Guidelines and Problems for Bracketing Chinese

Keh-Jiann Chen, Chu-Ren Huang, Zhao-Ming Gao
Chinese Knowledge Information Processing Group
Academia Sinica, TAIWAN

I. Framework: Information-based Case Grammar (ICG) (Chen and Huang (1990, 1997)

II. Abstract of Chen and Huang (1997)

The grammatical formalism of ICG stipulates that the lexical entry for each word contains both semantic and syntactic features. In the feature structure of a phrasal head, we encode syntactic and semantic constraints on grammatical phrasal patterns in terms of thematic structures, and encode the precedence relations in terms of adjunct structures. Such feature structures denote partial information which defines the set of legal phrases. They also provide sufficient information to identify thematic roles. With this formalism, parsing and thematic analysis can be achieved simultaneously. Due to the simplicity and flexibility of ICG, context dependent and discontinuous relations such as agreements, coordinations, long-distance dependencies, and control and binding, can be easily expressed.

III. Phrasal categories

CS　Chinese string such as 而言
DM　determiner measure phrase
GP　postpositional phrase (or localizer phrase), e.g. 睡覺前，夜裡，三百人以上
NP
PP
VP
N
S
A・的
N・的
V・的
S・的
DM・的　DM is determiner-measure phrase
GP・的

NP・的
PP・的
VP・的
ADV・的
A・地
N・地
V・地
S・地
DM・地
GP・地
NP・地
PP・地
VP・地
ADV・地
A・之
N・之
V・之
S・之
DM・之
GP・之
NP・之
PP・之
VP・之
ADV・之
到・NP
到・VP
到・S　　e.g. 吃到肚子發漲
個・VP　e.g. 吃個飽
得・NP
得・VP
得・S

III. The ICG framework and Its Implications for Bracketing Chinese

1. Parsing and the identification of thematic roles are done simultaneously.

The format of our bracketing
70. S(agent:NP (Head:Nca:中國)
|time:Dd:將 |Head:VF2:派
|goal:NP (quantifier:DM:34名 |Head:Nab:選手)
|theme:VP(Head:VC2:參加)))

2. There is no clear borderline between lexical and phrasal categories in our framework. For example, the node governing two conjoined nouns is labeled N rather than NP. The labeling of a XP is delayed until thematic role assignments are determined.

3. In general, only phrases that have a thematic role are labeled XP.

4. ICG is similar to dependency grammar. The tree structures under the ICG framework are thus flatter.

4. The node VP occurs when one of the following three conditions is met. (1). It is a subcategorized argument. (2). It is in a clause which lacks a subject NP. (3). It is conjoined with a VP.

5. There are no empty categories in our framework. Hence, Verbs like 覺得 'feel' can subcategorize for a VP or a S.

我 覺得 [VP [很 累]]
I feel very tired

我 覺得 [S [他 很 驕傲]]
I feel he very tired

6. Clitics such as 得, 地, 的, 個, 到 and their preceding phrases form a phrase whose phrase names are expediently expressed as VP・地, 到・VP, 到・S, etc.

7. There is no AP or ADVP in our framework because of 6.

8. Since a large portion of adjectives in Chinese can not only modify a noun but also function as verbs or adverbs, we do not give them different tags for different grammatical functions. In our framework, the tag VH is used to describe these adjectives. The tag A is used for Chinese adjectives which cannot be used as predicates.

9. Verbs that follow 的 is regarded as nominalization. They are annotated with the tag NV in parsing and bracketing.

10. Sentential subjects and clausal complements are labelled as S rather than NP.

11. Our experience shows that syntactic bracketing is not enough. To build a tree bank for Chinese NLP, semantic information such as thematic roles must be annotated.