A Probe into Ambiguities of Determinative-Measure Compounds

Shih-Min Li, Su-Chu Lin, Keh-Jiann Chen
CKIP, Institute of Information Science, Academia Sinica, Taipei
{shihmin, jess}@hp.iis.sinica.edu.tw; kchen@iis.sinica.edu.tw

Abstract

This paper aims to further probe into the problems of ambiguities in automatic identification of determinative-measure compounds (DMs) in Chinese. It is known that Chinese DMs are identifiable by regular expression rules. However, rule matching only partially solve structural and lexical ambiguities. In this paper, a deep analyses based on corpus data was studied. With the subtle analyses of error identification and disambiguation of DM compounds, we classified three types of ambiguities, i.e. structural, sense, and functional ambiguities. We also proposed resolution principles to eliminate the problems and thus to improve word segmentation and POS (Part-Of-Speech) tagging.

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: 104). Furthermore, Li and Thompson (1981) assert 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* (三個人), *zhe zhan* in *zhe zhan deng* (這盞燈), *ji jian* in *ji jian yifu* (幾件衣服), *liu li* in *liu li lu* (六里路), *na jin* in *na jin yangrou* (那斤羊肉) and *ji gang* in *ji gang cu* (幾缸醋) are classifier phrases/measure phrases, which are called as D-M 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: 389). Although the demonstratives, numerals and measures may be listed exhaustedly, their combination is inexhaustible. Certain constructions of DMs are ambiguous, for example:

(1) 廣告裡全篇多具聳人聽聞的口號
guanggao li quan pian duo ju hairentingwen de kouhao
A whole advertisement mostly has shocking slogans.

(2) 取此名
qu ci ming
choose this one (person) / name this name

(3) 二十五年的審核、排隊、等待
ershiwu nian de shenhe paidui dengdai
examining, lining up and waiting in the year of twenty five / for twenty five years

The morpheme *ju* in Academia Sinica Balanced Corpus (Sinica Corpus) has two parts of speech, VJ and Nf.[1] Thus the phrase *duo ju* in sentence (1) can be either a verb phrase with the meaning of 'mostly have' or a DM. When *ju* functions as a measure, it always modifies corpses, not slogans. Since *ju* never co-occurs with slogans, the phrase *duo ju* here is certainly a verb phrase and then the lexical ambiguity of *ju* is reduced. In example (2), *ming* can function as a measure as well as a noun so that this verb phrase has two meanings. In example (3), *ershiwu nian* can be a time point specifying the event-time of the verb, or denotes the period of time delimitating the time length of the event. The former temporal adverb is tagged as Nd; the latter is separated into two morphemes and individually tagged as Neu and Nf.[2] Examples (1) to (3) show the different degree of ambiguity.

[1] The symbol in Sinica Corpus, "VJ" stands for Stative Transitive Verb and "Nf" for Measure. The detailed parts of speech can be referred to Sinica Corpus website.

[2] The symbol "Nd" stands for Time Noun and "Neu" for Numeral Determinatives.

Due to the infinite of the number of possible DMs, Mo et al. (1991) propose to identify DMs by regular expression before parsing as part of their morphological module in NLP. The adoption of DMs rules really improves the accuracy of recognition, but we still have some difficulties in segmentation as the preceding examples. In this paper, the discussion and classification of ambiguities of DMs are the focus. In addition to the typical DM structure with the combination of one or more determinatives with a measure, the reduplicative DMs and the ellipsis of determinatives will be also included under investigation. After the analyses of multiple ambiguities, we try to find out resolution principles to reduce these ambiguities.

2 Literature Review

To deal with DMs, first we have to give a proper definition to DMs; thus we can delimit the scope of our discussion. There are numerous discussions on determinatives as well as measures, especially on the types of measures.[3] The classification of measures is not the issue in this paper. To avoiding confusion between classifiers and measures, we have to pay attention to the distinction between them. Tai (1994: 480) 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 measures. In his definition, a measure is a bound morpheme which forms a D-M compound with one of the determinative enumerated above (Chao 1968: 584). Classifiers are defined as ‘individual measures’, which is one of the nine kinds of measures. As we 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 D-M compounds have unlimited versatility. While Li and Thompson (1981) blend classifier with measure. They conclude not only does a measure word generally not take a classifier, but any measure word can be a classifier. In Tai’s opinion (1944: 481), 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. In this paper, since we adopt the CKIP DM rules and symbols of POS, we inherit the term determinative-measure compounds (DMs), which have been defined as the composition of one or more determinatives together with an optional measure (Mo et al. 1991: 111).

As for the linguistic ambiguity, Crystal (1991: 17) specifies 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 year is 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’). The definition of structural and lexical ambiguities can be referred to Prins (2005). Prins (2005: 1) mentions if we restrict our attention to the syntax in texts, then we 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 human can resolve most ambiguity, of both types, without even being consciously aware of alternatives. The ambiguity remains of which we are aware, 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. However, in our following analysis, we find out that only structural ambiguity and lexical ambiguity are not enough to obtain more detailed discussion on ambiguities of DMs. Structural ambiguity is caused by different segmentation of words. With the same segmentation, that string of words may still be ambiguous because the same string may have more than one meaning or may have different functions. Therefore, lexical ambiguity can be further divided into two types; the former is sense ambiguity and the latter is functional ambiguity.

In this paper, we examine and analyze Mandarin Chinese DMs in Sinica Corpus. In the subsections in section 3, we make a study of the structures and ambiguities of DMs, and then try to analyze and disambiguate these DMs.

[3] 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.

3 Structures and Ambiguities of DMs

The probe into DMs in this paper focuses on the typical DM, one or more determinatives following a measure, including the variant forms of DM, such as the ellipsis of the determinative and the insertion of an adjective into DM[4]. Besides, we will sketchily study the various reduplicative forms of DM, like the reduplication of only measures 'MM' and the numeral *y*i preceding the reduplicative measures '*yi*MM' (一MM). The following structures show the variants of DMs.

(4) 只有三十八位參選
zhiyou sanshiba wei canxuan
Only thirty eight persons take part in the election.

(5) 我們必須說句良心話
women bixu shuo ju liangxinhua
We have to give an absolutely honest speech.

(6) 爲一個問題作一大堆研究
wei yi ge wenti zuo yi da dui yanjiu
do a lot of research for the reason of an question

(7) 種種問題
zhong zhong wenti
all sorts of questions

(8) 一張張海報
yi zhang zhang haibao
each poster

(9) 如此一大口一大口地吃
ruci yi da kou yi da kou di chi
make such a mouthful of eating

The DM *sanshiba wei* in example (4) is the typical DM. In example (5), determinatives preceding the measure *ju* are omitted. Therefore, the demonstrative, specifying, numeral or quantitative functions of determinatives in (5) will flow away. Example (6) has the DM structure *yi da dui*, composed of DM *yi dui* and an insertion of an adjective *da*. The reduplication in (7), (8) and (9) are also our topical subjects.

As Chao (1968: 552) points out, a D-M compound is a substantive and can enter into constructions as subject, object, or attribute. In sentence (4) above, *sanshiba wei* is the subject of the verb *canxuan* and has the function as a pronoun. DMs in example (5) to (8) modify the amount of nouns, so they all have the function as an attribute. The reduplication of DMs in (9) describes the manner of eating and functions as an adverb.

3.1 Structural Ambiguities of DMs

When we identify DMs, structural ambiguity of them occurs as the following examples.

(10) 一服藥就見效
yi fu yao jiu jianxiao
Every time when he takes medicine, the illness is completely cured.
One dose is effective.

(11) 一串串珠飾品
yi chuan chuanzhu shipin
one string of beads

(12) 北市文昌國小五年一班
beishi wenchang guoxiao wunianyiban
the Fifth Grade Class One in Taipei Wen Chang Elementary school

(13) NC1 -> {NE1,NE2} {年} {NE1,NE2,ON} {班} ;

Example (10) is grammatically ambiguous and has two meanings. If *fu* functions as a verb, the meaning of (10) is the former one. If *fu* functions as a measure, the meaning of (10) is the latter. Because the combination of determinatives and measures are countless, the DM *yifu* won't be listed in the CKIP dictionary. Different word segmentation will bring structural ambiguity forth. However, Mo et al. (1991) list a resolution principle to reduce structural ambiguity. The principle asserts if

[4] The insertion of an adjective into DM has the structure of '*yi*AM'. The symbol 'A' stands for adjectives.

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, the former meaning in (10) has higher priority to the latter. Similarly, the DM in (11) may be segmented as *yi chuan* or *yi chuan chua*n if the measure *chuan* is followed by the same morpheme as it. The statistics of collocation of the context help to reduce the structural ambiguity existed in example (11). Example (12) also has structural ambiguity. If *nian* and *ban* are treated as measures, *wu nian yi ban* are segmented into four. By application of the resolution principle and DM rule (13), classes in elementary schools are viewed as a unit; therefore, *wunianyiban* is restricted be a unit whose POS is Nc[5].

Another structural ambiguity exists in the ellipsis of the determinatives. The phrase *you ge ren* in sentence (14) denotes to somebody. In (14), *ge* functions as a measure without any determinatives preceding it. The same phrase in (15), however, never refers to persons. The morpheme *geren* is viewed as a unit and tagged as Nh[6] with the meaning 'individual' in (15). The morpheme *you* in (15) is a verb and means 'have', whose function is not the same as the specifying determinative *you* in (14). The resolution principle and the collation help to resolve this kind of structural ambiguities.

(14) 有一次有個人瞥見我在街上拍照
you yi ci you ge ren piejian wo zai jie shang paizhao
Once somebody got a glimpse of my taking pictures in a street

(15) 有個人空間
you geren kongjian
have individual space

When dealing with addresses, we also encounter structural ambiguities, especially indicating the floor, number, alley, lane, section and neighbourhood. The following instances show the same forms with different segmentation between DMs and addresses.

(16) 屬於１１７號公路的一段
shuyu yiyiqi hao gonglu de yi duan
belong to a part of the 117th road

(17) 羅斯福路一段八號一樓
luosifulu yiduan bahao yilou
F 1, No. 8, Roosevelt Rd., Sec. 1

(18) 行經屏市長安里竹圍巷一之一０二號時
xing jing pingshi changanli zhuweixiang yizhiyilingerhao shi
when going through No. 1-102, Zhuwei Lane, Changan Village, Pingtung City

(19) NC2 -> {NE1,NE2} {鄰,巷,弄,樓} ;

(20) NC4 -> {NE1,NE2} {之,－} {NE1,NE2} {號 } ;

(21) 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 sanlou zhiyi
a few days ago, move to 3F-1, No. 7, XinYi Rd., Sec. 3, Taipei

(22) 物理研究所則列名於３５個研究機構之一
wuli yanjiusuo ze liemingyu sanshiwu ge yanjiu jigou zhi yi
The Institute of Physics is placed among 35 research centers.

(23) 等於是一日的三分之一
dengyu shi yi ri de ssanfenzhiyi
is equal to one day of thirds

(24) NE6 -> ({NE1,NE2} {又}) {NE1,NE2} {分之} {NE1,NE2,NE5} ;

In instance (16), *hao* and *duan* are both measures specifying the fixed amount or quantity of the road, so the numerals 117 and *yi* are separated from the measures. 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 conjoint 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, *yiduan*, *bahao* and *yilou* in (17) and *yizhiyilingerhao* in (18) are segmented as Nc in Sinica Corpus. The DM rules (19) and (20) help us tag locative DMs as Nc, which is different from the DM structures in (16). During the process of locative DMs, another concerned structure listed in (21), (22) and (23) derives. All the three instances have the same surface structure *zhi yi*, but the functions and segmentation are

[5] The symbol "Nc" stands for Place Noun in Sinica Corpus.

[6] The symbol "Nh" stands for Pronoun.

different. The morpheme *zhi yi* in (21) is tagged as a unit whose POS is Nc, in (22) is segmented into two units whose POS is individually DE[7] and Neu, while in (23) is the part of the whole quantitative determinative *sanfenzhiyi* tagged as Neqa[8]. To reduce these structural ambiguities, the DM rule (24) is necessary.

According to the above discussions, to resolve structural ambiguities, we conclude the following resolution principles which were implemented at the word segmentation system by Ma and Chen (2003).

a) D-M compounds are expressed and matched by regular expressions.
b) Lexical words have higher precedence than D-M compounds (cf. 11).
c) Long D-M has higher precedence than short D-M (cf. 12, 16, 17, 18, 21, 22, 23).
d) Covering ambiguities are resolved by collocation context (cf. 10, 14, 15).

The structural ambiguity is caused by different possible segmentation. Although example (10) has structural ambiguities, after the application of the resolution principles, the ambiguous segmentation is resolved and the correct segmentation has higher priority.

3.2 Sense Ambiguities of DMs

Senses and semantic functions of DMs are related to the types of measures. Li and Thompson (1981: 105) claim that Mandarin has several dozen classifiers, most of which can be found in Chao (1968: sec. 7.9). Chao (1968: 584-620) divides measures into nine kinds: (1) classifiers, or individual measures (Mc), (2) classifiers specially associated with V-O constructions (Mc'), (3) group measures (Mg), (4) partitive measures (Mp), (5) container measures (Mo), (6) temporary measures (Mt), (7) standard measure (Mm), (8) quasi-measures (Mq), and (9) measures for verbs (Mv). Briefly speaking, the function of DMs is to modify the amount and quantity of abstract and concrete things, to count the frequencies of events and actions, and to indicate the event time. To testify the functions of DMs, we analyze the semantic roles of DMs in Sinica Treebank. First we use “DM” as the keyword to retrieve the Sinica Treebank data and then calculate the frequencies of the semantic roles of these DMs. The most highly frequent semantic role is quantifier, whose frequency is 6434. Quantifier is mainly used to account for the amount of things. The statistics in Sinica Treebank show the frequencies of the semantic roles of DMs, and then we get the hierarchical order of semantics roles of DMs from high to low: quantifier > Head > DUMMY > range > time > property > frequency > goal > duration > theme > DUMMY1 > DUMMY2 > agent > quantity > topic > manner > apposition > location > instrument > experiencer. As expected, quantifier is the most common semantic role played by DMs. The semantic role “property” denoting attributes and “range” referring to amount both have high frequencies. “Quantity” is also used to modify the extent of actions. The measures such as *nian* (年), *ci* (次) and *tian* (天) are related to temporal concepts, whose semantic roles may be “time”, “frequency” or “duration”. The semantic roles “range”, “time”, “frequency” and “duration” are usually concerned with the measures classified into Mm and Mq in Chao’s classification of measures. The DMs always function as pronoun when their semantic roles are “goal”, “theme”, “agent”, “topic”, “apposition”, “location” and “experiencer”. If DMs play the semantic role “manner” and “instrument”, the measures are usually classified into Chao’s Mv. 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 information for resolving lexical ambiguities. Methods for word sense disambiguation are also applicable for DMs. Here we first discuss ambiguity about temporal adverbs to illustrate the sense ambiguity and possible resolution methods.

To represent the percentage, using Chinese characters like (23) is one form, and adopting mathematical symbols like (25) is another one. The form of mathematical symbols is sense ambiguous. It can refer to either the percentage like (25) or time point like (26). Without context, the symbol “１０／２１” can be a fractional number and read as “ten over twenty one” with the POS “Neqa” and as “ 10月21日 *shiyue ershiyiri*” with the POS “Nd” whose semantic role is time. To reduce this kind of sense ambiguities, we have the DM rules (28) and (29). The form in rule (28) is tagged as Neqa (numbers) while in (29) as Nd (time point). The mathematical symbol indicating a specific time usually denotes the year together so “2005／06／30” in (27) is tagged as Nd. Although we have rules (28) and (29) to help differentiate the meaning of percentage from that of time, we still have to have context to make (25) and (26) distinguishable.

[7] The symbol of “DE” is the POS of 的, 之, 得 and 地.

[8] The symbol “Neqa” stands for Quantitative Determinatives.

(25) ４０分的佔了２／３

sishi fen de zhan le sanfenzhier

Those of forty points occupy two-thirds.

(26) １０／２１召開全院網路工作小組第三次會議

shiyuershiyiri zhaokai quan yuan wanglu gongzuo xiaozu disan ci huiyi

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

(27) 2005／06／30更新

2005／06／30 gengxin

update on June 30, 2005

(28) NE5a -> {NE2} {—,／} {NE2} ;

(29) NE5b -> {NE2} {—,／} {NE2} {—,／} {NE2} ;

Chao (1968) gives an example about time words. 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. Below are examples with similar ambiguity as Chao mentions.

(30) 經過卡斯楚三十年的統治之後

jingguo sanshi nian geming de xili

after Castro’s thirty-year governance

(31) 三十年秋，緝私總隊復正名爲稅警總團

sanhinian qiu qisi zongdui fu zhongmingwei shuijingzongtuan

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

Examples (30) and (31) have the same temporal phrase *sanshi nian*, but their semantic functions and roles are different. The temporal phrase in (30) expresses time length and is segmented into two units as Neu and Nf. The semantic role of it is duration. However, *sanshinian* in (31) indicates time point and is tagged as Nd, whose semantic role is time. Although either a Chinese reign title or *Mingguo* (民國) preceding *sanshinian* is omitted, we still know *sanshinian* is a specific time, not a period, from the context. When the measures *nian* and *ri* are preceded by numerals, the temporal phrases always have sense ambiguity. Basically, we segment numeral and measure into two and then postprocess them by applying two tricks following. If DMs denote time point, 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 *Guanxu*, *Qianlong* (乾隆), *Tianbao* (天寶), *Jiajing* (嘉靖) and so on. Another trick helps to recognize DMs is its neighbouring temporal nouns. The temporal DMs usually co-occur with one or two temporal phrases such as *erlinglingwunian liuyue* (2005年6月), *liuyue sanshiri* (6月30日), *erlinglingwunian liuyue sanshiri* (2005年6月30日), etc.

Two tricks above can reduce some ambiguities of temporal phrases. But some ambiguities listed in the following examples cannot be reduced.

(32) 2005宜蘭童玩節

erlinglingwu yilan tongwan jie

I-Lan International Children’s Folklore & Folkgame Festival in 2005

(33) 一要有錢

yi yao you qian

first have to have money

(34) 九二一地震

jiueryi dizhen

the earthquake 921

(35) 鼓勵534人完成319鄉之旅

guli wubaisanshisi ren wancheng sanbaiyishijiu xiang zhi lu

encourage 534 persons to accomplish the travel around 319 villages

The composition of numeral mostly functions as numeral determinatives while sometimes doesn’t. The numeral 2005 in (32) refer to the year of 2005 AD; however, the numeral *yi* in (33) is a correlative conjunction. Furthermore, the similar numeral structures in (34) and (35) have different semantic meanings.

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

3.3 Functional Ambiguities of DMs

The semantic function of the temporal DM in (36) may be duration while (37) time. Same words and same word senses may play different semantic roles. The reason of making different assignment of semantic roles may be concerned with logical interpretation of sense collocations according to common sense and the real world knowledge.

(36) 18年的苦守
shiba nian de kushou
wait bitter for eighteen years

(37) 89年的反抗
bajiu nian de fankang
the revolt in the year of 89

When detecting DMs rules and Sinica Corpus data, we find out some interesting examples. The verb phrases (38) and (39) have the same morphemes except for the position of the temporal DM *sanshiba nian*. The semantic role of the DM in (38) is duration while in (39) is time. It seems the different position of temporal DMs will affect the meanings of sentences. Thus, we briefly calculate the data in Sinica Treebank. The totality of the semantic role time of NPs and PPs following the verb is close to that of the semantic role duration. But the totality of the semantic role time of NPs and PPs preceding the verb is much more than that of duration. It seems temporal DMs preceding verbs mostly function as time. Another different assignment of semantic role to the similar structure is shown by (40) and (41). The former DM is assigned the semantic role duration while the latter time. This kind of ambiguity has relation to situation types. The situation type of *fuxing* (服刑) is activity while that of *panxing* (判刑) is achievement. The feature [±Durative][9] of the events causes differences. The phrase *yixia* in (42) means 'for a while' and is assigned the role of duration, while in (43) is assigned frequency and composed of a numeral and a measure. The phrase in (44) is ambiguous with two meanings. One means 'bite him for a while' while another means 'bite him once'. Equal to the cause of differences between (40) and (41), the ambiguity in (44) is also due to the situation types.

(38) 親政三十八年
qinzheng sanshiba nian
take over reins of government upon coming of age for 38 years

(39) 三十八年親政
sanshibanian qinzheng
take over reins of government upon coming of age in the year of 38

(40) 34年的服刑
sanshisi nian de fuxing
serve a sentence for 34 years

(41) 34年的判刑
sanshisi nian de panxing
sentence a person in the year of 34

(42) 等我一下
deng wo yixia
Wait for me for a while.

(43) 敲他一下
qiao ta yi xia
strike him once

(44) 咬他一下
yao ta yixia
bite him for a while / bite him once

Semantic role assignment is not an easy task, since it requires not only linguistic knowledge but also world knowledge. In Yu and Chen (2004), they identify parameters of determining semantic roles and

[9] The more detailed discussion about situation types can be referred to Smith 1991.

proposed 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 semantic roles of DMs, the above parameters are further refined as the features of relative positions and situation types.

The examples above show that ambiguity is unavoidable when we deal with DMs. In addition to the typical DMs, some related structures like reduplicative DMs, numerals, the ellipsis of measures, etc. are also the topics for discussion. The composition of determinatives and measures brings about ambiguity. Some ambiguities are caused by different segmentations of words, some are due to the multiple meanings of words, and others are concerned with different functions. Therefore, ambiguities of DMs are classified into structural ambiguity, sense ambiguity and functional ambiguity. Here take *yi dian* (一點)for instance.

(45) 有一點要特別注意
zhe yi dian zhuyi shixiang hen zhongyao
This point for attention is very important.

(46) 一點心意你要收下
yidian xinyi ni yao shouxia
You must receive my little thanks.

(47) 一點集合
yidian jihe
assemble at one o'clock

(48) 漂亮一點
piaoliang yidian
a little bit beautiful

(49) 快一點
kuai yidian
nearly one o'clock
more quickly

(50) 慢一點
man yidian
more slowly

The phrase *yidian* has different structures in sentences (45) to (48). In (45), *yi dian* functions as a pronoun and is segmented into two units. In (46) to (49), *yidian* is viewed as one unit. It functions as a quantitative determinative modifying *xinyi* in (46), a time noun in (47), and a post-verb adverb of degree in (48). While in (49), *yidian* is lexically ambiguous depending upon context. However, when (49) has the former meaning, (49) and (50) are functionally ambiguous. The ambiguity of DMs is complex and it is possible that one DM compound has more than one classification of ambiguities

No matter the ambiguity is the structural one, sense one or functional one, the prescription of resolution principles and DM rules are helpful in disambiguating DMs. Besides, the neighbouring morphemes and context are one another tricks in reducing ambiguity. In some cases, ambiguity is not easily resolved. Furthermore semantic role ambiguities are concerned with common sense and the resolution features also include position of temporal DMs and the situation types. Such ambiguities have to be reduced by the application of parameters of context vector models.

4 Conclusion

In section 3, we discuss the ambiguity of DMs, which is mainly divided into structural ambiguity, sense ambiguity and functional ambiguity. These ambiguities can be reduced by applying resolution principles, DM rules, context sensitive rules, collocation bi-grams and parameters of context vector models. Because language reflects the human view of the world, different personal world knowledge may result in different explanation of sentences. Some reduction of ambiguities of DMs depends upon human's common sense knowledge.

During the process of segmentation, all DM candidates are matched and classified by regular expression rules. Then structure ambiguities will be resolved by applying resolution principles and segmentation models. Sense and function ambiguities are expected to be resolved by different approaches during postprocessing. Some DMs, such as *yidian*, are three way ambiguous. The resolutions of their structure ambiguities have to be delayed until sense ambiguities are resolved. For instance, temporal DMs are by default segmented into one unit first, which specifies time points and whose POS is Nd. If they are identified as sense of duration at postprocessing stages, the one unit DMs will be re-segmented into two units, i.e. a number followed by a measure. In future work, the debatable issue whether *yue* (月) is a quasi measure or an ordinary individual noun is our concerns. The insertion of adjectives into DMs and the reduplication of DMs are also worthy in investigation.

References

Chao, Yuen Ren. 1968. *A Grammar of Spoken Chines*e. Berkeley: University of California Press.

Chinese Knowledge Information Processing Group. 1996. *ShouWen JieZi - A Study of Chinese Word Boundaries and Segmentation Standard for Information Processing* [In Chinese]. CKIP Technical Report 96-01. Taipei: Academia Sinica.

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

He, Jie (何杰). 2002. *《現代漢語量詞研究》*.民族出版社.

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

Ma, Wei-Yun and Keh-Jiann Chen. 2003. 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*, pp168-171.

MDNA (Mandarin Daily News Association 國語日報出版社) and CKIP (中央研究院詞庫小組). 1997.《國語日報量詞典》.

Mo, Ruo-ping Jean, Yao-Jung Yang, Keh-Jiann Chen and Chu-Ren Huang. 1991. Determinative-Measure Compounds in Mandarin Chinese: Their Formation Rules and Parser Implementation. In *Proceedings of ROCLING IV (R.O.C. Computational linguistics Conference).* pp. 111-134.

Prins, Robbert Paul. 2005. *Finite-State Pre-Processing for Natural Language Analysis*. Art Dissertation.

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

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

You Jia-Ming and Keh-Jiann Chen, 2004. Automatic Semantic Role Assignment for a Tree Structure. In *Proceedings of 3rd ACL SIGHAN Workshop*. Barcelona Spain.

Website Resources

Sinica Corpus. http://www.sinica.edu.tw/SinicaCorpus/

Sinica Treebank. http://treebank.sinica.edu.tw/