation research. To achieve such a goal, our work is divided into four steps. The first step is to define collocation properly as there is no readily accepted common definition. Therefore, we first review and analyze the definitions of 'collocation' adopted in lexicology and corpus linguistics, and we then propose a definition with good linguistic consistency and good computational operability. Since collocations cover a wide-range of lexical usage, they have various characteristics. Some are very rigid, whereas others are quite flexible. Thus, the second step is to classify collocations according to their different characteristics and features. Each classified collocation type is expected with a better inner linguistics consistency. Thirdly, the guideline of corpus annotations collocates is determined. It restricts the annotation to the bi-gram (two words) and n-gram (multi words) collocations identification. For each identified collocation, its classification and syntactic dependency information are also annotated in the corpus. Following this annotation guideline, a corpus with five million words are annotated for the 3,643 common headwords from "The Dictionary of Modern Chinese Collocations" [Mei 1999]. Through circles of three-passes-annotation and verification, all the word bi-gram collocations and n-gram collocations corresponding to 3,643 headwords are annotated. Finally, 23,581 identical collocations are identified and annotated. This collocation bank has proven to be a useful resource for collocation dictionary enrichment and for learning in automatic collocation extraction systems.

The rest of this paper is organized as follows. Section 2 introduces the basic concepts, definition and properties of collocation. Section 3 describes the collocation classification principles and rationales. Section 4 describes the collocation bank annotation guideline. Section 5 describes the implementations issues of collocation bank annotation. Section 6 is the conclusions.

2. Concepts

2.1. Definition of Collocation

Even though collocation occurs in text naturally and can be easily understood by human beings, it is difficult to define clearly [Benson 1989; Manning et al. 1999]. Previous work on collocation proposed various definitions and brought much confusion. Within the area of corpus linguistics, collocation was always defined as '*a pair of words which co-occur more often than would be expected by chance'* [Church et al. 1990] or '*habitual and recurrent word combination*' [Lin 1998]. Generally speaking, these definitions emphasize the statistical

significance of collocation, yet, true collocations still cannot be distinguished from free word combinations since the syntactic and semantic relations are not apparent. Lexicographers, on the other hand, emphasized semantic associations within collocations. [Benson 1990] defined collocation as '*an arbitrary and recurrent word combination*'. Some researchers considered that the whole meaning of a collocated pair must not be predicted from a single component, whereas others considered such a restriction is too rigid [Allerton 1984]. Meanwhile, these definitions are too loose and cannot be directly applied in automatic collocation extraction. Furthermore, some researchers only considered uninterrupted consecutive word sequences without the consideration for interrupted collocations.

In this study, collocation is defined based on [Manning 1999] and is stated as follows:

Definition 1: A ***collocation*** is a recurrent and conventional expression containing two or more content word combinations that hold syntactic and/or semantic relations. More specifically, content words in Chinese include noun, verb, adjective, adverb, determiner, and so on.

In this study, we focus on lexically restricted word combinations with emphasis on semantically and syntactically collocated word combinations, such as '浓 茶'(*strong tea*) '热烈 欢迎' (*warm welcome*). That means, collocation in this research is '*semantic collocation*', as proposed by [Mckeown et al. 2000]. As for the proposed '*grammatical collocation*' which often contains prepositions or auxiliary words (e.g. 'put on' and 'by accident' in English and '吃 着'(*eating*) and '在 路上' (*on the way* in Chinese) are not included in this study.

The co-occurrence of two or more collocated words can appear adjacently, referred to as **uninterrupted collocation**, or distant, referred as **interrupted collocation**. According to the number of components, collocations are divided into **bi-gram collocation**, which contains two words, and **n-gram collocation**, which contains more than two words.

2.2. Properties of Collocations

Firstly, according to definition, a collocation is recurrent and habitual use which means that collocations should occur frequently in similar context because they are of conventional use. Thus word combinations with certain occurrence frequency are often used as a feature for collocation extraction. As collocation is a kind of habitually use with no general syntactic or semantic rules to apply, they must appear in certain fixed patterns. For instance, we only say "浓 茶"(strong tea), but not "烈茶" (powerful tea). Yet, the choice of which completely seem arbitrary. No syntactic or semantic reason can be given for the particular choice of words.

Secondly, collocation is limited compositional. [Brundage 1992] introduced the term "**compositional**" to describe the property where the meaning of an expression can be predicted from the meaning of its components. Generally speaking, free word combinations can be generated by linguistic rules and the meaning is the combination meanings of its components. Collocation should be limited compositional. In other words, collocations are expected to have certain additional meanings i.e., the meaning cannot be derived directly from the meaning of its components. On the other hand, for those word combinations that have little additional meaning over the combination of words, they are also regarded as collocations if they show close semantic restriction between their components.

Thirdly, collocation is limitedly substitutable and limitedly modifiable. Limitedly substitutable here refers that a word cannot be replaced freely by its synonyms in a context. For example, "包袱" (*baggage*) and 行李" (*luggage*) have similar meanings in Chinese. However, when collocate with "历史"(*historical*), we will only say "历史 包袱" rather than "历史 行李". Also, collocations cannot be modified freely by adding modifiers or through grammatical transformations. For example, a collocation of "斗志 昂扬" does not allow further modification or insertion.

Lastly, collocation is domain-dependent [Smadja 1993]. In some specific domain, many collocations tend to be terms. Furthermore, some word combinations are regarded as collocations only in certain domains where they tend to be free combinations with high co-occurrences.

3. Classifying Collocation

Different collocations have different characteristics. Some are rigid and some are flexible. Previous work does not distinguish different kinds of collocations which leads to weak linguistic consistency for the whole collection of collocations. It is also difficult to distinguish collocations from non-collocated word combinations. Based on linguistic characteristics and the co-occurrence statistics of typical collocations, this section proposes a classification scheme to divide Chinese collocations into four types according to the combined characteristics in terms of compositionality, substitutability, modifiability, and internal association.

Type 0: Idiomatic Collocation

Type 0 collocations must have fixed forms where their components cannot be shifted around or added to or alter. Also, their components are non-substitutable allowing no syntactic transformation and no internal lexical variation. Type 0 collocations are non-compositional as its meaning cannot be predicted from the meanings of its component parts such as in 缘木求鱼 (to climb a tree to catch a fish literally which is a metaphor for a fruitless approach). Most Type 0 collocations are already listed as idioms in the dictionary. Some terminologies are also Type 0 collocations is they are not considered one word/phrase after segmentation. For example, the term 蓝/a 牙/n (*Blue-tooth*), refers to a wireless communication protocol which is completely different from either 蓝(*blue colour*) or 牙 (*tooth*).

Type 1: Fixed Collocation

Type 1 collocations also have fixed forms which is

non-substitutable and non-modifiable. For example, in a collocation 外交/n 豁免权/n (*diplomatic immunity*), there is no additional meaning carried by this collocation different from its two component words. However, none of these two words can be substituted by any other words to retain the same meaning. Since compositionality is a semantic characteristics and it is difficult to verify by computational methods using monolingual resources, the fixed forms, non-substitutable and non-modifiable characteristics become the main discriminative features of Type 1 collocations.

Type 2: Strong Collocation

Type 2 collocations allow limited modifier insertion while the order of components must be kept unchanged. Furthermore, it is very limited-substitutable where its components can be substituted by very few synonyms and the newly generated word combinations have nearly the same meaning. For example, 缔结/v 同盟/n and 缔结/v 联盟/n (*form alliance*).

Type 3: Loose Collocation

A Type 3 collocation has loose restrictions. It allows modifier insertion and component order alteration. Its components may be substituted by some of the synonyms and the newly generated word combinations which usually have the same meaning. This means that more substitutions of its components are allowed but the replacement is not free. Furthermore, Type 3 collocations must be statistically significant. Here are some examples: 合法/v 收入/n (*lawful income*), 正当/v 收入/n (*legitimate income*), and 合法/v 收益 (*lawful income*).

Table 1 summarizes the differences among the four types of collocations in terms of compositionality, substitutability, modifiability, and internal association.

	Type 0	Type 1	Type 2	Type 3
Compositional	No	Limited to yes	Yes	Yes
Synonym substitutable	No	No	Very limited	Limited
Order alter	No	No	Yes	Yes
Modifiable	No	No	Very Limited	Limited
Statistic significance	Not required	Not required	Required	Strongly Required

Table 1. Comparison of different types of collocations

Compared with the previous research, collocations defined in [Benson 1990] and [Brundage et al. 1992] correspond to Type 0 to Type 2 collocations. Many previous researches used very loose restrictions. Thus their extraction algorithms can extract some word combines that would not be consider collocation by our definition because they are freely replaceable word combinations grammatically fitting. [Church 1988] used the most relaxed notions of collocation where all the strongly co-occurred word pairs like “doctor-nurse” and “plane-airport” are extracted as collocations which are considered pseudo collocations in this research.

By classifying collocations into different types, collocations of each type are more coherent which is helpful to collocation identification. Study has shown that collocation extraction algorithms targeted at different types of collocations are more effective [Xu et al. 2005].

4. Annotation Guideline

The establishment of annotation guideline is the first step in corpus annotation. The guideline details the annotation strategy and annotation tasks. Firstly, the annotation of this collocation bank follows headword-driven strategy. The annotation use selected headwords as the starting point. In each annotation cycle, collocations corresponding to one headword are manually identified and annotated. This makes a more efficient annotation as it is helpful to estimate and compare related collocations corresponding to each headword. Secondly, the guideline specifies the information to be identified and labels used in the annotation. The following annotation tasks are to be carried out in the annotated Chinese collocation bank.

1. For a given headword, identify its corresponding bi-gram collocations and n-gram collocations. For example, for a given headword 合作/n (*co-operation*), both the bi-gram collocations such as 全面/a 合作/n (*all-rounded co-operation*) and n-gram collocations such as 国/j 共/j 合作/n (*co-operation between GMT and CCP*) are identified.

2. Annotate and verify the co-occurrence of each collocation in the corpus. Collocation is a kind of close combination and their components must co-occur within in a short context. However, not all of the co-occurrences of these components are indeed collocated. For example, in the context 加强/v 两国/n 之间/f 全面/a 合作/n (*enhance the all-rounded co-operation between two nations*), 全面/a and 合作/n is a collocation. However, in the context 全面/ad 发展/v 两国/n 友好/a 合作/n 关系/n (*all-rounded develop the friendly co-operation relationship between two nations*), 全面/a and 合作/n are no not collocated words whereas 全面/ad is collocated with 发展/v, and 合作/n is collocated with 友好/a. The annotation and verification of co-occurrence of collocations is useful for the learning of their characteristics for collocation extraction algorithms

3. For each bi-gram collocation, we need to decide and annotate its type as well as its syntactic dependency relations. Since n-gram collocations normally have complex structure and they are mostly in some fixed pattern, we will not provide additional information about them. As for bi-gram collocations, we label the collocations according to the type it belongs to. Furthermore, information on syntactic dependency within the bi-gram collocations is also annotated because some collocation extraction algorithms based on dependency relationships can make use of such information. For example, 全面/a 合作/n is categorized as a Type 3 collocation since it does not carry any additional meaning and thus it is compositional. It also allows modifier insertion. The components can be substituted by some of their synonyms such as “协作/n”. Its co-occurrence

frequency is significant. Furthermore, 合作/n serves as the head while 全面/a serves as a modifier, a syntactic information, PZAN (*a noun and its nominal modifier*), is also manually annotated.

In the current annotation guideline, some typical syntactic dependency relations are defined and listed as follows.

PZA- Noun and its adjective modifier. E.g. 合法/a 收入/n (*lawful incoming*) and 私有/a 财产/n (*private property*).

PZN- Noun and its nominal modifier. E.g. 人身/n 安全/n (*personal safety*) while 安全/n is the head, and 道德/n 标准/n (*moral standard*) where 标准/n serves as the head.

SBI- Predicate and its object in which the predicate serves as head. E.g. 转换/v 机制/n (*change the mechanism*) and 保护/v 文物/n (*protect culture relic*)

SBU- Predicate and its complement in which the predicate serves as head. E.g. 医治/v 无效/v (*ineffectively treat*)

ZZ - Predicate and its adverbial modifier in which the predicate serves as head. E.g. 沉重/ad 打击/v (*heavily strike*)

SD - Serial verb constructions which indicates that there are serial actions and the last action is the cardinal action, E.g. 跟踪/v 报导/v (*trace and report*)

ZW - Predicate and its subject. E.g. 财产/n 转移/v (*property transfer*)

AA – Adjective and its adverbial modifier. E.g. 极其/d 惨痛/a (*greatly painful*)

5. Annotation of the Collocation Bank

5.1. Corpus Data Preparation

The People's Daily 1998 corpus, a segmented corpus with part-of-speech tags by Peking University, is used to as the raw data to construct the collocation bank. It is claimed that the accuracy of word segmentation and POS tagging was higher than 99.9% and 99.5%, respectively [Yu et al. 2001]. With this popular and accurate resource we have significantly reduced the cost of annotation in our research, and ensuring the sharing of our output.

5.2. Headeword Set Preparation

As discussed in Section 4, the annotation is headword driven. Thus, a headword set must first be prepared. Based on the linguistic resource, "The Dictionary of Modern Chinese Collocation" [Mei 1999] 3,643 headwords are selected to form the headword set. The selection is based both on the judgment of linguistic expert as well as statistical information that they are commonly used

5.3. Corpus Annotatoin

The annotation is conducted for each headword. The annotators are required to identify whether a co-occurred word combination is a collocation and the type of collocation. The annotation procedure, however, requires three passes. We use the headword 安全/an, as an example to illustrate the annotation procedure.

Pass 1. Concordance and dependent word identification

In the first pass, the concordance of the given headword is performed on the corpus. Sentences containing the headwords are obtains such as some listed below:

遵循/v 确保/v 安全/an 的/u 原则/n
确保/v 人民/n 群众/n 的/u 生命/n 财产/n 安全/an
确保/v 长江/ns 安全/an 度汛/v
... ...

The annotators then manually identifies all syntactically and semantically dependent words surrounding the observing headword 安全/an in its context. XML tags are used for the annotation as presented below.

```
<p> 遵循/v 确保/v 安全/an 的/u 原则/n</p>
<depend search= "安全/an "head="确保/v" depend ="安全/an" relation="SBI" ></depend>

<p>确保/v 人民/n 群众/n 的/u 生命/n 财产/n 安全/an</p>
<depend search= "安全/an "head="确保/v" depend ="安全/an" relation ="SBI" ></depend>
<depend search= "安全/an "head=" 安全/an" depend ="生命/n" relation ="PZN" ></depend>
<depend search= "安全/an "head=" 安全/an" depend ="财产/n" relation ="PZN" ></depend>

<p>确保/v 长江/ns 安全/an 度汛/v </p>
<depend search= "安全/an "head=" 度汛/v" depend =" 安全/an" relation ="ZZ" ></depend>
```

The dependency word combination is annotated with the help of the tag <depend> and several attributes gives its information:

-search indicates the current observing headword

-head indicates the head of the identified word dependency pair.

-depend indicates the dependent word (or called modifiers).

-relation gives the syntactic dependency relations labeled according to the dependency labels in the annotation guideline.

In the first two example sentences, the word combination 确保/v 安全/an has the dependency relationship (and are also collocated although this is identified in this pass), whereas in the third sentence, such dependency does not exist as 确保/v only determines 度汛/v.

Pass 2. n-gram collocations annotation

In the second pass, the annotators focus on the sentences where the headword has more than one dependent words. With the help of a simple statistical program [Xu et al. 2003], the word combinations which frequently co-occur in consecutive positions and in fixed order are extracted as n-gram collocations. No further analysis on the internal syntactic and semantic information of n-gram collocations is conducted. For the given headword, a

n-gram collocation “生命/n 财产/n 安全/an” is identified, and it is annotated as follows:

<ncolloc search=“安全/an ” w1=” 生命/n” w2=” 财产/n” w3=”安全/an”></ncolloc>

where,

-<**ncolloc**> indicates n-gram collocation

-***w1, w2,..wn*** gives the components of n-gram collocation according to the ordinal sequence.

Pass 3. bi-gram collocations annotation

In the third pass, all two-word combinations are examined to identify bi-gram collocations. If a dependent word combination is regarded as a collocation by the annotators, it is further labeled according to the type determined. The determination is examined based on the human knowledge and the use of several computational features [Xu et al. 2005] listed below:

Strenghth: Reflects the co-occurrence frequency significance of a collocation candidate among all of the word combinations with same headword

Spread: Reflects the co-occurrence distribution significance of a collocation candidate

Synonym Substitution Rate: Measures the substitutability of a collocation candidate

Distribution Similarity: Measures the distribution similarity between a collocation candidate and the statistically expected distribution.

Since these features are computational ones, a tool is developed to calculate each of them. We use the same program to two sets of statistic data sets. The first one is obtained from the dependency annotated 5-million-word corpus in **Pass 1**. Because the dependent word combination are manually identified and annotated in the first pass, the accurate statistical significance is useful to help identify whether it is a collocation and the type of collocation. However, data sparseness problem must be considered since 5-million-word is not large enough for collocation analysis. Thus, another set of statistical data are collected from a 100- million-word segmented and tagged corpus. Although data sparseness is no longer a serious problem for this set of data, the collected statistics are noise-prone since they are directly retrieved from text without any verification. By analyzing the statistical features coming from two sets of data and expert judgments, the annotators determines the collocation type. Annotation results of **Pass 3** for the example sentences are given below:

<bcolloc search=”安全/an” col=”确保/v” head= “确保/v” type=”2” relation=”SBI”></bcolloc>

<bcolloc search=”安全/an” col=”生命/n” head= “安全/an” type=”3” relation=”PZN”></bcolloc>

<bcolloc search=”安全/an” col=” 度汛/v” head= “度汛/v” type=”3” relation=”ZZ”></bcolloc>

where,

-<**bcolloc**> indicates a bi-gram collocation.

-col is for the collocated word.

-head indicates the head of an identified word dependency pair.

-type is the identified type.

-relation gives the syntactic dependency relations.

Even though the word pair 财产/n 安全/n in the second example sentence hold the dependency relation, they are not regarded as a collocation since it does not fulfill the statistical requirement. Therefore, it is not reserved in the annotation result.

5.4. Quality Assurance

The annotators of this work are three post-graduate students majoring in linguistics. 20% of the whole collocation bank annotation was annotated in duplicates by all three of them. Their outputs were checked by a program. Any annotated collocations and classified types that were accepted by either two or three annotators are reserved in the final data while the others are considered incorrect. The inconsistencies between different annotators were discussed to clear any misunderstanding the most appropriate annotation results and to help The remaining data were then divided into three parts and annotated separately.

5.5. Current Status

From the collocation bank, we have obtained 23,581 bi-gram collocations for the 3,643 headwords discussed in Section 5.2. We call this collection of bi-gram collocations as the *PolyU Collocation Collection*(*PCC*). For the same 3,643 headwords, The Dictionary of Modern Chinese Collocation [Mei 1999] provided 35,742 typical collocations, which we call *Mei's Collocation Collection*(*MCC*).. There are 19,967 common entries in *PCC* and *MCC*, which means that 84.7% collocations in *PCC* appear in *MCC indicating* good linguistic consistency between *PCC* and *MCC*. Furthermore, we obtained 3,614 collocations that are not recorded *MCC*. This means that the collocation bank provides some collocations that are dynamically obtained from corpus data which can be used to enrich a static collocation dictionary.

In a previous work the collocations in *MCC* were manually categorized into four types based on expert knowledge with reference to the use of statistical data from a 100-million-word corpus[ref]. Using the type information in *PCC* as the standard answer, we find that within the 19,967 common collocations, the classification accuracy provided in [Xu et al. 2005] for *MCC* are 94.1%, 91.5% and 89.4%, for Type0/Type 1, Type2 and Type 3 collocations, respectively, even though they are also manually categorized. The difference in accuracy is mainly because the classification of [Xu et al. 2005] was done based on statistical information which can be incorrect because some co-occurrence data may not be collocations. Statistical information collected from collocation bank is more accurate and can be more useful to linguistic research, and it is essential to improve the automatic collocation extraction systems.

6. Conclusions

In this study, a comprehensive definition of collocation and their classifications are proposed with good

operability. Based on a collocation annotation guideline, the collocations corresponding to 3,643 headwords are identified from a 5-million-word corpus and each occurrence is annotated and verified to construct a collocation bank. A total of 3,614 new collocations are identified compared to a published collocation dictionary. Furthermore, the annotated collocation corpus can be used to help analyzing the characteristics for different types of collocations, and to help selecting the discriminative features for automatic collocation extraction algorithms. Finally, the obtained collocation collection may be used as a standard answer set for evaluating the performance of the collocation extraction algorithms. In the future, collocations of all the unvisited headwords will be annotated to produce a complete 5-million-word Chinese collocation bank.

7. Acknowledgements

This research is supported by Hong Kong Polytechnic University (Project Code A-P203) and a CERG Grant (Project code 5087/01E). We also thank the valuable comments and suggestions from the anonymous reviewers.

8. References

Allerton D. J. (1984) Three or four levels of co-occurrence relations. *Linguistics*. no. 63, pp. 17-40

Benson M. (1989). The structure of the collocation dictionary. *International Journal of Lexicography*, vol.2, no.1, pp. 1-14

Benson M. (1990). Collocations and general-purpose dictionaries. *International Journal of Lexicography*, vol. 3, no.1, pp. 23-35

Brundage et al. (1992) Multiword lexemes: a monolingual and contrastive typology for natural language processing and machine translation, *Technical Report 232 -IBM*

Church K. et al. (1990). Word association norms, mutual information and lexicography. *Computational Linguistics*, vol.16, no.1, pp. 22-29

Church K. W. (1988). A stochastic parts program and noun phrase parser for unrestricted text, In *Proceedings of the Second Conference on Applied Natural Language Processing*, pp. 136-143

Gitsaki C. et al. (2000). *English Collocations and Their Place in the EFL Classroom*, pp. 121-126

Lin D.K. (1998). Extracting collocations from text corpora. In *Proceedings of First Workshop on Computational Terminology*, Montreal

Manning C. D. et al. (1999) *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts, 1999

Mckeown R. et al. (2000). Collocations, in *A Handbook of Natural Language Processing* (Dale, Robert et al. eds.), Marcel Dekker

Mei J.J. (1999). Mei, J. J., *Dictionary of Modern Chinese Collocations*, Hanyu Dictionary Press

Mitra et al. (1997). An analysis of statistical and syntactic phrases. In *Proceedings of RIAO*, pp. 200-214

Sinclair J. (1991). *Corpus, Concordance, Collocation*. Oxford University Press

Smadja. F. (1993). Retrieving collocations from text: *Xtract*, *Computational Linguistics*, vol. 19, no. 1, pp. 143-177

Xu R. F. et al. (2003), An automatic Chinese Collocation Extraction Algorithm based on Lexical Statistics, in *Proceedings of IEEE International Conference on Natural Language Processing and Knowledge Engineering*, pp.321-326, Beijing, China

Xu R. F. et al. (2005) A Multi-stage Chinese Collocation Extraction System, in *ICMLC 2005, LNAI 3930*, (Yeung D. S. eds.) Springer-Verlag Berlin Heidelberg: pp.740-749

Yu S. W. et al. (2001). *Guideline of People's Daily Corpus Annotation, Technical Report*, Peking University