A CIE Extraction System for CSL Learners

Shuang Xiao, [†] ***Hua Xiang,*** [†] ***Fuji Ren,*** [†, ††] ***and Shingo Kuroiwa*** [†]

[†] Faculty of Engineer, University of Tokushima 2-1 Minamijosanjima, Tokushima,770-8506, Japan
[††] School of Information Engineering, Beijing University of Posts and Telecommunications Beijing 100876, China

Summary
It is obliged to provide an effective reading support system for CSL (Chinese as Second Language) learners, for recognition and comprehension of CIE (Chinese Idiomatic Expression) in Chinese text are very difficult for them. Though most current research are focusing on Chinese reading support, there is still no perfect system that can provide a convenient support aiming at CIE. In this paper, we mainly propose how to help CSL learners recognize and comprehend CIE. We have created a CIE database with 2,305 idiomatic expressions of contemporary Chinese. At the same time we have analyzed the basic structures and using forms of these CIE. By the analysis we have presented an extraction approach which based on rules and characters of these CIE. In our extraction experiment of CCE, the recall achieved 81.65% and the precision achieved 94.34%.

Key words:
CIE, CSL, reading support, phrase extraction

Introduction

The field of CALL (Computer-Assisted Language Learning) is inherently multidisciplinary. It applies research from the fields of second language acquisition, sociology, linguistics, psychology, cognitive science, culture studies, and natural language processing to second language pedagogy, and it melds these disciplines with technology-related fields such as computer science, artificial intelligence, and media/communication studies [1]. Recently, the evolution of L2 (second-language) learning environment has been accelerated by improved wireless telecommunication capabilities, open networks, battery technology and computing hardware (such as Tablet PC). With these technologies, effective L2 learning environment can easily be embedded in everyday life [2] [3]. Though most of L2 learners are not in target language environment, they can still easily obtain authentic learning materials online and conveniently make the language learning activities. But the challenge of L2 learning support system is not only to make language learning activities available at any time, at any place, and in any form, but specifically to provide valuable approach to resolve learning problems and difficulties as well. However, in the exploring process of L2 learning support system, we need grasp the attributes of target language, predict the potential difficulties to the learner and simultaneously provide a valuable approach to resolve these potential problems [4].

Understanding language involves recognition and access to not only individual words, but also to a lot of fixed expressions such as idiomatic expressions [5]. The significance and difficulties of idiomatic expressions understanding have been discussed in many previous studies. In L2 (second language) learning process, recognition and comprehension of idiomatic expressions are always major hurdles for foreign learners. Though L2 learners usually manage to express themselves in plain non-idiomatic language, this is only an expression strategy that learners fall back on when their linguistic means falls short of achieving their communicative ends [6]. Actually L2 learners have to face idiomatic expressions in many reading activities. An idiomatic expression is a group of words which, as a whole, has a unitary/figurative meaning that is different from the dictionary definitions of the individual words. Hence, the meaning of the idiomatic expression is not the sum total of literal meanings of the words taken individually. The comprehension of idiomatic expressions in fact requires learners to go beyond a simple word-by-word comprehension strategy and to integrate figurative meaning into contextual information [7]. If a L2 learner can not recognize and understand idiomatic expressions in received language information, he/she may not comprehend even misunderstand the real meaning of the information. Also most of idiomatic expressions have the strong culture background. Many idiomatic expressions reflect definite living consuetude, value judgments, thinking manners and so on. An ‘idiomatic expressions nonuse’ L2 learner may be regarded as someone lacking in relevant knowledge and cultures. The idiomatic expressions grasp can be considered as comprehension of a kind of culture. By idiomatic expressions learning, the L2 learners can learn not only language knowledge but also the target culture as well.

With rapid increase of international exchanges, Chinese learning has aroused more and more interesting for foreigners. As a matter of fact CIE (Chinese Idiomatic Expression) is a big obstacle for CSL (Chinese as Second Language) learners. It is obliged to comprehend CIE in

Manuscript received April 5, 2006.
Manuscript revised July 26 , 2006.

Chinese way. In this paper, we present a CIE Reading Support System for CSL learners. The aims of system as follows: Helping learners to recognize CIE in Chinese texts.

We have created a database with 2,305 CIE. In section 2, we have analyzed the basic structures and using forms of CIE in our database. In Section 3 recognition processing of CIE has been described. In section 4 we have mainly discussed the results of the CIE recognition experiments. Finally, we gave the conclusions and perspective of our future work.

2. Linguistic Forms Analysis of CIE

2.1 Analysis of Structure

The system we presented is based on a database of 2,305 contemporary CIE. By the statistical analysis of these 2,305 idiomatic expressions, we have found that contemporary CIE may consist of different character quantity. Generally, contemporary CIE consist of three Chinese characters to twelve Chinese characters. And we have also found the idiomatic expressions with three Chinese characters are in majority in total quantities of contemporary CIE—they account for 65.2% (1,503) of all. The idiomatic expressions with four Chinese characters take up the second large quantities—about 16.5% (381). Furthermore, the idiomatic 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 CIE only account for 0.4% (10). Hence, according to situation of quantities, processing on idiomatic expressions with three, four, and five characters is the most important work for our research.

Simultaneously, we have analyzed the structures and unitary attributes of idiomatic expressions. We have classified four categories of idiomatic 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 CIE is mainly described by verbal phrase expressions and noun phrase expressions. For the reason they account up to 86.7% of whole idiomatic expressions, in our research we have focused on these two kinds of idiomatic expressions.

2.2 Form Analysis of CIE in Applications

We have created a large idiomatic expressions database of 21,018 example sentences. By statistics and comparing, we have found that there are various expressional situations exist in CIE. The detail analysis is depicted in table 2.

We have found that the CIE may remain their primary forms in most of time. These phenomena take up 82.4% in whole quantities of idiomatic expressions. As the second largest phenomena of idiomatic expression, the ‘inserted words’ CIE 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’ idiomatic 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 idiomatic expression can be replaced by the other characters (usually replaced by single character verb). After replacing, the novel idiomatic expression will retain the original meaning as before. These kinds of idiomatic expressions account for 0.5% of whole quantities. Besides, ‘order changing’ is a frequent expressional phenomenon too. ‘Order changing’ indicates that the order of idiomatic 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 1: The Structures and Attributes of CIE

Categ.	No.	Structure categories	Examples	Q.	Freq.
Verbal Phrase	1	‘predicate-object’ structure	吃白飯(fathead，to be a ‘good-for-nothing’)	1,056	45.8
	2	‘adverb-headword’ structure	鷄蛋里挑骨頭 (be captious，marked by a disposition to find and point out trivial faults)	50	2.2
	3	‘continuous predicates’ structure	見便宜就搶(to take an advantage of every opportunity, to gain extra advantage unfairly)	114	4.9
	4	‘predicate complement’ structure	活得不耐煩(the act or process of destroying oneself or itself)	5	0.2
Noun Phrase	5	‘attribute-headword’ struc.	過街老鼠(the universally condemned person)	759	32.9
	6	‘parataxis’ structure	半斤八两(all the same)	12	0.5
	7	‘character De’ structure	拿笔杆子的(the intellect)	4	0.2
‘Clause’	8	‘subject-predicate’ structure	井水不犯河水(none may encroach upon the precincts of another)	172	7.5
	9	‘complicate’ structure	胳膊折了往袖子里藏(to endure an humiliation by one’s own self)	35	1.5
Others	10	‘special’ structure	三一三十一(to divide equally, share alike)	18	0.8
	11	‘comma separate’ structure	拆東墻，補西墻(keep up in one place at the expense of others)	80	3.5
Total	11			2,305	100

Table 2 Form Categories of CIE in Application

Form in using	Examples	Q.	Freq.
‘unchanged’ form	此地無銀三百两(a very poor lie which reveals the truth)	17,325	82.4
Form with ‘inserted word’	抓别人的辮子(to seize on other people’s mistake or failure)	3,116	14.8
‘the first character repeating’	戴戴高帽子(the vain compliments)	123	0.6
Form of ‘character replaced’	趕/打/拿鴨子上架(force someone to do something)	102	0.5
‘order changing’ form (with the other insert words)	捡着便宜了(to get a bargain, to gain an extra advantage) 便宜都让他捡着了(He has gained all the advantages.)	352	1.7
Total		21,018	100

Table 3 Categories of CIE with Inserted Words

Structure categories	Examples	Q.	Freq.
‘predicate-object’ structure	打了一回漂亮的翻身仗(changed completely)	2521	80.9
‘subject-predicate’ structure	架子非常大(arrogant, haughty)	327	10.5
‘attribute-headword’ structure	掌上的明珠(a parent refers affectionately to a beloved daughter)	236	7.6
‘parataxis’ structure	鷄毛和蒜皮(tiny things, bits and pieces)	32	1.0
Total		3116	100

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

3. Recognition Processing of CIE

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 [9]. From 90’s, many researchers have tried to use shallow parsing technique to Chinese processing. At the same time, statistic method is adopted frequently [10] [11] [12]. Unfortunately, using the method which is based on frequency or collocation can not succeed fully for CIE. Because of that: many CIE have too small frequency in a corpus and some CIE don’t have strong collocation. Thus in this paper, we have proposed a way to extract CIE based on rules and characters. The detail procedure is described as the next five steps.

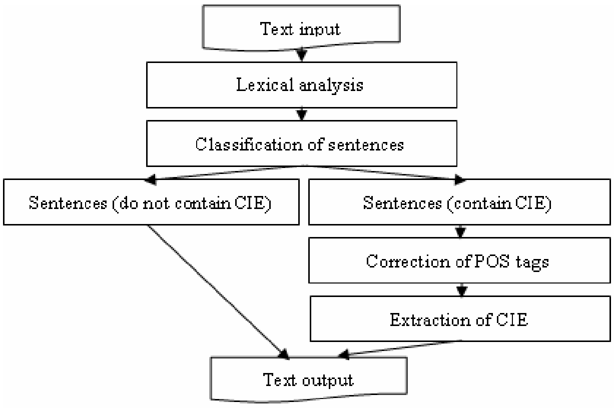

Fig. 1 The Recognition Process of CIE

3.1 Idiomatic Expression Registration

First of all we have registered almost all the idiomatic expressions we collected into the system. The registered CIE include not only the CIE themselves but also the structure, POS, and attributes of the CIE. All the CIE we put into register of the systems 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 CIE, the lexical analysis results of them will be output directly. The other part of sentences that contain CIE will be processed further. According to CIE in the sentences, we can divide the sentences with CIE into three types by matching the registration information of the system. The first type of sentences is that with ‘continuous words’ CIE. We can recognize these ‘continuous words’ CIE by using Maximum Matching Method. The second type of sentences is that with CIE of ‘comma separate’ structure. In this case the system will match the forward part of the comma, the comma, and the backward part of the CIE orderly. The third type of sentences is that with ‘Inserted Words’ CIE. The system will recognize these CIE 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 CIE, and at the same time if these word serials lap over each other, then the longest words serial will be taken as the candidate CIE in our system.

3.4 Correction of POS tags

By our test, we have found some POS tags of CIE are incorrect by using ICTCLAS. Therefore, we have corrected wrong POS tags of CIE 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 [13]. The TBL method can provide high precision of 98% and 95% for POS tags and chunk parsing respectively in English [14] [15]. 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 [16] [17]. There are three necessary requisite of using the TBL. 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:

(1) Initial POS Tags

We delete all the POS tags of CIE in training database then annotate them again by using ICTCLAS. The annotations of ICTCLAS are regarded as the basic results of processing.

(2) Generation of Candidature Rules

The candidature rules are generated from the wrong POS tags of CIE. Generation condition of the candidature rules is context environment basically.

(3) Acquisition of Transformation Rules

➢ *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 these results 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’.

➢ *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.

➢ *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:“他/r 像/v 一/m 只/q 出山/v 虎/n。” (He is brave and vigorous like a tiger.) Ex. 1 is the processing result of ICTCLAS. In this example, “出山/v 虎/n” and CIE of “出山/n 虎/n (NP+NP)” have the same word serial, but their POS tag of the word “出山” are different. According to fore-mentioned transformation rule, we can correct the POS tag of “出山” 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’ (idiomatic expression). This kind of transformation rule can be described as follows:

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

(‘P’ is POS tag. ‘K’ is a certain characters serial. ‘PN’ is the current POS tag. ‘DC’ is characters serial of registration CIE).

Ex. 2: “他/r 的/u 狐狸尾巴/I 被/p 抓住/v 了/y。” (his evil intention has been exposed.) In example 2, “狐狸尾巴” is a character serial of CIE. But its POS tag is not ‘cv’. Here, we transform its current POS tag from ‘i’ to ‘cv’.

The third kind of transformation rule is based on the characteristics of CIE.

Ex. 3: “大家/r 都/d 成/v 姥姥/n 不/d 疼/a， 舅舅/n 不/d 愛/v 了/y。” (Everyone has been ignored.) Example 3 contains the CIE—“姥姥/n 不/d 疼/a， 舅舅/n 不/d 愛/v”. But according to analysis result of ICTCLAS, the POS tag of the word ‘疼’ is ‘a’ (adjective). We can not correct its POS tag by syntactic environment in this sentence. In this occasion, we can correct its POS tag by using characteristics of CIE. Its transformation rule module can be described as follows:

```
    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 CIE

To insert “Tables” or “Figures”

(1) The Word with POS Tag ‘cv’

The word with POS tag ‘cv’ will be recognized as CIE. We can extract it directly.

Ex. 1: “那/r 件/q 事情/n 已/d [八九不離十/cv] 了/u。” (That thing is near success.)

(2) The CIE with ‘Comma Separate’ Structure

For the CIE with ‘comma separate’ structure, the extraction conditions are the ‘comma’, words serial and POS of every word. We can extract this kind of CIE from sentences by meeting these conditions.

Ex. 2: “他/r 一生/n 可以/v 叫/v [成/v 也/d 蕭/nr 何/nr, 敗/v 也/d 蕭/nr 何/nr] 。” (Both his success and failure are because of that man.)

(3) ‘Continuous Words’ CIE

The recognition knowledge of ‘continuous words’ CIE consist of attribute and context environments of the CIE. The attributes of CIE include the words, the order and the POS tags of the words. Environments of CIE indicate the attributes of the words those are located in front or back of the CIE.

➢ *‘Noun Phrase’ CIE*

There are three types of ‘noun phrase’ CIE in our system: ‘character De’ structure CIE, ‘attribute-headword’ structure CIE, and ‘parataxis’ structure CIE.

In the types of ‘character De’ structure only four CIE have been collected. The structures of them are: ‘verb + noun + (noun or proclitic word) + 的’. According to semantic relations, the CIE with ‘character De’ structure can be extracted by its attributes.

Ex. 3: “[拿/v 筆杆/n 子/k 的/b] 在/ 叫/v 他/r。” (He is an intellect.)

Comparing with ‘character De’ structure, the other two types of ‘noun-phrase’ CIE are far more complex. By analysis of the ‘noun-phrase’ CIE 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 CIE and ‘parataxis’ structure CIE by these expansion POS templates. For instance, the CIE ‘後勤部長’ can be extracted by No28 expansion template.

Ex. 4: “他/r 在/p 家/n 做/v [後勤/n 部長/n] 已/d 三/m 年/q 了/y。” (he has been a housekeeper for three years.)

➢ *‘Verbal Phrase’ CIE*

Four types ‘verbal phrase’ CIE have been collected in our system. There are ‘predicate-object’ structure CIE, ‘adverb-headword’ structure CIE, ‘continuous predicates’ structure CIE and ‘predicate-complement’ structure CIE. In this paper the ‘verb phrase’ CIE has been extracted by their structures, verb attributes or syntactic environment correspondingly.

Table 4 Expansion POS Templates

No.	P1	POS Temp.	P_1		No.	P1	POS Temp.	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
…	…	…	…		…	…	…	…

Some ‘verb phrase’ CIE, especially CIE 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 CIE 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 CIE, because the last word of them are intransitive verbs, so the boundary of these CIE can be recognized by attributes of the verbs(as ex. 8).

Ex. 5: “他/r 有/v 個/q [不/d 成器/v] 的/b 小子/n。” (He has a vain son.)

Ex. 6: “那/r 位/q 老兄/n 有/v 点/q [活/v 得/u 不耐煩/a] 了/y。” (It is seem that the man do not want to live anymore.)

Ex. 7: ‘李/nr 先生/n 是/v [吃/v 明/a 不/d 吃/v 暗/a]的/u 人/n。” (Mr. Li would rather fight face to face than infighting stealthily.)

Ex. 8: “部長/n 被/p [拿/v 下馬/v] 了/u。” (The Minister has been dismissed.)

It is very difficult to recognize the linguistic constituent of the last noun in some CIE 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 [18]. 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 CIE 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 CIE.

Ex. 9: “公司/n 需要/v 一/m 群/q 能/v [打/v 天下/n]的/u 人/n。” (The Company needs assiduous people.)

➢ *‘Clause’ CIE*

According to structural symmetry and semantic correlation of ‘Clause’ CIE, the boundary of most ‘Clause’ CIE can be recognized by their attributes directly (as ex. 10, 11).

Ex. 10: “社会/n 里/f 有/v 不少/m [大虫/n 吃/v 小虫/n] 的/b 事/n。” (In human society, there are many things operating as the law of jungle.)

Ex. 11: “不少/m 人/n 都/d [身/ng 在/p 福/n 中/f 不知/v 福/n] 。” (a lot of people neglect the happiness that they have owned.)

➢ *‘Special Structure’ CIE*

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

Ex. 12: “他/r [三/m 下/f 五/m 除/v 二/m] 就/d 把/p 那/r 件/q 事/n 干/v 完/v 了/y。 (He finished the work very quickly.)

(4) ‘Inserted Words’ CIE

➢ *‘predicate-object’ structure CIE*

‘predicate-object’ structure CIE 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’ CIE 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 CIE made of the DB and DY can finally be confirmed.

Based on analysis of Chinese verbal complement, the verbal complement can be classified into four types. They are possible (or impossible) complement, result complement, direction complement and movement complement.

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 ‘的’ 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 CIE 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 5 some POS templates of DY are described.

Ex. 13: “我/r [打/v 了/u 一個/m 漂亮/a 的/u 翻身仗/n] 。” (I have changed completely.)

Table 5 the Pos Templates of DY

No	POS template	No.	POS template
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	…	…

➢ *‘subject-predicate’ structure CIE*

The denial adverb and degree adverb are usually inserted in CIE with ‘subject-predicate’ structure. They can be recognized by composition rule of Chinese adverbial modifier. The extraction processing is similar to ‘Clause’ CIE which we mentioned before.

Ex. 14: “他/r [架子/n 非常/d 大/a] 。” (He is very arrogant.)

➢ *‘attribute-headword ’ structure CIE (noun+noun)*

CIE 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 ‘的’ or ‘之’. In this case, we can confirm the attributes of two parts which be separated and the inserted auxiliary word firstly. The boundary of the CIE can be recognized by the expansion template (same as expansion template of ‘noun phrase’ CIE.).

Table 6 Experimental Results of CIE Recognition

Categories of CIE	*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	82.62	95	88.37
Total	2,000	1,731	1,633	81.65	94.34	87.54

Ex. 15: "他/r 如同/v 一/m 只/q [過街/n 的/u 老鼠/n] 。(He is a universally condemned person)

- *'parataxis' structure CIE*

The 'parataxis' structure CIE with inserted part are usually made of nouns ('noun + noun'). The usual inserted words are paratactic conjunctions such as '和', '与', '又', '跟', '加', '同' and '对'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 CIE can be recognized by the expansion template (same as expansion template of 'noun phrase' CIE.).

Ex. 16: "[鶏毛/n 和/c 蒜皮/n] 的/u 小事/n。" (tiny things, bits and pieces)

4. Experiment of CIE Recognition

In current research 1,200 hypertext files (about 13,000 sentences and 2,000 CIE) 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.

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

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

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

In our experiment 1,731 multi-word units were extracted as CIE. Among them 1,633 were correct. We have achieved 81.65% in Recall, 94.34% Precision and 87.54% in F-measure. As table 8 shows. By the experiment results, we have found our extraction methods for CIE are successful. In this work the structural and semantic characters of CIE 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 CIE. The longer CIE is the more POS tag errors will be. So the extraction of long CIE is far more difficult than short ones. Thus the length of CIE must be taken in consideration in the future works. Besides, there were two causes leading the wrong extraction. One is boundary recognition of CIE. In this experiment, we successfully recognized the boundaries of 'noun phrase' CIE and 'verbal phrase' CIE. But the boundaries of some 'Clause' CIE were failed in recognition. The other cause is semantic judgment of CIE.

5. Conclusions and Future Works

In this paper, a CIE Reading Support System has been described, and an extraction method based on the rules and characters CIE has been presented. By the recognition experiment, the results of both Recall and Precision are over 80%. In the future, we will enlarge the analytical quantity of CIE. For improving recognition result of CIE, the length information, semantic analysis and the Particularity Words (name, address and thing etc.) of the CIE will be taken into consideration. Furthermore we will advance other reading support functions by relative techniques. Based on relevant researches of CIE, we will create a RDBMS of reading support resources. Besides a CIE comprehension support program which will connect the CIE recognition and reading support resources will be designed for effective comprehending CIE.

Acknowledgments

Our project was partly supported by The Education Ministry of Japan under Grant-in-Aid for Scientific Research B (No. 14380166) and the Outstanding Overseas Chinese Scholars Fund of the Chinese Academy of Sciences (No. 2003-1-1). We would like to thank all our colleagues participating in this project.

References

[1] Scholarly Activities in Computer-Assisted Language Learning, Development, Pedagogical Innovations, and Research, Joint Police Statements of CALICO, EUROCALL, and IALLT Arising from a Research Seminar at the university of Essen, Germany 30, April-1 May 1999.

[2] Lyytinen, K. and Yoo, Y.: Issues and Challenges in Ubiquitous Computing, CACM, Vol. 45, No.12, pp.63-65, 2002.

[3] Ogada, H., and Yano, Y.: How Ubiquitous Computing can Support Language Learning, Proc. of KEST 2003, pp.1-6, 2003.

[4] 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

[5] Dieter Hillert and David Swinney: The Processing of Fixed Expressions during Sentence Comprehension; Conceptual Structure, Discourse, and Language 4, SLI Press, Stanford, 1999

[6] Gass, S. and Selinker, L.: Language Transfer. In F. Eppert (ed) Transfer and Translation in Language Learning and Teaching. Singapore: SEAMEO.

[7] Maria Chiara Levorato, Barbara Nesi and Cristina Cacciari: Reading comprehension and understanding idiomatic expressions: A developmental study, Brain and Language 91 (2004) 303-314.

[8] Li Xinjian, Issues of Research and Standardization of ChineseIdiomatic Expressions, Applied Linguistics, 2002, 55-60

[9] C.Zong and F.Ren: "Chinese Utterance Segmentation in Spoken Language Translation", Computational Linguistics and Intelligent Text Processing, Ed. Alexander Gelbukh, Springer, LNCS2588, pp.516-525(2003)

[10] Patrick Pantel and Dekang Lin: "A Statistic Corpus-Based Term Extractor", Canadian Conference on AI2001, 36-46

[11] Tingting He, Jianzhou Liu, Donghong Ji: "A Statistical of Extracting Chinese Multi-Word Units", Journal of Chinese Language and Computing, 13(3)239-247, 2002

[12] Hideki Isozaki and Hideto Kazawa: "Speeding up Support Vector Machines for Named Entity Recognition", Information Processing Society of Japan, Vol.44, No.3, 2003

[13] Eric Brill: Transformation-Based Error-Driven Learning and Nature Language Processing: A Case Study in Part of Speech tagging, CLV21, 1995

[14] Eric Brill, Some Advances in Transformation-based Part of Speech Tagging, In: Proceedings of the Twelfth National Conference on Artificial Intelligence, 722-727, 1994

[15] Voutilainen. Atro: "NLPTool, a Detector of English Noun Phrases", In Proceedings of the Workshop on Very Large Corpora, ACL, 48-57, 1993

[16] 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

[17] Jun Zhao, Changning Huang: "A model based on transformation to recognize Chinese base-NP", Journal of Chinese Information Processing, Vol.13, No.2, 1999

[18] Yuehua Liu, Wenyu Pan and Wei Gu: "Chinese Grammer", Foreign Language Teaching and Research Press, 1986

[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

[21] 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

[22] Shiwen Yu: "specification for the Prase Structure Knowledge-based of Contemporary Chinese", Journal of Chinese Language and Computing, 13(2), 215-226, 2003

[23] 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)

[24] 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

[25] Bobrow, Samuel A and Susan M Bell: On Catching on to idiomatic expressions. Memory and Cognition 1, 343-346, 1973

[26] Reka Benczes: The semantics of idioms: a cognitive linguistic approach. The Even YearBook 5, 17-30, 2002

[27] Cacciari C., Glucksberg S.: Understanding idiomatic expressions: the contribution of word meanings. In G. Simpson ed. Understanding word and sentence, Amsterdam: North Holland, 217-240, 1991

[28] Gibbs Jr., Raymond W.: The poetics of mind: Figurative Thought Language and Understanding. Cambridge: Cambridge University Press, 1994

[29] She Xianjun, Song Ge, Zhang Biyin: The Effects of Predictability and Swmantic Bias in Idiom Comprehension. Acta Psychologica Sinica. Vol. 32 No. 2 P.203-209, 2000

[30] Carrell, P. and Eisterhold J. C.: Schema Theory and ESL Reading Pedagogy. TESOL Quarterly, 17(4), 553-573.

[31] Goodman, K.: On Reading: A Common-sense Look at the Nature Language and the Science of Reading. Portsmouth , NH: Heinemann

Shuan Xiao received the B.S. and M.S. degrees in Hunan Normal University and Kazan Technology University in 1994 and 2002 respectively. His current research interests are in NLP and E-learning. Now, he is completing his Ph.D. degree in Tokushima University of Japan.

Hua Xiang received the M.S. degrees in Kazan Technology University in 2003. Currently, she is completing her Ph.D. degree, and devoting the applied research of emotional information processing in Tokushima University of Japan.

Fuji Ren received the Ph.D. degree in 1991 from Faculty of Engineering, Hokkaido University, Japan. He worked at CSK, Japan, where he was a chief researcher of NLP. From 1994 to 2000, he was an associate professor in the Faculty of Information Sciences. From 2001 he joined the faculty of engineering, the University of Tokushima, Japan. His current research interests include NLP, Artificial Intelligence and Affective Computing.

Kuroiwa Shingo: received Ph.D. degree in electro-communications from the University of Electro Communications Tokyo, Japan, in 1986, 1988, and 2000 respectively. From 1988 to2001 he had been a researcher at KDD R&D Laboratories. Since 2001, he has been with the faculty of Engineering, Tokushima University, Japan. His current research is interests included Speech Recognition, NLP and Information Retrieval.