Computational Linguistics and Chinese Language Processing
Vol. 14, No. 1, March 2009, pp. 19-44

Automatic Sense Derivation for Determinative-Measure Compounds under the Framework of E-HowNet

Chia-Hung Tai*, Jia-Zen Fan*, Shu-Ling Huang*,+, and Keh-Jiann Chen*

Abstract

In this paper, we take Determinative-Measure Compounds as an example to demonstrate how the E-HowNet semantic composition mechanism works in deriving the sense representation for a newly coined determinative-measure (DM) compound. First, we define the sense of a closed set of each individual determiner and measure word in E-HowNet representation exhaustively. Afterwards, we make semantic composition rules to produce candidate sense representations for a newly coined DM. Then, we review development set to design sense disambiguation rules. We use these heuristic disambiguation rules to determine the appropriate context-dependent sense of a DM and its E-HowNet representation. The experiment shows that the current system reaches 89% accuracy in DM sense derivation and disambiguation.

Keywords: Semantic Composition, Determinative-Measure Compounds, Sense Representations, Extended How Net, How Net

1. Introduction

Building a knowledge base is time consuming work. The CKIP Chinese Lexical Knowledge Base has about 80 thousand lexical entries, and their senses are defined in terms of the E-HowNet format. E-HowNet is a lexical knowledge representation system. It extends the framework of HowNet (Dong *et al*., 2006) to allow semantic composition. Based on the framework of E-HowNet, we intend to establish an automatic semantic composition mechanism to derive sense of compounds and phrases from lexical senses (Chen *et al.,* 2005b),

* CKIP, Institute of Information Science, Academia Sinica
E-mail: {glaxy; kitajava; kchen} @iis.sinica.edu.tw
+ Department of Language and Literature Studies, National Hsinchu University of Education
E-mail: slhuang@mail.nhcue.edu.tw

[Received November 3, 2008; Revised May 27, 2009; Accepted May 28, 2009]

(Huang *et al.,* 2008). Determinative-Measure compounds (abbreviated as DM) are the most common compounds in Chinese. As a determiner and a measure normally coin a compound with unlimited versatility, the CKIP group does not define the E-HowNet representations for all DM compounds. Nevertheless, construction patterns for DMs are regular (Li *et al.,* 2006). Therefore, an automatic identification schema in regular expression (Li *et al.,* 2006) and a semantic composition method under the framework of E-HowNet for DM compounds were developed.

In this paper, we take DMs as an example to demonstrate how the E-HowNet semantic composition mechanism works in deriving the sense representations for all DM compounds. The remainder of this paper is organized as follows. Section 2 presents the background knowledge of DM compounds and sense representation in E-HowNet. We'll describe our method in Section 3 and discuss the experiment result in Section 4 before we present conclusions in Section 5.

2. Background

There are numerous studies on determiners as well as measures, especially on the types of measures[1]. Tai (1994) asserts that classifiers and measures words are often treated together under one single framework of analysis. Chao (1968) treats classifiers as one kind of measure word. In his definition, a measure is a bound morpheme which forms a DM compound with the determiners enumerated below.

i. Demonstrative determiners, *e.g.* 這 "this", 那 " that"…
ii. Specifying determiners, *e.g.* 每 "every", 各 "each"…
iii. Numeral determiners, *e.g.* 二 "two", 百分之三 "three percent", 四百五十 "four hundred and fifty"…
iv. Quantitative determiners, *e.g.* 一 "one", 滿 "full", 許多 "many"…

Measures are divided into nine classes by Chao (1968). Classifiers are defined as 'individual measures', which is one of the nine kinds of measures.

i. classifiers, *e.g.* 本 "a (book)",
ii. classifier associated with V-O constructions, *e.g.* 手 "hand",
iii. group measures, *e.g.* 對 "pair",
iv. partitive measures, *e.g.* 些 "some",

[1] Chao (1968) and Li and Thompson (1981) detect measures and classifiers. He (2002) 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.

v. container measures, *e.g.* 盒 "box",

vi. temporary measures, i. 身 "body",

vii. Standard measures, *e.g.* 公尺 "meter",

viii. quasi-measure, *e.g.* 國 "country",

ix. Measures with verb, *e.g.* 次 "number of times".

As mentioned in the introduction, Chao considered that determiners are listable and measures are largely listable, so D and M can be defined by enumeration, and that DM compounds have unlimited versatility. In this paper, we adopt the CKIP DM rule patterns and Part-of-Speeches for morpho-syntactic analysis, and, therefore, inherit the definition of determinative-measure compounds (DMs) in Mo *et al.* (1991). Mo *et al.* defined a DM as the composition of one or more determiners together with an optional measure. It is used to determine the reference or the quantity of the noun phrase that co-occurs with it. We use the definition of Mo *et al*. to apply to NLP and somewhat different from traditional linguistics definitions.

2.1 Regular Expression Approach for Identifying DMs

Due to the possible infinite number of DMs, Mo *et al.* (1991) and Li *et al.* (2006) proposed to identify DMs by regular expression as part of their morphological module in NLP. For example, when the DM compound is the composition of one determiner, *e.g.* numerals in (1), rules (2a), (2b), or (2c) will be first applied, and then rules (2d), (2e), or (2f) will be applied to compose complex numeral structures, and finally rule (2g) will generate the pos Neu of numeral structures. From the processes of regular expression, the numerals 534 and 319 in (1) are identified and tagged as Neu.[2]

(1) 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

[2] The symbol "Neu" stands for Numeral Determiners. Generation rules for numerals are partially listed in (2).

(2) 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 } ;

Regular expression approach is also applied to deal with ordinal numbers, decimals, fractional numbers and DM compounds for times, locations etc.. The detailed regular expressions can be found in Li *et al.* (2006). Rule patterns in regular expression only provide a way to represent and to identify morphological structures of DM compounds, but do not derive the senses of complex DM compounds.

2.2 Lexical Sense Representation in E-HowNet

Core senses of natural language are compositions of relations and entities. Lexical senses are processing units for sense composition. Conventional linguistic theories classify words into content words and function words. Content words denote entities and function words mainly serve grammatical functions which link relations between entities/events. In E-HowNet, the senses of function words are represented by semantic roles/relations (Chen *et al.* 2005a). For example, 'because' is a function word. Its E-HowNet definition is shown in (3).

(3) because|因爲 def: reason={};

which means reason(x)={y} where x is the dependent head and y is the dependent daughter of '因爲'.

In the following sentence (4), we'll show how the lexical concepts are combined into the sense representation of the sentence.

(4) Because of the rain, all the clothes are wet. 因爲下雨，衣服都濕了

In the above sentence, ‘濕 wet’, ‘衣服 clothes’ and ‘下雨 rain’ are content words while ‘都 all’, ‘了 Le’ and ‘因爲 because’ are function words. The difference of their representation is that function words start with a relation but content words have under-specified relations. If a content word plays a dependent daughter of a head concept, the relation between the head concept and this content word will be established after parsing process. Suppose that the following dependent structure and semantic relations are derived after parsing the sentence (4).

(5) S(reason:VP(Head:Cb:因爲|dummy:VA:下雨)|theme:NP(Head:Na:衣服) |
quantity: Da:都 | Head:Vh:濕|particle:Ta:了)。

After the feature unification process, the following semantic composition result (6) is derived. The sense representations of dependent daughters became the feature attributes of the sentential head ‘wet|濕’.

(6) def:{wet|濕:
theme={clothing|衣物},
aspect={Vachieve|達成},
manner={complete|整},
reason={rain|下雨}}

In (5), the function word ‘因爲 (because)’ links the relation of ‘reason’ between head concept ‘濕 wet’ and ‘下雨 rain’. The result of the composition is expressed as reason(wet|濕)={rain|下雨}, since, for simplicity, the dependent head of a relation is normally omitted. Therefore, reason(wet|濕)={rain|下雨} is expressed as reason={rain|下雨}; theme(wet|濕)={clothing|衣物} is expressed as theme={clothing|衣物} and so on in the expression (6).

2.3 The sense representation for determiners and measures in E-HowNet

The sense of a DM compound is determined by its morphemes and the morphemes of DMs are determiners and measures which are exhaustively listable. Therefore, in order to apply a semantic composition mechanism to derive the senses of DM compounds, we first need to establish the sense representations for all determiners and measures. Determiners and measures are both modifiers of nouns/verbs and their semantic relation with head nouns/verbs are well established. We, thus, defined them by a semantic relation and its value like (7) and (8) below.

(7) The definition of determiners in E-HowNet

this 這 def: quantifier={definite|定指}

first 首 def: ordinal={1}

one 一 def: quantity={1}

For measure words, we found that some measure words contain content sense, but for some measure words, such as classifiers, their content senses are not important and could be neglected. So, we divided the measure words into two types, with or without content sense, with their sense representations being exemplified below:

(8) The definition of measure words in E-HowNet

a) Measure words with content sense

碗 bowl def: container={bowl|碗}

米 meter def: length={meter|公尺}

月 month def: time={month|月}

b) Measure words without content sense

本 copy def:{null}

間 room def:{null}

樣 kind def:{null}

3. Semantic Composition for DM Compounds

To derive sense representations for all DM compounds, we study how to combine the E-HowNet representations of determiners and measures into a DM compound representation and make rules for automatic composition accordingly. Basically, a DM compound is a composition of some optional determiners and an optional measure. It is used as a modifier to describe the quantity, frequency, container, length, etc. of an entity. The major semantic roles played by determiners and measures are listed in Table 1. Since an E-HowNet sense representation is basically a feature value structure, we will apply feature unification process for semantic derivation of DMs. The basic feature unification processes (Duchier *et al.,* 1999) is as follows:

If a morpheme *B* is a dependency daughter of morpheme *A*, *i.e*. *B* is a modifier or an argument of *A*, then unify the semantic representation of *A* and *B* via the following steps.

Step 1: Identify the semantic **relation** between *A* and *B* to derive **relation**(*A*)={*B*}.

Step 2: Unify the semantic representation of *A* and *B* by insert relation(*A*)={*B*} as a sub-feature of *A*.

As exemplified in (9) and (10), a feature unification process can derive the sense representation of a DM compound if its morpheme sense representations and semantic head are known.

(9) one 一 def:quantity={1} + bowl 碗 def: container={bowl|碗} →
one bowl 一碗 def: container={bowl|碗:quantity={1}}

(10) this 這 def: quantifier={definite|定指} + 本 copy def:{null} →
this copy 這本 def: quantifier={definite|定指}

Table 1. Major semantic roles played by determiners and measures

Semantic Role	D/M
quantifier	*e.g.* 這、那、此、該、本、貴、敝、其、某、諸
ordinal	*e.g.* 第、首
qualification	*e.g.* 上、下、前、後、頭、末、次、首、其他、其餘、別、旁、他、另、另外、各
quantity	*e.g.* 一、二、萬、雙、每、任何、一、全、滿、整、一切、若干、有的、一些、部份、有些、許多、很多、好多、好幾、好些、少許、許許多多、幾許、多數、少數、大多數、泰半、不少、個把、半數、諸多
Formal={.Ques.}	*e.g.* 何、啥、什麼
Quantity={over, approximate, exact}	*e.g.* 餘、許、足、之多、出頭、好幾、開外、整、正
position	*e.g.* 桌子、院子、地、屋子、池、腔、家子
container	*e.g.* 盒(子)、匣(子)、箱(子)、櫃子、櫥(子)、籃(子)、簍(子)、爐子、包(兒)、袋(兒)、池子、瓶(子)、桶(子)、聽、罐(子)、盆(子)、鍋(子)、籠(子)、盤(子)、碗、杯(子)、勺(子)、匙(湯匙)、筒(子)、擔(子)、籮筐、杓(子)、茶匙、壺、盅、筐、瓢、鍬、缸

length	*e.g.* 公厘、公分、公寸、公尺、公丈、公引、公里、市尺、營造尺、台尺、吋(inch)、呎(feet)、碼(yard)、哩(mile)、 (海)浬、庹、噚、尺、里、釐、寸、丈、米、厘、厘米、海 哩、英尺、英里、英呎、英寸、米突、米尺、微米、毫米、 英吋、英哩、光年
size	*e.g.* 公畝、公頃、市畝、營造畝、坪、畝、分、甲、頃、平方公里、平方公尺、平方公分、平方尺、平方英哩、英畝
weight	*e.g.* 公克、公斤、公噸、市斤、台兩、台斤(日斤)、盎司(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、仟克、公毫
volume	*e.g.* 公撮、公升(市升)、營造升、台升(日升)、盎司、品脫(pint)、加侖(gallon)、蒲式耳(bushel)、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、立分公里、立方英尺、石、斛、西西
time	*e.g.* 微秒、釐秒、秒、秒鐘、分、分鐘、刻、刻鐘、點、點鐘、時、小時、更、夜、旬、紀(輪, 12 年) 、世紀、天(日)、星期(禮拜、週、周) 、月、月份、季、年(載、歲) 、週年、周歲、年份、晚、宿、世、輩、輩子、代、學期、學年、年代
address	*e.g.* 國、省、州、縣、鄉、村、鎮、鄰、里、郡、區、站、巷、弄、段、號、樓、衖、市、洲、地、街
place	*e.g* 部、司、課、院、科、系、級、股、室、廳
duration	*e.g* 陣(子)、會、會兒、下子

There are, however, some complications that must be resolved. First of all we have to clarify the dependent relation between the determiner and the measure of a DM in order to construct a correct feature unification process.

3.1 Head Morpheme of a DM Compound

In principle, a dependent head will take semantic representation of its dependent daughters as its features. Usually, determiners are modifiers of measures, such as '這 (this)' and '一 (one)' are modifiers of '碗 (bowl)' in the examples of 這碗, 一碗, 這一碗. For instance, Example (11) has the dependent relations of

NP(quantifier:DM(quantifier:Neu:一|container:Nfa:碗)|Head:Nab:麵)

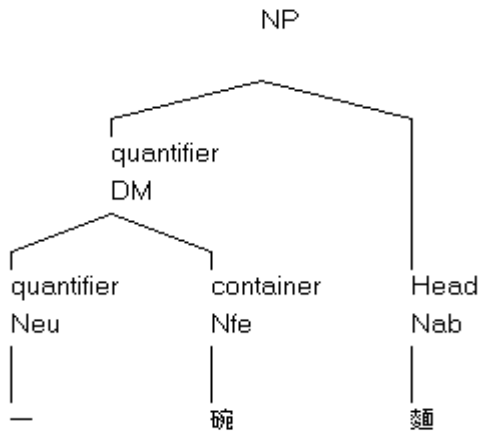

Figure 1. The dependent relations of 一碗麵 "a bowl of noddle".

After the feature unification process, the semantic representation of "一 def: quantity={1}" becomes the feature of its dependent head "碗 def: container={bowl|碗}" and derives the feature representation of "one bowl 一碗def: container={bowl|碗:quantity={1}}". Similarly, "one bowl 一碗" is the dependent daughter of "noodle|麵 def:{noodle|麵}". After the unification process, we derive the result of (11).

(11) one bowl of noodles|一碗麵 def:{noodle|麵:container={bowl|碗:quantity={1}}}

The above feature unification process, written in rule form, is expressed as (12).

(12) Determiner + Measure (D+M) → def: semantic-role(M) = {Sense-representation(M): Representation(D)}

The rule (12) says that the sense representation of a DM compound with a determiner D and a measure M is a unification of the feature representation of D as a feature of the sense representation of M as exemplified in (11).

Nevertheless, a DM compound with a null sense measure word, such as "this copy|這本" , "a copy|一本", or without measure word, such as "these three|這三", will be exceptions, since

the measure word cannot be the semantic head of DM compound. The dependent head of determiners become the head noun of the NP containing the DM and the sense representation of a DM is a coordinate conjunction of the feature representations of its morphemes of determiners only.

For instance, in (10), "copy" has weak content sense; therefore, we regard it as a null-sense measure word and only retain the feature representation of the determiner as the definition of "this copy|這本". The unification rule for DM with null-sense measure is expressed as (13).

(13) Determiner + {Null-sense Measure} (D+M) → def: Representation(D);

If a DM has more than one determiner, we can consider the consecutive determiners as one D and the feature representation of D is a coordinate conjunction of the features of all its determiners. For instance, "this one|這一" and "this one|這一本" both are expressed as "quantifier={definite|定指}, quantity={1}".

Omissions of numeral determiner occur very often while the numeral quantity is "1". For instance, "這本" in fact means "this one|這一本". Therefore, the definition of (10) should be modified as:

這本 def: quantifier={definite|定指}, quantity={1};

The following derivation rules cover the cases of omissions of numeral determiner.

(14) If both numeral and quantitative determiners do not occur in a DM,
then the feature quantity={1} is the default value of the DM.

Another major complication is that senses of morphemes are ambiguous. The feature unification process may produce many sense representations for a DM compound.

3.2 Sense Disambiguation

Multiple senses will be derived for a DM compound due to ambiguous senses of its morpheme components. For instance, the measure word "頭 (head)" has either the sense of {頭|head}, such as "滿頭白髮 full head of white hair" or the null sense in "一頭牛 a cow". Some DMs are inherent sense ambiguous and some are pseudo ambiguous. For instance, the above

example “一頭” is inherently ambiguous, since it could mean “full head” as in the example of “一頭白髮 full head of white hair” or could mean “one + classifier” as in the example of “一頭牛 a cow”. For inherently ambiguous DMs, the sense derivation step will produce ambiguous sense representations and leave the final sense disambiguation until seeing collocation context, in particular seeing dependent heads. Some ambiguous representations are improbable sense combination. The improbable sense combinations should be eliminated during or after feature unification of D and M. For instance, although the determiner “一” has ambiguous senses of “one”, “first”, and “whole”, “一公尺” has only the sense of “one meter”, so the other sense combinations should be eliminated.

The way we tackle the problem is that first we find all the ambiguous Ds and Ms by looking their definitions shown in Appendix A. We, then, manually design content and context dependent rules to eliminate the improbable combinations for each ambiguous D or M types. For instance, according to Appendix A, “頭” has 3 different E-HowNet representations while it functions as a determiner or measure, *i.e.* “def:{null}”, “def:{head|頭}”, and “def:ordinal={1}”. We write three content or context dependent rules below to disambiguate its senses.

(15) 頭 “head”, Nfa, E-HowNet: “def:{null}” : while E-HowNet of the head word is “動物({animate|生物})” and its subclasses.

(16) 頭 “head“, Nff, E-HowNet: “def:{head|頭}” : while pre-determiner is 一(Neqa) “one” or 滿 “full” or 全 “all” or 整 “total”.

(17) 頭 “first”, Nes, E-HowNet: “def:ordinal={1}” : while this word is being a demonstrative determiner that is a leading morpheme of the compound.

The disambiguation rules are shown in Appendix B. In each rule, the first part is the word and its part-of-speech. Then, the E-HowNet definition of this sense is shown, followed by the condition constraints for this sense. If there is still remaining ambiguity after using the disambiguation rule, we choose the most frequent sense as the result.

3.3 Simplification and Normalization for Sense Representation

Members of every type of determiners and measures are exhaustively listable except numeral determiners. Also, the formats of numerals are various. For example, "5020" is equal to "五零二零" and "五千零二十" and "五千二十". So, we have to unify the numeral representation into a standard form. All numerals are composed of basic numerals, as shown in the regular expressions (2). Their senses, however, are not possible to define one by one. We take a simple approach. For all numerals, their E-HowNet sense representations are expressed as themselves. For example, 5020 is expressed as quantity={5020} and we will not further define the sense of 5020. Furthermore all non-Arabic forms will be converted into Arabic expressions, *e.g.* "五千零二十" is defined as quantity={5020}.

The other problem is that the morphological structures of some DMs are not regular patterns. Take "兩個半(two and a half)" as an example. "半(half)" is not a measure word. So, we collect those words, like "多 (many), 半 (half), 幾 (many), 上 (up), 大 (big), 來 (more)" to modify the quantity definition. So, we first remove the word "半" and define the "兩個" as quantity={2}. As the word "半" means quantity={0.5}, we define the E-HowNet definition for "兩個半" as quantity={2.5}. For other modifiers such as "多 (many), 幾 (many), 餘 (more), 來 (more)," we use a function over() to represent the sense of "more", such as "十多個 more than 10" is represented as quantity={over(10)}.

In E-HowNet, complex word senses are expressed by some limit number of basic or primitive concepts. Nevertheless, some certain domain concepts can hardly be expressed by primitive concepts, for instance "焦耳 (joule)," "盧比 (rupee)," "五千零二十 (five thousand and twenty)," etc.. Therefore, we simplify our representations and consider many domain specific concepts as basic concept without further decomposing into primitive concepts.

Appendix A shows the determiners and measures used and their E-HowNet definition in our method. Now, we have the basic principles for compositing semantics of DM under the framework of E-HowNet.

The following steps show how we process DMs and derive their E-HowNet definitions from an input sentence.

I. Input: a Chinese sentence.
II. Apply regular expression rules for DM to identify all possible DM candidates in the input sentence.
III. Segment DM into a sequence of determiners and measures.
IV. Normalize numerals into Arabic form if necessary
V. Apply feature unification rules (12-14) to derive candidates of E-HowNet representations for every DM.

VI. Disambiguate candidates for each DM if necessary.

VII. Output: DM Compounds in E-HowNet representation.

For an input Chinese sentence, we use the regular expression rules created by Li *et al.* (2006) to identify all possible DMs in the input sentence. Then, for every DM compound, we segment it into a sequence of determiners and measures. If any numerals exists in the DM, every numeral is converted into decimal number in Arabic form. For every DM, we follow the feature unification principles to composite semantics of DM in E-HowNet representations and produce possible ambiguous candidates. Then, the final step of sense disambiguation will be carried out.

4. Experiments and Discussion

A corpus-based approach was adopted in developing our proposed method. We need a developing set to derive an exhaustive list of determiners and measures. We try to extract DMs and their morpheme components, *i.e.* determiners and measures, from the developing set and observe the instances of DM to decide their senses and sense representations. Furthermore, sense disambiguation rules will also be developed according to the context of sense ambiguous instances. First, we need to know how many DMs are sufficient to derive a list of determiners and measures with high coverage, if it is not exhaustive. Therefore, we extract DMs from different size subsets of Sinica Treebank and observe their character token coverage. The results are shown in Table 2 and Figure 2. We find that the set of determiners and measures extracted from more than 15000 sentences is sufficient to cover more than 99% of DM instances in the Sinica Treebank.

Table 2. The character token coverage of different subsets of Sinica Treebank

Sentences	0	5000	10000	15000	20000	25000	30000
DM char distribution coverage	0	0.971816	0.987363	0.994014	0.996259	0.997755	0.999169

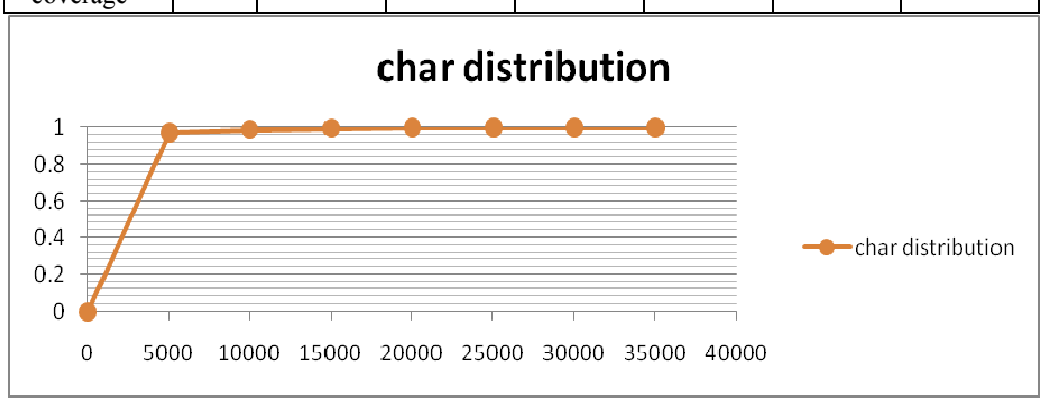

Figure 2. The growth diagram of DM character token coverage.

Therefore, we randomly selected 16070 sentences from Sinica Treebank as our development set and 10000 sentences as our testing set. The development set contained 3753 DM tokens and the testing set contained 1604 DM tokens. We used the development set to derive lexical sense representations and to design disambiguation rules. A total of 405 determiner types and 211 measure types were found, in which 367 out of the 405 determiners were numerals. Since the numbers of numeral determiners are infinite, all numerals will be converted into their Arabic form automatically instead of representing their E-HowNet sense representations individually. The rest of the determiners and measures are encoded with their E-HowNet sense representations manually. For words with ambiguous senses, we also derived their disambiguation rules according to their contextual information shown in development corpus. Finally, a total of 40 disambiguation rules were developed, as shown in Appendix B.

The sense representations of a DM compound will then be derived by a semantic composition process under the framework of E-HowNet. The evaluation of the sense derivation for DM compounds can be divided into two parts: the first part is the correctness of the semantic composition process, and the second part is the correctness of the sense disambiguation process.

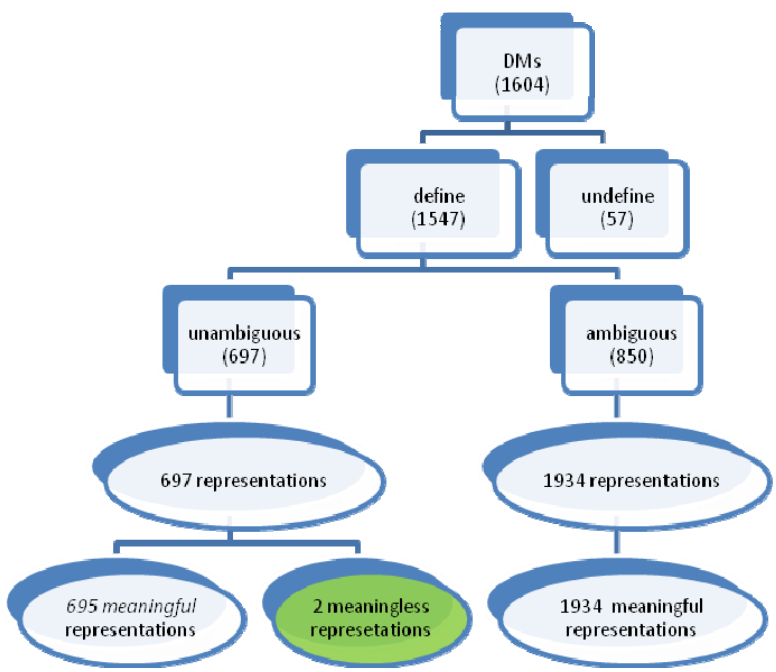

Figure 3. The evaluation result of the semantic composition process.

Figure 3 shows the evaluation result of the semantic composition process. The semantic composition process produced 2631 representations from 1604 words. The program failed to produce E-HowNet representations for the remaining 57 words because of undefined

morphemes. Ambiguous senses were found in 850 words out of the 1604 words. The quality of the result candidates is pretty good. Table 3 shows some sample results. For testing the correctness of our candidates, we checked the formats of 2631 candidates manually. Only 2 candidates out of 2631 displayed wrong or meaningless representations, with both coming from unambiguous words. Therefore, the covering ratio of semantic composition process, *i.e.* deriving meaningful representation without considering sense correctness, is 96% ((1547-2)/1604).

Table 3. Sample results of semantic composition for DM compounds.

DM Compounds	E-HowNet Representation
二十萬元	def:role={money|貨幣:quantity={200000}}
另一個	def:qualification={other|另},quantity={1}
二百三十六分	def:role={score|分數:quantity={236}}
前五天	def:time={day|日:qualification={preceding|上次}, quantity={5}}
一百一十六點七億美元	def:role={USD|美元:quantity={11670000000}}

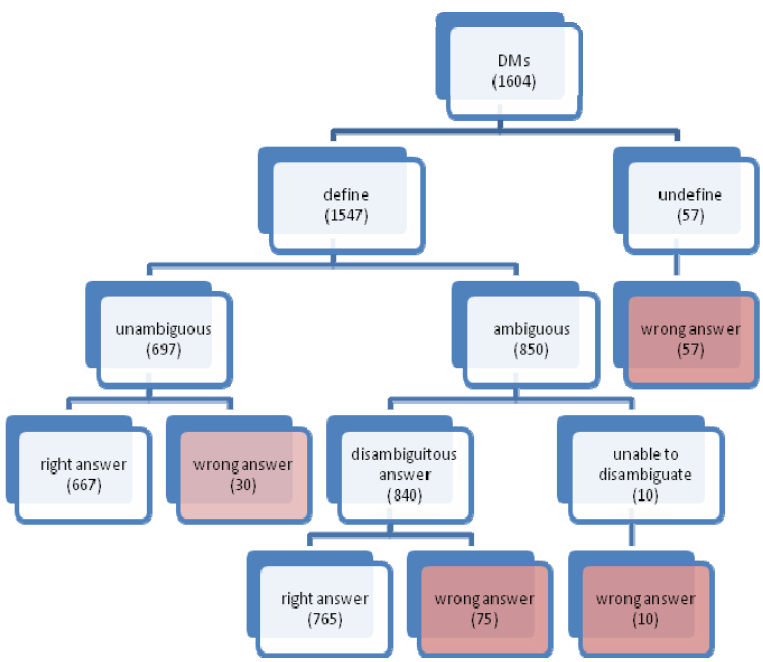

Figure 4. The accuracy of composed sense for DM compounds.

For checking sense correctness, after the disambiguation processes, the resulting E-HowNet representations of 1604 DM tokens in their context were judged manually. Among them, 850 token DMs were sense-ambiguous and the composition process failed to generate answers for 10 of them. Therefore, the composition rules cover 98.8% (840/850) of the ambiguous DM tokens and the precision of the disambiguation rules is 91% (765/840). In all, there are 1432 correct E-HowNet representations for 1604 DM tokens in both sense and format, *i.e*. the current model achieves 89% ((667+765)/1604) token accuracy. Among the 172 wrong answers, 57 errors are due to undefined ambiguous morpheme sense, 30 errors are unique but the wrong answer, and there are 85 sense disambiguation errors.

After data analysis, we conclude the following error types.

A. Unknown domain error:

七棒 "7th batter", 七局 "7th inning"

As there is no text related to the baseball domain in the development set, we get poor performance in dealing with text about baseball. The way to resolve this problem is to increase the coverage of sense representations and disambiguation rules for the baseball domain.

B. Sense ambiguities:

In the following parsed phrase, NP(property:DM:上半場 "first half "|Head:DM:二十分 "twenty minutes or twenty points"), the E-HowNet representation of 二十分 "twenty minutes or twenty points" can be defined as "def:role={score|分數:quantity={20}}" or "def:time={minute|分鐘:quantity={20}}". More contextual information is required to resolve such kinds of sense ambiguity.

For the type of unknown domain error, the solution is to expand the disambiguation rules and the sense representations for morphemes. For sense ambiguities, we need more information and better features to determine true senses.

5. Conclusion

E-HowNet is a lexical sense representational framework and intends to achieve sense representation for all compounds, phrases, and sentences through automatic semantic composition processing. For this purpose, we defined word senses of the CKIP Chinese lexicon in E-HowNet representation. Then, we tried to automate semantic composition for phrases and sentences. Nevertheless, many unknown words or newly coined compound words may occur in the target sentences. In fact, DM compounds are the most frequently occurring unknown words. Therefore, our first goal was to derive the senses of DM words automatically.

In this paper, we take DMs as an example to demonstrate how the semantic composition mechanism works in E-HowNet to derive the sense representations for all DM compounds. We analyze morphological structures of DMs and derive their morphological rules in terms of regular expression. Then, we defined the sense of all determiners and measures in E-HowNet format exhaustively. We created some simple composition rules to produce candidate sense representations for DMs. Then, we reviewed the development set to write some disambiguation rules. We used these heuristic rules to determine the final E-HowNet representation and reach 89% accuracy. The current version did not exhaustively collect all determiners and measures. The system, however, can be improved by gradual extension of the representations of new determiners and measures without retraining.

In the future, we will use similar methods to handle general compounds and to improve sense disambiguation and semantic relation identification processing. We intend to achieve semantic compositions for phrases and sentences in the future and we had shown the potential in this paper.

Acknowledgement

This research was supported in part by the National Science Council under a Center Excellence Grant NSC 96-2752-E-001-001-PAE and Grant NSC96-2221-E-001-009.

Reference

Chao, Y. R. (1968). *A grammar of Spoken Chinese,* University of California Press, Berkeley.

Chen, K. J., Huang, S. L., Shih, Y. Y., & Chen, Y. J. (2005a). Extended-HowNet- A Representational Framework for Concepts. In *Processing of OntoLex 2005 - Ontologies and Lexical Resources IJCNLP-05 Workshop*, Jeju Island, South Korea.

Dong, Z. D. & Dong, Q. (2006). *HowNet and the Computation of Meaning*, World Scientific Publishing Co. Pte. Ltd.

Duchier, D., Gardent, C., & Niehren, J. (1999). Concurrent constraint programming in Oz for natural language processing. Lecture notes, http://www.ps.uni-sb.de/~niehren/oz-natural-language-script.html.

Huang, S. L., Chung, Y. S., & Chen, K. J. (2008). E-HowNet- an Expansion of HowNet. In *Proceedings of 1st National HowNet Workshop*, Beijing, China.

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

Li, S. M., Lin, S. C., Tai, C. H., & Chen, K. J. (2006). A Probe into Ambiguities of Determinative-Measure Compounds. *International Journal of Computational Linguistics & Chinese Language Processing*, 11(3), 245-280.

Mo, R. P., Yang, Y. J., Chen, K. J., & Huang, C. R. (1991). Determinative-Measure Compounds in Mandarin Chinese: Their Formation Rules and Parser Implementation. In

Proceedings of ROCLING IV (R.O.C. Computational Linguistics Conference), National Chiao-Tung University, Hsinchu, Taiwan, 111-134.

Tai, J. H.Y. (1994). 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, 479-494.

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

陳怡君(Chen, Y. C.), (2005b), 黃淑齡, 施悅音, 陳克健, "繁體字知網架構下之功能詞表達初探", 收錄於*第六屆漢語詞彙語意學研討會論文集*, 廈門大學, 中國。

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

Appendix A. Determiner and measure words in E-HowNet representation

定詞(Determiners)

定指

D1->這、那、此、該、本、貴、敝、其、某、諸 def: quantifier={definite|定指}；這些、那些 def: quantifier={definite|定指}, quantity={some|些}

D2->第、首 def: ordinal={D4}

D3->上、前 def: qualification={preceding|上次}；下、後 def: qualification={next|下次}；頭、首 def:ordinal={1}；末 def: qualification={last|最後}；次 def:ordinal={2}

不定指

D4->一、二、萬、雙... def: quantity={1、2、10000、2...} or def:ordinal={1、2、10000、2...}

D5->甲、乙... def: ordinal={1、2...}

D6->其他、其它、其餘、別、旁、他、另、另外 def: qualification={other|另}

D7->每、任何、一、全、滿、整、一切 def: quantity={all|全}

D8->各 def: qualification={individual|分別的}

D9->若干、有的、一些、部份、有些、部分、些 def: quantity={some|些}

D10->半 def: quantity={half|半}

D11->多少、幾多、若干 def: quantity={.Ques.}

D12->何、啥、什麼 def: fomal={.Ques.}

D13->數、許多、很多、好多、好幾、好些、多、許許多多、多數、大多數、不少、泰半、半數、諸多 def: quantity={many|多}；少許、少數、幾許、個把 def: quantity={few|少}；幾 def:quantity={some|些}

D14->餘、許、之多、來 def: approximate()；足、整、正 def: exact()；出頭、數、好幾、幾、開外、多 def: over();

D15->0、1、2、3、4、5、6、7、8、9、０、１、２、３、４、５、６、７、８、９ def: quantity={1、2、3、4...}

量詞(Measure word)

有語意量詞(Measures with content sense)

Nff->暫時量詞—身、頭、臉、鼻子、嘴、肚子、手、腳 def:{身,頭, …}

Nff->暫時量詞—桌、桌子、院子、地、屋子、池、腔、家子 def: position={桌子,院子...:quantity={all|全}}

Nfe->容器量詞—

盒(子)、匣(子)、箱(子)、櫃(子)、櫥(子)、籃(子)、簍(子)、爐(子)、包(兒)、袋(兒)、池子、瓶(子)、桶(子)、罐(子)、盆(子)、鍋(子)、籠(子)、盤(子)、碗、杯(子)、勺(子)、匙(湯匙)、筒(子)、擔(子)、籮筐、杓(子)、茶匙、壺、盅、筐、瓢、鍬、缸 def: container={盒,匣,...}

Nfg->標準量詞—

表長度的，如：公厘、公分、公寸、公尺、公丈、公引、公里、市尺、營造尺、台尺、吋(inch)、呎(feet)、碼(yard)、哩(mile)、(海)浬、海里、庹、噚、尺、里、釐、寸、丈、米、厘、厘米、海哩、英尺、英里、英呎、英寸、米突、米尺、微米、毫米、英吋、英哩、光年。 def: length={公分,...}

表面積的，如：公畝、公頃、市畝、營造畝、坪、畝、分、甲、頃、平方公里、平方公尺、平方公分、平方尺、平方英哩、英畝。def: size={公畝,...}

表重量的，如：公克、公斤、公噸、市斤、台兩、台斤(日斤)、盎司(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、仟克、公毫。def: weight={公克,...}

表容量的，如：公撮、公升(市升)、營造升、台升(日升)、盎司、品脫(pint)、加侖(gallon)、蒲式耳(bushel)、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、立分公里、立方英尺、石、斛、西西。def: volume={公撮,公升,...}

表時間的，如：微秒、釐秒、刻、刻鐘、點、點鐘、更、旬、紀(輪, 12年) 、世紀、季 def:time={微秒,釐秒,…}；秒、秒鐘 def:time={second|秒}；分、分鐘 def:time={minute|分鐘}；時、小時 def:time={hour|時}；夜、晚、宿 def:time={night|夜}；天(日) def:time={day|日}；星期(禮拜、週、周) def:time={week|周}；月、月份 def:time={month|月}；年、載、歲、年份 def:time={year|年}；週年、周歲 def:duration={年}

表錢幣的，如：元(圓)、塊、兩 def:role={money|貨幣}；分、角(毛)、先令、盧比、法郎(朗)、辨士、馬克、鎊、盧布、美元、美金、便士、里拉、日元、日圓、台幣、港幣、人民幣。def: role={分, …,盧布…}

其他：刀、打(dozen)、令、綸(十條)、蘿(gross)、大籮(great gross)、焦耳、千卡、仟卡、燭光、千瓦、仟瓦、伏特、馬力、爾格(erg)、瓦特、瓦、卡路里、卡、仟赫、位元、莫耳、毫巴、千赫、歐姆、達因、兆赫、法拉第、牛頓、赫、安培、周波、赫茲、分貝、毫安培、居里、微居里、毫居里 def: quantity={刀,打,…,焦耳,...}

Nfh->準量詞—

指行政方面，如：部、司、課、院、科、系、級、股、室、廳。def: location={部,

司...}

指時間方面，如：世、輩、輩子、代、學期、學年、年代 def: time={學期, 年代,...} 會、會兒、陣(子) 、下(子) def: duration={TimeShort|短時間}

指方向的，如：面(兒)、方面、邊(兒)、方 def: direction={EndPosition|端}；頭(兒) def: direction={aspect|側}

指音樂的，如：拍、小節。def: quantity={拍,板...}

指分數，如：分 def:role={分數:quantity={D4,D15}}；

Nfi->動量詞—

指頻率的，如：回、次、遍、趟、下、巡、遭、響、圈、把、關、腳、巴掌、掌、拳頭、拳、眼、口、刀、槌(子)、板(子)、鞭(子)、棒、棍(子)、針、槍矛、槍、砲、度、輪、周、跤、回合、票 def:frequency={D4, D15}；步 def:{步}；箭 def:role={箭:quantity={D4,D15}}；曲 def:{曲:quantity={D4,D15}}

Nfc->群體量詞—

對、雙 def:quantity={double|複}；

列(系列)、排 def:quantity={mass|眾:manner={ InSequence|有序}}；

套 def:quantity={mass|眾:manner={relevant|相關}}；

串 def:quantity={mass|眾:dimension={linear|線}}；

掛、幫、群、伙(夥)、票、批 def: quantity={mass|眾}；

組 def: quantity={mass|眾: manner={relevant|相關}}；

窩 def: quantity={mass|眾:cause={assemble|聚集}}；

種、類、樣 def: {kind({object|物體})}；

簇 def:quantity={mass|眾:cause={assemble|聚集}}；

疊 def:quantity={mass|眾:cause={pile|堆放}}；

紮 def:quantity={mass|眾:cause={wrap|包紮}}；

叢 def:quantity={mass|眾:cause={assemble|聚集}}；

隊 def:quantity={mass|眾:manner={InSequence|有序}}；

式 def:{kind({object|物體})}

Nfd->部分量詞—

些 def:quantity={some|些}；

部分(份)、泡、綹、撮、股、灘、汪、帶、截、節 def: quantity={fragment|部}；

團 def: quantity={fragment|部:shape={round|圓}}；

堆 def: quantity={ fragment|部:cause={pile|堆放}}；

把 def: quantity={ fragment|部:cause={hold|拿}}；

層、重 def: quantity={ fragment|部:shape={layered|疊}}

Nfa->個體量詞

號 def:ordinal={}

無語意量詞(null-sense Measures)

Nfa->個體量詞—洞、號、渠、本、把、瓣、部、柄、床、處、期、齣、場、朵、頂、堵、道、頓、錠、棟(幢)、檔(檔子)、封、幅、發、分(份)、人份、紙、服、個(箇)、根、行、戶、件、家、架、卷、具、闕、句、屆、捲、劑、隻、尊、盞、張、枝(支)、樁、幀、只、株、折、炷、軸、口、棵、款、客、輛、粒、輪、枚、面、門、幕、匹、篇、片、所、艘、扇、首、乘、襲、頭、條、台、挺、堂、帖、顆、座、則、冊、任、尾、味、位、頁、葉、房、彎、班、員、科、丸、名、項、起、間、題、目、招、股、回、線、灣。def: {null}

Nfc->群體量詞—宗、畦、餐、行、副(付)、蓬、筆、房、綑(捆)、胎、嘟嚕、部、派、路、壟、落、束、席、色、攤、項、疊、紮。def: {null}

Nfd->部分量詞—口、塊、滴、欄、捧、抱、段、絲、點、片、縷、坨、匹、疋、階、抔、波、道。def: {null}

Nfb->述賓式合用的量詞—通、口、頓、盤、局、番。def: {null}

Nfi->動量詞—回、次、遍、趟、下、遭、聲、響、圈、把、仗、覺、頓、關、手、(巴)掌、拳(頭)、眼、口、槌(子)、板(子)、鞭(子)、棒、棍(子)、針、箭、槍(矛)、砲、度、輪、曲、跤、記、回合、巡、票。def: {null}

Nfh->準量詞

指書籍方面，如：版、冊、編、回、章、面、小節、集、卷。def: {null}

指筆劃方面，如：筆、劃(兒)、橫、豎、直、撇、捺、挑、剔、鉤(兒)、拐、點、格(兒)。def: {null}

其他：

程、作(例:一年有兩作)、倍、成。def: {null}

厘(例:年利五厘、一分一厘都不能錯)。def: {null}

毫(萬分之一)、絲(十萬分之一)(例:一絲一毫都不差)。

圍、指、象限、度。def: {null}

開(指開金)、聯(例:上下聯不對稱)。def: {null}

軍、師、旅、團、營、伍、班、排、連、球、波、端。def: {null}
樓、城(扳回一城)、回合、折、摺、流、等、桿、聲、次。def: {null}

Appendix B. The rules for candidate disambiguation

Head-Based Rules

Rule 1 一, Neu, def:quantity={1}, while part-of-speech of the head word is Na, except the measure word is 身 "body" or 臉 "face" or 鼻子 "nose" or 嘴 "mouth" or 肚子 "belly" or 腔 "cavity" .

Rule 2 塊, Nfg, def:role={money|貨幣}, while E-HowNet representation of the head word is "{money|貨幣}" or {null}, or the head word is 錢 "money" or 美金 "USD" or the suffix of word is 幣 "currency" and previous word is not D1.

Rule 3 塊, Nfd, def:{null}, otherwise, use this definition.

Rule 4 面,Nfa,def:{null}, while part-of-speech of the head word is Nab.

Rule 5 面,Nfh,def:direction={aspect|側}, otherwise use this one.

Rule 6 頭,Nfa,def:{null}, while the head word is Nab and E-HowNet representation of the head word is "動物{animate|生物}".

Rule 7 頭,Nfh,def:direction={EndPosition|端}, if the part-of-speech of the head word is Na, do not use this definition. The prefix determiners are 這 "this" or 那 "that" or 另 "another".

Rule 8 All Nfi, def:frequency={}, while the part-of-speech of the head word is Verb, i.e. E-HowNet representation of the head word is {event|事件} and it's subclass. Except POS of the head are V_2 and VG, and if the word is {次、回、口}, do not use this rule.

Rule 9 All Nfi, def:{null}, otherwise use this one. If the head word is {次、回、口}, do not use this rule.

Rule 10 部, 股…,Nfh,def:location={ }, if part-of-speech of the head word is Na or prefix determiner is 這"this" or 那"that" or 每"every", do not use this definition.

Rule 11 部,股…,Nfa,def:{null}, otherwise use this definition.

Rule 12 盤,Nfe,def:container={plate|盤},while the head word is food, *i.e*. E-HowNet representation of the head word is {edible|食物} and its subclasses.

Rule 13 盤,Nfb,def:{null},otherwise use this one.

Rule 14 分,Nfg, def:role={分}, while the head word is 錢 "money", *i.e*. E-HowNet representation of the head word is {money|貨幣} and its subclasses.

Rule15 分,Nfg, def:size={分}, while the head word is 地 "land", *i.e*. E-HowNet representation of the head word is {land|陸地} and its subclasses.

Rule 16 分,Nfa,def:{null}, while part-of-speech of the head word is Na or Nv. For example: 一分耕耘；十分力氣；五分熟.

Rule 17 點,Nfh;Nfd,def:{null}, while part-of-speech of the head word is Nab. If part-of-speech of the head word is V, Naa or Nad, do not use this definition.

Rule 18 度,聲, def:frequency={}, while part-of-speech of the head word is Verb.

Rule 19 度,聲, def:{null}, otherwise use this definition.

Collocation-Based Rules

Rule 20 分,Nfh,def:role={score|分數:quantity={D4,D15}}, while the sentence also contains the words 考 "give an exam" (E-HowNet representation is {exam|考試}) or 得 "get" (E-HowNet representation is {obtain|得到}) or 失 "lose" (E-HowNet representation is {lose|失去}) or E-HowNet representation of the head word is {hold|拿},{catch|捉住},{occupy|佔領},{rob|搶},{win|獲勝},{forming|形成},{add|增加},{suffer|遭受},{sink|下沉},{inferior|不如} and its subclasses, or the sentence contains the word 成績 "score", X 局 (for example,一局 "the first inning"), X 半場 (for example, 上半場 "the first half in game"), then use this definition.

Rule 21 分,Nfg,def:time={minute|分鐘}, if the sentence contains the word 時 "hour", 鐘頭 "hour", X 時(for example,五時 "5 o'clock") or X 秒 (for example,三十秒 "30 seconds").

Rule 22 兩,Nfg,def:weight={兩}, if the sentence contains the word 重 "weight" or 重量 "weight".

Rule 23 兩,Nfg,def:role={money|貨幣}, if the sentence contains the word 銀 "silver" or 錢 "money" or 黃金 "gold"

Pre-Determinant-Based Rule

Rule 24 頭, Nff,def:{head|頭}, while the pre-determinant is 一(Neqa) "one" or 滿 "full" or 全 "all" or 整 "total".

Rule 25 腳, Nff,def:{leg|腳}, while pre-determinant is 一(Neqa) "one" or 滿 "full" or 全 "all" or 整 "total" and the part-of-speech of the head word is not Na.

Rule 26 腳, Nfi,def:frequency={}, while part-of-speech combination is V+D4,D15+腳.

Rule 27 點,Nfg, def:time={點}, while part-of-speech of pre-determiners are D4 or D15(1~24) and part-of-speech of the previous word is not D1 or the previous word is not 有 "have".

Determinative-Based Rule

Rule 28 一、二...1、2...兩..., Neu, def:ordinal={}, the determiners are 第, 民國, 公元, 西元, 年號, 一九 XX or 12XX, (four digits number).

Rule 29 一、二...1、2...兩..., Neu,def:quantity={}, otherwise use this definition.

Rule 30 頭,Nes,def:ordinal={1},the word 頭 “head” is a determiner.

Rule 31 兩,Neu,def:quantity={}, the word 兩 “a unit of weight equal to 50 grams” is a determiner.

Measure Word-Based Rule

Rule 32 一,Neqa,def:quantity={all|全}, the part-of-speech of the measure word behind 一 is Nff, or the suffix of the measure word is 子, (for example, 櫃子 “cabinet”, 瓶子 “bottle”) or 籮筐 “large basket”.

Rule 33 一、二...1、2...兩..., Neu,def:ordinal={}, if measure word is 歲.

Head and Determinantive-Based Rule

Rule 34 次,Nfi,def:frequency={}, while part-of-speech of the head word is a Verb (Except POS V_2 and VG.), and determiners are not D1,D2,D3.

Rule 35 次,Nfi;Nfh,def:{null}, otherwise use this definition.

Rule 36 口,Nfa.def:{口:quantity={全}}, while the pre-determiners are 滿,全,or 整.

Rule 37 口,Nfi.def:frequency={}, while part-of-speech of the head word is Verb, and the pre-determiner is not 滿,全,or 整.

Rule 38 口,def:{null}, otherwise, while the pre-determiner is D4 or D15, use this definition.

Rule 39 回, def:frequency={}, while part-of-speech of the head word is Verb (Except POS V_2 and VG), and the determiner is not D1,D2,D3.

Rule 40 回, def:{null}, otherwise use this definition.