人工智能之NLP

NLP基础二

主讲人：GerryLiu

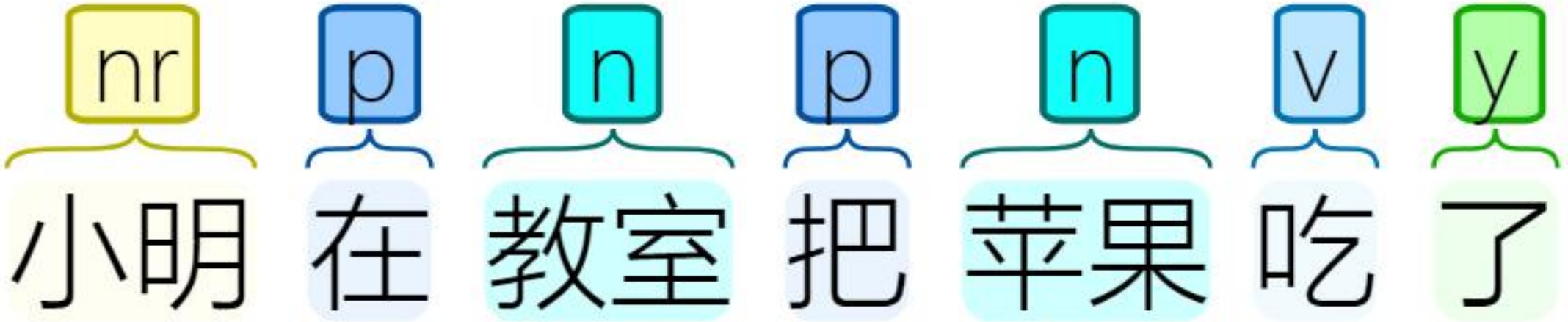
课程要求

- 课上课下“九字”真言
  - 认真听，**善摘录，勤思考**
  - **多温故，乐实践**，再发散
- 四不原则
  - 不懒散惰性，不迟到早退
  - 不请假旷课，不拖延作业
- 一点注意事项
  - 违反“四不原则”，不推荐就业

课程内容

- 词性标注
- 命名实体识别
- 句法分析
- 语法分析
- 关系抽取

NLP基础_词性标注



NLP基础_词性标注

- 词性是词汇的基本语法属性。词性标注是在给定句子中判定每个词的语法范畴，确定其词性并加以标注的过程。
- 在中文词汇中，一个词语一般只有1到2个词性，并且其中一个词性的使用频率会远远大于另一个，所以词性标注最简单的方式是从语料库中统计每个词对应的高频词性，作为默认词性。
- 当前主流的词性标注手段和分词一样，是将句子的词性标注当做一个序列标注问题来解决，比如：HMM、CRF等。
- 词性说明：
  - http://www.hankcs.com/nlp/part-of-speech-tagging.html#h2-8
  - https://github.com/hankcs/HanLP/blob/master/data/dictionary/other/TagPKU98.csv

NLP基础_命名实体识别

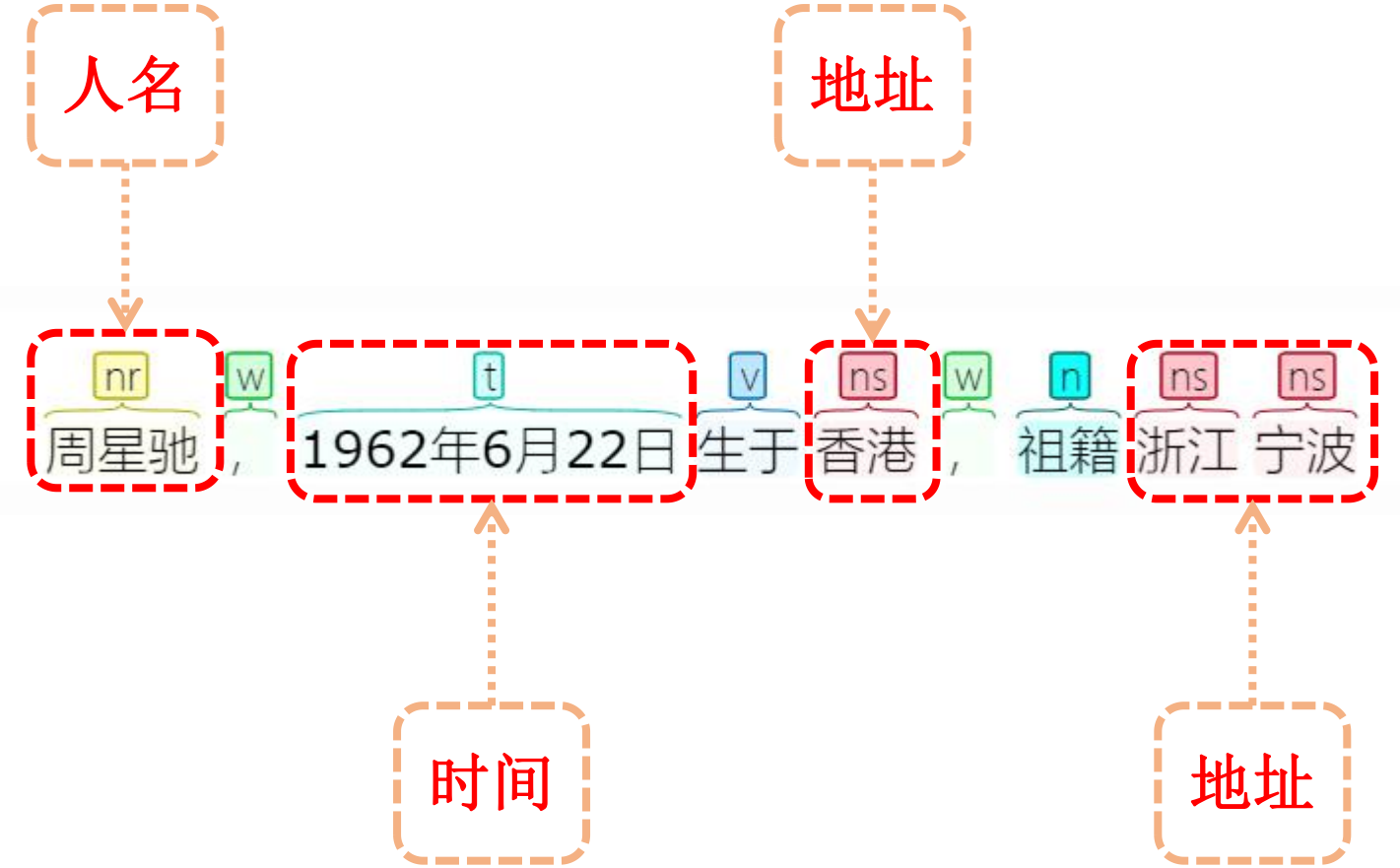
- 命名实体识别(**Named Entity Recognition，NER**)，是指识别文本中具有特定意义的实体，主要包括人名、地名、机构名、专有名词、时间、货币等信息的提取识别。通常包括两部分：
  - 实体边界识别
  - 确定实体类别（人名、地名、机构名等）
- 常用的实现方式有：
  - 基于规则的命名实体识别
  - 基于序列标注统计的命名识别识别：CRF、HMM、LSTM等

NLP基础_命名实体识别

NLP基础_命名实体识别

- NER标注方法：
  - BIO标注法： I(Inside)、O(Outside)、B(Begin)
    - I-xxx：在xxx类命名实体的内部(外开始外的所有位置)；
    - O：不属于实体；
    - B-xxx：是xxx类命名实体的开始；
  - BIOES标注法： B(Begin)、I(Inside)、O(Outside)、E(End)、S(Single)
    - B-xxx：是xxx类命名实体的开始；
    - I-xxx：在xxx类命名实体的内部；
    - O：不属于实体；
    - E-xxx：在xxx类命名实体的结尾；
    - S-xxx：单独属于xxx类命名实体。

NLP基础_关系抽取

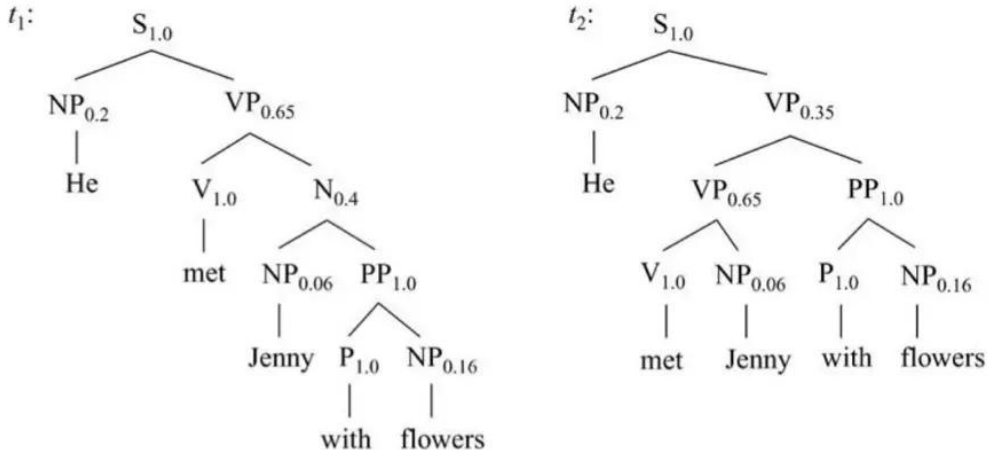


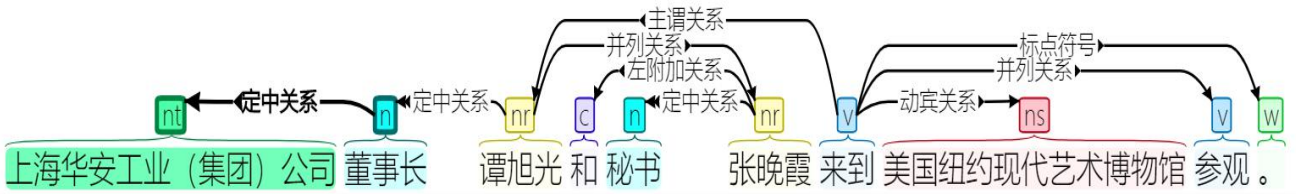
NLP基础_关系抽取

- 关系抽取是命名实体识别之后的具体应用，其应用主要分为两个方向：
  - 关系抽取：从一个句子中判断两个entity是否有关系，一般是一个二分类问题，指定某种关系。
  - 关系分类：一般是判断一个句子中两个entity是哪种关系，属于多分类问题。
  - NOTE: 一般情况下，我们所说的关系抽取实际上就是关系分类。
- 案例：
  - 文本: 周星驰，1962年6月22日生于香港，祖籍浙江宁波
  - 关系: 周星驰 --> 出生日期 --> 1962年6月22日、周星驰 --> 出生地 --> 香港、周星驰 --> 祖籍 --> 浙江宁波

NLP基础_句法分析

- 句法分析将句子分析成一颗依存句法树，描述出各个词语之间的依存关系。依存句法分析中句子的**核心**是**谓语动词**，围绕谓语找出其他成分词。
  - PCFG(Probabilistic Context Free Grammar)
  - 基于CRF的句法分析
  - 基于移进-归约的句法分析模型
- 参考：
  - http://hanlp.com/



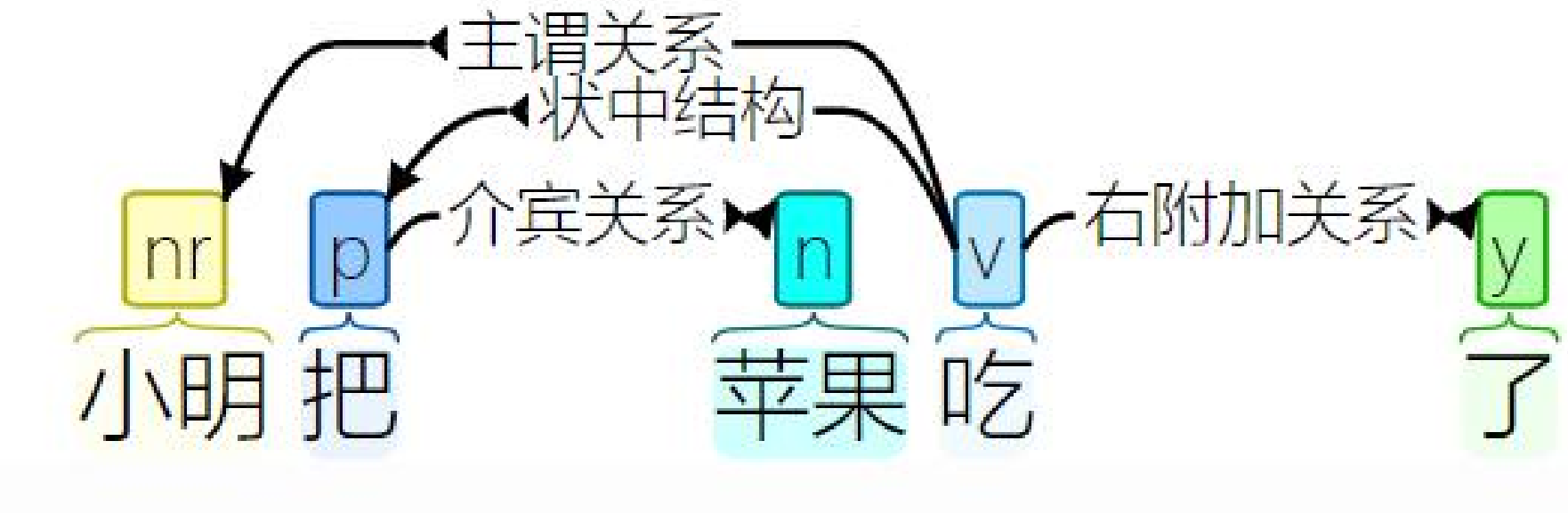


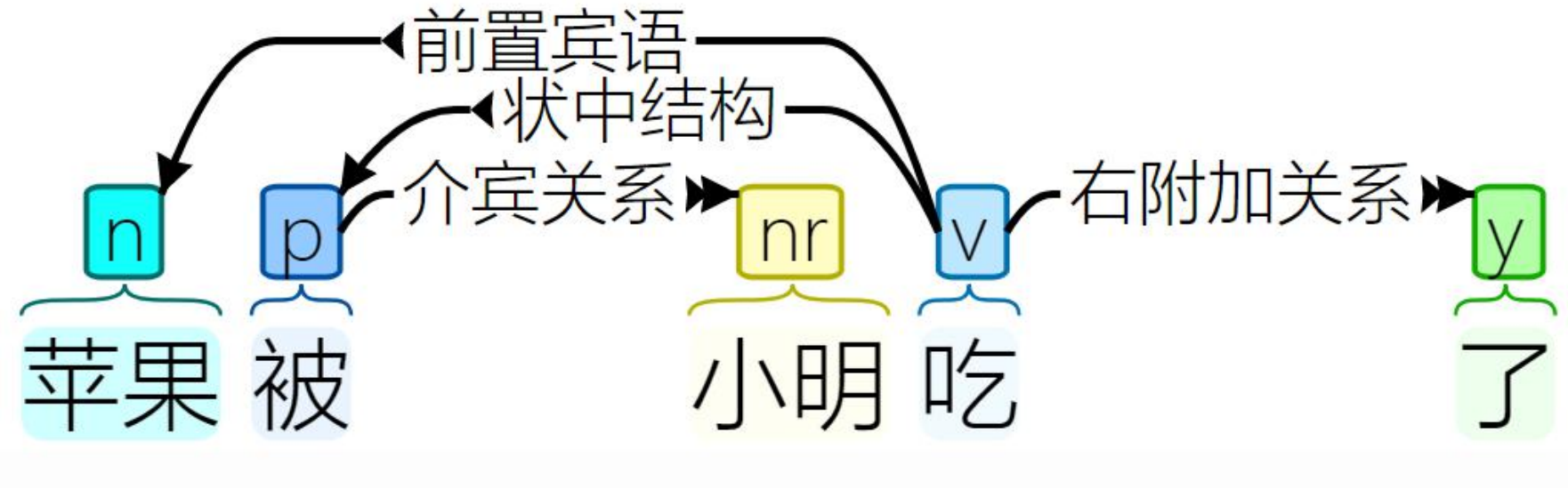
NLP基础_句法分析

Tag	关系	Description	Example
SBV	主谓关系	subject-verb	我送她一束花 (我 <-- 送)
VOB	动宾关系	直接宾语，verb-object	我送她一束花 (送 --> 花)
IOB	间宾关系	间接宾语，indirect-object	我送她一束花 (送 --> 她)
FOB	前置宾语	前置宾语，fronting-object	他什么书都读 (书 <-- 读)
DBL	兼语	double	他请我吃饭 (请 --> 我)
ATT	定中关系	attribute	红苹果 (红 <-- 苹果)
ADV	状中结构	adverbial	非常美丽 (非常 <-- 美丽)
CMP	动补结构	complement	做完了作业 (做 --> 完)
COO	并列关系	coordinate	大山和大海 (大山 --> 大海)
POB	介宾关系	preposition-object	在贸易区内 (在 --> 内)
LAD	左附加关系	left adjunct	大山和大海 (和 <-- 大海)
RAD	右附加关系	right adjunct	孩子们 (孩子 --> 们)
IS	独立结构	independent structure	两个单句在结构上彼此独立
WP	标点符号	punctuation	标点符号
HED	核心关系	head	指整个句子的核心

- http://www.hankcs.com/nlp/parsing/neural-network-based-dependency-parser.html

NLP基础_句法分析


主谓关系
状中结构
介宾关系
右附加关系
nr
p
n
v
y
小明
把
苹果
吃
了


前置宾语
状中结构
介宾关系
右附加关系
n
p
nr
v
y
苹果
被
小明
吃
了

NLP基础_语义分析

- 语义依存分析不受句法结构的影响，将具有直接语义关联的语言单元直接连接依存弧并标记上相应的语义关系。语义分析和句法分析的区别主要如下：
  - 句法依存某种程度上更重视非实词（如介词）在句子结构分析中的作用，而语义依存更倾向在具有直接语义关联的实词之间建立直接依存弧，非实词作为辅助标记存在。
  - 两者依存弧上标记的语义关系完全不同，语义依存关系是由论元关系引申归纳而来，可以用于回答问题，如我在哪里喝汤，我在用什么喝汤，谁在喝汤，我在喝什么。但是句法依存却没有这个能力。

NLP基础_语义分析

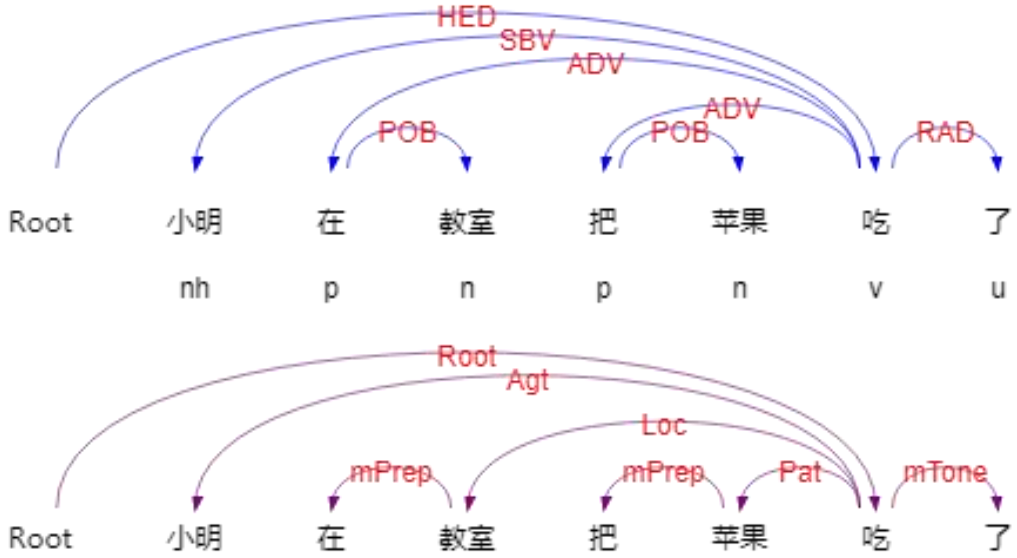
语义依存关系分为三类，分别是主要语义角色，每一种语义角色对应存在一个嵌套关系和反关系；事件关系，描述两个事件间的关系；语义依附标记，标记说话者语气等依附性信息。

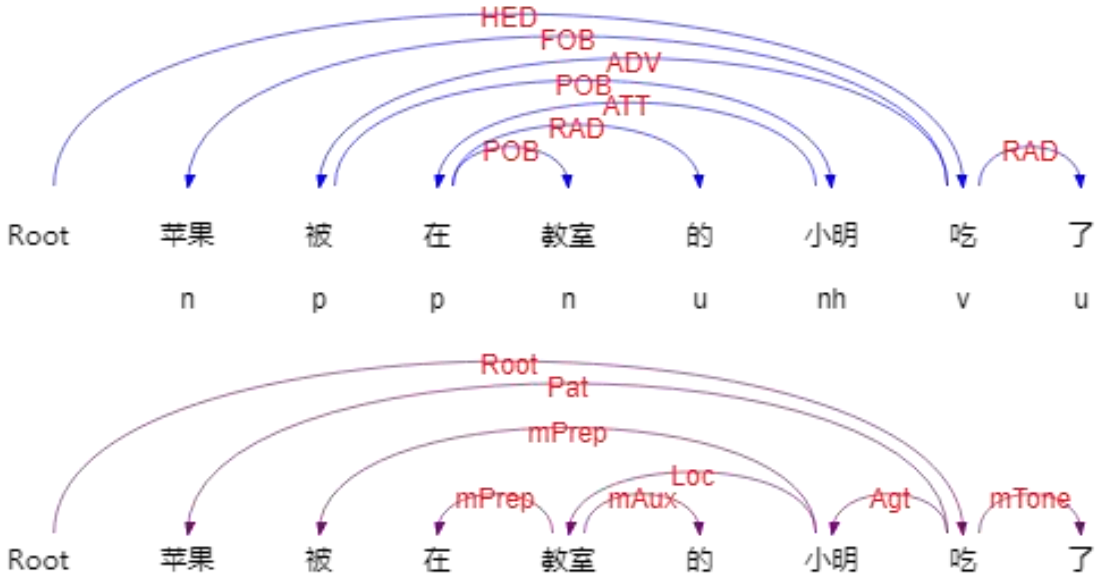
关系类型	Tag	Description	Example
施事关系	Agt	Agent	我送她一束花 (我 <-- 送)
当事关系	Exp	Experiencer	我跑得快 (跑 --> 我)
感事关系	Aft	Affection	我思念家乡 (思念 --> 我)
领事关系	Poss	Possessor	他有一本好读 (他 <-- 有)
受事关系	Pat	Patient	他打了小明 (打 --> 小明)
客事关系	Cont	Content	他听到鞭炮声 (听 --> 鞭炮声)
成事关系	Prod	Product	他写了本小说 (写 --> 小说)
源事关系	Orig	Origin	我军缴获敌人四辆坦克 (缴获 --> 坦克)
涉事关系	Datv	Dative	他告诉我个秘密 (告诉 --> 我)
比较角色	Comp	Comitative	他成绩比我好 (他 --> 我)
属事角色	Belg	Belongings	老赵有俩女儿 (老赵 <-- 有)
类事角色	Clas	Classification	他是中学生 (是 --> 中学生)
依据角色	Accd	According	本庭依法宣判 (依法 <-- 宣判)
缘故角色	Reas	Reason	他在愁女儿婚事 (愁 --> 婚事)
意图角色	Int	Intention	为了金牌他拼命努力 (金牌 <-- 努力)
结局角色	Cons	Consequence	他跑了满头大汗 (跑 --> 满头大汗)

语义角色类型

语义角色类型	说明
ADV	adverbial, default tag (附加的，默认标记)
BNE	beneficiary (受益人)
CND	condition (条件)
DIR	direction (方向)
DGR	degree (程度)
EXT	extent (扩展)
FRQ	frequency (频率)
LOC	locative (地点)
MNR	manner (方式)
PRP	purpose or reason (目的或原因)
TMP	temporal (时间)
TPC	topic (主题)
CRD	coordinated arguments (并列参数)
PRD	predicate (谓语动词)
PSR	possessor (持有者)
PSE	possessee (被持有)

NLP基础_语义分析


HED
SBV
ADV
POB
ADV
POB
RAD
Root
小明
在
教室
把
苹果
吃
了
nh
p
n
p
n
v
u
Root
Agt
Loc
mPrep
mPrep
Pat
mTone
Root
小明
在
教室
把
苹果
吃
了


HED
FOB
ADV
POB
ATT
RAD
POB
RAD
Root
苹果
被
在
教室
的
小明
吃
了
n
p
p
n
u
nh
v
u
Root
Pat
mPrep
Loc
mPrep
mAux
Agt
mTone
Root
苹果
被
在
教室
的
小明
吃
了

HanLP

- 自然语言处理， 功能如下：
  - 中文分词
  - 词性标注
  - 命名实体识别
  - 依存句法分析
  - 新词发现
  - 关键词短语提取
  - 自动摘要
  - 文本分类聚类
  - 拼音简繁

```
C:\Python3.5\Scripts>pip install pyhanlp -i https://pypi.tuna.tsinghua.edu.cn/simple/
Collecting pyhanlp
  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/db/78/5e20dadc4b0e63f0c0b8feb6b752
.48.tar.gz (57kB)
    100% |████████████████████████████████| 61kB 880kB/s
Collecting jpype1>=0.7.0 (from pyhanlp)
  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/28/63/784834e8a24ec2e1ad7f703c3dc6
0.tar.gz (470kB)
    100% |████████████████████████████████| 471kB 334kB/s
Installing collected packages: jpype1, pyhanlp
  Running setup.py install for jpype1 ... done
  Running setup.py install for pyhanlp ... done
Successfully installed jpype1-0.7.0 pyhanlp-0.1.48
You are using pip version 9.0.3, however version 19.2.1 is available.
You should consider upgrading via the 'python -m pip install --upgrade pip' command.
```

安装jpype如果失败，需要单独安装

- http://hanlp.com/
- https://github.com/hankcs/HanLP
- https://github.com/hankcs/pyhanlp
- http://www.hankcs.com/nlp/part-of-speech-tagging.html

HanLP

- 安装过程：
  - 1. 安装JDK1.8以上的版本，配置JAVA_HOME环境变量；
  - 2. pip install pyhanlp安装PyHanLP；
  - 3. 下载**data-for-1.7.4.zip**, 并将其解压放置在pyhanlp模块下的static文件夹下的data子文件夹中。
  - 4. 进入Python命令行，执行import pyhanlp(自动下载配置相关资源)；如果是网络原因，可以直接下载**hanlp-1.7.4-release.zip**解压后将文件放置在static文件夹下。
  - 5. 配置pyhanlp(配置hanlp.properties)。
  - NOTE: 下载路径：https://github.com/hankcs/HanLP/releases

jiagu

- 功能：
  - 中文分词
  - 词性标注
  - 命名实体识别
  - 知识图谱关系抽取
  - 关键词提取
  - 文本摘要
  - 新词发现

2. 命名实体说明（采用BIO标记方式）

B-PER、I-PER	人名
B-LOC、I-LOC	地名
B-ORG、I-ORG	机构名

```
C:\Users\ibf>pip install jiagu -i https://pypi.tuna.tsinghua.edu.cn/simple/
Looking in indexes: https://pypi.tuna.tsinghua.edu.cn/simple/
Collecting jiagu
  Downloading https://pypi.tuna.tsinghua.edu.cn/packages/fd/08/9b2862673e05d75e544f972fc131dc5(
.tar.gz (49.2MB)
    100% |████████████████████████████████| 49.2MB 417kB/s
Requirement already satisfied: tensorflow>=1.4.0 in c:\anaconda3\lib\site-packages (from jiagu)
Requirement already satisfied: numpy>=1.12.1 in c:\anaconda3\lib\site-packages (from jiagu) (1.
Requirement already satisfied: wheel>=0.26 in c:\anaconda3\lib\site-packages (from tensorflow>=
Requirement already satisfied: six>=1.10.0 in c:\anaconda3\lib\site-packages (from tensorflow>=
Requirement already satisfied: tensorflow-tensorboard<0.5.0,>=0.4.0rc1 in c:\anaconda3\lib\site
) (0.4.0)
Requirement already satisfied: protobuf>=3.3.0 in c:\anaconda3\lib\site-packages (from tensorf
Requirement already satisfied: enum34>=1.1.6 in c:\anaconda3\lib\site-packages (from tensorflo
Requirement already satisfied: werkzeug>=0.11.10 in c:\anaconda3\lib\site-packages (from tensor
rflow>=1.4.0->jiagu) (0.11.15)
Requirement already satisfied: markdown>=2.6.8 in c:\anaconda3\lib\site-packages (from tensorf
low>=1.4.0->jiagu) (2.6.11)
Requirement already satisfied: html5lib==0.9999999 in c:\anaconda3\lib\site-packages (from tens
sorflow>=1.4.0->jiagu) (0.9999999)
Requirement already satisfied: bleach==1.5.0 in c:\anaconda3\lib\site-packages (from tensorflo
w>=1.4.0->jiagu) (1.5.0)
Requirement already satisfied: setuptools in c:\anaconda3\lib\site-packages\setuptools-27.2.0-p
ow>=1.4.0->jiagu) (27.2.0)
Building wheels for collected packages: jiagu
  Running setup.py bdist_wheel for jiagu ... done
  Stored in directory: C:\Users\ibf\AppData\Local\pip\Cache\wheels\48\0d\e9\a6121302bd7130189b5
Successfully built jiagu
Installing collected packages: jiagu
Successfully installed jiagu-0.1.7
```

要求TensorFlow1.6版本以上

```
C:\Users\ibf>python
Python 3.6.6 |Anaconda, Inc.| (defaul
Type "help", "copyright", "credits" c
>>> import jiagu
```

1. 词性标注说明

n	普通名词
nt	时间名词
nd	方位名词
nl	处所名词
nh	人名
nhf	姓
nhs	名
ns	地名
nn	族名
ni	机构名
nz	其他专名
v	动词
vd	趋向动词
vl	联系动词
vu	能愿动词
a	形容词
f	区别词
m	数词
q	量词
d	副词
r	代词
p	介词
c	连词
u	助词
e	叹词
o	拟声词
i	习用语
j	缩略语
h	前接成分
k	后接成分
g	语素字
x	非语素字
w	标点符号
ws	非汉字字符串
wu	其他未知的符号

jiagu

```
>>> text = '姚明（Yao Ming），1980年9月12日出生于上海市徐汇区，祖籍江苏省苏州市吴江区震泽镇，前中国职业篮球运动员，司职中锋，现任中职联公司董事长兼总经理。'
>>> knowledge = jiagu.knowledge(text)
>>> print(knowledge)
[['姚明', '出生日期', '1980年9月12日'], ['姚明', '出生地', '上海市徐汇区'], ['姚明', '祖籍', '江苏省苏州市吴江区震泽镇']]
>>> text = "周星驰，1962年6月22日生于香港，祖籍浙江宁波"
>>> knowledge = jiagu.knowledge(text)
>>> print(knowledge)
[['周星驰', '出生日期', '1962年6月22日'], ['周星驰', '出生地', '香港'], ['周星驰', '祖籍', '浙江宁波']]
```

```
>>> text = '厦门明天会不会下雨'
>>> words = jiagu.seg(text) # 分词
>>> print(words)
['厦门', '明天', '会', '不会', '下雨']
>>> pos = jiagu.pos(words) # 词性标注
>>> print(pos)
['ns', 'nt', 'vu', 'vu', 'v']
>>> ner = jiagu.ner(text) # 命名实体识别
>>> print(ner)
['B-LOC', 'I-LOC', 'O', 'O', 'O', 'O', 'O', 'O', 'O']
```