Computational Linguistics and Chinese Language Processing
Vol. 11, No. 3, September 2006, pp. 245-280

A Probe into Ambiguities of Determinative-Measure Compounds

Shih-Min Li*+, Su-Chu Lin*, Chia-Hung Tai*, and Keh-Jiann Chen*

Abstract

This paper aims to further probe into the problems of ambiguities for automatic identification of determinative-measure compounds (DMs) in Chinese and to develop sets of rules to identify DMs and their parts of speech. It is known that Chinese DMs are identifiable by regular expressions. DM rule matching helps one solve word segmentation ambiguities, and parts of speech help one improve sense recognition and part-of-speech tagging. In this paper, a deep analysis based on corpus data was studied. With analyses of error identification and disambiguation of DM compounds, the authors classified three types of ambiguities, *i.e.* word segmentation, sense, and pos ambiguities. DM rules are necessary complements to dictionaries and helpful to resolve word segmentation ambiguities by applying resolution principles and segmentation models. Sense and pos ambiguities are also expected to be resolved by different approaches during postprocessing.

Keywords: Ambiguities, Word Segmentation Ambiguities, Sense Ambiguities, Part-of Speech Ambiguities, Determinative-Measure Compounds

1. Introduction

To a speaker of English, one of the most striking features of the Mandarin noun phrase is the classifier. A classifier is a word that must occur with a number and/or a demonstrative, or certain quantifiers before the noun [Li and Thompson 1981]. Furthermore, Li and Thompson [1981] assert that any measure word can be a classifier, so the combination of demonstrative and/or number or quantifier plus a classifier or a measure is defined as a classifier phrase or a measure phrase. For example, *san ge* in *san ge ren* 'three people' (三個人), *zhe zhan* in *zhe zhan deng* 'this lamp' (這盞燈), *ji jian* in *ji jian yifu* 'a few / how many garments' (幾件衣服), *liu li* in *liu li lu* 'six miles of road' (六里路), *na jin* in *na jin yangrou* 'that tael of lamb'

* Institute of Information Science, Academia Sinica, Taipei
E-mail: {shihmin, jess}@hp.iis.sinica.edu.tw; {glaxy, kchen}@iis.sinica.edu.tw
+ Graduate Institute of Linguistics, National Chengchi University, Taipei
E-mail: 95555501@nccu.edu.tw

[Received February 16, 2006; Revised November 27, 2006; Accepted November 28, 2006]

(那斤羊肉) and *ji gang* in *ji gang cu* ‘a few / how many vats of vinegar’ (幾缸醋) are classifier phrases / measure phrases, which are regarded as Determinative-Measure (DM) compounds in Chao [1968]. A determinative (D) and a measure normally make a compound with unlimited versatility and form a transient word of no lexical import [Chao 1968]. Although the demonstratives, numerals, and measures may be listed exhaustedly, their combination is inexhaustible. It is impossible to list thoroughly all combinations of DMs in dictionaries. Therefore, it requires a representational model to express DM compounds in Chinese NLP.

In Chinese, word segmentation, sense, and pos (*i.e.* part-of-speech) ambiguities commonly occur in certain constructions of DMs or DM-like structures. For examples:

(1) *三個*月餅舖的銷售量

a. *sange yuebingpu de xiaoshouliang*
three-M moon-cake store DE sales volume
three moon-cake stores’ sales volume

b. *sangeyue bingpu de xiaoshouliang*
three-M months cake store DE sales volume
the cake store’s three-month sales volume

(2) 取*此名*

a. *qu ciming*
choose this-M
choose this one (person)

b. *qu ci ming*
name this name
name it this name

(3) *二十五年*的審核、排隊、等待

a. *ershiwunian de shenhe paidui dengdai*
twenty five-M DE examine line up wait
examining, lining up and waiting in the year of twenty five

b. *ershiwunian de shenhe paidui dengdai*
twenty five-M DE examine line up wait
examining, lining up and waiting for twenty five years

The DM compound *sange* in example (1) modifies moon-cake stores as well as months. The string *sangeyuebingpu* can be segmented into either *sange yuebingpu* or *sangeyue bingpu*, which has word segmentation ambiguity. In example (2), *ming* functions either as a measure or as a noun. Although example (2) has two meanings and is sense ambiguous, the roles assigned to *ciming* in (2a) and (2b) are both the goal. In example (3), *ershiwunian* is a time referent, and it can either be a time point specifying the event-time of the verb or denote the period of time delimitating the time length of the event. In Sinica Treebank, no matter whether *ershiwunian* behaves as a time point or a time length, *ershiwunian* is tagged as a unit. When *ershiwunian* behaves as a time point, its pos is Ndaad and semantic role is time[1]. Furthermore, when *ershiwunian* behaves as a time length, its pos is DM and semantic role is duration. However, according to CKIP's word segmentation standard of Sinica Corpus, the temporal and locative DM structures are combined together when the meaning of the structure is not obtained from the composition of the components of the structure. Therefore, the temporal DM *ershiwunian* in (3a) is combined as a unit and tagged as Nd in Sinica Corpus while that in (3b) is segmented into two units, *i.e. ershiwu* and *nian*, and tagged as Neu and Nf individually. Therefore, example (3) has pos ambiguity. The different degrees of ambiguities are shown in examples (1) to (3).

In this paper, we examine and analyze Mandarin Chinese DMs in Sinica Corpus and Sinica Treebank. In section 3, we introduce the regular expression approach to identify DMs and their poses. In section 4, we make a study of the structures and ambiguities of DMs, and then try to analyze and disambiguate these DMs. Section 5 is for implementation and evaluation.

2. Literature Review

To deal with DMs, first one must give a proper definition to DMs. Thus, one can delimit the scope of the discussion. There are numerous discussions on determinatives as well as measures, especially on the types of measures.[2] The classification of measures is beyond the scope of this paper. To avoid confusion between classifiers and measures, one must pay attention to the relation between them. Tai [1994] asserts that in the literature on general grammar as well as Chinese grammar, classifiers and measures words are often treated together under one single framework of analysis. Chao [1968] treats classifiers as one kind of measure. In his definition, a measure is a bound morpheme which forms a DM compound with

[1] All the symbols such as Ndaad will be defined in the appendix of this paper.

[2] Chao [1968] and Li and Thompson [1981] detect measures and classifiers. He [2000] traces the diachronic names of measures and mentions related literature on measures. The dictionary of measures pressed by Mandarin Daily News Association and CKIP [1997] lists all the possible measures in Mandarin Chinese.

one of the determinatives enumerated above [Chao 1968]. Classifiers are defined as 'individual measures', which is one of the nine kinds of measures. As was mentioned in the section of introduction, Chao considers that determinatives are listable and measures are largely listable so D and M can be defined by enumeration, and that DM compounds have unlimited versatility. However, Li and Thompson [1981] blend classifiers with measures. They conclude that, not only does a measure word generally not take a classifier, but also any measure word can be a classifier. In Tai's opinion [1944], in order to better understand the nature of categorization in a classifier system, it is not only desirable but also necessary to differentiate classifiers from measure words. These studies on the distinction between classifiers and measures are not very clear-cut. In this paper, we discuss ambiguities of DMs in NLP as well as adopt the CKIP DM rules and symbols of poses, and therefore inherit the definition of determinative-measure compounds (DMs) in Mo *et al.* [1991]. Mo *et al.* define a DM as the composition of one or more determinatives together with an optional measure. The definition of Mo *et al.* is used to apply to NLP and somewhat different from traditional linguistics definition.

As for ambiguity, Crystal [1991] specifies that the general sense of ambiguity is a word or sentence which expresses more than one meaning. The most widely discussed type of ambiguity in recent years has been grammatical (or structural) ambiguity. In the structure *new houses and shops*, it could be analysed either as *new* [*houses and shops*] (*i.e.* both are new) or [*new houses*] *and shops* (*i.e.* only the houses are new). Furthermore, according to Crystal's assertion, ambiguity which does not arise from the grammatical analysis of a sentence, but is due solely to the alternative meanings of an individual lexical item, is referred to as lexical ambiguity, *e.g. I found the table fascinating* (= 'object of furniture' or 'table of figures'). Moreover, the definition of structural and lexical ambiguities can be referred to Prins [2005]. Prins mentions if one restricts his or her attention to the syntax in texts, then one may focus on ambiguity in two forms. The first is lexical ambiguity, the second is structural ambiguity. Lexical ambiguity arises when one word can have several meanings. Structural ambiguity arises when parts of a sentence can be syntactically combined in more than one way. Prins believes humans can resolve most ambiguity, of both types, without even being consciously aware of the alternatives. The remaining ambiguity, of which we are aware, is resolved when knowledge about the world is used in combination with what is known about the linguistic context of the ambiguity to arrive at the most likely analysis. Jurafsky and Martin [2000] define ambiguity as a sentence or words which can have more than one parse. Deciding which category a word belongs to can be solved by part-of-speech tagging. Deciding what sense a word has can be solved by word sense disambiguation. Resolutions of part-of-speech and word sense ambiguities are two important kinds of lexical disambiguation [Jurafsky and Martin 2000]. Furthermore, structural ambiguity occurs when the grammar assigns more than

one possible parse to a sentence. Structural disambiguation / syntactic disambiguation can be addressed by probabilistic parsing. Three particularly common kinds of structural ambiguity are attachment ambiguity, coordination ambiguity, and noun-phrase bracketing ambiguity [Jurafsky and Martin 2000]. In the following analysis, we find that Prins' division of ambiguity into structural ambiguity and lexical ambiguity is not enough to deal with ambiguities in NLP. One must apply Jurafsky and Martin's classification to obtain more detailed discussion on ambiguities of DMs in NLP. Word segmentation ambiguity caused by different segmentation of words is a kind of structural ambiguity. With the same word segmentation, that string of words may still be ambiguous because the string may either have more than one meaning or have different parts of speech, semantic roles and functions. Therefore, weone have sense ambiguity and pos ambiguity, which are the two subtypes of lexical ambiguities of DMs.

3. Regular Expression Approach for Identifying DMs

In this section, we introduce the regular expression approach to identify different types of DMs, their representational rules and their poses. Since this paper focuses on the DM defined in Mo *et al*., which is the composition of one or more determinatives together with an optional measure, the DM structure includes prototypical DMs, variant forms of DMs (*e.g*. the ellipsis of the determinative and the insertion of an adjective into a DM[3]), reduplicative forms of DMs (*e.g.* the reduplication of 'M', 'DM' or 'AM'), and forms of the numeral *yi* preceding the reduplicative measures (*e.g.* '*yi*MM' (一 MM) and '*yi*AM' (一 AM)). Due to the infinite of the number of possible DMs, Mo *et al.* [1991] proposed identification of DMs by regular expression before parsing as part of their morphological module in NLP. For example, when the DM compound is the composition of one determinative, *e.g.* numerals in (4), rules (5a), (5b) or (5c) will first apply, and then rules (5d), 5(e) or (5f) will apply to compose complex numeral structures, and finally rules (5g) will apply to generate the pos Neu of numeral structures. From the processes of regular expression, the numerals 534 and 319 in (4) are auto-tagged as Neu.

(4) 鼓勵***534***人完成***319***鄉之旅

guli	*wubaisanshisi*	*ren*	*wancheng*	
encourage	five hundred thirty four	person	accomplish	
sanbaiyishijiu	*xiang*	*zhi*	*lu*	
three hundred and nineteen	village	DE	travel	

encourage 534 persons to accomplish the travel around 319 villages

[3] The insertion of an adjective into a DM has the form of '*yi*AM'.

(5) a. NO1 = {○,一,二,兩,三,四,五,六,七,八,九,十,廿,卅,百,千,萬,億,兆,零,幾};

b. NO2 = {壹,貳,參,肆,伍,陸,柒,捌,玖,拾,佰,仟,萬,億,兆,零,幾};

c. NO3 = {１,２,３,４,５,６,７,８,９,０,百,千,萬,億,兆};

d. IN1 -> { NO1*, NO3*} ;

e. IN2 -> NO2* ;

f. IN3 -> {IN1,IN2} {多,餘,來,幾} ({萬,億,兆}) ;

g. Neu -> {IN1,IN2,IN3,IN4,IN5,DN} ;

Regular expression approach also applies in dealing with ordinal numbers, decimals, and fractional numbers. The ordinal number *diyi* in (6) applies rules (9) and (5g) so that it is regarded as a unit and tagged as Neu. Rules (10) and (5g) apply to decimals such as *sandianyi* in (7). Therefore, decimals are viewed as one unit and tagged as Neu. Depending upon the forms of fractional numbers, rules (11a) or (11b) apply to fractional numbers like *sanfenzhiyi* in (8) and are treated as single units. Then, after application of rules (11a) or (11b), rule (11c) applies to the fractional numbers. Thus the fractional numbers are tagged as Neqa.

(6) 假學術活動中心***第一***會議室召開

jia xveshu huodung zhongxin diyi hueiyishi zhaokai

at academic activity center first auditorium convene

convene at the first auditorium in the Center for Academic Activities

(7) 得分只有***三點一***

defen zhiyou sandianyi

point only three point one

get only 3.1 points

(8) 等於是一日的***三分之一***

dengyu shi yiri de sanfenzhiyi

equal SHI one day DE one third

is equal to one day of third

(9) IN5 -> {第} {IN1,IN2} ;

(10) DN -> IN1 { • , . ,·, · ,點} IN1 ;

(11) a. FN1 -> (IN1 {又}) IN1 {分之,一, / } {IN1,DN} ({強,弱}) ;
b. FN2 -> {DN,IN1} {％};
c. Neqa -> {NE5a,WQ,QQ,DQ1,DQ2,PQ,FN1,FN2,FN3,NOP3,RD13} ;

Below, DM structures of classes are also rule design and poses are identified by rule types.

(12) 北市文昌國小***五年一班***

*a. *beishi* *wenchang* *guoxiao* *wunian* *yiban*
Taipei City Wen-Chang elementary school five-M one-M
b. *beishi* *wenchang* *guoxiao* *wunianyiban*
Taipei City Wen-Chang elementary school Fifth Grade Class One
the Fifth Grade Class One in Taipei Wen Chang Elementary school

(13) a. CNP -> IN1 {年} {IN1,ON} {班} ;
b. Ncb -> {NC1,NC2,NC3,CNP,DSP1} ;

DM rules will generate/identify ambiguous DM compounds, *e.g.* (12). If *nian* and *ban* are regarded as measures, *wunianyiban* is segmented into two DMs, *i.e. wunian* 'five years' and *yiban* 'one run', like (12a). Therefore, when identifying DMs treated as classes, one has to apply the resolution principles and two DM rules (13a) and (13b). Then, classes in schools such as (12b) are viewed as a single unit, and the noun phrase *wunianyiban* is restricted to be a unit with the pos Ncb. The word segmentation algorithm will reduce semantic anomalies resulting from possible parses of sentences.

When dealing with addresses, especially indicating the floor, number, alley, lane, section and neighbourhood, we also adopt a regular expression approach for identification. The following instances show the same forms with different segmentation between DMs and addresses.

(14) 日前遷至台北市信義路*三段七號三樓之一*

riqian qian zhi taibeishi xinyilu sanduan qihao sanlouzhiyi
a few days ago move to Taipei City XinYi Rd. Sec. 3 No. 7 3F-1
a few days ago, moved to 3F-1, No. 7, XinYi Rd., Sec. 3, Taipei

(15) 行經屏市長安里竹圍巷*一之一0二號*時

xing jing pingshi changanli zhuweixiang yizhiyilingerhao shi
Go through Pingtung City Changan Village Zhuwei Lane No. 1-102 as
when going through No. 1-102, Zhuwei Lane, Changan Village, Pingtung City

(16) a. NC1 -> IN1 {鄰,巷,弄,段,號,樓};
b. NC2 -> IN1 {樓,號} {之,－} IN1 ;
c. NC3 -> IN1 {之,－} IN1 {號} ;

Normally, DMs such as *sanlouzhiyi* and *yizhiyilingerhao* are segmented into several units. The former is segmented into three units, *i.e. sanlou* 'the third floor', *zhi* 'DE' and *yi* 'one' while the latter is segmented into three units, *i.e. yi* 'one', *zhi* 'DE' and *yilingerhao* 'No. 102'. However, according to CKIP Technical Report 96-01 (1996: 50), the determinative measure structures expressing time and location will be combined together as a unit. The reason why the locative DMs are combined is because the first joint principle of segmentation stipulates that when the meaning of a string of words is not obtained from the composition of these components, this string should be segmented as a unit. Consequently, in (14), DM rule (16a) applies to *sanduan* and *qihao*, and DM rule (16b) applies to *sanlouzhiyi*. In (15), *yizhiyilingerhao* applies DM rule (16c). Then *sanduan*, *qihao*, *sanlouzhiyi* and *yizhiyilingerhao* are all processed by the application of DM rule (13b). Thus *sanduan*, *qihao* and *sanlouzhiyi* in (14) and *yizhiyilingerhao* in (15) are all segmented as a single unit and tagged as Ncb, not DM, in Sinica Treebank.

To deal with the reduplicative measures, *e.g.* '*yi*MM' (一 MM) and 'MM' (MM), we also adopt a context-sensitive regular expression to identify them. For example, *zhongzhong* in (17) and *yizhangzhang* in (18) are regarded as a single unit. First, the context-sensitive DM rule (19) is applied, where two measure words in MM are restricted to be equal, and then rule (11c), so the form of reduplicative measures with a preceding optional *yi* is tagged as Neqa.

(17) ***種種***問題

zhongzhong wenti

all kinds of question

all sorts of questions

(18) ***一張張***海報

yizhangzhang haibao

sheets of poster

sheets of posters

(19) RD13 -> ({一}) M M ;

From the examples and rules illustrated above, one knows that the regular expression approach helps people identify certain DMs. However, DMs still have ambiguities.

4. Ambiguities of DMs

The adoption of DMs rules really improves the recall of recognition, but one still has to resolve segmentation, sense, and pos ambiguities of DMs as shown in the preceding examples in Section 1.

4.1 Word Segmentation Ambiguities of DMs

There are two types of word segmentation ambiguities, *i.e*. covering ambiguities and overlapping ambiguities.

Type1: Covering ambiguities

(20) ***一服***藥就見效

a. *yi fuyao jiu jianxiao*

one take medicine then take effect

Every time when he takes medicine, the illness is completely cured.

b. *yifu yao jiu jianxiao*

one-M medicine then take effect

One dose is effective.

(21) 他們家*四口*人

*a. *tamen jia si kou ren*
they family four mouth person

b. *tamen jia sikou ren*
they family four-M person
Their family has four members.

Covering ambiguities are always associated with sense ambiguities, since different segmentations result in different sense interpretations. Goh *et al*. [2005] mention that the covering ambiguity is defined as follows: For a string w = xy, x ∈ W, y ∈ W, and w ∈ W. As almost any single character in Chinese can be considered a word, the above definition reflects only those cases where both word boundaries .../xy/... and .../x/y/... can be found in sentences. Example (20) is ambiguous in its meaning and has two different segmentations. When *fu* functions as a verb, *fuyao* is segmented as a unit, and the meaning of (20) is (20a). When *fu* functions as a measure, *yifu* is tagged as DM, and the meaning of (20) is (20b). Because the combinations of determinatives and measures are countless, DMs such as *yifu* won't be listed in the CKIP dictionary. Different word segmentation will bring structural ambiguities forth. Another segmentation ambiguity exists in the ellipsis of the determinatives. In (21), *kou* is sense ambiguous in that it functions as a noun or as a measure. When *kou* behaves as a noun in (21a), the sentence is semantically anomalous and syntactically ungrammatical. Only when *kou* functions as a measure, will the sentence (21b) be well-processed.

Type 2: Overlapping ambiguities

(22) *一串*串珠飾品

a. *yichuan chuanzhu shipin*
one-M beads accouterment
one string of beads

b. *yichuanchuan zhushipin*
several strings beady crystal
several strings of beady crystals

(23) 媽媽講了***一個***月亮的故事

a. *mama jiang le yige yueliang de gushi*
mother tell LE one-M moon DE story
Mother told a story about the moon.

b. *mama jiang le yige yue liang de gushi*
mother tell LE one-M month Liang DE story
Mother had told for a month about Liang's story.

(24) ***第***一派對分三部分

a. *diyi paiduei fen san bufen*
first party divide three part
The first party is divided into three sections.

b. *diyipai dueifen san bufen*
first group dichotomize three part
The first group dichotomized into three parts.

Goh *et al.* [2005] state that overlapping the ambiguity is defined as follows: For a string w = xyz, both w1 = xy ∈ W and w2 = yz ∈ W hold. Mo *et al.* [1991] list a resolution principle to reduce word segmentation ambiguity. The principle asserts if ambiguous word breaks occur between the words in the lexicon and the DMs, the words in the lexicon should have higher priority to get the shared characters. Therefore, (22a), (23a) and (24a) have higher priority to (22b), (23b) and (24b), individually.

According to the above discussions, to resolve word segmentation ambiguities, we propose the following resolution principles which were implemented in the word segmentation system of Ma and Chen [2003].

a) DM compounds are expressed and matched by regular expressions.
b) Lexical words have higher precedence than DM compounds (cf. 22, 23, 24).
c) Longest matching principle (*i.e.* 9, 10, 11, 13, 16, 19): a long DM has higher precedence than a short DM (*cf.* 6, 7, 8, 12 14, 15, 17, 18).
d) Covering ambiguities are resolved by collocation context (cf. 20, 21).

The word segmentation ambiguity is caused by different possible segmentations. Although the above examples have ambiguities, after the application of the resolution principles, the ambiguous segmentation is resolved and the correct segmentation has higher

priority.

4.2 Sense Ambiguities of DMs

Senses and semantic functions of DMs are sometimes related to the types of measures. The sense of certain types of DMs can be identified by types of measures. However, as usual, some DMs have ambiguous senses. Their ambiguity resolution is almost equivalent to word sense disambiguation. Therefore, context sensitive rules and collocation bi-grams are suitable information for resolving sense ambiguities. Methods for word sense disambiguation are also applicable for DMs.

As mentioned in Section 3, we have adopted a regular expression approach for identifying structures denoting addresses. For example, DM rule (16a) is applied to segment *yiduan*, *bahao*, and *yilou* in (25) as single units, and each of them is tagged as Ncb. However, in (26), *hao* and *duan* are both measures specifying the fixed amount or quantity of the road, so *yiyiqihao* and *yiduan* are both tagged as DM. Examples (25) and (26) have sense ambiguities during the processes of DM recognition, so they will be automatically segmented as (25a) and (25b) as well as (26a) and (26b), individually. According to CKIP Technical Report 96-01 (1996), locative DMs are combined so that one can disambiguate (25) and (26). Then, the correct segmentations (25b) and (26b) are resumed according to their contexts.

(25) 羅斯福路*一段八號一樓*

*a. *luosifulu* *yiduan* *bahao* *yilou*
Roosevelt Rd. one-M eight-M first floor

b. *luosifulu* *yiduan* *bahao* *yilou*
Roosevelt Rd. Sec. 1 No. 8 first floor
1 F, No. 8, Roosevelt Rd., Sec. 1

(26) 屬於*１１７號*公路的*一段*

*a. *shuyu* *yiyiqihao* *gonglu* *de* *yiduan*
belong No. 117 road DE Sec. 1

b. *shuyu* *yiyiqihao* *gonglu* *de* *yiduan*
belong 117-M road DE one-M
belong to a part of the 117th road

Similar to dealing with addresses, sense ambiguities occur when one refers to percentages. One can adopt two forms to represent percentage, *i.e.* Chinese characters in (8) and mathematical symbols in (27). Regular expression approach can help one identify word segmentation ambiguity existing in (8). The form of the mathematical symbols has sense ambiguity, which can refer to either percentage like (27) or a time point like (28). Without any context, the symbol "１０／２１" can be identified either as a fractional number and read as "*ershiyifenzhishi*" (二十一分之十 'ten over twenty one') with the pos Neqa or as a time point and read as "*shiyueershiyiri*" (10 月 21 日 'Oct. 21') with the pos Ndabd whose semantic role is time. Besides, the form of mathematical symbols is also used to refer to another kind of time point, such as (29). To reduce such kinds of sense ambiguities caused by mathematical symbols, DM rules (30a), (30b), (31a) and (31b) exist. DM rules (30a) and (30b) apply to the forms of mathematic symbols like (27) tagged as Neqa (numbers), while DM rules (31a) and (31b) apply to forms like (29) tagged as Nd (time point). Although DM rules (30) and (31) help disambiguate sense ambiguities between forms of mathematic symbols denoting percentage and time points such as (27) and (29), one still has to have context to make (27) and (28) distinguishable.

(27) ４０分的佔了***2／3***

sishifen de zhan le sanfenzhier

40-M DE occupy LE two-thirds

Those of forty points occupy two-thirds.

(28) ***１０／２１***召開全院網路工作小組第三次會議

shiyuershiyiri zhaokai quan yuan wanglu gongzuo xiaozu disanci huiyi

Oct. 21 convene whole faculty network work group third conference

convene the third conference of the network group on Oct. 21

(29) ***2005／06／30***更新

2005／06／30 gengxin

June 30, 2005 update

update on June 30, 2005

(30) a. NE5a -> {NE2} {—,／} {NE2} ;

b. Neqa -> {NE5a,WQ,QQ,DQ1,DQ2,PQ,FN1,FN2,FN3,NOP3,RD13} ;

(31) a. NE5b -> {NE2} {—,／} {NE2} {—,／} {NE2} ;

b. Nd -> {Ndabe,NE5b,ND3,ND5}

Take Chao's [1968] example to help disambiguate ambiguities similar to those between (27) and (28). As Chao discusses, the form *Guangxu sanshisinian* (光緒三十四年) can be either the phrase 'the thirty-fourth year of Guangxu (*i.e.*, 1908)' or the sentence 'Guangxu's reign was thirty-four years [long].' Chao believes that, in most cases, the context will resolve the ambiguity. The following examples in Sinica Corpus have similar ambiguities to those Chao mentions.

(32) 經過卡斯楚*三十年*的統治之後

jingguo kasichu sanshinian de tongzhi zhihou

after Castro 30-M DE governance afterward

after Castro's thirty-year governance

(33) *三十年*秋，緝私總隊復正名爲稅警總團

sanshinian qiu qisi zongdui fu zhengmingwei shuijingzongtuan

the year of thirty autumn anti-smuggling team again rectify tax policemen team

In the autumn in the year of thirty, the anti-smuggling team is rectified to the tax policemen team.

Examples (32) and (33) have the same temporal phrase *sanshinian*, but their senses, functions and roles are different. The temporal phrase in (32) denotes a time length and is tagged as DM. The semantic role of *sanshinian* in (32) is duration. However, *sanshinian* in (33) indicates a time point and is tagged as Ndaad, whose semantic role is time. Even though example (33) omits either a Chinese reign title or *Mingguo* (民國) preceding *sanshinian*, we still know *sanshinian* is a specific time, not a period, from the context. When the measures *nian* 'year' (年) and *ri* 'day' (日) are preceded by numerals, the temporal phrases always have sense ambiguities. We have two tricks to differentiate between time points and time lengths. If DMs

denote time points, they are usually preceded by key words of *Mingguo*, the Christian era like *Gongyuan* (公元) and *Xiyuan* (西元), or a Chinese reign title such as *Guanxu* (光緒), *Qianlong* (乾隆), *Tianbao* (天寶), *Jiajing* (嘉靖) and so on. When DMs are preceded by these key words, they are tagged as Ndaad (*i.e.* a time point). Another trick that helps one to recognize DMs as time points is their neighbouring temporal phrases. Time points, not time lengths, are usually combined with two or three co-occurring temporal phrases, *e.g.* *erlinglingwunian liuyue* (2005 年 6 月 'June 2005'), *liuyue sanshiri* (6 月 30 日 'June 30'), *erlinglingwunian liuyue sanshiri* (2005 年 6 月 30 日 'June 30, 2005'), etc. The two tricks mentioned above and context will help one reduce some ambiguities of phrases and mathematical symbols specifying time.

In conclusion, sense ambiguity resolution of DMs is almost equivalent to word sense disambiguation. Therefore, context sensitive rules and collocation bi-grams are suitable information for resolving sense ambiguities. Methods for word sense disambiguation are also applicable here.

4.3 POS Ambiguities of DMs

Here, we first discuss ambiguities about temporal adverbs to illustrate pos ambiguities and possible resolution methods. The temporal phrases in (34) and (35) are regarded either as time points or as time lengths. These temporal phrases have the same strings and word segmentation but have different parts of speech, semantic roles and functions. As is known, in Chinese a folktale, a woman called *Wang Baochuan* (王寶釧) went through 18 years of hardship for her husband's turning back home. In Mainland China, the Tiananmen Square massacre occurred in 1989. Therefore, the semantic roles of the temporal phrases *shibanian* in (34) and *bajiunian* in (35) will be labelled as duration and time individually. The pos of the former is DM while that of the latter is Ndaad. The reason for making different assignments of semantic roles is concerned with logical interpretation of sense collocations according to common sense and the real world knowledge.

(34) 18年的苦守

shibanian de kushou

18-M DE wait bitter

wait bitter for eighteen years

(35) 89年的反抗

bajiunian *de* *fankang*

the year of 89 DE revolt

the revolt in the year of 89

When detecting DMs in Sinica Corpus and Sinica Treebank, one finds some interesting examples. The verb phrases (36) and (37) have the same lexical items except for their linear word order. The pos of the DM structure in (36) is DM whose semantic role is duration while that in (37) is Ndaad whose semantic role is time. It seems that different positions of temporal DMs will affect the meanings of sentences. Therefore, we briefly reviewed the data in Sinica Treebank. The totality of the semantic role time of NPs and of PPs following a verb is close to that of the semantic role duration. But the totality of the semantic role time of NPs and of PPs preceding a verb is much more than that of duration. The statistics indicate that temporal DMs preceding verbs mostly function as time. Another problem with assignment of semantic role to a similar structure is illustrated by (38) and (39). The DM structure in (38) is tagged as DM and assigned the semantic role duration while, in (39), it is tagged as Ndaad and assigned the semantic role time. This kind of pos ambiguity has relation to situation types. The situation type of *fuxing* in (38) is an Activity while that of *panxing* in (39) is an Achievement. The feature [±Durative][4] of the events causes differences. As for the pos ambiguity of *yixi*a, *yixia* in (40) means 'for a while', which is tagged as Nddc and assigned the role of duration. However, *yixia* in (41) means 'once', which is tagged as DM and assigned the role of frequency. Nevertheless, *yixia* in (42) is POS ambiguous and has two senses. One is tagged as Nddc and means 'for a while' while another is tagged as DM and means 'once'. The former is labelled as duration and the latter is assigned as frequency. Equal to the cause of differences between (38) and (39), the ambiguities in (42) are due to situation types.

(36) 親政三十八年

qinzheng *sanshibanian*

hole the reins of government 38-M

hold the reins of government for 38 years

[4] More detailed discussion about situation types can be referred to Smith [1991].

(37) 三十八年親政

sanshibanian *qinzheng*

the year of 38 hold the reins of government

hold the reins of government in the year of 38

(38) 34年的服刑

sanshisinian *de* *fuxing*

34-M DE serve a sentence

serve a sentence for 34 years

(39) 34年的判刑

sanshisinian *de* *panxing*

the year of 34 DE sentence

sentence a person in the year of 34

(40) 等我一下

deng wo *yixia*

wait I for a while

wait for me for a while

(41) 敲他一下

Qiao *ta* *yixia*

strike he one-M

strike him once

(42) 咬他一下

a. *yao* *ta* *yixia*

bite he for a while

bite him for a while

b. *yao* *ta* *yixia*

bite he one-M

bite him once

Semantic role assignment is not an easy task, since it requires world knowledge as well as linguistic knowledge. In You and Chen [2004], they identify parameters of determining semantic roles and propose an instance-based approach to resolve ambiguities. They adopt dependency decision making and example-based approaches. Semantic roles are determined by four parameters, including syntactic and semantic categories of the target word, case markers, phrasal head, and sub-categorization frame and its syntactic patterns. The refinements of features extraction, canonical representation for certain classes of words and dependency decisions improve role assignment. To assign the semantic roles of DMs, the above parameters are further refined as the features of relative positions and situation types.

The examples above show that ambiguities are unavoidable when one deals with DMs. In addition to the typical DMs, some related structures like reduplicative DMs, numerals, the ellipsis of measures, etc. are also topics for discussion. During DM processing, certain DMs are ambiguous to automatic identification in word segmentations, senses as well as poses. Here, *yi dian* (一點) is taken as an example.

(43) 有一點要特別注意

you yidian yau tebie zhuyi
have one-M should special attention
There is a point very important for attention.

(44) 一點心意你要收下

yidian xinyi ni yao shouxia
little regard you should receive
You must receive my little thanks.

(45) 一點集合

yidian jihe
one o'clock assemble
assemble at one o'clock

(46) 漂亮一點

piaoliang yidian
beautiful a little
a little bit beautiful

(47) 快一點

a. *kuai* *yidian*
nearly one o'clock
nearly one o'clock

b. *kuai* *yidian*
fast a little
more quickly

(48) 慢一點

man *yidian*
slow a little
more slowly

The phrase *yidian* functions as a pronoun and tagged as DM in (43), functions as a quantitative determinative modifying *xinyi* and tagged as Neqa in (44), functions as a time noun and tagged as Ndabe in (45), and functions as a post-verb adverb of degree and tagged as Dfb in (46). While in (47), *yidian* has sense ambiguities depending upon context. In addition, *yidian* in (47a) and (48) is pos ambiguous. For another example, *qi* (起) functions as a measure in both *siqi anjian* (四起案件) and *yiqi mingan* (一起命案). In *fongyun siqi* (風雲四起) and *yiqi sikao* (一起思考), *siqi* and *yiqi* are tagged as VA11 and Dh individually. However, in *yiqi fanan* (一起犯案) and *fanan siqi* (犯案四起), the DMs *yiqi* and *siqi* are ambiguous. It is obvious that the ambiguities of DMs are complex and that a DM compound can have more than one classification of ambiguities.

No matter whether the ambiguity is from word segmentation, sense, or pos, the prescription of resolution principles and DM rules are helpful in disambiguating DMs. Besides, the neighbouring morphemes and context are other tricks in reducing ambiguities. Furthermore, pos ambiguities are concerned with common sense, and the resolution features also include positions of temporal DMs and situation types. Such ambiguities have to be reduced by the application of parameters of context vector models.

5. Implementation and Evaluation

We randomly chose 2035 sentences (11697 word tokens) from Sinica Treebank as our development set. In total, 545 tokens of the development data are processed by the revised DM rules (as shown in the appendix). Among the 545 tokens, 504 tokens are correctly

segmented, and 443 tokens are correctly pos tagged. The segmentation accuracy of the development data is 92.5%, the tag accuracy of the development data is 81.3%, and the tag accuracy with the correct segmentation of the development data is 88.0%. Contrastively, the segmentation accuracy and tag accuracy of the development set processed by the original DM rules are both lower than those applied the revised DM rules. The segmentation accuracy is 84.2%, and the tag accuracy is 71.0%. Then, to test the accuracy of the revised DM rules, we randomly chose 2111 sentences (12209 word tokens), which have no overlap with the development set, from Sinica Treebank as the testing set. In total 564 tokens of the testing data were processed using the revised DM rules. Among those 564 tokens, 508 tokens were correctly segmented, and 424 tokens were correctly pos tagged. By application of the revised DM rules, the segmentation accuracy of the testing data is 90.1%, the tag accuracy of the testing data is 75.2%, and the tag accuracy with the correct segmentation of the testing data is 83.5%. Contrastively, processed by the original DM rules, the segmentation accuracy and the tag accuracy of the testing data is 77.8% and 60.3% individually. Table 1 is the evaluation result.

Table 1. Accuracy of Development data and Testing data

data / accuracy		development set 2035 sentences (11697 word tokens)	testing set 2111 sentences (12209 word tokens)
original DM rules	segmentation accuracy	84.2%	77.8%
	tag accuracy	71.0%	60.3%
revised DM rules	segmentation accuracy	92.5%	90.1%
	tag accuracy	81.3%	75.2%
	tag accuracy with correct segmentation	88.0%	83.5%

Table 1 shows that both segmentation accuracy and tag accuracy of development set and of testing set processed by the revised DM rules are higher than those processed by the original DM rules, although the segmentation accuracy, tag accuracy and tag accuracy with correct segmentation of the testing set are a little bit lower than those of the development set.

After data analysis, we found that there were several reasons for inaccurate results. The most crucial factor resulting in inaccuracy is ambiguity, including word segmentation ambiguities, sense ambiguities and pos ambiguities. The word segmentation ambiguities, sense ambiguities, and pos ambiguities caused 11%, 32.4% and 27.2% of errors, respectively. For example, (49a) and (50a) are correct sentences. The contrastive sentences (49b) and (50b)

have errors in word segmentation ambiguities. Sentences (51b) and (52b) have errors in sense ambiguities, which are contrastive to the correct ones (51a) and (52a). Sentences (53b) and (54b) have errors in pos ambiguities. The total percentage of errors caused by ambiguities is about 70.6%.

(49) a. 汽車(Na) 可(D) 不(D) 是(SHI) ***一個(DM)*** 人(Na) 所(D) 發明(VC) 的(T)

qiche ke bu shi yige ren suo faming de

automobile should NEG SHI one-M person that which invent DE

Automobiles are not invented by one person.

*b. 汽車(Na) 可(D) 不(D) 是(SHI) 一(Neu) 個人(Nh) 所(D) 發明(VC) 的(DE)

qiche ke bu shi yi geren suo faming de

automobile should NEG SHI one individual that which invent DE

(50) a. ***有一次(DM)*** 要(VE) 刨(VC) 木(Na) 時(Ng)

youyici you pau mu shi

once should shave wood as

once when shaving wood

*b. 有(V_2) 一(Neu) 次要(A) 刨木(Na) 時(Ng)

you yi ciyou pau mu shi

have one secondary shave wood as

(51) a. 都(D) 有(V_2) ***一段(DM)*** 感人(VH) 的(DE) 故事(Na)

dou you yiduan ganren de gushi

all have one-M heart-stirring DE story

all have one heart-stirring story

*b. 都(D) 有(V_2) 一段(Nc) 感人(VH) 的(DE) 故事(Na)

dou you yiduan ganren de gushi

all have Sec. 1 heart-stirring DE story

(52) a. 但(Cbb) 也(D) 付出(VC) ***數位(DM)*** 創業(Nv) 先進(Na) 寶貴(VH) 的(DE) 生命(Na) 及(Caa) 損失(VJ) １４架(DM) 飛機(Na) 的(DE) 慘痛(VH) 代價(Na)

dan ye fuchu shuwei chuangye xianjin baoguei de shengming
but also pay several-M pioneering precursor valued DE life
ji sunshi shisijia feiji de cantong daijia
and lose 14-M airplane DE cruel cost

but also pay several pioneering precursors' valued lives and suffer the cruel costs of losing fourteen airplanes

*b.但(Cbb) 也(D) 付出(VC) 數位(A) 創業(VA) 先進(VH) 寶貴(VH) 的(DE) 生命(Na) 及(Caa) 損失(Na) １４架(DM) 飛機(Na) 的(DE) 慘痛(VH) 代價(Na)

dan ye fuchu shuwei chuangye xianjin baoguei de shengming
but also pay several-M digital precursor valued DE life
ji sunshi shisijia feiji de cantong daijia
and lose 14-M airplane DE cruel cost

(53) a. 10月(Nd) ***5日(Nd)***

shiyue wuri
Oct. fifth
Oct. 5

*b. 10月(Nd) 5日(DM)

shiyue wuri
Oct. five-M

(54) a. 對於(P) ***七十九年(Nd)*** 年終(Nd) 獎金(Na)

dueiyu qishijiunian nianzhong jiangjin
about the year of 79 year-end bonus
about the year-end bonus in the year of 79

*b. 對於(P) 七十九年(DM) 年終(Nd) 獎金(Na)

dueiyu qishijiunian nianzhong jiangjin
about 79-M year-end bonus

Other than errors in ambiguities, there are four kinds of errors bringing about inaccuracy. The first kind of errors is a result of the segmentation model (a Hidden Markov Model). In HMM, there are several possible paths, including the correct one. However, the result chosen was not the correct one. For example, (55a) and (55b) are the possible paths in HMM. For the result, the inaccurate one (55b) was chosen. The error of (56b) is also due to HMM. The percentage of errors made by HMM is 10.3%.

(55) a. 回想(VE) 起(Di) ***二十年(DM)*** ***前(Ng)*** 的(DE) 往事(Na)
hueixiang qi ershinian qian de wangshi
recall ASP 20-M before DE past
recall the past twenty years ago
*b. 回想起(VE) 二十(Neu) 年(Nf) 前(Ng) 的(DE) 往事(Na)
hueixiangqi ershi nian qian de wangshi
recall twenty M before DE past

(56) a. ***六十六歲(DM)*** 時(Ng)
liushiliusuei shi
66 years old as
as 66 years old
*b. 六十六(Neu) 歲(Nf) 時(Ng)
liushiliu suei shi
66 M as

The second kind of errors is because different contexts cause different tagging. In Sinica Treebank, one or more determinatives together with an optional measure will constitute a DM. However, certain determinatives and measures are tagged differently than usual because of the context. In (57a), *liangsanmiaozhong* (兩三秒鐘 'two and three seconds') is composed of *liangmiaozhong* 'two seconds' and *sanmiaozhong* 'three seconds'. The measure *miaozhong* 'second' is shared by determinatives *liang* 'two' and *san* 'three'. What is in (57c) is the tree structure of (57a). Both *liang* and *san* are tagged as Neu, and *miaozhong* is tagged as Nf. In (58a), *qibayue* (七八月 'July and August') is composed of *qiyue* 'July' and *bayue* 'August'. The diagram (58c) is the bracketed tree diagram of (58a). From the context, one knows that *qi* in (58a) is not the numeral 'seven' but a month 'July' so *qi* is tagged as Nd not Neu. The

percentage of errors resulting from contexts is 8.8%.

(57) a. 這(Nep) 姿勢(Na) 保持(VJ) ***兩(Neu) 三(Neu) 秒鐘(Nf)***

zhe zishi baochi liang san miaozhong

this pose keep two three M

This pose is keeping for two and three seconds.

*b. 這(Nep) 姿勢(Na) 保持(VJ) 兩三秒鐘(DM)

zhe zishi baochi liangsanmiaozhong

this pose keep two-three-M

c. S(theme:NP(quantifier:Nep:這|Head:Nac:姿勢)|Head:VJ1:保持
|duration:DM(Head:Neu(Head:Neu:兩|Head:Neu:三)|Head:Nfg:秒鐘))

(58) a. ***7(Nd)*** 、(Caa) ***8月(Nd)*** 爲(VG) 下午(Nd) ４點(Nd) 至(Caa) ４ 點(Nd) ３０分(Nd)

qi bayue wei xiawu sidian zhi sidian sanshifen

seven August is afternoon four o'clock to four o'clock thirty minutes

In July and August, it is from four to four-thirty o'clock.

*b. ７(Neu) 、(PAUSECATEGORY) ８月(Nd) 爲(P) 下午(Nd) ４點(DM) 至(Caa) ４點(DM) ３０分(DM)

qi bayue wei xiawu sidian zhi sidian sanshifen

seven August is afternoon four-M to four-M thirty minutes

c. S(theme:NP(Head:Ndabc(DUMMY1:Ndabc:７|Head:Caa:、
|DUMMY2:Ndabc:8月))|Head:VG2:爲|range:NP(property:Ndabe:下午
|Head:NP(DUMMY1:NP(Head:Ndabe: ４點)|Head:Caa:至
|DUMMY2:NP(property:Ndabe:４點|Head:Ndabe:３０分))))

The third kind of errors occurs when there is only one measure without any other determinatives, *e.g.* (59b) and (60b). The percentage of errors in one measure is 8.1%. The error in wrong tagging of one measure is because of the training data from Sinica Corpus. A sole measure is tagged as Nf in Sinica Corpus, but in Sinica Treebank it is viewed as a DM structure and tagged as DM. Therefore, a sole measure always has incorrect pos. This kind of error has to be dealt with during postprocessing.

(59) a. 最近(Nd) 聽到(VE) 了(Di) ***個(DM)*** 駭人聽聞(VH) 的(DE) 故事(Na)

zueijin tingdao le ge hairentingwen de gushi

recent hear LE M shocking DE story

recently hear one shocking story

*b. 最近(Nd) 聽到(VE) 了(Di) 個(Nf) 駭人聽聞(VH) 的(DE) 故事(Na)

zueijin tingdao le ge hairentingwen de gushi

recent hear LE M shocking DE story

(60) a. 有(V_2) ***顆(DM)*** 善良(VH) 的(DE) 心(Na)

you ke shanliang de xin

have M kindhearted DE heart

is kindhearted

*b. 有(V_2) 顆(Nf) 善良(VH) 的(DE) 心(Na)

you ke shanliang de xin

have M kindhearted DE heart

The last kind of error is because of unknown word identification such as (61) and (62). The unknown words in (61b) and (62b) are not identified correctly, so errors occur. The percentage of errors in unknown words is 2.2%.

(61) a. 誰(Nh) 言(VE) 寸(DM) 草(Na) 心(Na)

shei yan cun cao xin

who say inch grass heart

children like grass (cannot pay their parents back)

*b. 誰(Nh) 言(VE) 寸草(Na) 心(Na)

shei yan cuncao xin

who say grass heart

(62) a. 報(VC) 得(DE) 三(Neu) 春(Nd) 暉(Na)

bao de san chun huei

pay back DE three spring sunshine

pay parents back

*b. 報得三春暉(VH)

baodesanchunhuei

pay back

In our implementation and evaluation, the accuracy of segmentation, of tag, and of tag with correct segmentation of the development set processed by the revised DM rules are higher than those processed by the original DM rules. Through application of the revised DM rules, the segmentation accuracy, tag accuracy and tag accuracy with correct segmentation of the testing set are a little bit lower than those of the development set. The percentage of ambiguities causing inaccuracy is 70.6% while the total percentage of other factors is 29.4%. The high proportion of ambiguity shows that, although a regular expression approach was used in applying DM rules to deal with DMs, eventually, ambiguity is the most crucial issue one must confront. Therefore, the application of resolution principles, of DM rules, of context sensitive rules, of collocation bi-grams and of parameters of context vector models are necessary to help one disambiguate. Language reflects the human view of the world. Differing personal world knowledge may result in different explanations of sentences. Some reduction of ambiguities of DMs depends upon the human's common sense knowledge.

6. Conclusion

DMs are not a closed set, so one has to apply DM rules during the process of automatic identification of DMs. By observing Sinica Treebank, we had developed a set of regular expression rules to identify DMs and their parts of speech. Thus, all DM candidates can be matched and classified by regular expression rules. However, due to segmentation, pos and sense ambiguities of DMs, DM rules are necessary complements to dictionaries and helpful to resolve ambiguities by applying resolution principles and segmentation models. Sense and pos ambiguities are also expected to be resolved by different approaches during post-processing by applying context sensitive rules, collocation bi-grams and parameters of context vector models.

References

中央研究院詞庫小組(CKIP), "「搜」文解字—中文詞界研究與資訊用分詞標準," 技術報告 96-01, 中央研究院,台北, 1996.

何杰(He, J.), *現代漢語量詞研究*, 民族出版社, 北京市, 2002.

黃居仁, 陳克健, 賴慶雄(編著), *國語日報量詞典*, 國語日報出版社, 台北, 1997.

Academia Sinica, CKIP word segmentation system, http://ckipsvr.iis.sinica.edu.tw/.

Academia Sinica, Sinica Corpus, version 4.0, http://www.sinica.edu.tw/SinicaCorpus/, 2001.

Academia Sinica, Sinica Treebank, version 3.1, http://treebank.sinica.edu.tw/, 2006.

Chao, Y.-R., *A Grammar of Spoken Chines*e, University of California Press, Berkeley, 1968.

Crystal, D., *A Dictionary of Linguistics and Phonetics*, Blackwell, Massachusetts, Cambridge, 1991.

Goh, C.-L., M. Asahara and Y. Matsumoto, "Chinese Word Segmentation by Classification of Characters," *Computational Linguistics and Chinese Language Processing*, 10(3), 2005, pp. 381-96.

Jurafsky, D. and J. H. Martin, *Speech and Language Processing: An Introduction to natural Language Processing, Computational Linguistics, and Speech Recognition*, Prentice Hall, Upper Saddle River, N.J., 2000.

Li, C. N. and S. A. Thompson, *Mandarin Chinese: A Functional Reference Grammar*, University of California Press, Berkeley, 1981.

Ma, W.-Y. and K.-J. Chen, "Introduction to CKIP Chinese Word Segmentation System for the First International Chinese Word Segmentation Bakeoff," In *Proceedings of ACL, Second SIGHAN Workshop on Chinese Language Processing*, 2003, Sapporo, Japan, pp.168-171.

Mo, R.-P. J., Y.-J. Yang, K.-J. Chen and C.-R. Huang, "Determinative-Measure Compounds in Mandarin Chinese: Their Formation Rules and Parser Implementation," In *Proceedings of ROCLING IV (R.O.C. Computational linguistics Conference)*, 1991, National Chiao-Tung University, Hsinchu, Taiwan, pp. 111-134.

Prins, R. P., Finite-State Pre-Processing for Natural Language Analysis, Art Dissertation, 2005

Smith, C. S., *The Parameter of Aspect*, Kluwer Academic Publishers, Dordrecht, 1991.

Tai, J. H-Y, "Chinese classifier systems and human categorization," *In Honor of William S-Y. Wang: Interdisciplinary Studies on Language and Language Change*, ed. by M. Y. Chen and O J.-L. Tzeng, Pyramid Press, Taipei, 1994, pp. 479-494.

You, J.-M. and K.-J. Chen, "Automatic Semantic Role Assignment for a Tree Structure," In *Proceedings of 3rd ACL SIGHAN Workshop*, 2004, Barcelona, Spain, pp. 109-115.

Appendix. CKIP's DM Phrase Structure Rules

Abbreviation	Term
A	adjectives
DD	demonstrative determinatives
DESC	the adjectives that occur in the middle of a DM compound
DFa	pre-verbal degree adverbs
Dfb	post-verbal degree adverbs
Dh	manner adverbs
DM	determinative-measure compounds
DQ	quantitative determinatives denoting degree
DS	definite specific determinatives
M	measure words
NC	place nouns
Ncb	common place nouns
Nd	time nouns
Ndaad	time nouns indicating years
Ndabd	time nouns indicating days that are circular
Nddc	time nouns indicating future
Neqa	quantitative determinatives
Neu	numeral determinatives
Nf	measures
NO	numeral determinatives
ON	ordinal numerical determinatives
OS	ordinal specific determinatives
PNM	post nominal modifiers
PQ	quantitative determinatives denoting par relation
QO	interrogative quantitative determinatives
WQ	quantitative determinatives denoting totality
VA11	active intransitive motion verbs

NO1 = {〇,一,二,兩,三,四,五,六,七,八,九,十,廿,卅,百,千,萬,億,兆,零,幾} ;

NO2 = {壹,貳,參,肆,伍,陸,柒,捌,玖,拾,佰,仟,萬,億,兆,零,幾} ;

NO3 = {１,２,３,４,５,６,７,８,９,０,百,千,萬,億,兆} ;

NO3a = {百,千,佰,仟,萬,億,兆} ;

ON = {甲,乙,丙,丁,戊} ;

NC = {國,省,州,縣,鄉,村,鎮,鄰,里,郡,區,站,巷,弄,段,號,地,街,樓,衖,市,洲,部,司,課,院,科,系,級,股,室,廳} ;

Ndabe = {清晨,凌晨,早晨,早上,晚上,上午,中午,下午,晨間,午間,晚間,半夜,午夜,晨,午,晚,傍晚,深夜,晏午,子時,丑時,寅時,卯時,辰時,巳時,午時,未時,申時,酉時,戌時,亥時} ;

ND = {微秒,釐秒,秒,秒鐘,分,分鐘,刻,刻鐘,點,點鐘,點多鐘,更,旬,紀,輪,天,日,週,周,禮拜,季,年,年份,載,號,晚,宿,週年,周年,周歲,會,會兒,陣,世,輩,年期} ;

ND2 = {時,世紀,年度,月,月份,陣子,學期,學年,年代,下子} ;

DESC = {大,小,整} ;

PNM = {多,餘,半,出頭,好幾,開外,整,正,許,足,之多} ;

Nfa = {本,把,瓣,部,柄,床,處,期,齣,場,朵,頂,堵,道,頓,錠,闋,棟,幢,檔子,檔,封,幅,發,分,份,服,個,根,根兒,管,行,戶,件,家,架,卷,具,節,句,屆,箇,捲,劑,隻,尊,盞,張,枝,支,樁,幀,只,株,折,灶,軸,口,棵,款,客,輛,粒,片,輪,枚,面,門,幕,匹,所,艘,扇,首,乘,襲,頭,條,長條,台,臺,挺,堂,帖,顆,座,則,冊,任,尾,味,位,頁,葉,房,彎,班,員,介,丸,名,項,起,間,篇,題,目,招,股,回} ;

Nfb = {通,口,頓,盤,局,番} ;

Nfc = {對,雙,宗,番,畦,餐,行,身,列,長列,系列,排,長排,副,付,套,筆,串,長串,掛,幫,房,批,組,窩,綑,捆,群,胎,桌,嘟嚕,部,種,類,樣,樣兒,派,路,壟,落,伙,夥,束,簇,席,疊,紮,色,票,叢,隊,攤,式,蓬,項};

Nfd = {些,分,分兒,團,堆,泡,綹,撮,把,股,灘,汪,陣,口,口兒,抹,塊,滴,欄,捧,抱,層,重,帶,截,長截,長截兒,截兒,節,節兒,長節,長節兒,段,段兒,長段,長段兒,絲,絲兒,點,點兒,片,縷,部份,部分,坨,匹,疋,階,抔,波,道};

Nfe = {盒,盒子,匣,匣子,箱,箱子,櫃子,櫥,櫥子,籃,籃子,簍,簍子,爐子,包,包兒,袋,袋兒,池子,瓶,桶,聽,罐,罈,盆,鍋,籠,盤,碗,杯,勺,勺子,匙,湯匙,筒,擔,瓶子,桶子,罐子,罈子,盆子,鍋子,籠子,盤子,杯子,筒子,擔子,籮筐,杓,杓子,茶匙,壺,盅,筐,瓢,鍬,缸};

Nff = {身,頭,臉,鼻子,嘴,肚子,手,腳,桌子,院子,地,屋子,池,腔,家子};

Nfg = {公厘,公寸,公分,公尺,公丈,公引,公里,市尺,公釐,營造尺,台尺,吋,呎,碼,哩,浬,海浬,度,噚,尺,里,釐,寸,丈,米,厘,厘米,海哩,海里,英尺,英里,英呎,英寸,米突,米尺,微米,毫米,英吋,英哩,光年,公畝,公頃,市畝,營造畝,坪,畝,分,甲,頃,平方公里,平方公尺,平方尺,平方公分,平方英哩,英畝,公克,公斤,公噸,市斤,台兩,台斤,日斤,盎司,盎斯,磅,公擔,公衡,公兩,克拉,斤,兩,錢,噸,克,英磅,英兩,公錢,毫克,毫分,公毫,仟克,公撮,公升,市升,營造升,台升,日升,盎司,品脫,加侖,蒲式耳,公斗,公石,公秉,公合,公勺,斗,毫升,夸,夸特,夸爾,立方米,立方厘米,立方公分,立方公寸,立方公尺,立方公里,立方英尺,石,斛,西西,角,毛,元,圓,塊,先令,盧比,法郎,法朗,辨士,馬克,鎊,英鎊,美元,便士,里拉,日元,日圓,刀,打,令,綸,籮,大籮,焦耳,千卡,仟卡,燭光,仟瓦,千瓦,伏特,馬力,爾格,瓦特,瓦,卡路里,卡,馬克,仟赫,千赫,兆赫,赫,赫茲,位元,莫耳,歐姆,法拉第,安培,分貝,居里,微居里,毫居里,毫安培,毫米,毫巴,達因,牛頓,周波,歲,℃};

Nfh = {程,作,分,厘,毫,絲,圍,指,象限,度,開,開金,聯,師,旅,團,營,伍,班,排,連,球,波,回合,折,階,摺,等,票,流,棒,聲,次,股};

Nfi = {度,輪,回,次,遍,趟,下,下兒,遭,番,聲,聲兒,響,圈,圈兒,步,把,仗,覺,頓,關,手,手兒,腳,掌,巴掌,拳,拳頭,眼,口,刀,槌,槌子,板,板子,鞭,鞭子,棒,棍,棍子,陣,針,箭,槍,槍矛,砲,場,周,曲,跤,記,回合,票};

M = Nfa & Nfb & Nfc & Nfd & Nfe & Nfg & Nfh & Nfi & ND ;

TPNM = {半,多,許,整,正} ;

WQ = {一,全,滿,整,成,一切,所有} ;

QQ = {多少,若干,幾多} ;

DFa = {很,挺,怪,真,好,極,滿,更,再,頂,最,太,忒,多,夠,非常,異常,十分,尤其,有點,略爲,稍微,比較,不大,過分,過份,這麼,那麼} ;

DQ1 = {多,許多,許許多多,有些,好些,幾許,有的,少許,多數,少數,大多數,泰半,不少,部分,一部分,部份,個把} ;

DQ2 -> DFa {多,少} ;

PQ = {半,若干,有的} ;

DD = {這,那,哪};

OS = {上,下,前,後,頭,末,次,首,某,另,同} ;

DS = {本,貴,敝,什麼,啥,諸,何,別,旁} ;

IN1 -> { NO1*,NO3* } ;

IN2 -> NO2* ;

IN3 -> {IN1,IN2} {多,餘,來,幾} ({萬,億,兆}) ;

IN3a -> NO3a* ;

IN4 -> {上} IN3a ;

IN5 -> {第} {IN1,IN2} ;

DN -> IN1 { • , . ,、, · ,點} IN1 ;

NE5a -> {IN2} {一,／} {IN2} ;

NE5b -> {IN2} {一,／} {IN2} {一,／} {IN2} ;

FN1 -> (IN1 {又}) IN1 {分之,一,／} {IN1,DN} ({強,弱}) ;

FN2 -> {DN,IN1} {％} ;

FN3 -> {IN1} {成} ({IN1,PNM}) ;

FN = FN1, FN2, FN3 ;

NA1 -> IN1 {年級} ;

NC1 -> IN1 {鄰,巷,弄,段,號,樓} ;

NC2 -> IN1 {樓,號} {之,－} IN1 ;

NC3 -> IN1 {之,－} IN1 {號} ;

ND1 -> {IN1,這,那) ND ;

ND3 -> IN1 ND2 ;

ND4 -> IN1 ND (PNM,TPNM);

ND5 -> {這,那} {時,陣子,下子};

ONP -> ON M ;

NOP1 -> IN1 (DESC) ({半}) M ;

NOP2 -> DESC (半) M ;

NOP3 -> IN1 PNM ;

NOP4 -> M (PNM) ;

NOP5 -> {IN3,DN,FN,雙} M ;

NOP -> {FN,NOP1,NOP3,NOP4,NOP5} ;

WQP -> WQ M ;

WQP -> WQ Nff ;

WQP -> {整整,滿滿} NOP1 ;

QQP -> QQ NOP4 ;

DQP1 -> {好幾} {NOP1,NOP2,NOP4} ;

DQP2 -> {DQ1,DQ2} M ;

DQP -> {DQP1,DQP2} ;

PQP1 -> {數} {NOP1,NOP2,NOP4} ;

PQP2 -> PQ NOP4 ;

PQP -> {PQP1,PQP2} ;

XQP -> {WQP,QQP,DQP,PQP} ;

CNP -> IN1 {年} {IN1,ON} {班} ;

DSP1 -> {他} {國,省,州,縣,鄉,村,鎮,鄰,里,郡,區,站,巷,弄,段,號,地,樓,衖,市,洲} ;

DSP2 -> {該} {NOP,PQP} ;

DSP3 -> DS M ;

DSP -> {DSP1,DSP2} ;

OSP1 -> {第} NOP1 ;

OSP2 -> {每} {XQP,NOP,DSP2} ;

OSP2 -> {各} {XQP,NOP,DSP2} ;

OSP2 -> {逐} M ;

OSP3 -> {另外,近,將近} {PQP,NOP1,NOP5} ;

OSP4 -> OS {NOP,PQP} ;

DDP1 -> DD {WQP,DQP,PQP,NOP,NOP2} ;

DDP2 -> {此} {OSP1,NOP} ;

OSP -> {OSP1,OSP2,OSP4} ;

OHSP -> ({其它,其他,其餘}) {任何} {NOP1,DSP} ;

HOSP -> ({任何}) {其它,其他,其餘} {XQP,DDP1,OSP,NOP,ONP} ;

STDM -> IN1 {秒} IN1 ;

RNOP1 -> IN1 (DESC) M ;

RNOP2 -> {半} M ;

RNOP3 -> {DESC,成} M ;

RD13 -> ({一}) M M ;

Nac -> {NA1} ;

Ncb -> {NC1,NC2,NC3,CNP,DSP1} ;

Neqa -> {WQ,QQ,DQ1,DQ2,PQ,FN1,FN2,FN3,NOP3,RD13} ;

Neqb -> {PNM,TPNM} ;

Nep -> {DD} ;

Nes -> {OS,DS} ;

Neu -> {IN1,IN2,IN3,IN4,IN5,DN} ;

Nd -> {Ndabe,ND3,ND5}

DM ->{ND1,ND4,ONP,NOP1,NOP2,NOP4,NOP5,XQP,DSP,OSP,DDP1,DDP2,DSP3,STDM,RNOP1,RNOP2,RNOP3};