Construction of Chinese Semantic Resource Based on Feature Structure

Bo Chen

NLP lab, Computer school of Wuhan University

outline

- Background
- Feature Structure Theory
- Construction of Chinese Semantic resource
- Summary
- Application

Our work

- **Feature structure**
  - Semantic relations, types
  - Feature structure triple
  - formal representation
  - How to determine, criterion
- **Chinese Semantic resource**
  - sentences source
  - annotation criterion
  - annotation methods
  - annotation platform
- **Application**
  - Help to resolve the problems in linguistic fields
  - For Company, government

Background

Semantic analysis

Chinese characteristics:

Free order

Special sentence patterns

Many function words

It is difficult to represent completely the semantic relations of Chinese by traditional methods.

The examples of problems in Chinese:

Example 1

A:小王　　肚子　笑　　痛　　了。

- /xiaowang / /duzi/ /xiao/ /tong / /le/
- *stomach* to laugh painful

Example 2

B:小王　的　肚子　笑　　痛　　了。

- /xiaowang / /duzi/ /xiao/ /tong / /le/
- *'s stomach* to laugh painful

Free order

Example 3

C:小王　　笑　　痛　　了　　肚子。

- /xiaowang / /xiao/ /tong / /le/ /duzi/
- to laugh painful stomach

- *Xiaowang laughed too much to feel stomach painful.*

Chinese Verb-complement Structure

- **Example 4:**
- **衣服 洗 干净 了。**
- **/yifu/ /xi/ /ganjing/ /le/**
- **Clothes wash clean**
- ***Clothes are washed clean.***

- **Example 5:**
- **衣服 洗 完 了。**
- **/yifu/ /xi/ /wan/ /le/**
- **Clothes wash finished**
- ***Clothes are washed up.***

- **Example 6:**
- **衣服 洗 晚 了。**
- **/yifu/ /xi/ /wan/ /le/**
- **Clothes wash late**
- ***Clothes are washed too late.***

- **"S (*clothes*) + V (*to wash*) + C (*adjective*)".**

Complex Noun Phrase

- $N_1+N_2+N_3……+N_N$

Example 7:

▪今天　七一　建党节.

▪/jin tian/　/qi yi/　/jian dang jie/

▪Today　1st,July　the party's day

▪*Today is 1st July, the party's day.*

▪No verb

▪The predicate is two nouns, these are appositives

▪Which is the head??

Chinese Serial Verb Sentence pattern

- Example 8：
- 我　　买　了　　碗　　面　　吃.
- /wo/　/mai/ /le/　　/wan/　/mian/　　/chi/
- I　　buy　　bowl　noodle　eat
- *I bought a bowl of noodle to eat.*

More than 2 verb-phrases

- Example 9:
- 我　　开车　　去车站　　接他.
- /wo/　/kai/ /che /　　/qu / /chezhan /　　/jie / /ta /.
- I　drive car　　go to station　　pick up him
- *I drive a car to go to station to pick up him.*

Chinese subject-predicate predicate sentence pattern

- Example 10：
- 他 性格 坚强.
- /ta/ /xingge/ /jianqiang/
- He character strong
- *He is firm in character.*

性格(character) is a feature of 他 (he), can be omitted.

The predicate is a sentence with subject and predicate

The subject

▪离合词: the separable word

Example 11：

▪他　理　了　三　次　发。

▪/ta/　/li/　/le/　/san/ /ci/　/fa/

▪He　cut　three times hair

▪He took hair cutting 3 times.

▪理发　to take hair cutting

▪It's a word, also can be used separable.

▪理……发

▪Example 12：

▪他　结　了　三　次　婚。

▪/ta/　/jie/　/le/　/san/ /ci/　/hun/

▪He　married　three times

▪He married three times.

▪结婚　to marry

background

Language parse model

Syntax Structure

Dependency Structure

Labeled corpus

Syntactic treebank

Dependency treebank

Other relevant researches

SRL

Semantic orientation

Feature structure

Labeled corpus

- Syntactic Treebank, Dependency Treebank
- Languages：English, French, German, Hungarian, Portuguese, Czech, Chinese, Arabic, Japanese, Korean
- Label depth：word----single sentence---discourse
- Famous corpus：

▪English ：宾州树库（Penn Treebank）、Proposition Bank 、VerbNet ；FrameNet ，WordNet

▪Chinese：Penn Chinese Treebank、 汉语依存树库 Chinese dependency Treebank 、清华汉语树库（TCT）、知网（HowNet） ；台北中研院汉语树库（Sinica Treebank）、汉语框架语义知识库(Chinese FrameNet)

Feature structure in other fields

- **Not a new term**
- **Generative Phonology**
  - **to describe syllables**
  - **distinctive features**
- **GPSG (Generalized phrase structure grammar) , LFG (Lexical functional grammar)**
  - **the complexity of Feature Set**
  - **to describe the syntactic structure**

$$F = \begin{bmatrix} \text{FEATURE}_1 & \cdots & \text{VALUE}_1 \\ \vdots & & \vdots \\ \text{FEATURE}_i & \cdots & \text{VALUE}_i \\ \vdots & & \vdots \\ \text{FEATURE}_n & \cdots & \text{VALUE}_n \end{bmatrix} \qquad n \geqslant 1$$

“特征——值矩阵”(attribute-value matrix)

The difficulties

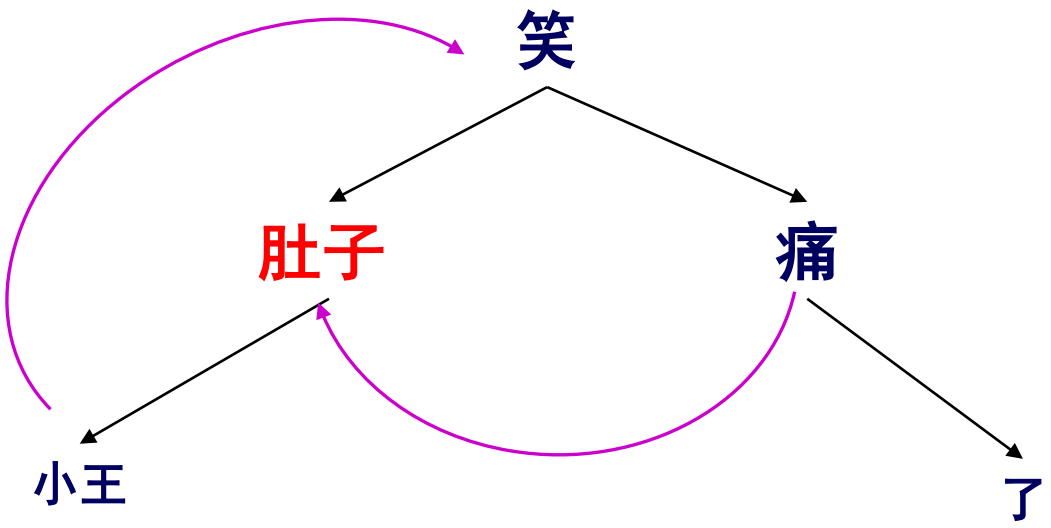
- **Semantic annotation methods**

---------how to annotate the Chinese special sentence patterns

- 北京地铁行为规范 Beijing subway action rule
- 肚子笑痛了。Belly laugh painful
- 今天星期天。Today Sunday
- 他走累了。He walk tired

- 小王　肚子　笑　痛　了。

stomach　to laugh　painful



The difficulties

Types of Dependency relations in different institutes in China

institutes		year	resources	Types of Dependency relations
Chinese academy of Social science, institute of language Li Wei		1990	Chinese dependency system for Netherland DLT	36
Tsinghua univeristy	Huang Changning Yuan Chunfa	1992	corpus, rule and statistics based parse(CRSP)	32
	Huang Changning Wu Sheng	1992	CRSP	65
	Huang Changning Zhou Qiang	1994	CRSP	106
	Huang Changning Zhou Qiang	1994	CRSP	44
	Zhou Qiang	2004	Tsinghua Chinese Treebank（TCT）	11
	Li Juanzi	2005	Semantic parse system	70
HIT-SCIR Harbin Institute of Technology		2006	Chinese dependency Treebank	34

Our work

- **Feature structure**
  - Semantic relations, types
  - Feature structure triple
  - formal representation
  - How to determine, criterion
- **Chinese Semantic resource**
  - sentences source
  - annotation criterion
  - annotation methods
  - annotation platform
- **Application**
  - Help to resolve the problems in linguistic fields
  - For Company, government

二、Feature structure

- Semantic relations, types
- Feature structure triple
- formal representation
- How to determine, criterion?

Feature structure

从广州飞到武汉

From Guangzhou fly to Wuhan

⟹ （飞，从，广州）(fly, from, Guangzhou)

（飞，到，武汉）(fly, to, Wuhan)

红围巾

红色围巾

red scarf

⟹ (围巾，颜色，红) (scarf, color, red)

大头皮鞋

shoes with big head

⟹ （皮鞋，部分，头）(shoes, part, head)

（头，形状，大）(head, shape, big)

大范围推广

Large-scale promotion

⟹ (推广，范围，大）

(promotion, scale, Large)

Feature structure triple

- **Generally, a phrase or sentence can be expressed as a set of triples of Entity, Feature and Value.**
- **We call the set "Feature Structure" of the phrase or sentence structure.**

Feature Triple: [Entity, Feature, Value]

从广州飞到武汉 ⟹ [飞，从，广州] [fly, from, Guangzhou]
[飞，到，武汉] [fly, to, Wuhan]

红围巾
红色围巾 ⟹ [围巾，$，红]
[scarf, , red]

大头皮鞋 ⟹ [皮鞋，$，头] [shoes, , head]
[头，$，大] [head, , big]

大范围推广 ⟹ [推广，范围，大]
[promotion, scale, Large]

中 高 级 职称 研究 人员

Middle high grade title research staff

[研究， ，人员]

[人员，职称，]

[职称， 级，中高]

The Feature, Value ,both can be another Entity in other FS triples.

老陈 说 明天 很 冷。

Chen says tomorrow very cold

[说， ，老陈]

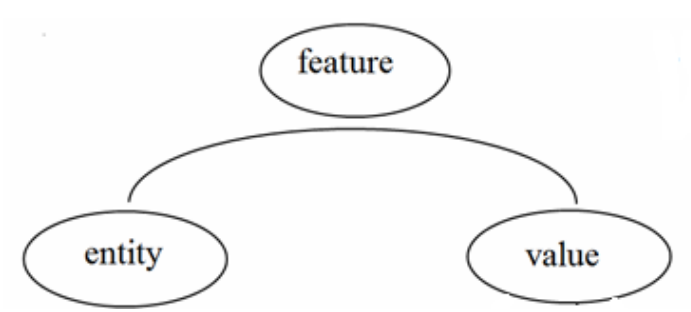
[说， ，[[冷， ，明天] [冷， ，很]]]

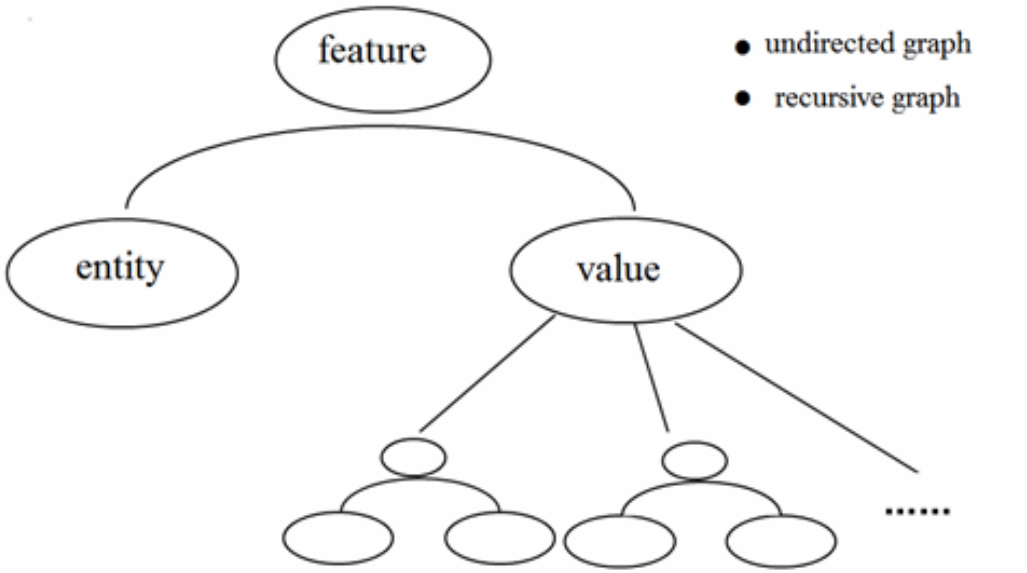
The Value can be one or many FS triples.

formal representation of Feature structure

Undirected graph

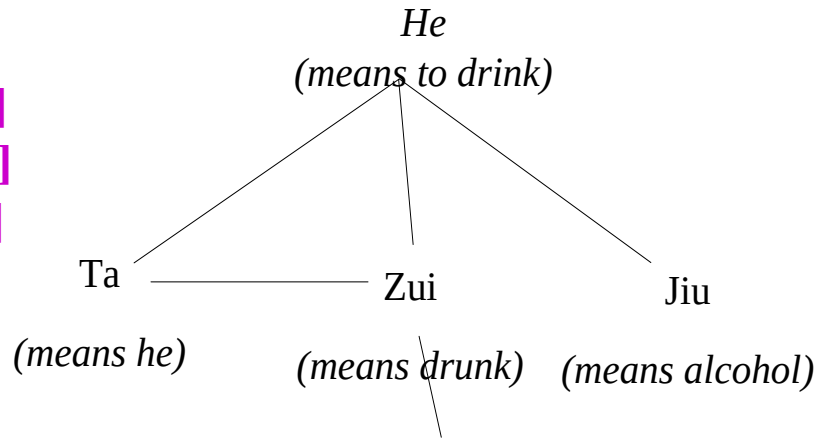
Recursive graph



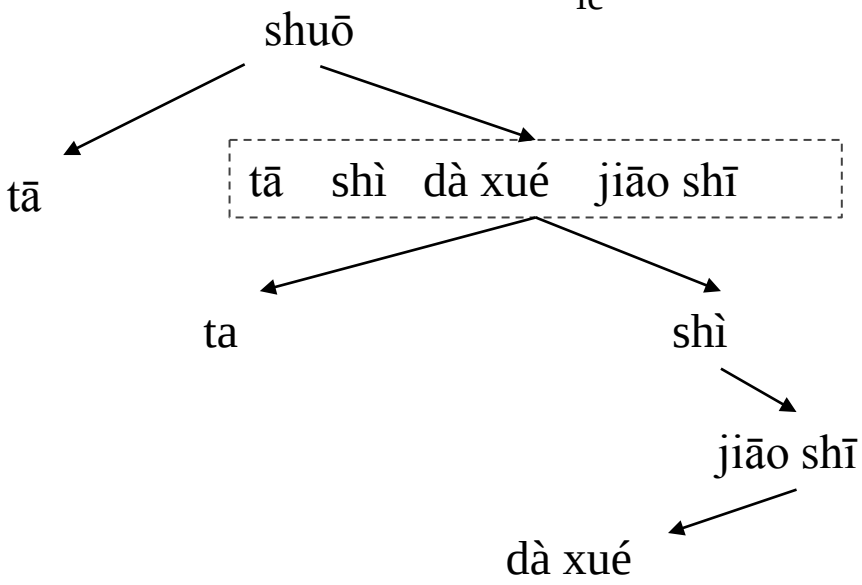


- 他　喝　醉　了　酒
- Tā　hē　zuì　le　jiǔ
- He　to drink　drunk　alcohol
- *He is drunk.*

[he, ,ta]
[he, ,jiu]
[he, ,zui]
[zui, ,ta]
[zui, ,le]



- 他　说　他　是　大学　教师。
- Tā　shuō　tā　shì　dà xué　jiāo shī
- He　say　he　is　university　professor
- *He says he is university professor.*



The determination of FS

Transformation

A 他喜欢文静女生。

He likes girls with quiet character.

B 他喜欢性格文静的女生。

character

C 他喜欢文静性格女生。

Interrogative

他喜欢性格怎么样的女生？

What kind of character does he like girls with?

他喜欢谁？

Whom does he like?

谁喜欢文静女生？

who likes girls with quiet character?

小王　　笑　　痛　了　肚子。

Xiaowang　to laugh　painful　　stomach

- 谁笑痛了肚子？ Who laughs?
- 哪里痛了？ Where feel painful?
- 肚子怎么了？ how about stomach ?
- 肚子怎么痛了？ What cause the stomach painful?

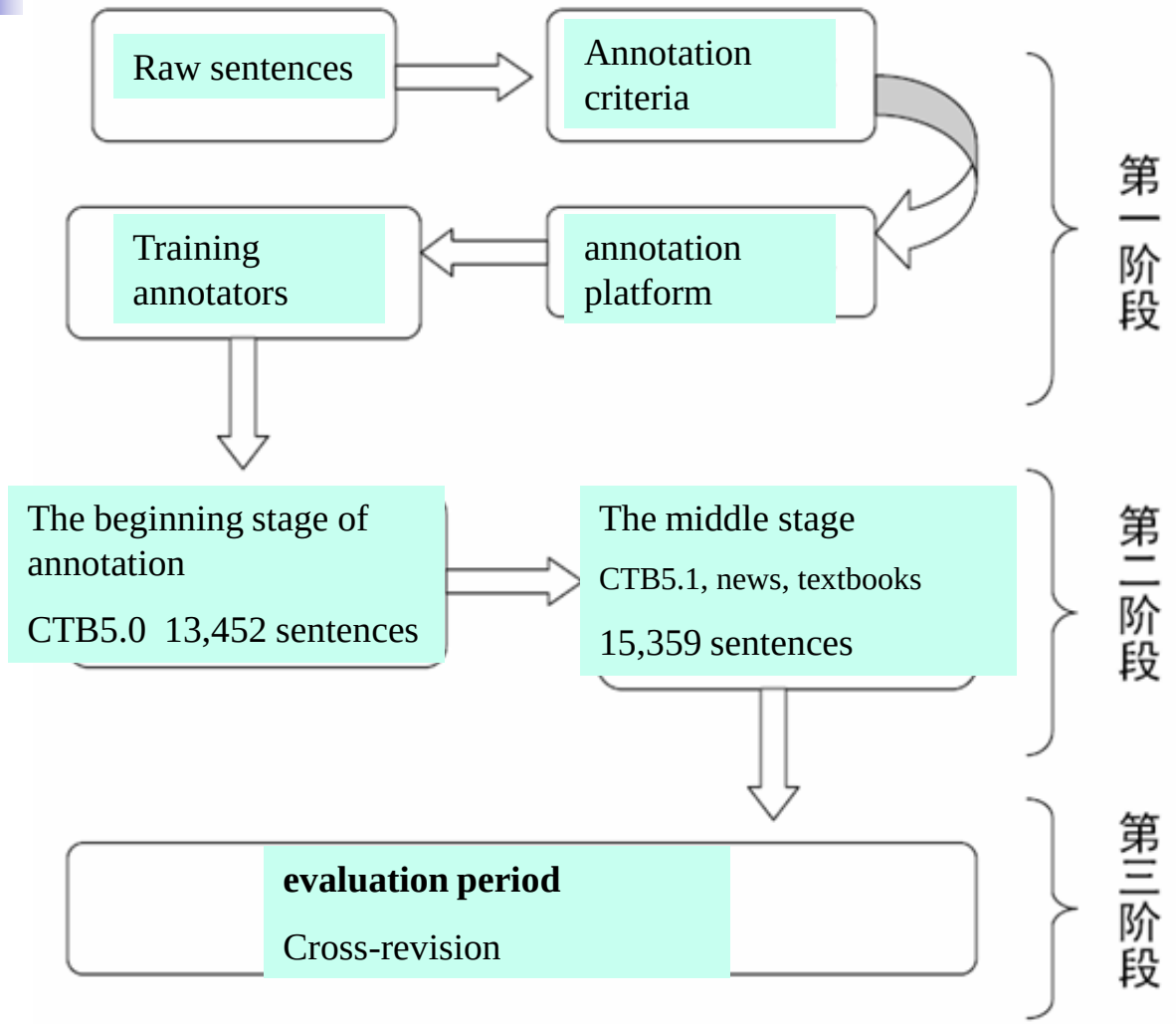
The types of FS triples

	Entity	Feature	Value	Examples
[Entity，Feature，Value]	+	+	+	新一代有比较多机会**用外在表现自我** [表现，用，外在]
[Entity，$，Value]	+	-	+	**柯庆明说** [说，，柯庆明]
[$，Feature，Value]	-	+	+	乡民**付出的**往往不是钱 [，的，付出]
[Entity，Feature，$]	+	+	-	外表都**被蓄意要求**「禁欲」 [要求，被，]
[Entity，$，$]	+	-	-	走！ 好！

三、Construction of Chinese Semantic resource

- sentences sources
- annotation criterion
- annotation methods
- annotation platform

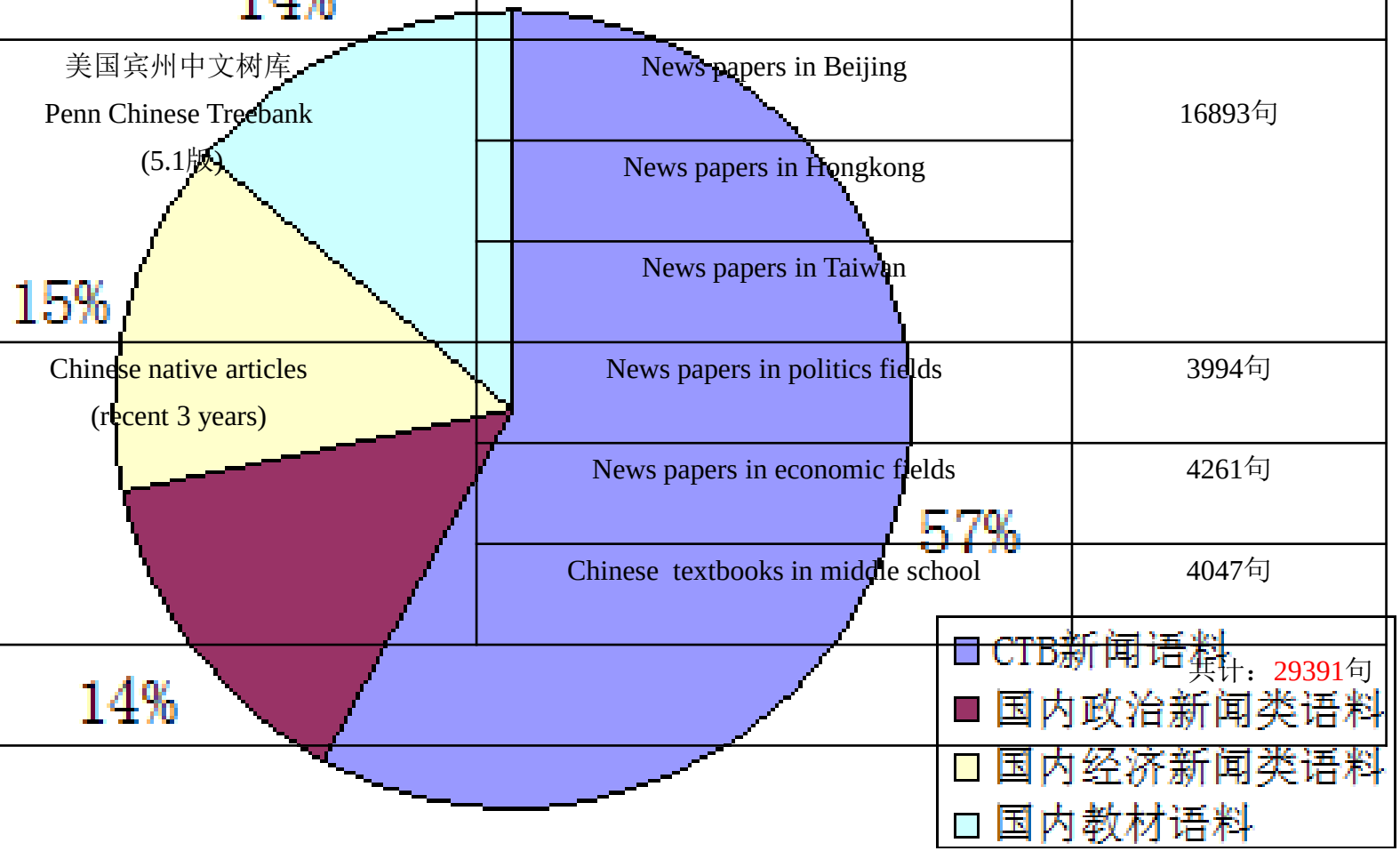
The flowchart



汉语语义标注语料库流程图

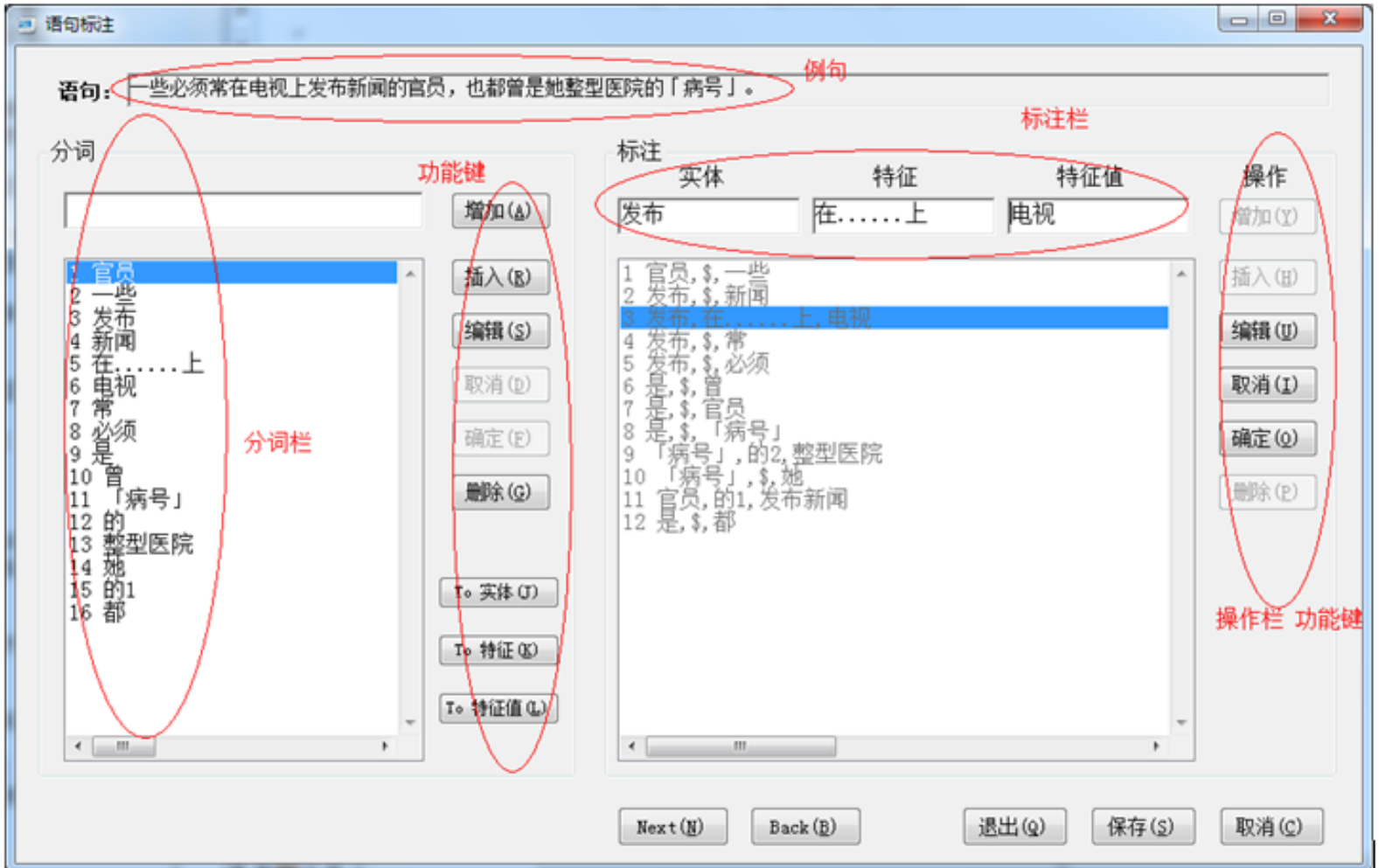
sources

语料来源	语料类别	数量
美国宾州中文树库 Penn Chinese Treebank (5.1版)	News papers in Beijing	16893句
	News papers in Hongkong	
	News papers in Taiwan	
Chinese native articles (recent 3 years)	News papers in politics fields	3994句
	News papers in economic fields	4261句
	Chinese textbooks in middle school	4047句
		共计：29391句



annotation

- manual annotation + annotation platform



FS criterion: question patterns

提问模板	例句	提问方式	答案
谁+实体（who）	我买苹果。	谁买？	[买， ，我]
实体+谁（who）	我去接小王。	接谁？	[接， ，谁]
特征+谁+实体（who）	与上海建立联系	与谁建立联系？	[建立，与，上海]
实体+什么(what)	买书	买什么？	[买， ，书]
特征+什么 (what)	因为A，所以B。	因为什么？所以什么？	[A，所以，B] [B，因为，A]
实体+特征+什么(what)	拉住我的手	拉住什么？	[拉，住，我的手]
实体+多少+量词？ (how many)	看过一两次	看过多少次？	[看，次，一两]
多少+实体(how much)	一些钱	多少钱？	[钱， ，一些]
实体+多久+特征(how long)	找了很长时间	找了多长时间？	[找， 时间， 长]
什么时候+实体(when)	那天晚上他喝酒喝醉了。	什么时候喝？	[喝， ，晚上]
实体+特征+什么时候(when)	弹琴弹到半夜	弹到什么时候？	[弹，到，半夜]
哪+实体(which)	今天晚上	哪个晚上？	[晚上，，今天]
实体+怎么样了？ (how)	那天晚上他喝酒喝醉了。	喝的怎么样了？	[喝， ， 醉]
为什么+实体？ (why)	那只鸟受惊飞走了。	为什么飞？	[飞， ，受惊]
特征+哪儿(where)	妹妹在大学念书。	在哪儿念书？	[念书，在，大学]
实体+吗？	放吗？	……吗？	[吗，，放]

Annotation criterion

- **Segmentation criteria**
- **Annotation criterion**

一	principles：
1、	Minimum unit
2、	Semantic relations
3、	recursive
4、	Non-head,　equal status
二	About Uncertain sentences
三	Compatibility with other institutes

Segmentation criteria

- **Basically, CTB 5.1**
- **add:**

一、Proper nouns, fixed phrases, habitual word combinations, etc., as a unit with no segmentation

二、Cardinal and ordinal numbers, as a unit with no segmentation.

如:“一亿三千万”、“三分之一”、“十”、“1989.98”、、“二十来岁”、“二十多岁”、“二十余岁”中的“二十来”、“二十多”、“二十余”等。

三、time as a unit with no segmentation.

如:“1999年10月10日12点50分”、“星期六”、“1965年”、“10月份”等。

四、if there are same words in a sentence, use “subscript” to be distinguished.

如:中国1，中国 2，中国3;的1，的2，的3等。

Annotation criterion

- **Syntactic**

句法成分充当特征三元素分布表

	实体	特征	特征值	例句
主谓	谓语		主语	[吃， ，他]
定中	中心语		定语	白布 [布， ，白]
状中	中心语		状语	很多 [多， ，很]
数量	可数名词/动词	量词	数词	一个人[人，个，一] 跑一趟 [跑，趟，一]
动补	动词		补语	做完[做 ， ，完]
动宾	动词		宾语	[喝 ， ，咖啡]

Annotation criterion

- **Part of Speech**

充当特征三元素的成分总结表

	常见词性	说明	例句
实体	名词 **noun**	定中短语中的中心词，数量短语的中心词	[书 ， 本 ，三]
	动词 **verb**	主要是实义动词	[发挥， ，作用]
	连词 **conj**	并列关系的，如：和、与、或者、或	[和， ，前景] [和， ，问题]
	形容词 **adj**	状中短语中的中心词	[多， ，很]
	成对出现的 **connectives**	既 A，又(也)B	
	顿号 **Sign of coordination**		[、， ，北京] [、， ，上海]
	语气词	吗、呢、吧、啊、的	[吧， ，走]

Entity

Modal partical

The feature word

充当特征三元素的成分总结表

	常见词性	说明	例句
特征	单个介词	为了、鉴于、关于、使、让、给、被	[统计，据 ，][誉为，被，]
	介词短语	在……中	[发挥，在……中，改善]
	量词	个，次	[台商，名，三万]
	结构助词	“的、地、得、之、所”等	[跑，地，飞快]
	单个出现的关联词	因为、由于、那么	[他没考好，因为，没有休息好]

Feature

preposition

Prep-phrase

Measure word

Structural auxiliary word

Connective words or phrases

The value

充当特征三元素的成分总结表

	常见词性	说明	例句
特征值	着、了、过	dynamic auxiliary	[吃， ，着]
	数词 number	数量短语中的数词	[人，个，一]
	形容词 adj	定中结构中的定语	[作用， ，显著]
	副词 adv	定中结构中的定语	[好， ，非常]
	代词 pronoun		[希望， ，我们]
	名词 noun	动宾结构中的宾语和主谓结构中的主语	[爱， ，我]
	助动词	能、能够、会	[游泳， ，敢]

Value

auxiliary verb

Summary：

- 2008-2012；
- manual annotation；
- 29，391 sentences;
- 1，301，974 characters
- universality
- Can be used directly to Relation Extraction, Event Extraction, Automatic Question & Answering

汉语句子级语义标注资源规模

标注资源构成	数量
句子	29，391 句
词语	601，947 词
字	1，301，974 字
特征三元组	397，115 个

汉语句子级语义标注资源特征三元

特征三元组构成[25]	语法单位(个)	个数百分比	总频次(次)
实体	60，556	33%	393,403
特征	2，650	1%	76,378
特征值	118，108	66%	393850

汉语句子级语义标注资源特征三元组类型分布

特征三元组类型	数量	百分比
[实体，特征，特征值]	71214	17.9%
[实体，　，特征值]	320408	80.6%
[　，特征，特征值]	2876	0.72%
[实体，特征，　]	2292	0.58%
[实体，　，　]	325	0.08%
共计	397，115	100%

Future work

- To extend the scope of the study of Chinese language phenomenon

  兼语句、是字句、存现句、把字句、被字句

- resource construction: from sentences to discourses

  Event chain

  Complex noun phrase

  Connective structure

- the composition and distribution of feature word set

- the sharing of the resource

Application

- **Public opinion monitoring system**
  - **For government For hospitals**
  - **For companies**

Thank you!