The Construction of a Chinese Collocational Knowledge Resource and Its Application for Second Language Acquisition

Renfen Hu
Institute of Chinese Information Processing, Beijing Normal University
19 Xinjiekouwai Street, Beijing, China
irishere@mail.bnu.edu.cn

Jiayong Chen and Kuang-hua Chen
Department of Library and Information Science, National Taiwan University
No.1, Sec. 4, Roosevelt Road, Taipei, Taiwan
{d04126001,khchen}@ntu.edu.tw

Abstract

The appropriate use of collocations is a challenge for second language acquisition. However, high quality and easily accessible Chinese collocation resources are not available for both teachers and students. This paper presents the design and construction of a large scale resource of Chinese collocational knowledge, and a web-based application (OCCA, Online Chinese Collocation Assistant) which offers free and convenient collocation search service to end users. We define and classify collocations based on practical language acquisition needs and utilize a syntax based method to extract nine types of collocations. Totally 37 extraction rules are compiled with word, POS and dependency relation features, 1,750,000 collocations are extracted from a corpus for L2 learning and complementary Wikipedia data, and OCCA is implemented based on these extracted collocations. By comparing OCCA with two traditional collocation dictionaries, we find OCCA has higher entry coverage and collocation quantity, and our method achieves quite low error rate at less than 5%. We also discuss how to apply collocational knowledge to grammatical error detection and demonstrate comparable performance to the best results in 2015 NLP-TEA CGED shared task. The preliminary experiment shows that the collocation knowledge is helpful in detecting all the four types of grammatical errors.

1 Introduction

Second language (L2) learners often have problems in appropriate use of word collocations (Bahns and Eldaw,1993; Farghal and Obiedat, 1995; Nesselhauf, 2003). For example, English speakers who are learning Chinese are likely to produce incorrect expressions e.g. 漂亮 的 歌 (piaoliang/beautiful de/* ge/song) because "beautiful song" is a reasonable collocation in English. However, 漂亮 (piaoliang/beautiful) is not an acceptable modifier for 歌 (ge/song) in Chinese (Lu, 1987). Bahns and Eldaw (1993) went deeply into this issue and suggested that leaner's collocational competence does not develop in parallel with general vocabulary knowledge, because words are usually acquired individually without taking note of their immediate environment (Siyanova and Schmitt, 2008).

This problem could be more serious for learning analytic languages such as Chinese, because apart from the negative transfer from the first language (L1), this typical analytic language does not have inflectional morphemes, and uses function words with little lexical meaning to express grammatical relationships. These function words are also highly involved in collocations. The following shows collocation examples with function words:

(a) 在 (zai/*) X 上 (shang/on) → on X
(b) 诚实 (chengshi/honest) 的(de/*) 人 (ren/person) → honest person
(c1) 买 (mai/buy) 东西 (dongxi/something) → buy something
(c2) 买 (mai/buy) 了(le/*) 东西 (dongxi/something) → have bought something
(c3) 买 (mai/buy) 着 (zhe/*) 东西 (dongxi/something) → be buying something

Example (a) is a two-word collocation in which the preposition 在 (zai) collocates with the postposition 上 (shang). The auxiliary word 的 (de) in Example (b) connects an attributive and a head noun. By comparing (c1), (c2) and (c3), the auxiliary words 了 (le) and 着 (zhe) are used to indicate perfective and progressive aspects. These examples illustrate at least three features of Chinese function words: (1) directly constituting specific collocations, e.g. a preposition and a postposition; (2) acting

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers,
pages 3254–3263, Osaka, Japan, December 11-17 2016.

as an auxiliary structural particle in a collocation, e.g. 的 (de); (3) indicating the aspect, e.g. 了 (le) and 着 (zhe). Therefore, teaching and learning Chinese collocations become a considerable challenge.

In order to offer informative reference of collocational knowledge to L2 teachers and learners, this paper discusses the construction of a Chinese collocational knowledge resource and its application for second language acquisition. We define nine types of collocations reflecting Chinese specific grammatical features, and automatically extract collocations from two source corpora: CTC[1], a Chinese text corpus for L2 learners (Yang and Xiao, 2015) and Simplified Chinese Wikipedia Corpus[2]. Sentences in the corpora are firstly processed by LTP-Cloud (Che et al., 2010), a Chinese NLP toolkit to do word segmentation, POS tagging and dependency parsing. Then we compile 37 rules based on word, POS and dependency relation features to extract the target word combinations. After building a collocation database containing over 1,750,000 collocations and the corresponding attributes, we design and construct the OCCA (Online Chinese Collocation Assistant) to offer free and convenient collocation search service to users. To examine the coverage, accuracy and efficiency of our collocation data, we compare it with two traditional collocation dictionaries and apply it to the grammatical error detection task. The experiment shows desirable results from both the comparison and the application.

The rest of the paper is structured as follows. In section 2 we review existing Chinese collocation resources and the automatic extraction methods. Section 3 discusses our definition and operational types for Chinese collocations. Section 4 describes the approach of extracting collocations from text corpora and building the Online Chinese Collocation Assistant. The comparisons and applications of our approach are discussed in Section 5 and 6. We draw brief conclusions in the last section based on the experimental results.

2 Related Work

The definition of collocation varies in previous studies, but most of them emphasize the importance of statistical frequency of word combinations (Firth, 1957; Church, 1990). Benson et al. (1986) defined collocation as “an arbitrary and recurrent word combination”, and classified collocations into a lexical group and a grammatical group. Xu et al. (2009) gave a more specific definition: “a collocation is a recurrent and conventional expression containing two or more content word combinations that hold syntactic and/or semantic relations.” This definition emphasizes the syntactic and semantic relations between words but excludes the function words.

Existing resources of collocational knowledge mainly fall into two types, i.e. collocation dictionary and collocation bank. Manually compiled dictionaries may have problems in coverage and consistency, and it is quite difficult to add new entries or collocations (Smadja, 1993). Besides, a list of typical collocations without contexts is not sufficient to support language learning. Xu et al. (2009) built Chinese Collocation Bank (CCB) with true collocations annotated in a large-scale news corpus. It is a valuable resource for collocation related research and NLP tasks, but might not be appropriate for language acquisition for two reasons. Firstly, collocations are domain dependent (Smadja, 1993). CCB annotated collocations from the People's Daily corpus (Yu et al., 2000), which consists of news articles of People's Daily, an official newspaper of the Chinese Communist Party. Frequent collocations in this corpus are mostly used in formal texts, and highly related to politics and economy. They might not be suitable to second language learning. Secondly, CCB defined collocations as content word combinations and did not deal with function words that are one of the most difficult parts for L2 learners.

Most previous studies used window-based methods to extract word combinations in a fixed window based on word co-occurrences or distribution scores (Church, 1990; Smadja, 1993; Sun et al., 1997; Xu and Lu, 2006). Obviously, these methods do not target at collocations with syntactic and/or semantic relations. Kilgarriff et al. (2004) used regular expressions over POS-tags to formalize rules of collocation patterns when building Word Sketch Engine. Huang et al. (2005) extended Sketch Engine to Chinese and found POS based rules were efficient in extracting grammatical information. However, they also addressed that the regular expression patterns faced challenges in long-distance collocation extraction and considerable risk of mis-classification, which is more serious in Chinese than in English.

[1] CTC (Corpus of Teaching Chinese as Second Language) contains text data from over 100 classic Chinese textbooks for L2 learning and New HSK (Chinese language proficiency Test) papers since 2009: http://www.aihanyu.org/basic.aspx.

[2] Chinese Wikipedia download: https://dumps.wikimedia.org/zhwiki/

Thus a better strategy should incorporate richer annotation of corpus and take more consideration of Chinese grammatical features.

3 The Proposed Approach

In contrast to definitions of collocation in the previous studies, We introduce function words into our collocation study, and identify four characteristics of collocations: (1) can be word combinations with more than two words; (2) can contain both content words and function words; (3) collocated words can be either adjacent or non-adjacent; (4) collocated words must hold syntactic or semantic relations.

We propose a syntax-based method based on dependency parsing and extract collocations from CTC, a text corpus for L2 learners to meet the domain needs. After preliminary studies, we find Chinese Wikipedia as a good supplement source because it has much bigger corpus size than CTC and the Wikipedia (WIKI hereafter) texts are from various domains. The collocations extracted from WIKI could serve as an important complement to CTC from both the coverage and the domain perspectives. In addition, complicated sentences in WIKI could be good resources for advanced-level learners.

In order to extract collocations based on the dependency trees efficiently and effectively, we define nine types of grammatical collocations as the extraction targets. As shown in Table 1, four of them have universal syntactic relations, while the other five types are Chinese unique collocations with specific syntactic or semantic relations that should be emphasized in second language acquisition. Function words are highly involved in all the types, and in addition to the most common two-word collocations, six types include three-word or four-word collocations.

Types	Language Independent	Function Words	No. of Words
Verb-Object (VO)	Y	direction verb, particle	2, 3, 4
Subject-Predicate (SP)	Y	direction verb, particle	2, 3, 4
Adjective-Noun (AN)	Y	particle	2, 3
Adverb-Predicate (AP)	Y	particle	2, 3
Classifier-Noun (CN)	N	classifier	2
Preposition-Postposition (PP)	N	preposition, direction noun	2
Preposition-Verb (PV)	N	prepostion, direction verb, particle	2, 3, 4
Predicate-Complement (PC)	N	direction verb, particle	2, 3, 4
Connective-Connective (CC)	N	conjunction, conjunctive adverb	2

Table 1. Collocation types

We will briefly introduce the five Chinese-dependent types in the following and the detailed descriptions for 26 forms of word combination based on the nine types are given in OCCA online user guide[3].

- **Classifier-Noun (CN):** Chinese classifier, also called measure word, is used to modify a noun after a number or quantifier. The translation of “a person" is 一个人 (yi/a ge/* ren/person), 个 (ge) is a measure word for 人 (ren/person). Most Chinese nouns require measure words based on their innate semantic and cognitive relations (Zhang, 2007; Her and Hsieh, 2010).
- **Preposition-Postposition (PP):** A preposition usually collocates with specific postpositions (also called direction nouns) to express spatial or temporal relations. They can constitute a prepositional phrase with a noun phrase in between, e.g. 在 桌子 上 (zai/* zhuozi/table shang/on) which means “on the table”.
- **Preposition-Verb (PV):** Both modern Chinese and English are S-V-O languages, however, word order in Chinese is often changed to S-P-O-V by some special prepositions to emphasize a part of the sentence, or to convey a nuance of the meaning (Hu et al., 2014). These prepositions directly introduce the objects of their collocated verbs, e.g. 把水喝了 (ba/* shui/water he/drink le/*) means “drink the water”. They only collocate with verbs that have certain semantic fea-

[3] OCCA user guide: http://occa.xingtanlu.cn/index.php?action=page&pid=11

tures, and most verbs must have one or two adjacent particles to indicate the aspect or direction (Wang, 1985; Lv, 1999; Zhang, 2001).

- **Predicate-Complement (PC):** Complement is a word, phrase or clause following the predicate (a verb or an adjective) to provide additional information, including result, direction, possibility, state, degree, quantity, duration, and location (Liu et al., 2001). To express these rich semantic features, complement has complicated internal structures. We summarize nine forms of Predicate-Complement collocation based on Liu et al. (2001)'s research.
- **Connective-Connective (CC):** Chinese conjunctions and adverbs can both serve as connectives between clauses to indicate various discourse relations, and these connectives are often used in a pair e.g. 因为 (yinwei/because) - 所以 (suoyi/so) for casual relation and 虽然 (suiran/although) – 但是 (danshi/but) for contrastive relation. We identify 45 connective pairs as word collocations based on Liu et al. (2001)’s research of conjunction and adverb usages.

After extracting collocations, we will construct a web-based collocation assistant (OCCA). In order to meet the practical requirements of learning and teaching which we have uncovered based on reviewing previous studies, the OCCA will provide users collocations with the following features: (1) given with context sentences; (2) defined and classified with full consideration of Chinese grammatical features, including syntactic structures and function words; (3) extracted from appropriate texts that are suitable for L2 learners in both domain and degree of difficulty; (4) updated easily with a change of corpus.

4 Extracting Collocations and Constructing OCCA

Figure 1 shows the steps in the process of extracting collocations and building OCCA. The details of each step will be explained as following.

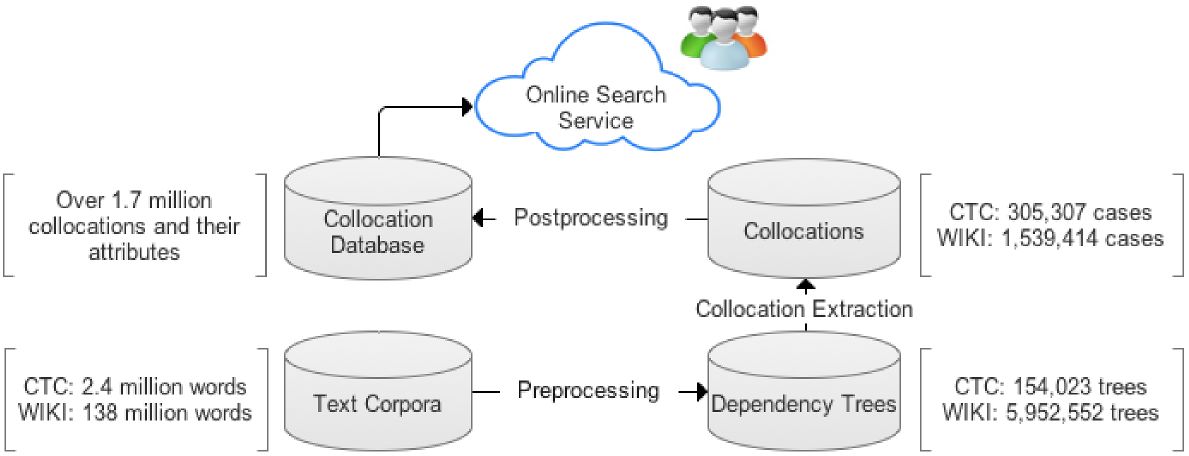
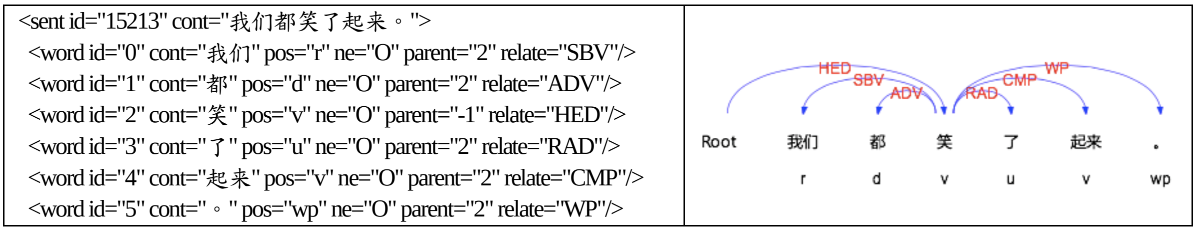


Figure 1. Procedure of our approach



```
<sent id="15213" cont="我们都笑了起来。">
 <word id="0" cont="我们" pos="r" ne="O" parent="2" relate="SBV"/>
 <word id="1" cont="都" pos="d" ne="O" parent="2" relate="ADV"/>
 <word id="2" cont="笑" pos="v" ne="O" parent="-1" relate="HED"/>
 <word id="3" cont="了" pos="u" ne="O" parent="2" relate="RAD"/>
 <word id="4" cont="起来" pos="v" ne="O" parent="2" relate="CMP"/>
 <word id="5" cont="。" pos="wp" ne="O" parent="2" relate="WP"/>
```

Figure 2. The pre-processing result of a Chinese sentence 我们都笑了起来 (wo’men/we dou/all xiao/laugh le/* qilai/*), which means “we all laugh”

4.1 Pre-processing

LTP-Cloud (Che et al., 2010), a Chinese NLP toolkit, is utilized to carry out word segmentation, POS tagging and dependency parsing for sentences in CTC and WIKI. We choose the latest released version (v3.3.1) of the toolkit to ensure the performance of each NLP module. We obtain over 6.1 million dependency trees from the two corpora and the results are presented in XML files. Figure 2 shows an XML tree and its dependency graph of a Chinese sentence.

4.2 Collocation Extraction

Dependency tree is an ideal carrier of various grammatical features including dependency relations as well as POS tags that can be used in automatic collocation extraction. As shown in Table 2, we firstly build mappings between dependency relations and collocation types. These relations exist between headwords and their direct modifiers. As they are not in one-to-one mappings with collocation types, POS and word location features are utilized as constraints. In addition to the dependency relations in Table 2, we use the RAD (right adjunct) relation to identify structural and aspect particles that are necessary units in three-word and four-word collocations. Since collocated words in Connective-Connective (CC) collocations do not have direct dependency relations between each other, we only adopt word and POS features for extraction of this type. After analysing all the collocation forms, we manually compile 37 rules to extract the nine types of collocations from the dependency trees. Most rules deal with two-word collocations by extracting dependency triples {headword; modifier; dependency relation}, and also collocations that have more than two words and multiple relations.

Collocation Type	Dependency Relation	Collocation Type	Dependency Relation
VO	VOB (object of verb)	AP	ADV (adverbial)
	IOB (indirect object)	CN	ATT (attribute)
	FOB (fronting object)	PP	POB (preposition-object)
SP	SBV (subject of verb)	PV	ADV (adverbial)
AN	ATT (attribute)	PC	CMP (complement)

Table 2. Mappings between dependency relations and collocation types

Taking the sentence in Figure 2 as an example, three colocations 我们 笑 (wo'men/we xiao/laugh), 都 笑 (dou/all xiao/laugh) and 笑 了 起来 (xiao/laugh le/* qilai/*) will be extracted in this step. The following rule illustrates how the three-word Predicate-Complement collocation 笑了起来 is extracted based on the words' location, POS and dependency features:

```
if word[relate] == 'CMP' & word[parent] == word[id]-2 & sent[word[id]-2][pos] == v & sent[word[id]-1][relate] == 'RAD' & sent[word[id]-1][parent] == word[id]-2:
    extract sent[word[id]-2][cont], sent[word[id]-1][cont], word[cont] as a three-word Predicate-Complement collocation.
```

4.3 Post-processing

In step 2, we extract 305,307 CTC collocations and 1,539,414 WIKI collocations. To refine the extraction results, we filter out 4,262 CTC and 84,893 WIKI collocations by seven filtering rules, e.g. exclude a collocation if there is a colon between the headword and its direct modifier in the sentence.

After that, we calculate the frequency of each collocation, and identify the relevance between two collocations if they have the same headword, direct modifier and dependency relation, e.g. examples (c1, c2, and c3) mentioned in Section 1. As the only difference is the structural particles, collocations like c2 and c3 are defined as variants of c1. The collocation relevance information could help users gain better understanding of grammatical and semantic roles of Chinese structural particles.

After post-processing, we obtain 1,755,566 collocations in the database. Unlike CCB (Chinese Collocation Bank) that only deals with content word combinations, our data covers a large number of collocations containing function words. Nearly 30% collocations are constituted by one or more function words including prepositions, conjunctions, direction verbs, direction nouns, structural and aspect particles. In addition to the most common two-word collocations, we also extract three-word and four-word collocations, which approximately account for 25% of the total number.

4.4 Building the Online Acquisition Services

We build OCCA based on Apache, PHP and MySQL. The website is http://occa.xingtanlu.cn/. Users can input a single keyword or multiple keywords to search for collocations from CTC (default) or WIKI database. The system outputs collocations with consideration of their relevance and ranks them based on frequency. Users can click a collocation for all the context sentences. The detailed user guide is available on OCCA website. Liu (2010) suggested that word collocations from a corpus can effectively assist language teachers to summarize typical usage patterns and important lexical properties of words. As OCCA offers collocations with statistics and context sentences, it can serve as a good assistant in vocabulary teaching. Students might also use it to retrieve word usages they are uncertain about.

5 Comparison with two Collocation Dictionaries

We evaluate the coverage (collocation type and quantity) and accuracy of OCCA by comparing its searching results with two classic Chinese collocation dictionaries: Modern Chinese Collocation Dictionary (Mei, 1999) and Modern Chinese Content Word Collocation Dictionary (Lin and Zhang, 1992) hereafter referred as D1 and D2.

D1 and D2 collect collocations for over 6,000 and 8,000 entries. In OCCA the CTC and WIKI databases contain 7,632 and 47,475 keywords with occurrences greater than 10 respectively. Considering CTC is a more comparable resource with the dictionaries in vocabulary size, we only use WIKI collocation data as reference in the comparison. As we are building a collocation assistant for second language acquisition, 100 words with the highest error occurrences in the HSK Dynamic Composition Corpus[4] are used as our test words.

After searching for these 100 words in OCCA (the default CTC database), and looking up them in D1 and D2, we find that OCCA contains 98 of them, with higher coverage than D1 (78 words) and D2 (24 words)[5]. The reason for this is that students make mistakes involving function words more often than content words. However, the main lexical items in the dictionaries are content words only, e.g. nouns, verbs, and adjectives.

We also analyse the collocation data of 19 words in the intersection of the three resources. The comparison is conducted from three perspectives: collocation type coverage, collocation quantity and collocation accuracy. As shown in Table 3, for these 19 entries, D1 and D2 both show six types of collocations in our definition, omitting Preposition-Verb (PV) and Preposition-Postposition (PP) collocations that deal with the syntactic and semantic relations of prepositions. As D1 and D2 do not include conjunction entries, the Connective-Connective collocations are not involved in this comparison either. We count the collocations of the 19 entries in these resources and find OCCA is much higher than D1 and D2 in collocation quantities. However, by analysing the collocation data, we also find that OCCA does not contain some collocations in D1 and D2, e.g. “外交 (waijiao/diplomatic) 问题 (wenti/problems)”, and “农村 (nongcun/village) 发展 (fazhan/develop)”, mainly because of the domain and size limit of the CTC corpus. However, the WIKI database could serve as a good supplement because it covers nearly 63% of the 278 collocations that cannot be retrieved in CTC.

Resources	Types	Quantities	Accuracy
D1	6	622	~100.00%
D2	6	3,234	~100.00%
OCCA	8	9,705	95.19%

Table 3. Collocation data of 19 entries

Assuming D1 and D2 both have 100.00% precision in collocation data, we manually label the correctness of the 9,705 collocations retrieved from the CTC collocation data. 467 of them are annotated as inappropriate collocations, thus the accuracy is 95.19%. Among the 467 collocations, about 6% of them are due to tokenization errors in pre-processing, 87% of them result from parsing errors e.g. 把 X 产生 (ba/* X chansheng/ produce), and 7% of them are extracted from correct parsing results but cannot be taken as appropriate word collocations, e.g. “三 表示” (san/three biaoshi/means). We also find that over 86% of these mistakes occur only one time, and over 8% of them occur two times. Thus OCCA now only outputs collocations with occurrences no less than three times to reduce the negative effect. We will collect collocation errors reported by OCCA users and develop better filtering module based on these data in future work.

[4] HSK dynamic composition corpus is a Chinese leaner corpus released by Beijing Language and Culture University: http://202.112.195.192:8060/hsk/index.asp

[5] Test word list and the retrieval results: http://101.200.121.46/OCCA_test_words.pdf

6 Experiment on Grammatical Error Detection

Collocation data has been proved helpful in a list of Computer Aided Language Learning (CALL) tasks for English learners (Shei and Pain, 2000; Futagi et al., 2008; Wu et al., 2010). To examine the effectiveness of our Chinese collocation resource, we conduct a preliminary experiment on grammatical error detection task using the OCCA collocation data.

NLP technologies for Chinese grammatical error diagnosis (CGED) have received a considerable amount of attention in recent years (Yu et al., 2014; Lee et al., 2015). Various linguistic and computational resources including L1 corpora, L2 corpora and web N-gram data have been employed to identify the grammatical errors (Chang et al., 2012; Cheng et al., 2014; Lee et al., 2014). However, Chinese collocation resource is rarely used in previous studies.

We use the test set of 2015 NLP-TEA CGED shared task in the experiment. This data set consists of 1,000 Chinese sentences collected from the essay section of the computer-based Test of Chinese as a Foreign Language (TOCFL), administered in Taiwan (Lee et al., 2015). Half of these sentences are correct, while the other half include a single defined grammatical error: redundant (132), missing (126), selection (110), and disorder (132).

As the test sentences are in traditional Chinese, while our collocation data is simplified, we carry out the language conversion with the Open Chinese Convert tool[6]. After that, we extract 7,262 word combinations from the 1,000 sentences by taking the same pre-processing and collocation extraction methods described in Section 4.

By looking up each collocation in OCCA including both CTC and WIKI databases, 496 instances in 401 sentences are identified with zero occurrence. We assume these sentences contain improper word combinations and manually label these collocations to check if they are valid to indicate grammatical errors. Table 4 shows examples we consider the zero occurrence tokens are related with the errors.

Type	Examples	Correction
Redundant	可是现在我把什么事都不记得。 (Preposition-Verb: 把 记得)	可是现在我什么事都不记得。 But I can’t remember anything now.
Missing	我们到现在都是以写信为连络。 (Verb-Object: 为 联络)	我们到现在都是以写信为连络方式。 We are still writing letters to each other.
Selection	排队了很多时间才轮到我。 (Predicate-Complement: 排队 很多 时间)	排队了很久才轮到我。 I queued for quite a while before my turn.
Disorder	他教书在那个很有名的大学。 (Predicate-Complement: 教书 在)	他在那个很有名的大学教书。 He is teaching at that famous university.

Table 4. Grammatical error examples with zero occurrence collocation

After manual analysis, we find the collocation data could successfully detect word errors in 212 sentences. Among them 41% of the errors are detected by non adjacent word combinations, e.g. 把- 记得 (ba/* jide/remember) in which 把 (ba/*) is a redundant word, and 29% of the errors are detected by 3-word or 4-word combinations, e.g. 排队 很多 时间 (paidui/queue henduo/much shijian/time) in which 很多 时间 is a word selection error.

We also calculate the sentence-level precision, recall and F1 score for four error types. Table 5 shows the test result and the performance of three teams that achieved the best precision (CYUT team run2), recall (NTOU team run1) and F1 score (NTOU team run2) in the error detection part of CGED shared task.

From the data in Table 5, we could see that the collocation data is a helpful resource in the detection of inadequate word combinations. Even though we adopt a very simple data retrieval method, it achieves higher Precision and F1 score than the average of the CGED teams. The precision of our method is very close to the best precision 0.7453 achieved by CYUT team, and we have higher recall and F1 score than they have. NTOU team (Lin and Chen, 2015) proposed to measure sentence likelihood scores with Chinese Web 5-gram data for error detection. Compared with the n-gram dataset that is widely used in grammatical error diagnosis, the most notable advantage of the collocation data is it

[6] Open Chinese Convert Project: https://github.com/BYVoid/OpenCC

captures not only adjacent words but also non-adjacent words with syntactic or semantic relations, thus it could be able to indicate errors that exist in a long distance, e.g. the preposition-verb example 把-记得(ba/* jide/remember) in Table 3.

Type	Precision	Recall	F1 score
Redundant (42)	0.8571	0.3182	0.4641
Missing (48)	0.8136	0.3810	0.5189
Selection (67)	0.7128	0.6091	0.6569
Disorder (55)	0.5392	0.4167	0.4701
Average	**0.7307**	**0.4312**	**0.5424**
CGED bestP Team (CYUT-run2)	**0.7453**	0.2400	0.3631
CGED bestR Team (NTOU-run1)	0.5000	**1.0000**	0.6667
CGED bestF1 Team (NTOU-run2)	0.5164	0.9760	**0.6754**
CGED average	**0.5600**	**0.6066**	**0.5327**

Table 5. Performance of grammatical error detection

It should also be noted that the recall of our method is lower than precision which means that not all the grammatical errors are related with word collocations, especially the word redundant and missing types. To understand the pros and cons of this collocation based method, we analyse the false positive (FP) and false negative (FN) cases and reach the following findings.

False positive cases are correct sentences that are identified with an error because they contain collocations that cannot be retrieved in our database. The most important reason is the word usage difference between traditional and simplified Chinese. Some common words in traditional Chinese e.g. 捷运(jieyun/subway) and 障碍者(zhang'aizhe/disabled person) have different representations e.g. 地铁(ditie/subway) and 残疾人(canjiren/disabled person) in Mandarin. There are also some word combinations incorrectly extracted because of the word segmentation and parsing errors.

False negative cases refer to the incorrect sentences whose errors are not detected by our method. By analysing these sentences, we find most of the errors are relevant with adverb and auxiliary usages. With an adverb or auxiliary missing or using incorrectly, word combinations in the sentence could still be grammatically correct but the whole sentence might not meet the strict requirement of word order or semantic pattern. The following sentence is an example with a word missing error that the collocation based method could not identify. In this sentence an adverb 都 (all) should be used to modify the adjective 差不多 (similar):

Sentence: 我们发现了每个国家人民*差不多。We find people in different countries are similar.

Correction: 我们发现了每个国家人民都差不多。We find people in different countries are all similar.

From the above analysis, we could see that the collocation based method could be effective in detecting a fairy proportion of grammatical errors, especially for those involved non adjacent words. However, the collocation data alone is far from enough, thus we need to combine the dataset with other language resources e.g. the Web 5-gram data and statistical models to explore better strategies.

7 Conclusions

This paper discusses the design and construction of Online Chinese Collocation Assistant (OCCA) for second language teaching and learning. We identify the important roles of function words in collocational knowledge and develop automatic extraction method to extract nine types of Chinese collocations from two text corpora. By searching for keywords in OCCA, users can easily obtain collocations with their statistics and context sentences that are helpful in the vocabulary teaching and learning. We compare the collocation data of OCCA with two traditional dictionaries, and find OCCA has much higher entry coverage and collocation quantity. It also has quite low collocation error rate at less than 5%. In order to investigate the helpfulness of OCCA, we conduct a preliminary experiment to apply the collocation resource to Chinese grammatical error detection. By implementing the simple data retrieval method, OCCA collocations are effective in detecting four types of grammatical errors and demonstrate comparable performance with comparison to the best results in 2015 NLP-TEA CGED shared task.

To minimize the negative impacts of mistakes and to offer much more reliable collocation resources, we will develop more efficient and effective methods for filtering and correcting collocations in the future. We also hope to conduct further study to verify the effectiveness of OCCA by combining the collocation data with other language resources and methods in CALL tasks.

Acknowledgments

This work was partially supported by the National High Technology Research and Development Program of China (No. 2012AA011104) and National Language Committee Research Program of China (No. YB125-124). This work was also partially supported by Ministry of Science and Technology, Taiwan under the grant number MOST-105-2410-H-002-118. We thank the anonymous reviewers for their insightful comments.

References

Jens Bahns and Moira Eldaw. 1993. Should we teach EFL students collocations?. System, 21(1): 101-114.

Morton Benson, Evelyn Benson and Robert F. Ilson. 1986. The BBI combinatory dictionary of English: A guide to word combinations. John Benjamins, Amsterdam and Philadelphia.

Ru-Yng Chang, Chung-Hsien Wu and Philips Kokoh Prasetyo. 2012. Error diagnosis of Chinese sentences using inductive learning algorithm and decomposition-based testing mechanism. ACM Transactions on Asian Language Information Processing, 11(1): Article 3.

Wanxiang Che, Zhenghua Li and Ting Liu. 2010. Ltp: A chinese language technology platform. In Proceedings of COLING (System Demonstrations), pages 13-16.

Shuk-Man Cheng, Chi-Hsin Yu and Hsin-Hsi Chen. 2014. Chinese Word Ordering Errors Detection and Correction for Non-Native Chinese Language Learners. In Proceedings of COLING, pages 279-289.

Ward K. Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. Computational linguistics, 16(1): 22-29.

Mohammed Farghal and Hussein Obiedat. 1995. Collocations: A neglected variable in EFL. IRAL-International Review of Applied Linguistics in Language Teaching, 33(4): 315-332.

John R. Firth. 1957. Papers in Linguistics, Oxford University Press, London, UK.

Yoko Futagi , Paul Deane , Martin Chodorow and Joel Tetreault. 2008. A computational approach to detecting collocation errors in the writing of non-native speakers of English. Computer Assisted Language Learning, 21(4):353-367.

One-Soon Her and Chen-Tien Hsieh. 2010. On the semantic distinction between classifiers and measure words in Chinese. Language and linguistics, 11(3): 527-551.

Renfen Hu, Zhiying Liu, Lijiao Yang and Yaohong Jin. 2014. Pre-reordering Model of Chinese Special Sentences for Patent Machine Translation. In Proceedings of COLING Workshop on Synchronic and Diachronic Approaches to Analyzing Technical Language, pages 40-47.

Chu-Ren Huang, Adam Kilgarriff, Yiching Wu, Chih-Ming Chiu, Simon Smith, Pavel Rychly, Ming-Hong Bai and Keh-Jiann Chen. 2005. Chinese Sketch Engine and the extraction of grammatical collocations. In Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing, pages 48-55.

Adam Kilgarriff, Pavel Rychly, Pavel Smrz and David Tugwell. The sketch engine. 2004. In Proceedings of Euralex, pages 105-116.

Lung-Hao Lee, Liang-Chih Yu and Li-Ping Chang. 2015. Overview of the NLP-TEA 2015 Shared Task for Chinese Grammatical Error Diagnosis. In Proceedings of the 2nd Workshop on Natural Language Processing Techniques for Educational Applications, pages 1–6.

Lung-Hao Lee, Liang-Chih Yu, Kuei-Ching Lee, Yuen-Hsien Tseng, Li-Ping Chang and Hsin-Hsi Chen. 2014. A Sentence Judgment System for Grammatical Error Detection. In Proceedings of COLING (System Demonstrations), pages 67-70.

Chuan-Jie Lin and Shao-Heng Chen. 2015. NTOU Chinese Grammar Checker for CGED Shared Task. In Proceedings of the 2nd Workshop on Natural Language Processing Techniques for Educational Applications, pages 15–19.

Dekang Lin. 1998. Extracting collocations from text corpora. In Proceedings of first workshop on computational terminology, pages 57-63.

Xingguang Lin and Shoukang Zhang. 1992. Dictionary of modern Chinese content word collocations. Commercial Press, Beijing, China.

Fengqin Liu. 2010. Corpus based collocation research and vocabulary teaching for second language learners. Modern Chinese, (6):115-117.

Yuehua Liu, Wenyu Pan and Weihua Gu. 2001. Pratical grammar of modern Chinese. Commercial Press, Beijing, China.

Jianji Lu. 1987. Vocabulary error analysis of Chinese second language leaners. Language Teaching and Research, (4): 122-132.

Shuxiang Lv. 1999. Eight hundred words in modern Chinese. Commercial Press, Beijing, China.

Jiaju Mei. 1999. Dictionary of modern Chinese collocations. Chinese Dictionary Press, Shanghai, China.

Nadja Nesselhauf. 2003. The use of collocations by advanced learners of English and some implications for teaching. Applied linguistics, 24(2): 223-242.

C.-C. Shei and Helen Pain. 2000. An ESL writer's collocational aid. Computer Assisted Language Learning, 13(2): 167-182.

Anna Siyanova and Norbert Schmitt. 2008. L2 learner production and processing of collocation: A multi-study perspective. Canadian Modern Language Review, 64(3): 429-458.

Frank Smadja. 1993. Retrieving collocations from text: Xtract. Computational linguistics, 19(1): 143-177.

Maosong Sun, Changning Huang and Jie Fang. 1997. Quantitative analysis of Chinese collocations. Studies of the Chinese Language, 256(1): 29-38.

Li Wang. 1985. Modern Chinese grammar. Commercial Press, Beijing, China.

Brent Wolter. 2006. Lexical network structures and L2 vocabulary acquisition: The role of L1 lexical/conceptual knowledge. Applied Linguistics, 27(4): 741-747.

Jian-Cheng Wu, Yu-Chia Chang, Teruko Mitamura and Jason S. Chang. 2010. Automatic collocation suggestion in academic writing. In Proceedings of the ACL, pages 115-119.

Ruifeng Xu and Qin Lu. 2006. A multi-stage chinese collocation extraction system. In Advances in Machine Learning and Cybernetics, pages 740-749.

Ruifeng Xu, Qin Lu, Kam-Fai Wong and Wenjie Li. 2009. Building a chinese collocation bank. International Journal of Computer Processing of Languages, 22(01): 21-47.

Lijiao Yang and Hang Xiao. 2015. Contextual information labeling in a corpus for second language learning. Applied Linguistics, (1):107-116.

Liang-Chih Yu, Lung-Hao Lee and Li-Ping Chang. 2014. Overview of grammatical error diagnosis for learning Chinese as a foreign language. 2014. In Proceedings of the 1st Workshop on Natural Language Processing Techniques for Educational Applications, pages 42-47.

Shiwen Yu, Xuefeng Zhu and Huiming Duan. 2000. Guideline of Large-scale Modern Chinese Corpus Annotation. Journal of Chinese Information Processing, 14(6): 58-64.

Hong Zhang. 2007. Numeral classifiers in Mandarin Chinese. Journal of East Asian Linguistics, 16(1): 43-59.

Wangxi Zhang. 2001. The displacement pattern of BA Sentence. Language teaching and research, (3): 1-10.