统计自然语言处理笔记

陈鸿峥

2020.01*

目录

*Build 20200109

1 简介

自然语言处理(natrual language processing, NLP)又叫计算语言学，是研究如何利用计算机技术对语言文本（句子、篇章、话语等）进行加工处理的一门学科，包括对词法、句法、语义、语用等信息的识别、分类、提取转换和生成等各种处理方法和实现技术。主要内容包括：机器翻译、信息检索、自动文摘、观点挖掘、问答系统、信息抽取、文档分类、文字编辑和自动校对、语音识别、文语转换、语音合成、说话人识别/认同/验证。

1.1 基本问题与主要困难

1.1.1 基本问题

- 形态学(morphology)：词由有意义的基本单位—**词素**（词根、前缀、后缀、词尾）的构成问题、单词识别/汉语分词问题
- 句法(syntax)问题：研究句子结构成分之间的相互关系和组成句子序列的规则（**主谓宾**）
- 语义(semantics)问题：研究如何从一个语句中推导出词的意义，以及这些词在语句句法结构中的作用来推导出该语句的意义（同一个词在不同语境下会有**不同意思**）
- 语用学(pragmatics)：研究在不同上下文语句的应用，以及上下文对语句理解产生的影响

1.1.2 主要困难

- 大量歧义：
  - 词法歧义

    研究所/取得

    研究/所取得
  - 词性歧义

    动物保护警察

    Time flies like an arrow.
  - 结构歧义

    今天中午吃馒头

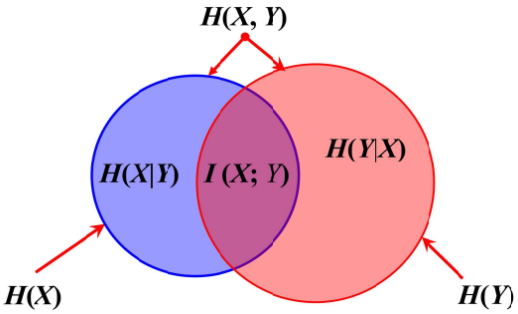
    今天中午吃食堂

    I saw [a man with a telescope].

    I [saw a man] with a telescope.
  - 语义歧义（缩略语、隐喻）

    意思意思是几个意思？
  - 语音歧义
- 大量未知语言现象：新词、人名、地名、术语、新含义、新句法、新句型

1.2 信息论基础



定义 1 (熵/自信息). 概率分布为$p(x) = P(X = x)$，则熵$H(X)$为

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x)$$

并约定$0 \log 0 = 0$，单位为二进制位比特*(bit)*

定义 2 (联合熵). X, Y为离散型随机变量$X, Y \sim p(x, y)$，联合熵定义为

$$H(X, Y) = -\sum_{x \in X} \sum_{y \in Y} p(x, y) \log_2 p(x, y)$$

是描述一对随机变量平均所需的信息量

定义 3 (条件熵).

$$H(Y \mid X) = -\sum_{x \in Y} \sum_{y \in Y} p(x, y) \log_2 p(y \mid x)$$

定义 4 (相对熵(KL距离)). 两个概率分布$p(x)$和$q(x)$的相对熵定义为

$$D(p||q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}$$

定义 5 (交叉熵). 若随机变量$X \sim p(x)$，$q(x)$用于近似$p(x)$的概率分布，则X与模型q的交叉熵定义为

$$H(X, q) = H(X) + D(p||q) = -\sum_{x} p(x) \log q(x)$$

其常用来衡量估计模型与真实概率分布之间的差异

定义 6 (互信息). 如果$(X, Y) \sim p(x, y)$，则X, Y之间的互信息$I(X; Y)$定义为

$$I(X; Y) = H(X) - H(X \mid Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log_2 \frac{p(x, y)}{p(x)p(y)}$$

互信息值越大，表示两个汉字之间的**结合越紧密**，越有可能成词。

例 1 (基于上下文分类的消歧方法). 基于贝叶斯分类器，假设某个多义词w所处上下文语境为C，其多个语义为s_i，则可以通过计算

$$\arg\max_{s_i} p(s_i \mid C) = \arg\max_{s_i} \left[\log p(s_i) + \sum_{v_k \in C} \log p(v_k \mid s_i) \right]$$

确定词义。由统计数据可得

$$p(v_k \mid s_i) = \frac{N(v_k, s_i)}{N(s_i)}$$
$$p(s_i) = \frac{N(s_i)}{N(w)}$$

其中$N(s_i)$为训练数据中词w用于语义s_i时的次数，而$N(v_k, s_i)$为w用于语义s_i时词v_k出现在w的上下文中的次数，$N(w)$为w在训练数据中出现的总次数。

1.3 语言学基础

定义 7 (词性(parts of speech, POS)/句法类/语法类). 相似的语法结构行为和典型的语义类型聚成不同的类，称为词性。

定义 8 (词法). 词法则是构词过程，包括**变形**（修改时态、数目等）、**派生**、**复合**。

短语结构可通过树来表达

[S[NP[AT The][NNS children]][VP[VBD ate][NP[AT the][NN cake]]]]

主要汉字编码标准：ASCII（英文）、GB2312（中文）、BIG5（繁体）、Unicode（统一编码）。

UTF(Unicode Transformation Format)是Unicode的实现方式，从Unicode码点到唯一字节序列的映射算法，一一映射，保证无损转换。

2 中文词法分析

2.1 基本概念

定义 9 (词). 能够独立运用的最小音义结合体。

若S是一个词，则满足

- S**内部字串**粘合度高
- S**外部环境**替换度高
- S**本身频度**高

2.2 文本分词中的歧义

- 交集性歧义

大学生|不看重|大城市|户口
大学生|不看|重大城市|户口

- 组合型歧义

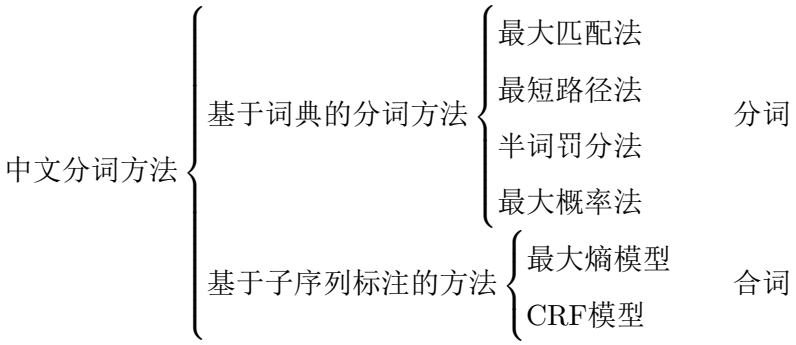
  你认为学生会|听老师的吗
  你认为学生|会听老师的吗

- 交集型歧义

  只有雷人|才能吸引人
  只有雷|人才|能吸引人
  只有雷|人|才能吸引人

中文文本中通常人名、地名、机构名、专业术语等等难被区分。

2.3 基本方法



2.3.1 最大匹配法

最大匹配法：有一个词典，设定最大词长，做字符串匹配

输入：S1= “计算语言学课程是两个课时"
输出：S2= " "
设定最大词长MaxLen = 5
W1= 计算语言学
……

词 典
…
计算语言学
课程
课时
…

大规模真实语料中99%的词例（token）的长度都在5字以内 [1]

2.3.2 最优路径法

- 最短路径法：词数最少最优
  优点：好于单向最大匹配方法
  缺点：同样无法解决大部分交集型歧义
- 半词罚分法（加权）：如果一个字不能单独使用则是半词，每个词罚1分，每个半词加罚1分，求罚分最低路径（词数最少、半词最少）
  缺点：依然无法解决有意见分歧的问题
- 最大概率法：在词图上选择词串概率最大的分词路径（动态规划），最优路径中第i个词W_i的累积概率等于左邻词W_{i-1}累积乘自身

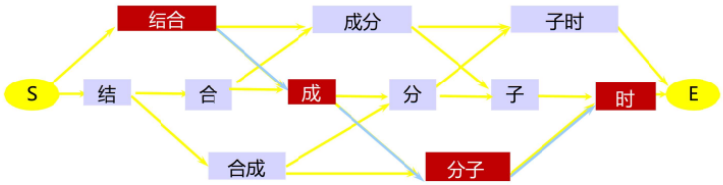
$$P'(w_i) = P'(w_{i-1}) \times P(w_i)$$

为方便计算，通常将概率转为路径代价，即取对数。

1. 对一个待分词的字串S，按照从左到右的顺序取出全部候选词$w_1, \ldots, w_i, \ldots, w_n$
2. 到词典中查出每个候选词的概率值$P(w_i)$，转换为代价$C(w_i)$，并记录每个候选词的全部左邻词
3. 按照$C'(w_i) = C'(w_{i-1}) + C(w_i)$计算累计代价，同时比较得到每个候选词的最佳左邻词
4. 如果当前w_n是字串S的尾词，且累积代价$C'(w_n)$最大，则w_n即为S的终点词
5. 从w_n开始，从右到左依次将最佳左邻词输出，即为S的分词结果

缺点：并不能解决所有交集型歧义，一般也无法解决组合型歧义

- 看待汉语词语切分问题的新视角：词图上的最优路径求解问题



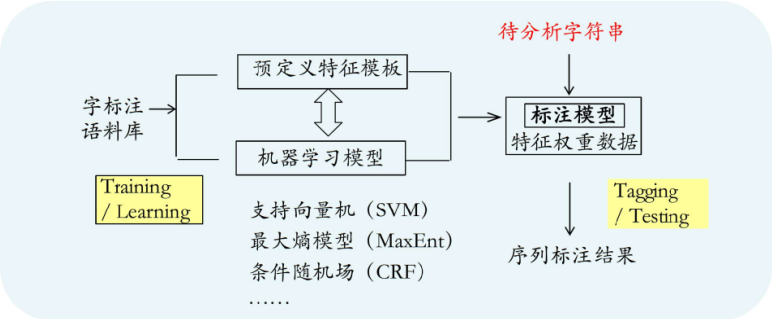
- 词图给出了一个字符串的全部切分可能性
- 分词任务：寻找一条起点S到终点E的最优路径

2.3.3 基于字序列标注的方法

字位标注法：分词可以看作是对字加“词位标注”的过程。根据字本身及其上下文特征来决定当前字的词位标注。

- 分词可以看做是对字加“词位标记”的过程
- “人”的词位分类示例：

B	E	M	S
词首	词尾	词中	独立词
人们	古人	小人国	听 人 说



优点：简单、鲁棒、效果好

- 能够平衡看待词表词和未登录词（未在词表中出现的词）的识别问题，因为都用统一字标注
- 学习架构上不必专门强调词表词信息，也不用设计特定未登录词（如人名、地名、机构名）识别模块，分词系统大大简化
- 字标注过程中，所有字根据预定义特征进行词位特性学习，获得一个概率模型。然后在待分字串上，根据字与字之间的结合紧密程度，得到一个词位标注结果

2.3.4 评价指标

$$\text{准确率}P = \frac{\#\text{切分正确}}{\#\text{切分结果所有分词}}$$

$$\text{召回率}R = \frac{\#\text{切分正确}}{\#\text{标准答案所有分词}}$$

$$\text{F评价}F_1 = \frac{2PR}{P+R}$$

3 语言模型

语言模型主要包括统计语言模型和神经语言模型。

3.1 统计语言模型

考虑语句的先验概率

$$p(s) = \prod_{i=1}^{m} p(w_i \mid w_1 \cdots w_{i-1}),\ p(w_1 \mid w_0) = p(w_1)$$

其中w_i可以是字、词、短语等，称为**统计基元**，通常用词代之。

为减少历史基元的个数，将$w_1 w_2 \cdots w_{i-1}$映射到等价类$S(w_1 w_2 \cdots w_{i-1})$，使等价类的数目远小于原来不同历史基元的数目，则有

$$p(w_i \mid w_1 \cdots w_{i-1}) = p(w_i \mid S(w_1 \cdots w_{i-1}))$$

n元文法(n-gram)模型

- 当$n = 1$时，出现在第i位上的基元w_i独立于历史，1元文法也被uni-gram或monogram
- $n = 2$时，2-gram(bi-gram)称为1阶马尔可夫链
- $n = 3$时，3-gram(tri-gram)称为2阶马尔可夫链，以此类推

实际操作加上句首<BOS>和句尾标记<EOS>。

❑ 举例：

给定句子：John read a book

增加标记：<BOS> John read a book <EOS>

Unigram: <BOS>, John, read, a, book, <EOS>

Bigram: (<BOS>John), (John read), (read a), (a book), (book <EOS>)

Trigram: (<BOS>John read), (John read a), (read a book), (a book <EOS>)

应用：

- 音字转换问题：给定拼音转为汉字串
- 汉语分词问题

对于n-gram，由最大似然估计求得

$$p(w_i \mid w_{i-n+1}^{i-1}) = f(w_i \mid w_{i-n+1}^{i-1}) = \frac{c(w_{i-1+1})}{\sum_{w_i} c(w_{i-n+1}^{i})}$$

其中$\sum_{w_i} c(w_{i-n+1}^{i})$是历史串$w_{i-n+1}^{i-1}$在给定语料中出现的次数，即$c(w_{i-n+1}^{i-1})$。

定义 10 (困惑度(perplexity)). 假定测试语料T由L个句子$(t_1, \ldots, t_L)$构成，则整个测试集的概率为

$$p(T) = \prod_{i=1}^{L} p(t_i)$$

为避免数据匮乏/稀疏导致的零概率问题，需要做数据平滑：调整最大似然估计的概率值，使零概率增值，使非零概率下调，消除零概率，改进模型的整体正确率。基本目标是测试样本语言模型的**困惑度越小越好**，基本约束是

$$\sum_{w_i} p(w_i \mid w_1, w_2, \ldots, w_{i-1}) = 1$$

- 加一法：

$$p(w_i \mid w_{i-1} = \frac{1 + c(w_{i-1}w_i)}{|V| + \sum_{w_i} c(w_{i-1}w_i)}$$

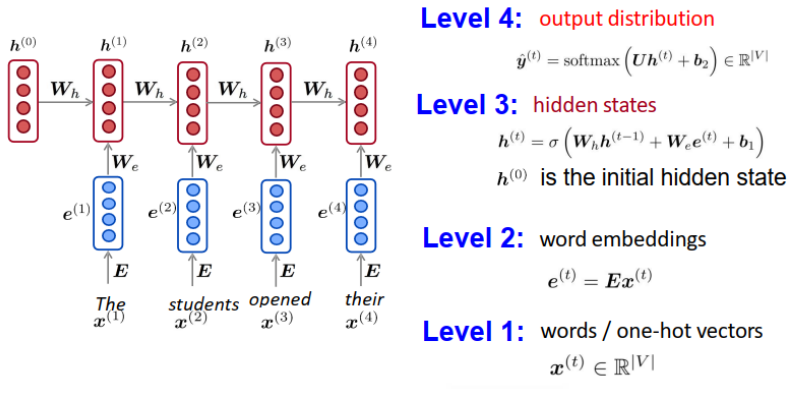
- 减值法/折扣法：将剩余概率量分配给未见概率
  - Good-Turing估计
  - 绝对减值法：从每个计数r中减去相同的量，剩余概率量由未见事件均分

$$p_r = \begin{cases} \frac{r-b}{N} & r > 0 \\ \frac{b(R-n_0)}{Nn_0} & r = 0 \end{cases}$$

其中n_0为样本中未出现的事件数目，$b \leq 1$为减去的常量，$b(R - n_0)/N$是由于减值而产生的剩余概率量

3.2 神经语言模型

对于固定窗口大小的神经网络显然是不适用的，需要采用一种能够处理任意长度输入的架构，即循环神经网络(RNN)，不同神经元共享相同的参数。



RNN的优点：

- 能够处理任意长度输入

- t时刻可以访问之前任意时刻的信息
- 对于较长输入，模型大小不变
- 权重共享

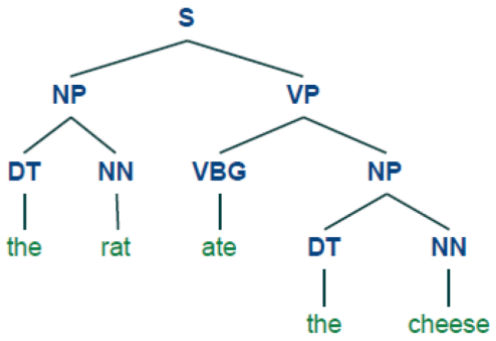
不足：循环计算过程慢，难以访问到距当前时刻较远的信息。

4 句法分析

句法分析是NLP中基础性工作，其分析句子的**句法结构**（主谓宾）和词汇间的**依存关系**（并列、从属），是语义分析、情感倾向、观点抽取等应用的基础。

4.1 上下文无关法

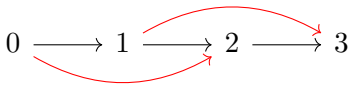
上下文无关法(Context-Free Gramma, CFG)由一系列规则组成，每条规则给出语言中某些符号可以被组织或排列在一起的方式。



比如上述第一条规则为$S \to NP + VP$，然后$NP \to DT + NN$，一直到叶子结点的单词为止。

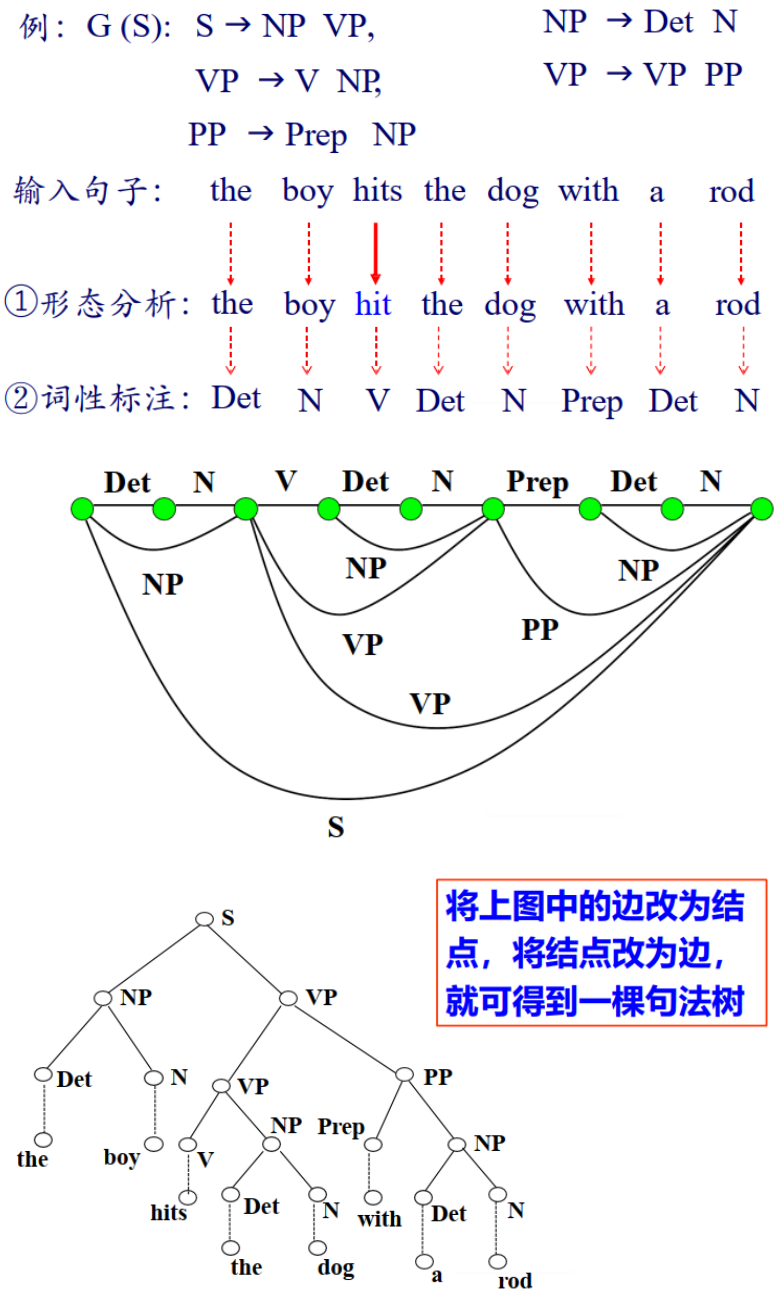
自底向上的线图分析法(chart parsing)[1]：

- 给定一组CFG规则：XP$\to \alpha_1 \cdots \alpha_n, n \geq 1$
- 给定一个句子的词性序列：$S = W_1 W_2 \cdots W_n$
- 构造一个线图：一组结点和边的集合



执行过程：查看任意相邻几条边上的词性串是否与某条规则的右部相同，若相同，则添加一条新的边跨越原来相应的边，新增加边上的标记为这条规则的头（左部）。重复这个过程，直到没有新的边产生。

[1] http://www.inf.ed.ac.uk/teaching/courses/icl/lectures/2006/earley-lec.pdf



CFG赋予语言层次化结构，但是根据CFG构建的语法分析树通常不止一个。

4.2 概率上下文无关法

为了应对产生多种语法分析结果的问题，引入概率上下文无关文法(PCFG)：为每棵树计算一个概率。

PCFG规则

$$\text{形式：} A \to \alpha \qquad [p]$$

$$\text{约束：} \sum_{\alpha} p(A \to \alpha) = 1$$

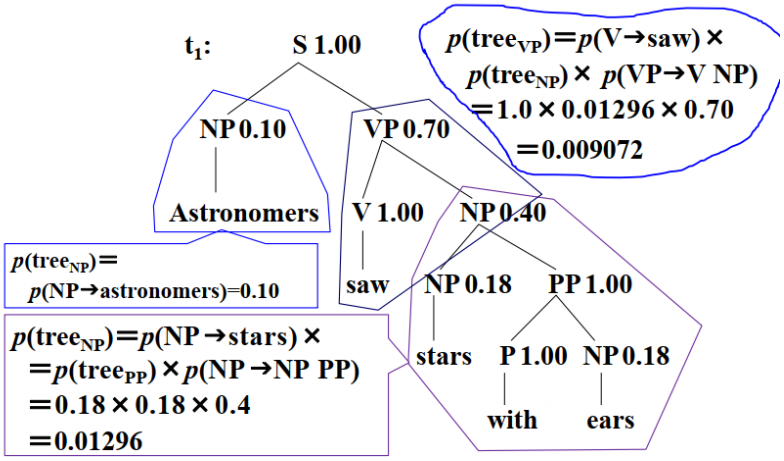
如

NP $\rightarrow$ DT NN $[p = 0.45]$

NN $\rightarrow$ leprechaun $[p = 0.0001]$

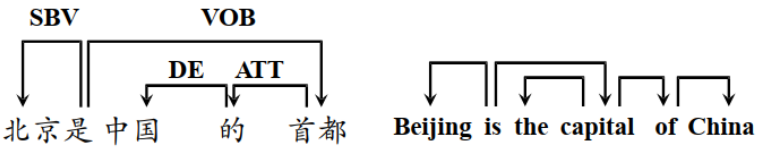
对于给定的语法分析树，可以计算其概率

$$P(T) = \prod_{i=1}^{n} P(RHS_i \mid LHS_i)$$



4.3 依存句法分析

定义 11 (依存). 依存是指词与词之间支配与被支配的关系，这种关系是不对等的，有方向的。处于支配地位的为支配者*(governor)*，被支配地位的为从属者*(modifier)*。



(a) 有向图-1　　(b) 有向图-2

两个有向图用带有方向的弧(或称边)来表示两个成分之间的依存关系，支配者在有向弧的发出端，被支配者在箭头端，我们通常说被支配者依存于支配者。

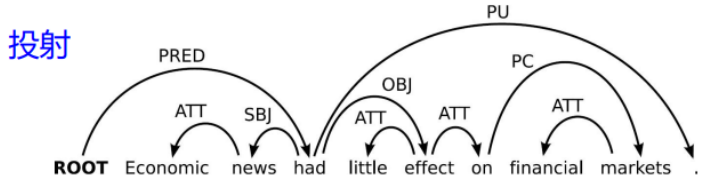
依存语法的优势：

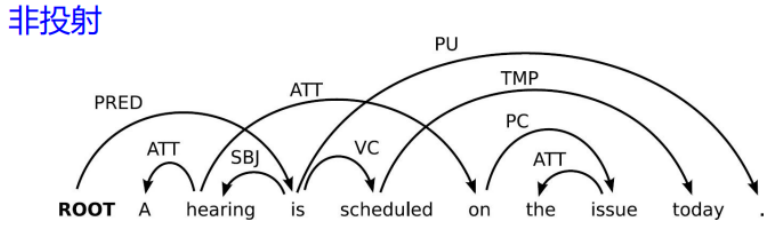
- 依存关系和实际的语义关系比较接近，有助于对句子的语义方面的理解
- 定义相对比较简单，有助于高效率的句法分析
- 能够有效建模**长距离依赖关系**，依存句法更适合词序列比较自由、灵活的语言

对依存图和依存树有约束：

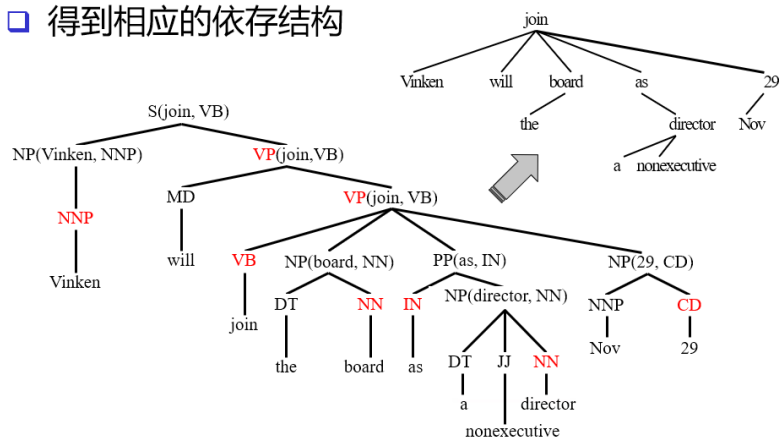
- 一个句子只有一个独立的成分
- 句子的其他成分都从属于某一成分
- 任何一成分都不能依存于两个或多个成分
- 如果成分A直接从属于成分B，而成分C在句子中位于A和B之间，那么，成分C或者从属于A，或者从属于B，或者从属于A和B之间的某一成分。

对应着对依存图和依存树的形式约束：单一父节点、连通、无环、可投射，由此来保证句子的依存分析结果是一棵有根的树结构。





依存结构表达的信息和短语结构句法树不一样，可以表达更长距离的信息依存关系。



短语结构可以转成依存结构：

- 定义中心词抽取规则，产生中心词表
- 根据中心词表，为每个节点选择中心子节点
- 将非中心子节点的中心词依存到中心子节点的中心词上，得到相应依存结构

5 问答与对话

定义 12 (问答系统). 输入：自然语言的问句，而非关键词的组合
输出：直接答案，而非文档集合

优点:

- 相对于基于知识推理的问答系统而言：不受知识库规模限制，不受领域限制，更加接近真实应用需求
- 相对于搜索引擎而言：问答式检索系统接受的是自然语言形式的提问，由于自然语言处理技术的应用，对用户意图的把握更加准确，呈现给用户的答案更加准确

缺点:目前问答式检索系统仅能处理有限的简单问题，如Factoid问题等

问答式检索方法：

- 信息检索+信息抽取
- 信息检索+模式匹配：离线阶段获取答案模式，在线阶段首先判断当前提问属于哪一类，然后使用这类提问的所有模式来抽取候选答案
- 信息检索+自然语言处理技术
- 基于统计翻译模型的问答技术：把提问句看作答案句在同一语言内的一种翻译

新方法基于深度学习：把回答问题的过程看作一个黑盒，通过复杂神经网络和超大规模数据集训练出一个拟合能力强大的模型。

对话管理（状态跟踪和学习）方法：

- 有限状态机(finite state machine)：太受限了
- 基于框架的方法(frame-based)：
  - 使用框架的结构指导对话过程：机器根据框架进行提问，人也根据框架进行回复
  - 问答过程就是一个槽-值填充过程：所有槽都填满了，就可以通过信息系统查询
  - 用户可以一次性回答多个系统问题
- 统计方法(MDP)：状态、动作、目标

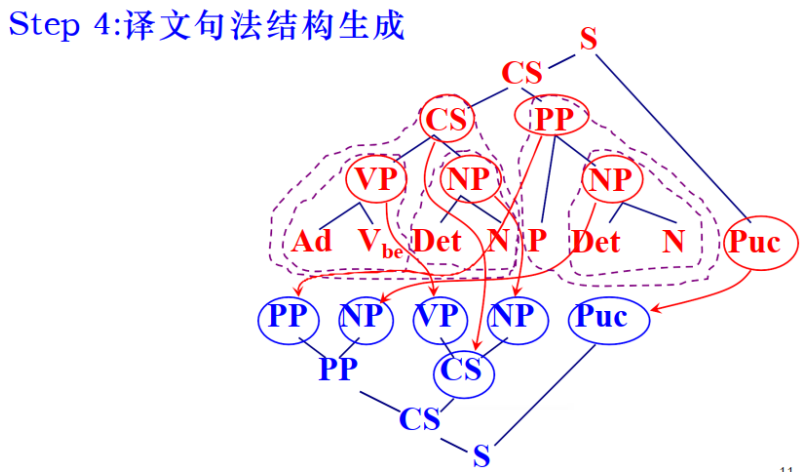
6 机器翻译

基本翻译方法

- 直接转换
- 基于规则的翻译方法
- 基于中间语言
- 基于语料库：基于事例、统计翻译、神经网络翻译

6.1 基于规则的翻译方法

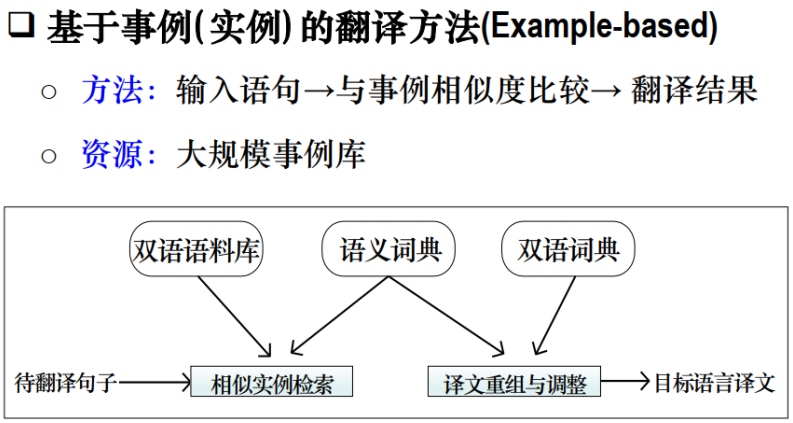
1. 对源语言句子进行词法分析
2. 句法/语义分析
3. 源语言句子结构到译文结构的转换
4. 译文句法结构生成
5. 源语言词汇到译文词汇转换
6. 译文词法选择与生成



优点：保持原文结构、规范语句

弱点：规则由人工编写，工作量大，主观性强，不利于系统扩充，灵活性低（对非规范语言现象无法处理）

6.2 基于语料库的翻译方法



优点：不需源语言句子符合语法规定，也不许对源语言句子做深入分析

缺点：两个不同句子之间相似性难以把握，难以处理陌生语言现象，且事例库庞大起来难以搜索

6.3 统计翻译

6.3.1 基本方法

- 源语言：$S = s_1^m = s_1 s_2 \cdots s_m$
- 目标语言：$T = t_1^l = t_1 t_2 \cdots t_n$

通过最大后验求解

$$\arg\max_T P(T \mid S) = \arg\max_T P(S \mid T) P(T)$$

其中$P(S \mid T)$为翻译模型(TM)，确定了单词和词语如何被翻译(fidelity)，从平行语料中学习； $P(T)$为语言模型(LM)，确定怎么写出好的目标语言的句子(fluency)，从单语语料中学习。

三个关键问题：

- 估计语言模型概率$P(T)$
- 估计翻译模型概率$P(S \mid T)$：关键问题是怎样定义目标语言句子中的词与源语言句子中的词之间的对应关系，基本原理是对位(alignment)模型
- 快速搜索最大值解

6.3.2 基于短语的翻译模型

基于词的翻译模型很难消除歧义，很难处理一对多、多对一问题。以短语为基本翻译单元，遵循短语划分、短语翻译、短语调序的步骤。

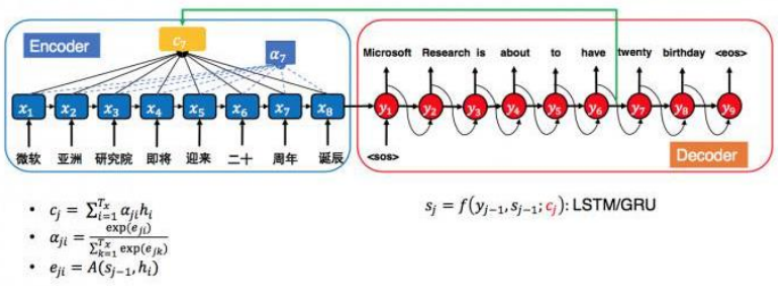
定义 13 (对齐一致性). S^j中每个词S_k，若$(k, k') \in A$，则$i' \le k' \le j'$，T_i^j中每个词$T_{t'}$，若$(t, t') \in A$，则$i \le t \le j$。

关键问题是学习短语翻译规则（双语句对词语对齐、短语翻译规则抽取）、估计短语翻译概率。

6.4 神经机器翻译

统计机器翻译都是人工设定的模块和特征，可解释性高、模块定制化、错误追踪，但是数据稀疏、不擅长复杂结构、依赖先验知识。

分布式的语义表示是统计机器翻译到机器翻译的核心



BLEU(BiLingual Evaluation Understudy)评价方法：统计同时出现在系统译文和参考译文中的n元词个数，最后把匹配到n元词的数目除以系统译文的n元词数目，得到评测结果。