Collocation and Trillocation[1]

Shaokang Qin, Hui Wang
Department of Chinese Studies, National University of Singapore, Singapore 117570
{qinshaokang, chswh}@nus.edu.sg

Abstract

In this paper we proposed that the neglected three words collocations (trillocation) should be emphasized in collocation study. From the point of view of colligations, more useful collocations could be covered by adding a third category. For a specific third word, it will help avoid the unnaturalness of a two words collocation. A statistic based automatic trillocation extracting system is proposed and achieves 45% correctness of trillocations. The system first extract two words collocations based on a statistic measure together with grammatical information and then extract the collocations (the third words) of the acquired two words collocations. Human experts also judged the ambiguity and naturalness of the collocations and the correspondent trillocations. The results show that the extracted trillocations could be utilized to help disambiguate word senses and generate natural language output.

Keywords:

Collocation; Colligation; Trillocation; Sense Disambiguation; Language Naturalness;

1. Introduction

Collocations are widely employed in natural language semantic processing researches, like word sense disambiguation, mainly because they nearly are the one and only information that indicates the meaning of a word. Because of the complexity of the natural sentences, in many cases, a third word, or even more words are involved in the

[1] The research was supported by Academic Research Fund (AcRF) Tier 1, Ministry of Education, Singapore (Grant No. R-102-000-046-112).

sentence generation and the meaning output. The paper suggests that in some cases a third word should be involved into the collocation and they even function as importantly, if not more than the first two words.

Although collocations of more words are more or less mentioned in the literature, it is no doubt that most studies focus on two words collocations and three or more words collocations are more or less neglected(Firth,1968; Halliday,1966; Aisenstadt,1981; Howarth,1998). The definitions of collocation generally do not exclude collocations of more words and can be categorized by the way people identify collocations: by statistics, by the word combination itself (Nesselhauf,2005). For the first way, Sinclair(1991) defines collocations as "*the occurrence of two or more words within a short space of each other in a text*"; for the second, Cowie(1981) considers that collocations are located somewhere in between of a continuum which the two ends are pure idioms and free combination (Cowie & Keith,2006). But Sinclair's collocation dictionary, *Collins COBUILD English collocations on CD-ROM*(CCEC, (Sinclair,1995)), only two words collocations are included. When A. P. Cowie suggested restricted collocations, like *perform a task*, actually his standard of "*the possibility of internal variation, or substitution of part for part*" (Cowie, Mackin, & McCaig,1975) were employed on his two words combinations. The focus on two words collocations can even traced back to Halliday(1966), and even Firth(1968), from Halliday's *strong tea* and *powerful car* to Firth's *silly ass*.

In natural language processing, most systems only employ the two words collocations((Yarowsky,1995) for English and (Ge & Li,2001) for Chinese) or even simple bigrams. Many collocation extracting methods only extract two words collocations (Smadja & McKeown,1990; Sun, Huang, & Fang,1997). In dictionary compilation, besides the above mentioned Sinclair(1995), there are some other influential dictionaries, *Oxford Collocations Dictionary for Students of English*(OCDSE, (Crowther, Dignen, & Lea,2002)), *The BBI combinatory dictionary of English*(BBI, (Benson,1986)) for English collocations, and *Xiandai Hanyu Shici Dapei Cidian*(XHSDC, (Zhang & Lin,1992)), *Xiandai Hanyu Dapei Cidian*(XHDC, (Mei,1999)) for Chinese collocations, which focus on two words collocations.

Undoubtedly, two words collocations are quite useful in natural language processing and also difficult to learn for foreign language learners. But is a third word not so important as the first two words? This paper only discusses the case of three words and, to make it simple, trillocation is used to refer to it hereafter.

2. Trillocation: The Third Class and Sense Disambiguation

The concept of colligation is developed in (Firth,1968), which can be simply explained

as "*the habitual co-occurrence of grammatical elements*" (Krishnamurthy & Keith,2006). The limited number of part-of-speech colligations are widely used to categorize the unlimited collocations (besides the following examples, see also (Aisenstadt,1981; Howarth,1998; Kjellmer,1994)). But some collocations could not be covered in the two-category system, while trillocations could imply them and their relation.

One of the reasons that two words collocations are emphasized is the dictionaries and scholars like to categorize the collocations into different groups by part-of-speech of the component words. For the above-mentioned dictionaries, except CCEC and XHDC, part-of-speech is used as the important information in the dictionary entries. The following is their categorizations (noun, verb and adjective only):

	OCDSE	BBI	XHSDC
noun	adj+n; q+n; v+n; n+v; n+n; prep+n; n+prep;	v+n; n+v; n1 of n2; adj+n;	n+v; n+adj; v+n; n+n; adj+n; mq+n
verb	v+n; n+v; adv+v; v+v; v+prep; v+adj;	v+n; n+v; v+pron; v+prep clause; v+adv;	v+v; v+adj; n+v; adj+v; v+n; v+mq; modal verv+v;
adjective	adj+n; v+adj; adj+prep;	adj+n; adj+adv;	v+adj; n+adj; adj+adj; adj+mq; modal v+adj; adj+n; adj+v;

Table 1. Part-of-speech Colligations in Three Dictionaries

The categorizations look perfect, and for English, the user can even easily create one to one correspondence between part-of-speech and different grammatical units. But there are two points worth questioning: 1) the lack of colligations of more component categories; 2) the missing of some underlying collocations. It is easy for us to find no colligations of more than two categories in Table 1. But there are some underlying collocations missing in this part-of-speech system.

Here is an example. For English collocations, BBI does not have a group of *n+n*. It seems quite good for XHSDC to have that group. But their *n+n* mainly means the first noun modifies the second one, for example, *问题/n 的/u 答案/n*. The structure of *n+n* in Chinese of course bears more relationships. Being one of the basic clause types, S+V+O is soundly to be considered as containing *n+v+n* type[2]. Let's take the

[2] Guo(2002)'s study shows that 99.8% of verbs in Chinese can be a predicate, 97% and 97.6% of

opinion that *n+v+n* is not suitable to be a colligation. Then *n+v+n* could be separated into two collocations sharing one verb: S+V and V+O. Reasonably we can ask whether there is certain relationship between S and O, which could well be *n* and *n*? The following example is from our corpus, *人类/n(human beings) 利用/v(to make use of) 海洋/n(ocean)*. Both*人类/n 利用/v* and*利用/v 海洋/n* can be collocations, while *人类/n* and*海洋/n* are not a collocation. In each collocation, *利用/v* is ambiguous even by human being if only the other word is provided. But if three words are provided simultaneously, *利用/v* can only mean to make use of (scrap material) rather than to exploit.

Even if we take the first noun *人类/n* and the second noun *海洋/n* as one collocation, there will be no place for them in the existing two-category system because one word is not modifying the other one. Therefore, if two-category colligations could not be enlarged to contain more collocations, three or more categories colligations should be accepted as one part in collocation study.

3. Trillocation: The Third Word and Naturalness

If we examine these dictionaries, many examples consist of more words to make them natural. *OCDSE* compilers actually provided the third part information in the collocate words and examples:

SCIENTIST + VERB **report sth**

VERB + FIND **report** *I reported my find to the landowner.*

Similarly, in *XHSDC*, the editors provide *利用（电脑）设备* rather than *利用 设备*. The arrangement of the entry shows the importance of the third part in a two word collocation. The problem how to use the correct third part may sometimes be solved by using grammatical knowledge. But this is not always the case in every collocation in every language because collocation might be unnatural in language output. Trillocation could help solve the problem.

Sometimes, two words are a collocation and could be extracted by the automatic extracting system. But the fact is that they are seldom used by themselves and the third part could not be added by grammatical rules either. The following is an example from Chinese:

1) *打/v 精神/n*

raise spirit

cheer up

2) *打/v 起/v 精神/n*

---

noun for subject and object respectively.

raise up spirit

cheer up

3) *强/a 打/v 精神/n*

forced raise spirit

cheer up

In the first example, *打/v* and *精神/n* are the two words in one collocation. When the two words are searched in the corpus we used (see 4.2), the only one hit is the second one. Among the first 20 hits of "*打精神*" in google.cn on August 11, 2008, there are 15 "*强打精神*" and all the others are those in which *打* and *精神* are not a collocation, like *打精神战*. In fact, if the combination of the two words is used by themselves, native speakers will feel it awkward or not acceptable. We could not ascribe the reason to grammar either because the two words used here *起/v* and *强/a* cannot be changed to any other verbs or adjectives. For these two points, *打/v* and *精神/n* should be one sort of specific collocation and *2)* and *3)* should be trillocations.

If trillocation could not be accepted, there is one alternative. We can take *打/v* and *精神/n* as collocation, then a third word could collocate with the collocation. (Hoey,2005) uses psychological term *priming* to explain collocations. Priming can be understood "As a word is acquired through encounters with it in speech and writing, it becomes cumulatively loaded with the contexts and co-texts in which it is encountered, and our knowledge of it includes the fact that it co-occurs with certain other words in certain kinds of context." The author thinks that one word can collocate with a second word, and then the collocation can have its own collocations which may not collocate with any word in the first two words collocation.

Whether it is trillocation or collocation of collocation, it is true that sometimes two words collocations are unnatural and the third word helps the naturalness of the sentence. In these cases, three words combinations have a stronger relation than any two words combined and have the more chance to be regarded as one part.

4. Extracting Collocations and Trillocations

As trillocation has not been emphasized, there are few projects on automatic Chinese trillocation extracting. We developed an automatic trillocation extracting system mainly based on statistics. We tested the system for three ambiguous words, *打/v, 培养/v* and *利用/v*, on the Textbook Corpus and analyzed the result.

4.1 Method

The collocation extracting system we developed is mainly on statistics and combined with some grammatical rules. The process of extracting was divided into two steps, extracting collocates of the node word (collocations acquired) first and then extracting collocates of the acquired collocations (trillocations acquired).

For the first step, there are many statistical measures available, for example, t-score, mutual information (MI), chi-square, fisher exact test (Manning & Schèutze,1999). All of them hold their own advantages and disadvantages. MI (Church, Gale, Hanks, & Hindle,1991) is a widely used measure in collocation extracting system (Church, et al.,1991; Sun, et al.,1997) and in dictionary compiling (Baugh, Harley, & Jellis,1996). Compared with MI, which gives too much weight to rare events, MI3 gives more weight to the frequent events(Oakes,1998). This is the statistical measure we used in the test to extract collocation candidates. The formula for calculating MI3 is:

$$\log_2 \frac{f_{AB}{}^3 N}{f_A f_B}$$

In this step, the system also avails itself of certain grammatical information of specific Chinese words occurred in the collocations. Many scholars agree that the words in collocations have some grammatical relations[3]. For each node and its candidate collocates, the system will look up the correspondent grammatical information to decide whether they are a collocation. We only use the information to reject a collocation. For example, there is the co-occurrence of *培养/v 激发/v* in the corpus with a high MI3 score. But the grammatical information that *培养/v* requires a nominal object gives the system more weight to reject them as collocation.

For the second step, we regard the acquired collocations as the node words and apply MI3 to extract the third word collocates in the same span.

The reason that only nouns, verbs and adjectives are chosen is that these three categories are the most common ones in Chinese. Most verbs, nouns and adjectives in Chinese can function as subject, predicate and object except nouns as predicates. This makes these three categories hard to process by machine or human being.

[3] Besides the above-mentioned part-of-speech colligations, Benson(1986)even divided collocations into grammatical collocations and lexical collocations.

4.2 Corpus Resource

The corpus we use is Textbook Corpus, a sub-corpus of Singapore Chinese Corpus, developed by Lexical Semantics and Computing Group, NUS. The Textbook Corpus consists of five versions of texts in primary school Chinese textbooks and three versions of texts in middle school Chinese textbooks. The texts range from the very beginning of Chinese to essays of more than 2500 words, which could be regarded as the fundamental Chinese for reading and writing.

The corpus has been word segmented and part-of-speech tagged and contains 1,370,095 words. The word segmentation and part-of-speech tags were proofread by human experts. Part-of-speech tags are the important grammatical information we adopted, especially when the word is of two or more part-of-speeches.

4.3 Result and Analysis

Three polysemous verbs are selected as the first words. Among them *打/v* is a most frequently occurred word , *利用/v* not so frequent as *打/v* and *培养/v* not frequent in the corpus. Table 2 shows five collocations with top MI3 scores of *利用/v* and their two trillocations with top MI3 scores.

Collocation	Co-occurrence	MI3	Trillocation	Co-occurrence	MI3
利用/v 开发/v	8	11.79	海洋/n	6	12.25
			资源/n	4	12.07
利用/v 课余/n	4	11.74	时间/n	4	10.48
利用/v 时间/n	16	11.08	课余/n	4	13.92
			走遍/v	2	10.74
利用/v 资源/n	7	10.88	开发/v	4	12.72
			海洋/n	3	10.31
利用/v 发电/v	4	10.52	来/v	4	8.49
			可以/v	2	7.20

Table 2. Results of Collocations and Trillocations

For ambiguity in a collocation/trillocation, we mean that human being could not judge the meaning of the node word by only knowing the collocate word(s). Naturalness means that human being feels that the collocation is not just an occasional co-occurrence but meaningful and can be used as a part of a sentence. We define the correct trillocations as those natural and of no ambiguity for each word. We tested the correctness of the automatically extracted trillocations with the help of human experts who checked whether the trillocation are correct ones. Because the third word may well be one collocate of the node word, there will be repeated trillocations. We removed all repeated trillocations here. The result is as the following Table 3.

	打/v	利用/v	培养/v	total
Correct	65	23	2	90
Total	146	38	14	198
Percentage of Correctness				45%

Table 3. Correctness of Trillocations

There are several cases for the other 55% trillocations: 1) The original collocation is natural and of no ambiguity and the third word may occurs in the context only by chance, for example, *打/v 人/n 压迫/v.* In the original corpus, it reads like *用/v 枪/n 来/v 打/v 这些/r 受/v 人/n 压迫/v*. 2) Some trillocations are still not natural, Chinese speakers do not use them as is, for example *打/v 烙印/n 脑海/n*. 3) Some third words collocate much more with the other words in the context rather than with the collocation. For example, *走遍/v* in *利用/v 时间/n 走遍/v*. 4) some node words could not be disambiguated in the trillocation, *利用/v* in *减轻/v 来/v 利用/v* for example. Part of the reason for the result could be the small size of the corpus used, although data sparseness could not be avoided in any corpus.

We also checked all the collocations and trillocations by hand to see how many collocations which are unnatural or of ambiguous node words (the first word) become natural and of no ambiguous node words. In order to see how well the third word functions, we did not remove the repeated trillocations in this test. The following Table4 is the result.

There are two numbers in No ambiguity and Natural columns, 19/68 in *打/v* row for example. This means that the 15 occurrences of *打/v* of ambiguity in collocations appears in 68 trillocations, and 19 of them could be disambiguated. Here is an example:

	Ambiguity	Unnatural	No ambiguity	Natural
打/v	15	22	19/68	22/73
利用/v	8	4	14/29	3/5
培养/v	3	2	4/6	1/4
total	26	28	37/103(36%)	26/82(32%)

Table 4. Ambiguity and Naturalness Comparison on Collocations and Trillocations

4) *打/v 用/v*

to strike to use

5) *用/v 鞭子/n 打/v*

to use whip to strike

to strike with a whip

用/v is one collocate of *打/v* by MI3, but the meaning of *打/v* could not be determined only by *用/v* as in *4)*. We then consider *打/v* as ambiguous in this collocation. Because *打/v 用/v* could not be used as is and bears no sense by themselves, we then regard this collocation as unnatural. Then we check all the trillocations derived from *打/v 用/v*. *鞭子/n* is one of the collocates of *打/v 用/v* and determines the meaning of *打/v* and *5)用/v 鞭子/n 打/v* is also natural in Chinese. By far, we identified *鞭子/n* as the third word that helps disambiguate *打/v* in *打/v 用/v* and make *打/v 用/v* natural. The above result shows that a third word has the great potential to improve the output of applications which only employ two words collocations.

5. Conclusion

Trillocations, or three word collocations, have the great potential in collocation research and applications. Firstly, compared with the two categories colligation system, three categories colligation system contains some useful relationships which are not covered by the former. This could be employed to disambiguate word sense. Secondly, the third word improves the naturalness of some two words collocations which might never be used as is. Our test proves the effects of the third word in trillocations and its potential usage in natural language processing applications.

We analyzed the necessity to extract not only collocations but also trillocations. A system which is based on statistics was also developed to automatic extract trillocations and analysis on the result of 45% correctness was also given. We also performed a test to present the potential a third word could add to a collocation.

This paper only suggests initial thought on collocations of more than two words and there is still large room to improve in the test. Further research should be on 1) thorough study on the inner relation between and among the three components of trillocations; 2) the improvement of the automatic collocation extracting system by using a larger corpus, like the whole SCC, and employing the recent combined statistic measures.

References

Aisenstadt, E. 1981. *Restricted collocations in English lexicology and lexicography.* ITL: Review of Applied Linguistics, 53, 53-61.

Baugh, S., Harley, A., & Jellis, S. 1996. *The Role of Corpora in Compiling* the Cambridge International Dictionary of English. International Journal of Corpus Linguistics, (11), 39-59.

Benson, M. (Ed.) 1986 *The BBI combinatory dictionary of English*. Amsterdam: John Benjamins Publishing Co.

Church, K., Gale, W., Hanks, P., & Hindle, D. 1991. *Using Statistics in Lexical Analysis.* In U. Zernik (Ed.), Lexical acquisition : exploiting on-line resources to build a lexicon (pp. 115-164). Hillsdale, N.J. : Lawrence Erlbaum Associates.

Cowie, A. P. 1981. *The Treatment of Collocations and Idioms in Learners' Dictionaries* Applied Linguistics II(4), 223-235.

Cowie, A. P., & Keith, B. 2006. *Phraseology Encyclopedia of Language & Linguistics* (pp. 579-585). Oxford: Elsevier.

Cowie, A. P., Mackin, R., & McCaig, I. R. 1975. *Oxford dictionary of current idiomatic English.* London: Oxford University Press.

Crowther, J., Dignen, S., & Lea, D. (Eds.). 2002 *Oxford Collocations Dictionary for Students of English*. Oxford: Oxford University Press.

Firth, J. R. 1968. *A Synopsis of linguistic theory,* 1930-55. In F. R. Palmer (Ed.), Selected Papers of J. R. Firth 1952-59 (pp. x, 209 p.). London and Harlow: Longmans, Green and Co Ltd.

Ge, R., & Li, J. 2001. *Design and Implementation of an Automatic System of W ord Sense Tagging.* Computer Engineering and Applications 37(17), 170-173.

Guo, R. 2002. *On Part of Speech of Modern Chinese Vocabulary.* Beijing: Commercial Press.

Halliday, M. A. K. 1966. *Lexis as a linguistic level.* In C. E. Bazell, J. C. Catford, M. A. K. Halliday & R. H. Robins (Eds.), In Memory of J. R. Firth (pp. 148-162). London: Lowe & Brydone (Printers) Ltd., .

Hoey, M. 2005. *Lexical priming : a new theory of words and language.* New York: Routledge.

Howarth, P. 1998. *The Phraseology of Learners' Academic Writing.* In A. P. Cowie (Ed.), Phraseology : theory, analysis, and applications (pp. 161-188). Oxford [England]; New York: Clarendon Press ;Oxford University Press.

Kjellmer, G. 1994. *A dictionary of English collocations : based on the Brown corpus : in three volumes.* Oxford; New York: Clarendon Press: Oxford University Press.

Krishnamurthy, R., & Keith, B. 2006. *Collocations Encyclopedia of Language & Linguistics* (pp. 596-600). Oxford: Elsevier.

Manning, C. D., & Schèutze, H. 1999. *Foundations of statistical natural language processing*. Cambridge, Mass.: MIT Press.

Mei, J. (Ed.) 1999 *Modern Chinese Collocation Dictionary* (ed.). Shanghai: Publishing House of an Unabridged Chinese Dictionary.

Nesselhauf, N. 2005. *Collocations in a learner corpus.* Amsterdam ; Philadelphia: J. Benjamins Pub. Co.

Oakes, M. P. 1998. *Statistics for Corpus Linguistics.* Edinburgh: Edinburgh University Press.

Sinclair 1995. *Collins COBUILD English collocations* on CD-ROM: Harper Collins.

Sinclair, J. 1991. *Corpus, concordance, collocation.* Oxford: Oxford University Press.

Smadja, F., & McKeown, K. 1990. *Automatically extracting and representing collocations for language generation* In Proceddings of the 28th Annual Meeting of the Association for Computational Linguistics. Pittsburgh.

Sun, M., Huang, C., & Fang, J. 1997. *Quantitative analysis of Chinese collocation.* Journal of Chinese Language, 256(1), 29-38.

Yarowsky, D. 1995. *Unsupervised word sense disambiguation rivaling supervised methods.* Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, 189–196.

Zhang, S., & Lin, X. 1992. *Collocation Dictionary of Modern Chinese Lexical Words*: Business Publisher, China.