The Recognition System of CCE (Chinese Conventional Expression) for CSL Learners

Shuang Xiao[1] Hua Xiang[1] Fuji Ren[1,2] and Shingo Kuroiwa[1]

[1] Faculty of Engineer, University of Tokushima 2-1 Minamijosanjima, Tokushima, 770-8506, Japan {sxiao, xianghua, ren, kuroiwa}@is.tokushima-u.ac.jp
[2] School of Information Engineering, Beijing University of Posts and Telecommunications Beijing 100876, China

Abstract. CSL (Chinese as Second Language) learners need relevant reading support because recognition of CCE (Chinese Conventional Expression) is a difficulty for them. In this paper, we mainly propose how to help CSL learners recognize CCE in Chinese text. We have created a CCE data-base with 2,305 conventional expressions of contemporary Chinese. At the same time we have analyzed the basic structures and application forms of these 2,305 conventional expressions. By the analysis we have presented an extraction approach which based on rules and characters of these CCE. In our extraction experiment of CCE, the recall achieved 81.65% and the precision achieved 94.34%.

1 Introduction

As a matter of fact CCE (Chinese Conventional Expression) is a big obstacle for the CSL (Chinese as Second Language) learners, it is obliged to recognize and understand CCE in Chinese way. CCE is a traditional, habitual expression that is widely used in daily life among Chinese. The applications of CCE are very flexible and unique. CCE also is a kind of special phrase which has the fixed structure [1]. Usually, they consist of two or more than two words, and most of them have the strong culture background. Because the natural meaning of CCE is not the simple combination of literal meaning of each word but is the extensional cultural implications of the phrase. In this case the learners maybe can't catch the real meaning of the term even if they can understand the every word of the CCE literally. For example: *'Yin1Wei2 Ta1 Diu1 Le Fan4Wan3, Quan2Jia1 De Sheng1Huo2 Jiu4 Bian4 De2 Fei1Chang2 Kun4Nan4 Le.'* (The living of his family has been very difficult because he lost his job.) In this sentence the underline parts-*'Diu1 Fan4 Wan3'* is a CCE which consist of the word *'Diu1'* (lose) and word *'Fan4 Wan3'* (rice bowl). Here the real meaning of the phrase is 'lose someone's job'. For CSL learners, it is a problem that how to recognize and understand the CCE in this sentence. (The italics are Chinese Pinyin, the Roman transliteration of Chinese characters, which is used throughout this paper for the convenience of English readers. The numbers are tone markers.)

In this paper, we discuss how to help CSL learners to recognize CCE in Chinese text. We have created a CCE database with 2,305 conventional expressions of contemporary Chinese. In section 2, we analyze the basic structures

© A. Gelbukh, S. Torres, I. López (Eds.)
Advances in Artificial Intelligence
Research in Computing Science 20, 2006, pp. 49-62

and application forms of these conventional expressions. In section 3, we present an extraction approach which based on rules and characters of these CCE to recognize them. Section 4 is the description of extraction experiment and some discuss of the result. Finally, we give conclusions and introduce part of our later research work.

2 Analysis of CCE

2.1 Analysis of Chinese Conventional Expressions

The system we presented is based on a database of 2305 contemporary CCE. By the statistical analysis of these 2305 conventional expressions, we have found that contemporary CCE may consist of different character quantity. Generally, contemporary CCE consist of three Chinese characters to twelve Chinese characters. And we have also found the conventional expressions with three Chinese characters are in majority in total quantities of contemporary CCE – they account for 65.2% (1503) of all. The conventional expressions with four Chinese characters take up the second large quantities-about 16.5% (381). Furthermore, the conventional expressions with five, six, seven, eight Chinese characters account for 7.7% (177), 4.9% (113), 3.7% (89) and 1.4% (32) respectively. Other CCE only account for 0.4% (10). Hence, according to situation of quantities, processing on conventional expressions with three, four, and five characters is the most important work for the research.

Simultaneously, we have analyzed the structures and unitary attributes of conventional expressions. We have classified four categories of conventional expressions by various external unitary attributes-verbal phrase expressions, noun phrase expressions, 'clause' expressions and the others expressions. The verbal phrase expressions include: 'predicate-object' structure, 'adverb-headword' structure, 'continuous predicates' structure and 'predicate-complement' structure. The noun phrase expressions include: 'attribute-headword' structure, 'parataxis' structure and 'character De' structure. The 'clause' expression including: 'subject-predicate' structure and 'complicate' structure. 'Complicate' here denotes that the structure is far more complicate than common phrases. It looks like a clause. The other expressions include: 'special' structure and 'comma separate' structure. 'Special' structure indicates the phrases with irrational structure which is absolutely different from general phrase structures. The 'comma separate' structure expression consists of 'in front part' and 'behind part'. Most 'in front part' and 'behind part' of 'comma separate' structure expression have symmetrical structure and equal quantities of characters.

From Table 1 we can learn that contemporary CCE is mainly described by verbal phrase expressions and noun phrase expressions. For the reason they account up to 86.7% of whole conventional expressions, in our research we have focused on these two kinds of conventional expressions.

Table 1. The Structures and Attributes of Chinese Conventional Expressions

Categ.	No.	Structure category	Examples	Q.	Freq.
Verbal phrase	1	'predicate-object' structure	*Chi1 Bai2Fan4* (fathead, to be a 'good-for-nothing')	1056	45.8
	2	'adverb-headword' structure	*Ji1Dan4 Li3 Tiao1 Gu2Tou2* (disposition to find and point out trivial faults)	50	2.2
	3	'continuous predicates' structure	*Jian4 Pian2Yi2 Jiu4 Qiang3* (to gain extra advantage unfairly)	114	4.9
	4	'predicate-complement' struc.	*Huo2 De2 Bu2Nai4Fan2* (the act or process of destroying oneself or itself)	5	0.2
Noun Phrase	5	'attribute-headword' struc.	*Hui1Se4 shou1Ru4* (illegal income)	759	32.9
	6	'parataxis' struc.	*Ban4Jin1 Ba1Liang3* (all the same)	12	0.5
	7	'character De' struc.	*Na2 Bi2Gan3 Zi3 De* (the intellect)	4	0.2
'Clause'	8	'subject-predicate' structure	*Jing3Shui3 Bu2 Fan4 He2Shui3* (none may encroach upon the precincts of another)	172	7.5
	9	'complicate' structure	*Ge1Bo1 She2 Le Wang3 Xiu4Zi3 Li3 Cang2* (to endure an humiliation by one's own self)	35	1.5
Others	10	'special' structure	*San1 Yi1 San1 Shi2 Yi1* (to divide equally, share alike)	18	0.8
	11	'comma separate' structure	*Chai1 Dong1Qiang2, Bu3 Xi1Qiang2* (keep up in one place at the expense of others)	80	3.5
Total	11			2305	100

2.2 Form Analysis of CCE in Applications

We have created a large conventional expressions database of 21,018 example sentences. By statistics and comparing, we have found that there are various expressional situations exist in CCE. The detail analysis is depicted in table 2. We have found that the CCE may remain their primary forms in most of time. These phenomena take up 82.4% in whole quantities of conventional expressions. As the second largest phenomena of conventional expression, the 'inserted words' CCE may take up 14.8% in whole quantities of our database. Another expressions is 'the first Chinese character repeating' phenomena. These phenomena indicate that a 'predicate-object' conventional expression with single verb Chinese character repeats at beginning. They account for 0.6% of the whole quantities. Next expressional situation is 'character replacing'. 'Character replacing' indicates that one certain character of the conventional expression can be replaced by the other characters (usually replaced by single character verb). After replacing, the novel conventional expression will retain the original meaning as before. These kinds of conventional expressions account for 0.5% of whole quantities. Besides, 'order changing' is a frequent expressional phenomenon too. 'Order changing' indicates that the order of conventional expression may be changed with the

other inserted words. They account for 1.7% of whole. Because 'order changing' expression is extremely complicate, and the database we collected about it is not large enough. Therefore in current work we will not make a detail analysis on these phenomena temporarily.

Table 2. Form Categories of Chinese Conventional Expressions in Application

Form in using	Example	Q.	Freq.
'unchanged' form	*Ci3Di4 Wu2 Yin2 San1Bai2 Liang3* (a very poor lie which reveals the truth)	17325	82.4
Form with 'inserted word'	*Zhua1 Bie2Ren2 De Bian4Zi3* (to seize on other people's mistake or failure)	3116	14.8
'the first Chinese character repeating' form	*Dai4 Dai4 Gao1 Mao4Zi3* (the vain compliments)	123	0.6
Form of 'character replaced'	*Gan3/Da3/Na2 Ya1Zi3 Shang4 Jia4* (force someone to do something)	102	0.5
'order changing' form (with the other insert words)	*Jian3 Zhao2 Pian2Yi2 Le* (to get a bargain, to get an extra advantage) *Pian2Yi2 Dou1 Rang4 Ta1 Jian3 Zhao2 Le* (He has gained all the advantages.)	352	1.7
Total		21018	100

Table 3. Categories of Chinese Conventional Expressions with Inserted Words

Structure categories	Example	Q.	Freq.
'predicate-object' structure	*Da3 Le Yi2 Ge4 Piao4Liang4 De Fan1Shen1 Zhang4* (changed completely)	2521	80.9
'subject-predicate' structure	*Jia4Zi3 Hen3 Da4* (arrogant, haughty)	327	10.5
'attribute-headword' structure	*Zhang3 Shang4 De Ming2Zhu1* ('a pearl on the palm', a parent refers affectionately to a beloved daughter)	236	7.6
'parataxis' structure	*Ji1Mao2 He2 Suan4Pi2* (tiny things, bits and pieces)	32	1.0
Total		3116	100

From table 3 we can learn that the change of the conventional expressions with 'predicate-object' structure is most active in daily applications. They take up 80.9% in entire 'inserted words' conventional expressions. The next frequent conventional expression forms are 'subject-predicate' and 'attribute-headword' structures, they take up 10.5% and 7.6% of whole 'inserted words' conventional expressions respectively. Besides, the 'parataxis structure' are relative less used, only account for 1.0%.

3 Recognition Processing of CCE

Comparing with English and Japanese, Chinese is a kind of 'isolated language'. It lacks of some characteristics (such as case-auxiliary word and changing of verb forms etc.) which English and Japanese do. In this case Chinese have more difficulties in word segment, lexical analysis and syntactic parsing than others [4][5]. From 90's, many researchers have tried to use shallow parsing technique to Chinese processing. At the same time, statistic method is adopted frequently [6][7][8]. Unfortunately, using the method which is based on frequency or collocation can not succeed fully for CCE. Because of that: many CCE have too small frequency in a corpus and some CCE don't have strong collocation. Thus in this paper, we have proposed a way to extract CCE based on rules and characters. The detail procedure is described as the next five steps.

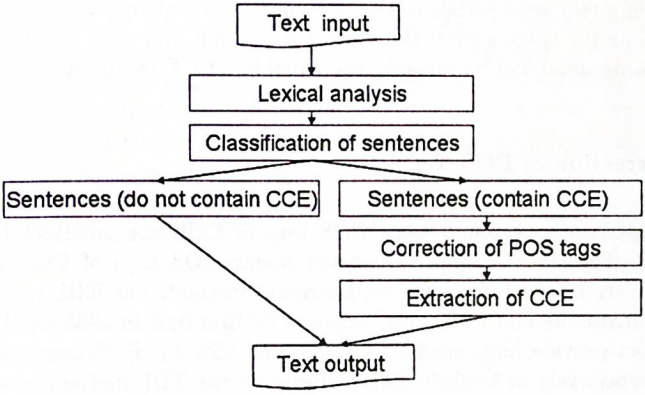

Fig. 1. The Recognition Process of CCE

3.1 Conventional Expression Registration

First of all we have registered almost all the conventional expressions we collected into the system. The registered CCE include not only the CCE themselves but also the structure, POS, and attributes of the CCE. All the CCE we put into register of the system consist of the four kinds of information.

3.2 Lexical Analysis

The lexical analysis system we adopted in our system is ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System) from Chinese Academy of Sciences. ICTCLAS have used the approach based on multi-layer HMM. It has a high segmentation precision of 97.58%. And the POS tagging we used is come from POS Tagging Collection from Beijing University.(http://www.nlp.org.cn/)

3.3 Classification of Sentences

After the lexical analysis, we can classify all the sentences of input text into two parts. One part of sentences doesn't contain CCE, the lexical analysis results of them will be output directly. The other part of sentences that contain CCE will be processed further. According to CCE in the sentences, we can divide the sentences with CCE into three types by matching the registration information of the system. The first type of sentences is that with 'continuous words' CCE. We can recognize these 'continuous words' CCE by using Maximum Matching Method. The second type of sentences is that with CCE of 'comma separate' structure. In this case the system will match the 'in front part' of the comma, the comma, and the 'behind part' of the CCE orderly. The third type of sentences is that with 'Inserted Words' CCE. The system will recognize these CCE by matching the every part of them orderly. In match processing, if there are two or more than two word serials can be matched with registration information of CCE, and at the same time if these word serials lap over each other, then the longest words serial will be taken as the candidate CCE in our system.

3.4 Correction of POS Tags

By our test, we have found some POS tags of CCE are incorrect by using ICTCLAS. Therefore, we have corrected wrong POS tags of CCE by TBL (Transformation-Based Error-Driven Learning) method. The TBL method is a kind of statistic method which was proposed by Eric Brill in 1995 [9]. The TBL method can provide high precision of 98% and 95% for POS tags and chunk parsing respectively in English [10][11]. Recently the TBL method is also used in Chinese processing. And the satisfied results of previous works show that the TBL method is effective for Chinese processing [12][13]. There are three necessary requisite of using the TBL method. They are a manually annotated database, an initial annotation program and a template of rule module. Firstly, the unannotated text is passed through an initial-state annotator giving an initial POS tags. Then the initial POS tags result is compared with the truth. A manually annotated corpus is used as our reference for truth. Based on this comparison, we can get a serial of candidature rules. By using the evaluation function, every candidature rule is given a score. We consider that the rule which was given the highest score is the best transformation rule. Next, the best transformation rule is used to correct POS tags of annotated text. Finally, the fore-mentioned processing is repeated until processing meet the finish condition. The detail algorithm will be explained in the following steps.

Initial POS Tags. We delete all the POS tags of CCE in training data-base then annotate them again by using ICTCLAS. The annotation of ICTCLAS is regarded as the basic results of processing.

Generation of Candidature Rules. Candidature rules are generated from the wrong POS tags of CCE. Generation condition of the candidature rules is context environment basically.

Acquisition of Transformation Rules. This is as follows:

(1) Evaluation Function: The evaluation standard for the candidature rule is the improvement of right recognition rate. The unannotated database is processed by the initial-state annotator, and this result in an annotated corpus with errors, determined by comparing the output of the initial-state annotator with the manually derived annotations for this corpus. Next, we apply each of the possible transformations in turn and score the resulting annotated corpus. At each interaction of learning, the transformation is found whose application results in the highest score according to the objective function being used, that transformation is then added to the ordered transformation list and the training corpus is updated by applying the learned transformation. We have provided the following formula of evaluation function: F(r)=C(r)−E(r). In this formula, 'r' is the rule. 'F' is the evaluation function. C(r) is the correct numbers which is obtained by using 'r', E(r) is the error numbers which is obtained by using 'r'.

(2) The End of Learning Process: Learning continues until no transformation can be found whose application results in an improvement to the annotated corpus. In our system, when $F(r) < 1$, learning process will be finished.

(3) Transformation Rules: We have obtained 156 transformation rules by the test. These transformation rules can be classified into three kinds. The first kind of transformation rules is based on POS tags. We use the rule module to describe them. In the rule module, 'P' is POS tag. 'T' is a word. 'PN' is the current POS tag of the word. 'P1' and 'P2'are POS tags of the first word and the second word which are located in the left side of the current word. 'P_1' and 'P_2' are POS tags of the first word and the second word which are located in the right side of the current word. We give a transformation rule and a correlative example as follows:

if {P1P2 is m/q && PN is not n} then { P of T from PN to n};

Ex. 1: '*Ta1/r Xiang4/v Yi4/m Zhi1/q ChuIShan1/v Hu3/n.*' (He is brave and vigorous like a tiger.) Ex. 1 is the processing result of ICTCLAS. In this example, '*ChuIShan1/v Hu3/n*' and CCE of '*ChuIShan1/n Hu3/n* (NP+NP)' have the same word serial, but their POS tag of the word '*ChuIShan1*' are different. According to fore-mentioned transformation rule, we can correct the POS tag of '*ChuIShan1*' from 'v'(verb) to 'n'(noun) easily.

If a character serial can be segmented as one word by ICTCLAS, we can transform POS tag of this segmented word from current POS tag to 'cv' (conventional expression). This kind of transformation rule can be described as follows('P' is POS tag. 'K' is a certain characters serial. 'PN' is the current POS tag. 'DC' is characters serial of registration CCE):

if {K can match DC && PN is not cv} then { P of K from PN to cv};

Ex. 2: '*Ta1/r De/u Hu2Li3Wei3Ba1/i Bei4/p Zhua1Zhu4/v Le/y.*' (The despicable fact of his action has been found.) In example 2, '*Hu2Li3Wei3Ba1*' is

a characters serial of CCE. But the POS tag is not 'cv'. We will transform its current POS tag from 'i' to 'cv'.

The third kind of transformation rule is based on characters of CCE. For example:

Ex. 3:'*Da4Jia1/r Dou1/d Cheng2/v Lao3Lao3/n Bu4/d Teng2/a, Jiu4 Jiu4/n Bu4/d Ai4/v Le/y.*"(Everyone has been ignored.) Example 3 includes the CCE '*Lao3Lao3/n Bu4/d Teng2/v, Jiu4Jiu4/n Bu4/d Ai4/v*'. But according to analysis result of ICTCLAS, the POS tag of word '*Teng2*' is 'a'(adjective). We can not correct its POS tag according to syntactic environment in this sentence. In this occasion, we can correct its POS tag by using characters of CCE. Its transformation rule module can be described as follow:

if { K can match DC && P1P2PN + comma + P_1P_2P_3 is n/d/a + comma + n/d/v} then { P of T from PN to v};

3.5 Extraction of CCE

The word with POS tag 'cv'. The word what has POS tag 'cv' will be recognized as CCE. We can extract it directly.

Ex. 1: '*Na4/r Jian4/q Shi4Qing2/n Yi3/d [Ba1Jiu3Bu4Li2Shi2/cv] Le/u.*' (That thing is near success.)

The CCE of 'Comma Separate' Structure. For the CCE of 'comma separate' structure, the extraction conditions are 'comma', words serial and POS of every word. If these conditions can be satisfied, we can extract this kind of CCE from sentences.

Ex. 2:'*Ta1/r Yi4 Sheng1/n Ke3Yi3/v Jiao4/v Cheng2/v Ye3/d Xiao1/nr He2/nr,Bai4/v Ye3/d Xiao1/nr He2/nr.*'(Both his success and failure are because of that person.)

'Continuous Words' CCE. The recognition knowledge of 'continuous words' CCE consist of attributes and environments of the CCE. The attribute of CCE include the words, the order and POS tags of the words. Environments of CCE indicate the attributes the words those are located in front or back of the CCE. (1) 'Noun Phrase' CCE: There are three types of 'noun phrase' CCE in our system: 'character De' structure CCE, 'attribute-headword' structure CCE, and 'parataxis' structure CCE. In the types of 'character De' structure only four CCEs have been collected. The structure of them are: 'verb + noun+(noun or proclitic word)+De'. According to semantic relations, the CCE with 'character De' structure can be extracted by its attributes.

Ex. 3: '*Ta1/r Shi4/v [Na2/v Bi2Gan3/n Zi3/k De/b].*'(He is an intellect.)

Comparing with 'character De' structure, the other two types of 'noun-phrase' CCE are far more complex. By analysis of the 'noun-phrase' CCE with 'attribute-headword' and 'parataxis' structure, we have found that there are five detail situations in both of these two types: 'noun + noun' structure, 'quantifier

+ noun' structure, 'noun + noun + noun' structure, 'quantifier + noun + quantifier + noun' structure and 'noun + prclitic' structure. Based on these different structures, we can generate five kinds of POS templates correspondingly. They are 'n-n', '(q)-m-n', 'n-n-n', '(q)-m-n-(q)-m-n', and 'n-k'. Thus the POS expansion templates can be obtained by combination of the interpunction and POS tag of the first forward or backward word next to the original templates. We have obtained 327 POS expansion templates totally. (a part of them is shown in table 4.) Consequently, we can extract 'attribute-headword' structure CCE and 'parataxis' structure CCE by these expansion POS templates. For instance, the CCE '*Hou4Qin2 Bu4Zhang3*' can be extracted by No28 expansion template.

Example 4: '*Ta1/r Zai4/p Jia1/n Zuo4/v [Hou4Qin2/n Bu4Zhang3/n] Yi3 Jing1/d San1/m Nian2/q Le/y.* '(He has been a housekeeper for three years.)

Table 4. Expansion POS templates

No.	P1	POS template	P_1		No.	P1	POS template	P_1
1	m+	n+n	+w		42	a+	n+n	+w
2	m+	n+n	+v		43	a+	n+n	+v
...	...	...	...		44	a+	n+n	+c
26	v+	n+n	+w		...	...	...	...
27	v+	n+n	+v		67	u+	n+n	+w
28	v+	n+n	+d		68	u+	n+n	+v
...	...	...	...		...	...	...	...

(2) 'Verbal Phrase' CCE: Four types 'verbal phrase' CCE have been collected in our system. There are 'predicate-object' structure CCE, 'adverb-headword' structure CCE, 'continuous predicates' structure CCE and 'predicate-complement' structure CCE. In this paper the 'verbal phrase' CCE has been extracted by their structures, verb attributes or context environment correspondingly.

Some 'verb phrase' CCE, especially CCE with 'adverb-headword' and 'predicate-complement' structure, can be extracted only by their attributes. For the adverbial modifier and complement in them case can be taken as the close adjunctive constituents here. (as ex.5,6). Furthermore, some 'continuous predicates' structure CCE with obvious structural symmetry and semantic correlation can be extracted by their attributes too (as ex.7). Besides, in the case of 'continuous predicates' structure CCE, because the last word of them are intransitive verbs, so the boundary of these CCE can be recognized by attributes of the verbs(as ex. 8).

Ex. 5: '*Ta1/r You3/v Ge4/q [Bu4/d Cheng2Qi4/v] De/b Xiao3Zi3/n.*' (He has a vain son.)

Ex. 6: '*Na4/r Wei4/q Lao3Xiong1/n You3/v Dian3/q [Huo2/v De2/u Bu2Nai4 Fan2/a] Le/y.*' (It is seem that the man do not want to live anymore.)

Ex. 7: '*Li3/nr Xian1Sheng1/n Shi4/v [Chi1/v Ming2/a Bu4/d Chi1/v An4/a] De/u Ren2/n.*' (Mr. Li would rather fight face to face than infighting stealthily.)

Ex. 8: '*Bu4Zhang3/n Bei4/p [Na2/v Xia4Ma3/v] Le/u.*' (The Minister has been dismissed.)

It is very difficult to recognize the linguistic constituent of the last noun in some CCE which with 'continuous predicates' (the last word is noun) and 'predicate-object' structures (as ex.9). According to composing rules of Chinese phrase, generally, whether a linguistic constituent is an object or not can be judged by two necessary conditions [14]. One of condition is that whether the linguistic constituent is an object of action. Based on combinability of the last noun and the words behind this noun, we can give some strict limited conditions to judge the back boundary of the CCE with 'continuous predicates' or 'predicate-object' structures. That is: if a noun and the word behind it can meet the limited conditions we given, the noun can be considered as a part of CCE.

Ex. 9: '*Gong1Si1/n Xu1Yao4/v Yi4/m Qun2/q Neng2/v [Da3/v Tian1Xia4/n] De/u Ren2/n.*' (The Company needs assiduous people.)

(3) 'Clause' CCE: According to structural symmetry and semantic correlation of 'Clause' CCE, the boundary of most 'Clause' CCE can be recognized by their attributes directly (as ex. 10, 11).

Ex. 10: '*She4Hui4/n Li3/f You3/v Bu4Shao3/m [Da4Chong2/n Chi1/v Xiao3 Chong2/n] De1/b Shi4/n.*' (In human society, there are many things operating as the law of jungle.)

Ex. 11: '*Bu4Shao3/m Ren2/n Dou1/d [Shen1/ng Zai4/p Fu2/n Zhong1/f Bu4Zhi1/v Fu2/n].*' (A lot of people neglect the happiness that they have owned.)

(4) 'Special Structure' CCE: 'Special structure' CCE indicate the CCE those disobey the Chinese grammar. In this case, it is impossible to recognize their boundary by context environment they are in. (18 CCE with 'special structure' have been collected in our work.) But the word serials of these CCE are very unique, so the possibility of co-occurrence of these words is very high correspondingly. As ex. 12 shows, the 'special structure' CCE is extracted by their attributes (the words and the order of these words).

Ex. 12: '*Ta1/r [San1/m Xia4/f Wu3/m Chu2/v Er4/m] Jiu4/d Ba3/p Na4/r Jian4/q Shi4/n Gan4/v Wan2/v Le/y.*'(He finished that work very quickly.)

'Inserted Words' CCE. These is as follows:

(1) 'Predicate-Object' Structure CCE: 'predicate-object' structure CCE with inserted words can be divided into four parts. They are verb (DC), verbal complement (DB), DingYu (DY) and object (BY). DingYu indicates the linguistic constituent that can be used to modify the noun object. The structure of this kind of 'Inserted Words' CCE as follows: DC-DB-DY-BY. Firstly the words, words order and POS tag of DC and BY can be confirmed. Then by matching corresponding registration information of the system, the DB and DY can be confirmed. Next a 'predicate-object' structure CCE which was inserted the DB and DY can finally be confirmed.

Based on analysis of Chinese verbal complement, the verbal complement can be classified into four categories. They are possible (or impossible) complement,

result complement, direction complement and movement complement. In table 5, some of DB which be inserted in 'predicate-object' CCE are shown.

The composition of DingYu is very complex. The general components of DingYu include quantifier, pronoun, noun, adjective, adverb, conjunction, onomatopoeic word and the auxiliary word 'De' etc. Besides, some complex phrases (such as 'subject-predicate' structure) and sentences can be considered as DingYu to modify noun too. In our database about 97% DingYu have relative simple structures. Thus the majority of DingYu with 'predicate-object' structure CCE can be recognized by their structures. In our work the POS templates were given to judge DingYu. When the POS tag order of DY matches the given POS templates, the DY is recognized successfully. In table 6, some POS templates of DY are described.

Ex. 13: '*Wo3/r Yi3/d [Da3/v Le/u Yi2 Ge4/m Piao4 Liang4/a De/u Fan1 Shen1 Zhang4/n].*'(I have changed completely.)

Table 5. Categories of Chinese verbal complement

No.	category	a part of inserted words
1	possible (or impossible) complement	*De2, De2Le, De2Zhao2, De2Hao3, De2Shang4, De2Xia4, De2Zhu4, De2Chu1, Bu4Le, Bu4Zhao2, Bu4Hao3, Bu4Shang4, Bu4Xia4, Bu4Zhu4, Bu4Chu1, ...*
2	result complement	*Chu1, Ru4, Cheng2, Dao4, Wan2, Hui4, Tong2, Jian4, Si3, Da4, Chang2, Duo1, Hao3, Kuai4, Xiao3, Shao3, Jin3, Huai4, Hei1, Zhong4, Ming2, Man3, ...*
3	direction complement	*Shang4, Xia4, Qi3*
4	movement complement	*Le, Zhe, Guo4*

Table 6. The POS templates of DY

No.	POS template of DY		No.	POS templates of DY
1	(q)+m		10	r+c+r+u+a+u
2	a+(u)		11	n+(u)
3	m+a+(u)		12	n+(u)+(q)+m
4	q+m+a+(u)		13	n+(c)+n+(q)+m
5	r+u		14	n+(u)+a+(u)
6	r+u+(q)+m		15	n+(u)+(q)+m+a+(u)
7	r+u+a+(u)		16	d+(d)+a+(u)
8	d+(d)+a+(u)		17	(q)+m+d+(d)+a+(u)
9	r+c+r+u		...	...

(2) 'Subject-Predicate' Structure CCE: The denial adverb and degree adverb are usually inserted in CCE with 'subject-predicate' structure. They can be

recognized by composition rule of Chinese adverbial modifier. The extraction processing is similar to 'Clause' CCE which we mentioned before.

Ex. 14: '*Ta1/r [Jia4Zi3/n Fei1Chang2/d Da4/a].*'(He is very arrogant.)

(3) 'Attribute-Headword ' Structure CCE (noun+noun): CCE with inserted part are usually made of nouns (noun+noun). Here the back noun is modified by the front noun. And the usual inserted part are the auxiliary words '*De*' or '*Zhi1*'. In this case, we can confirm the attributes of two parts which be separated and the inserted auxiliary word firstly. The boundary of the CCE can be recognized by the expansion template (same as expansion template of 'noun phrase' CCE.)

Ex. 15: '*Ta1/r Ru2Tong2/v Yi4/m Zhi1/q [Guo4Jie1/n De/u Lao3Shu3/n].*' (He is a universally condemned person.)

(4) 'Parataxis' Structure CCE: The 'parataxis' structure CCE with inserted part are usually made of nouns ('noun + noun'). The usual inserted words are paratactic conjunctions such as *'He2', 'Yu3', 'You4', 'Gen1', 'Jia1', 'Tong2'* and *'Dui4'*etc. In this case, we can confirm the attributes of two parts which be separated and the inserted conjunction word firstly. The boundary of the CCE can be recognized by the expansion template (same as expansion template of 'noun phrase' CCE.).

Ex. 16:'*[Ji1Mao2/n He2/c Suan4Pi2/n] De/u Xiao3Shi4/n.*'(tiny things, bits and pieces)

4 Experiment

In current research 1,200 hypertext files (about 13,000 sentences and 2,000 CCE) have been tested by our methods. All these test data are collected from internet, correlative books and papers. And the experiment evaluation was carried out by approaches of Recall, Precision and F-measure.

$$Recall = \frac{No.\ of\ extracted\ CCE\ correctly(C)}{No.\ of\ CCE\ in\ the\ test\ sentences(A)} \tag{1}$$

$$Precision = \frac{No.\ of\ extracted\ CCE\ correctly(C)}{No.\ of\ extracted\ as\ CCE(B)} \tag{2}$$

$$F-measure = \frac{Precision \times Recall \times 2}{Precision + Recall} \tag{3}$$

In our experiment 1,731 multi-word units were extracted as CCE. Among them 1,633 were correct. We have achieved 81.65% in Recall, 94.34% Precision and 87.54% in F-measure. As Table 7 shows. By the experiment results, we have found our extraction approach for CCE are successful. In this work the structural and semantic characters of CCE is taken as a key point of our research. We have also found that the error of word segment and POS tag are the main causes which produce failing Recall in a small part of CCE. The longer CCE is the more POS tag errors will be. So the extraction of long CCE is far more difficult

Table 7. Result of the Experiment

categories of CCE	A	B	C	Recall	Precision	F-measure
Noun Phrase	496	429	419	84.48%	97.67%	90.59%
Verbal Phrase	862	732	699	81.09%	95.49%	87.70%
'Clause'	169	146	125	73.96%	85.62%	79.37%
'special'	23	21	21	91.30%	100%	95.45%
'comma separate'	65	56	56	86.15%	100%	92.56%
'Predicate-Object'	273	248	227	83.15%	91.53%	87.14%
'Subject-Predicate'	57	51	41	71.93%	80.39%	75.93%
'Attribute-Headword'	32	28	26	81.25%	92.86%	86.67%
'Parataxis'	23	20	19	82.61%	95%	88.37%
Total	2,000	1,731	1,633	81.65%	94.34%	87.54%

than short ones. Thus the length of CCE must be taken in consideration in the future work. Besides, there were two causes leading the wrong extraction. One is boundary recognition of CCE. In this experiment, we successfully recognized the boundaries of 'noun phrase' CCE and 'verbal phrase' CCE. But the boundaries of many 'Clause' CCE were failed in recognition. The other cause is semantic judgment of CCE. In the system, the meanings of CCE still can not be judged very well.

5 Conclusion

In this paper we have discussed how to help CSL learners to recognize CCE in Chinese text. And an extraction approach based on the rules and character of CCE has been presented. By the extraction experiment, the results of both Recall and Precision are over 80%. In the future, we will enlarge the analytical quantity of CCE. For improving extraction result, the length information, semantic analysis and the Particularity Words (name, address and thing etc.) of the CCE will be taken into consideration. Furthermore we will advanced the reading support function by relevant technique.

References

1. Li Xinjian, Issues of Research and Standardization of Chinese Idiomatic Expressions, Applied Linguistics, 2002, 55-60
2. Mitsuko, Yamura Takei, Teruaki Aizawa, Miho FUJIWARA, Cognitive and SLA approaches to Computer-Assisted Reading, Making the Invisible Visible, Transaction of Japanese Society for information and System in Education, 2004
3. Susan M. Gass, Larry Selinker, Second Language Acquisition (An Introductory Course), Second Edition, Mahwah, N.J.: Lawrence Erlbaum Associates, 2001

4. F. Ren, Y. Miyazaki, and K. Tochinai: "An Algorithm for Estimating Chinese Supplementary Words in Japanese-Chinese Machine Translation System", Trans. Information Proce. Society of Japan, Vol.32, No.11, (1991-11), pp.1374-1382
5. C.Zong and F.Ren: "Chinese Utterance Segmentation in Spoken Language Translation", Computational Linguistics and Intelligent Text Processing, Ed. Alexander Gelbukh, Springer, LNCS2588, (2003),pp.516-525
6. Patrick Pantel and Dekang Lin: "A Statistic Corpus-Based Term Extractor", Canadian Conference on AI2001, 36-46
7. Tingting He, Jianzhou Liu, Donghong Ji: "A Statistical of Extracting Chinese Multi-Word Units", Journal of Chinese Language and Computing, 2002, 239-247
8. Hideki Isozaki and Hideto Kazawa: "Speeding up Support Vector Machines for Named Entity Recognition", Information Processing Society of Japan, Vol.44, No.3, 2003
9. Eric Brill: Transformation-Based Error-Driven Learning and Nature Language Processing: A Case Study in Part of Speech tagging, CLV21, 1995
10. Eric Brill, Some Advances in Transformation-based Part of Speech Tagging, In: Proceedings of the Twelfth National Conference on Artificial Intelligence, 722-727, 1994
11. Voutilainen. Atro: "NLPTool, a Detector of English Noun Phrases", In Proceedings of the Workshop on Very Large Corpora, ACL, 48-57, 1993
12. Sujian Li, Qun Liu, Chunk Parsing Based on Hybrid Model, in Maosong Sun, Tianshun Yao, Chunfa Yuan, eds., Advances in Computation of Oriental Languages, Proceedings of 20th International Conference on Computer Processing of Oriental Languages, Tsinghua University Press, pp.118-124, 2003
13. Jun Zhao, Changning Huang: "A model based on transformation to recognize Chinese base-NP", Journal of Chinese Information Processing, Vol.13, No.2, 1999
14. Yuehua Liu, Wenyu Pan and Wei Gu: "Chinese Grammer", Foreign Language Teaching and Research Press, 1986
15. Endong Xun, Changning Huang, and Ming Zhou: "A unified statistical model for the identification of English base-NP", ACL-2000: The 38th Annual Meeting of the Association for Computational Linguistics, Hong Kong, 3 - 6 October 2000
16. Shiwen Yu: "specification for the Prase Structure Knowledge-based of Contemporary Chinese", Journal of Chinese Language and Computing, 13(2), 215-226, 2003
17. Maria Chiara Levorate, Barbara Nesi, and Cristina Cacciari: "Reading comprehension and understanding idiomatic expressions: A developmental study", M.C. Levorato et al. l Brian and Language 91 303-314, (2004)
18. Hiroyuki Shinnou and Hitoshi Isanaha: "Automatic Acquisition of Idioms on Lexical Peculiarity", Trans. Information Proce. Society of Japan, Vol.36, No.8, pp.1845-1854, 1995
19. Andrew Hippisley, David Cheng and Khurshid Ahmad: "The head-modifier principle and multilingual term extraction", Cambridge University Press, Nature Language Engineering 11 (2): 129-157. 2005
20. Danny MINN, SANO Hiroshi: "A Study of Japanese Idioms for Learners of Japanese-A Statistic Approach", IPSJ SIJ Technical Report, 55-62, 2001