自然语言处理应用实践

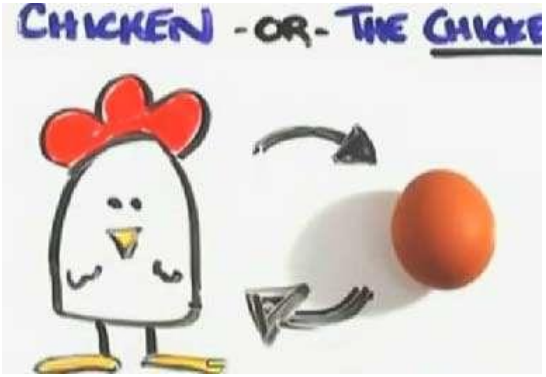
——暑期课程

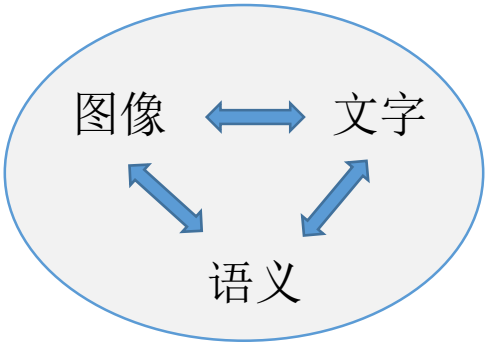
南京大学软件学院
李传艺
费彝民楼917

第一部分：自然语言处理基础理论与技术

- 自然语言特性
  - 从2个特性到7个特性
- 词汇与结构
  - 词汇语义学、结构语义学
  - 词汇歧义、结构歧义
- 正则表达式、最小编辑距离、最大公共子串（文本相似度）
- 文本特征表示（语言模型）
  - 词袋模型
  - 语义表示
- 中文分词
  - 基本理论
  - 工具使用

如何表达语义





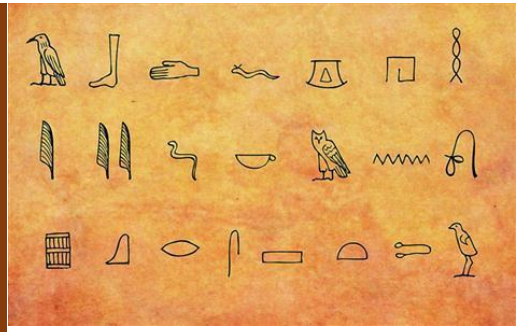
自然语言特性（1）

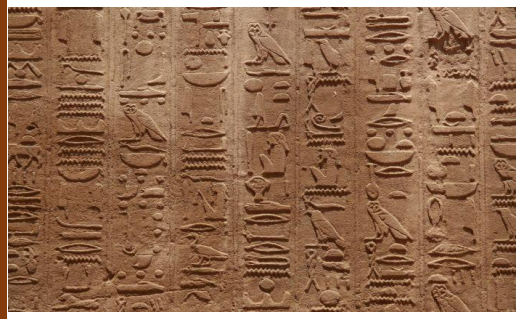
- 索绪尔（Ferdinand de Saussure）《普通语言学教程》
  - 语言：一种表达观念的符号系统
    - 词汇：符号集合，符号由能指和所指组成
      - 能指：声音和字形，符号本身
      - 所指：表达的概念和意义
    - 语法：词汇（符号）之间的关系，组合关系、选择关系
  - 言语：运用语言规则生产的具体话语
  - 言语行为：根据语言规则说话的活动
  - 语言学的研究对象——语言

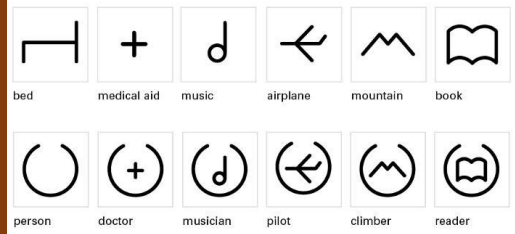
研究语言 *vs.* 研究言语

- 语言的特性1：符号的任意性
  - 能指和所指之间的关系是任意的，是随机的约定俗成
- 语言的特性2：能指的线条性
  - 符号只能在时间上展开，相继出现，构成一个链条 ？
  - 是语言单位的切分和替换的基本前提

一种规则存在于每个人意识中，是社会成员共有的表达媒介







自然语言特性（2）

- 电子计算机时代，特别是自然语言处理技术出现后，认识到的语言特性有：
  - 语言符号的层次性
    - 语言符号并非线条性的，而是立体的；从能指的发音上是立体的，从言语的结构上更是立体的。
    - 示例：The old man and women stayed at home.
  - 语言符号的非单元性
    - 每一个符号都不是不可分割的单元：一字多**义**，完全不相干的语义；语言符号是复杂的，可再分割的。
    - 语言符号不是最小不可分割单元，类似物理学中的分子
    - 对自然语言处理技术中“语义表示”具有很重要的启发：在一个表示中融合多种语义
  - 语言符号的离散性
    - 字符集的离散【区别于语义上的非单元性】：连续的言语是由许多离散的单元组成的
    - 通过停顿表达不同的语义
    - 典型应用
      - 汉语分词——利用词语之间的离散特征将相互连接的词语切开
      - 以词典的方式整理语言库
  - 语言符号的递归性
    - 语法规则（能指）是有限的，但是言语（所指）却是无限的
    - 《乔姆斯基语言理论介绍》：语言是有限手段的无限运用。

语言本身是离散性和连续性的统一体

使用连续性函数分析语言对应的言语

自然语言特性（3）

- 电子计算机时代，特别是自然语言处理技术出现后，认识到的语言特性有（续）：
  - 语言符号的随机性
    - 不是能指与所指关系的随机性，而是语言符号在言语中使用的随机性
    - 有些所指的能指使用的多，有些所指的能指使用的少——不确定性即是使用的随机性
    - 很多语法规则之外的句子——随机性的体现

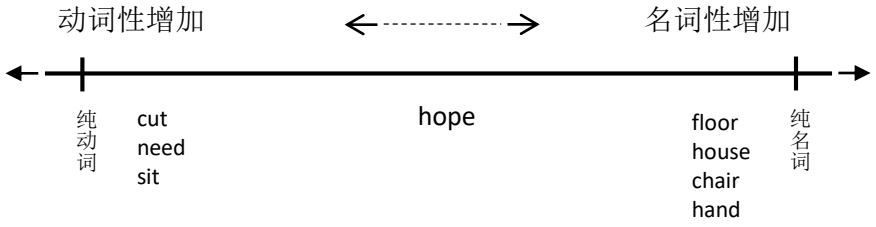
从符号本身定义语用规则 → 从言语库中统计语用现象

  - 语言符号的冗余性
    - 不是一切成分都是对于表达整体语义必不可少的，语言有能力将缺少的部分补充或恢复出来
    - 必要的、有益的：在不理想的环境下仍可以保证顺畅的沟通
    - 不同语言的冗余度
      - 英文：67%~80%
      - 中文：56%~74%

？

语言没有冗余更好

  - 语言符号的模糊性
    - “思想本身像是一团星云，没有必然划定的界限”，“没有符号的帮助，我们就没法清楚地、坚实地区分两个概念”。
    - 思想的模糊性➔语言的模糊性
    - 颜色构成的连续系统中使用单一词语描述固定颜色；红色？秃子？
    - 不仅语义模糊，语法也存在模糊性



自然语言特性（4）

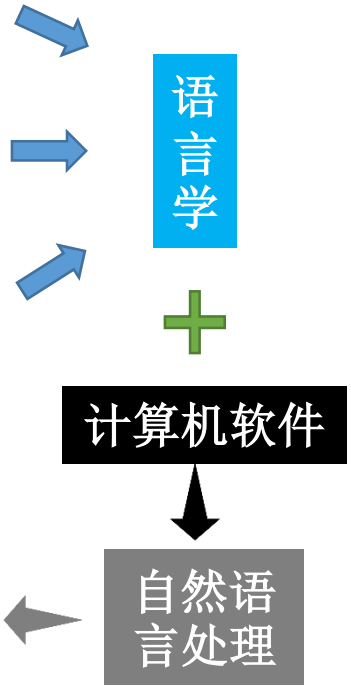
- 七大特性都是语言符号本身的特性?

- 多数是“言语”（语言符号实际使用过程中）体现出来的特性
  - 层次性
  - 非单元性
  - 离散性
  - 递归性
  - 随机性
  - 冗余性

  → 体现语言符号“物质-自然”的本质，使用自然科学的方法研究语言

- 属于语言符号本身的特性：人类心智活动和思维活动的特点
  - 模糊性 → 体现语言符号“智能-心理”的本质，使用思维科学的方法研究语言

- 能指与所指关系的随机性：社会约定性
  - 任意性 → 体现语言符号“社会-人文”的本质，使用社会科学的方法研究语言



研究的对象是言语

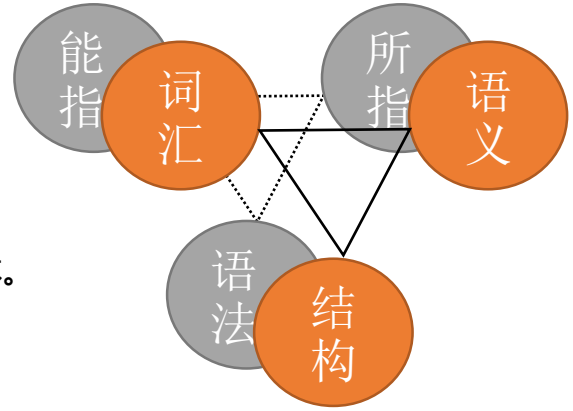
1. 根据语言规则自动化分析和理解言语
2. 自动从言语中剖析和发现语言潜规则

词汇与结构（1）

- 语言规则：能指、所指及其之间的关系；语法。

词汇是语言描述的中心

——英国功能语言学奠基人，弗斯



- 什么是词汇?
  - 搭配理论：某些词常常与某些词一起使用，“意义取决于搭配”
  - 中文：开车，开门，开水，开路，开关…
  - 英文：cow和milk，milk the cows，而不会用tigress和milk搭配
  - 词汇是组成句子的基本成分；是语言的建筑材料；是话语实现的主要载体。
- 在实现意义时，词汇与语法是交织在一起的，必须整合描述。
  - 生成语法学：词汇是所有语言之间所有差异的潜在所在，排除词汇的差异，人类语言应该只有一种。
  - 还原主义者：从较大“结构”探索较小“基元”的行为
  - 组成性原则：句子的意义是由成分的意义组合而成的，成分的意义决定了整个句子的意义。
- 什么是结构?
  - 简单说就是“语法”、“句法”
  - 各种不同语言学派的理念中，剥离“词汇”和“语义”后，剩余的就是“结构”
  - 句子中“词语”或“成分”之间的位置关系、组合规则、逻辑关系等等，都称为结构。

词汇是一个独立的语言学层面

词汇与结构（2）——语义学

- 什么是词汇语义学?
  - 词汇本身的语义信息是很重要的，在自然语言处理中，应该重视词汇语义的研究。
  - 词汇具有高度系统化的结构：单词与其意义之间的关系、个别单词的内部结构。

对这种系统化的、与意义相关的结构的词汇研究称为**词汇语义学, Lexical Semantics**

- 术语
  - 词位：词典中一个单独的条目，是一个特定的正字法形式、音素形式和一些符号的意义表示形式的组合。
  - 涵义：词位的意义部分。
  - 词典：是有限个词位的列表，也是无限的意义的生成机制。

一个词位有若干彼此关联的涵义

- 词位与涵义之间存在复杂的关系

同音异义 *vs.* 同形异义　　多义关系现象

  - 同形关系
    - 形式【发音/正词法形式】相同而意义上没有关系的词位之间的关系。
  - 同义关系
    - 可替换性：如果两个词位可以相互替换而不改变意思或句子可接受度，则同义。
  - 上下位关系
    - 一个词位是另一个词位的次类；特定性强（下位词）⇔概括性强（上位词）
  - 整体-部分关系
    - 手 和 （虎口、手臂、手掌）；汽车 和 （方向盘、车轮）
  - 集合-元素关系
    - 五岳 和 （泰山、华山、嵩山、恒山、衡山）

在具体的句子中讨论同义

中文詞彙網路 CHINESE WORDNET

WordNet A Lexical Database for English

涵义有无　意义色彩　搭配约束　社会因素

词汇与结构（3）——语义学

基础

- 什么是结构语义学？ 词汇语义学 ⟹ 独立于上下文语境 ⟹ 静态的
  - 求解句子中的单词之间的语义关系，是动态的，这种语义关系是随着单词在上下文语境而改变的。⟹ 动态的

- 问题1：题元角色关系，Thematic Role Relation
  - 句子中单词语义关系可以有多种不同表示方法
  - 题元角色就是一些范畴符号，可以作为描述动词论元的一种浅层的语义标记

“格”表示　配价语法表示　谓词论元表示

- 各种不同结构下
- 不同语义场景下

施事者，AGENT
经验者，EXPERIENCER
施力者，FORCE
主题，THEME
结果，RESULT

内容，CONTENT
工具，INSTRUMENT
受益者，BENEFICIARY
来源，SOURCE
目标，GOAL

事件

- 问题2：选择限制，Selection Restriction
  - 一个词位对于它的各个论元角色所施加的“语义约束”叫做选择限制
  - 针对的是词位的某个特定的涵义，而非整个词位
    - 使用选择限制根据上下文进行歧义消解
  - 不同词位、同一词位的不同涵义所施加的选择限制可能大不相同，有些选择很广，有些很窄
    - 如何表示选择限制？

Which airlines **serve** *Beijing*?
Which ones **serve** *breakfast*?

一阶谓词演算，FOPC

词网同义词集，Wordnet SYNSET

I cannot **imagine** *what this lady does all day*.
I often ask the musicians to **imagine** *a tennis game*.
To **diagonalize** *a matrix* is to find its eigenvalues.

词汇与结构（4）——歧义

John is a **bachelor**. → John is an unmarried man.

John holds a first university degree.

- 词汇歧义——一词多义（以英文为例）
  - 词义排歧非常重要：机器翻译、信息检索、文本分类、语音识别
  - 歧义的类型
    - 名词（多义词，同形异义词，相同的单复数，缩写）
    - 代词（指代不清）➔ 指代消解　　He shot **himself**.
    - 动词、形容词、连词、介词（多义词）

He looked at the river **bank**.

He looked at the **bank**.

He looked at the money **bank**.

John is **with** Tom.　　The damage was done **by** the river.

  - 基于知识的词义排歧方法：
    - 选择最常见的涵义的方法，Most Frequency Approach
    - 基于规则的方法：词类，选择限制（Semantic Frame和Semantic Distance），优选关系
    - 基于机器学习的方法：有监督，半监督，无监督
    - 基于词典的方法：机器可读的词典+**义项解释之间的相似度** → 文本相似度

- 结构歧义——一个以上的语法剖析
  - 附着歧义
  - 并列歧义　　他负责照顾年老的男人和女人.
  - 名词短语括号歧义——中文分词歧义　　南京市长江大桥。
  - 特殊歧义结构：V-ing， V-ed，to V，not to V...
  - 修饰语歧义　　这是一个很可爱的小女孩的裙子。
  - 状语歧义

【中英文的差异？】

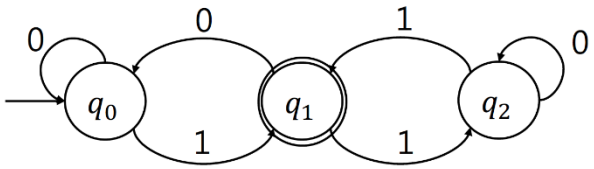
She knew that, before I met you, you had begun to study NLP.

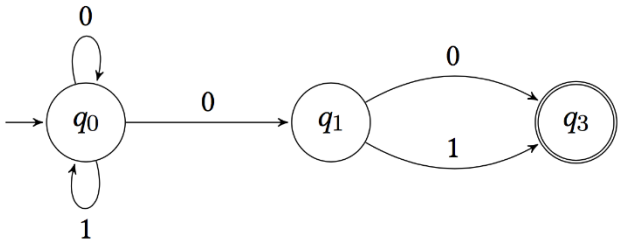
She knew that you had begun to study NLP before I met you.

Before I met you, she knew that you had begun to study NLP.

基础文本计算——词汇自动处理

- 正则表达式，Regular Expression
  - 有限自动机
    - 确定的有限自动机
    - 不确定的有限自动机
    - 相互转换（算法）

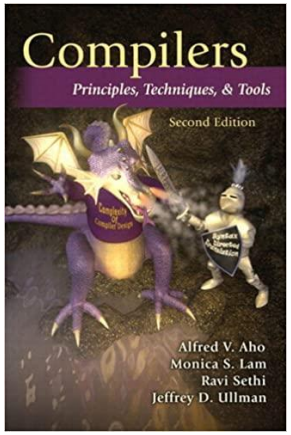


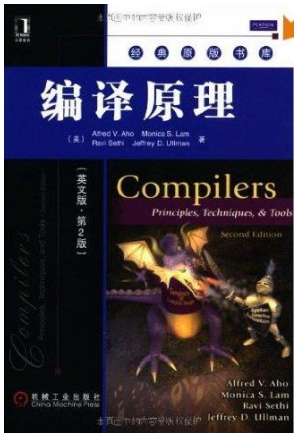


- 常用的正则表达式符号

https://www.runoob.com/regexp/regexp-tutorial.html 菜鸟教程

https://baike.baidu.com/item/%E6%AD%A3%E5%88%99%E8%A1%A8%E8%BE%BE%E5%BC%8F/1700215?fr=aladdin 百度百科





正则事件是可以被有限自动机表示的事件，而且有限自动机可以表示的事件也一定是正则事件

- 文本搜索（字符串匹配）——模式匹配

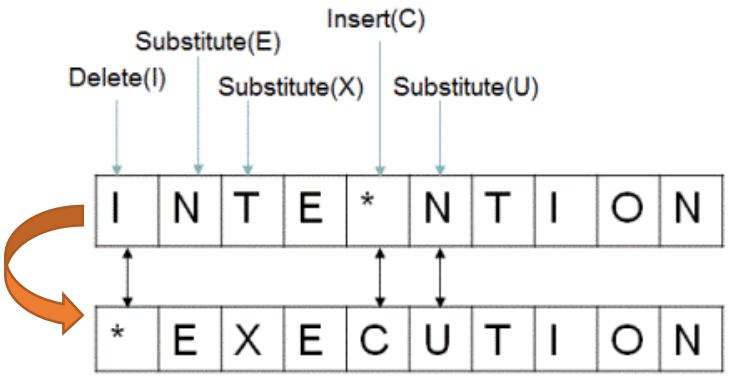
Horspool字符串匹配算法； Boyer-Moore字符串匹配算法

【如何搜索正则表达式？】

基础文本计算——词汇自动处理

- 最小编辑距离——Minimum Edit Distance
  - 判断两个单词中，哪一个在拼写上更接近第三个单词，是“**字符串距离**”的一种特殊情况。
  - 给定两个字符串A和B，求字符串A至少经过多少步字符操作变成字符串B。允许的操作有：
    - 删除一个字符
    - 插入一个字符
    - 替换一个字符

设定不同操作有不同权值



  - 很多种不同的操作序列，每一种称为Path，如：
    - 删除Intention所有字符，再插入Execution所有字符

Dynamic Programming
动态规划

【如何编程实现？】
如何对齐？
从哪里开始执行？

$$\mathrm{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0, \\ \min \begin{cases} \mathrm{lev}_{a,b}(i-1,j)+1 \\ \mathrm{lev}_{a,b}(i,j-1)+1 \\ \mathrm{lev}_{a,b}(i-1,j-1)+1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases}$$

基础文本计算——词汇自动处理

- 最长公共子串——Longest Common Subsequence
  - ADBCDACBA
  - ABCADBACB
  - 最长公共子序列：ABCDACB

- 蛮力解法
  - 对某一个序列(长度为m)的所有子串判断是否是另一个序列(长度n)的子串
  - 时间复杂度?
    - 每一次检查是O(n)
    - 共有2^m个子串
    - O(n*2^m)

- 动态规划（Dynamic Programming）

自底向上求解

递推关系

$$num[i][j]=\begin{cases}0 & i=0\text{或者}j=0\\ 1+num[i-1][j-1] & i,j>0,a[i]=b[j]\\ \max\{num[i][j-1],num[i-1][j]\} & i,j>0,a[i]\neq b[j]\end{cases}$$

i j	0	1	2	3	4	5	6	…	m
0	0	0	0	0	0	0	0	0	0
1	0								
2	0								
3	0								
4	0								
5	0								
…	0								
n	0								

文本特征及其表示

- 文本
  - **文本1**：1921年7月23日，中国共产党第一次全国代表大会在上海召开。由于会场受到法租界巡捕的搜查，最后一天的会议转移到浙江嘉兴南湖的游船上举行。
  - **文本2**：1921年8月3日黄昏，浙江嘉兴南湖的暑热逐渐散去。湖面上一艘中等大小的画舫内，气氛庄重肃穆。在“中国共产党万岁”的低声呼喊中，中国共产党第一次全国代表大会闭幕。
  - **文本3**：2021年是中国共产党百年华诞。中国站在“两个一百年”的历史交汇点，全面建设社会主义现代化国家新征程即将开启。世界将更多目光投向中国，聚焦中国共产党矢志不渝为人民谋幸福，为民族谋复兴，为世界谋大同。
  - **文本4**：巴勒斯坦人民斗争阵线总书记马吉达拉尼表示，中国共产党领导中国创造经济快速发展和社会长期稳定“两大奇迹”，促进中国实现高水平的繁荣和进步，并为世界和平发展和人类文明进步付出了巨大努力。

- 特征
  - 用于文本计算的变量及其取值

同一批文本共享文本特征定义
不同文本的特征取值不同

- 表示
  - 特征的形式：字符串、数值、向量、矩阵

判断文本内容是否与“中国共产党建党100周年”主题活动相关？

长度（字符级别、词语级别、句子级别）

与某个特定内容的关系

语义

正则表达式，
编辑距离，
最长公共子串

能想到哪些特征及其表示方法？

中文分词

- 传统切词方法——机械方法，基于词典
  - 正向最大匹配算法
    - 从左向右扫描文本，比对词典，直到最大长度
  - 逆向最大匹配算法
    - 自文本末尾，从右向左扫描文本，选择最大长度开始比对词典，直到成词
  - 最佳匹配算法
    - 过程不变，只是词表的顺序按照出现频率从高到低排列，优先匹配出现多的词语
  - 基于词频统计的切词法

动手实践

准备词典

实现算法

- 基于机器学习的切词方法——隐马尔可夫模型，条件随机场（CRF）算法
  - 足够的训练样本
  - 学习切词的概率模型

- 分词工具的使用
  - HanNLP
  - Jieba分词——Java版https://github.com/huaban/jieba-analysis
  - 哈工大TLP

知乎分词贴 https://zhuanlan.zhihu.com/p/86322679

使用工具 → 实现TFIDF提取关键词算法

实践汇总

- 算法1：在文本中检索给定的有限状态机表示的模式串
  - 输入：文本，有限状态机
  - 输出：所有串的列表
- 算法2：计算两个字符串的文本编辑距离
  - 输入：字符串1，字符串2
  - 输出：距离int型值
- 算法3：计算两个字符串的最长公共子序列
  - 输入：字符串1，字符串2
  - 输出：公共子序列字符串3

综合
使用完成的算法和程序处理《共产党宣言》
任务一：提取“资产阶级...无产阶级”句子
任务二：找到整个文本中最为相似的两个句子
任务三：统计词频并输出关键词
任务四：构建“共产党”概念图*******

- 程序1：使用Jieba分词对给定文本集合中所有文本进行分词，并实现TFIDF算法获取每个文本的关键词
  - 输入：文本集合
  - 输出：每一个文本分词结果+关键词列表

谢谢！