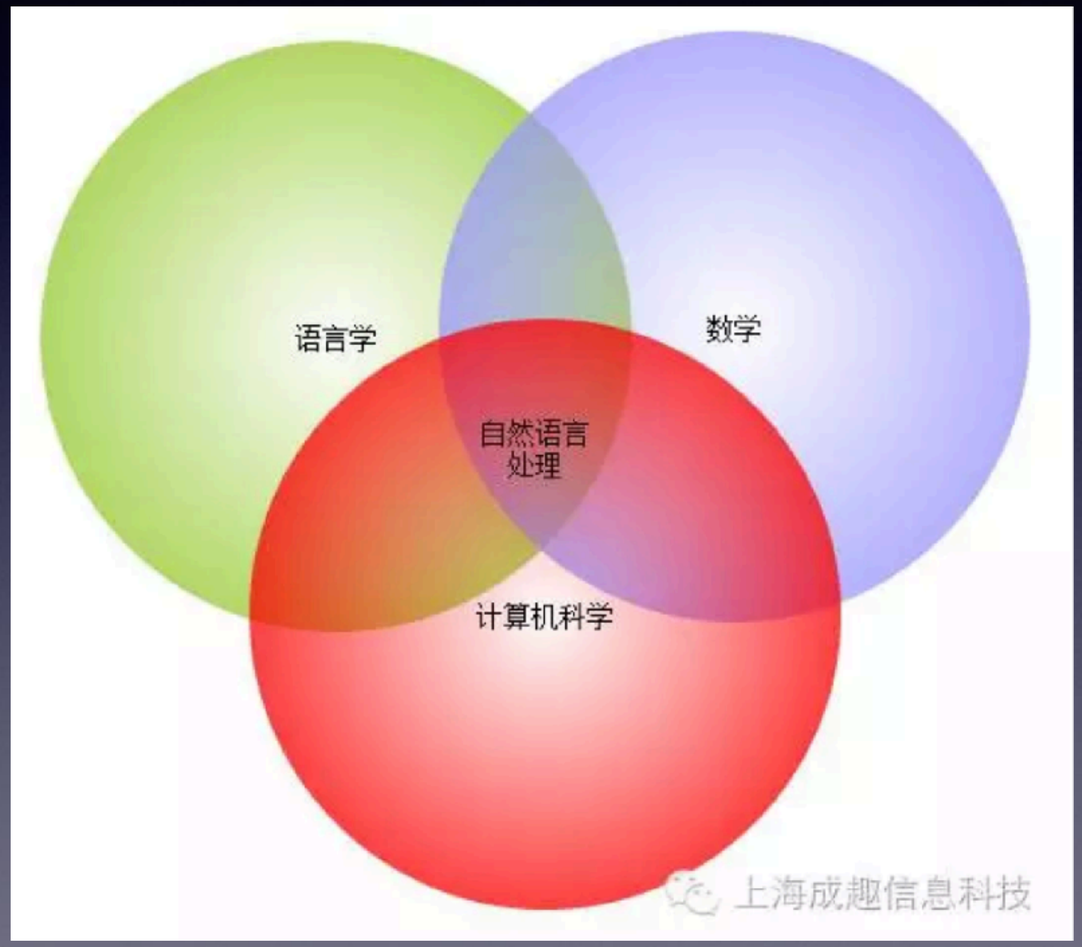
用LTP玩转NLP（自然语言处理）

Jane 上海成趣信息科技有限公司

- 自然语言处理的基本原理
- 自然语言处理的基本技能
- 分词
- 词性标注
- 命名实体识别
- 依存句法分析
- 语义角色标注
- 一起用哈工大LTP玩转NLP

自然语言处理

- 自然语言处理是用电子计算机处理和加工人类的书面和口头语言信息的技术；
- 是人工智能的一个主要内容，是人类用电子计算机模拟人类智能的一个重要尝试，也是人工智能领域的一大难题；
- 这项技术是一门专门的边缘性交叉性学科



自然语言处理系统

- 机器翻译系统
- 自然语言理解系统
- 信息自动检索系统
- 信息自动抽取系统
- 文本信息挖掘系统
- 术语数据库系统
- 计算机辅助教学系统
- 语音自动识别系统
- 语音自动合成系统
- 文字自动识别系统

自然语言处理的基本原理

- 基于句法—语义规则的方法解决了简单问题；
- 随着语料库建设和语料库语言学的发展以及计算机技术本身的发展，使用概率和数据驱动的方法，自然语言处理可以专注于大规模真实文本的处理；
- 语料库语言学（英语：corpus linguistics）是基于语言运用的实例（即语料库）的语言研究。 语料库语言学可以对自然语言进行语法与句法分析，还可以研究它与其他语言的关系；
- 在句法剖析、词类标注、参照消解、话语分析、机器翻译这些技术中引入概率，并且采用从语音识别和信息检索中借鉴的基于概率和数据驱动的评测方法。

自然语言处理的基本原理（续）

- 机器自动学习：计算机自动地从语料库中获取准确的语言知识
- 工作重点：建设机器词典和大规模语料库
- 由于建造标注语料库需要花费较高的成本，相比有监督的机器学习方法，无监督的机器学习技术会得到更广泛的应用。支持向量机技术、最大熵技术、多项逻辑回归、图式贝叶斯模型技术广泛应用于自然语言处理研究。
- 随着高性能计算机的发展和应用，机器学习系统可以得到很好的训练，系统性能也能得到提高。
- 使用机器学习方法开发的基于语料库的自动分析软件是独立于具体语言的。研究者不需要懂相关语言，只要基于训练语料库使用自动分析软件就可以得出不错的分析结果。

自然语言处理的基本技能

- 机器学习的相关知识以及统计学习的主要方法，特别是监督学习方法：
包括感知机、k近邻法、朴素贝叶斯法、决策树、逻辑斯谛回归与最大熵模型、支持向量机、提升方法、EM算法、隐马尔可夫模型和条件随机场等

- 理解自然语言处理的基本线路
分词、词性标注、解析

分词

- https://ltp.readthedocs.io/zh_CN/latest/appendix.html#id2

基于字的序列标注问题。对于输入句子的字序列，模型给句子中的每个字标注一个标识词边界的标记。

- 基于词典的方法
- 基于统计的方法
- 基于规则的方法

标记	含义	举例
B	词首	_中_国
I	词中	哈_工_大
E	词尾	科_学_
S	单字成词	的

分词方法：基于词典的方法

- 定义:按照一定策略将待分析的汉字串与一个“大机器词典”中的词条进行匹配，若在词典中找到某个字符串，则匹配成功。
- 按照扫描方向的不同：正向匹配和逆向匹配
- 按照长度的不同：最大匹配和最小匹配

分词方法：基于统计的方法

- 主要思想：上下文中，相邻的字同时出现的次数越多，就越可能构成一个词。因此字与字相邻出现的概率或频率能较好的反映词的可信度。
- 主要统计模型为：N元文法模型（N-gram）、隐马尔科夫模型(Hidden Markov Model, HMM
- 第n个词的出现只与前面N-1个词相关，而与其它任何词都不相关，整句的概率就是各个词出现概率的乘积

分词方法：基于规则的方法

通过模拟人对句子的理解，达到识别词的效果。

基本思想：语义分析，句法分析，利用句法信息和语义信息对文本进行分词。自动推理，并完成对未登录词的补充。

具体概念:有限状态机\语法约束矩阵\特征词库

词性标注

- 863词性标注集https://ltp.readthedocs.io/zh_CN/latest/appendix.html#id3
- viterbi算法原理及适用情况
- 当事件之间具有关联性时，可以通过统计两个以上相关事件同时出现的概率，来确定事件的可能状态。
- 给出一个句子后，我们需要给这个句子的每个词确定一个唯一的词性，实际上也就是在若干词性组合中选择一个合适的组合。

Tag	Description	Example	Tag	Description	Example
a	adjective	美丽	ni	organization name	保险公司
b	other noun-modifier	大型, 西式	nl	location noun	城郊
c	conjunction	和, 虽然	ns	geographical name	北京
d	adverb	很	nt	temporal noun	近日, 明代
e	exclamation	哎	nz	other proper noun	诺贝尔奖
g	morpheme	茨, 甥	o	onomatopoeia	哗啦
h	prefix	阿, 伪	p	preposition	在, 把
i	idiom	百花齐放	q	quantity	个

命名实体识别

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

https://ltp.readthedocs.io/zh_CN/latest/appendix.html#id4

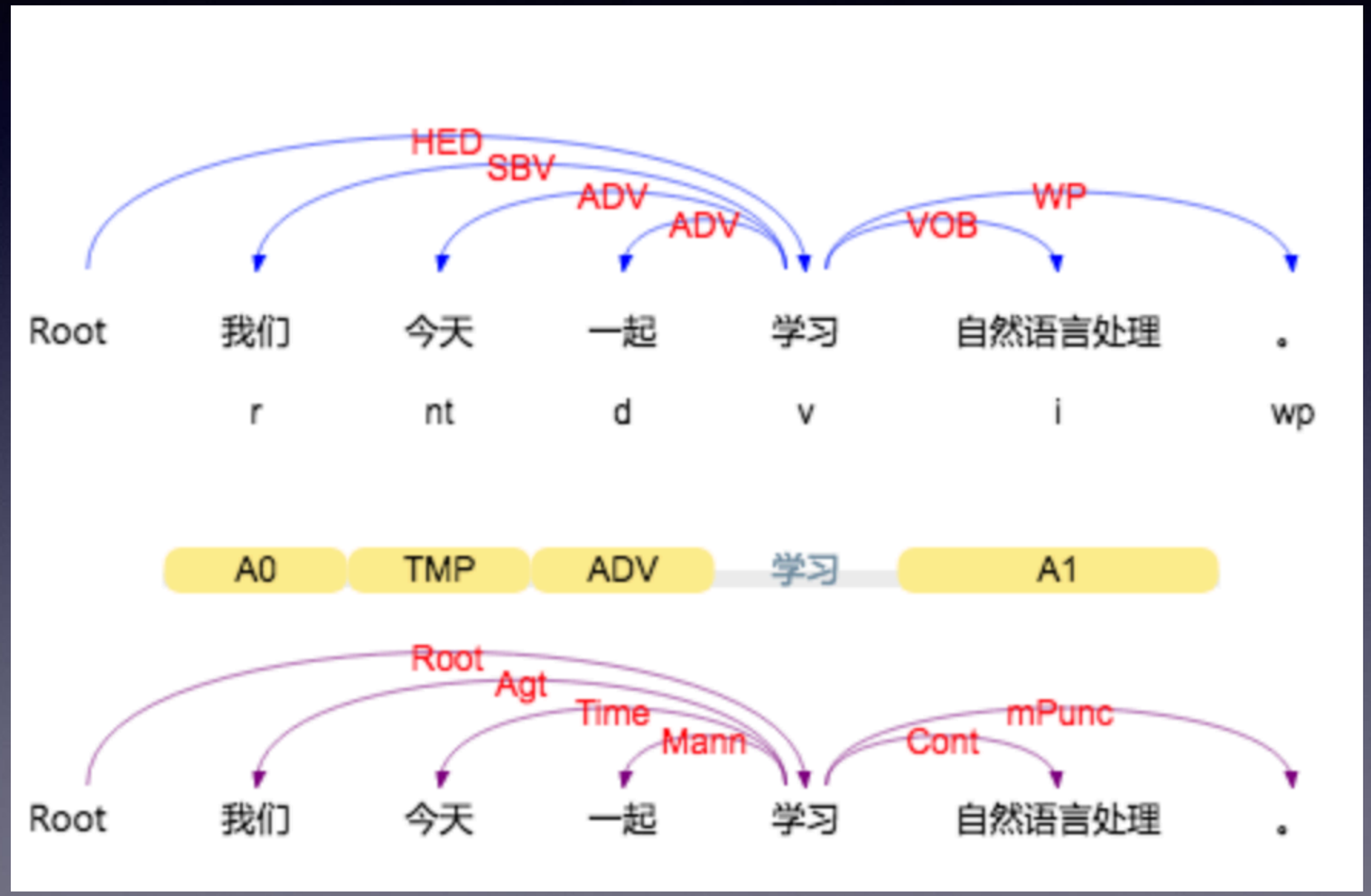
分类器方法

标记	含义
Nh	人名
Ni	机构名
Ns	地名

依存句法分析

依赖解析 (Dependency Parsing, DP) 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。即分析识别句子中的“主谓宾”、“定状补”这些语法成分，并分析各成分之间的关系。

https://ltp.readthedocs.io/zh_CN/latest/appendix.html#id5



依存分析算法

- 依据神经网络依存句法分析算法，Chen and Manning (2014)；
- 同时加入丰富的全局特征和聚类特征；
- 在模型训练时，参考了Yoav等人关于dynamic oracle的工作

语义角色标注

- https://ltp.readthedocs.io/zh_CN/latest/appendix.html#id6

- 主要方法：机器学习的监督算法

支持向量机 (SVM),

最大熵 (Maximum Entropy,

SNoW (Sparse Network of W i n n o w s)

- 丰富有效的特征对语义角色标注来说更加重要

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)

语义角色标注系统的过程

1. 使用启发式规则过滤多数不可能是语义角色的句法成分

2. 识别语义角色，用二元分类器将角色候选为语义角色和非语义角色

3. 使用多类分类器将第二阶段识别的语义角色

注：有些系统会加入基于启发式规则的后处理阶段

参考论文：刘怀军等中文语义角色标注的特征工程

哈工大LTP

LTP提供了一系列中文自然语言处理工具，用户可以使用这些工具对于中文文本进行分词、词性标注、句法分析等等工作。从应用角度来看，LTP为用户提供了下列组件：

- 针对单一自然语言处理任务，生成统计机器学习模型的工具
- 针对单一自然语言处理任务，调用模型进行分析的编程接口
- 使用流水线方式将各个分析工具结合起来，形成一套统一的中文自然语言处理系统
- 系统可调用的，用于中文语言处理的模型文件
- 针对单一自然语言处理任务，基于云端的编程接口

LTP相关信息

- 官网：http://www.ltp-cloud.com/
- 项目地址：https://github.com/HIT-SCIR/ltp
- 快速安装：https://github.com/HIT-SCIR/ltp/blob/master/doc/install.rst
- 安装LTP的Python 封装：https://github.com/HIT-SCIR/pyltp
- 在线演示：http://ltp.ai/demo.html
- 常见问题：http://ltp.ai/faq.html
- 讨论区：https://github.com/HIT-SCIR/pyltp/issues

LTP实现原理与性能

- https://ltp.readthedocs.io/zh_CN/latest/theory.html#id2

用哈工大LTP玩转NLP

- http://ltp.ai/docs/index.html

ACM-W China

支持、庆祝和倡导中国女性充分
参与计算领域的各个方面

DevHub开发者社区

分享、启发、探索

传播IT知识文化
陪伴探索者前行

杨晓春
上海成趣信息科技有限公司
独立顾问

产品设计
技术开发、技术管理
人工智能、数据分析解决方案
物联网解决方案
医疗养老产品
DevHub开发者社区