Construction of the Basic Sentence-pattern Instance Database Based on the International Chinese Textbook Treebank

Shuqin Zhu[1,2], Yinxia Zhang[2],Jihua Song[1] ,Weiming Peng[1]
1.College of Information Science and Technology, Beijing Normal University, Beijing, 100875, China
2.Teachers?College of Beijing Union University, Beijing, 100011, China
sftzhushuqin@buu.edu.cn, sftyinxia@buu.edu.cn, songjh@bnu.edu.cn, pengweiming@bnu.edu.cn

***Abstract*—In order to better serve the international Chinese teaching, according to the sentence-based syntactic system and Chinese syntax structure characteristics, the paper builds a basic sentence-pattern instance database based on the international Chinese textbook Treebank. Based on the instance database, the paper adds an index for the predicate, and then extracts the relevant knowledge and information of predicate keywords. Finally, a certain amount of complex natural language sentences are converted into more instances with teaching values, which provide basic resources for international Chinese teaching.**

Keywords—Sentence-based Syntactic; International Chinese Textbook Treebank; Basic Sentence-pattern Instance; Single-predicate Core

1 Introduction

Jinxi Li proposed the sentence-based syntactic system in 揑ew mandarin grammar?, which has a close connection with the teaching grammar system. Based on the sentence-based idea, Language and Language Resources Research Center of Beijing Normal University has constructed the international Chinese textbook Treebank. From the existing international Chinese textbook Treebank the paper extracts sentence-pattern instances with single-predicate core, which belong to the different basic sentence patterns. Although these basic sentence-pattern instances are not complete natural language sentences, they have the teaching value. This paper adds a middle layer between the Treebank and teaching grammar, namely sentence- pattern instance layer. A certain scale Treebank is extended to a larger scale basic sentence-pattern instance database based on the international Chinese textbook Treebank, which provides an idea for the application of the international Chinese textbook Treebank in the international Chinese teaching.

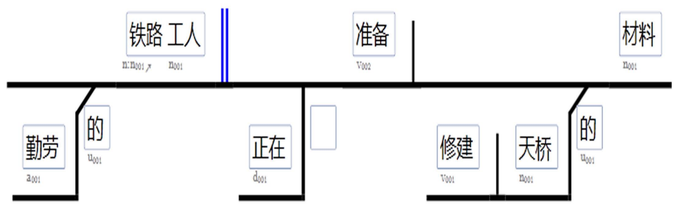
2 Basic Sentence-pattern Instancs

The basic sentence-pattern instance database only contains the single-predicate core sentences which reflect the argument structure of the central verb.

Four corresponding basic sentence-pattern instances can be extracted from the (What do you think of the rule that the state does not allow the celebrity to appear in the advertisement as a patient) ".

- **"您|||对 规定]看"**
- **"您|||怎么]看"**
- **"国家|||不]允许|名人"**
- **"名人|||以 身份]出现<在 广告 中>?**

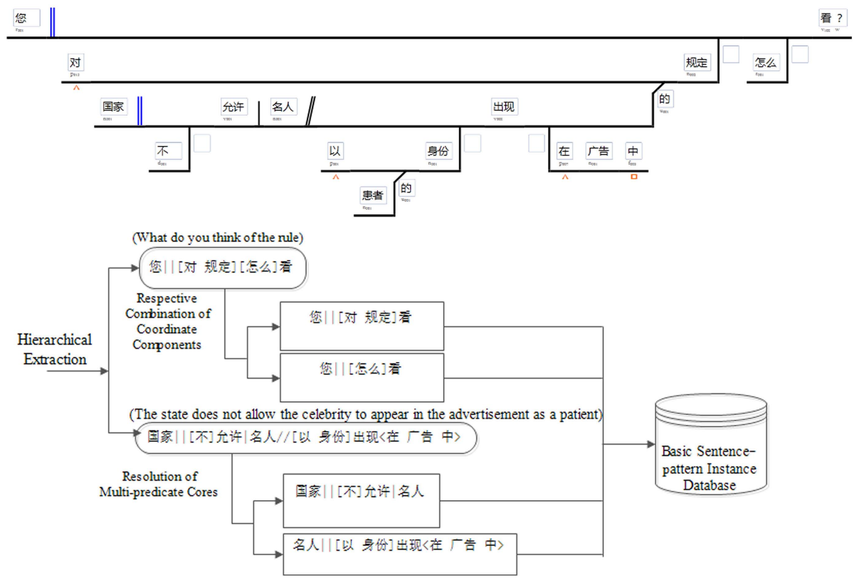
3 International Chinese Textbook Treebank




```
<xj>
  <sbj>                                   Subject
    <att>                                 Attribute
      <a sen="001">勤劳</a>
      <uu> <u sen="001">的</u> </uu>
    </att>
    <n mod="n2↗n2">
      <n sen="001">铁路</n> <n sen="001">工人</n>
    </n>
  </sbj>
  <adv> <d sen="001">正在</d> </adv>      Adverbial
  <prd scp="VO"><v sen="002">准备</v> </prd>   Predicate
  <obj>                                   Object
    <att>                                 Attribute
      <prd scp="VO"> <v sen="001">修建</v> </prd>
      <obj> <n sen="001">天桥</n> </obj>
      <uu> <u sen="001">的</u> </uu>
    </att>
    <n sen="001">材料</n>
  </obj>
</xj>
```

4 Extraction Process of the Basic Sentence-pattern Instance

4.1 Extraction Principle of the Basic Sentence-pattern Instance



4.2 Extraction Method of the Basic Sentence-pattern Instance

Extract the subject

- According to the subject?situation, the algorithm of extracting the subject is: if <sbj> contains the predicate component <prd> nodes, and the subject needs the further analysis, otherwise find the <sbj> child nodes which are not the <x/> node and the name length is 1, then save it in the subject array

Extract the predicate

- If there is only one <prd> node in the clause <xj>, it needs to take out all the corresponding adverbial parts of the node, and according to the value of the <prd> property 揾 scp? the other corresponding components of the <prd> node can be gotten, and the basic sentence-pattern instance of the single-predicate core can be obtained after the respective combination.
- If the clause <xj> contains more than one <prd> node, all the corresponding adverbial parts of every <prd> node need to be extracted firstly, and then according to the value of the <prd> property 揾 scp?, its corresponding object or complement components of the <prd> node can be gotten, and lastly according to the relationship between predicates, the basic sentence-pattern instance about every predicate can be obtained after the respective combination. Note that if there is a pivotal relationship, the subject is updated as the object of the previous predicate.

the complement extraction module, the adverbial extraction module, and the object extraction module

- It is need to analyze whether there is the predicate in these components, and if the predicate is included, it needs to conduct the further in-depth analysis. If these components don揾contain <prd>, they still needs to be analyzed whether there is the coordinate <cc> node, and if the coordinate component is included, each part will be analyzed individually, and if the coordinate component isn揾ncluded, the central words will be extracted.

4.3 Storage of the Basic Sentence-pattern Instance Database

- **"S" represents the subject component**
- **"D" represents adverbial component**
- **"V" indicates predicate component**
- **"O" indicates object component**

5 Overview of the Basic Sentence-pattern Instance Database

SVO-TYPE BASIC SENTENCE-PATTERN INSTANCES

C	Predicate	Object
你们_r001(you)	离 开 _v001(are leaving)	母 校 _n001(the university)
n: 杭 州 _n_ 人 _n001(Hangzhou people)	v: 常 _d003 显 _v002(are)	温 文 尔 雅 _a001(gentle and cultivated)
眼 神 _n001(The eyes)	v:写_v001 满_a001 了 _u001(are full of)	留 恋 _v001(nostalgia)
我_r001(I)	打 开 _v001(opened)	门_n002(the door)
n: 壁 _n001_ 龛 _n001(in the niche)	v: 点 _v019 上 _v302(light)	蜡烛_n001(a candle)
人 类 _n001(Human)	破 译 _v001(decode)	语 言 _n001(language)
他_r001(He)	是_v201(is)	绅 士 _n001(a gentleman)

6 Conclusion

Basic sentence-pattern instances are extracted from the international Chinese textbook Treebank. Based on the instance database, through establishing the keyword index, all related knowledge and information about the predicate keywords are extracted. The knowledge base with three-tier structure is realized which makes a certain amount of complex natural language sentences to be converted into more instances with the teaching value.