COMPLEX FEATURES IN THE DESCRIPTION OF THE CHINESE LANGUAGE

ZHWEI FENG
University of Trier

1. Multiple-Value Label Function

The phrase structure grammar (PSG) has been used extensively in the parsing of natural language. A PSG can be expressed by tree graph where every node has a correspondent label. The relationship between the node x and its label y can be described by a monovalue label function L:

$$L(x) = y$$

For every value of node x, there is only one corresponding value of label y.

In 1981, we designed a multilingual automatic translation system FAJRA (from Chinese to French, English, Japanese, Russian, German). In 1985, we designed two automatic translation systems GCAT (from German to Chinese) and FCAT (from French to Chinese). In the FAJRA system, we must do automatic analysis of Chinese, in the GCAT and FCAT systems, we must do automatic generation of the Chinese language.

We found that the linguistic features of Chinese expressed by the mono-value label function of PSG are rather limited. In the automatic analysis of the Chinese language, PSG can not properly treat the inherent ambiguity, hence the analysi often yields a lot of ambiguous structures. In the automatic generation of the Chinese language, the generative power of PSG is so strong that a large amount of ungrammatical sentences is generated. This is the chief drawback of PSG. In order to overcome this drawback of PSG, in the FAJRA system we proposed a multiple-value label function to replace the mono-value label function, and in the GCAT and FCAT systems, we further improved this approach.

A multiple-value label function can be described as below:

$$L(x) = \{y1, y2, ..., yn\}$$

In this description a label x of a tree can correspond to several labels {y1, y2, ..., yn}. By means of this function, the generative power of PSG was re-

stricted, the number of ambiguous structures was reduced, and the drawback of PSG was efficiently overcome.

Beginning with the augmented transition network (ATN) concept and inspired by J. Bresnan's work on lexically oriented non-transformational linguistics, the lexical functional grammar (LFG) framework of J. Bresnan and R. Kaplan was shaped. Simultanously, M.Kay devised the functional unification grammar (FUG), and G. Gazdar, E. Klein and G. Pullum proposed the generalized phrase structure grammar (GPSG). Implementation of GPSG at Hewlett-Packard led C. Pollard and his colleagues to design the head driven-phrase structure grammar (HPSG) as a successor to GPSG. In all these formalisms of grammars the complex features are widely used to replace the simple features of PSG and it is thus an improvement of PSG. Therefore, the concept of complex features is very important for the current development of computational linguistic.

In the FAJRA, GCAT and FCAT systems, the values of labels must be the features of language, thus the multiple-value labels must also be the complex features of language. In fact, the concept of multi-value labels and the concept of complex features is very similar. Historically, all these concepts are the results of various strains of research in computational linguistics aiming at an improvement of PSG.Thus we can take our multiple-value labels as complex features.

Famous linguist De Saussure (1857-1913) had pointed out in his in General Linguistics : "Language in a manner of speaking, is a type of algebra consisting solely of complex terms" (p 122, English version, 1959). He takes the flexion of "Nacht:Nächte" in German as the example. The relation "Nacht: Nächte" can be expressed by an algebraic formula a\b in which a and b are not simple terms but result from a set of relations -- complex terms. The complex terms of Nacht are: noun, feminine gender, singular number, nominative case; its principal vowel is "a". The complex terms of Nächte are: noun, feminine gender, plural number, nominative case; its principal vowel is "ä" , its ending is "e", the pronunciation of "ch" changes from /x/ to /ç/. De Saussure said: "But language being what it is, we shall find nothing simple in it regardless of our approach; everywhere and always there is the same complex equilibrium of terms that mutually condition each other" (p 122, 1959). So-called "complex terms" mentioned here by De Saussure are nothing but the "complex features" or the "multiple-value labels" in computational linguistics. After all, De Saussure was a scholar with great foresight. He has proved himself to be a pioneer of modern linguistics. The concept of "complex terms" of De Saussure has served as a source of inspiration for us to propose the concept of "multiple-value labels". However, the properties of the Chinese language are more important than the concept of De Saussure in the development of our "multiple-value labels" (or "complex features"), because without the "multiple-value labels (or "complex features") we cannot adequately describe the Chinese language in automatic translation.

2. Necessity of Complex Feature for the Description of Chinese Language

If there is a necessity of complex features in the description of English, then this necessity is still more obvious for the description of Chinese.

In this paragraph we are going to explain the reason and to give three arguments.

2.1 Argument One

In Chinese sentences there is not a one-to-one correspondence relation between the phrase types (or parts of speech) and their syntactic functions.

Chinese is different from English. In English, the phrase types generally correspond to their syntactic functions.For example, we have

$$NP + VP \longrightarrow S$$

in English, whereby NP corresponds to subject, VP corresponds to predicate, and S is a sentence, becoming a "subject+predicate" construction. There is one-to-one correspondence between the phrase types and their syntactic functions. In Chinese, the structure "VP+NP" can form a sentence, so we also have

$$NP + VP \longrightarrow S$$

E.g. in the phrase "Xiao Wang Kesou" (little Wang coughs), "Xiao Wang" (little Wang) is a NP, "kesou" (to cough) is a VP, forming a "subject+predicate"

construction. However, in many cases, the NP, does not correspond to the subject, the VP does not correspond to the predicate. for example, "chengxu sheji" (programming) in chinese, "chengxu" (program) is a NP, "sheji" (to design) is a VP, but this NP is a modifier, and this VP is the head of the structure "NP+VP". The structure "VP+NP" cannot form a sentence, but forms a new noun phrase NP1:

$$NP + VP \longrightarrow NP1.$$

So this noun phrase becomes a "modifier+head" construction. Similar examples are: "yuyan xuexi" (language learning), "wuli kaoshi" (physics examination), etc. In these phrases, the phrase types "NP+VP" cannot form a "subject+predicate" construction, but form a "modifier+head" construction. In this case, the "NP+VP" is a syntactically ambiguous structure, the simple features "NP+VP" cannot distinguish the differences between the construction "subject+predicate" and the construction "modifier+head". We must use complex features to describe these differences.

The structure "NP+VP" which forms a "subject+predicate" construction can be formularized as the following complex feature set:

$$\left\| \begin{array}{l} K = NP \\ CAT = N \\ SF = SUBJ \end{array} \right| \quad + \quad \left| \begin{array}{l} K = VP \\ CAT = V \\ SF = PRED \end{array} \right\|$$

where K is the feature of the phrase type, NP and VP are the values of this

feature; CAT is feature of parts of speech, N and V are the values of this feature; SF is the feature of syntactic function, SUBJ and PRED are the values of this feature.

With the complex features the structure "NP + VP" which forms a "modifier + head" construction can be formularized as following:

$$\left|\begin{array}{l|l} \left|\begin{array}{l} K = NP \\ CAT = N \\ SF = MODF \end{array}\right| & + \quad \left|\begin{array}{l} K = VP \\ CAT = V \\ SF = HEAD \end{array}\right| \end{array}\right|$$

where MODF and HEAD are the values of the feature SF. Obviously, the structure "NP + VP" is ambiguous; there are two syntactically different constructions included in this structure; their phrase types are identical, but their syntactic functions are different. In order to describe the differences between them adequately, we have to use the complex features.

In English, we have

$$VP + NP \longrightarrow VP1,$$

the VP corresponding to the predicate, the NP to the object, and VP1 is a new verb phrase; it is a "predicate + object" construction. In Chinese, the structure "VP + NP" can form a new verb phrase, so we also have

$$VP + NP \longrightarrow VP1.$$

For example, "taolun wenti" (to discuss a problem) in Chinese, "taolun" (to discuss) is a VP, "wenti" (problem) is an NP, forming a new "predicate + object" verb phrase. However, in many cases, VP doesn't correspond to predicate, NP does not correspond to object. For example, "chuzu qiche" (taxicab) in Chinese, "chuzu" (to hire) is a VP, "qiche" (automobile) is a NP, but this VP is a modifier and this NP is the head of the "VP + NP" structure. This structure cannot form a new verb phrase; it forms a new noun phrase, so we have:

$$VP + NP \longrightarrow NP1.$$

The synctatic function of this noun phrase is a "modifier + head" construction; similar examples are: "yanjiu fangfa" (aproach to research), "xuexi zhidu" (regulation for study), "kaifang zhengce" (open door policy). The phrase type structure "VP + NP" cannot form a "predicate + object" construction, but forms a "modifier + head" construction. In this case, the "VP + NP" structure is syntactically ambiguous. The simple features "VP + NP" are not enough to distinguish the differences between the "predicate + object" construction are the "modifier + head" construction. We must use complex features to describe these differences.

The structure "VP + NP" which forms a "predicate + object" construction can be formularized as the following complex feature set:

$$\left|\begin{array}{l|l} \left|\begin{array}{l} K = VP \\ CAT = V \\ SF = PRED \end{array}\right| & + \quad \left|\begin{array}{l} K = NP \\ CAT = N \\ SF = OBJE \end{array}\right| \end{array}\right|$$

where PRED and OBJE are the values of feature SF.

The structure "VP + NP" which forms a "modifier + head" construction can be formularized as the following complex feature set:

$$\left|\begin{array}{l|l} \left|\begin{array}{l} K = VP \\ CAT = V \\ SF = MODF \end{array}\right| & + \quad \left|\begin{array}{l} K = NP \\ CAT = N \\ SF = HEAD \end{array}\right| \end{array}\right|$$

where MODF and HEAD are the values of the feature SF.

Obviously the use of simple features is insufficient for the description of the structure "VP + NP"; instead, it is necessary to use complex features.

2.2. *Argument Two*

For the constructions which have the same phrase type structure and the same syntactic function structure, their syntactic relations may be different. So there is no simple one-to-one correspondence between syntactic function and its semantic relationship.

The values of semantic relation features are the following: Agent, Patient, Instrument, Scope, Aim, Result, ... etc. In English, there is not much correspondence between the syntactic function of a sentence element and its semantic relation. In Chinese, the correspondence is more complicated and less specified than English. In the construction "subject + predicate" with corresponding phrase type structure " NP + VP", the subject may be Agent, but it may also Patient, or Instrument, etc. For example, in the sentence "wo dule" (I have read), the subject "wo" (I) is the Agent, but in the sentence "shu dule" (the book has been read), the subject "shu" (book) is the Patient, and the verb "dule" (to read) doesn't change its form, it always takes the original form. In most European languages, if the subject is the Agent, then the verb must take an active form, and if the subject is Patient, then the verb must take the passive form. However, in Chinese the verb always keeps the same form no matter whether the subject is an Agent or Patient. In an overwhelming majority of cases, the passive form of the verb is seldom or never used in Chinese. For this reason, in automatic information processing of the Chinese language, it is insufficient to use only the features of phrase types and syntactic functions, we must also use the features of semantic relations. Thus the features for description of the Chinese language will become more complex than they are for European languages.

In the structure "NP + VP" , if the syntactic function of NP is subject, and semantic relation of NP is Agent, then the complex features of this structure can be formularized as following:

$$\left|\left|\begin{array}{l}K = NP\\CAT = N\\SF = SUBJ\\SM = AGENT\end{array}\right| + \left|\begin{array}{l}K = VP\\CAT = V\\SF = PRED\end{array}\right|\right|$$

where SM is the feature of semantic relation, AGENT is the value of the feature SM. In the structure "NP + VP", if the syntactic function of NP is subject and the semantic relation of NP is Patient, then the complex features of this structure can be formularized as follows:

$$\left|\left|\begin{array}{l}K = NP\\CAT = N\\SF = SUBJ\\SM = PATIENT\end{array}\right| + \left|\begin{array}{l}K = VP\\CAT = V\\SF = PRED\end{array}\right|\right|$$

where PATIENT is the value of the feature SM.

In the structure "VP + NP" with corresponding "predicate + object" syntactic construction, the object may be a Patient, but it may also be an Agent, or an Instrument, or a Scope, or an Aim, or a Result, ..., etc. In the sentence "ca chuangzi" (to wipe the window), "ca" (to wipe) is a predicate, "chuangzi" (window) is the object, and its semantic feature is the Patient. The complex features of this sentence can be formularized as follows:

$$\left|\left|\begin{array}{l}K = VP\\CAT = V\\SF = PRED\end{array}\right| + \left|\begin{array}{l}K = NP\\CAT = N\\SF = OBJE\\SM = PATIENT\end{array}\right|\right|$$

where PATIENT is the value of the feature SM.

But we also have the following sentences where the semantic relation of object is not the Patient. There are many very interesting phenomena in Chinese: in the sentence "sile fuqin" (the father died), the object "fuqin" (father) is the Agent of the verb "sile" (to die). This structure

can be formularized as the following complex feature set:

$$\left|\left|\begin{array}{l}K = VP\\CAT = V\\SF = PRED\end{array}\right| + \left|\begin{array}{l}K = NP\\CAT = N\\SF = OBJE\\SM = AGENT\end{array}\right|\right|$$

where AGENT is the value of the feature SM.

In the sentence "chi dawan" (to eat with a big bowl), the object "dawan" (big bowl) is the Instrument of the verb "chi" (to eat). This structure can be formularized as the following complex feature set:

$$\left|\left|\begin{array}{l}K = VP\\CAT = V\\SF = PRED\end{array}\right| + \left|\begin{array}{l}K = NP\\CAT = N\\SF = OBJE\\SM = INST\end{array}\right|\right|$$

where INST is the value of the feature SM.

In the sentence "kao shuxue" (to be examined in mathematics), the object "shuxue" (mathematics) is the Scope of the verb "kao" (to examine). This structure can be formularized as the following complex feature set:

$$\left|\left|\begin{array}{l}K = VP\\CAT = V\\SF = PRED\end{array}\right| + \left|\begin{array}{l}K = NP\\CAT = N\\SF = OBJE\\SM = SCOPE\end{array}\right|\right|$$

where SCOPE is the value of the feature SM.

In the sentence "kao yanjiusheng" (to be examined in order to become a graduate student), the object "yanjiusheng" (graduate student) is the Aim of the verb "kao" (to examine). This structure can be formularized as the following complex feature set:

$$\left|\left|\begin{array}{l}K = VP\\CAT = V\\SF = PRED\end{array}\right| + \left|\begin{array}{l}K = NP\\CAT = N\\SF = OBJE\\SM = AIM\end{array}\right|\right|$$

where AIM is the value of the feature SM.

In the sentence "kao manfen" (to pass the examination and get excellent marks). The object "manfen" (excellent marks) is the Result of the verb "kao" (to examine). This structure can be formularized as the following complex feature set:

$$\left|\left|\begin{array}{l}K = VP\\CAT = V\\SF = PRED\end{array}\right| + \left|\begin{array}{l}K = NP\\CAT = N\\SF = OBJE\\SM = RESULT\end{array}\right|\right|$$

where RESULT is the value of the SM.

Thus we can see very clearly that only with complex features can the differences between these sentences be revealed sufficiently.

2.3. Argument Three

The grammatical and semantic features of the words play an important role in putting forword the rules of parsing. These features can be used as the conditions

in the rules, and they must be included in the system of complex features. Thus complex features are the basis for setting up the rules for the parsing of the Chinese language. Without the complex features, we cannot parse the Chinese language adequately.

In the structure VP+NP, if the grammatical feature of VP is an intransitive verb, then the syntactic function of VP must be modifier, and the syntactic function of UP must be head. The syntactic function of this structure VP+NP will be "modifier + head". Thus we can have a rule which is described with complex features as the following:

$$\left|\begin{array}{l}K = VP\\CAT = V\\TRANS = IV\end{array}\right| + \left|\begin{array}{l}K = NP\\CAT = N\end{array}\right| \longrightarrow \left|\left|\begin{array}{l}K = VP\\CAT = V\\TRANS = IV\\SF = MODF\end{array}\right| + \left|\begin{array}{l}K = NP\\CAT = N\\SF = HEAD\end{array}\right|\right|$$

where TRANS represents the feature of verb transitivity, IV represents intransitive verb and it is a value of the TRANS feature.

This rule means: In the structure VP+NP, if the value of TRANS of VP is IV, then the syntanctic function of VP can be given the value MODF, and the syntanctic function of NP can be given the value HEAD.

The semantic features of words can also be used to put forword the rules of parsing. In the structure VP+NP, if VP is a transitive verb, then we have to use the semantic features of NP to decide the value of the syntactic function of this structure.

Generally speaking, if VP is a transitive verb, the semantic feature of NP is "abstract thing", or "title of technical or professional post", then the syntanctic function of VP is modifier, and the syntanctic function of NP is head. For example, in the phrase "xunlian mudi" (purpose of training), the VP "xunlian" (to train) is a transitive verb, and the semantic feature of NP "mudi" (purpose) is "abstract thing", we can decide that the syntanctic function of "xunlian" (to train) is modifier, and the syntanctic function of "mudi" (purpose) is head. In the phrase "jinxiu jiaoshi" (teacher with advanced training), the VP "jinxiu" (to engage in advanced studies) is a transitive verb, the NP "jiaoshi" (teacher) is the title of professional post. Thus we can decide that the syntanctic function of VP "jinxiu" (to engage in advanced studies) is modifier, the syntanctic function of NP "jiaoshi" (teacher) is head.

Therefore we have the two following rules for parsing:

(1)

$$\left|\begin{array}{l}K = VP\\CAT = V\\TRANS = TV\end{array}\right| + \left|\begin{array}{l}K = NP\\CAT = N\\SEM = ABS\end{array}\right| \rightarrow \left|\left|\begin{array}{l}K = VP\\CAT = V\\TRANS = TV\\SF = MODF\end{array}\right| + \left|\begin{array}{l}K = NP\\CAT = N\\SEM = ABS\\SF = HEAD\end{array}\right|\right|$$

(2)

$$\left|\begin{array}{l}K = VP\\CAT = V\\TRANS = TV\end{array}\right| + \left|\begin{array}{l}K = NP\\CAT = N\\SEM = PRF\end{array}\right| \rightarrow \left|\left|\begin{array}{l}K = VP\\CAT = V\\TRANS = TV\\SF = MODF\end{array}\right| + \left|\begin{array}{l}K = NP\\CAT = N\\SEM = PRF\\SF = HEAD\end{array}\right|\right|$$

where TV represents transitive verb (a value of TRANS), ABS represents abstract thing (a value of SEM), PRF represents the title of a technical or professional post (another value of SEM). On the basis of these complex features, two new values of feature SF can be deduced:

SF = MODF and SF = HEAD.

With complex features, we obtain good results in the FAJRA, GCAT and FCAT machine translation systems.

REFERENCES

Feng, Z. 1983. "Multiple-branched and multiple-labeled tree analysis of Chinese sentences". *Journal of Artificial Intelligence* 2.

De Sausure, F. 1959. *Course in General Linguistics*,

Kay, M. 1985. "Parsing in functional unification grammar", *Natural Language Parsing: Psychological, Computational and Theoretical Perspectives*. 1985.