Semantic Annotation of Chinese Phrases Using Recursive-Graph

Ji Donghong
Kent Ridge Digital Labs,
Singapore, 119613
dhji@krdl.org.sg

Abstract

In this paper, we propose a recursive graph based scheme for semantic annotation of Chinese phrases. Compared with others, this scheme can fully differentiate those Chinese phrases that comprise the same content words but hold different meanings due to their different word order or some involved function words, and capture the hierarchical conceptual structure of Chinese phrases, which underlies their main semantic information. We also give the guidelines for annotating various commonly used types of Chinese phrases.

1 Introduction

Semantically annotated linguistic data are important resources for natural language processing, and have been used in many NLP areas, e.g., parsing, word sense disambiguation, co-reference resolution and information extraction, etc. But due to huge efforts needed in building them and general difficulties in dealing with semantic information, such resources are scarcely seen for most languages including Chinese, if not all.

In this paper, we focus on Chinese phrases and present a recursive graph based scheme for their semantic annotation. Compared with the same task for sentences, it only involves a relatively small data set and simple semantic information but has potential generality and application.

As a specific semantic annotation task, two problems should be made clear at first. One is at which goal the annotation is aimed. In our case, the goal is established to be to help disclose the correspondence between linguistic forms and meaning, which is also the primary goal in both traditional linguistic research (Chomsky, 1968) and natural language understanding (Allen, 1995).

The other problem is whether the annotation can significantly contribute to the established goal. In our case, the answer is of course positive. First, the phrases themselves are a specific kind of linguistic forms, thus their semantic annotation directly provides the correspondence between them and their meanings. Second, by some kinds of analogy rules, these annotated phrases can be examples for deriving the correspondence between new phrases and their meanings. Third, as compared with western languages, Chinese language has a specific feature that its sentences form in roughly the same way as its phrases (Zhu, 1982), which shows that by some kinds of combinatory rules, the mapping between sentences and their meanings can also be determined based on the examples.

The remainder of this paper is organised as the following. In section 2, we describe the motivation of introducing recursive graph. In section 3, we formally define what is a recursive graph. In section 4, we specify how commonly used types of Chinese phrases are annotated using recursive graph. In section 5, we give the conclusion and discuss some future work.

2 Motivation

In general, semantic annotation of linguistic forms is to associate with them their semantic information being represented in some formal languages or diagrams. With the semantic information involved varying, the formal languages or diagrams may be different.

One commonly used diagram for semantic annotation of linguistic forms is dependent tree, in which the *dependence* or *control* relationship between constituents of a linguistic form is depicted (Langacker, 1997). But such trees may be not powerful enough to differentiate those Chinese phrases that comprise the same content words but hold different meanings due to their word order or involved function words. As an example, consider 1) and 2)[1].

1) 走私 汽车
/zousi/ /qiche/
smuggle car

2) 汽车 走私
/qiche/ /zousi/
car smuggle

Notice 1) and 2) contain the same content words, but hold different word order. Regarding their meanings, 1) is an ambiguous phrase, corresponding with two English translation phrases as 3) and 4).

3) to smuggle cars

4) smuggled cars

The translation phrase for 2) is 5).

5) the smuggling of cars

So, there are altogether three meanings held by the two phrases. But the two content words can only form two dependent trees, listed in 6) and 7).

6) 汽车 (/qiche/, car)
↑
走私(/zousi/, smuggle)

7) 走私(/zousi/, smuggle)
↑
汽车 (/qiche/, car)

Obviously, these two dependent trees cannot code the three meanings listed in 3), 4) and 5). Only if we could see the same word 走私(/zousi/, smuggle) semantically different in 1) and 2), we could add another dependent tree 8) to code 5), with 6) and 7) corresponding with 3) and 4) respectively.

8) 走私'(/zousi/, smuggle)[2]
↑
汽车 (/qiche/, car)

But this view is quite unintuitive, and will lead to contradictory. Consider other two phrases 9) and 10).

9) 走私 集团
/zousi/ /jituan/
smuggle bloc
smuggling bloc

10) 汽车 走私 集团
/qiche/ /zousi/ /jituan/
car smuggle bloc
car smuggling bloc

Intuitively, the word 走私(/zousi/, smuggle) in 2) holds the same meaning as the word in 10), which subsequently is equivalent with the same word in 9). On the other hand, there is no reason to treat the same word 走私(/zousi/, smuggle) semantically differently in 1) and 9), two typical noun

[1] In this paper, whenever listing a Chinese word, we always list its Pinyin included within two symbols '/', and its English translation. For a Chinese phrase, we furthermore list its English translation when necessary.

[2] To differentiate the two words, we add one quotation mark.

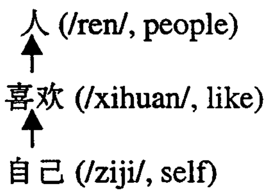
phrases. Particularly, they are both followed by a typical noun in the two phrases.

The second problem for dependent tree to semantically annotate linguistic forms is that due to its tree nature, it cannot represent *multi-dependence* relationship, in which one node is *controlled* by several nodes. For example, consider 11).

11) 喜欢 自己 的 人
/xihuan/ /ziji/ /de/ /ren/
like self of people
the people who like oneself

Conventionally, its dependent tree should be 12).

12)



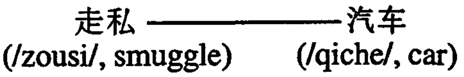
But intuitively, there should be some *dependence* relationship between 人(/ren/, people) and 自己 (/ziji/, self). If we add this relationship, it will become a graph.

2.2 Conceptual Graph

Conceptual graph is another diagram for semantic annotation of linguistic forms, which comprises *concepts* and *conceptual relationship* denoted by linguistic forms (Eklund, 1996). Although it is claimed to be a directed graph in its original form, it is equivalent to an undirected graph in nature, with its relationship nodes and their directed edges replaced with an undirected edge to directly denote the relationship.

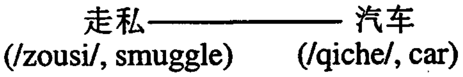
One problem with this diagram for semantic annotation is that it cannot code the information about *head*, if any, in a linguistic form, which intuitively specifies the main information carried by a linguistic form. This will lead to severe problems when using the graph to represent linguistic forms. For example, both phrases 1) and 2) with all three meanings 3), 4) and 5) would be represented by the same diagram as in 13)[3].

13)

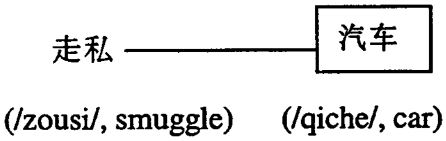


To differentiate the three meanings held by the two phrases, we suggest using the following two weighted graphs and one unweighted graph, i.e., 14), 15) and 16) to represent 3), 4) and 5) respectively.

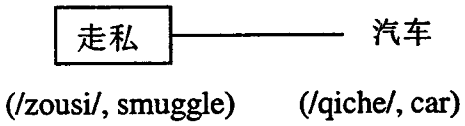
14)



15)



16)



Here we basically use undirected graphs to annotate phrases, and introduce a rectangle to denote the *head* of a linguistic phrase, if any. Notice that we don't mark a head in 14), which means that we don't take the verb 走私(/zousi/, smuggle) as the head of the verb phrase as usual. In general, for most verb phrases like 14) in Chinese, they correspond with two modifier-center phrases like 15) and 16) that comprise the same content words but with different meanings. For such phrases, we generally use a headword to differentiate between the verb phrase and the two modifier-center phrases, and then use different headword to distinguish the two modifier-center phrases.

Another problem with conceptual structure is concerned with its ability to deal with hierarchical structures. Although *nested conceptual structure* is introduced to

[3] Unless necessary, we don't list relationship in the conceptual graph.

describe *nested belief*, one particular kind of hierarchical structures (Geneviève, 1998), some simple hierarchical structures cannot be distinguished or annotated appropriately. As an example, consider 17) and 18).

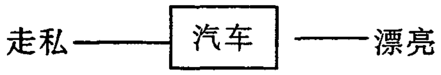
17) 漂亮 的 走私 汽车
/piaoliang/ /de/ /zousi/ /qiche/
beautiful of smuggle car
beautiful smuggled cars

18) 走私 漂亮 的 汽车
/zousi/ /piaoliang/ /de/ /qiche/
smuggle beautiful of car
to smuggle beautiful cars

Using conceptual graph, we can annotate both of them as 19), in which case the two phrases cannot be distinguished.

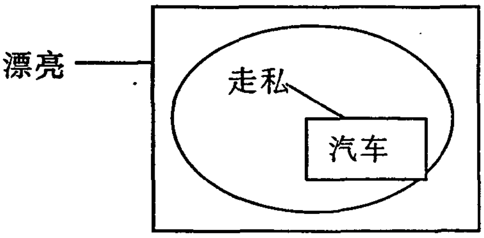
19) 走私——汽车——漂亮

If based on undirected graph plus head, 17) can be annotated as 20).

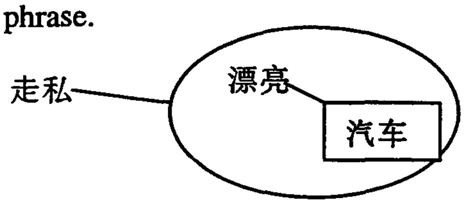
20) 

But there will be no appropriate annotation for 18), because on the one hand, 汽车(/qiche/,car) should be coded as a *head* due to its relationship with 漂亮(/piaoliang/, beautiful), on the other hand, it's role as the object of the verb 走私(/zousi/, smuggle) makes it illegal to be a *head*.

To differentiate 17) and 18), we further suggest specifying the embedded structures in linguistic forms in some way, and use *circles* to denote them. In such opinion, 17) can be annotated as 21) and 22) respectively.

21) 

In this diagram, the smaller rectangle denotes the head of the modifier-center phrase 走私汽车 (/zousiqiche/, smuggled cars), the circle codes the phrase as an embedded structure of the whole phrase, while bigger rectangle denotes the head of the whole phrase.

22) 

In this diagram, the circle denotes the embedded structure 漂亮的汽车(/piaoliang de qiche/, beautiful cars), while the rectangle denotes the head of the embedded structure. Notice the phrase on the whole is a verb phrase, so there is no head coded here.

3. Recursive Graph

One major concern for semantic annotation of linguistic forms is what semantic information will be coded, or more generally, what is their semantic information, which has long been a quite controversial question. From the point of view of *concepts*, linguistic forms including phrases semantically refer to *concepts*, which we think generally fall within four categories:

a) *preliminary concept*. For example, we may see the word 物(/wu/, thing) denotes a preliminary concept.

b) *compositional concept or situation*, which consists of some *concepts* and their *relation*[4]. For example, phrase 1) with the meaning 3) denotes a *situation* including the concepts the two content words, 走私(/zousi/, smuggle) and 汽车(/qiche/, car), denote respectively, and their relationship, 受事(/shoushi/, patient).

[4] *relation* is also a specific kind of concepts, e.g., 受事(/shoushi/, patient) is a *relation* between 走私(/zousi/, smuggle) and 汽车(/qiche/, car).

c) *subordinate concept or specific concept within a situation*, for example, phrase 1) with the meaning 4) refers to the concept denoted by 汽车(/qiche/, car) in the *situation* mentioned in *b*).

d) *subordinate feature, or specific feature of a situation or a concept,* which generally stands for the *relationship* between a *situation* or a *concept* and another unknown *concept*. Consider 23).

23) 姑娘 的 外貌
/guniang/ /de/ /waimao/
girl of appearance
the appearance of girls

In this phrase, the word 外貌 (/waimao/, appearance) denotes a *feature* of the concept the word 姑娘 (/guniang/, girl) denotes, and the *value* of this *feature* doesn't occur in this phrase.

Notice that the concepts within one situation themselves may be compositional concepts, subordinate concepts or subordinate features, so the concepts that linguistic forms including phrases denote generally presents a kind of hierarchical structure. In fact, this hierarchical structure in return represents the main semantic information of linguistic forms.

We introduce *recursive graph* as formal diagram to represent the hierarchical structure of linguistic forms. Let P_0 be the set of *preliminary points*, we call $p \in P_0$ a *0-level graph*. Suppose $P_1 \subseteq P_0$, $E_1(\subseteq P_0)$ is the set of the *edges* between *points* in P_1, $R_1(\subseteq(P_1 \times E_1))$ is the set of *relations* between *points* in P_1 and *edges* in E_1[5], then:

i) *$<P_1, E_1, R_1>$ is a 1-level compositional graph;*
ii) *$<<P_1, E_1, R_1>, p>$ ($p \in P_1$) is a 1-level point-headed graph;*
iii) *$<<P_1, E_1, R_1>, e>$ ($e \in E_1$) is a 1-level edge-headed graph;*
iv) *1-level concepts comprise 1-level compositional graphs, 1-level point-headed graphs, and 1-level edge-headed graphs.*

Let $\Sigma_0 = P_0$, Σ_{n-1} be the set of *(n-1)-level graphs*, suppose $P_n \subseteq (\Sigma_0 \cup \Sigma_1 \cup, \ldots, \Sigma_{n-1})$, $(P_n \cap \Sigma_{n-1}) \neq NIL$, $E_n(\subseteq P_n)$ is the set of the *edges* between *points* in P_n, $R_n(\subseteq(P_n \times E_n))$ is the set of *relations* between *points* in P_n and *edges* in E_n, then:

v) *$<P_n, E_n, R_n>$ is a n-level compositional graph;*
vi) *$<<P_n, E_n, R_n>, p>$ ($p \in P_n$) is a n-level point-headed graph;*
vii) *$<<P_n, E_n, R_n>, e>$ ($e \in E_n$) is a n-level edge-headed graph;*
viii) *n-level concepts comprise n-level compositional graphs, n-level point-headed graphs, and n-level edge-headed graphs.*

Intuitively, *0-level graph* corresponds with *preliminary points*, *compositional graph* corresponds with *situation*, *point-headed graph* corresponds with *subordinate concept*, and *edge-headed graph* corresponds with *subordinate feature*.

4. Annotation Guidelines

In general, Chinese phrases can roughly be classified into five categories, i.e., sub-predicate, verb-object, modifier-center, verb-complement, and coordinate. We give some examples in the following for each category.

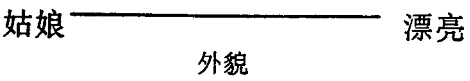
4.1 Sub-predicate

[5] Here, Edges are also points. An edge point connecting two other points here equals an edge in a traditional definition of graph.

In general, the phrase in this category denotes a *compositional concept*. For example, 24) can be annotated as 25).

24) 姑娘 漂亮
/guniang/ /piaoliang/
girl beautiful
Girls are beautiful

25)

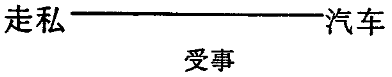


Intuitively, the concept the word 姑娘(/guniang/, girl) denotes has a feature, 外貌(/waimao/, appearance), and its value is 漂亮(/piaoliang/, beautiful). In other words, there exists a relationship, i.e., 外貌(/waimao/, appearance), between the two concepts denoted by 姑娘(/guniang/, girl) and 漂亮(/piaoliang/, beautiful) respectively[6].

4.2 Verb-object

Similar with sub-predicate phrases, the phrase in this category also denotes a *compositional concept*. Phrase 1) with the meaning 3) is an example, and can be annotated as 26).

26)



Intuitively, there exists a relationship, i.e., 受事(/shoushi/, patient), between the two concepts denoted by 走私(/zousi/, smuggle) and 汽车(/qiche/, car) respectively.

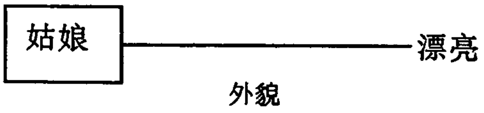
4.3 Modifier-center

The phrase in this category generally denotes a *subordinate concept*. The center here in the phrase can be a verb, a noun or an adjective. As examples, 27) and 28) are annotated as 29) and 30).

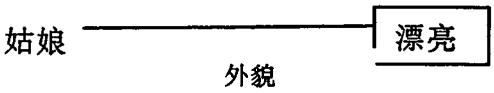
27) 漂亮 的 姑娘
/piaoliang/ /de/ /guniang/
beautiful of girl
beautiful girls

28) 姑娘 的 漂亮
/guniang/ /de/ /piaoliang/
girl of beautiful
the beauty of girls

29)



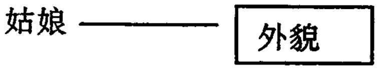
30)



Intuitively, the two concepts denoted by the two content words 姑娘(/guniang/, girl) and 漂亮(/piaoliang/, beautiful) in 27) and 28) hold the same relationship, i.e., 外貌(/waimao/, appearance) as in 24). The difference lies in that the phrases 27) and 28) both have a *head*, i.e., 姑娘(/guniang/, girl) and 漂亮(/piaoliang/, beautiful) respectively.

Another example is 23), it denotes an *subordinate feature*, annotated as 31)[7].

31)



Intuitively, the concept denoted by 姑娘(/guniang/, girl) has a feature 外貌(/waimao/, appearance), which is the head of this phrase.

4.4 Verb-complement

For the phrase in this category, the semantic relationship between its two parts, i.e., verb and complement, is very complicated (Ma, 1987); even there is no direct semantic links

[6] Unless necessary, we don't list the link between concepts and relation names.

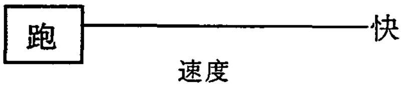
[7] The link between 姑娘(/guniang/, girl) and 外貌(/waimao/, appearance) denotes a relation between concepts and relationships.

between them sometimes. Consider 32) and 33).

32) 吃 饱
/chi/ /bao/
eat full
to eat and be full

33) 跑 快
/pao/ /kuai/
run fast
to run fast

In 32), 饱(/bao/, full) has a direct link with the agent of 吃(/chi/, eat), in which sense the two concepts denoted by 饱(/bao/, full) and 吃(/chi/, eat) has no direct semantic link. We don't consider such phrases in our annotation temporarily. In contrast, in 33), 快(/kuai/, fast) acts as the value of one feature 速度(/sudu/, speed) of 跑(/pao/, run), in which sense it has a direct semantic link with 跑(/pao/, run). They form a subordinate concept, annotated as 34).

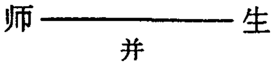
34)


4.5 Coordinate

In general, the phrase in this category denotes a compositional concept. As an example, consider 35).

35) 师 生
/shi/ /sheng/
teacher student
teacher and student

Intuitively, there is a relationship, i.e., 并(/bing/, and), between the two concepts denoted by the two words. So, it can be annotated as 36).

36)


5 Conclusion and Future Work

In this paper, we propose a recursive graph based scheme for semantic annotation of Chinese phrases. This scheme can fully differentiate those Chinese phrases that comprise the same content words but with different meaning due to their different word order or some involved function words. It can also capture the hierarchical conceptual structures of Chinese phrases, which determine their main semantic information.

Now, we have annotated about 5,000 Chinese phrases using this scheme. One methodological issue is that we also choose multi-character Chinese words as candidate annotation phrases. The reason is that unlike western languages, there is no clear-cut between words and phrases in Chinese, for most multi-character Chinese words, their components may also be words, and the meaning of the multi-character words tends to be strongly related with that of their component words. In this sense, the words can be treated as *basic* phrases. More importantly, such basic annotation examples can be more complete in the sense of coverage.

Future work will extend to 50,000 Chinese phrases. Another future work is to learn the rules that determine the mapping between linguistic forms and meanings based on this annotation, and apply the rules to new phrases and sentences. If this can be successful, we can develop a semantic analyzer for Chinese without any syntactic analysis. In particular, we can avoid part-of-speeches or syntactic structures that have long been difficult notions for Chinese languages.

Acknowledgments

I would like to thank Prof. Dong Zhendong and Prof. Huang Changning for their insightful discussions, and the anonymous reviewers for their comments.

References

J.Allen, Natural Language Understanding, Second Edition, the Benjamin/Cumming Publishing Company, INC, 1995.

Chomsky, 1968, Language and Mind, Enlarged edition, Harcourt Brace Jovanovich, Publishers.

Eklund, Peter W., Gerard Ellis, & Graham Mann, eds. (1996) Conceptual Structures: Knowledge Representation as Interlingua, Lecture Notes in Artificial Intelligence 1115, Springer-Verlag, Berlin.

Geneviève Simonet, 1998, Two FOL Semantics for Simple and Nested Conceptual Graphs, in proceedings of international conference on Conceptual Graph.

R. W. Langacker, 1997, Constituency, dependency, and concept grouping, Cognitive Linguistics, 8(1): 1-32.

Ma Xiwen, 1987, Some sentential patterns related with verb-result verbs, Zhong Guo Yu Wen (Chinese in China), 201 (6): 424-441. (in Chinese)

Zhu Dexi, Yu Fa Da Wen (in Chinese), Beijing University Press, 1982.