Analysis of Chinese Morphemes and Its Application to Sense and Part-Of-Speech Prediction for Chinese Compounds

You-shan Chung and Keh-Jiann Chen
Institute of Information Science
Academia Sinica
Taipei, Taiwan
{yschung, kchen}@iis.sinica.edu.tw

***Abstract*—Since new compounds are generated very productively in Chinese, an automatic scheme is required to predict their part-of-speeches and senses in order to automate computer language processing. To this end, we analyzed the morpho-syntactic behaviors of about 4,025 morphemes (characters) in our Affix database. We found that semantic and logical compatibility are more important than syntactic constraints in compounding. Hence, we classified morphemes into four major semantic types: object, act, attribute and value, and use semantic composition rules to predict the meaning and part-of-speech of compounds. Some morpheme types and composition rules are ambiguous. We propose constraint-based resolutions to deal with them.**

Keywords- Chinese compounds, semantic type, affix, part-of-speech prediction, sense disambiguation

I. INTRODUCTION

Compounding by joining two monosyllabic morphemes is highly productive in Chinese [1]. In the Sinica Corpus [2], about 2% of the words are unknown compounds whose part-of-speeches (henceforth POS) and meaning have to be figured out. The current study deals with the compounds with lexical transparency, namely those whose meaning is a composition of its components. The first step toward understanding an unknown word is to know its POS [3]. Although it is possible to predict the POS of a word without knowing the POS of its components [3], this approach cannot attain high prediction accuracy as the factors that determine the compound's POS are not investigated. Tseng et al. [4] thus combined information from the meaning of the modifying component of compounds and from morphemes or composition patterns suggesting membership of a certain POS to predict a compound's POS. In another analytic effort, Tseng and Chen [5] study the relationship between the POS of a compound and the component morphemes, enabling an analyzer that gives the morpho-syntactic structure of compounds. With the analyzer, Tseng [6] and Shi et al. [7] predict the meaning of a compound by measuring the semantic distances between the modifying component of the target compound and those of other compounds based on the semantic distance between them on the taxonomies of CiLin[1] and E-HowNet [8]. However, a pitfall of this approach is that compounds for comparison are sometimes lacking.

Departing from a POS-based approach and thus avoiding reliance on similar compounds for comparison, Liu [9] explains the meaning and syntactic behaviors of compounds in a cross-POS analysis of component morphemes as object,

This research was supported in part by the National Science Council of Taiwan under Grants NSC 98-2631-H-001-013 and NSC 96-2661-E-001-023-MY3

[1] A thesaurus that divides words into 12 main semantic categories and other subcategories, reviewed in [4].

act, attribute and value.

The types and order of morphemes interact to create compounds of various types that differ in POSes and meaning. Knowing the types of the components makes automatic understanding possible. Therefore, we labeled morphemes in our Affix database [10] as denoting objects, acts, attributes and values.

The paper is organized as follows. In Section 2, we explain the definitions of the types and their morpho-syntactic properties. Section 3 presents various morpho-syntactic constructions and the corresponding rules that govern the combination of morphemes. In Section 4, the issue of ambiguity is addressed. Summarization and conclusion are drawn in Section 5.

II. Object, Act, Attribute and Value- Definitions and Their Morpho-Syntactic Properties

When it comes to morphological constructions, semantic and logical compatibility are more important than syntactic compatibility. Liu's [9] framework contains four major semantic types, i.e. *object*, *act*, *attribute*, and *value*, which explain how things are described in Chinese. The semantic types do not completely coincide with POSes. Different semantic types tend to occur in different morphological positions and play different semantic roles.

A. Objects

Objects refer to entities, concrete (e.g. pencil, person, air, Taipei) or abstract (e.g. culture, thinking, physics, joy), and are common nouns. They usually play the roles of noun-phrase (NP) heads and modifiers of nouns. When an *object* modifies another object, their semantic relation could be very complicated, e.g. telic, agentive, material, part, location, etc. The POS and semantic type of the resulting compound are usually the same as those of the head morpheme, which tends to occur in the suffix position.

B. Acts

Acts refer to actions, e.g. run, write, encourage and calculate. Acts are generally verbs. They play head roles of verb phrase (VP) and the resulting compounds are verbs. However, nominalized acts as suffixes result in nominal compounds, e.g. 海*釣* 'sea fishing', 聯*考* 'joint exam', 國*防* 'national defense'. Directional actions, such as 起來 and 進去, function as complements as in 站起來|stand up, 走進去|walk in.

C. Attributes

Attributes refer to a special class of nouns with distinct meaning and syntactic behaviors. They cannot independently denote things as ordinary object nouns do but denote particular properties of entities. For example, a cup must have properties like *weight*, *color* and *shape*, etc. We call these properties 'attributes'. Attributes are less independent than objects in the sense that they are vague if not associated with their host entities and associated values.

Attribute nouns seldom modify another attribute or entity because attributes represent aspects to be described by values but not values themselves.

In our Affix database, prefixes that denote attributes only account for 1.33% of the 1,870 prefix morphemes listed. By contrast, suffix attributes are productive enough to be freely coined with prefix objects, acts and values, mostly to form attribute-type nominal compounds, e.g. 車速|car speed, 紅色|red and 展期|exhibition-period.

D. Values

A morpheme that modifies an attribute or object belongs to the *value* type. In most cases, values are adjectives which refer to descriptive properties. For example, the attribute *appearance* has values like 美|good-looking and 醜|ugly. But values are not restricted to adjectives as some nouns can also denote values, such as values of measurement. For example, 高度|height is an attribute that has to be specified by nominal measurement values like 三公尺|3 meters. Sometimes, values refer to the content of an attribute. For example, *material* is an attribute whose values can be nouns like cotton, wool, or polyester in 衣料|fabric. Based on whether they modify object or event attributes, values can be accordingly classified into object and event values. For example, since the attribute *duration* describes the temporal properties of an event, 長期|long-term, which refers to a long period of time, is an event-value.

III. The Major Morphological Combinations of the Four Types

With morphemes in the Affix database labeled based on Liu's theory [9], we present the major combinations and their meaning and POS prediction below, which are determined by the head of the compound. We also provide the constraints for rule application to resolve rule ambiguities.

A. *Coordinate Construction*

Object1+Object2→Object3 (1)

Act1 + Act2→ Act3 (2)

Value1 + Value2→ Value3 (3)

Constraint: Semantic-type (M1) = Semantic-type (M2); M1 M2 are near-synonyms

POS: POS (M1 or M2);

Semantics: And (M1, M2)

Examples: 手腳|hand and foot, 打擊|hit and attack, 高大|tall and big

Value1+Value2→Attribute1 (4)

Constraint: Value1 and Value2 are antonyms

POS: noun

Semantics: Attribute1, whose values are Value1 and Value2

Examples: 好壞|good or bad, 興衰|rich or poor, 是非|right or wrong

B. *Modifier-Head Construction*

Combinations (5-9) are head-final constructions, and (10-11) are head-initial construction.

Value1 + Object1→Object2 (5)

Constraint: Object1 must have an attribute that has Value1 as one of its values

POS: noun

Semantics: Object1 has an attribute modified by Value1.

Examples: 紅花|red flower, 新衣|new clothes, 快鍋|pressure cooker

Object1 + Object2 → Object3 (6)

Constraint: few if not none

POS: noun

Semantics: determined by world knowledge

Examples: 手錶|watch, 桌布|table cloth

Object1 + Attribute1 → Attribute2 or Value1 (7)

Case 1: Object1 +Attribute1→Attribute2

Constraint: more than one possible value of Object1 for Attribute1

POS: noun

Semantics: Object1's Attribute1.

Examples: 雲量|quantity of clouds, 花期|blooming season, 車速|car speed, 腳程|walking speed

Case 2: Object1 +Attribute1→Value1

Constraint: a unique value of Object1 for Attribute1

POS: adjective

Semantics: the value of Attribute1 of Object1

Examples: 星形|star-shaped, 陶質|pottery, 法式|French-style

Act1 + Event-Attribute1 → Event-Attribute2 (8)

Constraint: Act1 has Event-Attribute1

POS: noun

Semantics: Event-Attribute1 of Act1

Examples: 展期|exhibition period, 航速|cruising speed

Value1 + Act1 → (nominalized) Act2 (9)

Case 1: Value1 + Act1 →Act2

Constraint: Value1 is a value of one of Act1's attributes

POS: the POS of Act

Semantics: Act1 happens in the way described by Value1

Examples: 靜思|think quietly, 快煮|fast cook

Case 2: Value1 +Act1 →nominalized Act2

Constraint: Act1 as a suffix acts mostly as nominalized verbs, e.g. 跑|run, 舞|dance, 釣|fish.

POS: noun (nominalized verb)

Semantics: Act2 seen as a nominal entity

Examples: 定存|deposit, 長跑|long-distance running, 熱吻 |passionate kiss

Value1 + Attribute1 → Value2 (10)

Constraint: Value1 is one of Attribute1's own values

POS: adjective

Semantics: the value of Attribute1 is Value1

Examples: 軟質|soft, 新型|new-style, 橫式|horizontal-style, 假性|pseudo

Object1 + Value1 → Object2 (11)

Constraint: Value has to be one of a set of shape values, such as 串|cluster, 粉|powder, 圈|ring, 丸|ball, etc.

POS: noun

Semantics: Object has a shape denoted by Value

Examples: 肉丸|meat ball, 磚塊|brick

C. *Verb-Complement Construction*

There is a closed set of resulting and directional actions. These actions serve as verb-complements, most of which being stative verbs and belonging to value type.

Act1+Act2→Act3 (12)

Constraint: Act2 is one of a closed set of directional acts, such as 失|lost, 動|move, 來|come, 去|go, 進來|come in

POS: verb

Semantics: Act1 results in Act2

Examples: 尋獲|find, 走失|get lost, 走來|come

Act1 + Value1→Act2 (13)

Constraint: Value1 belongs to stative verbs

POS: verb

Semantics: Act1 results in Value1

Examples: 跑累|tired from running, 氣昏|faint from anger

D. *Verb-Argument Construction*

Act1 + Object1→ Act2/Object2 (14)

Constraint: Object1 satisfies the selectional restriction of Act1

POS: verb, noun, nominalized noun

Semantics: Act1 on Object1 (the result is a verb or a nominalized noun)/Object2 to which Act1 is done onto (the result is noun)

Examples: 買書|buy books, 炒蛋|fry eggs/fried eggs, 跳舞 |dance/dancing

$$\text{Act1+Attribute1} \rightarrow \text{Act2} \qquad (15)$$

Constraint: Attribute1 satisfies the selectional restriction of Act1

POS: verb

Examples:升溫｜temperature rises/raise temperature, 整容|have plastic surgery

$$\text{Object1 + Act1} \rightarrow \text{Act2} \qquad (16)$$

Constraints: Object1 is the instrument or subject of Act1.

POS: verb

Semantics: Act1 is done with Object1 as an instrument or subject

Examples: 槍殺|kill with a gun 火烤|grill, 鳥叫|bird chirps/bird chirping.

IV. AMBIGUITY

While the above-postulated morphological rules can derive the meaning and POS of a new compound, there are several kinds of ambiguities that need to be solved.

A. Ambiguity of Morpheme Senses

Many Chinese morphemes are polysemous. Such cases are given multiple senses in our data base. To determine which sense is relevant, prefix/suffix position resolves sense ambiguity for many morphemes. For example, 單 has at least two meanings when pronounced as *dan*: (a) a leaflet and (b) singular. As a suffix, 單 means ‘a leaflet’, e.g. 成績單|transcript, 帳單|bill. As a prefix, it means ‘singular’, e.g. 單身|single person, 單人床|single bed.

Some morphemes appear to have different senses at the same positions. For example, 機 has many different senses as either prefix or suffix. As a suffix, it can refer to ‘machine’ (e.g. 果汁機|juicer) ‘airplane’ (e.g. 戰鬥機|fighter plane) or ‘opportunity’ (e.g. 商機|business opportunity). To figure out the relevant sense, one of the major approaches is by analogy. For example, to deal with a unknown compound, firstly, we have to find the most similar known words which also have the suffix 機 and whose modifiers are the nearest synonyms of the target compound’s modifier. The sense of the unknown word is predicted to be that of the most similar examples [7].

B. Rule Ambiguity

Another type of ambiguity has to do with different results of the same type combination. For instances, the “object + object” pattern occurs in the coordinate construction (1) and the modifier-head construction (6) at the same time, yielding different readings. The same type combination belonging to the same construction could still lead to different meaning and POS, e.g. rule (3) and (4). In these cases, constraints help identify which rule applies. However, some constraints cannot readily apply. For example, in rule (7), despite the constraint that the result of object+attribute is value when there is only one possible value of object for the attribute and is attribute elsewhere, the question remains as to how to determine whether one value or more is involved.

The semantic role of the object determines how many values are involved. If the semantic role of the object is the host of the attribute, then the compound denotes attribute, otherwise it denotes value. After all, however, it requires world knowledge to know the semantic role of the object. World knowledge tells us that 詞|word and 髮|hair are, respectively, the hosts of property|性 and 型|style and may have many different 性|property and 型|style, while 法|French and 星|star are not the hosts in 法式|French-style and 星形|star-shaped.

Another example where ambiguity still remains regarding which rule to apply is in (14). We are still working to find out how to distinguish whether the result will be act or object.

Inherent POS or sense ambiguities also occur. For instance, the object-attribute construction 木質 could mean either ‘wooden’ as in 木質地板|wooden floor or ‘material of wood’ as in 木質好壞 |wood quality. Such examples can also be seen in the act-object and object-act construction (cf. (14) and (16)). For example 炒蛋 and鳥叫 could have either nominal readings of 'fried eggs' and 'bird chirping' or verbal readings of 'fry eggs' and 'bird sings' respectively

V. CONCLUSION

By manually labeling morphemes in our Affix database as objects, acts, attributes and values and giving them formalized meaning representations, we can automate a major part of understanding the meaning and syntactic behaviors of combinations of morphemes. Such a semantic analytic approach also provides methods for resolving ambiguous interpretations. In the future, we will continue to test whether the current rules suffice to automatically extract the meaning and POSes or not.

REFERENCES

[1] C. N. Li, S. A. Thompson, Mandarin Chinese: A Functional Reference Grammar. Taipei: Crane, 1981.

[2] Sinica Corpus, http://dbo.sinica.edu.tw/SinicaCorpus/

[3] C. J. Chen, M. H. Bai, K. J Chen, "Category guessing for Chinese unknown words," Proceedings of the Natural Language Processing Pacific Rim Symposium 44, pp. 35-40, 1997.

[4] H. Tseng, C. L. Liu, Z. M Gao, "A hybrid approach for automatic classification of Chinese unknown verbs," (以構詞律與相似法為本的中文動詞自動分類研究) Computational Linguistics and Chinese Language Processing, pp.1-28, 2002.

[5] H. Tseng, K. J Chen, "Design of Chinese morphological analyzer," Proceedings of the first SIGHAN Workshop on Chinese Language Processing, pp. 1-7, 2002.

[6] H. Tseng, "Semantic classification of Chinese unknown words," Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, pp. 72 -79, 2003.

[7] Y. Y. Shih, S. L. Huang, K. J. Chen, "Semantic representation and composition for unknown compounds in E-HowNet," Proceedings of PACLIC 20, pp. 378-382, 2006.

[8] K. J. Chen, S. L. Huang, Y. Y. Shih, Y. J. Chen, "Extended-HowNet- A representational framework for concepts," OntoLex 2005 - Ontologies and Lexical Resources IJCNLP-05 Workshop, 2005.

[9] C. H. Liu, Xiandai Hanyu Shuxing Fanchou Yianjiu (現代漢語屬性範疇研究). Chengdu: Bashu Books, 2008.

[10] Affix database, http://140.109.19.103/affix/.