# 成分句法分析

主讲人: 张栋

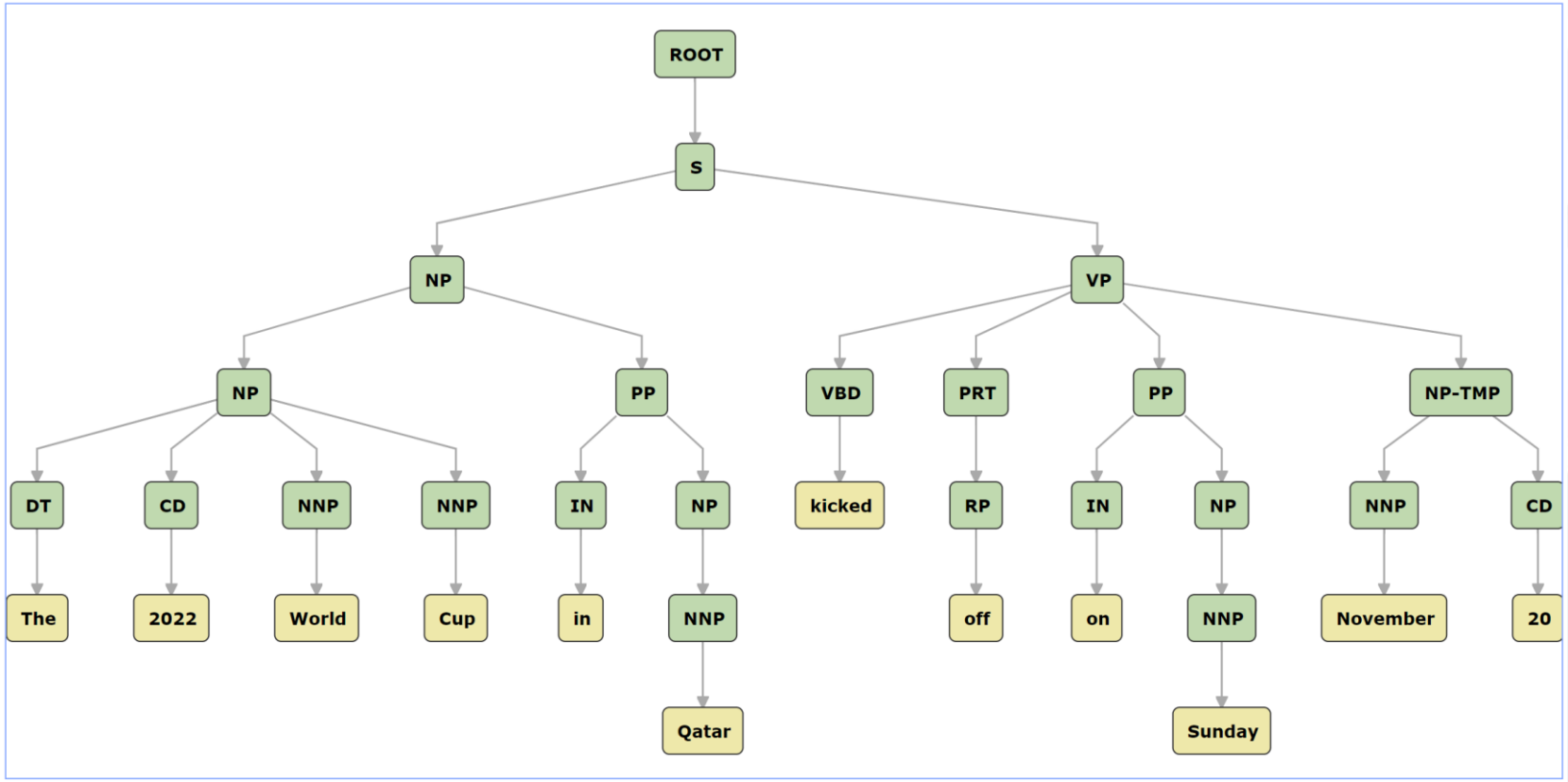
# 参考资料

- 统计自然语言处理 (第2版). 宗成庆著
- https://web.stanford.edu/~jurafsky/slp3/13.pdf
- https://web.stanford.edu/~jurafsky/slp3/14.pdf
- https://zhuanlan.zhihu.com/p/414241465
- https://zhuanlan.zhihu.com/p/51186364
- https://chmx0929.gitbook.io/machine-learning/zi-ran-yu-yan-chu-li/zi-ran-yu-yan-chu-li/ju-fa-fen-xi

- **句法分析** (Syntactic Parsing)

  - 确定句子的句法结构 或 句子中词汇之间的依存关系
    - 成分句法分析 (constituency parsing)
    - 依存句法分析 (dependency parsing)
  - 不是自然语言处理任务的最终目标，实现最终目标的重要环节，甚至关键环节

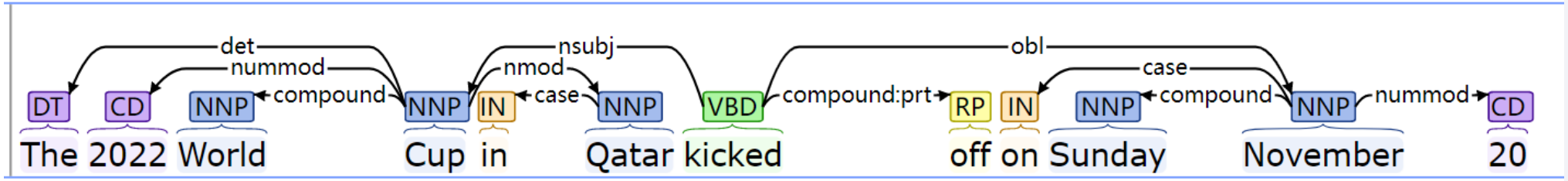
- # 成分句法分析




The 2022 World Cup in Qatar kicked off on Sunday November 20

摘自https://corenlp.run/

- # 依存句法分析




The 2022 World Cup in Qatar kicked off on Sunday November 20

摘自https://corenlp.run/

# • 成分句法分析

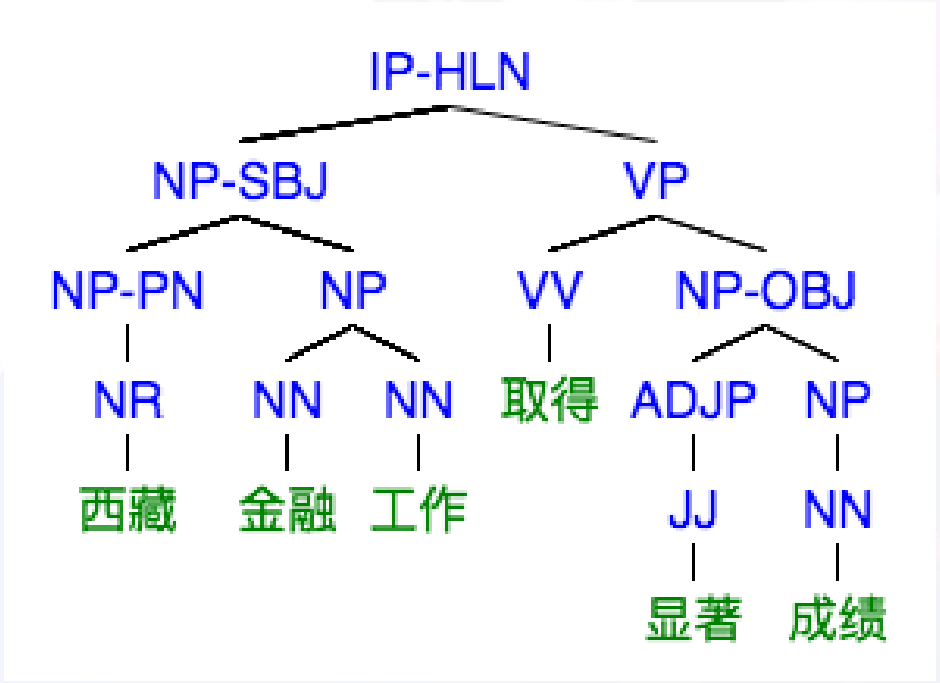
## 句法结构分析

- 基本概念
  - 指对输入的单词序列（一般为句子）判断其构成是否合乎给定的语法，分析出合乎语法的句子的成分句法结构
- 成分句法结构
  - 用树状数据结构表示，称为句法分析树 (syntactic parsing tree)，或简称分析树 (parsing tree)
- 句法分析器/分析器
  - 完成句法分析过程的程序模块

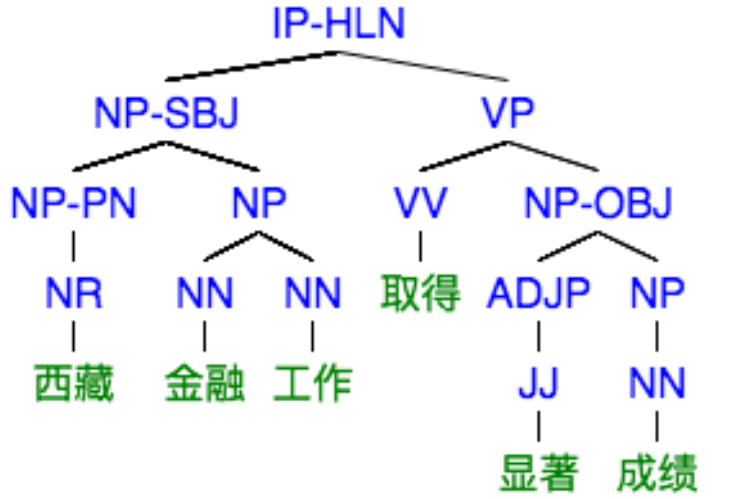
- **句法树中包含的信息**

1. 分词结果及每个单词的词性

  - NR = 专有名词
  - NN = 普通名词
  - VV = 普通动词
  - JJ = 形容词



- **句法树中包含的信息**




2. 短语

名词短语(NP): “西藏”、“西藏 金融 工作”、“显著 成绩”

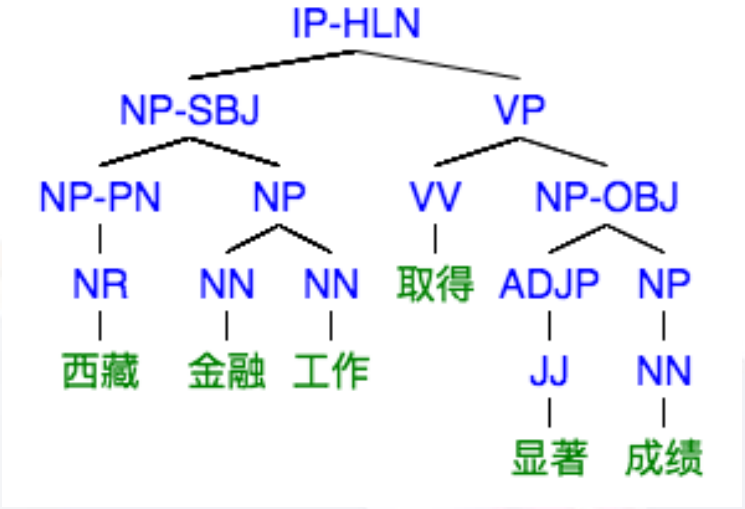
动词短语(VP): “取得 显著 成绩”

形容词短语(ADJP): “显著”

句子/子句/从句(IP): “西藏 金融 工作 取得 显著 成绩

- **句法树中包含的信息**

3. 其他有用信息




NP(西藏 金融 工作) 是 动词 VV(取得)的 主语 (Subject).

NP(显著 成绩) 是 动词 VV(取得)的 宾语 (Object).

中心词: 每个短语都有一个中心词, 如 NP(西藏 金融 工作)、VP(取得 显著 成绩)

- **成分句法分析两个难点**

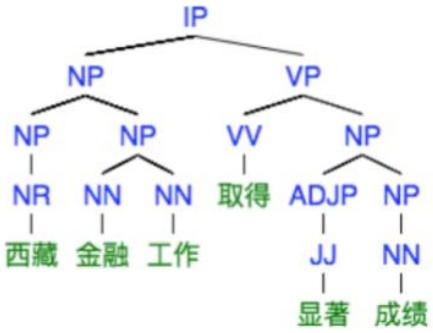
- 歧义:
- 搜索空间

# • 成分句法分析评测

- PARSEVAL评测体系:
  - 准确率: 分析正确的短语个数，占**分析结果**中所有短语个数的比例
  - 召回率: 分析正确短语个数，占**标准分析树**全部短语个数的比例
  - 交叉括号数: 分析得到的某一个短语的**覆盖范围**与**标准分析树**的某个短语的覆盖范围存在**重叠又不包 含**关系，即构成一个交叉括号

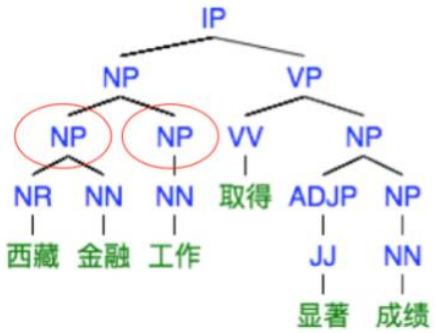
分析正确: 短语类型一致 && 短语的跨度一致

*Evalb: This is a bracket scoring program. It reports precision, recall, F-measure, non crossing and tagging accuracy for given data. http://nlp.cs.nyu.edu/evalb/*




正确句法树

| 类型 | 跨度 |
|---|---|
| IP | [1 6] |
| NP | [1 3] |
| NP | [1 1] |
| NP | [2 3] |
| VP | [4 6] |
| NP | [5 6] |
| ADJP | [5 5] |
| NP | [6 6] |




自动句法树

| 类型 | 跨度 |
|---|---|
| IP | [1 6] |
| NP | [1 3] |
| NP | [1 2] |
| NP | [3 3] |
| VP | [4 6] |
| NP | [5 6] |
| ADJP | [5 5] |
| NP | [6 6] |

Precision: 6/8 = 0.75

Recall: 6/8 = 0.75

F1: 6/8 = 0.75

注: 词性结点不参与计算

# • 成分句法分析语料

- 英文
  - Penn Treebank (PTB)
- 中文
  - Penn Chinese Treebank (CTB)
  - **清华树库** (Tsinghua Chinese Treebank)
  - **台湾中研院树库** (Sinica Treebank)

# 谢谢