文本挖掘之句法分析

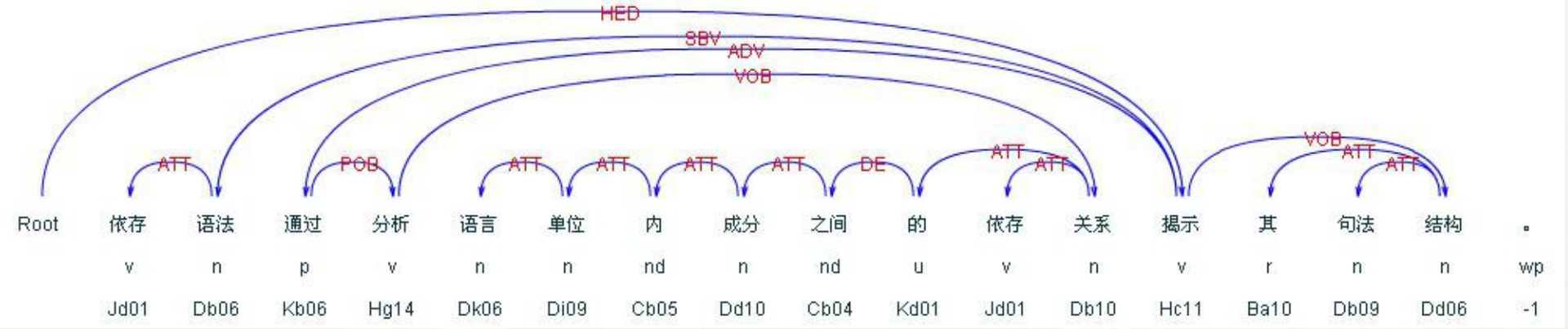
杨展富

7.13

- 一、句法分析简介
- 二、树库简介
- 二、CTB简介
- 三、基于依存语法构建多视图汉语树库
- 四、Dependency Tree Representations of Predicate-Argument Structures
- 五、句法分析一些接口

- 树库构建体系的正确性和完善性进行验证
- 搜索引擎用户日志分析和关键词识别,比如信息抽取、自动问答、机器翻译等

- 依存文法最早由法国语言学家L.Tesniere在其著作《结构句法基础》（1959年）中提出，对语言学的发展产生了深远的影响，特别是在计算语言学界备受推崇。



- 依存语法通过分析语言单位内成分之间的依存关系揭示其句法结构。

五条公理：

- 1、一个句子中只有一个成分是独立的；
- 2、其它成分直接依存于某一成分；
- 3、任何一个成分都不能依存与两个或两个以上的成分；
- 4、如果A成分直接依存于B成分，而C成分在句中位于A和B之间，那么C或者直接依存于B，或者直接依存于A和B之间的某一成分；
- 5、中心成分左右两面的其它成分相互不发生关系。

- 一、句法分析简介
- 二、树库简介
- 二、CTB简介
- 三、基于依存语法构建多视图汉语树库
- 四、Dependency Tree Representations of Predicate-Argument Structures
- 五、句法分析一些接口

树库简介

- 树库(*treebank*)就是一种经过了结构标注的语料库。

- 如果考虑歧义，那么一个句子可能对应多棵树。大量句子以及其对应的树结构的集合就构成树库。

树库简介

- 首先，它可为基于统计的自动句法分析器提供必要的训练数据和统一的测评平台；
- 其次，它能为汉语句法学研究提供真实文本标注素材，便于语言学家从中总结语言规则和规律；
- 第三，它是进一步进行句子内部的词语义项和语义关系标注的基础。

- 一、句法分析简介
- 二、树库简介
- 三、CTB简介
- 四、基于依存语法构建多视图汉语树库
- 五、Dependency Tree Representations of Predicate-Argument Structures
- 六、句法分析一些接口

CTB简介

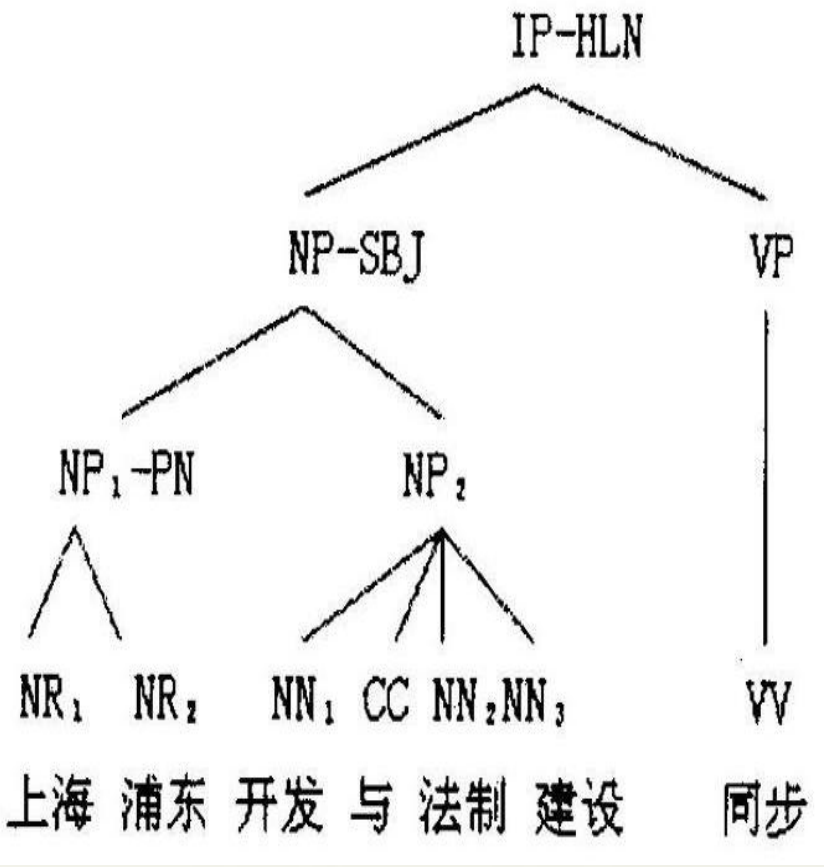
- 宾州大学汉语树库(CTB)的目标是建立一个100万词的经过句法标注的语料库。
- 它是基于短语结构的，进行了短语结构、短语功能、空元素、指数的标注。
- CTB 到目前发展至7. 0版。CTB7.0语料取材于中国新闻，新闻杂志，各种广播新闻和广播谈话节目，新闻和博客网站。

CTB简介

- LDC中文树库(CTB)属于短语结构树库，采用句子的结构成分描述句子的结构。

CTB简介

```
(IP-HLN (NP-SBJ (NP-PN (NR 上海)
                       (NR 浦东))
                (NP (NN 开发)
                    (CC 与)
                    (NN 法制)
                    (NN 建设)))
        (VP (VV 同步)))
```



CTB优点

- （1）短语结构树可以表示句子较全面的句法信息。
- （2）采用短语结构可以有效地结合现有研究成果。
- （3）按照不同的应用需求，树结构可以转换为骨架分析树和依存关系树等。

CTB缺点

- 运用英语的语法框架来分析汉语，有的时候跟汉语为母语的语感不符。
- 标注的颗粒度有时候比较粗，在向依存结构树库转换时就会出错。有的地方的层次还应该细分等。

- 一、句法分析简介
- 二、树库简介
- 三、CTB简介
- 四、基于依存语法构建多视图汉语树库
- 五、Dependency Tree Representations of Predicate-Argument Structures
- 六、句法分析一些接口

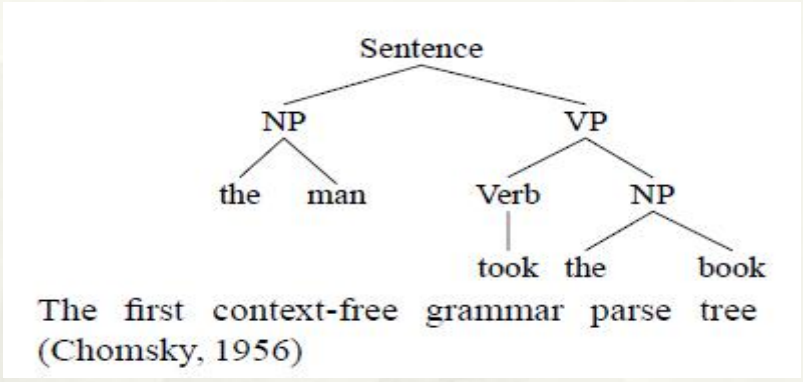
基于依存语法构建多视图汉语树库

邱立坤、金澎、王厚峰

北京大学计算语言学研究所

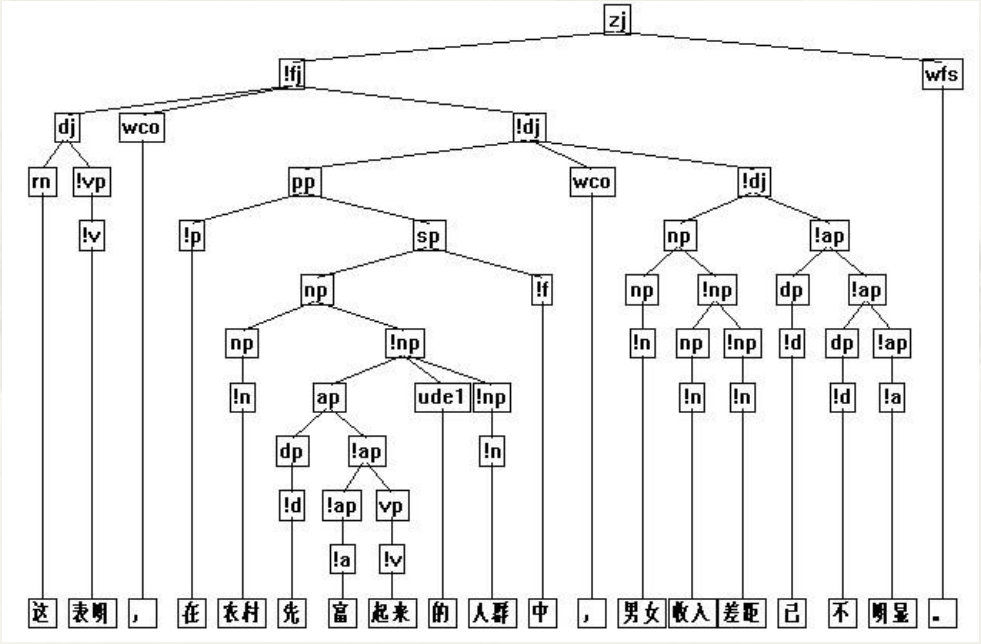
树库的类型1：短语结构树

- 可以显示的信息
  - 层次
  - 中心语，
  - 短语整体功能
  - 语法结构关系



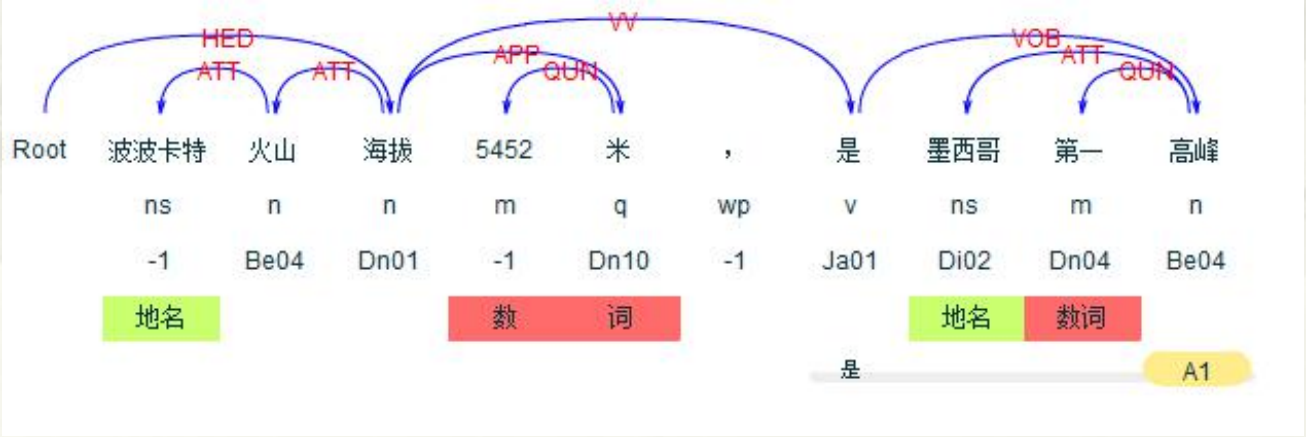
The first context-free grammar parse tree (Chomsky, 1956)

```
(IP (NP-SBJ (NR 中国))
    (VP (PP-LOC (P 在)
                (LCP (NP (DNP (NP-PN (NR 西门子))
                              (DEG 的))
                         (NP (NP-PN (NR 亚太))
                             (NP (NN 发展)
                                 (NN 战略))))
                     (LC 中)))
        (VP (VV 处于)
            (NP-OBJ (ADJP (JJ 重要))
                    (NP (NN 地位))))))
```



树库的类型2：依存树

- 可以显示的信息
  - 中心语
  - 语法结构关系（语法角色）
  - 语义结构关系（语义角色）



树库的类型3:组合范畴语法树及其它

- 可以显示的信息：
  - 整体功能（组合范畴）
  - 中心语
  - 谓词论元关系（组合范畴语法*CCG*区别于传统上下文无关文法的一个显著特性）

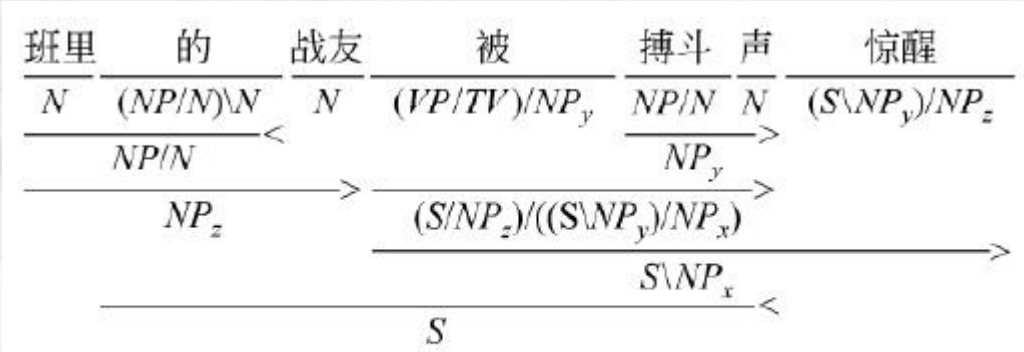


图 4　被动结构的 CCG 分析结果

树库转换中的问题

- 树库转换中面临着各种歧义问题
- 从短语结构语法到依存语法
  - 从整体功能信息生成语法角色信息
- 从依存语法到短语结构语法
  - 从中心语和语法角色生成整体功能和层次信息
- 从短语结构语法到组合范畴语法
  - 如何生成谓词论元关系

多视图树库的提出

- 本文提出多视图（Multi-view）树库的概念
- 多视图树的“多”首先体现在构建阶段
- 多视图树的“多”还体现在使用阶段

基于依存语法的多视图树库框架

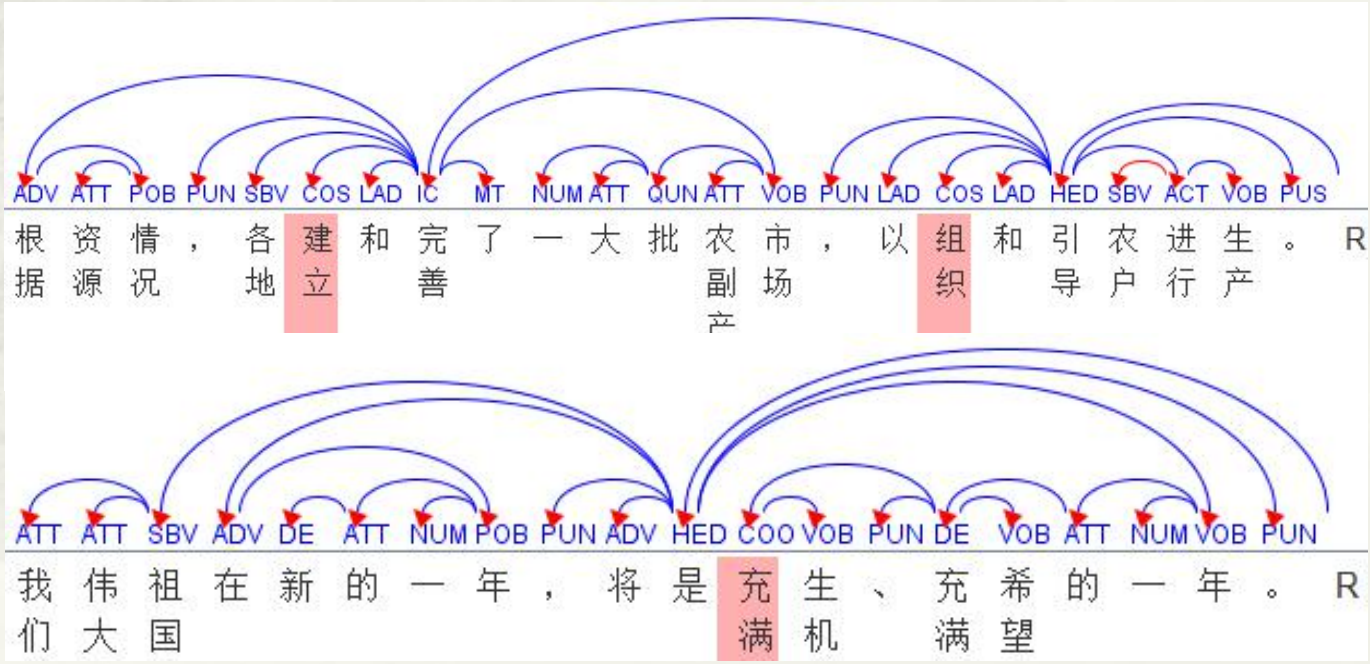
- 在本文中，我们主要讨论含有短语结构语法和依存语法两个视图的多视图树
  - 以依存视图为核心
  - 在句法层面上仅仅标注中心语和语法角色两类信息，自动转换出相应的短语结构树
  - 需要解决的关键问题是短语整体功能的推导和层次信息的推导。

短语整体功能推导的具体实现

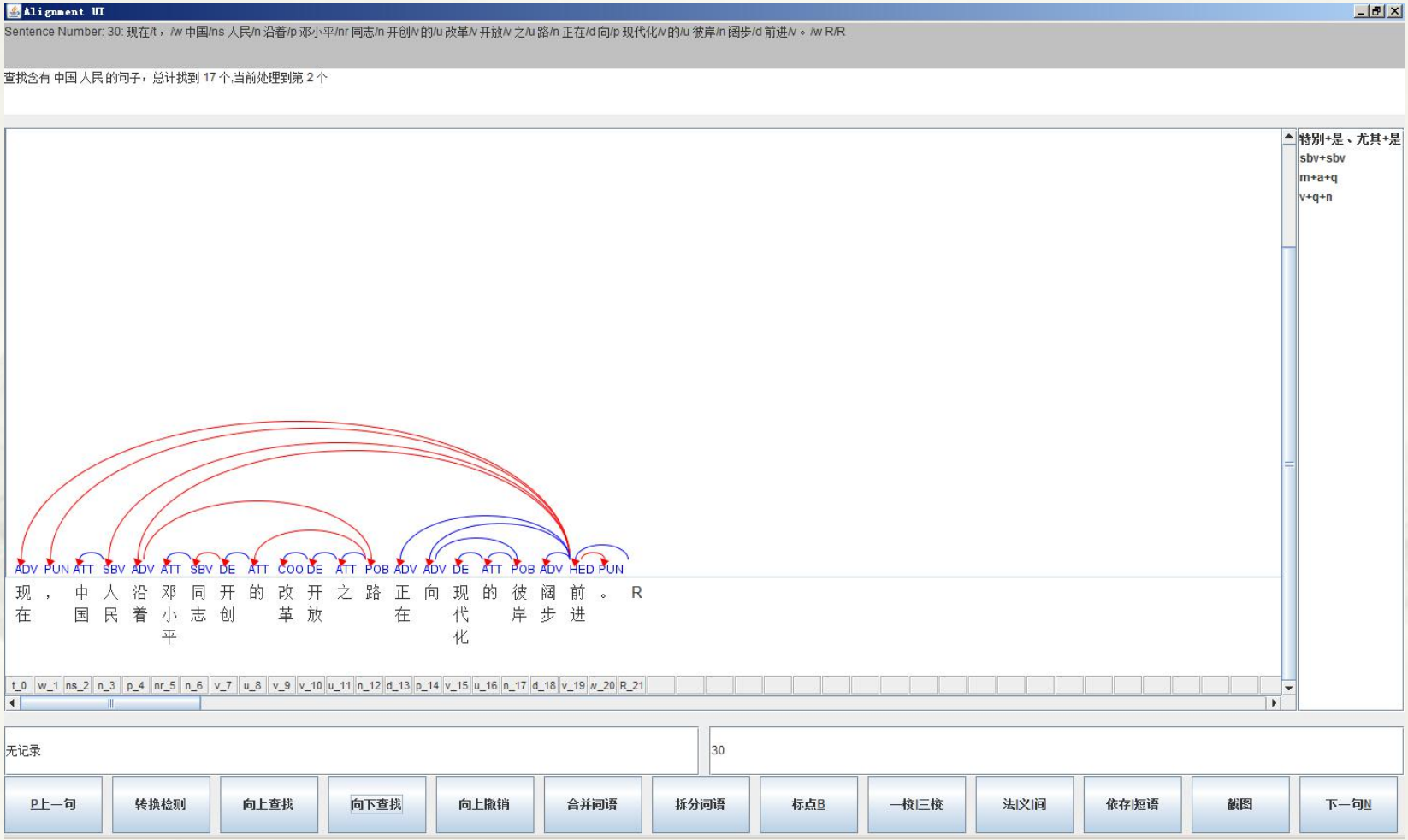
- 基于规则的推导方式：
  - 父结点词类+子结点词类+语法角色=>短语整体功能标记
  - 通过递归的方式，可以依次获得各短语的直接成分的整体功能标记，以取代上述规则中的词类

层次信息的可推导性：例证

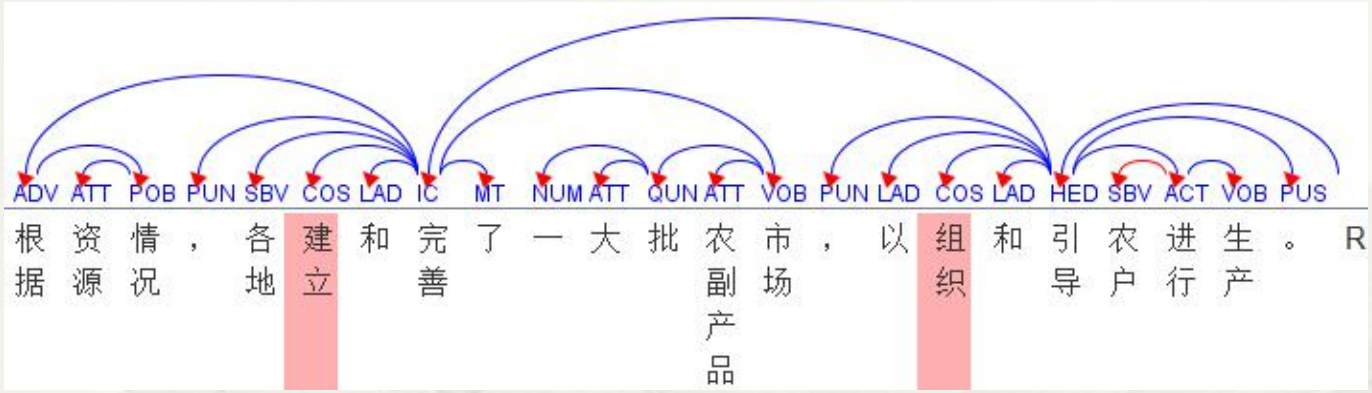
- 共享并列（*COS*），一般并列（*COO*）。
  - 共享并列先与并列成分归并
  - 一般并列先与子结点归并

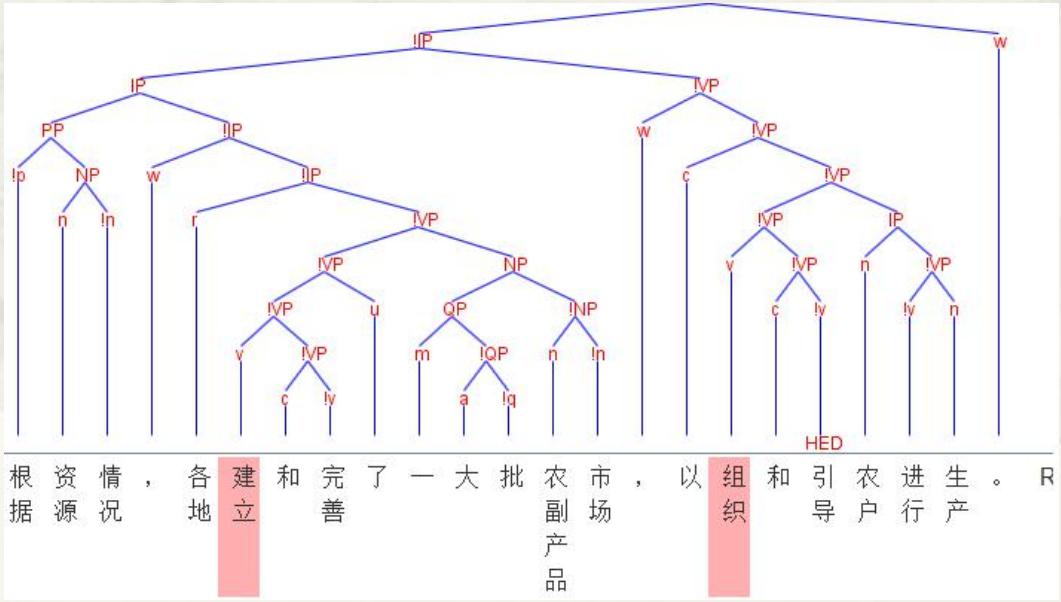


多视图树库标注工具

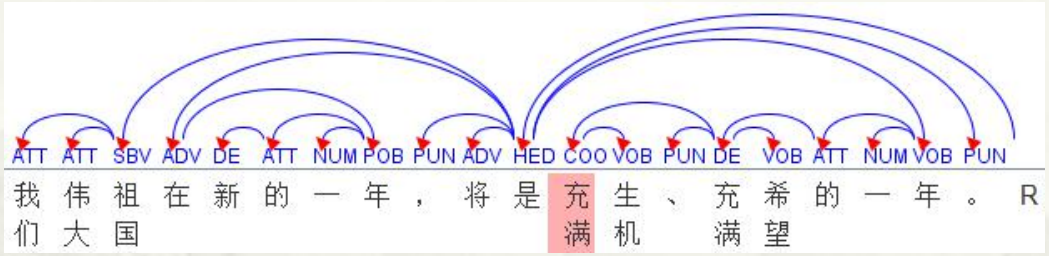

Alignment UI
Sentence Number: 30: 现在/t ，/w 中国/ns 人民/n 沿着/p 邓小平/nr 同志/n 开创/v 的/u 改革/v 开放/v 之/u 路/n 正在/d 向/p 现代化/v 的/u 彼岸/n 阔步/d 前进/v 。/w R/R
查找含有 中国 人民 的句子，总计找到 17 个,当前处理到第 2 个
特别+是、尤其+是
sbv+sbv
m+a+q
v+q+n
ADV PUN ATT SBV ADV ATT SBV DE ATT COO DE ATT POB ADV ADV DE ATT POB ADV HED PUN
现 在 ， 中 国 人 民 沿 着 邓 小 平 同 志 开 创 的 改 革 开 放 之 路 正 在 向 现 代 化 的 彼 岸 阔 步 前 进 。 R
t_0 w_1 ns_2 n_3 p_4 nr_5 n_6 v_7 u_8 v_9 v_10 u_11 n_12 d_13 p_14 v_15 u_16 n_17 d_18 v_19 /v_20 R_21
无记录
30
P上一句
转换检测
向上查找
向下查找
向上撤销
合并词语
拆分词语
标点B
一校|三校
法|义|间
依存|短语
截图
下一句N

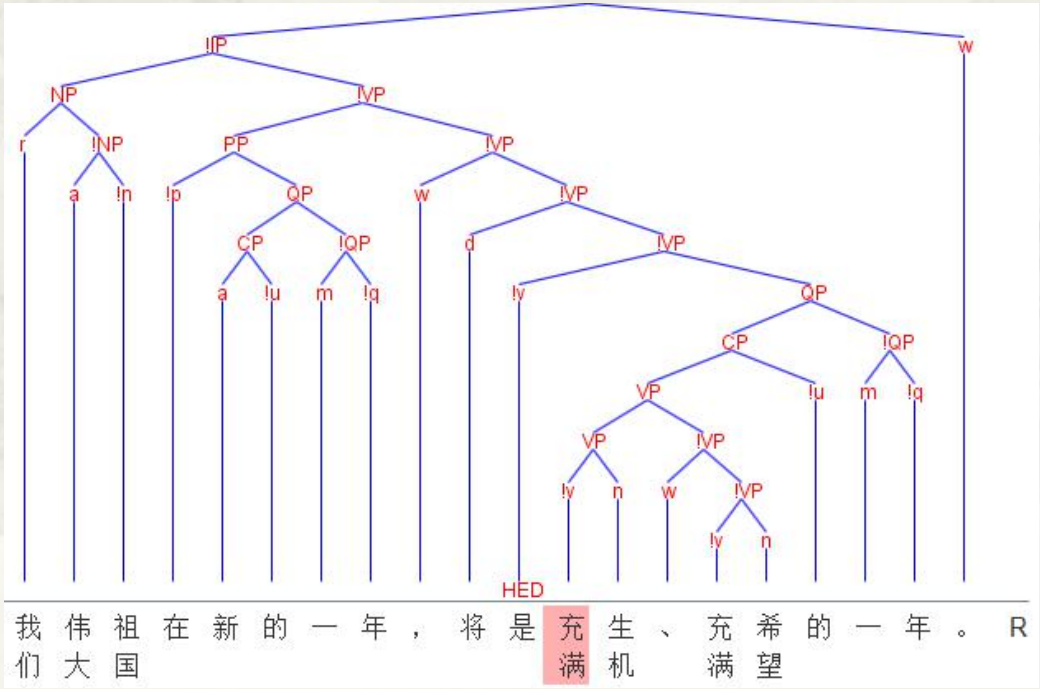
转换示例1


ADV ATT POB PUN SBV COS LAD IC MT NUM ATT QUN ATT VOB PUN LAD COS LAD HED SBV ACT VOB PUS
根 据 资 源 情 况 ， 各 地 建 立 和 完 善 了 一 大 批 农 副 产 品 ， 以 组 织 和 引 导 农 户 进 行 生 产 。 R


根 据 资 源 情 况 ， 各 地 建 立 和 完 善 了 一 大 批 农 副 产 品 ， 以 组 织 和 引 导 农 户 进 行 生 产 。 R
HED

转换示例2


ATT ATT SBV ADV DE ATT NUM POB PUN ADV HED COO VOB PUN DE VOB ATT NUM VOB PUN
我 伟 祖 在 新 的 一 年 ， 将 是 充 生 、 充 希 的 一 年 。 R
们 大 国 满 机 满 望


我 伟 祖 在 新 的 一 年 ， 将 是 充 生 、 充 希 的 一 年 。 R
们 大 国 满 机 满 望
HED

- 一、句法分析简介
- 二、树库简介
- 三、CTB简介
- 四、基于依存语法构建多视图汉语树库
- 五、Dependency Tree Representations of Predicate-Argument Structures
- 六、句法分析一些接口

Dependency Tree Representations of Predicate-Argument Structures

Likun Qiu, Yue Zhang and Meishan Zhang

AAAI会议A类文

Abstract

- 作者提出一种新颖的谓词论元结构（Predicate-argument structure）表达体系——语义角色传递机制，借助该机制以简单的依存树结构即可完整地表达谓词论元结构信息，在降低语义角色自动标注复杂度的同时还使得句法和语义分析一体化计算更为自然、更加便捷。

Introduction

- Argument（论元）：指带有论旨角色的名词短语。
- Thematic role（论旨角色）：由谓词根据其与相关的名词短语之间语义关系而指派(assign)给这些名词短语的语义角色，主要有施事agent、受事patient、客体theme、经验者experiencer、受益者beneficiary、工具instrument、处所location、目标goal和来源source。
- Argument Position（论元位置）:论元在句中所占的位置

Introduction

- 今天的风儿真喧嚣啊！
- 今天的风儿真喧嚣啊！ 论元
- 今天的风儿真喧嚣啊！ 论旨角色（经验者experiencer）
- 今天的风儿真喧嚣啊！ 谓词

Introduction

	他	去过	北京	，	后来	离开	了
去过.01	A0		A1				
离开.01	A0				ADV		

Figure 1: PB-style structures. ("他 (he) 去过 (went to) 北京 (Beijing) ， (,) 后来 (then) 离开 (left) 了 (le; a function word)" (He went to Beijing, and then left.))

Annotation Framework 注释框架

- 我们的标注：从北大的多视图语料库.33个子节点充当父节点的宾语（VOB）标签，32个句法标签.vob是一个简化版的标签集合.

- 句法标签集, including SBV (subject), TPC (topic), VOB (direct object), ACT (action object), POB (prepositional object),COS (sharing-object coordinate), COO (non-sharing-objectcoordinate), DE (modifier of “的” (special function word)) and HED (root of a sentence)句子的中心, 是设计注释框架的中心.

语义角色定义

- Rule 1:如果在任何一个框架中，有两个语义角色共同出现，它们都是对立的。否则，它们是互补的。
- Rule 2:如果2个语义角色是互补的，他们可以被视为相同的语义角色。

（A）规范的语义角色

- 主　谓　宾

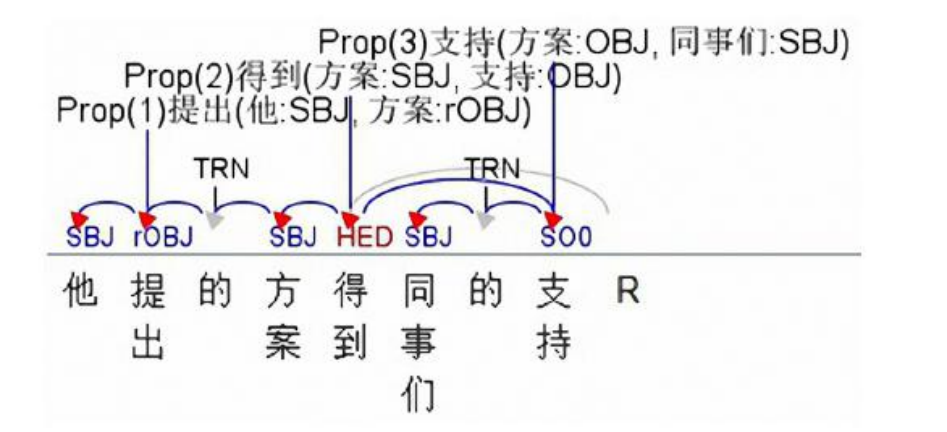
	Tag	Content	Freq
c	ACT	action object	1803
o	DAT	contrast, beneficiary, partner, target	5546
r	FIN	location finite, state finite, time finite	1644
e	OBJ	content, isa, patient, possession, result, OfPart	31169
	POS	possessor acted by topic	538
	SBJ	agent, coagent, existent, experiencer, possessor, relevant, whole	28523

	Tag	Content	Freq	Tag	Content	Freq
p	CAU	cause	381	COS	cost	22
e	DEG	degree	10	DIR	direction	127
r	BSS	basis	585	INS	instrument	279
i	LOC	location	3283	MAN	manner	385
p	MAT	material	13	PUR	purpose	331
h	QUN	quantity	702	RAN	range	164
e	SCO	scope	700	THR	location through	108
r	TIM	time, duration	5043			
a						
l	INI	location initial, state initial, time initial				618

Table 1: Canonical semantic roles.

（B）反向语义角色

- 用于在关系从句，在谓词修改参数（直接或间接）语法



规范或反向

- Rule 3：如果一个动词或形容词的语义角色标注、句法标记是DE，并且其母词的句法标注是ATT那么是对立语义角色；否则，是规范角色。

（C）外联动词传输标签

- （1）介词 于，向
- （2）动词 是
- （3）助词 的
- （4）轻动词 给予，进行
- （5）位词 后，来
- （6）名词性语素 时

（D）内联动词传输标签

（1）在源谓词上所传递的短语的句法作用

（2）在目标谓词上所传递的短语的语义作用

（3）传输方向（即向内或向外）

我们使用标签，如SS1和SSO表示谓词间传输，其中三个字母分别对应以上的（1）、（2）和（3）

其他各种结构表示

- 在主语控制结构中：

  主从句的主语与一个嵌入谓语的空元是同指的。

- “约翰不愿离开”

- 在对象控制结构中：

  主子句的对象与一个嵌入谓词的空元是共同引用的。

- “汤姆说服汤姆离开”

- 在动词并列结构中：

共享对象的并列结构（*COS*）和非共享对象并列结构（*COO*）。他们的区别在于，两个并列的动词是否在右边分享一个对象。

- 在特殊组合结构中：

在某些情况下，几种类型的特殊结构共同发生在一个子句中将会对语言分析产生巨大的影响

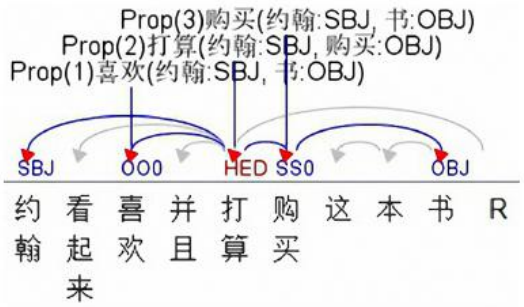


Figure 6: Special constructions. (约翰 (John) 看起来 (seems) 喜欢 (to like) 并且(and) 打算 (plan) 购买 (to buy) 这 (this) 本 (a classifier) 书 (book) (John seems to like and plan to buy this book.)

experiment

采用北大多视图树库1.0进行试验

2400个句子作为测试集，12000个句子作为训练集

采用基于记忆的学习方法

experiment

	Syntax	LAS	UAS	PF1	SF1 (%)
Baseline	auto	82.82	85.69	43.35	74.07
	gold	—	—	61.42	83.87
CRF	auto	82.82	85.69	44.69	74.16
	gold	—	—	64.00	84.50
MLN	auto	82.82	85.69	**46.33**	**75.21**
	gold	—	—	**66.50**	**85.73**

前景

1，提供了语义资源
2，自动标注语义角色
3，倒排索引
4，搜索引擎

- 一、句法分析简介
- 二、树库简介
- 三、CTB简介
- 四、基于依存语法构建多视图汉语树库
- 五、Dependency Tree Representations of Predicate-Argument Structures
- 六、句法分析一些接口

工业应用句法分析平台

- 1.腾讯API

域名：wenzhi.api.qcloud.com

接口名: TextDependency

- 输入参数

参数名称	必选	类型	描述
content	是	String	待分析的文本（只能为utf8编码）

2.语言云（哈工大）

参数名	含义	说明
api_key	用户注册语言云服务后获得的认证标识	
text	待分析的文本。	请以UTF-8格式编码，GET方式最大10K，POST方式最大20K
pattern	用以指定分析模式，可选值包括ws(分词)，pos(词性标注)，ner(命名实体识别)，dp(依存句法分析)，sdp(语义依存分析)，srl(语义角色标注)，all(全部任务)	plain格式中不允许指定全部任务
format	用以指定结果格式类型，可选值包括xml(XML格式)，json(JSON格式)，conll(CONLL格式)，plain(简洁文本格式)	在指定pattern为all条件下，指定format为xml或json，返回结果将包含sdp结果，但conll格式不会包含sdp结果；
xml_input	用以指定输入text是否是xml格式，可选值为false(默认值)，true	仅限POST方式
has_key	用以指定json结果中是否含有键值，可选值包括true(含有键值，默认)，false(不含有键值)	配合format=json使用
only_ner	用以指定plain格式中是否只需要ner列表，可选值包括false(默认值)和true	配合pattern=ner&format=plain使用
callback	用以指定JavaScript调用中所使用的回调函数名称	配合format=json使用