NN Methods for NLP

Chapter 6-7 Sharing

Textual Features and Related Cases

Reporter: Liu Yang

神经自然语言处理

6-7章 分享

文本特征和相关任务

主讲人：刘洋

NLP数据拓扑结构

- 研究对象，目标的格式。

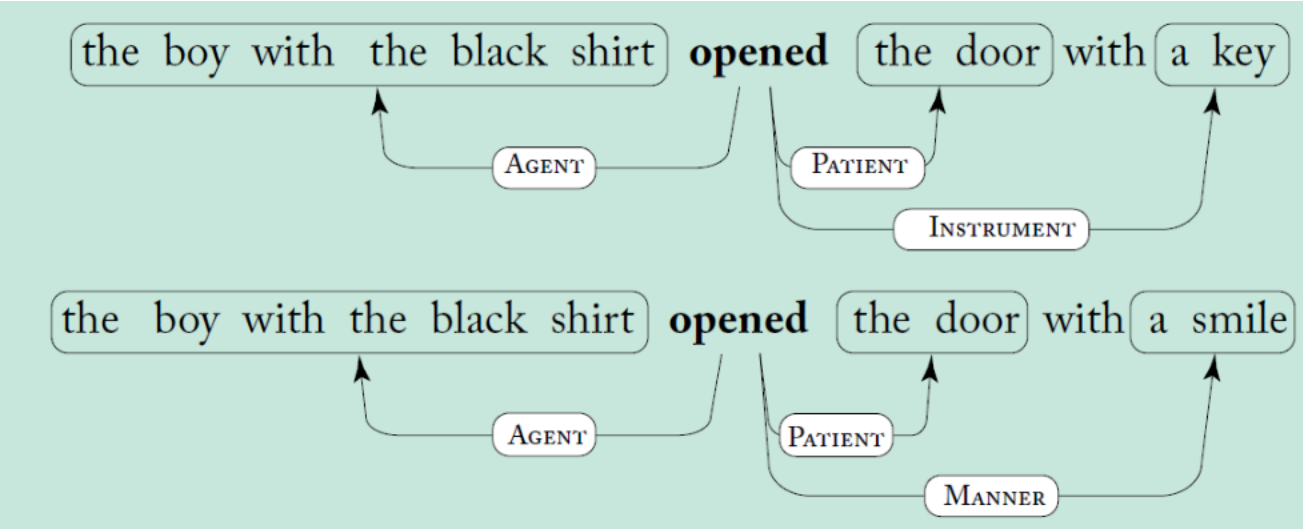
- 词（误拼）
- 文本（文本分类、情感倾向、主题分类）
- 成对文本（翻译、含义推断）
- 上下文中的词（谓词识别，动词匹配，词性标注）
- 词之间的关系（主语检测，句法，语义角色标注）

NLP特征-直接可测特征

- 依赖数据格式与关注点，语言学概念。
- 特征的可计数性，词袋特征。
- 直接可测特征
  - 单独词特征 （词，长度，字符特征，前后缀）
  - 词元lemma和词干*stem*（book-books、 pictur-picture/pictures/pictured）
  - 词汇资源 （阴阳性、词型、格、体、数）
  - 分布信息
  - 文本特征（词袋、权重）
  - 上下文词特征（相对窗口、绝对位置）

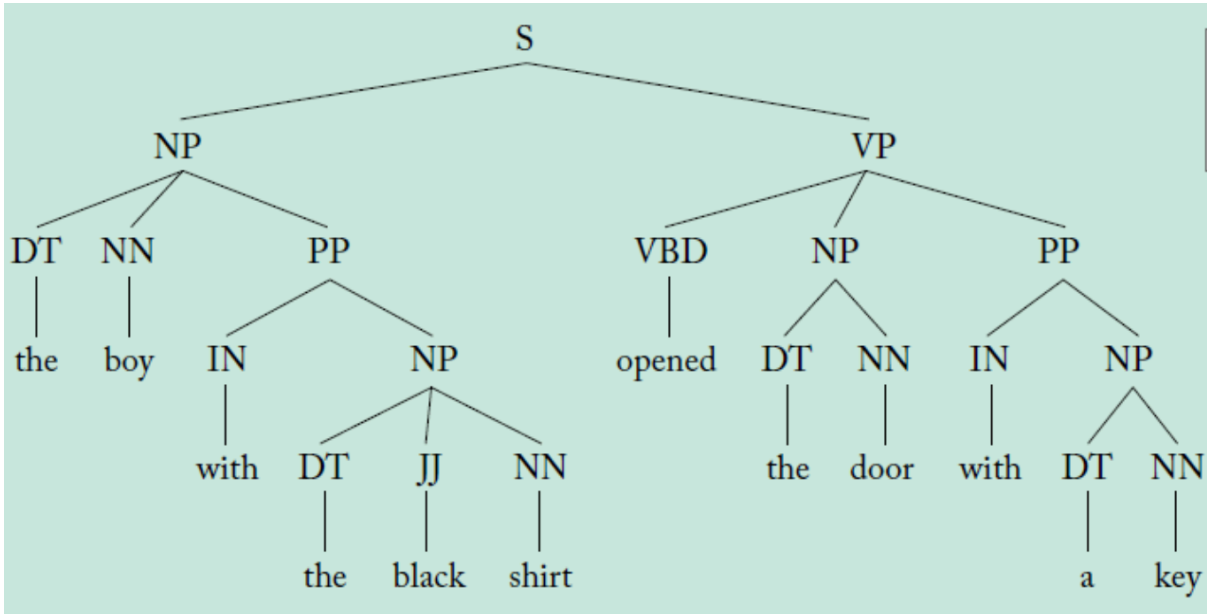
NLP特征-可推断的语言学特征

- 可推断的语言学特征
  - 词性
    - [NP the boy] [PP with] [NP the black shirt] [VP opened] [NP the door] [PP with] [NP a key]
  - 短语结构树与依存句法树（下一页）（句法关系，句法路径，深度，公共父节点）
  - 语义信息：论元角色信息

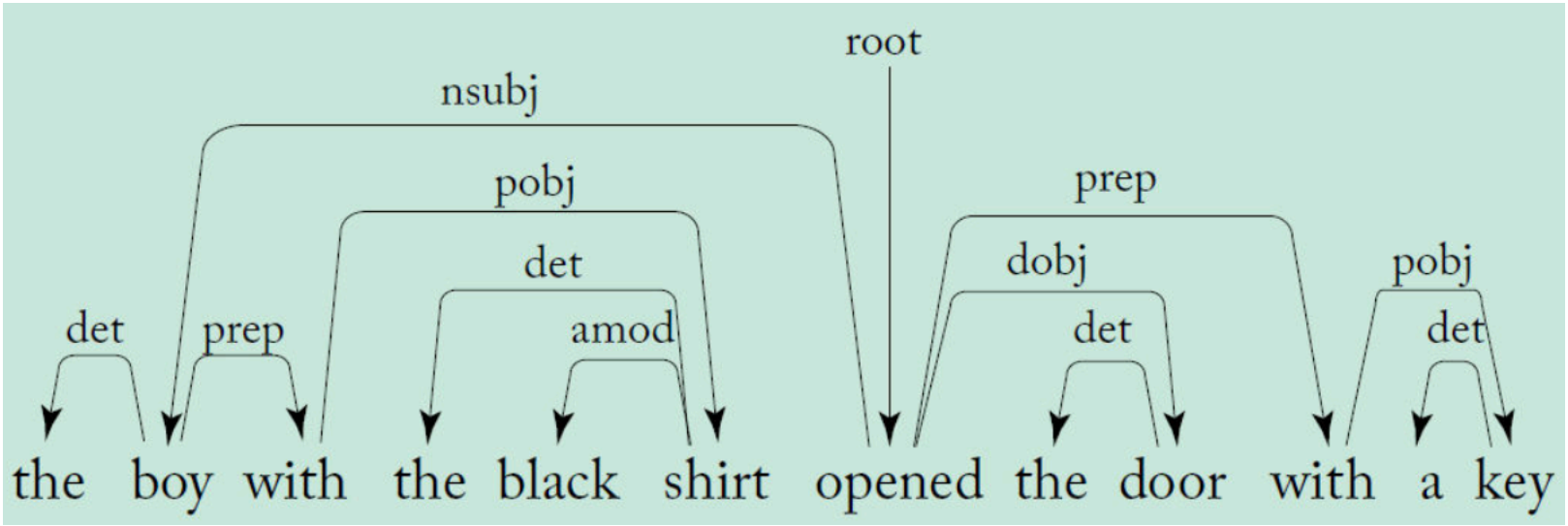


NLP特征

- 短语结构树
- 依存句法树



- 特征：（句法关系，句法路径，深度，公共父节点）

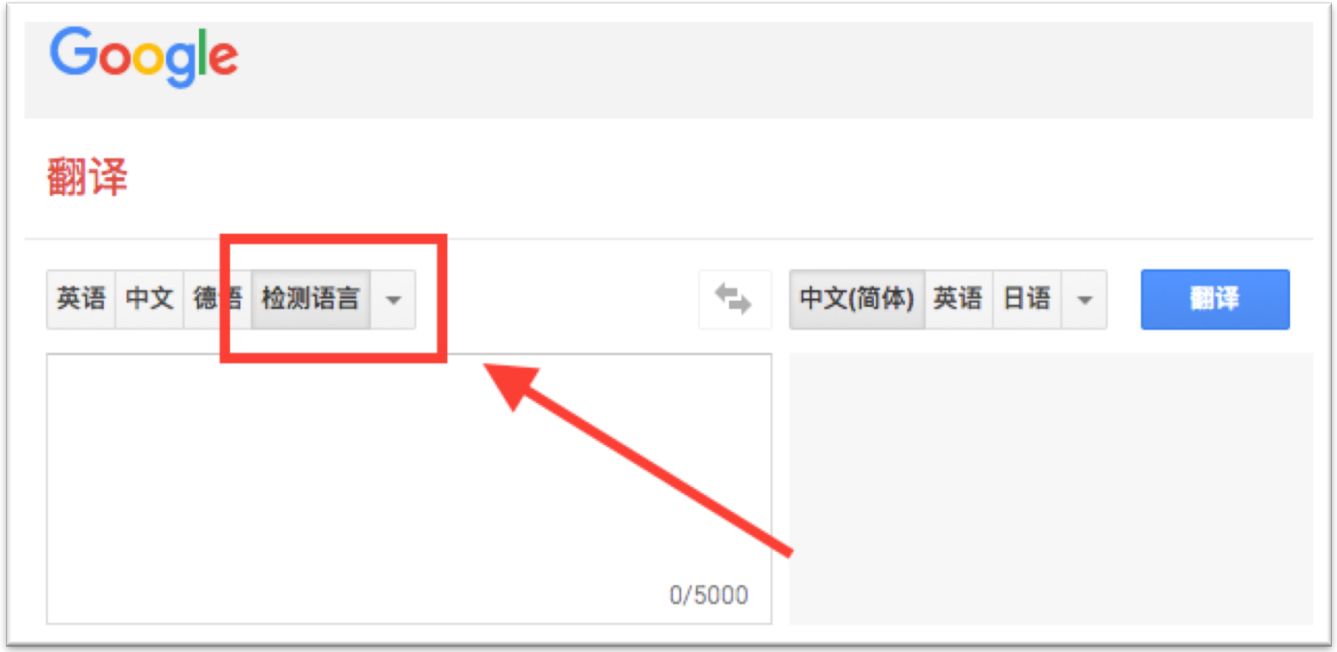


NLP特征-其他

- 核心特征 **vs 组合**特征
  - 增加了设计者先验知识。
  - 在神经网络之中可以在一定程度上学习特征组合。
- N-gram特征
  - 元词能够比单独的词富含更多信息（New York，not good，和 Paris Hilton ）
  - 学到多元特征较难。尤其是在bag-of-word输入情况下。
- 分布特征
  - 刻画词相似性
  - 词向量表示和word Embedding（10,11章）

NLP特征案例-文档分类

- 语言识别
  - 较强特征：二元字母词袋特征



NLP特征案例-文档分类

- 主题分类
  - 给定的文档，需要将它分类为一预定的主题 (如 政治 体育 休闲 八卦 生活方式等)
  - 较强特征：一、二元文法词袋特征
  - 训练样本少：lemma或词向量等分布特征
  - 给词袋加权：安装信息量加权如tf-idf
- 作者归属
  - 任务：文本=>推测作者的身份 、属性（性别、年龄、母语）
  - 线索微妙，涉及文体属性而不是内容
  - 较强特征：词性一二三四元（pos），功能词（on of the）一二元、距离

NLP特征案例-上下文中的单词

- 词性标注（POS-tag）
  - 句子=> 每个单词分配正确的词性

他	叫	汤姆	去	拿	外衣	。
r	v	nh	v	v	n	wp

  - 较强特征：
    - 词本身：前后缀、词元、大词表频率
    - 外部：周围单词及其前后缀、前面单词的预测结果
  - OOV，词表之外的词：
    - 依靠一些词本身前后缀信息作为补偿
    - 依靠词外部特征
  - 特征模板

- 单词=X
- 2字母后缀=X
- 3字母后缀=X
- 2字母前缀=X
- 3字母前缀=X
- 单词是否大写
- 单词是否包含连字符
- 单词是否包含数字
- P 值[-2，-1，+1，+2]:
  - 位于位置 P 的单词=X
  - 位于位置 P 的单词的2字母后缀=X
  - 位于位置 P 的单词的3字母后缀=X
  - 位于位置 P 的单词的2字母前缀=X
  - 位于位置 P 的单词的3字母前缀=X
  - 位于位置 P 的单词是否大写
  - 位于位置 P 的单词是否包含连字符
  - 位于位置 P 的单词是否包含数字
- 位于位置-1 的单词的词性=X
- 位于位置-2 的单词的词性=X

NLP特征案例-上下文中的单词

- 命名实体识别（NER）
  - 句子=> 找到人名地名，机构名等。

  - 任务转化为序列标注

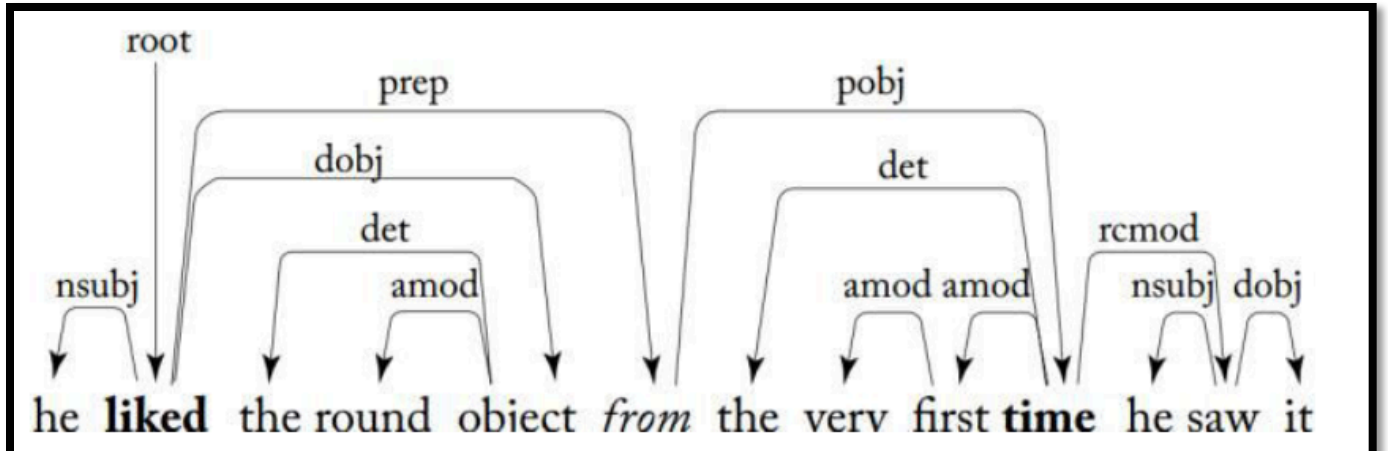
    I-PER O I-LOC O B-ORG I-ORG O O O

  - 较强特征：
    - 与pos任务基本相同
    - 共现单词，分布式特征，是否在列表中存在

标签	含义
O	不是命名实体的一部分
B-PER I-PER	人名的开始词 人名的继续词
B-LOC I-LOC	地名的开始词 地名的继续词
B-ORG I-ORG	机构的开始词 机构的继续词
B-MISC I-MISC	其他类别命名实体的开始词 其他类别命名实体的继续词

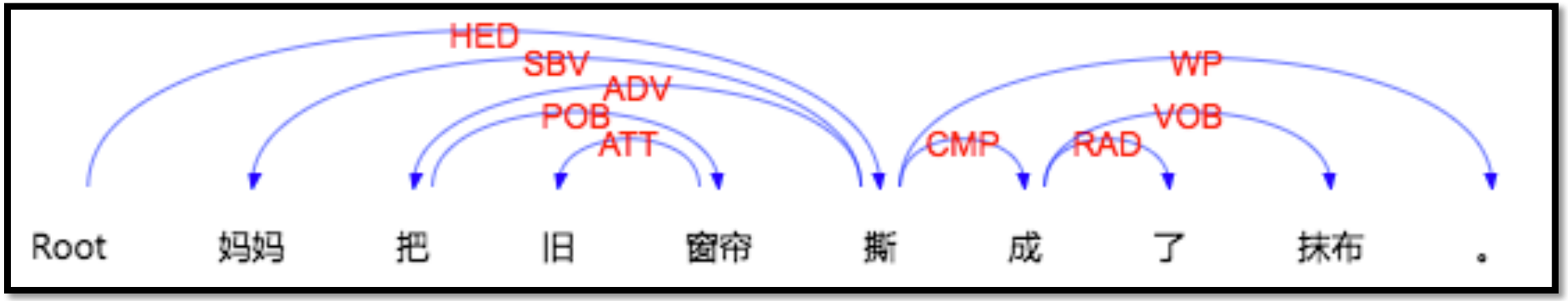
NLP特征案例-上下文中的单词语言特征

- 介词词义消歧
  - 例子：
    - We went there **for** lunch. 目的
    - He paid **for** me. 受益者
    - We ate **for** two hours. 时间
    - He would have left **for** home, but it started raining. 地点
  - 如何选择较大的上下文窗口
    - 左边第一个谓词、右边第一个名词
    - 依存分析器信息



NLP特征案例-上下文中的单词关系

- 弧分解分析，句法分析
- 句子=> 句法依赖关系树



- 方法：
  - 搜索总体分数最大的树 ∑arcscore(父节点头词,子节点修饰词,句子)
    - 利用头词、修饰词及其pos，窗口内单词及词性
  - 基于转移的方法

Thanks and F&Q.

主讲人: 刘洋

WordNet, FrameNet 和 PPDB