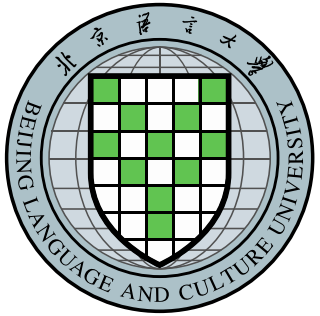


句式结构树库的自动构建

谢晨晖 胡正升 杨麟儿 廖田昕 杨尔弘

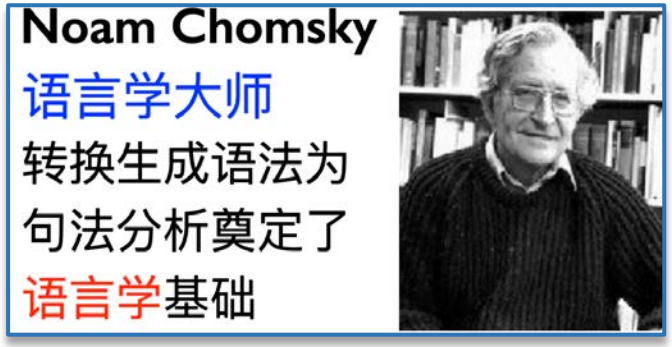
北京语言大学

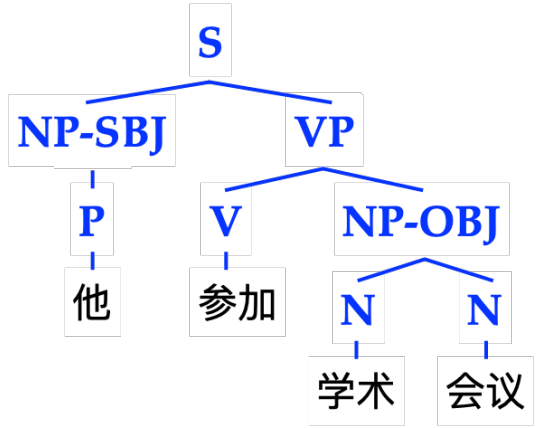
2022年10月29日

ICALL
智能语言学习

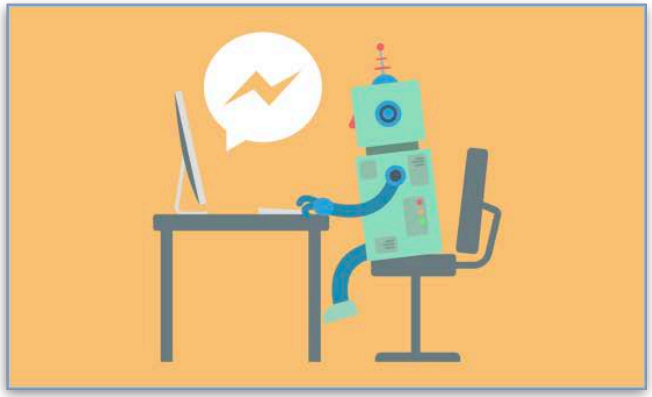
句法分析重要性

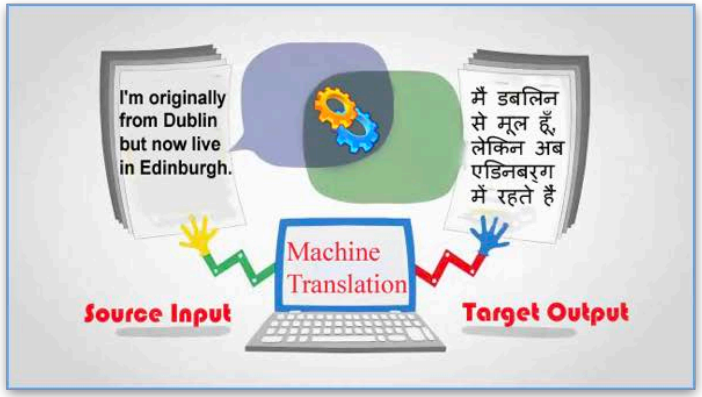
- **理论：揭示人类语言的产生机制**





- **应用：**





句法分析是自然语言处理的核心任务之一

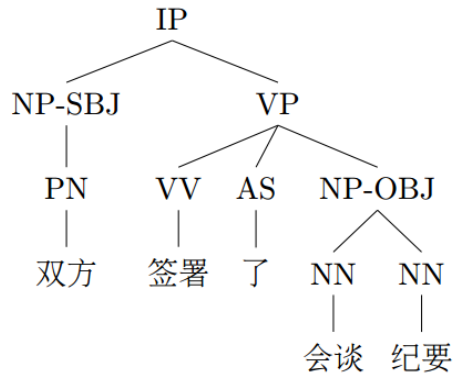
ICALL

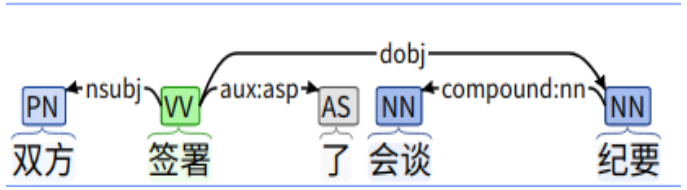
常用树库类型

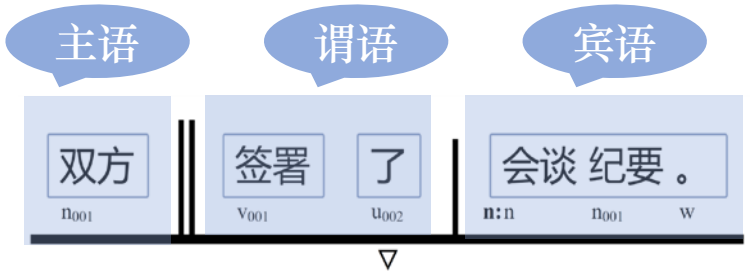
短语树库：宾州中文树库、清华短语树库等

依存树库：哈工大依存树库、通用依存树库等

句式结构树库：北师大句本位树库

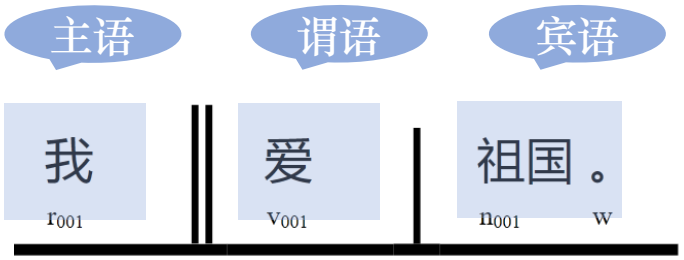






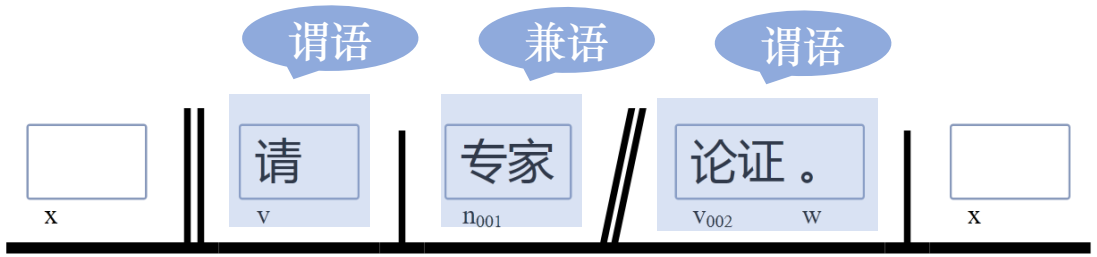
ICALL

句式结构树库



句本位语法

主干成分



句式

ICALL

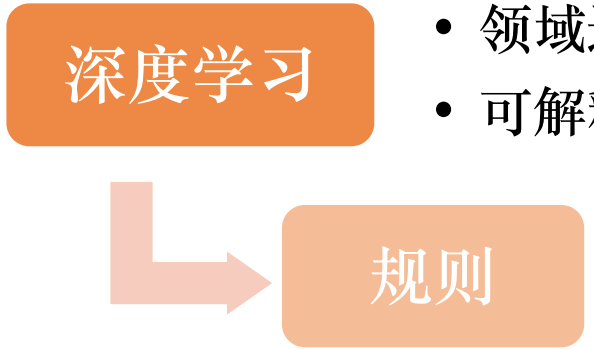
问题的提出

■句式结构树库

- 语料主要来源于国际汉语教材、中小学教材与文学作品，其他领域数据量少

■树库自动转换

- 短语结构 ⇌ 依存结构
  - Lin(1995)；Xia(2001)等
- 短语结构⟶句式结构
  - 张引兵等人（2018）



- 领域迁移问题
- 可解释性差

ICALL

问题的提出

■句式结构树库

- 语料主要来源于国际汉语教材、中小学教材与文学作品，其他领域数据量少

■树库自动转换

- 短语结构 ⇌ 依存结构
  - Lin(1995)；Xia(2001)
- 短语结构⟶句式结构
  - 张引兵等人（2018），**转换规则未开源**

清华短语树库

- 结构主义语言学的层次分析法
- 数据量较少

宾州中文树库

- 转换生成语法理论，层次更深
- 句法信息更丰富
- 通用性更强

ICALL

问题的提出

■句式结构树库

- 语料主要来源于国际汉语教材、中小学教材与文学作品，其他领域数据量少

■树库自动转换

- 短语结构 ⇌ 依存结构
  - Lin(1995)；Xia(2001)；党政法等人（2005）
- 短语结构⟶句式结构
  - 张引兵等人（2018）

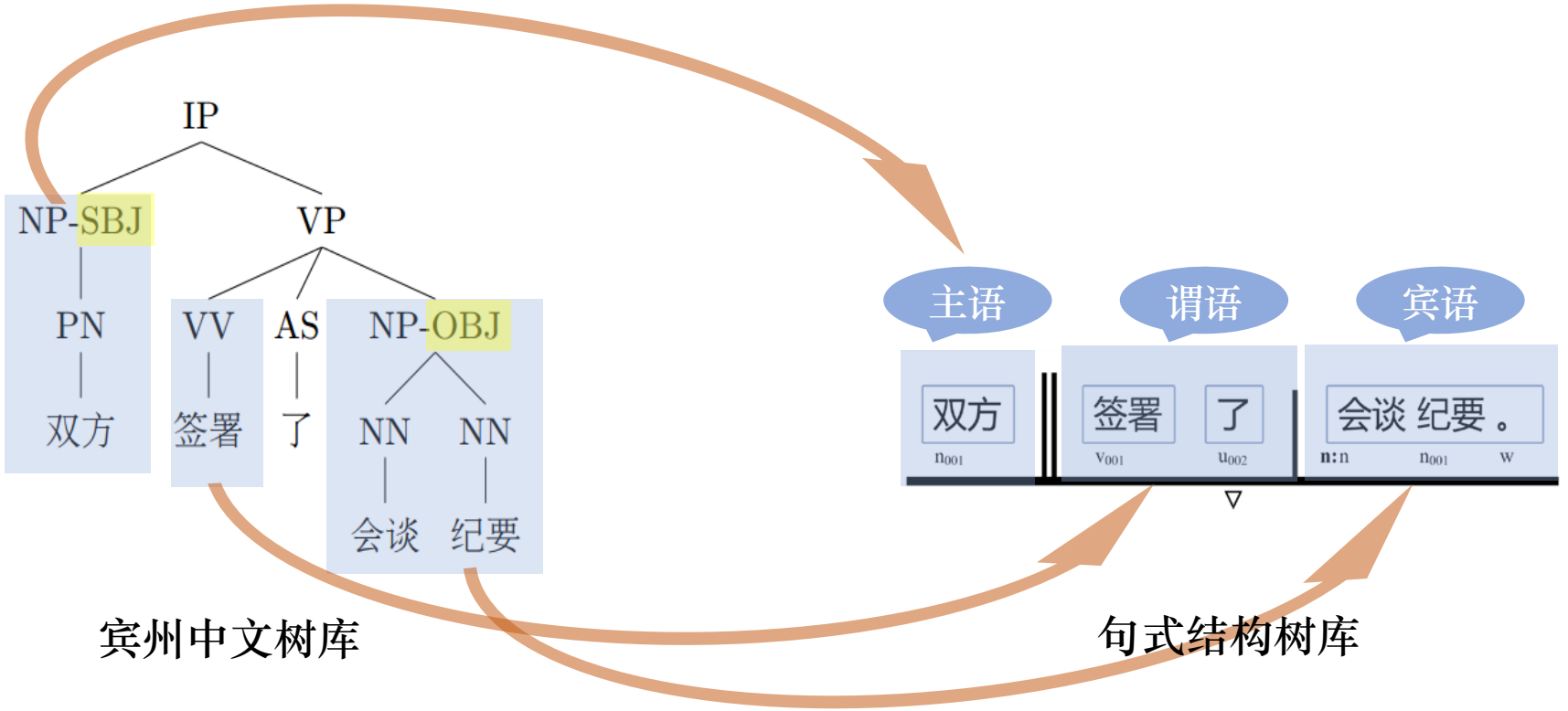
ICALL

构建句式结构树库

- **依据句式结构树库和宾州中文树库标签之间的关系，制定树库转换规则**

源树库：宾州中文树库 → 目标树库：句式结构树库

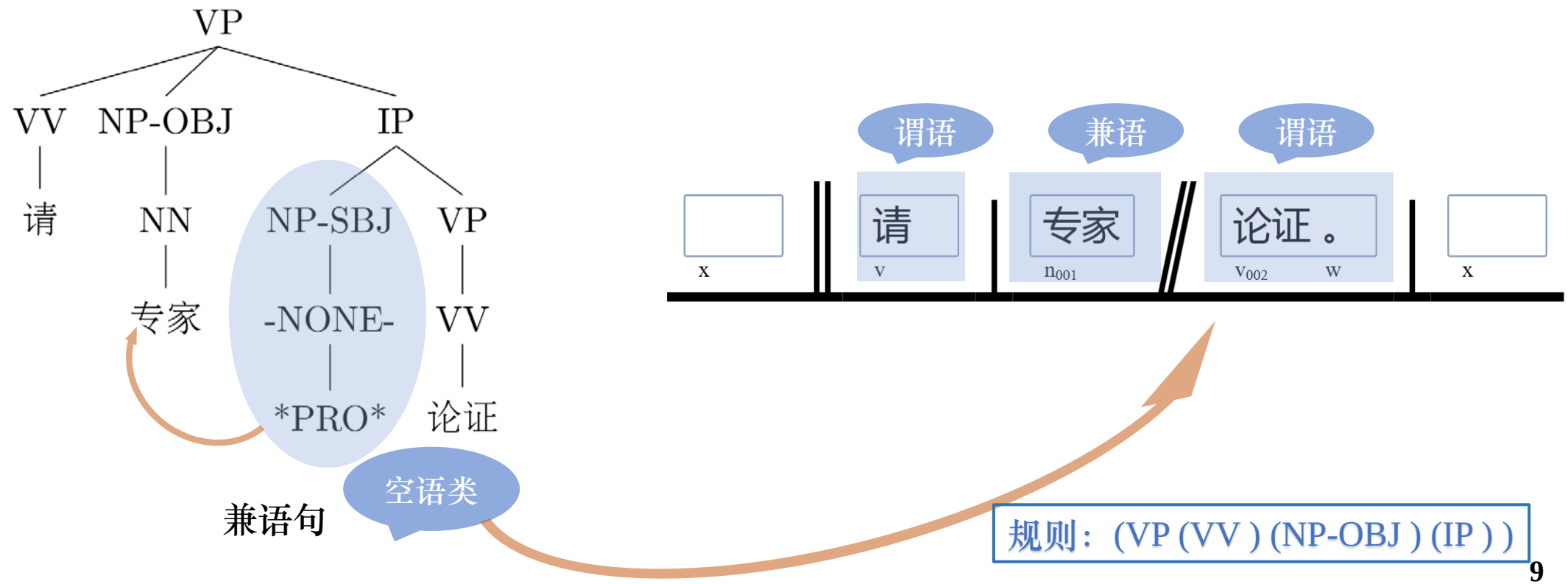
SBJ：主语
OBJ：宾语



构建句式结构树库

■句式转换规则

- 将宾州中文树库的语言单位转换成句式结构树库的句式



构建句式结构树库

■词性转换规则

- 根据句式结构树库和CTB的词性标签体系的对应关系直接转换

句式结构树库		宾州中文树库		句式结构树库		宾州中文树库	
标记	词性	标记	词性	标记	词性	标记	词性
n	名词	NN	普名	v	动词	VV	普通动词
		NR	专名			VE	“有”
		FW	外来词			VC	系动词
		URL	网页链接	a	形容词	VA	表语形容词
		NN-SHORT	略缩普名			JJ	区别词/紧缩形容词
		NR-SHORT	略缩专名				

ICALL

实验设置

实验方法

- 训练句式结构自动句法分析器自动生成树库
  - 借鉴Kitaev(2018)的基于自注意力机制的神经网络模型
- 将短语结构自动句法分析与转换算法相结合
  - 借鉴Kitaev等人(2018)(2019) 提出的基于自注意力机制的神经网络方法
  - 训练得到可以分析功能标签的短语句法分析器
- 树库自动转换规则

数据集划分

数据集	训练集	开发集	测试集
北师大句式结构树库	67,558	1,352	-
宾州中文树库 (CTB 5)	17,554	352	318

人工标注

- 标注员为语言学专业同学
- 语料来源于 CTB 5 测试集共318句

ICALL

实验结果

短语结构自动句法分析+转换规则

- 优于句式结构自动句法分析的方法，进一步说明转换规则的有效性
- 该方法不依赖于人工标注的宾州中文树库作为源树库来构建句式结构树库，具有更强的通用性

方法	F_1值				准确率	召回率	F_1值
	小句	成分	虚词位	句式			
句式结构自动句法分析	95.21	80.61	88.74	**87.44**	83.85	85.01	84.43
短语结构自动句法分析 + 转换规则	95.09	86.01	94.33	74.03	88.40	86.73	87.56
宾州中文树库 + 转换规则	**95.70**	**88.43**	**95.72**	78.87	**90.68**	**88.79**	**89.72**

ICALL

实验结果

宾州中文树库结合转换规则

- 整体效果最优，说明基于规则的转换算法在树库自动构建上具有一定优势
- 规则对同位、并列、联合谓语结构的处理存在不足

方法	F_1值				准确率	召回率	F_1值
	小句	成分	虚词位	句式			
句式结构自动句法分析	95.21	80.61	88.74	**87.44**	83.85	85.01	84.43
短语结构自动句法分析 + 转换规则	95.09	86.01	94.33	74.03	88.40	86.73	87.56
宾州中文树库 + 转换规则	**95.70**	**88.43**	**95.72**	78.87	**90.68**	**88.79**	**89.72**

ICALL

实验结果

基于人工标注的自动转换结果评估

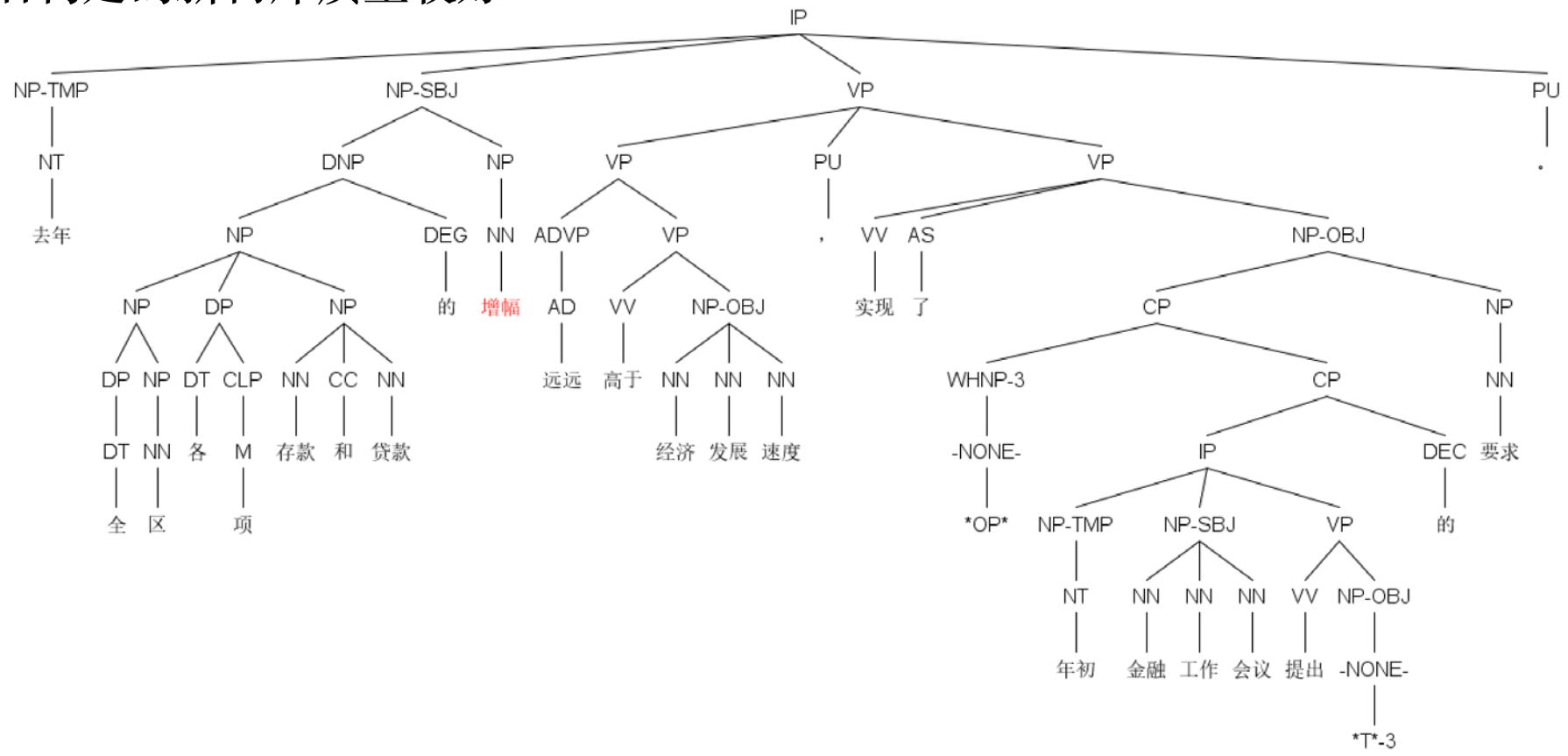
大类	小类	P	R	F1	大类	小类	P	R	F1
句子	小句	93.81	93.37	93.59	虚词位	助词位（定状补）	93.43	100.00	96.60
成分	主语	91.67	92.17	91.92		助词位（附NP）	100.00	66.67	80.00
	谓语	92.30	91.33	91.81		助词位（附VP）	94.07	87.40	90.61
	宾语	91.75	91.40	91.57	句式	并列	96.89	63.39	76.64
	定语	74.84	72.18	73.49		同位	69.23	79.41	73.97
	状语	93.11	93.23	93.17		合成谓语	95.29	77.14	85.26
	补语	85.71	85.71	85.71		联合谓语	72.46	72.46	72.46
虚词位	介词位	95.87	99.69	97.74		兼语	97.22	92.11	94.59
	连词位	93.10	75.00	83.08		连动	90.00	90.00	90.00
	方位词位	99.08	99.08	99.08					
整体精确率					90.68				
整体召回率					88.79				
整体F1					89.72				

ICALL

实验结果分析

基于人工标注的自动转换结果评估

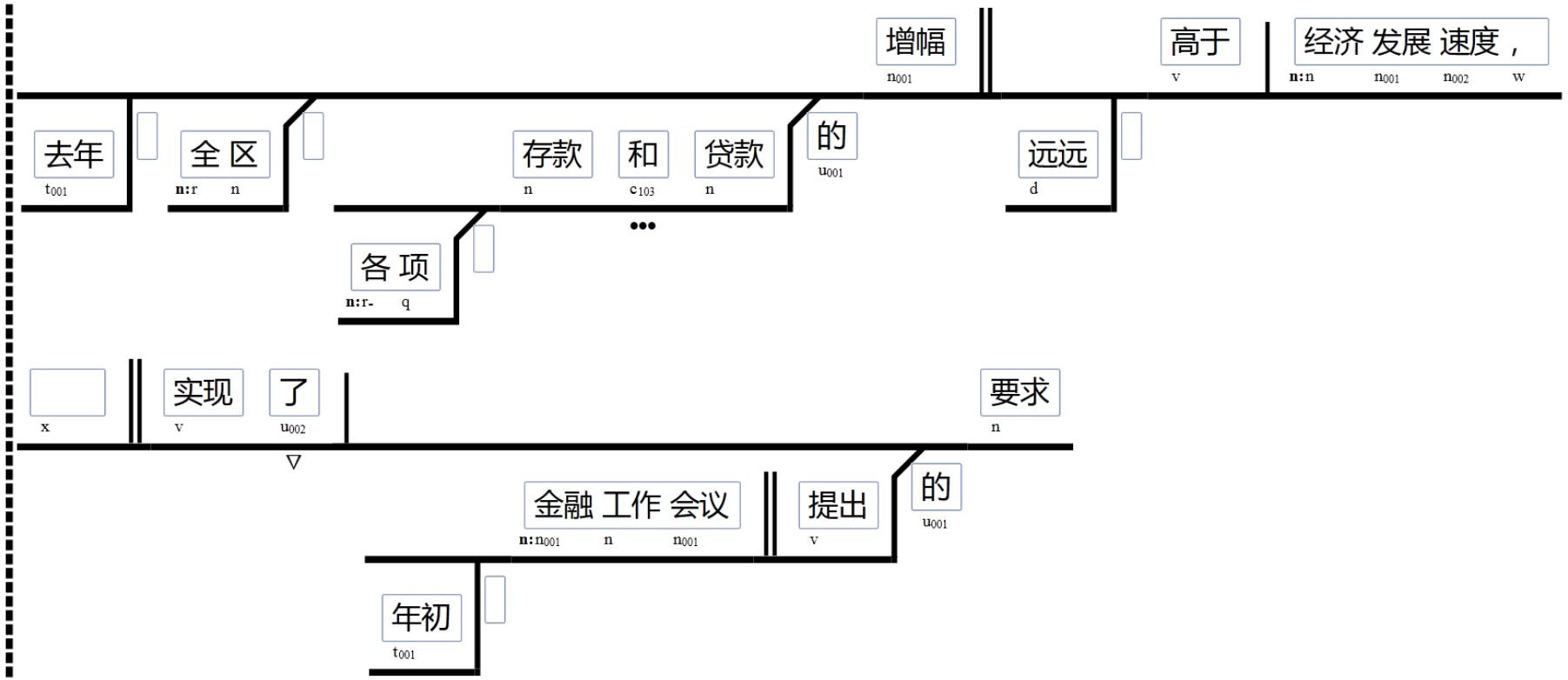
- 转换规则对于小句切分、主谓宾定状补的处理结果较好，这两步对于句子分析非常重要，说明最后构建的新树库质量较好



实验结果分析

基于人工标注的自动转换结果评估

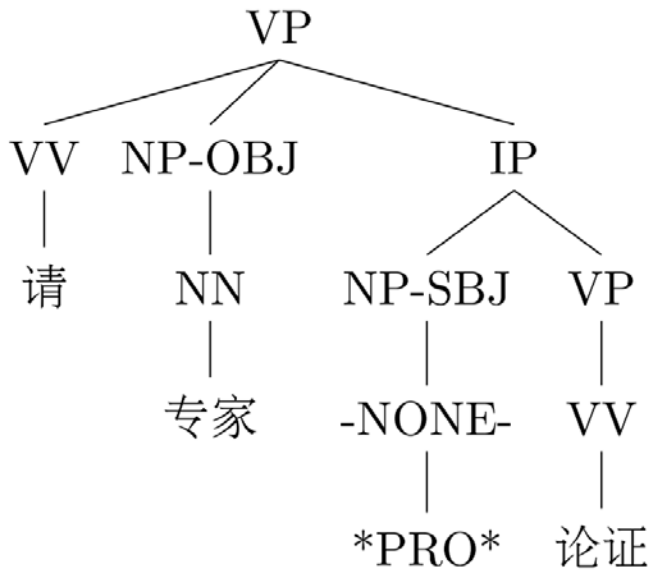
- 转换规则对于小句切分、主谓宾定状补的处理结果较好，这两步对于句子分析非常重要，说明最后构建的新树库质量较好

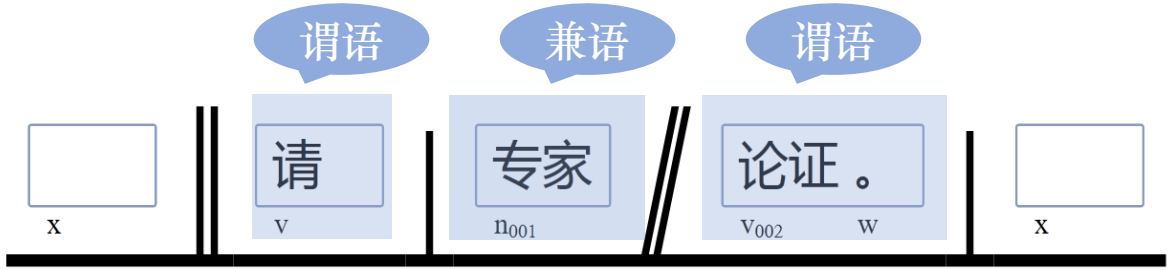


实验结果分析

基于人工标注的自动转换结果评估

- 兼语结构、连动结构在句式的转换结果中效果最好
- 两者是根据特定的CTB的短语树结构进行转换的，说明树结构能基本对应相应的句式结构





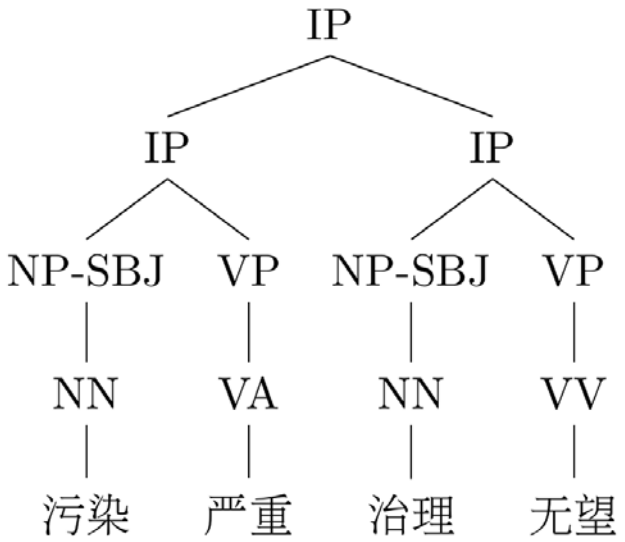
兼语句： (VP (VV) (NP-OBJ) (IP))

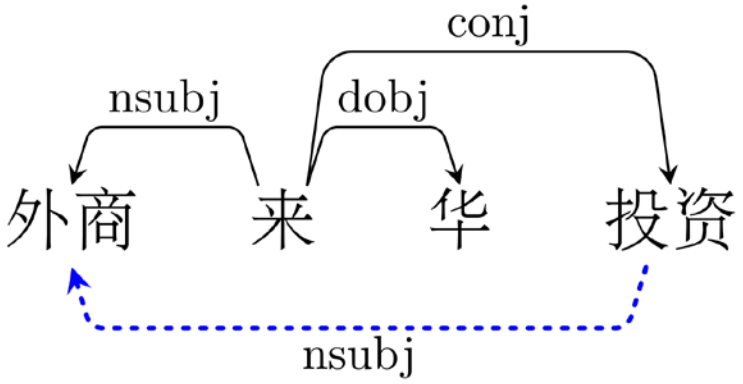
句式结构树库呈现的兼语句

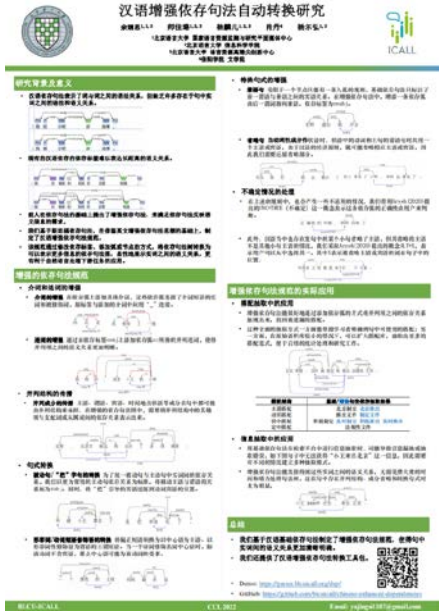
实验结果分析

基于人工标注的自动转换结果评估

- 两个谓词性成分之间的关系是“连动”还是“联合”？
- 如 “污染严重治理无望”为联合谓语结构，但处理成了连动结构







通过语义逻辑表示的联合谓语结构

汉语增强依存句法：连动句的增强

总结

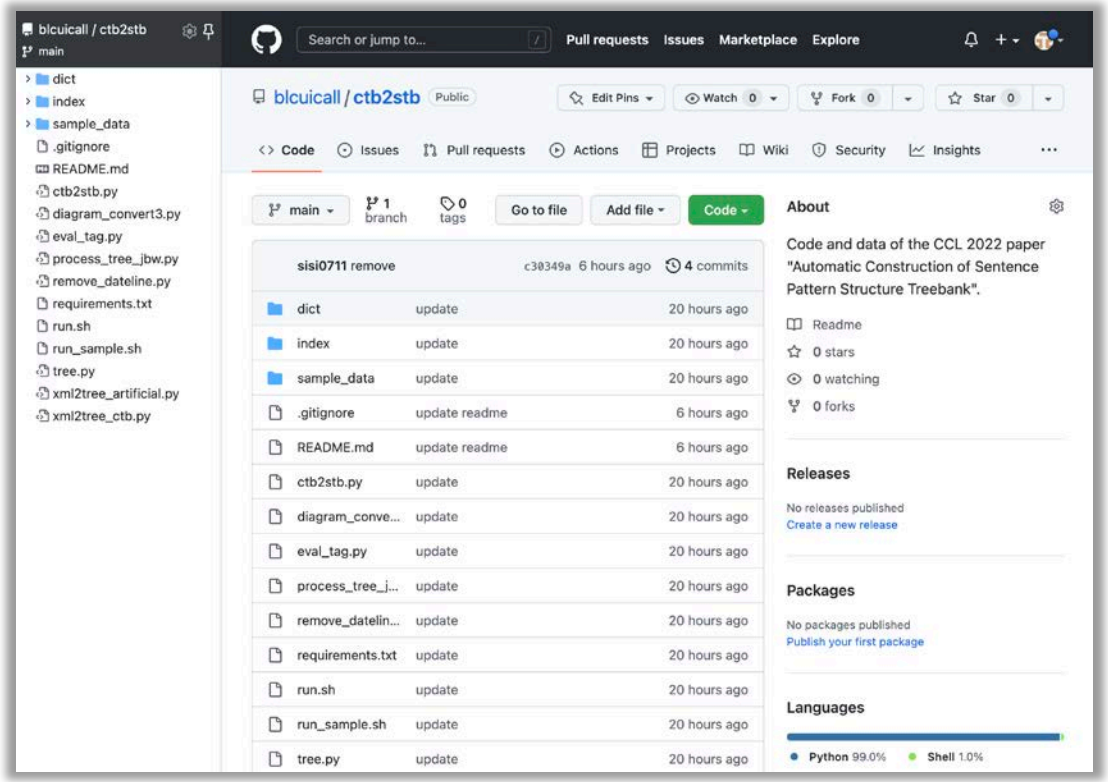
- **我们的工作**
  - 提出短语树库向句式结构树库的自动转换规则
  - 扩充了句式结构树库的数据
  - 丰富了宾州中文树库的句法标注体系
- **展望**
  - 设计句式结构向短语/依存结构的转换规则
  - 探索句式结构树库及句法分析器在**语言教学、辅助写作、教材编写**等方面的应用

ICALL

转换规则代码开源

- 简介
  - 转换规则包括小句切分、句法成分转换规则、句式转换规则、词性转换规则等
  - 支持宾州中文树库 CTB 各个版本
- 地址
  - https://github.com/blcuicall/ctb2stb

开源代码 →

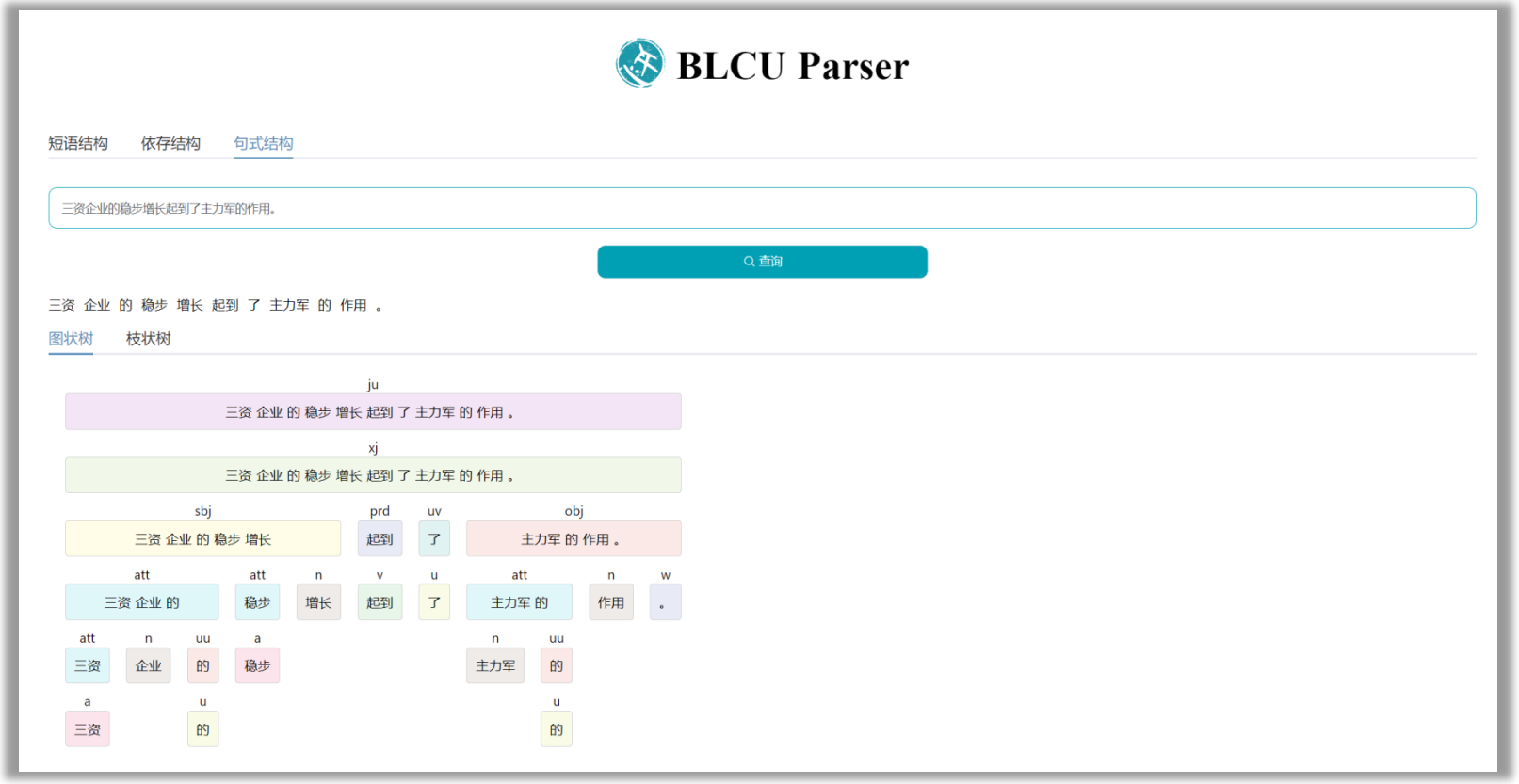


ICALL

欢迎大家使用和反馈问题！

■句法分析平台

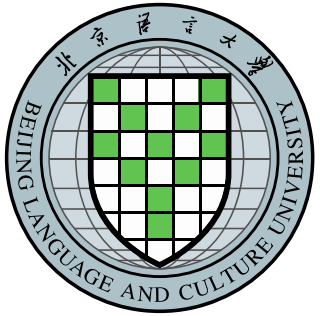
- 输入一个中文句子，获得相应的句式结构树



演示系统：
https://parser.blcuicall.org

BLCU Parser

ICALL



请大家批评指正！

ICALL
智能语言学习