译者序

计算机对自然语言的研究和处理需要经过以下4个过程：

1. 把需要研究的问题在语言学上加以形式化，使之能以一定的数学形式，严密而规整地表示出来；
2. 把这种严密而规整的数学形式表示为算法，使之在计算上形式化；
3. 根据算法编写计算机程序，使之在计算机上加以实现；
4. 对于所建立的自然语言处理系统进行评测，使之不断地改进质量和性能，以满足用户的需要。

Bill Manaris对自然语言的定义：

自然语言处理可以定义为研究在人与人交际中以及在人与计算机交际中的语言问题的一门学科。自然语言处理要研制表示语言的能力（linguistic competence）和语言应用（linguistic performance）的模型，建立计算框架来实现这样的语言模型，提出相应的方法来不断地完善这样的语言模型，根据这样的语言模型设计各种实现系统，并探讨这些实用系统的评测技术。

自然语言处理可以定义为：研究在人与人交际中以及在人与计算机交际中的语言问题的一门学科。

自然语言处理要研制：语言的表示能力（linguistic competence）和语言的应用模型（linguistic performance），通过建立计算框架来实现这样的语言模型，提出相应的方法来不断地完善这样的语言模型，设计各种系统来实现这样的语言模型，并探讨这些实用系统的评测技术。

建立自然语言处理模型需要的知识：

* 声学和韵律学的知识：描述语言的节奏、语调和声调的规律，说明语音是怎样形成音位的。
* 音位学的知识：描述音位的结合规律，说明音位是如何形成语素的。
* 形态学的知识：描述语素的结合规则，说明语素是怎样形成单词的。
* 词汇学的知识：描述词汇系统的规律，说明单词本身固有的语义特性和语法特性。
* 句法学的知识：描述单词（或词组）之间的结构规则，说明单词（或词组）是怎样组成句子的。
* 语义学的知识：描述句子中各个成分之间的语义关系，这样的语义关系是与情景无关的，说明怎样从构成句子的各个成分中推导出整个句子的语义的。
* 话语分析的知识：描述句子与句子之间的结构规律，说明是怎样由句子形成话语或对话的。
* 语用学的知识：描述与情景有关的情景语义，说明是怎样推导出句子具有的与周围话语有关的各种含义的。
* 外界世界的常识性知识：描述关于语言使用者与语言使用环境的一般性常识，例如：语言使用者的信念和目的，说明是怎样推导出这样的信念和目的内在结构的。

P8 图0.1 自然语言处理系统中的知识平面：

→（话语）→音位学平面→（音位）→形态学平面→（语素）→词汇学平面→（单词）→句法学平面→（句法结构）→语义学平面（意义表示）→（语义解释）→语用学平面→（语用解释）

[译者序 1](#_Toc29395144)

[第1章 导论 22](#_Toc29395145)

[1.1 语音与语言处理中的知识 22](#_Toc29395146)

[1.2 歧义 23](#_Toc29395147)

[1.3 模型和算法 23](#_Toc29395148)

[1.4 语言、思维和理解 25](#_Toc29395149)

[1.5 学科现状与近期发展 25](#_Toc29395150)

[1.6 语音和语言处理简史 25](#_Toc29395151)

[1.7 小结 25](#_Toc29395152)

[第一部分 词汇的计算机处理 27](#_Toc29395153)

[第2章 正则表达式 与 自动机 27](#_Toc29395154)

[2.1 正则表达式 27](#_Toc29395155)

[基本的正则表达式模式 27](#_Toc29395156)

[正则表达式的基本算符 28](#_Toc29395157)

[正则表达式的简单例子和复杂例子 28](#_Toc29395158)

[正则表达式的高级算符 28](#_Toc29395159)

[正则表达式的实际操作 29](#_Toc29395160)

[2.2 有限状态自动机（FSA） 29](#_Toc29395161)

[FSA的表示方式 30](#_Toc29395162)

[FSA中的形式语言 30](#_Toc29395163)

[非确定的FSA（NFSA） 31](#_Toc29395164)

[NFSA接收符号串 31](#_Toc29395165)

[FSA识别就是搜索 31](#_Toc29395166)

[DFSA与NFSA之间的关系 32](#_Toc29395167)

[2.3 正则语言与有限状态自动机 32](#_Toc29395168)

[2.4 小结 33](#_Toc29395169)

[第3章 词 与 转录机 34](#_Toc29395170)

[3.1 英语形态学概观 34](#_Toc29395171)

[屈折形态学（inflectional morphology） 35](#_Toc29395172)

[派生形态学（derivational morphology） 36](#_Toc29395173)

[附着（cliticization） 36](#_Toc29395174)

[非毗连形态学 36](#_Toc29395175)

[一致关系（agreement） 37](#_Toc29395176)

[3.2 有限状态形态剖析 37](#_Toc29395177)

[3.3 有限状态词表的建造 37](#_Toc29395178)

[3.4 有限状态转录机（FST） 37](#_Toc29395179)

[定序转录机和确定性 38](#_Toc29395180)

[3.5 用于形态剖析的有限状态转录机 39](#_Toc29395181)

[3.6 转录机和正词法规则 39](#_Toc29395182)

[3.7 结合有限状态转录机的词表与规则 39](#_Toc29395183)

[3.8 Porter词干处理器（不使用词表） 40](#_Toc29395184)

[3.9 单词和句子的词例还原 40](#_Toc29395185)

[中文的自动切词 40](#_Toc29395186)

[3.10 拼写错误的检查与更正 40](#_Toc29395187)

[3.11 最小编辑距离（minimum edit distance） 41](#_Toc29395188)

[3.12 人的形态处理方式 41](#_Toc29395189)

[3.13 小结 41](#_Toc29395190)

[第4章 N 元语法 42](#_Toc29395191)

[4.1 语料库中单词数目的计算 43](#_Toc29395192)

[4.2 简单的（非平滑的）N元语法 43](#_Toc29395193)

[4.3 训练集与测试集 44](#_Toc29395194)

[未知词：开放词汇和封闭词汇 44](#_Toc29395195)

[4.4 N元语法的评测：困惑度 44](#_Toc29395196)

[4.5 N元语法的平滑算法 45](#_Toc29395197)

[Laplace平滑 45](#_Toc29395198)

[Good-Turing打折法 45](#_Toc29395199)

[Good-Turing估计 46](#_Toc29395200)

[4.6 N元语法的插值法 46](#_Toc29395201)

[4.7 N元语法的回退法 46](#_Toc29395202)

[4.8 实际问题：工具包和数据格式 46](#_Toc29395203)

[4.9 语言模型建模中的高级专题 46](#_Toc29395204)

[Kneser-Ney平滑法 46](#_Toc29395205)

[基于类别的N元语法 47](#_Toc29395206)

[语言模型的自适应 47](#_Toc29395207)

[长距离信息的使用 47](#_Toc29395208)

[4.10 信息论背景 48](#_Toc29395209)

[用于比较模型的交叉熵 48](#_Toc29395210)

[4.11 英语的熵和熵率均衡性 48](#_Toc29395211)

[4.12 小结 48](#_Toc29395212)

[第5章 词类标注 49](#_Toc29395213)

[5.1 英语词的分类 50](#_Toc29395214)

[5.2 英语的标记集 52](#_Toc29395215)

[5.3 词类标注 52](#_Toc29395216)

[5.4 基于规则的词类标注算法 52](#_Toc29395217)

[5.5 基于HMM模型的词类标注算法 53](#_Toc29395218)

[寻找最可能的标记序列 53](#_Toc29395219)

[HMM标注算法的形式化说明 53](#_Toc29395220)

[使用Viterbi算法进行HMM标注 53](#_Toc29395221)

[使用三元语法的HMM标注算法 54](#_Toc29395222)

[5.6 基于转换的词类标注算法 54](#_Toc29395223)

[TBL规则的应用 54](#_Toc29395224)

[TBL规则的学习 54](#_Toc29395225)

[5.7 评测和错误分析 55](#_Toc29395226)

[5.8 词类标注中的高级专题 56](#_Toc29395227)

[标记的不确定性 56](#_Toc29395228)

[未知词 56](#_Toc29395229)

[高度屈折语和黏着语的词类标注 57](#_Toc29395230)

[标注算法的结合 57](#_Toc29395231)

[5.9 拼写中的噪声信道模型 58](#_Toc29395232)

[5.10 小结 58](#_Toc29395233)

[第6章 HMM 与 MaxEnt 59](#_Toc29395234)

[6.1 马尔可夫链 60](#_Toc29395235)

[6.2 隐马尔可夫模型（HMM） 61](#_Toc29395236)

[6.3 似然度问题：向前算法 62](#_Toc29395237)

[6.4 解码问题：Viterbi算法 62](#_Toc29395238)

[6.5 HMM的参数学习问题：向前——向后算法 62](#_Toc29395239)

[6.6 最大熵模型的背景知识 63](#_Toc29395240)

[线性回归 63](#_Toc29395241)

[逻辑回归 64](#_Toc29395242)

[逻辑回归（分类） 64](#_Toc29395243)

[逻辑回归（训练） 64](#_Toc29395244)

[6.7 最大熵模型（MEMM） 65](#_Toc29395245)

[6.8 最大熵马尔可夫模型 66](#_Toc29395246)

[6.9 小结 67](#_Toc29395247)

[语音处理 68](#_Toc29395248)

[第7章 语音学 68](#_Toc29395249)

[7.1 言语语音和语音标音法 68](#_Toc29395250)

[7.2 发音语音学 69](#_Toc29395251)

[发音器官 69](#_Toc29395252)

[辅音：发音部位 69](#_Toc29395253)

[辅音：发音方法 70](#_Toc29395254)

[元音 71](#_Toc29395255)

[音节 71](#_Toc29395256)

[7.3 单位范畴与发音变异 72](#_Toc29395257)

[语音特征 72](#_Toc29395258)

[语音变异的预测 73](#_Toc29395259)

[影响语音变异的因素 73](#_Toc29395260)

[7.4 声学语音学和信号 73](#_Toc29395261)

[波 73](#_Toc29395262)

[语音的声波 73](#_Toc29395263)

[频率（音高）与振幅（响度） 74](#_Toc29395264)

[音子的波形解释 74](#_Toc29395265)

[声谱和频域 74](#_Toc29395266)

[声源滤波器模型 75](#_Toc29395267)

[7.5 语音资源 75](#_Toc29395268)

[7.6 发音音系学与姿态音系学 75](#_Toc29395269)

[7.7 小结 75](#_Toc29395270)

[第8章 语音合成 77](#_Toc29395271)

[8.1 文本归一化 77](#_Toc29395272)

[8.2 语音分析 78](#_Toc29395273)

[8.3 韵律分析 79](#_Toc29395274)

[8.4 双音子波形合成 81](#_Toc29395275)

[8.5 单元选择（波形）合成 82](#_Toc29395276)

[8.6 评测 83](#_Toc29395277)

[第9章 语音自动识别 83](#_Toc29395278)

[9.1 语音识别的总体结构 84](#_Toc29395279)

[9.2 HMM在语音识别中的应用 84](#_Toc29395280)

[9.3 特征抽取：MFCC矢量 85](#_Toc29395281)

[预加重 86](#_Toc29395282)

[加窗 87](#_Toc29395283)

[DFT（离散傅里叶变换） 87](#_Toc29395284)

[Mel滤波器组和对数 87](#_Toc29395285)

[倒谱：逆向DFT 88](#_Toc29395286)

[Delta特征与能量特征 88](#_Toc29395287)

[9.4 声学似然度的计算 89](#_Toc29395288)

[矢量量化 89](#_Toc29395289)

[概率密度函数 89](#_Toc29395290)

[9.5 词典和语言模型 90](#_Toc29395291)

[9.6 搜索与解码 90](#_Toc29395292)

[9.7 嵌入式训练 91](#_Toc29395293)

[9.8 评测：词错误率 92](#_Toc29395294)

[9.9 小结 92](#_Toc29395295)

[第10章 语音识别：高级专题 93](#_Toc29395296)

[10.1 多遍解码：N-最佳表和格 93](#_Toc29395297)

[10.2 A\*解码算法（“栈”解码算法） 95](#_Toc29395298)

[10.3 依赖于上下文的声学模型：三音子 96](#_Toc29395299)

[10.4 分辨训练（判别） 97](#_Toc29395300)

[10.5 语音变异的建模 97](#_Toc29395301)

[10.6 元数据：边界、标点符号和不流利的现象 99](#_Toc29395302)

[10.7 人的语音识别 100](#_Toc29395303)

[10.8 小结 101](#_Toc29395304)

[第11章 计算音系学 102](#_Toc29395305)

[11.1 有限状态音系学 102](#_Toc29395306)

[11.2 有限状态音系学：高级专题 103](#_Toc29395307)

[11.3 计算优选理论（Optimality Theory，OT） 103](#_Toc29395308)

[11.4 音节切分 104](#_Toc29395309)

[11.5 音位规则和形态规则的机器学习 105](#_Toc29395310)

[11.6 小结 106](#_Toc29395311)

[句法处理 108](#_Toc29395312)

[第12章 英语的形式语法 108](#_Toc29395313)

[12.1 组成性 109](#_Toc29395314)

[12.2 上下文无关语法 109](#_Toc29395315)

[12.3 英语的一些语法规则 111](#_Toc29395316)

[句子一级的结构 111](#_Toc29395317)

[句子和子句 112](#_Toc29395318)

[名词短语 112](#_Toc29395319)

[一致关系 113](#_Toc29395320)

[动词短语和次范畴化 114](#_Toc29395321)

[助动词 114](#_Toc29395322)

[并列关系 115](#_Toc29395323)

[12.4 树库 115](#_Toc29395324)

[12.5 语法等价与范式 116](#_Toc29395325)

[12.6 有限状态语法和上下文无关语法 117](#_Toc29395326)

[12.7 依存语法 117](#_Toc29395327)

[12.8 口语的句法 118](#_Toc29395328)

[12.9 人的语法处理 118](#_Toc29395329)

[12.10 小结 118](#_Toc29395330)

[第13章 句法剖析 120](#_Toc29395331)

[13.1 剖析就是搜索 120](#_Toc29395332)

[13.2 歧义 121](#_Toc29395333)

[13.3 面对歧义的搜索 122](#_Toc29395334)

[13.4 动态规划剖析算法 122](#_Toc29395335)

[CKY剖析算法 122](#_Toc29395336)

[Earley剖析算法 123](#_Toc29395337)

[线图剖析算法 124](#_Toc29395338)

[13.5 局部剖析 124](#_Toc29395339)

[基于规则的有限状态组块分析 125](#_Toc29395340)

[基于机器学习的组块分析 125](#_Toc29395341)

[组块分析系统的评测 126](#_Toc29395342)

[13.6 小结 126](#_Toc29395343)

[第14章 统计剖析 127](#_Toc29395344)

[14.1 概率上下文无关语法（PCFG） 127](#_Toc29395345)

[14.2 PCFG的概率CKY剖析 128](#_Toc29395346)

[14.3 PCFG的规则的概率的学习 128](#_Toc29395347)

[14.4 PCFG存在的问题 128](#_Toc29395348)

[14.5 使用分离非终极符号来给结构依存建模 128](#_Toc29395349)

[14.6 概率词汇化的CFG来解决词汇依存建模问题 129](#_Toc29395350)

[14.7 剖析器的评测 129](#_Toc29395351)

[14.8 判别（分辨）再排序 130](#_Toc29395352)

[14.9 基于剖析器的语言模型 130](#_Toc29395353)

[14.10 人的剖析 131](#_Toc29395354)

[14.11 小结 131](#_Toc29395355)

[第15章 特征 与 合一 132](#_Toc29395356)

[15.1 特征结构 132](#_Toc29395357)

[15.2 特征结构的合一 133](#_Toc29395358)

[15.3 语法中的特征结构 134](#_Toc29395359)

[15.4 合一的实现 136](#_Toc29395360)

[合一的数据结构 136](#_Toc29395361)

[合一的算法 137](#_Toc29395362)

[15.5 带有合一约束的剖析 137](#_Toc29395363)

[Earley剖析器结合合一的剖析 137](#_Toc29395364)

[基于合一的剖析 137](#_Toc29395365)

[15.6 类型与继承 138](#_Toc29395366)

[15.7 小结 139](#_Toc29395367)

[第16章 语言和复杂性 140](#_Toc29395368)

[16.1 Chomsky层级 140](#_Toc29395369)

[16.2 正则语言的判定 141](#_Toc29395370)

[抽吸引理（pumping lemma） 141](#_Toc29395371)

[16.3 自然语言是上下文无关的吗？ 142](#_Toc29395372)

[16.4 计算复杂性和人的语言处理 142](#_Toc29395373)

[16.5 小结 143](#_Toc29395374)

[语义处理 与 语用处理 145](#_Toc29395375)

[第17章 什么是意义？ 145](#_Toc29395376)

[17.1 意义表示的计算要求 146](#_Toc29395377)

[17.2 模型论语义学 147](#_Toc29395378)

[17.3 一阶逻辑（First-Order Logic，FOL） 148](#_Toc29395379)

[17.4 事件与状态的表示 150](#_Toc29395380)

[时间表示 151](#_Toc29395381)

[体 151](#_Toc29395382)

[17.5 描述逻辑 152](#_Toc29395383)

[描述逻辑 153](#_Toc29395384)

[逻辑推理 153](#_Toc29395385)

[Web本体语言和语义网络 154](#_Toc29395386)

[17.6 意义的具体化 与 情境表示方法 154](#_Toc29395387)

[17.7 小结 155](#_Toc29395388)

[第18章 计算语义学 155](#_Toc29395389)

[18.1 句法驱动的语义分析 156](#_Toc29395390)

[18.2 句法规则的语义扩充 156](#_Toc29395391)

[18.3 量词辖域歧义及非确定性 157](#_Toc29395392)

[存储与检索方法 157](#_Toc29395393)

[基于约束的方法 158](#_Toc29395394)

[18.4 基于合一的语义分析方法 159](#_Toc29395395)

[18.5 语义与 Earley 分析器的集成 159](#_Toc29395396)

[18.6 成语（俗语）和组成性 159](#_Toc29395397)

[18.7 小结 160](#_Toc29395398)

[第19章 词汇语义学 160](#_Toc29395399)

[19.1 词义 161](#_Toc29395400)

[19.2 含义间的关系 162](#_Toc29395401)

[同义关系（synonymy） 162](#_Toc29395402)

[反义关系（antonymy） 162](#_Toc29395403)

[上下位关系（hypernymy） 163](#_Toc29395404)

[语义场（semantic field） 163](#_Toc29395405)

[19.3 WordNet：词汇关系信息库 164](#_Toc29395406)

[19.4 事件参与者 164](#_Toc29395407)

[题旨（主题的）角色（Thematic Roles） 164](#_Toc29395408)

[因素交替 164](#_Toc29395409)

[题旨（主题）角色存在的问题 165](#_Toc29395410)

[命题树库（The Proposition Bank，PropBank） 165](#_Toc29395411)

[框架网络（FrameNet） 165](#_Toc29395412)

[选择限制（selectional restrictions） 166](#_Toc29395413)

[19.5 基元分解 166](#_Toc29395414)

[19.6 隐喻（metaphor） 167](#_Toc29395415)

[19.7 小结 167](#_Toc29395416)

[第20章 计算词汇语义学 168](#_Toc29395417)

[20.1 词义排歧（WSD）：综述 168](#_Toc29395418)

[20.2 有监督词义排歧 169](#_Toc29395419)

[监督学习的特征抽取 169](#_Toc29395420)

[朴素贝叶斯分类器 和 决策表分类器 170](#_Toc29395421)

[20.3 WSD评价：方法、基准线、上限 170](#_Toc29395422)

[20.4 WSD：字典方法和同义词库方法 171](#_Toc29395423)

[Lesk算法 171](#_Toc29395424)

[选择限制和选择优先度 171](#_Toc29395425)

[20.5 半监督WSD（最小化监督的WSD）：自举法 172](#_Toc29395426)

[20.6 词语相似度：语义字典方法 173](#_Toc29395427)

[20.7 词语相似度：分布方法 175](#_Toc29395428)

[定义词语的共现向量 175](#_Toc29395429)

[度量与上下文的联系 175](#_Toc29395430)

[定义两个向量之间的相似度（向量计算） 176](#_Toc29395431)

[评价分布式词语相似度 176](#_Toc29395432)

[20.8 词语关系：下位关系及其他关系 176](#_Toc29395433)

[20.9 语义角色标注 178](#_Toc29395434)

[20.10 无监督语义排歧 179](#_Toc29395435)

[20.11 小结 179](#_Toc29395436)

[第21章 计算话语学 180](#_Toc29395437)

[21.1 话语分割 181](#_Toc29395438)

[无监督话语分割 181](#_Toc29395439)

[有监督话语分割 183](#_Toc29395440)

[话语分割的评价 183](#_Toc29395441)

[21.2 文本连贯性 183](#_Toc29395442)

[修辞结构理论 184](#_Toc29395443)

[自动指派连贯关系 185](#_Toc29395444)

[21.3 指代消解 186](#_Toc29395445)

[21.4 指代现象 187](#_Toc29395446)

[指示语的五种类型 187](#_Toc29395447)

[信息状态 188](#_Toc29395448)

[21.5 代词回指消解所使用的特征 189](#_Toc29395449)

[用来过滤潜在指代对象的特征 189](#_Toc29395450)

[代词解释中的优先关系（显著性） 189](#_Toc29395451)

[21.6 代词回指消解的三种算法 190](#_Toc29395452)

[21.7 共指消解 191](#_Toc29395453)

[21.8 共指消解的评价 191](#_Toc29395454)

[21.9 基于推理的连贯判定 192](#_Toc29395455)

[21.10 所指的心理学研究 192](#_Toc29395456)

[21.11 小结 193](#_Toc29395457)

[自然语言处理的应用 195](#_Toc29395458)

[第22章 信息抽取 195](#_Toc29395459)

[22.1 命名实体识别（NER） 196](#_Toc29395460)

[NER中的歧义 196](#_Toc29395461)

[基于序列标注的NER 196](#_Toc29395462)

[NER的评价 197](#_Toc29395463)

[实用的NER框架 197](#_Toc29395464)

[22.2 关系识别和分类 197](#_Toc29395465)

[用于关系分析的有监督学习方法 198](#_Toc29395466)

[用于关系分析的弱监督学习方法（？） 199](#_Toc29395467)

[关系分析系统的评价 199](#_Toc29395468)

[22.3 时间和事件处理 199](#_Toc29395469)

[时间表达式的识别 200](#_Toc29395470)

[时间的归一化 201](#_Toc29395471)

[事件检测和分析 202](#_Toc29395472)

[TimeBank（语料库） 202](#_Toc29395473)

[22.4 模板填充 203](#_Toc29395474)

[模板填充的统计序列标注方法 203](#_Toc29395475)

[有限状态机模板填充系统 204](#_Toc29395476)

[22.5 生物医学信息的抽取 204](#_Toc29395477)

[22.6 小结 204](#_Toc29395478)

[第23章 问答 和 摘要 205](#_Toc29395479)

[23.1 信息检索 206](#_Toc29395480)

[向量空间模型 206](#_Toc29395481)

[词语权重计算 207](#_Toc29395482)

[词语选择和建立 207](#_Toc29395483)

[信息检索系统的评测 207](#_Toc29395484)

[同形关系、多义关系和同义关系 208](#_Toc29395485)

[改进用户查询的方法 208](#_Toc29395486)

[23.2 事实性问答 209](#_Toc29395487)

[问题处理 210](#_Toc29395488)

[段落检索 210](#_Toc29395489)

[答案处理 211](#_Toc29395490)

[事实性答案的评价 213](#_Toc29395491)

[23.3 摘要 213](#_Toc29395492)

[23.4 单文档摘要 215](#_Toc29395493)

[无监督的内容选择 215](#_Toc29395494)

[基于修辞分析的无监督摘要 216](#_Toc29395495)

[有监督的内容选择 216](#_Toc29395496)

[句子简化 216](#_Toc29395497)

[23.5 多文档摘要 217](#_Toc29395498)

[多文档摘要的内容选择 217](#_Toc29395499)

[多文档摘要的信息排序 217](#_Toc29395500)

[多文档摘要的句子实现 218](#_Toc29395501)

[23.6 主题摘要和问答 218](#_Toc29395502)

[23.7 摘要的评价 219](#_Toc29395503)

[23.8 小结 219](#_Toc29395504)

[第24章 对话 和 会话智能代理 220](#_Toc29395505)

[24.1 人类会话的属性 221](#_Toc29395506)

[话轮与话轮的转换 221](#_Toc29395507)

[语言作为行动：言语行为 222](#_Toc29395508)

[语言作为共同行动：对话的共同基础 223](#_Toc29395509)

[会话结构 224](#_Toc29395510)

[会话隐含 224](#_Toc29395511)

[24.2 基本的对话系统 225](#_Toc29395512)

[ASR（自动语音识别）组件 225](#_Toc29395513)

[NLU（自然语言理解）组件 225](#_Toc29395514)

[生成和TTS组件（自然语言生成器） 226](#_Toc29395515)

[对话管理器 227](#_Toc29395516)

[任务管理器的错误处理：确认和拒绝 228](#_Toc29395517)

[24.3 VoiceXML 229](#_Toc29395518)

[24.4 对话系统的设计与评价 229](#_Toc29395519)

[对话系统的设计 229](#_Toc29395520)

[对话系统的评价 230](#_Toc29395521)

[24.5 信息状态和对话行为 231](#_Toc29395522)

[使用对话行为 232](#_Toc29395523)

[解释对话行为 232](#_Toc29395524)

[检测纠正行为 233](#_Toc29395525)

[生成对话行为：确认和拒绝 234](#_Toc29395526)

[24.6 马尔可夫决策过程架构 234](#_Toc29395527)

[24.7 基于规划的对话行为 235](#_Toc29395528)

[规划推理的解释和生成 235](#_Toc29395529)

[对话的意图结构 235](#_Toc29395530)

[24.8 小结 236](#_Toc29395531)

[第25章 机器翻译 237](#_Toc29395532)

[25.1 机器翻译面临的困难 238](#_Toc29395533)

[类型学 238](#_Toc29395534)

[其他的结构差异 239](#_Toc29395535)

[词汇差异（翻译中词汇的问题） 240](#_Toc29395536)

[25.2 经典的机器翻译方法与Vauquois三角形 240](#_Toc29395537)

[直接翻译 240](#_Toc29395538)

[转换方法 241](#_Toc29395539)

[直接翻译与转换方法的融合 241](#_Toc29395540)

[中间语言的思想与使用意义 242](#_Toc29395541)

[25.3 统计机器翻译 243](#_Toc29395542)

[25.4 基于短语的翻译模型 243](#_Toc29395543)

[25.5 翻译中的对齐 243](#_Toc29395544)

[IBM模型1 244](#_Toc29395545)

[HMM对齐模型 244](#_Toc29395546)

[25.6 对齐模型的训练 244](#_Toc29395547)

[训练对齐模型的EM算法 244](#_Toc29395548)

[25.7 基于短语的翻译模型的对齐 244](#_Toc29395549)

[25.8 基于短语的统计机器翻译的解码 245](#_Toc29395550)

[25.9 机器翻译的评价 245](#_Toc29395551)

[人工评价 246](#_Toc29395552)

[自动评价：BLEU 246](#_Toc29395553)

[25.10 机器翻译的句法模型 246](#_Toc29395554)

[25.11 IBM模型3和繁衍度 247](#_Toc29395555)

[IBM模型3的训练 247](#_Toc29395556)

[25.12 机器翻译的MEMM模型 247](#_Toc29395557)

[25.13 小结 248](#_Toc29395558)

# 导论

会话代理（conversational agents）或者对话系统（dialogue systems）的组成部分

* 语言输入
  + 自动语音识别（automatic speech recognition）
  + 自然语言理解（natural language understanding）
* 语言输出
  + 自然语言生成（natural language generation）
  + 语音合成（speech synthesis）

推理（inference）：对于需要抽取嵌入到网页的其他文本中的信息才能回答的那些更加复杂的问题，需要根据已经知道的事实推出结论，或者从多重的信息源或网页中对信息进行综合或摘取。

自然语言理解系统：拼写歧正、语法检查

## 语音与语言处理中的知识

语音处理的基础知识：

* 语音学（phonetics）：单词是怎样发出音来而成为声音序列的。
* 音系学（phonology）：每一个声音是怎样在语音学上实现的。

语言处理的基础知识：

* 形态学：关于词的有意义的组成成分的知识
* 句法学：关于词与词之间结构关系的知识。
* 语义学：关于意义的知识
  + 词汇语义学（lexical semantics）：单词的意义
  + 组合语义学（copositional semantics）：单词组合的意义
* 语用学（pragmatic）或对话（dialogue）：关于意义与说话人的目的和意图之间的关系的知识
* 话语学：关于比一个单独的更大的语言单位的知识。

## 歧义

消解（resolve）或者排歧（disambiguation）的模型与算法。

* 词汇排歧（lexical disambiguation）：
  + 词类标注（parts-of-speech tagging）
  + 词义排歧（word sense disambiguation）
* 句法排歧（syntactic disambiguation），也叫语法排歧。
  + 概率剖析（probabilistic parsing）
  + 言语行为解释（speech act interpretation）

## 模型和算法

模型：

* 状态机器（state machine），就是形式模型。形式模型包括：状态、状态之间的转移以及输入表示等等。
  + 确定的有限状态自动机（deterministic finite-state automata，DFSA）
  + 非确定的有限状态自动机（non-deterministic finite-state automata，NFSA）
  + 有限状态转录机（finite-state transducers，FST）
* 形式规则系统（formal rule system）
  + 正则语法（regular grammars）
  + 正则关系（regular relations）
  + 上下文无关语法（context-free grammars）
  + 特征增益语法（feature-augmented grammars）
* 基于逻辑（logic）的模型
  + 一阶逻辑（first order logic），即谓词演算（predicate）
  + λ运算（lambda-calculus）
  + 特征结构（feature structure）
  + 语义基元（semantic primitives）
* 概率模型（probabilistic models）
  + 加权自动机（weighted automaton）
  + 隐马尔可夫模型（Hidden Markov Models，HMM）
  + 向量空间模型（vector-space models）

算法：

* 动态规划（dynamic programming）算法的状态空间搜索（state space search）
  + 深度优先搜索（depth-first search）
  + 最佳优先搜索算法（best-first search）
  + A\*搜索算法（A\* search）（Ref：Ch10）
* 分类器（classifiers）和序列模型（sequence models）
  + 分类器把一个单独的客体指派到一个单独的类别中
    - 决策树（decision trees）
    - 支持向量机（support vector machines）
    - 高斯混合矩阵（Gaussian mixture models）
    - 逻辑回归（logistic regression）
  + 序列模型对于一个客体序列进行分类，并将它指派到一个类别序列中
* 期望最大化算法（Expecctation-Maximization，EM）
* 统计技术
  + 交叉验证（cross-validation）

## 语言、思维和理解

图灵测试（Turing test）

## 学科现状与近期发展

## 语音和语言处理简史

研究领域

* 语言学中的计算语言学（computational linguistics）
* 计算机科学中的自然语言处理（natural language processing）
* 电子工程中的语音识别（speech recognition）
* 心理学中的计算机心理语言学（computational psycholinguistics）

## 小结

本章重点：语音处理和语言处理

* 语音和语言处理技术与音系学、语音学、形态学、句法学、语义学、语用学和话语分析等不同平面上的语言知识的形式模型和形式表示方法有着依赖关系。使用包括状态机、形式规则系统、逻辑等在内的形式模型以及概率模型。
* 语音和语言处理的基础是计算机科学、语言学、数学、电子工程和心理学。
* 语言和思维之间的密切联系使语音和语言处理技术成为了机器智能的一部分。
* 语音和语言处理的应用已经越来越丰富多彩。

第一部分 词汇的计算机处理

# 正则表达式 与 自动机

正则表达式（regular expression）：描述文本序列的标准记录方式。

## 正则表达式

正则表达式的搜索需要提供的数据：搜索的模式（pattern）和被搜索的文本语料库（corpus）。

### 基本的正则表达式模式

正则表达式是由简单字符构成的一个序列，表达式是区分大小写的。

方括号“[]”内部的字符符号串表示字符是析取（disjunction）的。

* 连字符“-”：表示在某一范围内的任何字符。
* 脱字符“^”：如果是方括号内使用，表示否定后续的模式，即不出现某些字符；如果在方括号外使用，只表示字符本身。
* 通配符（wildcard）：点号“.”。表示任何与单个字符（回车符除外）相匹配的字符。点号常常与星号共用。
* 计数符：
  + 问号“?”：表示前面一个字符存在或者不存在。
  + 星号“\*”：称为“Kleene \*”，其直接前面的字符或正则表达式出现零次或者多次。
  + 加号“+”：称为“Kleene +”，其直接前面的字符或正则表达式出现一次或者多次。
  + 大括号“{m,n}”：表示前面的字符至少出现m次，至多出现n次。
* 锚号（anchors）：把正则表达式锚在符号串中某个特定位置的特殊字符。
  + 脱字符“^”：与行的开始相匹配；
  + 美元符“$”：与行的结束相匹配。
  + “\b”：表示词界；
  + “\B”：表示非（不是）词界。

### 正则表达式的基本算符

析取符（pipe symbol）“|”：析取算符（disjunction operator）。

圆括号“()”：表示优先关系。

算符优先层级（operator precedence hierarchy）：圆括号＞计数符＞序列与锚＞析取符

贪心模式：正则表达式尽可能与最长的符号串匹配。

### 正则表达式的简单例子和复杂例子

错误类型：

* 正面错误（false positives）：错误地匹配的字符串
* 负面错误（false negatives）：错误地遗漏的字符串

处理目标：（两个目标是矛盾的）

* 增加准确率（accuracy）：把正面错误减少到最低限度。
* 增加覆盖率（coverage）：把负责错误减少到最低限度。

### 正则表达式的高级算符

高级算符可以用简单算符来表达，高级算符可以简化表达式的内容，更加方便阅读和理解。

* \d = [0-9]：任何数字字符
* \D = [^0-9]：任何非数字字符
* \w = [a-zA-Z0-9\_]：任何字母字符、数字字符或者空白字符
* \W = [^\w]：一个非字母字符、数字字符或者空白字符
* \s = [ \r \t \n \f]：空白区域（空白字符、制表字符、换行字符、回车字符）
* \S = [^\s]：非空白区域

### 正则表达式的实际操作

替换（substitution）是正则表达式的用途之一。

使用数字算符（number operator）“\1”可以参照前面的模式，从而实现替换。

数字算符使用数字存储器保存前面的模式，这些数字存储器称为寄存器（registers）。

## 有限状态自动机（FSA）

正则表达式

* 是一种用于文本搜索的元语言。
* 是描述有限状态自动机（Finite-State Automaton，FSA）的一种方法。
  + FSA是语言计算的理论基础，任何正则表达式都可以使用有限状态自动机来实现
* 是刻画正则语言（regular language）的一种方法。
  + 正则语言是形式语言中的一种。正则语言可以使用正则表达式、有限状态自动机和正则语法（regular grammar）进行描述，即三者是等价的。

### FSA的表示方式

自动机（automaton），也叫有限自动机，或者叫有限状态自动机，能够识别由符号串构成的集合。

自动机的表示方法：

* 图表示：结点表示状态，弧表示转移；
* 状态转移表（state-transition table）表示。

自动机需要5个参数实现形式化定义：

* Q：N中状态的有限集合
* Σ ：有限的输入符号字母表
* q0：初始状态
* F：终极状态集合，F ⊆ Q
* σ（q, i）状态之间的转移函数或者转移矩阵

### FSA中的形式语言

形式语言（formal language）：是一个用来生成或者识别满足形式语言定义的要求的符号串的模型。

形式语言是符号串的集合，而每一个符号串是由字母表（alphabet）中的有限的符号的集合组合而成的。基于形式语言定义的自动机可以在封闭的形式中表示无限的集合。

生成语法（generative grammar）：表示形式语言的语法，即自动机定义的能够生成一切可能的符号串的语言。

### 非确定的FSA（NFSA）

非确定的有限自动机（Non-deterministic FSA，NFSA），也可以称之为概率的有限自动机（Probabilistic FSA，PFSA），相对应的自动机也可以称之为确定的有限自动机（Deterministic FSA，DFSA）。

NFSA或PFSA存在概率转移（ε-转移，ε-transition）。

### NFSA接收符号串

三种非确定问题的解决方案（solution to the problem of non-determinism）：

* 回退（backup）：在选择点做记号，记录位置和状态，当遇到错误的选择时可以回退到选择点，从而试探其他的路径
  + 在每个选择点上，需要记住所有不同的选择
  + 对于不同的选择，需要存储足够的信息
  + 结点与位置的结合体称为识别算法的搜索状态（search-state）。
  + 自动机的状态称为结点（node）或者机器状态（machine-state）。
* 前瞻（look-ahead）：在输入中向前看，预测性地判断应该选择哪条路径
* 并行（parallelism）：在选择点上并行探查每条不同的路径。

### FSA识别就是搜索

状态空间搜索（state-space search）算法：系统地探索自动机中所有可能的路径，从而正确地识别正则语言中的符号串。

* 深度优先搜索（depth-first search）策略，即后进先出（Last In First Out，LIFO）策略。当状态空间是无限的时候，搜索可能永远无法停止。
* 广度优先搜索（breadth-first search）策略，即先进先出（First In First Out，FIFO）策略
* 规模较大的问题，更加复杂的搜索技术：动态规划（dynamic programming）（Ref：Ch13）和A\*搜索算法（Ref：Ch10）。

### DFSA与NFSA之间的关系

使用算法可以把NFSA转换为等价的DFSA，因此NFSA与DFSA是完全等价的。

## 正则语言与有限状态自动机

正则语言的形式化定义：

* Φ是正则语言
* ∀ a ∈ Σ ∪ ϵ ，{a}是正则语言
* 如果L1和L2是正则语言，那么
  + L1和L2的毗连也是正则语言
  + L1和L2的并或者析取也是正则语言
  + L1的Kleene闭包也是正则语言

正则语言的运算：

* 交：如果L1和L2是正则语言，那么L1 ∩ L2也是正则语言
* 差：如果L1和L2是正则语言，那么L1 - L2也是正则语言
* 补：如果L1是正则语言，那么Σ\* - L1也是正则语言
* 逆：如果L1是正则语言，那么L1^(-1)也是正则语言

正则表达式与自动机的等价转换，即正则表达式的每个基本操作都可以通过自动机模拟。

## 小结

本章重点：有限自动机（finite automaton）和正则表达式（regular expression）。

* 正则表达式语言是模式匹配的有力工具。
* 正则表达式的基本运算
  + 符号的毗连
  + 符号的析取（[]，|，.）
  + 记数符（\*，+，{m,n}）
  + 锚号（^，$）
  + 前于运算符（(，)）。
* 正则表达式等价有限状态自动机，任何正则表达式都可以转换为有限自动机。
* 存储器（\1和()）是一种高级运算，经常作为正则表达式的一部分，可以实现替换功能，但是不能实现为有限自动机。
* 形式语言被隐含地定义为在任何的词汇（符号集）中自动机所接收的符号串的集合。
* 确定的自动机（DFSA）的行为完全由它的状态决定。
* 非确定的自动机（NFSA）的行为有时必须在多条路径之间依据概率进行选择。
* 任何一个NFSA都可以转换成为等价的DFSA，即NFSA和DFSA是等价的。
* NFSA的进程表中搜索下一个状态的顺序决定了搜索策略：
  + 深度优先或者后进先出（LIFO）策略相当于把进程表看成堆栈；
  + 广度优先或者先进先出（FIFO）策略相当于把进程表看成队列。
* 任何的正则表达式都可以自动地编译为NFSA，即可以自动地编译为DFSA。

# 词 与 转录机

正词法规则（orthographic rules）：即拼写规则（spelling rules）。当两个语素结合时，拼写变化的规则。

形态规则（morphological rules）：关于形态顺序的模型，用于解释在一个词内，语素与语素之间的联系规则。

剖析（parsing）就是取一个输入并且产生出关于这个输入的各类语言结构。剖析产生的结果可以是形态结构、句法结构、语义结构或者话语结构。剖析产生的形式可以是符号串、树或者网络。

形态剖析（morphological parsing）：把单词剖析为多个词素。

形态剖析的任务：

* 词干还原（stemming）
* 词目还原（lemmatization）
* 词例还原（tokenization）
* 单词切分（word segmentation）

形态剖析测量两个单词在正词法上的相似度：最小编辑距离

## 英语形态学概观

形态学：是研究语素（小的意义单位）构成词的方法。

语素（morpheme）：是语言中负荷意义最小的单位。分为词干（stem）和词缀（affix）。

词缀包括：

* 前缀（prefix）：位于词干之前
* 后缀（suffix）：紧接词干之后
* 中缀（infix）：插入到词干中间
* 位缀（circumfix）：同时处于词干的前面和后面

语素构成单词的四种方法：

* 屈折（inflection）：把词干和一个表示语法的语素结合起来，所形成的单词一般与原来的词干属于同一个词类，还会产生一些如“一致关系”类的语法功能。
* 派生（derivation）：把词干和一个表示语法的语素结合起来，所形成的单词一般属于不同的词类，产生的新意义经常难于精确地预测
* 合成（compounding）：把多个词干结合在一起
* 附着（cliticization）：把一个单词与一个附着成分（clitic）结合起来。附着成分也是一个语素，它的句法作用像一个形式简化了的单词，按照音系学的规则或者正词法的规则附着在其他单词上。

### 屈折形态学（inflectional morphology）

英语的屈折系统相对简单：只有名词、动词和部分形容词有屈折变化，可能的屈折词缀的数目也相对较少。屈折的能产性（productive）比较高。

英语名词的屈折变化：

* 复数（plural）
* 领属（possessive）

英语动词分类：

* 主要动词（main verbs）：eat, sleep, impeach
* 情态动词（modal verbs）：can, will, should
* 基础动词（primary verbs）：be, have, do

英语动词的屈折变化：

* 规则的屈折动词：
* 不规则的屈折动词：

注：-ing是一个能产性高的后缀，即它可以用于英语中所有的动词。

### 派生形态学（derivational morphology）

英语的派生系统相对复杂：通过对动词和形容词的变化产生名词，即名词化（nominalization）。派生的能产性比较低。

### 附着（cliticization）

附着成分是处于词缀和单词之间的语言单位。

* 位于单词前面的附着成分称为前附着成分（proclitics）
* 跟在单词后面的附着成分称为后附着成分（enclitics）。

附着成分的音系学功能相当于词缀，一般比较短，也没有重读。

附着成分的句法功能更像一个单词，作用经常相当于代词、冠词、连接词或者动词。

### 非毗连形态学

毗连形态学（concatenative morphology）：单词是由彼此毗连的语素构成的符号串。

非毗连形态学（non-concatenative morphology），也称为模板形态学（template morphology）或者词根与模式形态学（root-and-pattern morphology）：语素组合方式极为复杂。

### 一致关系（agreement）

数的一致，性的一致。

## 有限状态形态剖析

形态剖析器：

* 词表（lexicon）：词干和词缀表以及它们的基本信息
* 形态顺序规则（morphotactics）：关于形态顺序的模型，用于解释在一个词内，语素与语素之间的联系规则。
* 正词法规则（orthographic rules）：即拼写规则（spelling rules）。当两个语素结合时，拼写变化的规则。

## 有限状态词表的建造

计算机词表的构造：

* 列出语言中的每个词干和词缀
* 表示出形态顺序规则，即词干和词缀的结合规则
  + 基于有限状态自动机为形态顺序规则建模

形态识别（morphological recognition）问题：使用FSA判断由字母构成的输入符号串是否合法。

## 有限状态转录机（FST）

有限状态转录机（Finite-State Transducer，FST）：用来进行两个层之间的映射的自动机，即可以进行两个符号集合之间的映射的有限自动机。

FST的用途：

* 作为识别器（recognizer）：取一对符号串S1和S2作为输入和输出，如果S1作为输入得到输出是S2，或者S2作为输入得到输出是S1，则识别成功；否则识别失败。
* 作为生成器（generator）：如果FST能够输出一对符号串S1和S2，则输出成功并同时输出这对符号串，否则输出失败。
* 作为翻译器（translator）：FST输入符号串S1，输出符号串S2
* 作为关联器（relater）：计算机符号串的两个集合之间的关系

FST需要7个参数定义：

* Q：状态的有限集合N
* Σ：对应于输入字母表中的符号的有限集合
* Δ：对应于输出字母表中的符号的有限集合
* q0 ∈ Q：初始符号
* F ⊆ Q：终极状态的集合
* δ（q,w）：转换函数或者状态之间的转换矩阵
* σ（q,w）：输出函数。对应于每一个状态和输入，给出可能的输出符号串的集合。

FSA与正则语言同构（isomorphic），FST与正则关系（regular relation）同构。

正则关系是符号串偶对的集合，作为符号串集合的正则语言的自然扩充。

### 定序转录机和确定性

定序转录机（sequential transducer）：是转录机的一个次类，输入是确定的。

后继转录机（subsequential transducer）是定序转录机的泛化。

p-后继转录机（p-subsequential transducer）：是后继转录机的泛化。

## 用于形态剖析的有限状态转录机

有限状态形态学（finite-state morphology）的范式（paradigm），把一个单词表示为词汇层（lexical level）和表层（surface level）之间的对应。

* 词汇层：表示组成该词的语素之间的毗连关系
* 表层：表示该词实际拼写的字母之间的毗连关系。

## 转录机和正词法规则

使用正词法规则来处理英语中在语素边界发生拼写变化的问题。

正词法规则可以在转录机上实现。

## 结合有限状态转录机的词表与规则

把转录机的词表和规则结合起来进行剖析和生成。

双层形态学结构，既能用于剖析，还能用于生成。

词表转录机把表示词干和形态特征的词汇层面映射于表示语素简单毗连的中间层面。

若干个正词法转录机并行地运行各种不同的拼写规则。

有限状态转录机从词汇带子生成表层带子时，或者从表层带子剖析词汇带子时，都可以使用带有同样状态序列的同样的层叠式转录机。

剖析比生成要复杂一些，因为在剖析中存在歧义的问题，而排歧需要某些外部的证据。

运行层叠式转录机时，可以通过组合（composing）和交合（intersecting）转录机的方式使之更加有效。

## Porter词干处理器（不使用词表）

形态剖析的标准算法：使用词表加规则的方法来建立转录机。缺点是需要大规模的联机词表。

Porter词干处理器是相对简单的算法，可以应用于信息检索中。（Ref：Ch23）

## 单词和句子的词例还原

词例还原（tokenization）：把文本切分成单词和句子。

* 单词切分（word segmentation）
* 句子切分（sentence segmentation）

### 中文的自动切词

中文切分算法：最大匹配算法（maximum matching，maxmatch），是一种贪心搜索算法，作为基准算法，需要配备一部语言的词典（词表）。缺陷在处理未知词或者未知组合时。

## 拼写错误的检查与更正

拼写错误更正的标准算法是概率算法。

拼写错误的检查和更正分解为三大问题：

* 非词错误检查（non-word error detection）：检查会导致非词的拼写错误。
* 孤立词错误更正（isolated-word error correction）：更正会导致非词的拼写错误。
* 依赖于上下文的错误检查和更正（context-dependent error detection and correction）：如果错误的拼写恰好是一个英语中真实存在的单词，就需要使用上下文来检查和更正这样的拼写错误。

有限状态形态剖析器提供了实现大规模词典的技术手段。对于每个单词都可以给出形态剖析，因此FST剖析器在本质上就是单词的识别器。如果使用投影操作把FST下侧的语言图抽取出来，那么FST形态剖析器就可以转化为有效的FSA单词识别器。FST词典就会表示那些能产性的形态变化。

最小编辑距离算法用于计算来源和表层错误之间的距离。

## 最小编辑距离（minimum edit distance）

字符串距离（string distance）：两个符号串之间的距离是这两个符号串彼此相似程度的度量。

最小编辑距离是符号串距离算法的基础，说明了两个符号串之间对齐的情况。

Levenshtein距离是加权的最小编辑距离。

最小编辑距离使用动态规划进行计算。

最小编辑距离算法可以用来做两个符号串之间的最小代价对齐（alignment）。

## 人的形态处理方式

完全枚举法（full listing）：假定人的心理词表中，不管单词内部形态结构如何，只会把语言中的全部单词都一一枚举出来。形态结构只是一种没有因果关系的现象。

最小羡余法（minimum redundancy）：假定人的心理词表中，只表示那些有组合能力的语素，因此当处理单词中必须把与单词相关的所有语素组成起来。

## 小结

* 形态剖析是发现在词中所包含的连续的语素的过程
* 英语主要使用前缀和后缀来表示屈折形态和派生形态
* 英语的屈折形态比较简单，包括：人称和数的一致关系以及时态标志
* 英语的派生形态比较复杂，包括：前缀和后缀
* 英语的形态顺序规则（可容许的语素的顺序）可以用有限自动机来表示
* 有限状态转录机是能生成输出符号的有限自动机的扩充。FST的重要运算包括：组合、投影和交运算
* 有限状态形态学和双层形态学是有限状态转录机在形态表示和剖析中的应用
* 转录机的自动编译程序是存在的，并且对于任何简单的重写规则都能够选出一个转录机。词表和拼写规则可以通过组合和交合不同的转录机的方式结合起来
* Porter算法是词干还原的简单方法，可以帮助词干剥离词缀。精确度不如包含了词表的转录机，但是可以应用在不需要做精确形态剖析的工作中，例如：信息检索
* 单词的词例还原可以使用简单的正则表达式替换或者使用转录机来实现
* 拼写错误检查通常可以通过发现那些没有在词典中出现的单词的办法来实现；为此可以使用FST的词典
* 两个符号串之间的最小编辑距离是把一个符号串编辑为另一个符号串时所需要的最少的操作次数。最小编辑距离可以使用动态规划的方法来计算结果，也可以用来作两个符号串的对齐。

# N 元语法

N元语法模型（N-gram model）是概率模型。

2元语法（bigram）是包含2个单词的序列；

3元语法（trigram）是包含3个单词的序列；

语言模型（Language Models，LM）是单词序列的概率模型。

## 语料库中单词数目的计算

自然语言中统计计算需要依赖于语料库（单数：corpus，复数：corpora）。

语料库是计算机可读的文本或者口语的集合体。

阻断（disfluencies）：切断和有声停顿。

切断（fragment）：一个单词在中间被拦腰切开就形成切断。

有声停顿（filled pauses）：也称为过滤成分（filters）。单词的停顿。

词目（lemma）：具有相同的词干和相同的词义并且主要的词类也相同的词汇形式

词形（wordform）：是一个单词的全部的屈折或者派生形式。

语言的“型”（type）和“例”（token）。“型”就是语料库中不同单词的数目，或者是词汇容量的大小，记为V；“例”就是使用中的全部单词数目，记为N。

## 简单的（非平滑的）N元语法

计算某个单词的概率，只考虑最接近该单词的若干个单词，近似地逼近该单词的历史，这是N元语法模型的直觉解释。

一个单词的概率只依赖于它前面单词的概率的这种假设称为马尔可夫假设（Markov assumption）

使用最大似然估计（Maximum Likelihood Estimation，MLE）估计N元语法模型的概率。

从语料库中得到的计数加以归一化（normalize）。

用前面符号串（prefix）的观察频率来除这个特定单词序列的观察频率，就得到N元语法概率的估计值，这个比值称为相对频率（relative frequency）。

## 训练集与测试集

* 训练集（training set）或者训练语料库（training corpus）：
* 测试集（test set）或者测试语料库（test corpus）：
* 初始的测试集又称为调试测试集，又称为开发集（development test set，devset）：用于分析和调整模型的参数
* 保留集（held-out set）：附加的数据源来增强训练集

在一些数据上训练，在另一些数据上测试，还可以用来评估不同N元语料的总体结构。

N元语法及其对训练语料库的敏感性：确保训练语料库与测试语料库有相似性。

### 未知词：开放词汇和封闭词汇

未知词（unknown words），或者表外词（Out Of Vocabulary，OOV）：不在系统中的单词。

* 在测试集中出现的表外词OOV的百分比称为表外词率（OOV rate）。
* 在封闭词汇（closed vocabulary）系统中，假定不存在未知词。
* 在开放词汇（open vocabulary）系统中，给测试集加上一个伪词（pseudo-word）来给这些潜在的未知词建模，这个未知词的模型称为<UNK>。

## N元语法的评测：困惑度

* 外在评测（extrinsic evaluation）称为端到端（end-to-end）的评测，也称为现实评测（in vivo evaluation）或者叫体内评测，就是将语言模型嵌入到某种应用中，然后通过测试这个应用的总体性能来评价语言模型的质量。
* 内在评测（intrinsic evaluation）：与任何应用无关的模型质量的评测方法。

困惑度（perplexity，PP）：是对于N元语法模型的一种最常见的内在评测的度量指标。困惑度可以用来快速地检验算法，困惑度的改进也可以由端对端的评测来加以确认。

在一个测试集上语言模型的困惑度是该语言模型指派给测试集的概率的函数。

对于测试集，困惑度就是用单词数归一化之后的测试集的概率。

语言的加权平均转移因子（weighted average branching factor）是语言中的任何一个单词后面可能接续的单词的数目。

## N元语法的平滑算法

最大似然估计过程的主要问题就是训练N元语法的参数，而最大似然估计是建立在特定的训练数据集上，因此会产生数据稀疏（sparse data）问题。

“平滑”（smoothing）是用来填补零计数的概率导致的概率计算问题。

### Laplace平滑

Laplace平滑（smoothing）或Laplace定律（law）：也称为加一平滑（add-one smoothing）。语法模型的计数矩阵，所有的计数加1，再对概率进行归一化。实际应用效果不好。

### Good-Turing打折法

各种打折算法（Good-Turing打折法，Witten-Bell打折法，Kneser-Ney平滑法）都是利用看过一次的事物的计数来帮助估计从来没有看到过的事物的计数。

只出现过一次的单词或者N元语法（或者任何事件）都可以称为单元素（singleton），或者称为罕用语（hapax legomenon），即只出现过一次的单词。

Good-Turing打折法就是使用单元素的频率作为零计数的一元语法的频率来重新估计概率量的大小。

### Good-Turing估计

在对N元语法进行打折时，不仅使用Good-Turing打折法，还需要使用回退和插值算法。

## N元语法的插值法

使用“有层次”的N元语法的两种途径：

* 回退法（back off）：当高阶语法计数存在零概率值时，就回退到低阶的N元语法中，使用非零概率值进行计算。
* 插值法（interpolation）：把所有的N元语法估计中的概率值混合起来，即一元语法、二元语法、……、N元语法的计数进行加权插值。

## N元语法的回退法

Katz回退法是使用了Good-Turing打折法的回退法。

## 实际问题：工具包和数据格式

回退N元语法模型一般用ARPA格式存储。

建立语言模型的工具包：

* SRILM工具包
* Cambridge-CMU工具包

## 语言模型建模中的高级专题

### Kneser-Ney平滑法

Kneser-Ney平滑法基于绝对折扣（absolute discounting）的打折方法。

Kneser-Ney打折法使用更加精致的方法分摊回退值，从而提升绝对折扣

Kneser-Ney算法使用插值的形式比使用回退的形式效果更好。

### 基于类别的N元语法

基于类别的N元语法（class-based N-gram）或聚类N元语法（cluster N-gram）是使用单词的类别信息或者聚类信息的N元语法的变体。

基于类别的N元语法对于处理训练集中的数据稀疏问题效果很好。

IBM聚类N元语法，是一种硬聚类模型；

### 语言模型的自适应

语言模型的自适应（adaptation）：当某个领域内只有数量很少的训练数据，但是其他领域又存在大量数据时，就会面临语言模型的自适应问题。可以使用领域之外的大量的数据集进行训练，设法使训练得到的模型与某个领域内的小量数据产生自适应。

### 长距离信息的使用

N元语法建模中，长距离的上下文信息基本没用。为了更好地利用长距离的上下文信息，可以使用的模型：

* 隐藏语言模型（cache language model），基于缓存的语言模型。
* 跳跃式的N元语法（skip N-grams）：上下文可以“跳跃过”某些中间的单词。
* 可变长的N元语法（variable-length N-gram）。

语言模型中结合语言结构的方法参考基于统计剖析的句法结构的语言模型（Ref：Ch14）

基于对话中言语行为的语言模型（Ref：Ch24）

## 信息论背景

困惑度是建立在信息论（information theory）中关于交叉熵（cross-entropy）概念的基础上。

熵（entropy）是信息量的度量。

* 可以度量在一个特定的语法中的信息量的多少
* 可以度量给定语法和给定语言的匹配程度的大小
* 可以预测一个给定的N元语法中下一个单词是什么

计算序列（sequences）的熵。

熵率（entropy rate）：用单词数来除序列的熵所得的值，即每个单词的熵。

平稳（stationary）的随机过程：随着时间的推移，随机过程指派给序列的概率是不变的。

### 用于比较模型的交叉熵

交叉熵（cross-entropy）。

单词序列W上的模型的困惑度形式地定义为交叉熵的指数。

## 英语的熵和熵率均衡性

英语的熵为概率语法试验提供了可靠的下界；英语的熵帮助信息量最大的内容。

计算英语熵值的方法：

* Shannon计算的是英语中每个字母的熵，熵值偏低；
* 使用随机模型，在很大的语料库上训练模型，使用很长的英语序列指派对数概率

## 小结

* N元语法概率是一个单词在前面给定的N – 1 个单词的条件下的条件概率。
  + N元语法概率可以通过在语料库中简单地计数，并且使之归一化的方法进行计算（最大似然估计MLE），或者也可以通过更加复杂的算法计算
  + N元语法的优点是可以使用丰富的词汇知识
  + N元语法的缺点是对训练语料库的依赖太强
* 平滑算法为N元语法概率的估计提供了比最大似然估计更好的解决办法
  + 依赖于低阶N元语法计数的常用的N元语法的平滑算法是回退法和插值法
* 常用的打折算法
  + Kneser-Ney打折法
  + Witten-Bell打折法
  + Good-Turing打折法
* 评测N元语法的语言模型时，要把语料库分为训练集和测试集两个部分
  + 训练集用于训练模型，测试集用于评测模型
  + 测试集上语言模型的困惑度用于对不同的语言模型进行比较

# 词类标注

词类（Part-of-Speech，POS），又称为单词类别（word classes）、形态类别（morphological classes）或者词汇标记（lexical tags）。

词类标注（part-of-speech tagging）：把词类指派给单词。

词类标注的常用方法：

* 基于规则的标注（rule-based tagging）：手写规则
* 基于统计机器学习模型的标注
  + 基于HMM（隐马尔可夫）模型的标注
  + 基于MEMM（最大熵）模型的标注
* 基于转换的标注（transformation-based tagging）：是基于规则的标注与基于统计机器学习模型的的结合
* 基于记忆的标注（memory-based tagging）

## 英语词的分类

* 封闭类（closed class）：包含的单词成员相对固定的词类，封闭类的单词又称为虚词（function words）。
  + 介词（prepositions）：语义上表示关系。通常是空间或时间的关系。On, under, over, near, by, at, from, to, with
  + 限定词（determiners）：与名词一起出现，常常作为名词短语开始的标记。a, an, the
    - 冠词（article）
      * 无定冠词：a, an
      * 有定冠词：the（有定性（definiteness）是话语和语义的一个特性）
  + 代词（pronouns）：she, who, I, others
  + 连接词（conjunctions）：用来连接两个短语、分句或者句子。and, but, or, as, if, when
    - 并列连接词（coordinating conjunction）：连接地位平等的两个成分
  + 助动词（auxiliary verbs）：can, may, should, are
    - 系动词（copula）：be, do, have
    - 情态动词（modal verb）：
  + 小品词（particles）：与动词结合起来一起使用，把意义加以扩展。up, down, on, off, in, out, at, by
    - 短语动词（phrasal verb）：动词与小品词结合形成一个独立的句法或者语义单位的组合。
  + 数词（numerals）：one, two, three, first, second, third
  + 叹词（interjections）
    - 否定词（negatives）
    - 礼貌标志词（politeness markers）
    - 问候词（greetings）
    - 表示存在的there
* 开放类（open class）：包含的单词成员经常变化。
  + 名词（nouns）
    - 专有名词（proper noun）
    - 普通名词（common noun）
      * 可数名词（count noun）
      * 物质名词（mass noun）
  + 动词（verbs）
    - 助动词（auxiliaries）
  + 形容词（adjectives）
  + 副词（adverbs）
    - 方位副词（directional adverbs）或地点副词（locative adverbs）：说明某个行为的方向或地点
    - 程度副词（degree adverbs）：说明某个动作、过程或性质延伸的程度
    - 方式副词（manner adverbs）：描述某个行为或者过程的方式
    - 时间副词（temporal adverbs）：描述某个行为或者事件发生的时间

## 英语的标记集

* Brown语料库的87个标记
* Penn Treebank的45个标记，是小型标记集
* CLAWS的C5标记集，61个标记，是中型标记集

## 词类标注

词类标注（Part-of-speech tagging，POS tagging）简称为标注（tagging），给语料中的每一个单词指派一个词类或者其他句法类别标记的过程。

先进行词例还原，才可以完成词类标注，词类标注的困难就是歧义消解（ambiguity resolve）。

## 基于规则的词类标注算法

词类自动标注算法使用两阶段的体系结构：

* 使用一部词典给每一个单词指派一个潜在的词类表；
* 使用一个手工书写的排歧规则筛选原来的潜在词类表，使每个单词得到一个单独的词类标记。

EngCG标注算法（EngCG tagger）是最全面的基于规则的词类标注算法，是使用约束语法的方法。EngCG中的词典ENGTWOL是建立在双层形态学（Ref：Ch3）基础上的。

SV、SVO、SVOO表示动词的次范畴化（subccategorization）或补语化（commplementation）模式。（Ref：Ch12，Ch15）

* SV表示只带主语的动词
* SVO表示只带主语和宾语的动词
* SVOO表示带一个主语和两个补语的动词

## 基于HMM模型的词类标注算法

基于HMM模型的词类标注就是以贝叶斯推理（Bayesian inference）为理论基础的。

HMM标注算法的两个简化假设：

* 单词出现的概率只与其本身的词类标记有关；
* 标记的出现概率只与它前面一个标记有关（即二元语法假设）

使用二元语法的标注算法包含两种概率：标记的转移概率和单词的似然度。

### 寻找最可能的标记序列

寻找最可能的标记序列需要计算两个项乘积的最大值：

* 标记序列的概率最大值
* 每一个标记生成单词的概率最大值

### HMM标注算法的形式化说明

### 使用Viterbi算法进行HMM标注

解码（decoding）：对于包含隐藏变量的模型，确定隐藏在观察序列之下的源变量的变量序列。

Viterbi算法是经典的动态规划算法（dynamic programming algorithm），是解码的常用算法。

### 使用三元语法的HMM标注算法

* 让标记的概率依赖于前面两个标记
* 增加一个与序列结尾有关的记号
  + 解决数据稀疏问题可以使用删除插值法

## 基于转换的词类标注算法

Brill标注就是机器学习中的基于转换的学习（Transformation-Based Learning，TBL）标注算法。

* TBL是基于规则的
* TBL的规则是利用机器学习技术从数据中自动推导出来的
* TBL是有指导的学习技术，在标注之前需要预先标注过的训练语料库

TBL算法的标注规则：

* 基于语料库比较普遍的，但是宽泛的规则来标注
* 使用特殊的规则来修正原来的某些标记

### TBL规则的应用

在应用规则之前，标注系统已经给每个单词都标注上了最可能的标记。在选择了最可能的标记之后，Brill标注算法应用转换规则。

### TBL规则的学习

Brill标注算法的三个阶段：

* 把每个单词标上最可能的标记；
* 检查每一个可能的转换，并且选择那个能够最大程度地改善标注的转换
* 应用这个转换对数据进行重新标注

注：反复应用后面两个阶段，直到不能再继续充分地改善前一轮的结果为止。

TBL是一种有监督的学习算法（supervised learning algorithm）。

TBL的输出过程是一个转换的有序表，这些转换组成了一个“标注过程”（tagging procedure），并且这个“标注过程”还可以应用于新的语料库。

TBL的转换规则可以有无限多个，即解空间是无穷的。因此为了限制转换集合，可以设计一个称为模板（templates）的小集合，这个模板也就是转换的摘要（abstracted transformation）。

## 评测和错误分析

算法基于数据集的评测：

* 利用不同数据集评测： 数据集分成训练集、测试集和第二测试集（也称为调试测试集）
* 利用数据集交叉验证评测：随机地将数据分成训练集与测试集，分成N次就做N次训练和测试，然后计算这N次的平均错误率。

算法评测的基准：

* 人类顶线（human ceiling）：人类黄金标准，即人做标记时可以达到的标准，这个标准就是分类算法的上界顶线。
* 最大频率类别底线：训练集中最大频率出现的标记所占的比重就是评价分类器性能的底线。
  + Gale对每一个歧义单词选择一元语法最可能的标记。每一个单词的最可能的标记可以根据手工标注的语料库来计算。

算法的错误分析：

* 含混矩阵（confusion matrix）或者列联表（contingency table）
  + 含混矩阵的“行”所表示的是正确的标记
  + 含混矩阵的“列”所表示的是假定的标记
  + 含混矩阵的每个“单元”所表示的是标注错误的百分比

## 词类标注中的高级专题

### 标记的不确定性

标记的不确定性问题（tag indeterminacy）：当一个单词在多个标记之间出现歧义，并且排除歧义非常困难。

* 可以使用多重标记（multiple tags）处理这个问题。
* 可以使用下面的方法来解决这个问题：
  + 采用某种方式选择一个标记来替代不确定的标记
  + 训练时，使用这个单词中的某个标记作为训练数据；
  + 测试时，只要多重标记中有一个标记属于正确标记，就认为这个标记是正确的
  + 把不确定的标记作为一个单独的复杂标记来处理

先词例还原，再词类标注。

### 未知词

* 最简单的未知词算法：假定每个未知词在所有可能的标记中都是有歧义的，具有相同的概率。因此，标记器必须根据上下文相关的词类三元语法给未知词一个恰当的标记。
* 复杂的未知词算法：在未知词上的标记的概率分布与在训练集中只出现一次的单词的标记的概率分布非常相似。
  + 一次性罕用词（单数：hapax legomenon，复数：hapax legomena）：只出现一次的单词。
* 最强的未知词算法：使用了单词的形态信息和正词法，通过分析单词的形态形式和拼写方式来猜测单词所属的词类。
* 最新的未知词算法：最大熵马尔可夫模型（Maximum Entropy Markov Model，MEMM）

### 高度屈折语和黏着语的词类标注

* 未知词的标注是很困难的问题：增加专门处理未知词的词典能够改善语言标注器的性能
* 单词形态编码的信息十分丰富：使用标记集给单词指派标记时必须综合解决词类标注和形态排歧的问题。因此，形态剖析对于未知词的处理起着关键性的作用。

### 标注算法的结合

* 并行的结合方法：标注算法之间的误差互补时有效
  + 在同一个句子上并行地运行多个标注算法
  + 再把输出结果综合考虑
    - 通过投票的方法
    - 通过训练另外一个分类器的方法
  + 选出一个最可信的标注算法
* 串行的结合方法
  + 对于每个单词先使用基于规则的方法去掉那些不可能标记的概率
  + 然后再使用HMM标注算法从剩下的标记中把最好的标记选择出来

## 拼写中的噪声信道模型

噪声信道模型（noisy channel model）：可以作为用于词类标注的贝叶斯推理模型的另一种解释。是语音识别和机器翻译的关键工具，还可以用来更正单词拼写错误。

单词拼写错误：

* 非词拼写错误（non-word spelling errors）：输入的单词是非词，即输入的单词不是英语词典中存在的单词。通过查询英语词典就可以探测出这些非词，使用最小编辑距离就可以提供候选的更正单词
* 真词拼写错误（real-word spelling errors）：输入的单词是英语词典中存在的单词，但是不是应该输入的单词（例如：three输入为there）。使用噪声信道模型进行上下文拼写检查（contextual spell checking）就可以更正真词拼写错误

噪声信道模型的解释：一个正确拼写的单词通过噪声通信的信道时因为“噪声”（noise）的原因给“扭曲”（distorted）了。

解决方案：建立一个模拟的噪声信道，然后用词典中的单词通过这个噪声信道，然后模拟的噪声信道输出的噪声词与原始文本中的噪声词对比，找出最相似的噪声词，再返回词典中的输入单词，就是需要发现的真正的单词。

噪声信道模型也是基于贝叶斯推理模型来实现的。

上下文有关的拼写更正（contextual-sensitive spell correction）：是基于噪声信道模型来探测和更正真词拼写错误。将操作错误作为输入信道中的噪声，产出的真词拼写错误是输出的结果。

## 小结

* 各种语言都有一个相对较小的封闭类的词集合
  + 封闭类的词多是高频词，多是虚词，多是词类标注中歧义较强的词
  + 开放类的词多是名词、动词和形容词
  + 词类编码方案，即标记集合。拥有40个~200个标记
* 词类标注是给单词序列中的每个单词指派词类标记的过程
  + 基于规则的标注算法使用手写规则来区分标记的歧义
  + 基于HMM的标注算法选择单词似然度与标记序列概率的乘积最大的标记序列作为标注结果
  + 其他的机器学习模型使用最大熵模型和对数线性模型、决策树、基于记忆的学习、基于转换的学习等方法来进行标注
* 基于HMM标注算法的概率是在手工标注的训练集上进行训练的，使用删除插值法与不同级别的N元语法相结合，并使用了复杂的未知词的识别模型
* 对于给定的HMM模型和输入符号串，Viterbi算法可以进行最优标记序列的解码
* 评价标注器的常规方法是基于测试集对比系统和人的标注结果。
  + 错误分析可以精确地确定标注器的不足之处，方便改善标注器的性能

# HMM 与 MaxEnt

HMM和MEMM：都是序列分类器（sequence classifier）或者序列标号器（sequence labeler），是给序列中的某个单元指派类或标号的模型。基于概率序列分类器，计算在可能的标号上概率分布，并且选择出最好的标号序列。

序列分类问题：

* 词类标注，序列中每个单词都被指派一个词类的标记。（Ref：Ch5）
* 语音识别（Ref：Ch9）
* 句子切分和字素——音位转换（Ref：Ch8）
* 局部剖析或者语块分析（Ref：Ch13）
* 命名实体识别和信息抽取（Ref：Ch22）

序列分类器（本章重点）：

* 隐马尔可夫模型（Hidden Markov Model，HMM）
* 最大熵马尔可夫模型（Maximum Entropy Markov Model，MEMM）

非序列的分类器（非本章重点）：

* 最大熵模型（Maximum Entropy，MaxEnt）
* 高斯混合模型（Gaussian mixture model）（Ref：Ch9）
* 朴素贝叶斯（naïve Bayes）和决策表（decision list）分类器（Ref：Ch20）

## 马尔可夫链

一个有限自动机由状态集和状态之间的转移集来确定，而这个状态之间的转移则是根据输入观察来进行的。

加权有限自动机（weighted finite-state automaton）：是有限自动机的扩充，其中的每一个孤与一个概率相关联，说明通过该弧所指的路径的可能性的大小。

马尔可夫链（Markov chain），也称为显式马尔可夫模型（observed Markov model）。

马尔可夫链是有限自动机的扩充，是加权有限自动机的特殊情况，输入序列唯一地确定了自动机将要通过的状态。因为马尔可夫链不能表示固有的歧义问题，因此只能应用于没有歧义的序列的概率指派。

概率图模型（graphical model）是表示图中概率假设的一种方法。

马尔可夫链可以看作概率图模型的一种。

马尔可夫链需要三个参数描述：

* 状态N的集合
* 转移概率国
* 初始状态和终结状态（最后状态）
  + 初始概率分布
  + 接收状态

马尔可夫链的重要假设（马尔可夫假设：）：在一个一阶马尔可夫链中，一个特定状态的概率只与它的前面的一个状态相关。

## 隐马尔可夫模型（HMM）

HMM是有限自动机（Ref：Ch3，Sec5.5）的扩充

HMM需要五个参数说明：（Ref：Ch6，P150）

* N个状态的集合
* 状态转移矩阵
* 观察序列
* 观察似然度
* 开始状态与终极状态（最后状态）
  + 初始概率分布
  + 接收状态

一阶隐马尔可夫模型有两个假设：

* 马尔可夫假设：一个特定状态的概率只与它前面一个状态有关
* 输出独立性假设：输出“观察”的概率只与产生这个“观察”的状态有关

HMM模型的分类：

* 全连通HMM（fully connected HMM）或者遍历HMM（ergodic HMM）
* 从左到右的HMM（left-to-right HMM），也称为Bakis HMM。一般用于含有时间进程的语音建模（Ref：Ch9）

HMM的三个基本问题：

1. 似然度问题（模型选择问题）：给定HMM模型参数和观察序列，计算给定序列下不同模型参数的似然度
2. 解码问题（序列标注问题）：给定HMM模型参数和观察序列，找出最好的隐藏状态序列
3. 学习问题（参数估计问题）：给定HMM中的状态集合和观察序列，通过学习得出HMM的模型参数

## 似然度问题：向前算法

向前算法（forward algorithm）是一种动态规划算法，当得到观察序列的概率时，使用一个表来存储中间值，从而解决了计算复杂度呈指数增长的问题

## 解码问题：Viterbi算法

Viterbi算法是一种动态规划算法，使用动态规划网格解决计算复杂度呈指数增长的问题。

## HMM的参数学习问题：向前——向后算法

向前——向后算法（forward-backward algorithm）或者Baumm-Welch算法，是一种期望最大化算法（Expectation-Maximization algorithm）。

* 反复地估计所得的计数
* 对于一个观察，计算它的向前概率，从而得到估计概率，然后把这个估计概率在所有的路径上进行分摊

## 最大熵模型的背景知识

最大熵模型是一种统计机器学习框架，是一种多元逻辑回归模型，是非序列分类工具。

分类：是取一个单独的观察，抽出描述这个观察的某些有用的特征，然后根据这些特征对这个观察进行分类，使之归于离散类别集合中某个类中。

概率分类器（probabilistic classifier）：不仅需要指派标记和类别，还需要给出观察在不同类别中的概率。对于一个给定的观察，概率分类器将要给出在所有类别上的概率分布。

非序列分类的应用：

* 文本分类
* 情感分析

MaxEnt属于指数分类器（exponential classifier）或者对数线性分类器（log-linear classifier）。先从输入中抽取某些特征，再把这些特征线性地组合起来（即每个特征乘以权值，然后相加）。

* 线性回归（linear regression）：使用特征进行预测的基础性工作
* 逻辑回归（logistic regression）：使用指数模型

### 线性回归

在统计学中，把某些输入特征映射到某个输出值：

* 输出是实数值（real-valued），就使用“回归”（regression）
* 输出是离散值（即类别离散集合中的某个值），就使用“分类”（classification）

用于预测的因素称为特征（feature），使用多个特征的线性回归称为多元线性回归（multiple linear regression），训练回归模型基于代价函数（cost function）。

### 逻辑回归

优势率（odds）：预测观察处于类别中的概率与观察不处于类别中的概率的比率

分对数函数（logit function）：优势率的对数

在回归模型中，使用线性函数来估计概率的分对数，而不是估计概率，这样的回归称为逻辑回归（logistic regression）。

逻辑函数（logistic function）的一般形式（逻辑回归由这个函数得名）

注：结合本章可以更容易理解逻辑回归的推导

### 逻辑回归（分类）

分类（classification），也称为推论（inference）：给定一个特定的观察，确定其属于两个类别的哪一个，也称为二元分类问题。

超平面（hyperplane）：用于分别两个类别的平面。

### 逻辑回归（训练）

在线性回归中，训练就是选择合适的权值，使得训练集上的代价函数（误差平方和）最小。

在逻辑回归中，使用条件最大似然度估计（conditional maximum likelihood estimation）来估计权值，这个问题属于凸优化（convex optimization）问题。

凸优化问题常用算法：

* 准牛顿法（Quasi-Newton method）：L-BFGS
* 梯度上升法（gradient ascent）
* 共轭梯度法（conjugate gradient）
* 迭代定标算法（iterative scaling algorithm）

## 最大熵模型（MEMM）

最大熵模型是对数线性方法（log-linear approaches）家庭中的一员，主要用于分类。其中要计算很多特征以便对单词进行标注，并且所有的特征都结合到一个基于多项逻辑回归（multinomial logistic regression）的模型当中。

对数线性模型也应用于汉语的词类标注中。汉语的单词比较短，形态特征有助于改进汉语未知词的标注效果。汉语的未知词大多是普通名词和动词，因为合成构词法（compounding）在汉语中是一种普遍的形态构词手段。

指示函数（indicator function）：如果一个特征只取值0和1.

MaxEnt在进行分类时，类别上的概率分布也会作为输出的分类结果。MaxEnt分类是（布尔）逻辑回归中的分类的泛化。在布尔逻辑回归中，分类时需要建立一个线性回归，把在该类别中的观察与不在该类别中的观察分离开来。在MaxEnt中的分类与此相反，分类时对于C中的每一个类别都要建立一个分离的线性回归。

在MaxEnt对于各种特征的布尔组合需要用手工编码为一个特征，因此成功使用MaxEnt关键在于设计恰当的特征与特征组合。

训练MaxEnt的关键是对于权值进行平滑，称为正则化（regularization）。

正则化的目的是惩罚那些大的权值。

等概率分布具有最大熵（maximum entropy），在建立最大熵模型时，默认可能的假设减至最少。“最大熵”就是通过不断地增加特征的方法来建立分布。每一个特征是一个指示函数，这个指示函数从训练的观察集合中抓取一个子集。对于每一个特征，在总的分布中增加一个约束，从而表示对于这个子集的分布与训练数据中看到的经验性的分布是匹配的。

为了从所容许的概率分布的集合中筛选出一个模型，就要选择具有最大熵的模型。这个最优化问题的解恰恰就是多元逻辑回归的概率分布，权值把训练数据的似然度最大化。当根据最大似然度来训练时，多元逻辑回归的指数模型能够找到最大熵分布，这个最大熵分布服从于来自特征函数的约束。

## 最大熵马尔可夫模型

MEMM把Viterbi算法与MaxEnt结合起来。

* MEMM是分辨模型（discriminative model，也称为判别模型）：因为模型通过训练可以在各种可能的标记序列中进行分辨（判别）
* HMM是生成模型：把似然度最优化，把似然度和先验概率结合起来估计后验概率，生成后验概率最大的标记序列
* MEMM使用输入观察中的任何有用的特征作为条件，而HMM则无法使用所有特征，而是基于似然度，即计算观察中的每个特征的似然度
* MEMM计算给定观察的状态序列的概率公式（6.96），HMM计算给定观察的状态序列的概率公式（6.95）

MEMM的训练

* 有监督学习模式。给定观察序列、特征函数和相应的隐藏状态，训练权值使得训练语料库的对数似然度最大化。
* 半监督学习模式（类似于HMM）。使用EM算法来解决训练数据的标记序列在某种程度上的缺失。

注：MEMM和HMM的对比（P175，图6.20）

* MEMM以前面的状态和当前的观察为条件，在每个状态只计算一个单独的概率函数，差别当前的观察的最可能状态
* HMM对于观察似然度和先验概率分别计算两个不同的概率，使用Viterbi算法解码得到最有可能的状态序列

## 小结

* 隐马尔可夫模型（HMM）是把观察序列与解释这些观察的隐藏类别或者隐藏状态序列联系起来
* 对于给定的观察序列，发现其隐藏状态序列的过程称为解码（decoding）或者推理（inference）。
  + Viterbi算法通常用来进行解码
* HMM的参数是转移概率矩阵和观察似然度矩阵。
  + 参数学习可以使用Baumm-Welch算法或者向前——向后算法来训练。
* MaxEnt模型是一个分类器，根据观察特征的加权集合的指数函数来计算概率，MaxEnt模型把类别指派给观察
* MaxEnt模型可以使用凸优化领域中的方法来训练
* 最大熵马尔可夫模型（MEMM）是MaxEnt序列模型的提升，
  + MEMM算法使用Viterbi解码
  + MEMM的训练可以使用EM算法来提升MaxEnt的训练算法

语音处理

# 语音学

口语词由言语的最小单位组成而成，是现代音系学（phonology）理论的最基础的原始理论（Ur-theory）。把言语和单词拆分为较小的单位的这种思想是语音识别（speech recognition）现代算法的基础，也是语音合成（speech synthesis）或者文语转换（text-to-speech）的基础。

音系学（phonology）：语言学的一个领域，描写语音在不同的环境中得到不同的实现的系统性的途径，并研究这个语音系统怎样与语法的其他部分相联系。

变异（variation）：语音在不同的上下文环境中发音不同的现象。

## 言语语音和语音标音法

语音学（phonetics）：是研究使用于世界语言中的语音的科学。

语音学（phonetics）：研究语言的声音，研究语音如何从人类发音器官产生出来，研究声音如何在声学上数字化、如何通过计算机处理。

* 发音语音学（articulatory phonetics）：研究言语的声音是如何通过口腔中的发音器官产生出来的
* 声学语音学（acoustic phonetics）：研究如何对言语的声音进行声学分析

音子（phones）：言语的发音；语段（segments）；语音字母表（phonetic alphabets）

* 国际音标（International Phonetic Alphabet，IPA）
* ARPAbet：为美国英语标音而特别设计的，使用ASCII字符。等价于IPA的ASCII表示法。

## 发音语音学

发音语音学（articulatory phonetics）：研究口腔、咽喉和自爱等不同的器官是如何改变从肺中来的气流而产生语音的。

### 发音器官

气管（trachea）、喉头（larynx）、声带（vocal folds）、声门（glottis）。

声腔（vocal tract）：口腔（oral tract）、鼻腔（nasal tract）

浊音（voiced）：声带合在一起并且发生振动时而产生的语音

清音（unvoiced或者voiceless）：声带不振动时产生的语音

鼻音（nasal sounds）：通过鼻腔而产生的语音，发鼻音时需要同时使用口腔和鼻腔作为共鸣腔（resonating cavities）。

辅音（consonants）：产生辅音时需要以某种方式限制和阻挡气流的运动。

元音（vowels）：产生元音时空气受到的阻挡较小，一般是浊音，比辅音响亮，且延续时间较长。

半元音（semivowels）：兼具辅音和元音的某些性质。

### 辅音：发音部位

辅音的发音部位（place of articulation）：最大阻挡形成的部分

* 唇音（labial）：阻挡主要在双唇的发音部位形成的辅音，也叫双唇音（bilabial）
* 唇齿音（labiodental）：发音时上齿压住下唇，气流从上齿缝流出而形成的辅音
* 齿音（dental）：舌头顶住牙齿而形成的辅音
* 齿龈音（alveolar）：齿龈是口腔顶部上齿后面的部分。发音时用舌头顶住齿龈
* 舌面前音（coronal）同时表示齿音和齿龈音
* 上腭音（palatal）：上腭（palate）处于齿龈后面口腔的顶部。发龈腭音（palato-alveolar）时，舌叶向齿龈后部的上腭隆起。
* 软腭音（velar）：软腭（velum）是口腔顶部最后面的可去的肌肉盖面部分。发音时，舌头后部向上隆起接近软腭。
* 喉音（glottal）：发音时，声带合起，喉头关闭。

### 辅音：发音方法

辅音可以通过气流的阻挡方式的不同来区分，这样的特征称为发音方法（manner of articulation）。把发音部位与发音方法充分结合就能鉴别辅音。

辅间的发音方法：

* 塞音（stop）：发音时，气流在短时间内被完全的阻塞。当空气解除阻塞时，就会发出爆破的声音，也称为爆破音（plosives）。阻塞阶段称为成阻（closure），爆破阶段称为除阻（release）。
* 鼻音（nasal）：发音时，软腭下降，气流通过鼻腔流出
* 擦音（fricative）：发音时，气流被压缩，但是并没有被完全切断，压缩而震动的气流会产生特殊的“摩擦音”
  + 声音较高的擦音称为咝音（sibilants）
  + 先摩擦而后阻塞的塞音称为塞擦音（affricates）
* 半元音（approximant）：发音时，参与发音动作的两个发音部位十分接近，但是还没有接近到能够引起强烈气流的程度
  + 边音（lateral）：气流是从舌头的边上流出来
* 颤音（tap或者flap）：发音时，舌头顶住齿龈做快速运动。

### 元音

元音可以通过发音部位来描述。元音的三个重要参数：

* 发音时舌位的高低，即舌头最高部分所处的位置
  + 高、中、低元音（high、mid、low vowels）：舌位最高点的值处于高、中、低的元音
* 元音的前后（frontness or backness）：舌头的最高位置是指向口腔的前部还是口腔的后部
  + 前元音（front vowels）：舌位处于前面的元音
  + 后元音（back vowels）：舌位处于后面的元音
* 发音时嘴唇圆的形状（rounded）：圆唇和不圆唇
* 双元音（diphthong）：在发音过程中舌位发生变化的元音

### 音节

音节（syllable）：辅音和元音结合在一起，是一个元音和其周围联系紧密的辅音结合而成的。

* 音节核（nucleus）：处于音节的核心部位的元音
* 音节头（onset）：处于音节开始的可随选的辅音或者辅音集合
  + 复杂音节头（complex onset）：音节头上的辅音超过一个
* 音节尾（coda）：音节核之后的可随选的辅音或者辅音序列

音节头+音节尾，称为韵（rime）或者韵脚（rhyme）。

音节切分（syllabification）：把一个单词自动地分割为音节

音子配列（phonotactics）：语言中音子彼此之间跟随关系的约束条件

词汇重音和非重读央元音

重读音节（accented syllable）：某些音节显得比其他音节具有更高的突显度（prominent）。与这种突显度关联的语言标志，称为音高重音（pitch accent），也称为句子重音（sentence stress）。被突显的单词或者音节，称为音高重音的负荷者（bear）。

词重音（lexical stress）是重读中需要在语音词典中表示出来的重要因素。

非重读的元音可能被弱化成为弱化元音（reduced vowels）。非重读央元音就是弱化元音。

突显度（prominence）的级别：

* 重读（accented）
* 重音（stressed）
* 第二重音（secondary stress）
* 实足元音（full vowel）
* 弱化元音（reduced vowel）

## 单位范畴与发音变异

音位（phoneme）：抽象的类别

音位变化（allophones）：音位在不同的上下文环境中的表层实现

### 语音特征

区别特征（distinctive features）把引起协同发音的各种音子进行总体性的泛化。

发音特征：发音部位和发音方法来描述

常用的发音特征：P193

### 语音变异的预测

预测效果不好的原因：

* 变异是一个随机过程，
* 在预测语音变体时，还受到许多与语音环境没有关系的因素影响

### 影响语音变异的因素

影响语音变异的重要因素：

* 语速（rate of speech）：每秒钟发出的音节数来度量
* 单词的频率或者单词的可预见性

## 声学语音学和信号

声学波形的数字化。

### 波

声学分析的基础是：正弦函数和余弦函数

波的重要特征：频率（frequency）和振幅（amplitude）。

频率是波在每1秒之内重复振动的次数，也是周数（cycle）。使用每秒钟内的周数（cycles per second）来度量频率。每秒内的周数称为赫兹（hertz，简写为Hz）。

波的周期为波在完成一周的振动时所用的时间。

### 语音的声波

模拟信号到数字信号的转换（analog-digital conversion）的两个步骤：

* 抽样（sampling）
* 量化（quantization）

抽样率（sampling rate）：每秒钟提取的样本数目 。

### 频率（音高）与振幅（响度）

基音频率（fundamental frequency，简称基频），简写为F0，声带振动的频率，或者复杂波的频率。

基音踪迹（pitch track）

RMS振幅（RMS amplitude）：振幅的均方根。

音高（pitch）和响度（loudness）是两个重要的感知特性。

* 音高是对基音频率的感受
  + “美”（mel）是音高的单位
* 响度是对信号强度的感受

基频抽取（pitch extraction）：抽取基音频率的算法

### 音子的波形解释

浊音表现为振幅上有规则的波峰（peaks）。

### 声谱和频域

傅里叶分析（Fourier analysis）：每个复杂波都可以表示为很多频率不同的正弦波的总和

信号的声谱（spectrum）：信号的频率成分和这些频率成分的振幅

频谱（spectrogram）：把声波的三个维（时间维、频率维、振幅维）可视化的一种非常有用的办法

每一条音色条纹（或者声谱峰）称为共振峰（formant）。

### 声源滤波器模型

声源滤波器模型（source-filter model）是一种解释声音的声学特性的方法，模拟声门（即声源）产生脉冲和声腔（滤波器）使脉冲成型的过程。

谐波（harmonics）的频率是基本波的倍数。

## 语音资源

* 发音词典（Pronunciation Dictionaries）
* 语音标注语料库（phonetically annotated corpus）

## 发音音系学与姿态音系学

发音音系学（articulatory phonology）把发音看成语音产生的基础。

发音姿态（articulatory gesture）：是音位抽取的基础，可以定义为参数化的动态系统

## 小结

* 音子（phones）表示单词的发音
  + 表示音子的标准系统是国际音标（International Phonetic Alphabet，IPA）
  + 最常用的英语语音计算机转写系统是ARPAbet，可以很方便地使用ASCII字符来表示英语的语音
* 音子可以由发音器官的发音方式来描述
  + 辅音的描述
    - 发音的部位
    - 发音的方法
    - 声带是否振动
  + 元音的描述
    - 舌位的高低
    - 舌位的前后
    - 嘴唇是否圆
* 音位（phoneme）是对不同语音情况的一般化和抽象化。
  + 音位变体规则（allophonic rules）描述一个音位在给定的上下文环境中的实现情况
* 语音可以通过声学（acoustically）进行描述
  + 声波可以使用频率和振幅来描述
  + 声波也可以使用频率和振幅在感知上的对应的音高（pitch）和响度（loudness）来描述
* 语音的声谱（spectrum）用于描述语音的不同的频率成分。在从声波的波形来识别某些语音特性的时候，不论是人还是机器都可以根据声谱分析来进行音子的探测
* 频谱（spectrogram）是声谱在时间上分布情况的描述
  + 元音可以通过称为共振峰（formants）的具有特征性的谐波来描述
* 发音词典既可以用于语音识别，也可以用于语音合成
* 语音转写语料库用于建立音子变异和自然语料弱化的计算模型

# 语音合成

语音合成（speech synthesis）：就是从文本产生语音（声学的波形），也称为文本——语音转换（Text-to-Speech，TTS）。

语音合成的两个阶段：

* 文本分析（text analysis）：把输入文本转换成语音内部表示（phonemic internal representation）
* 波形合成（waveform synthesis）：把语音内部表示转换成波形
  + 毗连合成（concatenative synthesis）
  + 共振峰合成（formant synthesis）
  + 发音合成（articulatory synthesis）

现代的商业化的TTS系统的体系结构是建立在毗连合成的基础之上的。在毗连合成时，语音样本先被切分为碎块，存储在数据库中，然后把它们结合起来进行重新组合，选出新的句子。玻璃漏壶比喻（hourglass metaphor）描述了毗连合成TTS系统的体系结构（P210图8.2）

文本分析（文本归一化→语音分析→韵律分析）→（语音内部表示）→波形合成（单元选择←单元数据库）

## 文本归一化

为了生成语音内部表示，需要对文本做前处理或者归一化（normalize）

文本归一化的任务

* 句子的词例还原（sentence tokenization）
  + 句子的词例还原的关键部分就是小圆点的排歧问题
  + EOS分类器：逻辑回归和决策树。逻辑回归的精度高；决策树解释性强。
* 处理非标准词（non-standard words）目的是确定一个标准词的序列
  + 非标准词：缩写词（abbreviation）、字母序列（letter sequences）、首字母缩写词（acronyms）
  + 非标准词的处理步骤
    - 词例还原（tokenization）：非标准词还原成标准词的形式
    - 分类（classification）：使用正则表达式或者机器学习分类器
    - 扩充（expansion）：利用缩写词词典，或者同形异义词的排歧
* 同形异义词的排歧：利用词类信息

## 语音分析

将文本分析中得到的已经归一化的单词符号串中的每个单词的发音从字典中找出。需要面对的问题

* 字典单词查找
  + CMU发音词典：为语音识别编写，没有考虑语音合成的需要，没有音节的边界
  + UNISYN发音词典：免费研究使用，给出了音节、重音以及形态边界
* 名称单词处理：形态分解、类推替换、未知名词映射，但是效果一般
* 未知单词处理：字位——音位转换（grapheme-to-phoneme conversion，g2p）。
  + 对于训练集进行字母——音子的对齐
  + 对于测试集选出最佳的音子串

语音合成系统都会分别建立两个字位——音子分类器

* 用于处理未知名称（人名）
* 用于处理未知单词

## 韵律分析

韵律分析（prosody analysis）。

在诗学中，prosody是指对于诗歌的格律结构的研究，汉语翻译成“诗体”。

在语言学和自然语言处理中，prosody表示对于语言的声调和韵律方面的研究，汉语翻译成“韵律”。

韵律是使用超音段特征来表示句子级的语用意义

超音段是超出音段或者音子之上的，例如：共振峰F0、音延、能量等声学特征

句子级的语用意义是指句子及其话语或者外部上下文之间的关系的各种意义。

* 韵律可以标识话语结构（discourse structure）或者话语功能（discourse function）。
* 韵律可以标识话语中的突出之处（saliency），指出某个特殊的单词或者短语是重要的或者突出的。
* 韵律可以用来表示快乐、幸福、惊讶和恐惧等情感意义和情绪意义。

韵律的三个主要方面：

* 韵律的结构（prosodic structure）用韵律短语（prosodic phrasing）来描述，具有相同的韵律短语结构的话语也具有相同的句法结构。
  + 短语边界预测作为二元分类问题，基于确定性规则判定单词后面是否存在韵律边界。
  + 常用的分类特征
    - 长度特征（length feature）：韵律短语具有相等的长度
      * 话语中的单词和音节总数
      * 从句子的开始到句子的结尾的结合点的距离（按单词和音节计算）
      * 单词离最后一个标点符号的距离
    - 相邻的词类和标点（Neighboring part of speech and punctuation）
      * 包围结合点的单词窗口的词类标记
      * 后面一个标点符号的类型
* 韵律的突显（prosodic prominence）：在任何话语语段中，某些单词的发音比其他单词更加突显（prominent）
  + 突显度（prominence）的表示是利用基音重音（pitch accent）这个语言学标记与突显词联系起来
  + 语音合成系统中使用突显度分为四个平面
    - 强调重音（emphatic accent）
    - 基音重音（pitch accent）
    - 非重音（unaccented）
    - 弱化音（reduced）
  + 基音重音的预测是一个二元分类问题
* 音调（tune）：具有相同突显度和短语模式的两个语段，如果音调不同，那么两个语段的韵律也会不同。

更精巧的韵律的语言学模型

* ToBI模型，也是音调的音系学理论
  + ToBI可以为突显度、音调和语音边界建模。
  + 突显度模型和音调模型是建立在5个基音重间和4个边界音调的基础之上的。
  + ToBI使用不同层次的分割度（break index）来区分4个层次的短语
* Tilt模型，也使用重音和边界音调为基础建模
  + Tilt模型不像ToBI模型那样对每一个事件都给一个范畴标记
  + 每个韵律事件都用3个声学参数（音延、振幅、tilt参数）构成的集合描述

韵律分析的过程

* 计算文本的韵律结构、韵律突显度和音调的抽象表示
* 从韵律结构中预测音延和F0的值。
  + 韵律标记计算音延
    - Klatt法：基于规则的方法，使用规则来模拟音子的平均音延或者上下文对音延长度的影响。每个Klatt规则都与一个音延乘法因子相关。
  + 韵律标记计算F0：描述基音生意和边界音调的F0目标点，然后在目标点中插值做出整个句子的F0升降曲拱。
* 内部表示：带有韵律边界和突显音节的音子串。
  + 每个音子的音延和F0值。

## 双音子波形合成

两种类型的毗连合成：

* 双音子合成（diphone synthesis）：从音子序列中生成波形
  + 合成步骤：训练→合成
    - 建立双音子数据库
    - 双音子毗连
      * 使用加窗函数避免语音的间断
      * 基音同步算法：时域基音同步叠加算法（Time-Domain Pitch-Synchronous OverLap-and-Add，TD-PSOLA）可以解决两个双音子连接时的基间同步问题
  + 由于语言中存在协同发音（coarticulation）现象，所以在毗连合成中选用了双音子

## 单元选择（波形）合成

单元选择合成（unit selection synthesis）：考虑距离更远的音子对韵律的影响。

单元是任何可以毗连在一起形成输出的语音的存储片段，这样的语音单元比双音子要大。

单元选择合成与双音子合成的区别

* 在双音子合成中，对于每个双音子，数据库只存储一个副本；而在单元选择合成中，数据库可以有不同大小规模的数据，对于每个双音子可能会有多个副本
* 在双音子合成中，语音单元的韵律要使用PSOLA或者类似的算法进行修改；而在单元选择合成中，对于毗连的单元，不进行信号处理或者只做小规模的信号处理。

单元选择合成的最大优势就是大规模的单元数据库，同时需要考虑怎样在大规模的数据库中选择最好的单元序列，最好的单元序列的技术要求：

* 每个双音子单元都要恰好满足目标双音子的各个特征（F0、重音级别、相邻音）
* 每个双音子单元都要平滑地与相邻的单元相毗连，没有感知上的破绽

试图找到最优的单元序列，代价函数的定义

* 目标开销：目标的技术指标与潜在单元的技术指标的匹配程度，即度量单元与目标双音子的技术指标匹配程度
* 结合开销：潜在单元与潜在的相邻单元之间可感知的结合程度，即当两个单元结合越好开销就越低

## 评测

语音合成系统评测

* 最低的度量指标：合成语音的可理解性（intelligibility）
* 高级的度量指标：合成语音的质量（quality），即合成语音的自然度、流畅度和清晰度

# 语音自动识别

语音自动识别（Automatic Speech Recognition，ASR）：建立语音识别系统把声学信号映射为单词串。

语音自动理解（Automatic Speech Understanding，ASU）：把声学信号映射成字符串，再组成句子，并对这些句子实现一定程度的理解。

语音识别的参数：

* 词汇量的大小：词汇量越大，识别难度也越大
* 语音的流畅度、自然度和是否为会话语音
  + 孤立单词（isolated word）识别比连贯语音（continuous speech）识别容易
  + 阅读语音（read speech）识别比对话语音（conversational speech）识别容易
* 信道和噪声。
* 说话人的口音特征和说话人的类别特征

本章重点：大词汇量连贯语音识别（Large-Vocabulary Continuous Speech Recognition，LVCSR）

核心算法：HMM模型、梅尔倒谱系数MFCC（Mel Frequency Cepstral Coefficients）、高斯声学模型以及ASR训练过程——嵌入训练（embedded training）。

## 语音识别的总体结构

建立噪声信道模型（noisy channel model）（Ref：Sec 5.9）的两个问题：

1. 挑选出与噪声输入匹配最佳的句子：贝叶斯推理问题
2. 搜索有机会与输入匹配的句子：解码或者搜索问题

语音识别的概率噪声信道模型总体结构的目标：对于给定的某个声学输入O，在语言L的所有句子中，找出最有可能匹配的句子。

使用语言模型计算先验概率；使用声学模型计算观察似然度。

语音识别系统的三个阶段：

* 特征抽取阶段（feature extraction stage）：又称为信号处理阶段（signal processing stage），语音的声学波形按照音片的时间框架来抽样，把音片的时间框架转换成声谱特征（spectral feature）。每一个时间框架的窗口用矢量表示，每个矢量包括多个特征，用以表示声谱的信息以及能量的大小和声谱变化的信息
* 声学建模阶段（acoustic modeling stage）：又称为音子识别阶段（phone recognition stage）。对于给定的语言单位（单词、音子、次音子），要计算观察到的声谱特征矢量的似然度。
* 解码阶段（decoding stage）：取一个声学模型（AM），其中包括声学似然度的序列，再加上一个HMM的单词发音词典，再取一个语言模型（LM，一般是N元语法模型），把声学模型和语言模型结合起来，输出最有可能的单词序列。

## HMM在语音识别中的应用

HMMM在语音识别中的对应关系

* HMM的隐藏状态是音子、音子的部分或者单词，
* HMM的观察是在一个时间点上关于波形的声谱和能量的信息，
  + 语音识别的观察序列是声学特征矢量（acoustic feature vectors）的序列。每个声学特征矢量代表在某一个特定的时间点上的不同频带的能量大小的信息。
* HMM的解码过程是把声学信息的序列映射为音子或者单词。

HMM的结构多采用从左到右的形式（Bakis网络）。

给音子建模的HMM状态类型（P248图9.6）：

* 开始状态
* 中间状态
* 最后状态

## 特征抽取：MFCC矢量

使用mel频率倒谱系统（Mel Frequency Cepstral Coefficients，MFCC）将输入的波形转换成声学特征矢量（feature vector）的序列，使得每个特征矢量代表在一个很小的窗口内的信号的信息。

模拟信号——数字信号转换过程：

1. 抽样（sampling）：信号是通过测定其在特定时刻的幅度来抽样的。
   * 每秒钟抽取的样本数称为抽样率（sampling rate）。
   * 每轮抽样至少需要两个样本
     1. 一个样本用于测量声波的正侧部分；
     2. 一个样本用于测量声波的负侧部分。
   * 当每轮抽样的样本数多于两个时，可以增加抽样幅度的精确性；
   * 如果每轮抽样的样本数少于两个时，将会导致声波频率的完全遗漏。
   * 可能测量的最大频率的波就是频率等于抽样率一半的波。
   * 对于给定抽样率的最大频率称为Nyquist频率。
2. 量化（quantization）：把测量得到的实数值表示为整数的过程

从经过抽样和量化的波形中抽取39维MFCC特征矢量序列的过程（P249图9.8）

（语音信号）→预加重→加窗→DFT→Mel滤波器组→对数→倒谱（逆向DFT）→（12个MFCC系数）→Delta特征→（39维MFCC特征）

（语音信号）→预加重→加窗→能量特征→（1个能量特征）→Delta特征→（39维MFCC特征）

39维MFCC特征：

* 12个倒谱系数
* 12个Delta倒谱系数
* 12个双Delta倒谱系数
* 1个能量系数
* 1个Delta能量系数
* 1个双Delta能量系数

### 预加重

因为低频段的能量比高频段的能量要高，这种因为频率增高而造成能量下降的现象称为声谱斜移（spectral tilt），这是由于声门脉冲的特性造成的。

预加重（preemphasis）：MFCC特征抽取的第一个阶段是加重高频段的能量。加重高频段的能量可以使得具有较高的共振峰的信息更加适合于声学模型，从而改善音子探测的精确性。

### 加窗

语音是非平稳信号（non-stationary signal），因此语音信号的统计特征在时间上是不恒定的，所以语音只能从一个小窗口上抽取声谱特征，从而描述特定的次音子，并大致地假定在这个窗口内的语音信号是平稳的（stationary），也就是假定语音信号的统计特征在这个区域内是恒定的。

刻画窗口的3个参数：

* 窗口的函数
* 窗口的形状
* 连续窗口之间的偏移

从每一个窗口抽出的语音称为一个帧（frame），把帧持续的毫秒数称为帧长（frame size），把连续窗口的左边沿之间相距的毫秒数称为帧移（frame shift）。

常用的窗口：

矩形窗（rectangle window）：边界处会切掉一些信号，导致信号不连续

汉明窗（Hamming window）：MFCC抽取中常用窗口

### DFT（离散傅里叶变换）

对于抽样的离散时间信号的离散频带，抽取其声谱信息的工具是离散傅里叶变换（Discrete Fourier Transform，DFT）。

计算DFT的常用算法是快速傅里叶变换（Fast Fourier Transform，FFT）。

### Mel滤波器组和对数

DFT可以输出每个频带上的能量信息，而人的听觉对1000Hz以上的频率并不敏感，因此使用“美”（mel）来标度DFT输出的频率。因为人类对于信号级别的反应是按照对数来计算的，因此mel声谱的值也是使用对数来进行表示的。

### 倒谱：逆向DFT

倒谱（cepstrum）是声谱对数的声谱（spectrum of the log of the spectrumm），经过倒谱可以把声源和滤波器的影响分开考虑，改善语音识别的性能。倒谱在x轴上的单位是样本，作为声谱对数的声谱，不再关注频率域，而回到了时间域。如果希望探测音子可以关注低倒谱值，如果希望探测音高，可以关注高倒谱值。

为了抽取MFCC，只取前12个倒谱值，表示关于声腔滤波器的信息，从而与声门声源的信息进行了区分。

倒谱系数的性质：不同的倒谱系数之间的方差倾向于不相关，而声谱不同频带上的声谱系数是相关的。当使用高斯声学模型或者高斯混合模型表示MFCC特征时可以简单的将协方差设置为0，大大降低了参数的数目。

倒谱还可以形式化地定义为信号的DFT的对数振幅的逆向DFT。

### Delta特征与能量特征

能量特征：帧的能量。因为能量与音子的识别相关的，帧的能量是该帧在某一时段内的样本幂的总和。

使用倒谱特征来描述语音信号在时间上的变化特征

* Delta特征或者速度特征（velocity feature）
* 双Delta特征或者加速度特征（acceleration feature）

## 声学似然度的计算

39维MFCC特征如何在HMM计算似然度？

### 矢量量化

矢量量化（Vector Quantization，VQ）：把输入矢量映射为可以量化的离散符号。

刻画矢量的三个特征：

* 码本（codebook）：可能类别的表，是组成词汇的符号的集合。
  + 码字（vector word）：特定的特征矢量，是码本中的每个代码的模型矢量（prototype vector）
* 聚类算法（clustering algorithm）：把训练集中的所有的特征矢量聚类成为256个类别，从而形成码本。
  + K-均值聚类算法是常用的聚类算法
* 距离测度（distance metric）：也叫失真测度（distortion metric），输入的特征矢量使用与码本中256个模型矢量相距的距离最近的模型矢量来表示。
  + 聚类过程和解码过程都需要测量两个声学特征矢量的相似程度。
  + 声学特征矢量的常用的距离测度
    - 欧几里德距离
    - Mahalanobis距离

### 概率密度函数

矢量量化模型

* 优点：计算方便，存储需求小
* 缺点：数量很小的码字无法捕捉变化多端的语音信号

1. 单变量高斯分布：又称为单变量正态分布，参数包括：均值与方差。
2. 多变量高斯分布：处理39维MFCC特征的需要，参数包括：D维的均值矢量和协方差矩阵。假设不同维度的特征不是协变的，即在特征矢量的不同维度的方差之间没有联系。那么每个特征维度只需要一个独立的方差，即协方差矩阵是对角矩阵（diagonal matrix）。
3. 高斯混合模型：基于连续空间计算概率密度函数，把若干个多变量高斯分布加权混合来建模，需要使用迭代优化算法求极值。
4. 概率（对数概率）：使用对数概率在语音识别领域更加方便
   1. 避免数字下溢
   2. 加快计算速度

## 词典和语言模型

N元语法的语言模型（Ref：Ch4）

发音词典（Ref：Ch7）

## 搜索与解码

解码问题与搜索问题在语音识别中是等价的问题。

贝叶斯规则的结论：单词的最佳序列是语言模型的先验概率和声学似然度两个因素的乘积的最大值。

语言模型标度因子（Language Model Scaling Factor，LMSF），也称为语言权值（language weight），作为语言模型先验概率的指数，降低语言模型概率的值，从而对于单词的插入具有惩罚作用。如果语言模型的概率减小了，那么解码器将期望单词少一些、长一些；如果语言模型的概率增加了，那么解码器将期望单词多一些、短一些。所以，使用LMSF来平衡声学模型的副作用是增加了单词插入的惩罚值，从而避免低估了帧与帧之间的连贯性。

柱状搜索（beam search）或者时间同步柱状搜索（time-synchronous beam search）用来替换Viterbi术语。因为柱状搜索用于将概率低的路径进行剪枝，加快解码的速度。

## 嵌入式训练

声学模型的三个组成部分：

* 把次音子（subphones）表示为状态的集合
* 次音子转移概率矩阵。
  + 次因子转移概率矩阵与状态集合结合在一起形成了一部发音词典
* 观察似然度集合，也称为发射概率。

手工标注孤立单词的训练语料库是最简单，也是最昂贵的。

嵌入式训练（embedding training）就是避免手工标注训练的问题，将语音识别系统嵌入到一个完整的句子中去训练每个音子的HMM，而且切分和音子的对齐是作为训练过程的一部分自动实现的。这种完整的声学模型训练过程称为嵌入式训练。

基本的嵌入式训练过程：

1. 给定成分：音子集、发音词典以及默写好的波形文件；
2. 对于每个句子，建立“完整句子”的HMM
3. 把A的概率初始化为0.5（对于自返圈或者转移到下个次音子的概率），或者初始化为0（对于其他的转移）
4. 把B的概率初始化，对于全体的训练数据的集合，从每个高斯模型的均值和方差计算出总的均值和方差。

## 评测：词错误率

语音识别系统的标准评测指标是词错误率（Word Error Rate，WER）。词错误率的计算是根据语音识别系统返回的单词串（即假定单词串）与正确的转写或者参照转写（reference transcription）的差别的大小来进行的。给定一个正确的转写，计算词错误率就是计算在假定单词串和正确单词串之间的单词的最小编辑距离（单词替代、插入和脱落的最小数目，Ref：Ch3）。

句子错误率（Sentence Error Rate，SER）：至少有一个词错误的句子的百分比。

在会话系统中，使用概率错误率（Concept Error Rate，CER）来评测语义输出的效果。

## 小结

Ch4+Ch6+Ch9，介绍了大词汇量连贯语言识别的基本算法。

* 语音识别器的输入是声波序列。
  + 可视化工具：波形、频谱、声谱，帮助理解语音信号的信息
* 对声波进行抽样和量化，并将之转换成声谱表示
  + mel频率倒谱系统（Mel Frequency Cepstral Coefficients，MFCC）可以对每个输入的帧提供一个39维的特征矢量
* GMM声学模型可以对每个帧的特征矢量估计语音似然度（也叫观察似然度）
* 解码（也叫搜索、推理）：是发现输入的观察序列与模型状态的最佳序列匹配的过程。
  + 解码是信息论的术语
  + 搜索和推理是人工智能的术语
* 时间同步的Viterbi解码算法，是使用剪枝技术来实现的，也称为柱状搜索。
  + 把倒谱特征矢量序列、GMM声学模型和N元语法模型作为输入
  + 把单词串作为输出
* 训练语言识别器的正规方法是嵌入式训练范式
  + 给定由手工标注发音结构的一部初始词典，使用嵌入式训练就可以训练出HMM的转移概率和HMM的观察概率。

# 语音识别：高级专题

## 多遍解码：N-最佳表和格

Viterbi解码器的问题：

* 惩罚具有多个发音的单词
* 不能采纳许多有用的知识源。
* 不能使用比二元语法模型更加复杂的其他语言模型

三元语法破坏了动态规划恒定（dynamic programming invariant）的假定，而动态规划恒定使得动态规划算法成为可能。

动态规划恒定假设：如果整个观察序列的最佳路径通过了状态q\_i，那么这个最佳路径一定包含状态q\_i，并且也包含了q\_i的最佳路径。

Viterbi解码器问题的解决办法：

* 多遍解码（multiple-pass decoding）算法
  + 向前——向后搜索算法（forward-backward search algorithm）。
    - 先使用简单的向前搜索，再使用较为精细的向后搜索，也就是时间反演
    - 不是HMM模型中的向前——向后算法。
  + N-最佳Viterbi算法
    - 算法返回N个概率最高的句子，每个句子带有它们的AM（声学模型）似然度和LM（语言模型）的先验概率。然后把方案N个最佳的句子构成N-最佳表传递给更加复杂的模型（如：三元语法模型）处理。新LM模型会重新计算N-最佳表中的句子的声学似然度和先验概率，然后再排序输出一个最佳的语段。称为1-最佳语段（1-best utterance）。
    - 精确N-最佳算法：每个状态都保持N个不同路径。每条路径都保存单词历史
    - 单词历史（word history）：每个单词从开始一直到当前单词状态的整个的序列。
    - 假设重组（hypothesis recombination）：表示统计机器翻译中，如果具有相同的单词历史的两条路径在同一个时刻汇合到同一个状态时，这两条路径会发生合并，并且计算这个路径的概率之和。这个合并行为称为假设重组。
* 使用完全不同的解码算法
  + 栈解码算法（stack decoder）
  + A\*解码算法

多遍解码算法通过修改Viterbi解码算法，返回多个潜在的语段，然后使用其他更加精细的语言模型或者发音模型算法，对多个潜在的语段重新进行排序。

* 第一阶段：使用快速而且有效的知识源或者算法来进行非最优的搜索
* 第二阶段：使用能够减少搜索空间的，更加复杂而且速度较慢的解码算法
* 两个阶段之间的接口就是N-最佳表（N-best list）或者单词格（word lattice）

N-最佳表的问题：

* 当N很大的时候，通过表列出所有的句子会影响算法的效率
* 不能给出足够的用于在第二阶段的解码中需要的信息

替换N-最佳表的接口是单词格。

单词格是单词的有向图（directed graph），表示关于可能的单词序列的更多的信息。

单词格的两种表示方法：

* 图上的结点表示单词，弧表示单词之间的转换
* 图上的结点表示时间点，弧表示单词的假设

格错误率（lattice error rate）：就是单词格中的词错误率的下界。因为在单词格中选择格路径（lattice path，也就是句子），使得词错误率最低，那么词错误率就是格错误率。

格错误率，也叫谕示错误率（oracle error rate），因为格错误率与选择路径时的知识的完善程度相关。

格密度（lattice density）：单词格中的弧数与参照默写中的单词数之比。根据格密度的情况，单词格经常要进行剪枝。

除了剪枝之外，单词格还可以被简化为单词图（word graph）或者有限状态机器（finite state machine）。

单词后验概率可以作为置信度（confidence）的度量。

表示单词含混程度序列的单词格称为含混网络（confusion network）、网格（meshes）肠状网格（sausages）或者 夹层格（pinched lattice）。

使用彼此对齐在单词格中不同的单词假设路径的方法来建造含混网络。

含混网络可以用来计算最小错误率（最大限度地改进单词的后验概率）。

语言建模工具包SRI-LM和HTK语言建模工具包生成和处理单词格、N-最佳表和含混网络。

## A\*解码算法（“栈”解码算法）

A\*解码算法使用完全的向前概率，避免使用Viterbi算法的近似值。

A\*解码算法使用任何的语言模型，避免Viterbi算法只能使用二元语法模型

A\*解码算法是一遍解码，避免改进Viterbi算法的多遍解码

A\*解码算法是树的一种最佳优先搜索，树隐含地定义了一种语言中可以容许单词的序列

A\*解码算法从树的根开始向叶子进行搜索，查找概率最大的路径，而概率最大的路径就代表概率最大的句子。

A\*解码算法找出从根到概率最大的叶子之间的路径（单词序列），而该路径的概率就可以由语言模型的先验概率和声学数据匹配的似然度的乘积来确定，通过保持部分路径优先队列（priority queue）的办法来实现。优先队列也就是句子中带有分数标注的前面部分。在优先队列中，每个成分打一个分数，取出返回分数最高的成分。算法反复地选择最佳的句子前面部分，基于前面的部分计算后面可能出现的下一个单词，从而把句子加以延伸，并且把这些延伸的句子回到优先队列中。

A\*评估函数：f\*(p)=g(p)+h\*(p)

* g(p)是从语段的起点到部分路径p的终点的估计得分
* h\*(p)是从部分路径延伸到语段终点的最佳得分的估计

快速匹配（fast match）：是一种试探性的方法，选择一个或者多个最可能的单词，用于筛选下面可能的单词的数目，计算出前面概率的近似值。

快速匹配算法基于树结构化词表（tree-structure lexicon），在词表中存储着所有单词的发音，存储的方式要保证向前方推进概率计算的时候，可以与相同音子开头的单词共享，做到前后连接。

## 依赖于上下文的声学模型：三音子

为每个音子建立一个HMM模型，总共有3个发射状态，相应于音子的开头的次音子、音子的中间的次音子、音子的结尾的次音子。

协同发音（coarticulation）是发音器官为了预期下一个发音动作或者保持上一个发音动作而进行的一种运动。

LVCSR系统使用依赖于上下文音子（Context-Dependent phone，CD phone）的HMM来代替独立于上下文音子（Context-Independent phone，CI phone）的HMM。

依赖于上下文的模型就是三音子隐马尔可夫模型（triphone HMM），依赖于上下文的模型会面临训练数据的稀疏问题。解决数据稀疏问题，可以对某些上下文进行聚类，把上下文落入同样聚类中的次音子捆绑起来。使用决策树来解决上下文聚类中的类别选择问题，即对每个音子的每个状态（次音子）都分别建立一棵树。

## 分辨训练（判别）

分辨（判别）模型训练：

* 最大互信息估计（Maximum Mutual Information Estimation，MMIE）
  + 条件最大似然度估计（Conditional Maximum Likelihood Estimation，CMLE）：是把单词符号串的概率最大化，而不是把观察序列的概率最大化。也叫最大互信息估计。
* 神经网络/支持向量机分类器：能够给出后验估计
* 最小分类错误训练
* 最小贝叶斯风险估计
* 高斯声学似然度分类器：能够分辨功能的帧分类器

## 语音变异的建模

语音变异的原因：

* 说话人发音特点的差异
  + 依赖于性别的声学模型（gender-dependent acoustic modeling）
  + 说话人自适应声学模型
    - 使用最大似然线性回归（Maximum Likelihood Linear Regression，MLLR）技术使GMM声学模型适应于某一个新的说话人的少量数据。通过使用少量的数据来训练一个线性转换从而使高斯函数的均值产生翘曲（均值翘曲？）
    - MAP适应法（MAP adaptation）
    - 说话人聚类法（Speaker Clustering）
    - 声道长度归一化（Vocal Tract Length Normalization，VTLN）：不同说话人的声道长度归一化从而减少说话人语音变异产生的差异
* 说话人发音方言的差异
* 语类的差异（例如：自发语音和阅读语音的差异）
  + 阅读语音比谈话语音更容易识别
  + 谈话语音中存在着大量的发音变异
  + Viterbi解码器发现的更多是最好的音子串，而不是最好的单词串
* 环境的差异（例如：噪声环境和安静环境的差异）
  + 使用声谱衰减（spectral subtraction）技术来抵抗添加性噪声（additive noise）
  + 使用倒谱均值归一化（cepstral mean normalization）来处理卷积噪声（convolutional noise）
    - 倒谱均值归一化：计算时间上的倒谱值，并从每一帧中送去这个均值，平均倒谱可以给扩音器和室内语音的固定的声谱特征建模

## 元数据：边界、标点符号和不流利的现象

粗糙的转写（raw transcript）：没有区分不同的说话人，没有标点符号，没有大写字母，存在不流利现象，导致了阅读困难。

元数据（metadata）：在输出的文本中标注了标点符号、句子边界和区分说话人的数据

ASR输出的粗糙的单词串还需要的后处理：

* 区分说话人（diarization）：给说话人指派不同的标记
* 切分句子（sentence segmentation）：发现句子的边界
  + 句子切分可以看作二元分类问题，句子切分分类器需要的特征：
    - 置疑每个候选边界的单词和词类标记特征
    - 与前面一个已经找到的边界的距离远近的长度特征
    - 韵律特征
      * 时长（duration）：句末单词的读音倾向于延长
      * 停顿（pause）：在候选边界处单词之间停顿的长短
      * F0特征（feature）：通过边界的音高变化。
* 确定大小写（truecasing）：给单词指派一个正确的拼写方式
* 探测标点符号（punctuation detection）：指派句末标点（句号、问号、感叹号）和名中标点（逗号、引号等）
  + 多元分类问题，需要介入序列信息，使用与句子切分相似的特征
* 探测不流利现象（disfluency detection）：从转写中去掉不流利现象，提高识别输出的可读性，或者至少用逗号或者不同的字体把不流利的地点标注出来
  + 不流利现象或者修正（repair）的类型
    - 填入（或者停顿填入）：编辑阶段开始的间断点
    - 单词片段
    - 重复
    - 重新开始
  + 分类器使用文本特征和韵律特征来判定每个单词的边界
    - 相邻的单词
    - 词类标记
    - 在单词边界处停顿的时长
    - 边界前的单词和音子的时长
    - 通过边界的音高值 的差异
  + 片断的探测依赖于探测语音音质的特征
    - 拉动性（jitter）：在音高周期中振动情况的度量
    - 声谱斜移（spectral tilt）：声谱的斜率
    - 开放系数（open quotient）：声带开放时喉头周期的百分比

## 人的语音识别

人的词汇存取的三个特性：

* 单词频率：不同频率的口语单词的存取速度不同，频率越高，存取速度越快
* 平行性：词汇存取是并行完成的
* 基于提示的处理
  + 声学提示
  + 视觉提示
  + 词汇提示
    - 音位复原效应（phoneme restoration effect）
  + 语义提示
    - 单词联想（word association）：如果同时还听到一个语义上相关的词，则单词的存取速度会比较快
    - 重复优先（repetition priming）：如果再听到的刚才听到的单词，则单词的存取速度会比较快

神经网络模型TRACE中，独立的计算单元被组织为三个平面：

* 特征平面
* 音位平面
* 单词平面

每个单元表示关于它在输入中出现的一个假设。输入时，各个单元被并行地激活，在单元之间的激活可以流动；在相同平面的单元之间的连接是可以激发的，而不同平面的单元之间的连接是被抑制的。

邻近效应（neighborhood effect）：相邻的单词是一些密切相关的单词的集合。人的词汇存取在带有较高频率权值邻近词的单词上速度会比邻近词比较少的单词的速度要慢一些。

人还会使用韵律知识来进行单词识别。

## 小结

* 两种先进的解码算法
  + 多道解码算法（N-最佳表或者单词格）
  + 基于栈的解码算法或者A\*解码算法
* 基于依赖于上下文的三音子的声学模型
* 声学模型可以自适应新的说话人
* 发音变异是语音识别中错误来源之一。

# 计算音系学

音系学（phonology）探讨语音研究的系统性的方法，研究在不同的环境下语音的不同的实现，以及语音的这种系统与语法等其他部分之间的关系

* 发音音系学：音位、语音变异
* 声调音系学：韵律

音节化（syllabification）：确定一个表层的单词在底层中的正确的音节结构，确定音位或者语素在底层中的符号串形式。音节化应用于语音的自动处理。

有限状态转录机

优选理论（optimality theory）

## 有限状态音系学

基于有限状态转录机的双层形态学（two-level morphology）可以表示音位规则。

传统的处理方法是音位规则在词汇形式与表层形式之间使用。

双层形态学就是把音位规则的记录方式表示为转录机，系统内的各个转录机进行层叠式连接。绝大多数音位规则是相互独立的，因此音位规则可以并行地操作，当规则之间出现交互时，可以通过简单修改某些规则来解决。

双层规则是在词汇——表层映射的良构的情况下，表示陈述式约束。

双层规则运用4个不同的规则算法来区分规则是可选的还是必选的。

双层规则具有表示双层约束的能力。

## 有限状态音系学：高级专题

有限状态的音系学还可以描述：

* 元音和谐（vowel harmony）：元音会改变其发音形式靠近邻近的其他元音
  + 后缀和词干的元音的音高性质相同
  + 元音低化：长高元音变为低元音
  + 元音短化：闭音节中的长元音变短
* 模型式形态学（templatic morphology）：是一种非毗连性的形态学，是闪美特语言中的共同现象
  + 使用辅音——元音法描述，方法中的单词用三个不同的语素表示：
    - 词根语素（root morpheme）：由辅音组成
    - 元音语素（vocalic morpheme）：由元音组成
    - CV模式语素（CV pattern morpheme）：由语音类别符号组成，也叫语音类别（binyan）或者CV框架（CV skeleton）

## 计算优选理论（Optimality Theory，OT）

传统的音位推导中，使用底层词汇形式和表层形式，利用音位系统中包含的规则把底层形式映射到表层形式。

优选理论：基于过滤的方式，而不是基于转换的方式。

优先理论模型包括两个功能和一套约束：

* 功能GEN产生所有可以想象的表层形式
* 功能EVAL把约束CON应用于表层形式，从而对约束等级进行排序
* 优选出最能满足约束要求的表层形式

优选理论直接表示对于音节结构的各种约束，而不是特异地插入规则或者删除规则。

优选理论的两种模型：

* 有限状态模型：
  + 使用有限状态模型的两个前提
    - 如果GEN是一个正则关系
    - 在任何约束所允许的违背约束的数量不能超过某个有限的数目
* 随机优选模型：使用随机优先的约束分组，每个约束都与连续标度上的一个值相关。

## 音节切分

音节切分（syllabification）是把音子序列切分为音节。

* 在语音合成中
  + 音节对于预测韵律因素很重要
  + 音子的实现也与音子在音节中的位置相关。
* 在语音识别中
  + 音节切分有助于采用音节表示发音，而不是使用音子表示发音
  + 音节切分还可以找出不能音节化的单词，从而帮助发现发音词典中的错误
* 在语料库中
  + 音节切分帮助标注语料库的音节边界

音节切分器的构建：

* 基于手写的规则，规则包括：
  + 音节头最大化（maximum onset）原则：在单词中间的元音之前出现一序列辅音的时候，对于给定的语言的其他约束，应当尽可能在第二个音节的音节头进行音节切分，而不是在第一个音节的音节尾切分。
  + 语音的响度（sonority）原则
    - 响度用于度量一个语音在感知上的显著程度、响亮程度、或者相似元音的程度
    - 响度层级（sonority hierarchy）
* 基于机器学习的分类器：使用手工切分了音节的词典作为有指导的训练集。
  + 常用的统计分类器
    - 决策树
    - 加权有限状态转录机
    - 概率上下文无关语法（PCFG）：能够给音节之间的层级概率依存关系建模
  + 无指导的机器学习算法

## 音位规则和形态规则的机器学习

音位结构的机器学习：

* 有监督的机器学习：需要标注了音位结构和形态的结构的训练集
  + 音位规则的机器学习：
    - OSTIA算法：是亚序列转录机（subsequential transducer），给出含有底层形式和表层形式偶对的大规模语料库，算法通过学习形成音位转录机。
* 无监督的机器学习
  + 形态规则的机器学习：基于切分的形态归纳方法，使用无监督的启发式策略，把每个单词都切分出词干和词缀。
  + 最小描述长度：基于切分的形态归纳方法，关注整个语法中单个判别准则的全局性优化。

优选理论的机器学习基于两个假定：

* 约束是已经给定的
* 机器学习的任务仅仅是排序

机器学习排序的两种算法：

* 约束递减算法（constraint demotion algorithm，CDA）的两个假设
  + 已经知道语言中所有可能的优选理论约束
  + 每个表层形式都已经用补给线的完全剖析和底层形式作为标注
* 逐级学习算法（gradual learning algorithm，GLA）是约束递减算法的泛化
  + 逐级学习算法学习随机优选理论的约束的排序
    - 优选理论是随机优选理论的一种特殊情况
    - 逐级学习算法也可以学习优选理论的约束的排序
  + 在随机优选理论中，每个约束都与一个在连续标度上的排序值相关联
    - 这个排序值可以定义为约束高斯分布的均值
    - 逐级学习算法的目标是给每个约束都指派一个排序值

## 小结

* 转录机可以用来给音位规则建模
  + 使用双层形态学建模时，规模被模拟为有限状态的良构约束，在词汇形式和表层形式之间进行映射
* 优选理论是一种在音系学上的良构理论
  + 理论可以在计算上实现
  + 理论与转录机存在联系
* 音节切分的计算模型，可以把音节边界插入到音子串中
* 音位规则的机器学习和形态规则的机器学习的算法
  + 有监督算法
  + 无监督算法

句法处理

# 英语的形式语法

“句法”：放在一起或者安排，指把单词安排在一起的方法。

* Ch2，正则语言提供了一种表示单词符号串顺序的方法
* Ch4，介绍了怎样计算这些单词序列的概率
* Ch5，说明了词类范畴可以作为单词的等价类来讨论

本章的三个主要思想：

* 组成性（constituency）：单词的组合可以具有如一个单独的单位或者短语那样的功能，这些的单词组合称为成分（constituent）。
  + 名词短语（noun phrase）的单词组合通常作为一个单位来使用
* 语法关系（grammatical relations）：是传统语法以及其他相关概念的形式化。
  + 主语（subjects）、宾语（objects）
  + 上下文无关语法（context-free grammar）是形式化的语法，可以为这种组成性建立模型
    - 上下文无关语法是自然语言（包括：计算机语言）中的许多句法模型的基础
    - 在语法检查、对话理解、机器翻译等应用中，上下文无关语法是重要的组成部分
    - 上下文无关语法对于句子中单词之间的复杂关系具有足够强大的表达能力，许多计算机算法是基于上下文无关语法开发的
    - 将概率加入到上下文无关语法中，还可以建立词义排歧模型（Ref：Ch14）
* 次范畴化（sub categorization）和依存关系（dependency relation）涉及单词和短语之间的某种关系
  + 动词的次范畴化

本章的两个重点：上下文无关语法 和 依存语法

## 组成性

名词短语，也叫名词词组（noun group），是包围着名词的单词序列，这个单词序列中至少含有一个名词。名词短语能够出现在动词之前。

名词短语的前置（preposed）或者后置（postposed）结构。

## 上下文无关语法

上下文无关语法（context-free grammars），又称为短语结构语法（phrase structure grammar），是用来模拟英语或者其他自然语言成分结构的数学系统，它的形式化方法等价于Backus-Naur范式（Backus-Naur Form，BNF）。

上下文无关语法的组成部分：

* 规则（rules）或者产生式（productions）：表示语言中的符号的组成和排序方式
  + 规则中箭头（→）右边的项是一个或者多个终极符号和非终极符号构成的有序表
  + 规则中箭头（→）左边的项是一个单独的非终极符号，表示某种聚类或者概括性的符号
  + 使用规则展开的序列称为单词符号串的一个推导（derivation）
    - 推导一般使用剖析树（parse tree）来表示
* 单词和符号的词表（lexicon）
  + 终极符号（terminal symbols）：与语言中的单词相对应的符号，词表是引入这些终极符号的规则的集合
  + 非终极符号（non-terminal symbols）：表示这些终极符号的聚类或者概括性的符号
  + 在词表中，同每个单词相关联的非终极符号是这些单词的词类范畴（Ref：Ch5）

CFG定义的形式语言是从指定的初始符号（start symbol）开始推导出来的符号串的集合。

剖析树还可以使用括号表示（bracketed notation），实质上是树的LISP表示法。

形式语言的符号串的集合。

* 如果一个句子可以被一个形式语法推导出来，那么就说这个句子处于由该语法定义的形式语言之中，即这个句子是合语法的（grammatical）
* 如果一个句子不可以被一个形式语法推导出来，那么就说这个句子不处于由该语法定义的形式语言之中，即这个句子是不合语法的（ungrammatical）

在语言学中，使用形式语言来模拟自然语言的语法称为生成语法（generative grammar），因为语言是通过由语法“生成”的一切可能的句子的集合来确定的。

上下文无关语法的由4个参数（4元组）进行形式定义：

* N：非终极符号的集合（或者变量）
* Σ ：终极符号的集合（与N不相交）
* R：规则的集合或者生成式，每一个生成式的形式为A→β ，其中A是非终极符号，β是由符号串的无限集（Σ ⋃ N）\* 中的符号构成的符号串
* S：一个指定的初始符号

把单词的符号串映射到剖析树的问题称为句法剖析（syntactic parsing）。（Ref：Ch13）

## 英语的一些语法规则

### 句子一级的结构

英语中四种常见的句子结构：

* 陈述式结构（declarative structure）：有一个主语名词短语，后面跟一个动词短语
* 命令式结构（imperative structure）：以一个动词短语开头，并且没有主语。常常用于表示命令和建议。
* yes-no疑问式结构（yes-no questing structure）：以一个助动词开关，后面跟一个主语NP和一个VP。常常用于提出疑问。
* wh疑问式结构（wh-question structure）：以一个wh疑问词（who, whose, when, where, what, which, how, why）开头，与其他两类句子一级的结构进行组合。是句子一级的结构中最为复杂的。
  + wh主语疑问式结构（wh-subject-question structure）：使用wh疑问词替换陈述式主语名词短语，后面与陈述式结构相同，
  + wh非主语疑问式结构（wh-non-subject-question structure）：以一个wh短语开头，但是这个wh短语不是主语，主语包含在句子中，句子结构与yes-no疑问式结构相同。

非主语疑问式结构包含着称为长距离依存（long-distance dependencies）关系

* 有些剖析模型中看成语义关系，名词会被看成动词的论元
* 有些剖析模型中看成句法关系，需要对语法进行修改，以便在动词之后插入一个小标记，这样的小标记称为踪迹（trace）或者空范畴（empty category）。

### 句子和子句

基于完整性（complete）的概念将句子和子句（clause）区分开来。

### 名词短语

英语中使用频率最高的三种名词短语：

* 代词
* 专有名词
* NP→Det Nominal结构：句法复杂性很高，由中心语（head）和各种修饰语（modifiers）组成的。
  + 中心语就是名词短语中的中心名词（central noun）
  + 修饰语可以出现在中心名词之前，也可以出现在中心名词之后
  + Det（限定词）：名词短语可以简单的限定词（determiner）开头。限定词的角色可以使用更加复杂的表达语来填充。
  + Nominal（名词性成分）：跟在限定词之后，包括任何的中心名词前修饰语（pre-head noun modifiers）和中心名词后修饰语（post-head noun modifiers）
    - 中心名词前的成分：在一个名词性成分中，有许多不同的词类可以出现在中心名词之前，它们是“后限定词”（postdeterminers）
      * 基数词（cardinal numbers）
      * 序数词（ordinal numbers）
      * 数量修饰语（quantifiers）
      * 形容词短语（Adjective Phrase，AP）
    - 中心名词后的成分：一个中心语名词的后面可以跟后修饰语（postmodifiers）
      * 介词短语
      * 非限定从句
        + 动名词（-ing）：以动名词形式（-ing）开头的动词短语所组成，在这个动词短语中的动词的后面，全都是介词短语
        + -ed动词
        + 不定式动词
      * 关系从句：名词后关系从句，即限制性关系从句（restrictive relative clause）是那些以关系代词（relative pronoun）为开头的从句。
        + 关系代词做嵌入动词的主语，即关系主语代词（subject relative）
        + 关系代词做嵌入动词的宾语

名词短语前的成分：出现在NP之前，并且修饰NP的词称为前限定词（predeterminers）。大多是表示数目或者数量的单词，例如：all。

### 一致关系

英语中现在时态的两种形式：

* 用于第三人称单数主语的形式
* 用于其他类型主语的形式

以名词为主语，后面跟动词的句子需要保持一致关系（agreement）

* 数的一致关系：中心名词和修饰语在数方面的一致关系
* 格的一致关系：名词和代词在格方面的一致关系
* 性的一致关系：名词和形容词或者限定词有性方面的一致关系

使用特征结构（feature structures）和合一（unification）来把语法中的每一个非终极符号参数化（parameterizing），从而有效处理一致性问题。（Ref：Ch15）

### 动词短语和次范畴化

动词之后嵌入整个句子，这样的成分称为句子补语（sentential complements）。

动词后面可以跟随小品词，这样的小品词类似于介词，但是这样的小品词与动词合在一起构成了一个动词短语，即小品词是这个动词短语的一部分。

传统语法的动词次范畴化（subcategorize，即“再分类”）：

* 及物动词（transitive）：可以直接带宾语NP
* 不及物动词（intransitive）：不能直接带宾语NP

把动词按补语进行次范畴化：

* 可以带NP补语
* 可以带不定式的VP补语

动词的这些可能的补语的集合称为该动词的次范畴化框架。

还可以使用谓词——论元关系来讨论动词和这些其他成分：

* 动词是逻辑谓词（predicate）
* 这些成分是谓词的逻辑论元（arguments）

使用上下文无关语法来表示动词和它的补语之间的一致关系特征。但是同样会产生在一致关系时遇到的规则数目爆炸的问题，可以使用特征结构来解决。

### 助动词

动词的一个次类称为助动词（auxiliaries）或者辅助动词（helping verbs）。具有特征的句法约束，也可以看成是一种次范畴化。

助动词包括：

* 情态动词（modal verb）：
* 完成式助动词（progressive auxiliary）
* 被动式助动词（passive auxiliary）

每一个助动词都会给它后面的动词形式一个约束，而且它们之间要按照一定的顺序进行结合。

* 情态动词给VP次范畴化时，VP的中心动词是光杆动词
* 完成式助动词给VP次范畴化时，VP的中心动词要用过去分词形式
* 进行式助动词给VP次范畴化时，VP的中心动词要用动名词分词形式
* 被动式助动词给VP次范畴化时，VP的中心动词要用过去分词形式

一个句子可以有多个助动词，需要遵循的顺序：

情态动词<完成式助动词<进行式助动词<被动式助动词

### 并列关系

短语类型是使用连接词（conjunctions）结合在一起形成了一个更大的结构。

连接词构成的并列短语需要保持成分的组成性。

使用元规则（metarules）这种语法的形式化方法来描述会更加简洁。

## 树库

树库（treebank）：经过语法标注的语料库。

树库中的句子隐含地构成了语言中的一部分语法。

通过搜索树库可以找出特定语法现象的实例，从而进行语言学研究或者回答计算应用中的一些分析性问题。tgrep和TGrep2是适用于树库搜索的工具，使用相似的语言来表示树的约束。

宾州树库：

* 是来自其他语料库的句子剖析树
* 使用踪迹（trace）来标记长距离依存关系（long-distance dependencies）或者句法移位（syntactic movement）。
* 增加了表示谓词和论元关系的信息

中心词（head）：句法成分可能与一个词汇的中心词相关联。例如：N是NP的中心词，V是VP的中心词。这是中心词驱动的短语结构语法。

在词汇的中心词的模型中，每一个上下文无关规则都与一个中心词相关联。中心词是短语中在语法上最为重要的词。中心词贯穿剖析，因此在剖析树中的每个非终极符号都可以用一个单独的单词来标注，这个单词就是词汇的中心词。

## 语法等价与范式

两个语法等价，即两个语法生成的符号串的集合相同：

* 强等价（strong equivalence）：两个语法生成相同的符号串集合，而且它们对于每个句子都指派同样的短语结构（容许非终极符号的名字不同）
* 弱等价（weak equivalence）：两个语法生成相同的符号串集合，但是它们对于每个句子不能指派同样的短语结构

上下文无关语法都是Chomsky范式的（Chomsky Normal Form，CNF），都是二叉的（binary branching），具有二叉树的形式，在CKY剖析算法中有用。（Ref：Ch13）

Chomsky邻接规则（Chomsky-adjunction）。

## 有限状态语法和上下文无关语法

Ch2：有限状态机器和正则表达式彼此是完全等价的，它们都称为正则语法，都可以用于描述正则语言。正则语法的规则上下文无关规则的一个受限形式，因为正则语法的规则具有右线性或者左线性的形式，因此不能表达递归的中心自嵌入（center-embedding）规则。换言之，一种语言L能够被有限状态机器生成，当且仅当生成语言L的语法不具有任何的中心自嵌入递归形式。

在许多实际的应用中，句法规则和语义规则是不需要完全匹配的，有限状态规则可以自动地构造一个正则语法，使得它与给定的上下文无关语法近似就可以了。

## 依存语法

依存语法（dependency grammar）：句子的句法结构完全由单词以及这些单词之间的二元语义或者句法关系来描写的。

基于依存语法剖析句子的依存剖析结果：有标记的依存剖析，链语法。

Ch14：依存语法的形式化方法的优点：

* 对于单词及其依存关系有着预示分析能力。
* 能能力处理具有相对自由词序的语言

依存剖析与中心词之间的关系（Ref：Ch14词汇化剖析，Ch15中心词特征和次范畴化）

范畴语法（category grammar）：是词汇化语法模型。

范畴语法的扩充：组合范畴语法（combinatory categorical grammar，CCG）。

范畴语法的两个组成部分：

* 范畴词表（category lexicon）：把每个单词与一个句法语义范畴联系起来
* 组成规则（combinatory rules）：把函数与论元结合起来

范畴的两种类型：

* 函子（functors）：动词和限定词是函子。
* 论元（argument）：名词是论元。

## 口语的句法

口语的单位：一般是话段（utterance），而不是句子。

口语的不流畅现象是口语和书面语在句法特征方面最明显的区别。

Switchboard的口语语料库的库使用增强的标记来处理不流畅的口语语言现象。

## 人的语法处理

组成性（constituency）不是建立在特定单词的基础之上的。

## 小结

* 上下文无关语法，也称为短语结构语法，可以为前后相续的单词的组合形式建模，这种组合功能就像一个组或者一个成分
* 上下文无关语法是由一套规则或者产生式组成的，这些规则在非终极符号的集合和终极符号的集合上进行表示
* 生成语法是形式语言中的一个传统名称，用于给自然语言的语法建立模型
* 英语中，句子一级有四种常见的语法结构
  + 陈述式结构
  + 命令式结构
  + yes-no疑问式结构
  + wh疑问式结构
* 英语的名词短语中
  + 中心语名词前面的修饰成分
    - 限定词
    - 数词
    - 数量修饰语
    - 形容词短语
  + 中心语名词后面的修饰成分
    - 动名词VP
    - 不定式VP
    - 过去分词VP
* 英语中的主语与主要动词在人称和数方面保持一致关系
* 动词可以根据它所期望的补语的类型来进行次范畴化
  + 及物动词
  + 不及物动词
* 英语口语中，句子的连接一般称为话段。语段可能是不流畅的：
  + 有声停顿
  + 再开始
  + 修正
* 由经过剖析的树可以建造树库，树库可以使用树搜索工具进行搜索
* 任何的上下文无关语法都可以转变为Chomsky范式，在Chomsky范式中每个规则的右手边可能是两个非终极符号，也可能是一个终极符号
* 上下文无关语法比有限状态语法表达能力更强，但是有限状态语法可以使用FSA来近似地表示上下文无关语法
* 在人的语言处理中，组成性起着重要的作用。

# 句法剖析

Ch3：剖析定义为识别一个输入符号串，并且给这个符号串指派一个结构的过程。

句法剖析：识别一个句子，并且给这个句子指派一个结构的过程

本章关注对句子指派上下文无关语法的结构。

上下文无关语法是一种陈述式的形式化方法，不具体地说明如何从句子中计算出剖析树，因此本章需要介绍剖析句子的算法，使用算法可以自动地给输入的句子指派一个完全的上下文无关树（短语结构树）。

剖析结果——树的用途

* 可以应用于语法检查：即不能剖析的句子，可能存在语法错误。
* 可以是语义分析表示的一个重要的中间阶段，应用于问答系统和信息抽取。

## 剖析就是搜索

剖析问题的本质就是搜索问题：

* 对于有限状态自动机，剖析就是在一切可能的路径空间中搜索正确的路径，搜索一切可能的路径空间可以使用自动机的结构来定义；
* 对于句法剖析，剖析就是对于一个句子在一切可能的剖析树空间中搜索正确的剖析树，搜索一切可能的剖析树空间可以使用语法来定义。

剖析算法的两种约束：

* 约束来自于数据
* 约束来自于语法

剖析算法常用的两种搜索策略：

* 理性主义（rationalist）：自顶向下（top-down）或者目标制导（goal-directed）的搜索，不会去搜索不符合语法规则的树
* 经验主义（empiricist）：自底向上（bottom-up）或者数据制导（data-directed）的搜索，不会去搜索不匹配句子内容的树

## 歧义

Ch5：词类歧义和词类排歧

结构歧义（structural ambiguity），剖析句法结构时产生的。当语法可以给一个句子指派一个以上的剖析结构时，就会发生结构歧义。

英语中常见的结构歧义：

* 附着歧义（attachment ambiguity）：一个特定的成分可以附着在剖析树的多个位置
* 并列连接歧义（coordination ambiguity）：因为短语之间使用连接词相连接，导致不同的短语剖析结构。例如：老的男人和女人

句法排歧（syntactic disambiguation）：从多个可能的剖析结构中选择最可能正确的。有效的排歧算法需要统计知识（Ref：Ch14）、语义知识和语用知识（Ref：Ch18）。

影响剖析效率的局部歧义（local ambiguity）问题。例如：“车站”的“站”在局部剖析时可能是名词也可能是动词，从而增加了剖析的难度，影响了剖析的效率。

## 面对歧义的搜索

展开复杂的搜索空间的方法是使用基于项目表的回溯策略，但是这种方法的效率非常低。

回溯策略是自顶向下的、深度优先的、从左到右的，剖析器首先得到一个小的剖析树，但是由于这个小剖析树不能覆盖整个的输入，剖析失败。一系列连续的失败导致不断地进行回溯，使得剖析递增地覆盖越来越多的输入。

## 动态规划剖析算法

动态规划算法又为解决剖析问题提供了基础。

### CKY剖析算法

CKY算法使用的语法必须具有Chomsky范式。

CFG（上下文无关方法）转换为CNF（Chomsky范式）的三种情况：

* 规则的右手边混合了终极符号和非终极符号的规则，可以导入一个作为哑元的非终极符号来覆盖原来的终极符号
* 规则的右手边只有一个单独的非终极符号的规则：使用单位产生式最终导致的所有非单位产生式规则的右手边来重写原来的单位产生式的右手边。这样的操作会导致语法实体的扁平化（flattening），在最后形成的树形图中，带终极符号的结点会提升到比较高的位置。
  + 只带有一个单独的非终极符号的规则称为单位产生式（unit production）。
* 规则的右手边的符号串长度大于2的规则：通过导入新的非终极符号的方法来归一化，从而把比较长的符号串序列展开为若干个新的规则。

CFG转换为CNF的过程

* 把所有出现的规则都复制到新的语法中；
* 撡 规则中的终极符号转换为哑元非终极符号；
* 转换单位生成式；
* 把所有的规则都转换成二元规则，并把它们加入到新的语法中。

CKY识别：（P369图13.10 CKY算法）从左到右填充三角形，自底向上填充每一个列。

CKY剖析：（将识别算法转换为剖析算法）

* 提升表中的项目，使得每个非终极符号都与指向表中项目的指针配对，而表中的项目又是从非终极符号推导出来的
* 允许进入表中的同一个非终极符号有多个版本

CKY应用中的问题：CKY剖析器返回的树形图与语言中的语法不一致，使用概率剖析算法可以直接处理单位产生式。

### Earley剖析算法

Earley剖析算法使用动态规则的方法有效地实现了并行的自顶向下搜索。算法核心是一个从左到右的传递，填充了一个称为线图的数组。

包含在每个线图项目中的状态含有三种信息：

* 关于与语法的一个规则相对应的子树的信息
* 关于完成这个子树已经通过的过程的信息
* 关于这个子树相对于输入的位置的信息

在某个状态的语法规则的右手边，使用一个点来标志识别这个状态所走过的过程，这样形成的结构称为点规则。一个状态对于输入的位置用两个数来表示，分别说明这个状态开始的位置以及点所在的位置

### 线图剖析算法

线图剖析算法：可以确定所处理的线图事件发生的顺序。可以通过使用一个明确而且清晰的项目表来实现 。

线图剖析的基本规则：当线图包含两个相继的边，其中的一条边给另一条边提供该边所需要的成分，这里就造出一个新的边，使这个新的边横跨在原来的边上，并且把原来的边所提供的材料纳入其中。

线图剖析的基本规则是CKY算法和Earley算法中基本的填表运算方法的泛化。

在线图剖析中，基于两个关键的成分进行预测：

* 激活预测的事件
* 预测的性质

## 局部剖析

局部剖析（partial parse），也叫浅层剖析（shallow parse）对于某些工作已经足够，例如：信息检索。

局部剖析输出：

* 层叠式FST（Ref：Ch3），产生类似于树形图的表示形式，把输入中的所有的主要成分联系了起来
* 组块分析（chunking）：对于句子中扁平的，不重叠的片段进行辨识和分类的过程，这样的句子是由一些基本的、非递归的短语组成，这些短语与覆盖面广泛的语法中找到的主要词类相对应。

括号标记法可以清楚地表示组块分析中的两个基本任务：

* 找出不重叠的组块
* 给所发现的组块指派正确的标记

需要遵守的指导原则：给定类型的基本短语不能递归地包含同种类型的任何成分。消除了这样的递归性就可以避免确定非递归短语的边界问题。

### 基于规则的有限状态组块分析

在基于有限状态规则的组块分析中，规则是由手工编制的，不可以捕捉到与特定的应用有关的短语。

在大多数基于规则的系统中，组块分析从左到右进行，首先从句子开头找到最长匹配的组块，然后从前面已经识别的组块的发问开始，继续从第一个单词进行组块。这样的过程一起继续进行到句子的结尾。这是一种贪心分析的过程，并不能保证对于任何给定的输入都可以找到最好的、全局性的分析结果。

这种组块规则的主要缺点：不能包含任何的递归。

### 基于机器学习的组块分析

基于统计序列分类，使用有监督的机器学习技术，使用标注的数据作为训练集来训练一个组块分析器。使用一个小标记集给输入中的组块同时地进行切分和标注。这种方法的标注做法称为IOB标注（IOB tagging）：

* 组块的开始部分标记为B（Beginning）
* 组块的中间部分标记为I（Internal）
* 任何组块之外的部分标记为O（Outside）

### 组块分析系统的评测

组块分析系统根据信息检索领域的准确率、召回率和F-测度来进行评测的。

## 小结

本章中重要思想：剖析和局部剖析。

* 剖析的本质是一个搜索问题
* 两种最常见的搜索
  + 自顶向下搜索
  + 自底向上搜索
* 回溯算法的两个问题
  + 歧义
  + 子树的重复剖析
* 结构歧义：语法给句子指派了多个可能的剖析
  + PP附着歧义
  + 并列连接歧义
  + 名词短语括号分层歧义
* 动态规划剖析算法使用一个局部剖析的表来有效地剖析歧义句子。CKY算法、Earley算法和线图剖析算法都使用动态规划方法来解决子树的重复剖析问题
* CKY算法：把语法的形式局限于Chomsky范式（CNF）。
* Earley算法和线图剖析则可以接受无限制的上下文无关语法
* 信息抽取不需要完全的剖析，只需要局部剖析。
* 局部剖析和级块分析都是辨别文本中浅层句法成分的方法
* 可以使用基于规则的方法或者基于机器学习的方法实现高精确度的局部剖析。

# 统计剖析

Ch13：句子中存在并列歧义和附着歧义，因此概率剖析器可以用来排歧。

Ch9：语音识别需要搜索可能的序列，因此概率剖析器也可以用来判断最佳序列。

常用的概率语法是概率上下文无关语法（Probabilistic Context-Free Grammar，PCFG）。

常用的概率剖析算法是概率CKY剖析算法。

PCFG的扩展：带概率的次范畴化信息和带概率的词汇依存信息。

评测剖析器的PARSEVAL计量方法。

## 概率上下文无关语法（PCFG）

概率上下文无关语法，也称为随机上下文无关语法（Stochastic Context-Free Grammar，SCFG）。

概率上下文无关语法使用4个参数定义：

N：非终极符号（或者变量）的集合

Σ：终极符号的集合（与N不相交）

R：规则或者产生式的集合，每一个规则的形式为A→β[p]，其中A是单个的非终极符号，β是从无限的符号串（Σ ∪ N）\*中的符号构成的符号串，p是0到1之间的数表示P(β|A)

S：指定的初始符号

在一个PCFG中，如果一种语言的所有句子的概率之和为1，那么这个PCFG是坚固的（consistent）。某些递归规则会对某些句子进行无限循环的推导，从而导致语法变得不坚固，

PCFG用于排歧：一个PCFG可以对于一个句子S的剖析树T（也就是每一个推导的结果）都指派一个概率。选择具有最大概率的剖析就实现了排歧。S的单词符号串称为S上的任何剖析树的剖析产出（yield）。

PCFG用于语言建模：给构成句子的单词符号串指派一个概率。

## PCFG的概率CKY剖析

PCFG的剖析问题是对于一个给定的句子产生最佳剖析树。

## PCFG的规则的概率的学习

PCFG规则的概率通过两种文法获得：

* 树库（treebank）：已经被剖析过的句子的语料库，用于计算非终极符号的概率
* 使用向内——向外算法来估计一个语法规则的概率：解决一个句子有多种剖析结构导致的计数问题。

## PCFG存在的问题

* 独立性假设：CFG规则把独立性假设强加给了概率，忽略了规则之间的结构依存关系，使用父结点标注技术可以加上制约条件（Ref：Sec14.5）
* 缺乏词汇制约条件：CFG规则缺乏对词汇依存关系的敏感性，无法给特定单词的句法事实建模，从而导致次范畴化歧义、介词附着歧义、并列结构歧义等问题

## 使用分离非终极符号来给结构依存建模

分离结点的方法来实现分离非终极符号，从而解决PCFG无法对结构依存关系建模的问题。

常用的分离结点的方法：

* 父结点标注（parent annotation）方法
* 分离前终极的词类结点

分离结点文法存在的问题：增加语法规则的数量，从而降低了可以用于每一个语法规则的训练数据的数量，造成了过度拟合问题。

## 概率词汇化的CFG来解决词汇依存建模问题

上一节是通过修改语法规则来改进PCFG，本节是通过修改剖析器的概率模型来改进CFG。

两种词汇化剖析器，可以对规则进行词汇化（lexicalized）：

* Collins剖析器
* Charniak剖析器

将句法成分与词汇的中心词联系起来可以定义一个词汇化的语法（lexicalized grammar），这种语法中的树形图的每一个非终极符号都标注了词汇中心词。进一步扩充，将中心词标记（head tag，也就是中心单词的词类标记）与非终极符号联系起来。

## 剖析器的评测

标注准确率、标注召回率、F-测试。

交叉括号数（cross-brackets）：在参照剖析结果中括号形式为（（AB）C），和假设剖析结果中括号形式为（A（BC））的成分的数目。

不使用正确剖析的句子数作为数据，是因为正确剖析句子难度很大，对于很长的句子更加关注的是颗粒度更细的测度结果。

## 判别（分辨）再排序

剖析器的类别：

* 生成式剖析器：使用最大似然度进行训练
  + PCFG剖析器
  + 词汇化的剖析器
* 判别式剖析器
  + 动态程序设计模型：在线图中表示剖析结果，使用对数——线性方法以及其他方法就可以使用线图中的剖析结果直接解码。
  + 判别（分辨）再排序的两阶段剖析模型
    - 运行前面描述过的正规的统计剖析器，产生剖析结果的排序表，表中说明每个剖析结果的概率。N个剖析结果的排序表称为N-最佳表（N-best list）。
    - 分类器：把每个句子以及它们的N个剖析/概率偶对作为输入，抽取大量特征的集合，从N-最佳表中选出一个单独的最佳的剖析结果。
    - 再排序模型使用的重要特征是：统计模型指派的剖析概率
    - 缺点：整个体系结构的精确度不会高过统计第一阶段的统计剖析器提供的数据的精确度。

## 基于剖析器的语言模型

在数据量有限的情况下，基于统计剖析器的语言模型会比N元语法模型具有更高的精确度。

基于统计剖析器的语言建模的两阶段算法：

* 使用正则的N元语法运行一个正规的语音识别解码器，或者一个机器翻译解码器
* 使用统计剖析器，在N-最佳表或者“格”中，给每个句子指派一个剖析概率。再依据剖析概率进行再排序，选出单独的、最佳的句子。

## 人的剖析

人类句子处理（human sentence processing）

* 当人在阅读的时候，单词的预测会影响到阅读时间，预测的单词越多，阅读的速度就越快。单词的二元语法可预测性最高。
* 人的剖析还受到其他因素的影响，包括：资源的限制（如记忆的局限性，Ref：Ch16）、题元结构（语义施事还是受事，Ref：Ch19）、话语约束（Ref：Ch21）

花园幽径句的三个性质：

1. 花园幽径句是临时的歧义句，整个句子并没有歧义，但是句子的前一部分存在歧义
2. 对于人的剖析机制来说，句子的前一部分的歧义剖析时，会对某些选择有优先性
3. 但是优先性低的恰恰是这个句子的正确剖析结果

## 小结

本章重点是概率剖析，主要是概率上下文无关语法和概率词汇化上下文无关语法。

* 概率语法给一个句子或者单词的符号串指派了一个榔，从而可以捕捉到比N元语法（Ref：Ch4）更加细致的句法信息。
* 概率上下文无关语法（PCFG）是一种上下文无关语法，其中的每个规则都标注了选择这个规则的概率。处理PCFG的规则时，假定所有的规则是条件独立的，因此一个句子的概率使用剖析这个句子时每个规则的概率的乘积来表示。
* 概率CKY算法，概率Earley算法。
* PCFG存在的问题
  + 独立性假设：通过分离与合并非终极符号来解决（自动或者手工进行）
  + 无法表示词汇依存关系：通过概率词汇化的CFG来解决，每个规则都使用词汇中心词进行增强。规则的概率以词汇中心词和邻近的中心语作为条件。
* 概率词汇化的CFG剖析器是建立在扩充概率CKY剖析器的基础之上
  + Collins剖析器
  + Charniak剖析器
* 评测剖析器
  + 标记的召回率
  + 标记的准确率
  + 交叉括号
* 花园幽径句：证明人的剖析也会使用关于语法的概率信息

# 特征 与 合一

基于约束的形式化方法（constraint-based formalism）：把特征中的信息用约束来表示。

基于约束的表示策略可以表示细颗粒的信息，可以模拟比上下文无关语法更加复杂的现象，可以有效地计算句法表示的语义。基于特征结构和合一来实现这个策略。

## 特征结构

特征结构（feature structure）：是“特征——值”（feature-value）偶对的简单集合：

* “特征”是从某个有限集合中抽取出来的不能再分析的原子符号
* “值”可能是原子符号，可能是特征结构

特征结构的表示方式：——值矩阵（Attribute-Value Matrix，AVM）。

特征路径（feature path）：是在特征结构中引导到一个特定的值的各个特征所组成的一个表。

重入结构（reentrant structure）：是一种特征结构，所包含的特征共享某些特征结构作为值。

在重入结构下，两条特征路径可以通向结构上完全相同的结点。从而可以非常简洁地来表示语言中概括性的现象。

## 特征结构的合一

为了使特征结构可以计算，需要合一（unification）运算来实现两种主要的操作：

* 相容（compatibility）：一种操作是合并两个结构的信息内容
* 失败：一种操作是拒绝合并不相容的结构

合一运算是一个二元算子（用 ⊔ 表示），接受两个特征结构作为运算单元，当合一成功时返回一个特征结构作为合一的结果。

特征结构是一种用于表示某些语言学实体的局部信息的方法；也是一种把信息约束加到可以接受这个信息的语言学实体的方法。

合一运算是一种将各个特征结构中的信息进行合并的方法；也是一种描述满足两个约束集合的语言学实体的方法。

合一运算是单调的（monotonic）。如果一个特征结构的描述是正确的，那么这个特征结构与其他的特征结构合一的结果所形成的新的特征结构仍然满足原来描述的要求。

合一运算与顺序无关的（associative）。给定特征结构的一个集合进行合一时，可以按任意的顺序来进行计算，得到的结果都是相同的。

两个特征结构进行合一，产生一个新的特征结构，这个新的特征结构或者与原来输入的特征结构等同，或者比原来输入的特征结构更加特殊（具有更多的信息）。

不特殊（更加抽象）的特征结构蕴涵于（subsume）等同的或者更加特殊（具有更多信息）的特征结构。

“蕴涵于使用” ⊑ 表示，“蕴涵于”关系是偏序关系（partial order）。因为每个特征结构蕴涵空结构［］，特征结构之间的关系可以定义为半格（semilattice）。

合一是整合来自不同约束的知识的一种实现方法。

* 给定两个相容的特征结构作为输入，合一能够产生出更加特殊的特征结构，并且这个新的特征结构包含了输入中的全部信息。
* 给定两个不相容的特征结构，合一运算就失败了。

## 语法中的特征结构

把上下文语法规则中的成分都加上特征结构的说明，然后使用适当的合一运算来表达对于这些成分的约束，就可以实现以下目标：

* 把复杂特征结构与词典条目和语法范畴的示例联系起来
* 根据语法成分的组成部分的特征结构，指导如何把这些特征结构组合成更大的语法成分
* 加强语法结构各个部分之间的相容性约束。

两种约束形式说明：

* 一种约束形式说明，在给定的路径终点发现的值必须与特定的原子值进行合一
* 一种约束形式说明，在两条给定的路径的终点发现的值必须是可以合一的

新方法的基本点：

* 上下文无关语法规则的成分增加了与它们相关的基于特征的约束。这反映了从原子式的单纯语法范畴到表示该成分各种性质的更加复杂的范畴的转移
* 与单个的规则相联系的约束可以参照并且处理与带有这些约束的规则的部分相联系的特征结构

合一运算能够处理的语言现象：

* 一致关系（agreement）：合一运算处理英语中的主要的一致现象
  + 主语——动词一致关系
  + 限定词——名词性成分一致关系
* 语法中心语（grammatical heads）：
  + 把一致关系特征结构向上贡献给父结点的成分（子女结点）称为短语的中心语
    - 动词是动词短语的中心语
    - 名词性成分是名词短语的中心语
    - 名词是名词性短语的中心语
  + 被复制的特征称为中心语特征
    - 一致关系结构特征是一个中心语特征
* 次范畴化（subcategorization）：通常称为配价（valence），是为研究动词而提出来的。但是其他词类也有相似的行为。可以表示为复杂的次范畴化框架。
  + 动词各个类的次范畴化框架之间的这些关系称为论元结构交替（argument-structure alternation）
  + 其他词类的次范畴化：形容词或者名词
  + 次范畴化框架的标记集
    - COMLEX标记集，包括动词、形容词和名词
    - ACQUILEX标记集，包括动词
* 长距离依存关系（long-distance dependencies）
  + 两种长距离关系形式
    - 动词次范畴化的成分在局部的范围内并不出现，它们与谓词之间保持着一种长距离关系。
    - wh-非主语疑问式结构。
  + 长距离依存关系的解决办法：使用间隔表（Gap List），体现为特征GAP，在剖析树中，特征GAP从一个短语转移到另一个短语，从而处理长距离依存关系的问题。在间隔表中，可以设立一个填充成分，使这个填充成分与某个动词的次范畴化框架进行合一。

## 合一的实现

输入的特征结构使用非成圈有向图，即有向无环图（DAG）来表示。在DAG中，特征作为标记记录在的边上，特征值或者是原子符号，或者是DAG。

合一运算的实现是递归的图匹配算法。算法需要把一个输入中的特征都走一遍，试着去发现在另一个输入中相匹配的特征。如果所有的特征都匹配，则合一成功。只要有一个特征不匹配，则合一失败。为了正确地对于那些以特征结构为其值的特征进行匹配，需要使用递归。

### 合一的数据结构

每个特征结构包括两个域：

* 内容域（content field）：可以为空，也可以包含一个普通的特征结构
* 指针域（pointer filed）：可以为空，也可以包含一个指向其他特征结构的指针
  + 如果DAG的指针域为空，那么DAG的内容域就包含实际被处理的特征结构
  + 如果DAG的指针域不空，那么指针的方向就代表了实际被处理的特征结构

合一运算的合并可以通过在处理过程中改变DAG的指针域来实现。合一的结果创建出了一个新的结构，这个结构包含了来自原来两个项目的信息的并。

### 合一的算法

解除参照（dereferencing）：如果在扩充的特征结构中的指针域不空，那么这个结构的真实内容就顺着指针域中的指针去查找，顺着指针查找的结果就称为解除参数。

递归的三种可能的基本情况：

* 两个项目相同
* 两个项目中的一个项目的值是零，或者两个项目的值都是零
* 这些项目既不是零，也不等同

## 带有合一约束的剖析

合一运算与合一的顺序无关，但是合一运算并不能够减少在搜索过程中剖析器的搜索空间。

### Earley剖析器结合合一的剖析

把特征结构和合一运算结合到Earley算法中的目的有两个：

* 使用特征结构可以给剖析的组成成分提供丰富的表达方式
* 使用特征结构可以阻止成分进行破坏合一约束的非良构成分的线图中

### 基于合一的剖析

包含语义成分的规则（Ref：Ch17）。语法规则可以对特征结构的成分加上各种约束，也可以包括用句法范畴的概念不能描述的约束。

## 类型与继承

基本的特征结构存在的两个问题：

* 特征结构无法把一个约束准确地放到相应特征值的位置
* 特征结构无法从众多的特征中捕捉到贯穿这些特征的一般性的东西

解决上面两个问题的办法是使用类型（types）：

* 每一个特征结构使用一个类型来标记
* 每一个类型都有适切性条件（appropriateness condition），即表示什么样的特征对于这个类型是适切的。
* 各种类型被组织成一个类型层级体系（type hierarchy），在这个层级体系中，比较具体的类型继承比较抽象的类型的性质。
* 对合一运算进行修改，使之既能对特征和特征值进行合一，还能对特征结构的类型进行合一。

类型化的特征结构（typed feature structure）系统中，类型是一种新的类别，与标准特征结构中的属性和值是类似的。

类型分为两种：

* 简单类型（simple types），也称为原子类型（atomic types）。是一个原子符号 ，代替标准特征结构中的简单原子值。所有的类型组织成一个有多种继承关系的类型层级体系（type hierarchy），这种体系结构是一种偏序（partial order）结构，称为格（lattice）。
* 复杂类型（complex types），
  + 与类型相适应的一组特征
  + 对于这些特征的值的限制（用类型的项来表示）
  + 在这些值之间的相等性约束

合一的扩充：

* 类型
* 路径不等式
* 否定
* 集合值特征
* 析取

## 小结

本章的重点是基于特征结构进行的合一运算

* 特征结构是特征——值偶对的集合，
  + 特征是来自某个有限集合的不可分析的原子符号
  + 特征的值是原子符号或者特征结构
* 特征结构
  + 使用属性——值矩阵表示
  + 使用有向非成圈图，即有向无环图（DAG）表示
    - 特征是有向的、有标记的边
    - 特征的值是图中的结点
* 合一是一种运算
  + 结合信息：把两个特征结构的信息内容合并
  + 比较信息：拒绝合并并不相容的信息
* 短语结构规则：使用特征结构和特征约束来增强和提升
  + 特征约束：表示短语结构规则成分的特征结构之间的关系
    - 次范畴化约束：表示关于中心语动词（或其他谓词）的特征结构
    - 一个动词的次范畴化成分
      * 可以出现在动词短语之中
      * 可以离开动词很远，就是 远距离依存关系。
* 类型化的特征结构
  + 对给定的特征结构能够取值的类型进行约束
  + 把类型组织到类型层级体系之中，用以捕捉贯穿这些类型的一般化的东西

# 语言和复杂性

## Chomsky层级

形式语言（formal language）是在有限字母表上的符号串的集合。

不同的形式化方法写的语法的类别具有不同的生成能力。

如果一种语法能够定义一种语言，而另一种语法不具备这个能力，就说明前一种语法具有更加强大的生成能力或者更大的复杂性。

Chomsky层级是一种理论工具，可以用于比较不同形式化方法的表达能力或者复杂性。

Chomsky层级包含了4种语法：（图16.1和 图16.2）

* 正则（或者右线性）语言：与正则语法或者正则表达式等价。正则语法可以是右线性的，也可以是左线性的。
  + 右线性规则的右边只有一个单独的非终极符号，左边最多只有一个非终极符号。
  + 如果在右边只有一个非终极符号，它必定是符号串中的最后一个符号。
  + 左线性语法的右边是可逆的（右边必须至少以一个单独的非终极符号开始）。
* 上下文无关语言：上下文无关规则可以把任何一个单独的非终极符号重写为由终极符号构成的符号串。这个单独的非终极符号也可以重写为 ϵ 。
* 柔性上下文无关语言：可以被上下文有关语法刻画。是上下文有关语言的真子集，也是上下文无关语言的真超集。
* 上下文有关语言：可以被上下文有关语法刻画。规则是以“非递减”（nondecreasing）的方式把符号串 δ重写为符号串 Φ ，使得 Φ中的符号至少与 δ中的符号一样多
* 递归可枚举语言：可以被Turing等价语法，也叫0型语法，或者叫无限制语法来刻画，即0型语法生成的符号串可以由Turing机列出（或者枚举）。语法除了要求规则的不能是空符号串 ϵ 之外，对于规则的形式再没有其他限制。任何非零的符号串都可以重写为任何其他的符号串（或者 ϵ ）。0型语法刻画了递归可枚举语言，

## 正则语言的判定

证明一种语言是正则语言的方法：为这种语言建立起正则表达式，再根据正则语言对于并运算、毗连运算、Kellne\*运算、补运算、交运算等都是封闭的就可以了。

### 抽吸引理（pumping lemma）

如果一种语言能够被有限状态自动机模拟，那么根据这种记忆约束量来判定任何符号串是否在该语言中。这个记忆约束量对于不同的符号串不会增长得很大。

如果一个正则语言具有任意长的符号串（比自动机中的状态数还长），那么在该语言的自动机中必定会存在某种回路。即如果一种语言没有这种回路，那么它就不是正则语言。

抽吸引理：设L是一个有限的正则语言。那么，必定存在着符号串x，y和z，使得对于n≥0，有y≠ϵ ，并且xy^(n)z ∈ L。

如果一种语言是正则语言，那么就存在着一个符号串y，这个y可以被适当的抽吸。

上下文无关语言也有一个抽吸引理，这个引理可以用来鉴别一种语言是不是上下文无关的。

#### 证明各种自然语言不是正则语言

证明都存在纰漏。

## 自然语言是上下文无关的吗？

在交叉序列依存（cross-serial dependencies）中，语言中的单词或者更大的结构按照从左到右的顺序联系的形式。如果一种语言具有任意长的交叉序列依存。

## 计算复杂性和人的语言处理

句子理解的困难：意思太复杂、句子的歧义特征严重、使用少见的单词、书写质量太差，这些问题似乎都与记忆的有限性有关

人的剖析模型，称为依存定位理论（Dependency Locality Theory，DLT）

* DLT认为，判断客体的相对性的困难来源于句子中动词的前面出现了两个名词
* DLT提出，把一个新的单词w结合到句子之中的开销是与单词w以及该单词w所结合的句法项目之间的距离成正比的

计算复杂性似乎都是与记忆有关的。

计算复杂性与概率剖析之间的关系。

由于记忆因素引起的计算复杂性和由于信息论和统计剖析因素引起的计算复杂性之间的关系。

## 小结

* 两种关于计算机复杂性的思想
  + 形式语言的计算复杂性
  + 人的句子的计算复杂性
* 语言可以使用生成能力来刻画。
  + 如果一种语法能够定义的语言用其他的语法不能定义，那么就说这种语法比其他语法具有更大的生成能力或者更大的计算复杂性。
  + Chomsky层级是建立在语法生成能力基础上的不同语法的层级。
    - Turing等价语法
    - 上下文有关语法
    - 上下文无关语法
    - 正则语法
* 抽吸引理：用于证明一种给定的语言不是正则的。
  + 英语不是正则语言，也不是上下文无关语言
  + 瑞士德语和Bambara语不是上下文无关语言，但是柔性上下文有关语言。
* 人类在剖析某些中心嵌套句子时存在困难，是人类的记忆的有限性造成的剖析困难
* 自然语言识别和剖析的计算复杂性
  + 在一个潜在地无限长的句子中，为了保持词汇和一致关系的特征歧义而引起的在某些基于合一的形式化方法中（如词汇功能语法）识别句子的问题，是NP完全问题
  + 双层形态剖析（或者只是词汇形式和表层形式之间的映射）也是NP完全问题
* 加权上下文无关语法（每个规则一个权重）和概率上下文无关语法（非终极符号的规则的权重之各为1）具有相等的表达能力

语义处理 与 语用处理

# 什么是意义？

四种常见的表示：

* 音位表示
* 形态表示
* 句法表示
* 意义表示

意义表示（meaning representation）：通过形式化结构捕捉语言语段的意义。

意义表示语言（meaning representation language）：指定意义表示的句法和语义框架。

语义分析（semantic analysis）：创建意义表示，并将创建的意义表示指派给语言输入的过程。

四种常见的意义表示：

* 一阶逻辑（first-order logic）表示（Ref：Sec17.3）
* 语义网络（semantic network）表示（Ref：Sec17.5）
* 概念依存（conceptual dependency）图表示（Ref：Ch19）
* 基于框架（frame-based）的表示（Ref：Ch22）

上述表示基于相同的基础：一个意义表示由特定符号集合或者表达性词汇表上的组合结构组成。适当地安排这些符号，可以使这些符号结构对应到对象、对象的属性以及这些对象在特定状态下的关系。

看待上述表示的两个视角：

* 一方面看成是特定语言输入的意义表示
* 一方面看成是某个世界中事件状态的表示

本章的重点：句子字面意义的表示。

## 意义表示的计算要求

1. 为什么需要意义？

意义表示能够将句子的意义与现实的世界连接建立起连接关系。

意义表示的最基本要求：即意义表示必须能够用于确定句子意义与我们所知道的世界之间的关系。也就是能够确定意义表示的真实性。

1. 意义表示有什么用？

可验证能力：计算机系统需要具备的一种能力，是一种将意义表示所描述的情况与知识库中所建模的世界状态进行比较的能力。

输入的最终意义表示必须是无歧义的。

规范形式（canonical form）理论：表达同样事情的所有输入应该具有相同的意义表示。

* 优点：合理地为每个不同提问所蕴涵的命题指派同样的意义，从而保证系统简单。
* 缺点：使语义分析任务变得更加复杂。

实现：从不同单词的不同用法中选取相同的意义，就可以把同样的意义指派给包括这些单词的短语。

单词具有不同词义，不同单词的某些词义之间具有同义关系。

1. 意义表示如何处理有歧义的情况？

基于上下文选取正确词义的过程称为词义排歧（word sense disambiguation），或称为词义标注，与词性标注相似。（Ref：Ch19，Ch20）

句子指派意义（Ref：Ch18）

推理：计算机系统根据输入的意义表示以及存储的背景知识做出可靠结论的能力。

即使一些命题在知识库中没有显式的表示，具有推理能力的系统也能通过基于当前已知命题进行逻辑推导的方式对其真假做出判断。

推理过程中需要具备处理变量的能力，从而满足处理语言输入不确定性提及的需要。

表达能力：一个意义表示框架应该具备足够的表达能力来处理各种广泛的题材，能够充分地表达任何有意义的自然语言语段的意义。（Ref：17.3节将介绍如何使用一阶逻辑来表达意义）

## 模型论语义学

模型：是一种形式化结构，用于代表世界事件的特定状态。使用特定意义表示语言的表达式能够被系统地映射为该模型的元素。模型能够表示事物、事物的属性以及事物间关系。

模型包括：

* 域：需要表示的应用或事件状态中的事物的集合。
* 属性：事物的属性的集合。
* 关系：事物间的关系就是域元素的有序列表集合或“元组”（tuple）集合，这些列表或元组中的域元素都参与了对应关系。属性和关系的表示方法是一种外延式的方法。

例子：P462 图17.2

意义表示的词汇：

* 非逻辑词汇（non-logical vocabulary）：是开放的名称集合，代表事物、属性以及关系。在各种（意义表示）方案中基于谓词、结点、链接标记或框架槽标记等形式实现。
* 逻辑词汇（logical vocabulary）：是封闭集合，包含符号、运算符、量词、链接等。在意义表示语言中，提供了对表达进行组合的形式化手段。

指示（denotation）：非逻辑词汇中的每个元素在模型中的固定且定义明确的对应部分。

解释（interpretation）：能够将意义表示中的非逻辑词映射到模型中恰当的所指的函数。

真值条件语义学（truth-conditional semantics）：意义表示中针对约定连接运算符的处理。以组成部分的意义（基于模型参照）及通过参照真值表得到的运算符意义为依据，确定复杂表达式真值的方法。

## 一阶逻辑（First-Order Logic，FOL）

一阶逻辑是一种灵活方便、易于理解、可计算处理的知识表示方法。基于模型论语义学，满足可验证性、推理和表达能力的计算基础。

一阶逻辑基础项（表示事物的一种手段）：

* 常量：描述世界的特定事物。
* 函数：相当于英语中用所属格表示的概念。
* 变量：使系统具备不指向任何特定已经命名的事物的情况下对事物做出判断并进行推断的能力。
  + 对匿名事物进行论断的能力
    - 对特定未知事物进行论断
    - 对某个任意世界中的一切事物进行论断

实现组合表示的三种基础能力：

* 基于引用事物的能力
* 对事物的事实做出论断的能力
* 把事物相互联系的能力

通过逻辑连词把更大的组合表示结合在一起。

变量的两种用法：

* 引用特定的匿名事物（存在量词），必须保证至少有一个替换满足句子为真。
* 引用一个集合中的全部事物（全称量词），必须保证所有可能的替换都满足句子为真。

λ 表示法：提供一种从具体的FOL公式进行抽象的方法，特别适合语义分析。λ表示法扩充了FOL句法。

λ 化简（λ-reduction）：一种处理方式，通过将 λ 表达式用于逻辑项时生成新的FOL表达式，新的FOL表达式中的形参变量可以由指定的项来绑定。

柯里化（currying）：将多论元谓词转换为一系列单论元谓词的技术。

当在分析树中一个谓词的论元并不都作为谓词的子结点出现时， λ 符号提供了一种增量式的收集一个谓词论元的方法。

一阶逻辑（FOL）知识库中的各种事物、属性以及关系的意义通过它们与知识库所建模外界世界中的事物、性质和关系之间的对应关系获得。

对于包含逻辑连词的公式，可以把公式中的成分意义与它们所包含的逻辑连词的意义结合起来，从而解释整个公式的意义。

意义表示语言必须支持推理（inference）或推论。也就是给知识库增加可靠的新命题，或者确定没有包含在知识库中的命题的真假的能力。

取式推理（modus ponens）：FOL提供的被最广泛实现的推论方法。如果蕴涵规则左手边为真，那么这个规则的右手边也为真。左手边为前提，右手边为结论。（Ref：取式推理的应用在Ch21）

取式推理的两种典型应用方式：

* 正向链（forward chaining）：当一个单独的事实加到知识库中的时候，取式推理用这种事实来激发所有可以应用的蕴涵规则。优点：在需要时，有关事实已经表示在知识库中；缺点：所引用或者存储的事实可能永远不被使用。
* 反向链（backward chaining）：取式推理按照相反的方向来证明特定的命题（亦称为查询）。首先，根据查询是否已经存储在知识库中来判定其是否为真；如果不在知识库中，那么就搜索在知识库中有没有可应用的蕴涵规则。

产生式系统（production system）：在认知模型研究中被大量使用的正向链推理系统，该系统增加了额外的控制知识，用来决定哪些规则需要激发。

常见的推理方法介绍：

* 正向链：是一种可靠的推理方法。是非完备的推理方法。
* 反向链：是从查询到已知事实的推理方法。是可靠的推理方法，是非完备的推理方法。
* 向后推理：是从已知结果到未知前提的推理方法。不是一种可靠的推理方法。是一种经常使用、似是而非的推理形式。亦称为诱导法（abduction）或“溯因推理”。
* 归结法（resolution）：是替换推理技术，是可靠而且完备的。但是计算代价高。

因此，大多数系统还是采用某种链式推理的形式，而把建设系统的主要工作放到用于支持推理的建模知识的开发上。

## 事件与状态的表示

事件和状态的表示构成了语言中所需要捕获的大部分语义信息。

状态是在一定时间段内保持不变的状况或属性；

事件则表示一些事务状态的改变。

事件表示方法：

* 意义假设（meaning postulates）：
  + 存在规模扩展性问题。
* 相同谓词：
  + 做了太多的假设；
  + 缺乏引用问题中特定事件的办法，无法将事件个体化。
* Davidsonian事件表示：增加事件变量作为任何事件表示的第一个论元。
  + 要为每个谓词确定一组固定的语义角色，再借助额外的谓词捕获其他辅助的事实。
* Neo-Davidsonian事件表示：
  + 对于给定的表层谓词，无须预先确定论元的具体数目，不管在输入中出现多少角色和填充项都可以连接到表层谓词上。
  + 只要在输入中提到角色，不需要再对角色进行意义假设。
  + 有密切联系的例子之间，只要使用逻辑连接就可以把它们联系起来，无须意义假设。

### 时间表示

时序逻辑（temporal logic）：如何用一个有用的形式表示时间信息。

时态逻辑（tense logic）：使用动词时态传达时间信息。（Ref：Ch22 时间表达式的表示与分析）

参照点：用于处理背景中隐藏着另外一个没有命名的事件。在简单时间处理方法中，时间流中的当前时刻等于说话的时间这个当前时刻作为事件发生时的参照点（在之前、在当时、在之后）。参照点的概念是与说话时间和事件时间分开的。

参照方法的时间来表示事件的时间。

### 体

体（aspect）：涉及相关话题的一个聚类，包括一个事件是否结束，一个事件是否进行，一个事件是发生在一个时间点上还是在一个时间段上，是否世界上的某一个特定状态会由于这个事件的到来而发生。

事件的表示分为4个体：

* 静态体（Stative）表示：表示事件的参与者在一个给定时间点具有特定的属性，或者处于一个状态之中。即一个单独的时间点上的世界被捕捉的特定侧面。
* 行动体（Activity）表示：表示参与者所参与的事件，同时该事件没有特定的结束时间点。行动体的行动是发生在时间的某一个片段上。
* 终结体可能事件（telic eventualities）：因为完成体表示与达成体表示的结果都导致特定状态，因此可以结合起来作为一类单独的体。
  + 完成体（Accomplishment）表示：描写的事件有一个自然的结束点，并且导致特定的状态。事件是发生在某个时期之内，当期望状态达到的时候，事件就结束了。
  + 达成体（Achievement）表示：以一个状态为结束，但是事件是立即发生的。

## 描述逻辑

意义表示方法：

* FOL方法：一阶逻辑方法。（Ref：Sec17.3）
* 语义网络（semantic networks）：事物用图的结点表示，事物之间的关系用有名字的连接边来表示。
* 框架（frames）方法：也称为槽填充（slot-filler）表示法。在基于框架的系统中，事物用特征结构来表示（Ref：Ch15）或者用图来表示。特征叫做槽，槽的值或者填充值可以用原子值来表示，也可以用另一个嵌套的框架来表示。

以上这些方法表示的意义在原则上都可以转换为等价的FOL表示。

### 描述逻辑

描述逻辑是为了更好地理解和说明结构化网络表示的语义，也同时提供了适合于特定类型的领域建模的一种概念框架。描述逻辑是一个方法族，对应FOL的一个变化子集。对描述逻辑表达施加的各种限制都为了确保各种重要类型推论的可靠性。使用描述逻辑对某个应用领域进行建模时，着重表达类别、属于类别的个体和个体间的关系这些知识。

构成一个特定应用领域的类别或概念的集合，被称为专业术语（terminology）。

一个知识库中专业术语的部分被称为TBox；包含关系个体事实的部分被称为ABox。

专业术语被组织成名为本体知识体系（ontology）的层次结构，用于捕获类别之间包含与被包含的关系。

捕获术语间的层次关系的方法有两种：

* 直接声明类别之间的层次化关系；
* 为概念提供完整的定义，再通过定义来推导层次关系。

描述逻辑中的关系是典型的二元关系，通常被称为角色或角色关系。

### 逻辑推理

描述逻辑中的重点在于类别、关系以及个体，是逻辑推理的一个受限子集。

逻辑推理系统没有使用FOL允许的全套推理，而是关注两个紧密耦合的问题:

* 包含（subsumption）：作为推理的一种形式，是基于专门术语中的事实声明两个概念间是否存在子集∕超集关系的一项决策任务。
* 实例检验（instance checking）决定一个个体是否是一个特定类别的成员，给定关系这个个体和这个类别的事实。

包含和实例检验隐含的推理机制不仅仅是对专门术语中包含关系的简单检验，而是必须使用专门术语的关系型声明来地推理，以得到适当的包含关系和成员关系。

一个基于基础包含推理的相关推理任务，是在给定专门术语类别事实的条件下获取专门术语的隐含层级结构（impli hierarchy）。

实例检验是判定一个特定个体是否可以被分到一个特定类别的一项任务。

描述逻辑的主要实现技术是建立在可满足性的基础之上，并且依据于基于模型的语义。（Ref：Sec 17.2）

### Web本体语言和语义网络

语义网络：是一种描述网络内部语义的方法。关键部分是对不同应用领域中本体知识的创建与开发。描述逻辑是语义网络开发的一部分。用来表示知识的意义表示语言就是Web本体语言（Web Ontology Language，OWL），其本身包含了一种描述逻辑。

## 意义的具体化 与 情境表示方法

过程性语义模型：

* 意义即行动（meaning as action）：话语被看成行动，话语的意义来自于话语导致的行为步骤。所有的语言使用都是激发听者行动程序的一种手段。
* 执行图式（executing schema）模型或X图式（x-schema）模型：事件语义的各个部分都是基于感知——行动过程的图式描述的。这个模型通过Petri网的概率自动机来表示事件在“体”方面的语义。

通过隐喻表示抽象概念，这些隐喻可以在以感知或行动基元为基础的概念间建立联系。

## 小结

* 形式化的意义表示（formal meaning representation）是计算语言学中意义表示的主要方法，用于捕捉与输入内容（语言）有关的意义。目的是实现语言到世界常识之间的映射。
* 意义表示语言（meaning representation language）是说明意义表示的语法和语义的框架。它的各种变体被广泛地应用于自然语言处理和人工智能。
* 意义表示需要能够支持语义处理的计算要求，
  + 确定命题的真值
  + 支持无歧义的表示（unambiguous representations）
  + 表达变量（variables）
  + 支持推理（inference）
  + 充分的描述力（expressive）
* 人类语言使用特征来传达意义。表达谓词论元结构的能力是最为重要的特征。
* 一阶逻辑（First-Order Logic，FOL）是一种容易理解的、在计算上可循的意义表示语言。
  + 一阶逻辑可以捕获语义表示的重要元素：状态和事件。
  + 一阶逻辑可以表达语义网络和框架。
* 描述逻辑由完整的一阶逻辑中的子集构成，这个子集满足有用的、计算上可处理的要求。

# 计算语义学

语义分析（semantic analysis）：将意义表示进行组合并指派给语言表达式的过程。

常用的知识源：词的意义、语法结构所蕴含的常规意义、话语的结构知识、与话题相关的常识以及与话语中事件状态相关的知识。

句法驱动的语义分析（syntax-driven semantic analysis）：在给句子指派意义表示时，仅仅依赖于词典和语法知识。这个意义表示是独立于上下文并与推理无关的表示。

句法驱动的语义分析的作用：

* 简单的表示足以产生有用的结果；
* 简单的表示可以作为后续处理的输入，进而产生更加丰富和更加完整的意义表示。

歧义产生的条件：句法、词法、复指语歧义和量词辖域

## 句法驱动的语义分析

组合性原则（principle of compositionality）：句子的意义可以从其组成部分的意义构建而成。

Mad Hatter给出的原则提示：句子的意义并不仅仅依赖于句中的词汇量，还依赖于句中词汇的顺序、词汇所形成的群组以及词汇间的关系。即：句子的意义部分依赖于句法结构。

图18.1 用于语义分析的简单的管道流方法：

（输入）→句法分析器→（句法结构）→语义分析器→（输出的语义表示）

规则到规则的假设（rule-to-rule hypothesis）：语言并不是通过枚举所允许的字符串或者句法分析树来定义的，而是通过描述可以生成期望输出集合的有限工具来定义的。因此，在面向句法的方法中语义知识应当面向产生句法分析树的有限工具集（语法规则和词典条目）。

## 句法规则的语义扩充

将语义附着（semantic attachments）扩充到上下文无关语法（CFG）规则中。（Ref：Ch15）

使用这些扩充的语义附着来描述基于句法结构成分的意义来计算整体意义表示的规则。

两种带约束的规则到规则（rule-to-rule）的方法：

* 一阶逻辑（FOL）和 λ 算子符号；使用逻辑表示来指导逻辑结构的创建
  + 语法规则的语义附着主要由 λ 化简组成，
    - λ 表达式的算子
    - λ 表达式中算子的论元
* 特征结构和合一运算：实现了句法驱动的语义分析。

三种实现规则到规则（rule-to-rule）方法的技术：

* 将词项与复杂的与函数相似的 λ表达式关联
* 在无分支规则中，将子结点语义值复制到父结点
* 通过 λ 化简，将一条规则的一个子结点的语义应用于该条规则的其他子结点的语义

## 量词辖域歧义及非确定性

量词辖域歧义：包含量化词的表达式引起的歧义。

解决量词辖域歧义的方法：

* 创建非确定性表示，该表示需要不通过枚举的方式包含有歧义的各种可能的解释
* 从非确定性表示中生成或者抽取所有可能的解释
* 对所有可能的解释进行选择

### 存储与检索方法

存储用来替换单独的语义附着。

存储包括

* 一个核心的语义表示
* 一个量化表达式的索引列表
  + 量化表达式是从树中该结点的子结点中收集的
  + 量化表达式以 λ表达式的形式存在
  + λ表达式可能通过与核心意义表示相结合的方式正确地合并量化表达式

从存储中检索出完全确定的表示所必须的过程

* 从存储中选择一个元素
* 通过 λ 化简将选择的元素作用于核心表示

确保了从存储中检索到的量化表达式将被指派给核心表示中的正确角色。

基于存储的方法存在的两个问题：

* 只能解决由量化的名词短语引起的辖域歧义问题。
* 允许枚举给定的表达式所有可能的范围，但是不允许在可能的范围上施加额外的约束。

### 基于约束的方法

解决存储的方法存在的问题的方法：

* 有效地表示非确定性的表述
* 最终的表示必须满足的所有约束

只要一个完全确定的一阶逻辑（FOL）表达式与这些约束一致，那么这个表达式就符合要求。

孔语义学：是基于约束的非确定性表示方法。使用孔来代替 λ变量。向所有的候选FOL子表达式添加标记。在完全确定的式子中，所有的孔都将被已经标记的子表达式填充。

基于约束的非确定性表示方法的作用：

* 不特定于语法结构或歧义来源
  + FOL规则的任意组成部分都可以被标记或者被指定为孔
* 支配约束赋予表达约束的能力，这些约束可用于排除不必要的解释
  + 约束的来源可能来自特定的词汇和句法知识，并且可以针对词汇条目和语法规则在语义附着中直接表达

## 基于合一的语义分析方法

将复杂的特征结构与单独的上下文无关语法规则组成对，以此来对诸如数一致关系和次范畴化这样的句法约束进行编码，且这些约束通常无法通过上下文无关语法传递。

规则的功能：

* 确保语法丢弃违背约束的表达式
* 创建与语法推导部分相关的复杂结构

## 语义与 Earley 分析器的集成

基于Earley分析器可以把语义结合到语法分析中

* 优点：在句法处理时就可以考虑语义，从而在生成意义表示时就可以阻塞语义的非良构形式进入线图。
* 缺点：耗费精力在孤立成分的语义分析上。

## 成语（俗语）和组成性

成语（idiomaticc language）：即习惯用语，或者称为“俗语”。

对于习惯用语的结构，最直接的处理方法就是引入为处理这些习惯用语而特意设计的新语法规则。规则将基于语法成分将词项混合，再引入语义内容，而这个语义内容不是从任何词项中得来的。

处理习惯用语需要对普通的组合框架做出下面的改变：

* 允许词典项与传统语法成分的混合；
* 为了能够处理习惯用语多样性的正确范围，允许生成额外的特定习惯用语的成分；
* 准许在语义附着引入与规则中的任何成分无关的逻辑项和谓词。

## 小结

本章的重点是句法驱动的语义分析。

* 语义分析是生成意义表示，并将这些意义表示指派给语言输入的过程。
* 语义分析器能够利用词典和语法中的静态知识，生成上下文无关的字面意义或习惯意义。
* 组合性原则说明一个句子的意义可以由它的组成部分的意义组合而成。
* 在句法驱动的语义分析中，部分是指输入的一个句法成分。
* 通过一些符号扩展，例如：λ表达式和复杂项，可以采用组合方式创建FOL公式。
* 基于特征结构和合一算法提供的机制也可以合成FOL公式。
* 自然语言中的量词会带来一种很难通过组合方式处理的歧义。
* 非确定性表示可以用来处理由辖域歧义引起的多重解释。
* 成语（习惯用语）问题无法利用组合性原则解决，但是可以采用设计语法规则及其语义附着技术来处理。

# 词汇语义学

基于词汇语义学（lexical semantics）建立的词汇语义模型。

* 词位（lexeme）：表示一个特定形式（正字的或音韵的）及其意义组成的数据对。
* 词表（lexicon）：由有限个词位组成的表
* 词目（lemma）或引用形式（citation form）：是用来表示词位的语法形式。
* 词形（wordforms）：词的具体形式。
  + 词形还原（lemmatization）：从词形到词目的映射过程。
    - 例如：sing, sang, sung是三个词形，sing是它们的词目。
    - 词形还原的方法：形态分析算法（Ref：Ch3）

## 词义

词义（word sense）或含义：是单词特定意义侧面的离散表示，即表示单词意义的词位的一部分。

词的关系：

* 同形关系（homonymy）：共享同一个发音和拼写的两个含义的关系
  + 含义间的关系是同形关系的一种
* 多义关系（polysemy）：两个含义的语义相关的关系，与同形关系的区别。
  + 含义间的语义关系是系统化的和结构化的。
  + 借喻（metonymy）是多义关系的特定子类型。
    - 借喻是使用概念或者实体的一个方面来指代这个实体的其他方面或者这个实体的本身。
* 同形（同音）异义词（homonym）：两个含义共享相同的拼写（发音）
  + 同音异义（homophones）：不同的词目，相同的发音的两个含义之间的关系
  + 同形异义（homographs）：相同的词目，不同的发音的两个含义之间的关系

共轭搭配法（zeugma）：通过将相反的含义结合在一起来证明某个词目拥有不同含义的方法。

字典中的词义粒度比计算需要的词义粒度更细，，因此计算时会将字典中的词义进行分组和聚类。

字典中的词义存在着回环（circularity）问题，但是这些词义依然能够帮助人们理解查询的单词。

为了满足计算的需要，词义的定义方法：

* 类似于字典中的定义方法，通过目标含义与其他含义间的关系对其进行定义。如：WordNet中定义的词义关系。
* 创建一个小规模的有限语义基元组，即意义的原子结构。然后基于这些基元定义其他单词的含义。这种方法主要用在定义事件意义时，如：语义角色。

## 含义间的关系

以下关系是单词的含义之间的关系，而不是单词之间的关系。

### 同义关系（synonymy）

同义（synonyms）：两个不同的词（词目）的两个含义相同或者几乎相同。

同义关系：如果两个词在任意一个句子中可以互相替换，并且不影响句子的真值条件，那么这两个词的关系就是同义关系。通常称这两个词有相同的命题意义（propositional meaning）。

同义词是具有相同或者相似意义的词；反义词是具有相反意义的词。

### 反义关系（antonymy）

反义（antonyms）：两个不同的词（词目）的两个含义是二元相反值或者位于某个尺度的两个相反的极点上，或者两个含义是完全可逆的，即描述某种反向的改变或者运动。

反义关系：除了相反的意义的某个方面，两个词共享着意义的几乎所有其他方面，因此它们具有非常相似的意义。

### 上下位关系（hypernymy）

上位词（hypernym，superordinate）与 下位词（hyponym）。

* 上位词是下位词的抽象，下位词是上位词的具体；
* 上位词是下位词的超类，下位词是上位词的子类；
* 上位词表示的类在外延上包含了下位词表示的类。

如果A中的所有对象都是B的对象，则称A蕴涵（entailment）了B，即含义A是含义B的下位词。

本体（ontology）是指对单一领域或者微世界（microworld）进行分析而获得的不同客体的集合。（Ref：Ch17）

分类体系（taxonomy）是指把本体知识体系中的元素排列成树状分类结构的一种特别方式。

分类体系是上下位关系的子类型。

### 语义场（semantic field）

部分——整体（part-whole）关系，称为整体部分关系（meronymy），包括：整体词（holonym）和部分词（meronym）。

同义关系、反义关系和上下位关系都是两个含义之间的二元关系。

语义场是一个针对某个特定领域所有词间的关系集合的更加综合、更加整体的模型。使用的工具有：框架、模型、脚本等。

FrameNet（框架网）提供了一个健壮的框架知识的计算资源。在FrameNet表示中，框架中的每个词都针对不同的框架定义，并且与框架中的其他词共享意义的各个方面。

## WordNet：词汇关系信息库

同义集（synset，synonym set）：一个义项的一组近乎同义词；

同义集是WordNet的重要的基础性成分。

WordNet将概念表示为可以用来诠释概念的词义列表。

WordNet中的根结点被称为独立起始概念（unique beginner）。

## 事件参与者

事件论元的两种语义约束：语义角色（semantic roles） 和 选择限制（selectional restrictions）。

### 题旨（主题的）角色（Thematic Roles）

题旨（主题）角色：特定的语义角色模型。

深层角色特定于不同的事件。

题旨（主题）角色试图捕获不同词之间的语义共性。

参与者的题旨（主题）角色就是主题（theme）。

### 因素交替

题旨（主题）角色帮助泛化论元的不同表层实现。

动词支配的题旨角色论元组被称为题旨格（thematic grid）或者θ格或者格框架（case frame）。

多论元结构的实现被称为动词交替（verb alternations）或者因素交替（diathesis alternations）。

格交替（dative alternations）与动词的特定语义类同时出现。

### 题旨（主题）角色存在的问题

题旨（主题）角色存在的问题是难以正式地定义语义角色，问题的原因：

* 很难提出一组标准的角色
* 创建具体角色的正式定义

解决问题的办法是广义语义角色（generalized semantic roles）的定义，广义语义角色是对特定题旨（主题）角色的抽象。

两种基于语义角色的词汇资源为语义角色标注算法提供训练数据：

* 命题树库（The Proposition Bank，PropBank）同时使用了原型角色（proto-roles）和动词特定的（verb-specific）语义角色。
* 框架网络（FrameNet）使用框架特定的（frame-specific）语义角色。

### 命题树库（The Proposition Bank，PropBank）

PropBank：标注了语义角色的句子的资源库。标记的是宾州树库（Penn Treebank）中的句子。

PropBank中语义角色特定于动词；FrameNet中的语义角色特定于框架。

### 框架网络（FrameNet）

框架是一个类似于脚本的结构，实例化了一组特定于框架的语义角色，这些语义角色称之为框架元素（frame elements）。

每个词引起一个框架，并且描述这个框架及其元素的一些方面。

框架中的语义角色（即框架元素）包括

* 核心角色（core roles）
* 非核心角色（non-core roles）

FrameNet还编码了框架和框架元素间的关系。

* 继承关系：框架间可以彼此继承
* 泛化关系：不同框架的元素间的泛化关系也可以通过继承获得
* 因果关系

### 选择限制（selectional restrictions）

语义角色通过论元与谓词之间的关系来表示论元的语义。

选择限制是一种语义类型限制，表示一个动词对允许填充到它的论元角色的概念类别的限制。

使用事件表示来捕获选择限制的语义。

表示语义角色的选择限制的方法是使用WordNet的同义集而非逻辑概念。每个谓词指定WordNet的一个同义集作为每个论元的选择限制。如果填充语义角色的词是同义集的上位词，那么这个意义表示就是良构的。

## 基元分解

基元分解（primitive decomposition）或者成分分析（componential analysis）的模型应用于词义定义。

基元分解可以解释状态和行为之间或者使动与非使动谓词之间的相似性，但是需要依赖于拥有的大量的谓词。由于难以提出表示所有可能意义的基元集合，所以语义基元在现有的计算机语言学工作中应用不多。

语义特征（semantic features）：表示某种基元意义的符号。

概念依存（Conceptual Dependency，CD）是动词性谓词分解方法。

## 隐喻（metaphor）

隐喻：与借喻（metonymy）类似，即用意义来自完全不同领域的词或短语来提及或探讨另一个领域及其概念。

## 小结

* 词汇语义学（lexical semantics）研究词的意义以及词之间系统化的意义关联的关系。
* 词义（word sense）是词的意义的体现，定义以及关系通常在词义的层面定义，而不在词形层面定义。
* 同形关系（homonymy）是指两个含义共享一个词形但是意义之间没有关联。
* 多义关系（polysemy）是指两个含义共享一个词形并且意义之间有关联。
* 同义关系（synonymy）是指具有相同意义的不同词间的关系。
* 上下位关系（hyponymy）是指具有类别包含（class-inclusion）关系的词间的关系。
* 语义场（semantic field）被用于捕捉某个单独领域的某个词位集之间的语义关系。
* WordNet：是一个大规模的英语词汇信息库。
* 语义角色（semantic roles）从特定深层语义角色出发，通过归纳各类动词之间的相似角色抽象得出。
* 题旨角色（thematic roles）是基于一个有限角色列表的语义角色模型。
  + PropBank：实现了动词特定的语义角色，以及原型施事（proto-agent）∕原型受事（proto-patient）
  + FrameNet：实现了框架特定的角色列表
* 语义选择限制（selectional restriction）：容许词（特别是谓词）对论元词设置某些语义限制。
* 基元分解（primitive decomposition）是词意义表示的一种方法，是基于词元词汇的有限集合。

# 计算词汇语义学

语境的相似性可以用于计算语义的相似性。

计算词汇语义学（computational lexical semantics）：词义计算。

词义排歧（word sense disambiguation，WSD）：检查语境中的词例，并决定每个单词在该语境下的义项。

词语相似性（word similarity）的计算

* 依靠词语之间的关系（上位词、下位词、部分词）
* 依靠语料相似性
* 依靠类似WordNet的结构化资源

语义角色标注（semantic role labeling），也叫格角色指派（case role assignment）或者题旨角色指派（thematic roles assignment）：这些算法是从句法分析中得到的特征去指派语义角色。

## 词义排歧（WSD）：综述

两种不同的WSD任务：

* 词汇采样（lexical sample）：一小组预先定好的词语被选择出来，同时每一个词语在特定词典中的目标语义集合也被选择出来。因为词的集合和词义的集合都很小，可以使用监督机器学习算法来处理。先手工标注选出的词语，然后训练分类系统，接着就可以标注没有手工处理的词语。
* 全词排歧（all-word disambiguation）：系统输入为整个文本，以及对每个单词都标注了对应词义目录的词典，系统对文本中每一个词都需要排歧。问题是标记集很大，因为数据稀疏问题无法构建大量可用的训练数据。

## 有监督词义排歧

使用监督学习方法解决词义排歧的问题：

* 从文本中抽取对于预测特定词义有帮助的特征
* 然后利用这些特征训练一个分类器用来给词语指定一个正确的词义
* 训练的结果是使用分类器给文本中未标注词语指定词义标签。

使用的语料库：

* 对于词汇采样任务，使用面向单个词语的标注语料库。
  + 语料包含目标词语的上下文句子及该目标词语的正确语义标注。
* 对于全词排歧任务，使用语义一致性（semantic concordance）语料
  + 语料中每个句子的开放性词语都标注有来自特定字典或者同义词词典的正确语义

### 监督学习的特征抽取

监督训练需要抽取对词义具有预测性质的特征。

特征向量（feature vector）由数值构成，编码了语言学信息，是机器学习算法的输入。

从邻近的上下文中可以抽取两类特征：

* 搭配特征（collocational feature）是指与目标词语有特定位置关系的词语或短语，包含了目标词语左右特定位置的信息。
  + 从上下文词语中抽取的典型特征包括：单词本身、单词的原形以及该单词的词性。
  + 特征能够有效地包含局部词汇和语法信息，而这些信息能够准确地区分给定的词义。
* 词袋特征（bag-of-words feature）包含了邻近词语的词袋的信息。
  + 词袋是词语的无序集合，忽略了词语的位置信息。
  + 最简单的词袋方法：将目标词语的上下文表示为特征向量。
  + 词汇集从训练集中预先选择词语的有用子集。
  + 目标词语附近的上下文区域是以目标词语为中心的对称的固定大小的窗口。
  + 词袋特征能够有效地捕捉目标词语所在的上下文的一般主题信息，很容易地确定属于特定领域词语的意思。

词义排歧方法同时使用了搭配特征和词袋特征。

### 朴素贝叶斯分类器 和 决策表分类器

朴素贝叶斯分类器是基于特征独立的假设，即给定词义特征之间是条件独立的。

使用朴素贝叶斯分类器进行词义排歧时通常利用拉普拉斯平滑方法将概率平滑。

决策表分类器（decision list classifiers）：产生了一系列选择条件。

注：朴素地（naively）即假设各特征之间相互独立

## WSD评价：方法、基准线、上限

评价WSD这样的组件技术是很复杂的，目标是对WSD在端到端应用中的效果评价。

外在评价（extrinsic evaluation）：也叫基于任务（task-based）的评价，或者叫端到端（end-to-end）的评价，或者叫体内（in vivo）评价，是对嵌入到端到端应用的组件NLP任务的评价。

由于外在评价非常困难，并且对应用来说非常耗时，不易推广，因此WSD系统通过采用内在评价。

内在（intrinsic）评价，也叫体外（in vitro）评价，是将WSD组件看成一个独立于任何给定应用的单独系统。系统通过其精确匹配词义准确率（sense accuracy）来评价，即在测试集中系统标注与人工标注一致的词义所占的百分比。

SENSEVAL已经对语义评价进行了标准化。提供了共享任务以及该任务的训练和测试语料，并建立了多种语言下的词汇采样和全词排歧任务的语义清单（sense inventories）。

基准线（baseline）标准、上限（ceiling）标准。

## WSD：字典方法和同义词库方法

使用字典或同义词库的无监督方法。

### Lesk算法

Lesk算法：是基于字典的词义排歧算法。

Lesk算法泛指一系列算法。算法通过计算词义的字典注释或者定义和目标词语邻近词语的交集，然后把交集最大的词义赋给目标词语。

Corpus Lesk算法：不仅仅计算重叠词语的个数，还为每个重叠词语赋予一个权重。权重是逆文档频率（inverse document frequency，IDF）（Ref：Ch23）。

Lesk算法与监督方法的结合可以通过添加类似Lesk词袋特征的方式实现。

### 选择限制和选择优先度

选择限制（Ref：Ch19）可以作为词义排歧的知识资源。

谓词通过排除那些违背其某一选择限制的词义来判断歧义词语的正确意思。

由于硬性限制导致合法的句子与选择限制的冲突，因此选择限制常常作为参考条件而不是必要条件。

在词义排歧方面，选择优先性的Resnik模型和其他无监督方法一样好，但是不如Lesk或有监督方法。

Resnik将选择优先度（selectional preference strength）定义为一个谓词提供了关于其变量语义类别的大体信息量。

选择优先度可以定义为两个分布之间的信息差异：

* 期望语义类别P(c)的分布（直接宾语落入到类别c的可能性）
* 给定特定怕期望语义类别P(c|v)的分布（动词v的直接宾主落入到类别c的可能性）。

这两个分布的差异越大，动词提供关于其宾语的信息就越多。两个分布之间的差异可以用相对熵（relative entropy）或者Kullback-Leiber散度距离来衡量。

特定类和动词的选择关联性（selectional association）作为该类别对动词一般选择优先度的相对贡献，是一个概率度量，用来度量谓词和支配谓词变元的类别的关联程度。

## 半监督WSD（最小化监督的WSD）：自举法

WSD的有监督方法和基于字典的方法都需要大量的手工构建的资源

* 有监督方法需要监督训练集
* 基于字典的方法需要大规模的词典。

自举（bootstrapping）算法，也被称为半监督学习（semi-supervised learning）或最低限度的监督学习（minimally supervised learning），只需要非常小的人工标注训练集。

Yarowsky算法：

* 算法的目标是为特定目标词语建立一个分类器（在词汇采样任务中）。
* 算法的关键是由小的种子集合构造较大训练集的能力。
* 算法需要一个准确的初始种子集合以及一个好的置信度衡量，从而能够选出好的新的例子添加到训练集中。
  + 算法产生初始种子的方法
    - 手工标注一部分实例
    - 启发式地选择正确的种子。

启发式地选择正确的种子的方法：

* 一个搭配一个词义（one sense per collocation）假设：和目标词义有很强联系的特定词语或短语不可能与其他词义共现。
* 一段话语一个词义（one sense per discourse）假设：一个特定词语在一段正文或一篇文章中多次出现，那么通常具有相同的意思。这个假设的有效性依赖于语义的粒度，大部分情况下语义粒度越粗越有效。

## 词语相似度：语义字典方法

词语计算主要使用：

* 同义关系：词语之间的二元关系。
* 相似度

词语相似度（word similarity）或语义距离（semantic distance）度量来代替同义关系。

两个词拥有的相同意思的特征越多或者两个词是近义词，则两个词的相似度就越高或者两个词的语义距离越近。

度量词语相似度的两种算法：

* 基于语义字典（thesaurus-based）算法：使用类似WordNet或MeSH的在线语义字典来度量两个义项之间的距离。
* 通常使用上位关系（继承关系）∕下位关系（包含关系）的层次结构。
  + WordNet算法只能计算名词和名词之间的或动词或动词之间的相似度，而不能计算名词与动词、形容词或其他词性之间的相似度。
* 基于分布（distributional）算法（Ref：Sec20.7）

词语相似度（word similarity）和词语相关度（word relatedness）的差别：

* 词语相似度是指两个词是近义词或在上下文中可以近似替代。
* 词语相关度是指一大类词语之间的潜在关系。
* 词语相似度是词语相关度的子情况。因此本节中算法统称为相似度度量。

在语义字典层次结构图中两个词语或者两个词语的义项之间的路径越短就越相似。

基本的路径长度算法的隐含假设是网络中每个链接代表的距离相同。

信息量词语相似度（information-content word-similarity）算法仍然依赖于语义字典的结构，但是添加了从语料库中提取出来的概率信息，属于细粒度的衡量。

遵循基本的信息理论，定义一个概念的信息量（IC）。

两个概念的最低公共包含结点（lowest common subsume，LCS）。

* Resnik提出利用两个结点的最低公共包含结点的信息量去估计它们共同的信息量

基于字典的方法（注释是字典的属性而不是语义字典的属性）：如果字典中两个概念或者两个义项的注释包含相同的词语，则它们就相似。

基于语义字典的相似度评价（Evaluating Thesaurus-Based Similarity）

* 内在评价方法是计算算法得出的词语相似度分数和人工标注的词语相似度排序的相关序数；
* 外在评价方法是把相似度度量方法嵌入到某些终端应用中。

## 词语相似度：分布方法

分布（distributional）方法直接为NLP任务提供词语相关性度量

* 分布方法可以用来自动生成语义字典（automatic thesaurus generation）
* 分布方法可以自动地给在线语义字典（如：WordNet）添加新的同义关系和其他关系（如：下位关系、部件关系等）。

分布方法基于的理念：一个词语的意思与它周围词语的分布相关（由词之伴可知其意！）

分布相似度度量方法需要确定三个参数：

* 共现的词语；（Ref：Sec20.7.1）
* 词语赋予权重；（Ref：Sec20.7.2）
* 向量距离度量方法。（Ref：Sec20.7.3）

### 定义词语的共现向量

使用和目标词语具有某种语法关系（grammatical relation）或依存关系（dependency relation）的词语。

实体的意义以及实体间语法关系的意义，和这些实体相对于其他实体的结合限制相关。

### 度量与上下文的联系

目标词语和给定特征之间的权重（weights）或关联度（association）。

点间互信息（Pointwise Mutual Information，PMI）用来度量两个事件的共现频数与假设二者互相独立时它们共现出现的期望频数的比值。

### 定义两个向量之间的相似度（向量计算）

两个向量之间的距离的度量方法：

* 曼哈顿距离（Manhattan distance），也被称为Levenshtein距离，或者L1范数。
* 欧几里德距离（Euclidean distance），也被称为L2范数。

相似度计算：点乘（dot product）或者内积（inner product）。

向量长度需要归一化。

Jaccard度量方法，也称为Tanimoto度量方法，或者最小∕最大度量方法。

分布相似度度量方法：KL散度距离或相对熵。

Jenson-Shannon散度距离。

### 评价分布式词语相似度

分布式算法使用内在方式评价：

* 与一个标准的语义字典进行比较
* 使用t检验对关联度进行加权
* 使用Dice或Jaccard算法去度量向量相似度

## 词语关系：下位关系及其他关系

词语语义关系：相似关系、下位关系、上位关系、反义关系、部件关系。

自动学习语义关系利用词汇——句法模式（lexico-syntactic pattern）。

用来推断下位关系的五种模式：图20.14（P547）

在关系模式挖掘中发现新模式的方法有：

* 自举法：在大规模语料中具备某种关系的词语能够同时出现在这种关系的多种不同模式中。
  + 因此自举法仅仅需要从一小部分准确的模式去获取具备给定关系的词语集合
  + 这些词语可以用来在大规模语料中查询以某种依赖关系包含这些词语的句子
  + 新的模式可以从这些新的句子中抽取出来
  + 这个过程可以一直重复直到模式集合足够大。
  + 使用自举法时需要避免语义漂移。（Ref：Ch22）
* 使用大规模的词汇资源（如WordNet）作为训练信息的来源，每一对上位∕下位关系提供关系中词语的某些信息，然后训练分类器用来寻找具备这种关系的词语。

使用WordNet解决上位关系问题的方法：

* 选择未知词语在一个完整层次结构中的插入位置。
* 类似命名实体识别的标注任务。
  + 使用WordNet中的广义分类标签，抽取特征，使感知器分类器。

寻找部分关系（meronyms）很困难，因为表示该关系的词汇——句法模式极具歧义。

学习词语之间的关系是字典归纳（thesaurus induction）任务的重要组成部分。

在字典归纳中，词语间的相似度估计、上下位关系和其他关系被整合起来构造成一个完全的知识本体或者字典。

两阶段字典归纳算法：

* 自底向上的聚类（clustering）算法，将语义上类似的词语聚集到不带标签的词语层次结构中；
* 给定不带标签的层次结构，使用基于模式的下位关系分类器去为每一类中的词语指定上位关系标签。

## 语义角色标注

语义角色标注（semantic role labeling），也称为主题（题旨）角色标注（thematic role labeling）或者格角色赋值（case role assignment）或者浅层语义分析（shallow semantic parsing）。

语义角色标注：将词语的意义与句子的语义进行链接。目的是自动发现句子中谓语的语义角色。即找出句子中给定谓语的语义变元，为每个变元选择合适的角色。能够改善所有语言理解应用的质量。

主要的语义角色标注方法：基于有监督机器学习，需要大量的训练和测试资源，主要利用FrameNet和PropBank资源。

语义角色标注都从句法分析开始，通过遍历句法分析的结果来寻找所有的谓语承接语。遍历句法分析树来确定成分与谓语角色的对应，从而将句法成分描述成谓语对应的特征集合，然后利用训练得到的分类器来基于这些特征做出判断。常用特征有：

* 管辖谓语：PropBank，谓语总是动词；FrameNet还包括名词和形容词
* 句法成分的短语类型：句法分析树中这一成分的结点名字
* 句法成分的中心词：由标准的中心词规则计算获得，特定的中心词（如：代词）可以对受中心影响的语义角色施加很强的限制
* 句法成分的中心词词性
* 在句法分析树中从成分到谓词之间的路径：作为成分和谓词之间多种语法功能关系的简要表示，路径具有非常大的作用
* 成分所在从句的时态（主动时态和被动时态）：两种时态的句子中语义角色与表面形式间接链接有很大的不同
* 句法成分和谓词之间的二元线性位置，其值或是前面或是后面
* 谓语的次范畴化：使用谓语的直接父结点的短语结构展开规则来抽取信息
  + 动词的次范畴化：是指出现在动词词组中的期望变元的集合（Ref：Ch12）

特征按照上面提到的顺序进行排列。

多阶段分类器：

* 剪枝：基于简单规则排除一些角色的成分以便加快运行速度
* 识别：对每个结点进行二元分类
* 分类：对前一阶段标上ARG标签的成分使用1-of-N分类器

语义角色标注需要注意成分重叠问题，因为成分的语义角色是相互依赖的，因此可以采用N-best重打分技术。先利用分类器为每个成分指定多个标签，每个标签都有一定的概率，然后再使用全局最优的算法从中挑选最好的标签序列。

利用组块技术可以直接对原始（或者带有词性标注）文本进行语义角色标注。

在评价语义角色标注系统时要求把每个变元正确地指派到相应的词语序列或句法分析成分上。因此可以计算准确率、召回率和F值。

一个简单的基于规则的系统可以用来作为基准系统。

## 无监督语义排歧

在语言应用中使用较多的是凝聚式聚类（agglomerative clustering）算法。

评价无监督的词义排歧方法，最好做外在或体内测试。

## 小结

* 词义排歧（Word-Sense Disambiguation，WSD）判定特定上下文中词语的正确义项。
  + 监督方法利用单个词语（单词任务）或所有词语（全词任务）出现的句子，这些句子使用WordNet中的义项进行了标注。
  + 监督WSD使用朴素贝叶斯分类器、决策列表分类器以及其他分类器进行训练和预测，分类器是在描述词语上下文的词中的搭配特征和词袋特征上进行训练。
* WSD的重要基准系统是选用最频繁词义，等价于WordNet中取词语的第一个意义。
* Lesk算法通过对比字典的定义和目标词语的上下文，选择重叠词语最多的义项作为目标词语的词义。
* 词语相似度可以通过度量语义字典中的链接距离或者语义字典中的信息含量，使用语料中的分布相似度或者通过使用信息论方法来计算。
* 分布式相似度的关联度度量方法包括：PMI、Lin和t-test。
* 向量相似度的度量方法包括：余弦、Jaccard、Dice和Jiang-Conrath。
* 词汇间的关系（如：下位关系）可以通过词汇——句法模式发现并识别。
* 语义角色标注通常从对句子进行句法分析开始，然后自动地为句法分析树中的每个结点标识语义角色。

# 计算话语学

话语（discourse）：由搭配在一起、具有一定结构并且连贯（coherent）的句子群组成。

话语的类型：

* 独白（monologue）
* 对话（dialogue）
  + 人和人的对话（human-human dialogue）
  + 人机对话（human-computer dialogue）

指代消解（Reference Resolution）：决定代词以及其他名词短语指代的内容。

* 代词回指消解（anaphora resolution）
* 共指消解（coreference resolution）

话语结构：

* 简单话语分割（discourse segmentation）
  + 简单话语分割：就是把一篇文档分割成线性序列的多个段落的篇章。
* 连贯性关系（coherence relation）
  + 连贯关系（coherence relations）：决定话语中句子间的连贯结构。连贯话语中的句子之间必须有语义上的联系。
  + 基于实体的一致性（entity-based coherence）是一种连贯性，说明连贯的话语与涉及在话语中实体之间必须表现出一定的关系。

篇章→关系→实体。

## 话语分割

话语分割（discourse segmentation）：把一篇文档切分成一个线性的子主题序列。

使用内聚机制寻找话语结构。

### 无监督话语分割

线性分割（linear segmentation）：把文本分割成多个多段单元的任务，其中每个单元表示原文中的一个子主题或者段落。

内聚性（cohesioin）：是指用一定的语言学手段将文本单元联系或者连接在一起。

* 词汇内聚性（lexical cohesion）：是指两个语言单元中基于词语间关系表现出来的内聚性。
* 非词汇内聚性
  + 指代关系
    - 使用回指（anaphora），也称指代或者首语重复。
* 内聚链：通过相关词语的一个序列表现出来的内聚性。

内聚性（cohesion）与连贯性（coherence）的区别。

* 内聚性：指的是文本单元联系在一起的方式，把两个单元聚成一个单元。
* 连贯性：用来解释不同文本单元的意义如何结合在一起以表达一个更大粒度的话语意义，是两个单元意义之间的关系。

线性话语分割的无监督算法基于内聚性。

* 基于内聚性的分割算法的都是同一个子主题中的句子或者段落之间具有内聚性，而相邻的子主题之间的段落则没有这种内聚性。
* 度量内聚性时目标：子主题内部的内聚性强，邻近子主题的内聚性弱。

TextTilling算法（基于内聚性的算法）的三个部分：

* 分词（tokenization）
* 词汇分值确定（lexical score determination）
  + 词汇内聚性得分（lexical cohesion score）：边界前伪句子中词语到边界后伪句子中词语的相似度的平均。
* 边界识别（boundary identification）
  + 深度分数（depth score） ：度量边界“相似度山谷”的深度，是计算从山谷两边的山峰到谷底的深度
  + 自顶向下的划分式聚类

### 有监督话语分割

基于多种分类器完成有监督话语分割。

在话语分割的预处理阶段需要进行命名实体识别

在有监督话语分割算法中使用的特征（是无监督分类特征的超集）：

* 内聚性特征
* 词语重叠度
* 词语的余弦距离
* LSA
* 词汇链
* 共指，等等
* 话语标记（discourse markers）或提示词（cue word）：表现话语结构的词语或者短语。
  + 话语标记是领域特定的，可以利用手写规则或者正则表达式去确定特定领域的话语标志

### 话语分割的评价

WindowDiff通过在系统输出的分割上滑动一个探测器，即大小为k的滑动窗口，来对自动标注的边界和人工标注的边界进行比较。

因为准确率、召回率以及F值对分割边界的距离误差不敏感，因此不使用它们来评价分割算法。

## 文本连贯性

连贯关系（coherence relations）：话语的话段之间所有可能的连接

* 结果（Result）：推测A声明的状态或者事件导致了B声明的状态或者事件。
* 说明（Explanation）：推测B声明的状态或者事件导致了A声明的状态或者事件。
* 平行（Parallel）：推测A声明的和推测B声明的都是类似的。
* 细化（Elaboration）：推测A声明和推测B声明的是同一个命题。
* 时机（Occasion）：推测从A声明的状态到B声明的最终状态的状态变化，或者推测从B声明的状态到A声明的最初状态的状态变化。

树中每个结点代表一组局部连贯的从句或者句子，称之为话语片断（discourse segment）。

### 修辞结构理论

修辞结构理论（Rhetorical Structure Theory，RST）：是连贯关系理论中的一种，是一种文本组织模型，应用在文本生成领域。

修辞结构理论包含了23种修辞关系，用于表示话语中不同跨度的文本之间的关系。

大部分修辞关系保持在两个文本跨度之间，一个作为核心（nucleus），一个作为外围（satellite）

* 核心是更接近作者意图的并且能够独立解释的单元
* 外围是离作者意图远些并且通常需要和对应的核心一起解释。

在证据关系（Evidence）中，外围为核心表述的观点或情况提供证据。

修辞结构理论（RST）中的关系：

* 细化（Elaboration）：外围对核心做进一步的补充说明
* 属性（Attribution）：外围给出核心中转述语实例的属性来源
* 对照（Contrast）：多核心关系，两个或多个核心在某些重要的维度上进行对比
* 并列（List）：多核心关系，两个或多个核心不进行比较
* 背景（Background）：外围给出解释核心的上下文

### 自动指派连贯关系

连贯关系的指派（coherence relation assignment）：给定一个句子串，自动确定句子之间的连贯关系。

话语分析（discourse parsing）：抽取能够表示整个话语的树或图。

基于提示短语（cue phrases）的浅层算法。

1. 识别文本中的提示短语。

* 提示短语（cue phrase），或者话语标志（discourse marker），或者提示词（cue word）是能够指示话语结构的词或者短语，特别是能够把话语片段联系在一起。
* 连接语（connectives）是一种提示短语，通常是连词或者副词，提供了两个片段之间存在的连贯关系的线索。

1. 基于提示短语把文本分割成话语片断。

* 话语片断的大小是从句或者类似从句的单元。
* 基于单个提示短语手工编写分割规则。
* 利用句法分析器，可以利用句法短语制定更加复杂的分割规则。

1. 利用提示短语对连续话语片段间的关系进行分类。

* 为话语标志撰写规则，但是注意提示短语的歧义性带来的问题
* 使用自举法对大规模语料进行连贯关系自动标注，从而满足训练分类器需要的大量数据。
* 使用正则表达式抽取包围提示短语的话语片段对，然后移除提示短语。最终的句子对不带有提示短语，就可以用于抽取连贯关系的监督训练集。

## 指代消解

指代消解（reference resolution）：决定哪些实体被哪些语言表述所指代。

提示语（referring expression）：用于实现指代的自然语言表达。

所指对象（referent）：指向的实体。

共指（corefer）：两个指示语用于指向同样的实体。

先行词（antecedent）：以一种方式准许使用另一个提示语。

复指（anaphora）或者回指：提及一个先前已经被引入话语的实体。使用的指示语是复指语（anaphoric）。

指向实体的方式依赖于描述实体的话语上下文（discourse context），还依赖于话语的情境上下文（situational context）。

话语模型（discourse model）：具有特定地位的信念子集形成了听话人对正在进行的话语的心理模型，包括本话语中所指向实体的表示以及它们参与的关系。

指代消解系统的两个部分：

* 构造话语模型的方法，该模型能够随着所表示的话语的动态变化而演化；
* 各种指示语暗含的信息到听话人的信念集之间的映射方法，包括该话语模型。

话语模型的两个基本操作：

* 当话语中首次提及所指对象时，表示对象被唤起（evoke）而进入模型；
* 当话语中再次提及所指对象时，从模型中访问（access）它的表示。

指代消解的两种任务：

* 代词回指消解（pronominal anaphora resolution）：找出一个代词的先行词
  + 可以人称代词回指消解看成共指消解的子任务
* 共指消解（coreference resolution）：找出文中所有的指向同一实体的指示词
  + 即找出所有具有共指（corefer）关系的表述
  + 一系列的指示语称为共指链（coreference chain）

## 指代现象

### 指示语的五种类型

1. 不定名词短语（indefinite noun phrase）：不定所指将一个新的实体引入了话语环境
2. 有定名词短语（definite noun phrase）：有定所指表示指示对象为听话人可以确认的实体
   1. 实体在文本中已经被提起，并且也被表示于话语模型中
   2. 实体包含在听话人关于世界的信念集中
   3. 实体本身的描述就包含了唯一性。
3. 代词（Pronoun）：使用代词的所指相比有定名词短语受到更强的约束，要求在话语模型中所指对象具有高度的活力或者显著性。
   1. 代词指示的实体被引入的位置相比有定名词短语的要近。
   2. 代词也可以参与提前指代（cataphora），即在代词所指对象出现之前就提及代词。
4. 指示代词（Demonstrative）：既可以单独出现，也可以作为限定词。
   1. “this”是近端指示词（proximal demonstrative）：表示文字上或者隐喻上比较接近；
   2. “that”是远端指示词（distal demonstrative）：表示文字上或者隐喻上相隔较远（例如：时间上相隔较远）
5. 名字（Names）：包括人名、地名、机构名。在话语中可以用来指代新的或者旧的实体。

### 信息状态

相同的指示语（如许多不定名词短语）能够用来表示新的指示对象，其他的指示语（如许多确定名词短语）可以用来指向旧的所指对象。

信息状态（information status）或信息结构（information structure）：对不同所指形式提供新的或者旧的信息的方式。

话语中不同种类的所指形式和所指对象的信息度或者显著性之间的关系：

* 约定层级（givenness hierarchy）：是一种表示6种信息状态的尺度，每一种信息状态由不同的指示语指示。
* 相关接受度尺度（accessibility scale）：
  + 显著的实体需要较少的语料内容来指代就可以帮助听者恢复所指对象
  + 不显著的实体需要较长的和较显著的指示语来帮助听者恢复所指对象。
* 听者状态和话语状态来分析信息状态。
  + 听者状态表明所指对象对听者来说是已经知道的或者是新的；
  + 话语状态表明所指对象在话语的前面部分是否已经被提起。

指示语形式和信息状态之间的关系：

* 推理对象（inferrables）：也叫桥接推理（bridging inferences），或者中间物（mediated）。
  + “指示语”不指向文中已经被明显唤起的实体，而是指向与唤起实体具有推理性关系的实体。
* 类属指代（generics）：
  + “指示语”不指向文中已经被明显唤起的实体，而是指向与唤起实体具有同类关系的实体。
* 无所指形式（non-referential forms）：某些无所指形式与指示语在表面上很相似。

## 代词回指消解所使用的特征

给定代词及代词前面的上下文，从上下文中找出代词的先行词。

### 用来过滤潜在指代对象的特征

4种相关的固定不变的构词特征：

* 数的一致（number agreement）：指示语和所指对象在数上保持一致。
* 人称一致（perosn agreement）：代词的先行词与代词在数上保持一致。
* 性的一致（gender agreement）：指示语和所指对象在性别上保持一致。
* 约束理论限制（binding theory constraints）：指示语和先行名词短语出现在相同句子中，所指关系受到指示语和先行名词短语之间句法关系的约束。

### 代词解释中的优先关系（显著性）

用来预测代词指代对象的特征：

* 新近性（recency）：新近的话段所引入的实体比先前较远的话段所引入的实体具有较高的显著性。
* 语法角色（grammatical role）：通过实体表示的语法位置来排序的实体显著性层级。
  + 主语位置的实体的显著性最高
  + 宾主位置的实体的显著性其次
  + 后续位置的实体的显著性最低
* 重复提及（repeated mention）：已经作为焦点的实体，在后面的话语中更可能成为焦点，它们的所指也更可能被代词化。
* 平行（parallelisom）：平行效果会带来明显的优先关系。
* 动词语义（verb semantics）：动词会对某个位置的论元产生强调，而影响指代优先级。
* 选择限制：语义知识可以影响指代优先级。

## 代词回指消解的三种算法

* 代词回指基准系统：Hobbs算法
  + 算法最简单，包括：句法分析器、形态性别检查器和数字检查器，常常用作评价的基准系统。
  + 输入：代词所在句子及代词之前的几个句子的句法表示；
* 代词回指消解的中心算法
  + 中心理论（centering theory）显式采用话语模型表示。是实体一致的模型。在话语中的任何给定点都有一个单独的实体被作为“中心”，该实体与被唤起的其他实体有所不同。需要句法分析器和形态性别检查器。
* 代词回指消解的对数线性模型
  + 对数线性分类器：基于手工标注的训练语料，语料由标有先行词的代词组成。

代词回指消解的特征

* 严格的数匹配（strict number）：真或假。
* 相容的数匹配（compatible number）：真或假。
* 严格的性别匹配（strict gender）：真或假。
* 相容的性别匹配（compatible gender）：真或假
* 句子距离（sentence distance）：代词和潜在的先行词之间的句子数目。
* Hobbs距离（Hobbs distance）：从代词开始回溯找到潜在先行词之前，Hobbs算法必须跳过的名词组的数目。
* 语法角色（grammatical role）：潜在先行词的角色——句法中的主语、直接宾语或者PP中的一个嵌入成分。
* 语言学形式（linguistic form）：潜在先行词的形式——专有名词、确定描述、不定描述或者代词。

## 共指消解

共指消解算法：通过二元分类器，以指代和潜在先行词作为输入，判定是否共指。

共指消解常用特征：

* 回指编辑距离（anaphor edit distance）：从潜在先行词到指代的字符最小编辑距离
* 先行词编辑距离（antecedent edit distance）：从指代到先行词的最小编辑距离
* 别名（alias）：真或假。命名实体标签的多重特征。
* 同位语（appositive）：真或假。指代语和先行词是否处于语法中的同位关系。
* 语言学形式（linguistic form）：潜在回指的形式——专有名称、确定描述、不定描述或者代词。

## 共指消解的评价

模型理论的共指评价（model-theoretic coreference evaluations）

共指链（reference chain）或真实链（true chain）：是实体出现的正确的或者真实的共指链。

假设链（hypothesis chain）：共指消解算法为实体指派的链或者类。

B-CUBED算法，用来评价MUC-6方法的扩展，依赖于手工标注的指代短语间的共指语料。计算的是相对于共指链实体在假设链中的准确率和召回率。

## 基于推理的连贯判定

如果分析器能够为每个从句指派合理的语义，那么连贯判定方法就可以依靠和每个连贯关系相关的语义限制来执行。

推理类型：

* 演绎（deduction）就是依赖这些限制进行推理的方法，是向前推出隐含关系的推理方法，是一种可靠推理形式（即前提为真，结论必为真）。
* 溯因推理或者诱导（abduction）推理：是后向推理方法，即从结果中寻找可能的原因，是不可靠的推理形式，是可以废止的（defeasible），但是提供了更大范围的推理能力。

溯因推理所用的证据的质量的判定方法：

* 概率模型
* 启发式策略：优先选择假设数目最少的解释或者最具体的解释
* 基于代价（cost-based）的策略：结合了概率特征和词性式方法。世界知识和领域知识被用于确定话段间最合理的连贯关系。

注：基于代价的策略可以被应用于语言理解中的许多问题。

## 所指的心理学研究

阅读时间实验：研究代词解释中句子新近性的影响。

不同任务环境下对比代词与语法角色在指代下的优先级：

* 问题回答任务：解释代词的方式
* 所指对象指定任务：代词的所指对象

基于句子完成任务研究动词的“隐含的因果关系”对代词判定的影响。

基于中心理论原则研究重复名称惩罚仅仅出现在主语位置的所指对象。

## 小结

本章的重点是自然语言处理在话语层（句子之间）的运算

* 话语具有层级结构，与句子一样。
  + 通过假设话语具有简单的线性结构，可以基于词汇内聚性、话语标志或者提示短语对文本进行主题条件分割。
* 话语是连贯的。
  + 保持话语连贯的因素是句子间的连贯关系以及基于实体的连贯关系。
* 修辞关系是连贯关系理论中的一种。
  + 用于检测修辞关系的算法可以使用表层提示（例如：提示短语、句法信息）
* 话语模型：为话语状态建立的一种可以演变的话语表示方式
  + 已经提及的实体
  + 实体之间关系的表示。
* 指向实体的方法将世界知识集和话语模型一起加工后传递给听话人。
* 代词所指被用于话语模型中具有足够显著度的所指对象。各种话语因素都会影响显著性。
* 代词回指消解的三种算法：提供了不同使用和结合限制的方式。
  + Hobbs算法
  + 中心算法
  + 对数线性模型
* 完整的NP共指任务必须处理名称和确定的NP。
  + 任务中的字符串编辑距离是重要的特征。
* 建立连贯关系的高级算法使用一个或者多个连贯关系构成的限制
  + 基于不完全的逻辑溯因规则的推理方法能够推出说话者的潜在信息

自然语言处理的应用

# 信息抽取

本章重点是从文本中提取限定类别的语义内容的方法。

信息提取（Information Extraction，IE）过程，也被称为文本分析（text analytics），将文本中包含的非结构化数据转换为结构化数据的。

信息提取过程如下：

* 发现并分类文本中所提及的专有名词，叫做命名实体识别（Named Entity Recognition，NER）
  + 将文本中的专有名词的实例划分到具体类别下，称为命名实体提及（named entity mentions）。
* 在确定文本所有命名实体提及之后，对这些命名实体提及进行链接或者将其聚类，使每一个提及集合指向其真实所指的实体，叫做指代消解（reference resolution）
* 关系识别和分类（relation detection and classification）是找出给定文本中实体间的语义关系，并将这些关系进行分类。
* 事件识别与分类（event detection and classification）：找出实体参与的事件，并对其进行分类。
* 时间表达式识别（temporal expression recognition）和时序分析（temporal analysis）：找出文本中事件发生的时间以及这些事件之间的时序关系
  + 时序分析就是将时间表达式映射到一个确定日期或者一天中的某个时间上，然后使用这些信息建立事件的时序关系
    - 确定时间表达式相对于锚定日期或者锚定时间的关系
    - 找出时间表达式与文中事件的关联关系
    - 将事件按照完整一致的时间线排序
* 模板填充（template filling）是找出描述一些固定的场景发生的文档，并且使用适当的语言单元完成填充模板的任务。这类槽填充可以从文本中直接抽取的文本段组成，也可以通过一些额外处理得到的文本元素推断出来的概念组成。

## 命名实体识别（NER）

命名实体（named entity）：是任何一个可以被专有名称指代的事物。

命名实体识别（named entity recognition）过程：

* 找出构成专有名称的文本片段
* 对于这些文本片段所指向的实体类别进行分类

命名实体识别有时也可以包括：时间表达式和数值表达式的识别

### NER中的歧义

* 同一个名称可以指向同一个类型的不同实体
* 同一个名称可以指代完全不同类型的实体

### 基于序列标注的NER

基于统计序列标注的命名实体识别方法，需要训练一个分类器标注文本中的词，该分类器使用指示标签说明特定类别实体是否出现，指示标签与句法组块分析中的IOB编码方式相同。

* 标记I：表示词在一个块中
* 标记B：表示块的开始
* 标记O：表示块之外的词

NER中的重要特征：

* 形态特征（shape feature）：包括常见的大写体、小写体和首字母大写，以及用于捕捉表达式更加复杂模式的特征
* 存在于命名实体列表中（presence in a named entity list）：是具备调试预测能力的特征。
* 上下文窗口中基于预测词和N-grams的特征

### NER的评价

准确率、召回率和F测度（P382，Sec 13.5.3）也可以用来评价NER系统。

### 实用的NER框架

NER系统是一个基于词典、规则和有监督的机器学习的实用组合，运行过程：

* 使用高正确率的规则标注无歧义的实体提及
* 基于字符串概率化匹配相似度，搜索能匹配之前检测出的名称的子字符串
* 查看特殊领域的名称列表，确定该领域可能的命名实体提及
* 应用概率序列标注技术，使用前述步骤的标签作为额外的特征

## 关系识别和分类

识别文本中已经检测出来的实体之间所存在的关系。

关系是由基于某个领域元素的有序变量集合组成。（Ref：Ch17模型理论）

### 用于关系分析的有监督学习方法

* 基于一个人工选择出来的关系集合，文本所有存在的关系都被标注出来，建立特征集合
  + 命名实体本身的特征
    - 两个候选参数的命名实体类别
    - 两个实体类别的拼接
    - 关系角色的头词
    - 每个关系角色的词袋子表示
  + 文本中被分类的词的特征
    - 实体间文本的词袋和二元词袋
    - 上述特征的词干还原版本
    - 直接位于实体之前和之后的词及其词干
    - 两个参数之间的词间距
    - 两个参数之间的实体数量
  + 利用句法结构标注的实体之间的关系特征
    - 在某成分结构中，一些特定的结构是否出现
    - 基于短语的组块分析路径
    - 组块的中心词词袋
    - 依存树路径
    - 成分树路径
    - 两个参数之间的树距离
* 标注过的文本会用于训练系统
* 系统被用于对新的文本进行类似的标注
  + 检测两个实体之间是否存在关系，属于二元判断问题
  + 对于已经检测出手关系进行分类，标注实体对之间的关系

### 用于关系分析的弱监督学习方法（？）

简单的关系抽取方法是使用正则表达式来匹配感兴趣的关系表达式的文本段。

* 泛化模式：能够捕获包含所需信息的表达式。
* 松驰模式：可以在匹配中忽略了部分文本的结果，可能引入更多不正确的结果
* 高正确率模式：在一个大规模并且离散度比较高的文档集合上，一个模式的扩展集合会获取更多所需信息。需要一些具备领域知识的专家，构造更多具有更高覆盖率的模式
* 使用较涉集合的种子模式进行搜索，再从结果中结合使用自举技术引入新的模式

### 关系分析系统的评价

两种评价方法：

* 关注在给定文本中系统识别并且分类所有关系提及的表现
  + 带标签的评价要求系统能够正确地将关系分类
  + 无标签的评价只衡量系统是否检测出关联的实体
* 只关注从文本中抽取的元组，而不关注关系的提及。
  + 多用于评价无监督方法

## 时间和事件处理

抽取文本内容最为重要的工作就是：找出和分析文本中的事件，并判断这些事件之间的时序关系。

### 时间表达式的识别

时间表达式是指向绝对时间点、相对时间、时段，或者它们的集合的表达式。

* 绝对时间表达式（absolute temporal expression）：是可以直接映射到日历日期、在一天中的时间或者二者都映射到的表达式
* 相对时间表达式（relative temporal expression）：需要通过其他的参考时间点指向特定的时间
* 时段（duration）是指在不同粒度等级上的时间跨度。包括：秒、分钟、小时、天、星期、月、年、世纪等等。

从语句构造上看，时间表达式以表示时间的词汇触发词作为其中心词。

词汇触发词（lexical triggers）可以是名词、专有名词、形容词和副词；而完整的时间表达式使用这些词汇组成的短语来表示，包括：名词短语、形容词短语和副词短语。

识别时间表达式：就是找出所有时间表达式文本段的开始和结束的位置。

识别时间表达式的方法：

* 基于规则的系统，使用部分句法分析或者组块分析技术。
  + 使用串联式的自动机，识别复杂度是逐渐增长的模式。
  + 识别过程：标注词性；识别句法块（句法块中包含了时间触发词）
* 基于统计序列分类的系统，使用标准的逐词IOB标注技术。
  + 三个标签I、O、B被用来标注词是在时间表达式的内部、外部或者开始位置。
* 基于成分分类的系统，使用与标注语义角色相同的技术。
  + 结合了句法组块分析和逐词标注，使用二元分类器完成，分类器使用已经标注过的数据进行训练，标注数据的特征与IOB-style模型的特征相同。
    - 完整的成分分析会通过自动的方法产生
    - 根据是否包含时间表达式对结果对中的结点逐个分类

识别时间表达式的方法存在的问题：

* 如何达到合理的覆盖率
* 如何正确地识别出时间表达式的范围
* 如何处理触发器误判的表达式
  + 误判：是当专有名词中的某个部分属于时间触发词时所引发的问题

### 时间的归一化

时间的归一化（temporal normalization）：指将时间表达式映射到某一特定的时间点或者时间段的任务。

* 时间点：是日历日期、一天中的某一时刻或者二都都对应
* 时间段：是由一段时间组成的，但是当提供了相关信息时，也就包含了时间段的开始点和结束点。

时间归一化的方法：基于规则的方法，结合语义分析程序和匹配特定时间表达式的模式。一个成分的意义是由其部件的意义计算而来的，用于计算意义的方法又特定于被构造的成分。（Ref：Ch18）

为了归一化时间表达式需要的四类表达式规则：

* 完全确定的时间表达式（fully qualified date expressions）：包括常规格式的年、月、日。
  + 表达式里面的特定时间单位必须被检测出来，然后放置在相应的ISO模式的正确位置上。
  + 完全确定的时间表达式在实际文本中几乎不存在，大多数都是隐式的表达，通常相对于新闻报道的标题时间给出，称之为新闻报道的时间锚（temporal anchor）。
* 绝对时间表达式
* 相对时间表达式
* 时间段

### 事件检测和分析

事件检测和分类（event detection and classification）是识别出文本中的事件提及，并将这些事件划分到某个类别中。

事件提及：是指任意一个指代某个事件的表达式，或者指代这个事件在某一时间点或者时间段的状态的表达式。

事件检测多使用表层信息来解决问题，表层信息包括：词性信息、特殊词汇的出现、动态时态信息等。

事件检测的常用方法有：

* 基于规则的方法
* 基于统计机器学习的方法

事件检测的主要特征：P611图22.23

### TimeBank（语料库）

TimeBank语料库通过TimeML标注方式显式地标注出了新闻报道文本中的时间表达式和时间提及。

## 模板填充

脚本（scripts）里有子事件、参与者、角色和道具的原形序列。

脚本可以表示为模板（templates），由固定的槽（slots）集合组成，这些槽使用属于特定类别的值进行槽填充（slot-fillers）。模板填充（template filling）就是找出包含特定脚本内容的文档，然后用从文本中抽取的填充物，填充相关模板的槽。

槽填充对应着可以检测到的各种类别的命名实体，因此模板填充基于命名实体识别、时间表达式和共指算法提供的标注来识别候选的槽填充。

槽填充常基于序列标注技术实现，模板填充常基于层叠式有限状态自动机实现。

### 模板填充的统计序列标注方法

模板填充问题等价于统计序列标注问题。训练好的系统将词序列标注为某个特定槽的可能填充物：

* 为每个槽分别训练独立的序列分类器，然后将整个文本发给每个标注器，槽用其对应分类器识别出的文本段进行填充
  + 常用的特征：词、词形、词性、句法组块分析标签以及命名实体标签
  + 冲突：多个文本段贴上相同的槽标签
    - 指向同一个实体但是使用不同的指代表达式的冲突文本段
    - 指向不同的实体的可能候选的冲突文本段
  + 解决冲突的办法是采用置信度最高的文本段；基于马尔可夫方法的选择标注概率最高的文本段
* 训练一个大的分类器（通常是HMM），为每个待识别的槽赋予标签。（Ref：Ch24）

### 有限状态机模板填充系统

信息理解会议（Message Understanding Conference，MUC）是一系列美国政府组织的信息抽取评测会议，其介绍的任务描绘了一个更为复杂的模板填充问题。评测任务要求系统产生层次链接的模板，这样模板需要以大量的转换机为基础。转换机的的每一级语言学处理过程都从文本中抽取一定的信息，然后传递给更高一级的处理过程。

## 生物医学信息的抽取

生物学命名实体识别的常用技术：

* 基本名词词组（base-noun-group）的句法组块分析器
* 数量非常少的启发式规则
* 缩写定义检测算法

基因归一化：识别出文本中所有的生物学实体提及之后，将它们映射到数据库或者本体中的唯一标识。可以看作词义消歧，介于目标词词义消歧和所有词词义消歧。

实体对间关系的分类：

* 生物学关系问题：发现文本中实体之间的二元关系并对其分类
* 生物学角色问题：根据文本的中心事件，识别和分类实体的角色

## 小结

从文本中抽取限定格式的语义内容的技术。

* 命名实体：使用统计序列标注技术识别和分类
* 实体间的关系
  + 使用已经标注的训练数据，通过有监督的学习方法检测和分类
  + 当有少量种子元组或者种子框框苛用时，使用轻量级有监督的自举方法
* 时间表达式：使用时间推理，结合基于规则和统计学习的方法来进行检测和归一化
* 事件：基于规则和统计学习的方法可以检测、分类和按时间排序。
* 模板填充应用程序：能够识别文本中固定模式的情形，并将文本中的元素赋值给由确定的槽集合确定的角色
* 信息抽取技术在处理生物学领域的文本上是极其有效的
* 语义角色标注问题中算法的选择没有特征的选择重要，许多特征都来源于精确的句法分析。

# 问答 和 摘要

问题（Question Answer）任务和摘要（Summarization）任务，都以生成特定的短语、语句或者短小段落为目标，以回答用户用自然语言所表达的信息需求为结果。

信息检索（Information Retrieval，IR）是返回与特定自然语言查询相关方法的任务。包括对各类媒体，包括：文本、图片、音频和视频等数据进行存储、分析和检索。本章重点是基于词的查询的文本文档存储和检索。

这些子任务背后的核心思想是：从文档或者类似于Web这样的文档集中直接抽取（extracting）满足用户信息需求的段落。

文本摘要（text summarization）：是为了生成包含重要或者相关信息的删减版的文本。复杂问题的回答都是通过拼接来自较长文档的摘要片段得到。

## 信息检索

信息检索（Information Retrieval，IR）是对各种媒体存储和检索相关的任务。当今IR系统主要基于组合语义，即文档中的含义仅仅由它所包含的词的集合所决定，而忽略了句法信息，俗称为“词袋子”（bag of words）模型。

在信息检索领域的几个重要名词：

* 文档（document）泛指被系统索引以及提供给检索的文本单元。
* 文档集（collection）表示用于满足用户需要的一组文档。
* 检索词（term）表示文档集中出现的词汇项，也可以是短语
* 查询（query）表示由一组检索词表达的用户信息需求
* 特征型检索（ad hoc retrieval）：特定信息检索任务，假定独立用户向检索系统发出一个查询请求，系统返回一个按可能有用性排序的文档集合。

### 向量空间模型

在信息检索的向量空间模型（vector space model）中，文档和查询都被表示为一个特征向量，其中特征表示在文档集中出现的词语

每个特征的值被称为词语权重（term weight），通常是检索词在文档中出现频率以及其他因素的一个函数。

在基于向量的信息检索中，使用余弦相似度（cosine metric）来度量两篇文档之间的距离。

余弦值也可以看作为归一化的点积（normalized dot product），即两个向量的点积（dot product）除以每个向量的长度。

把文档表示为词语向量的权重就可以把整个文档看成一个（稀疏的）权重矩阵，称为词语——文档矩阵（term-by-document matrix）。

### 词语权重计算

TF-IDF：将词频与倒排文档频率相结合构成的加权方案。偏好于在当前文档出现次数较多，而在整个文档集合中出现较少的词。出现在少数文档中的检索词拥有更高的权重。

* 词频（term frequency，TF）：检索词在一个文档中出现的原始频率
  + 如果一个词在文档里出现的次数越多，那么权重越高
* 倒排文档频率（Inverse Document Frequency，IDF）的检索词权重是一种给高判别性词赋予更高权重的方法。
  + IDF的定义为分数，其中N是集合中文档总数，是词i出现的文档数目。
* 将TF与IDF结合在一起的加权方案：w\_(i,j)=tf\_(i,j) x idf(i)

TF-IDF加权余弦值：测量两个文档之间的距离

### 词语选择和建立

词干化（stemming）是对单词的屈折和形态变化做出的还原处理。采用词干化处理可以将特定的查询词与任何包含这个词形态的各种变体的文档相匹配。

停用词表（stop list）：需要被排除的索引词表。一般是高频词的简单列表，将高频词排除在文档和查询表示之外的原因是：几乎不具有语义权重。

### 信息检索系统的评测

性能评价的两个基本工具：正确率（precision）和召回率（recall）。

* 正确率=返回的相关文档/所有的返回文档，是指返回的文档中，相关文档所占的比率
* 召回率=返回的相关文档/所有的相关文档，是指相关的文档中，返回文档所占的比率

返回的文档分成两个部分：相关文档（与检索目的有关的文档）和无关文档。

性能评价的两个方法：

* 基于绘制正确率/召回率曲线的方法，比较两个系统或者方法的曲线来评价性能
  + 利用插值正确率方法
    - 提供了在一系列查询之上计算平均性能的方法
    - 提供了一种平滑原始数据中的不规则的正确率的合理方法
* 基于平均正确率（Mean Average Precision，MAP）的方法：比较两个系统或者方法的平均正确率来评价性能
  + 优点：提供了一个单独的清晰的指标，容易偏好相关文档排在较前的系统，即返回较少并且高正确率的文档
  + 缺点：忽略了系统的高召回率和对返回结果的全面性考虑为代价

### 同形关系、多义关系和同义关系

同形关系和多义关系可能导致系统返回与用户需求无关的文档，从而造成准确率的降低。

同义关系和上下位关系可能导致系统错误与用户需求相关的文档，从而造成召回率的降低。

### 改进用户查询的方法

改善向量空间模型中信息检索系统性能的方法：

* 相关反馈（relevance feedback）：用户向系统提交一个查询条件，系统返回给用户少量的检索结果文档，然后要求用户指出满足其需求的文档，系统则根据用户指定的相关文档和不相关文档中的检索词分布，重组用户的原始查询条件，重组后的查询条件作为一个新的查询条件提交给系统并给用户返回新的检索结果
  + 对采用相关反馈的系统进行评测，建立使用剩余文档集来计算正确率和召回率。
  + 剩余文档集（residual collection）：是去除原始文档集中任何一轮提交给用户判断的文档后剩余的文档集合。
* 查询扩展（query expansion）：用户的原始查询条件用原始检索词的同义词或者一些与原始检索词相关的检索词进行扩展
  + 以提高召回率，特征正确率来改善系统性能
  + 添加到查询的词都是从类属词典（thesaurus）中找出来的。
    - 类属词典生成方法，根据集合中的文档自动生成类属词典
    - 对集合中的词聚类来构建词典，称为词语聚类（term clustering）。

## 事实性问答

问答（question answering）：使用一段特定信息来回答用户问题的任务。

事实性问答：使用命名实体来回答用户的问题

事实性问答系统是从网络或者其他文档集合中通过查找可能包含答案的较短的文本片段，并对其进行重构来最终呈现用户的任务

问题系统的二个步骤：

* 使用信息检索的方法来检索较少数量的可能文档
* 使用代价较高的技术（句法分析、角色标注等）在相关文档上进行排序

现代事实性问答系统的三个阶段：

* 问题处理
* 段落检索和排名
* 答案处理

### 问题处理

问题处理阶段从问题中抽取出的两项内容：

* 符合IR系统输入要求的关键词查询，即查询构建
  + 创建一个关键词列表，构造一个IR查询
* 构成该问题合理答案的实体类别，即问题分类（question classification）或者答案类型识别（answer type recognition）。
  + 根据问题所期望的答案类型（answer type）对其分类
  + 可以使用半自动方式动态构建一个答案类型分类体系（answer type taxonomy）或者问题本体（question ontology）。
  + 问题分类器
    - 手工编写的规则来实现
    - 通过机器学习来实现
    - 手工编写和机器学习融合来实现
  + 问题分类器的特征
    - 问题中的特定词：能为答案类型提供额外的信息，也叫中心词（head-word）或者核心词或者答案类型词（answer type word）
    - 问题中词的语义信息
    - 问题中词的同义词、上位词、下位词

### 段落检索

文档检索阶段将返回一个文档集，这个文档集将提交给信息检索系统，或者是私有索引文档集上的IR引擎，或者是私有索引文档集上的网络搜索引擎。这个文档集一般是按照相关性排序，但是排名最高的文档可能并非需要的答案，因此可以从文档集中提取一系列可能的答案段落，然后进行段落检索从而过滤掉不需要的文档，并对剩下的文档进行排序。

段落（passage）：一般是包含了节、段落和句子的文档，具体由系统决定。

段落检索（passage retrieval）：过滤返回文档中不包含潜在答案的段落，然后对剩下的段落根据包含答案的可能性进行排序

段落检索使用的特征集：

* 段落中该类型的命名实体的个数
* 段落中问题关键词的个数
* 段落中与问题精确匹配的最长的关键词串
* 提取该段落的源文档的排名
* 原始查询中关键词之间的距离（proximity）
  + 对每个段落，确定其中能够包含该段落中所含关键词的最短文本跨度。
  + 系统更加偏好在更短的跨度中包含更多的关键词
* 段落和问题之间的N-gram重叠（N-gram overlap）
  + 答案段落中的N-gram和问题中的N-gram的重叠度越高，段落的权重就越高

### 答案处理

答案抽取任务：从段落中抽取特定的答案提供给用户。

常用的答案抽取算法

* 基于答案类型模型的抽取算法：使用期望的答案类型和正则表达式模板
  + 基于特定问题类型的答案模板，既可以手工编写，也可以自动学习。
  + 用于答案排序的自动学习的分类器使用的特征
    - 答案类型匹配：如果候选答案包含一个短语，其符合正确答案的类型，则为真
    - 模板匹配：匹配候选答案的模板编号
    - 与问题关键词匹配的数量：候选答案中包含多少个问题关键词
    - 关键词距离：候选答案与查询关键词之间的距离（用平均词汇数或者候选答案中出现在同一个语法片段的关键词数量来表示）
    - 新颖性因子：如果候选答案中至少有一个单词是新的，即未出现在查询中，则为真
    - 同位特征：如果候选答案是一个包含很多问题检索词的短语的同位语，则为真
    - 标点位置：如果候选答案后面紧跟一个逗号、句号、引号、分号或者感叹号，则为真
    - 问题检索词的序列：候选答案中出现的最长的问题检索词的序列长度
  + 评测答案模板精度的方法
    - 用户问答模板匹配的方法是保留具有高正确率的模板
* 基于N-gram的拼接算法（N-gram tiling），也称为基于冗余的方法（redundancy-based approach），从搜索引擎返回的重构查询结构摘要开始。
  + N-gram挖掘（N-gram mining）：抽取摘要中出现的所有unigram、bigram、trigram，并且赋予权重。
    - 权重是包含该N-gram的摘要的数量，以及返回该N-gram的查询重构模板权重的函数
  + N-gram过滤（N-gram filtering）：根据N-gram与预测的答案类型的匹配度，对N-gram打分。这些分数由为每个答案类型编写的过滤器计算。
  + N-gram拼接算法把重叠的N-gram片段连接成更长的答案
    - 贪心算法从得分最高的候选答案开始，试图将该答案与其他候选答案拼接，移除得分低的答案，直到剩下唯一的答案。

### 事实性答案的评价

TREC使用的主要度量：

* 平均逆排名（Mean Reciprocal Rank，MRR）：是一个内在的（intrinsic）评价指标，也叫体外的（in vitro）的评价指标。
  + MRR假设有一个已经标注好正确答案的测试问题集
  + MRR假设系统返回了排序后的答案列表或者包含答案的段落列表
  + MRR把每个问题的首个正确答案所处的排序号取倒数，作为该问题的得分。

## 摘要

文本摘要（text summarization）是从文本中提炼最重要的信息，并且根据特定和用户生成一个缩略版本的过程。

* 文档的大纲
* 科技文献的摘要
* 新闻报道的新闻提要
* 搜索引擎对结果网页的摘录
* 商务会议的实施条目或者其他摘要
* 电子邮件往来的摘要
* 压缩语句以生成简单的或者压缩的文本
* 通过摘录多篇文档生成复杂问题的答案

文本摘要的分类

* 单个文档摘要 和 多个文档摘要
  + 单文档摘要（single-document summarization）对单个文档生成摘要，用于生成新闻提要或者大纲，最终目的都是为了描述单个文档的内容
  + 多文档摘要（multiple-document summarization）对一组文档生成摘要，目的是为了浓缩整个文档集的内容
* 一般性摘要 和 针对查询的摘要
  + 一般性摘要（generic summary）是一种不需要考虑特定用户或者特定信息需求的摘要，只需要给出文档的重要信息即可。
  + 针对查询的摘要（query-focused summarization），也称为主题摘要（focused summarization）、基于主题的摘要（topic-based summarization）以及针对用户的摘要（user-focused summarization）。这种摘要是为了满足特定的用户查询才生成的，是一种回答用户问题的较长的非事实性答案。

文本摘要系统的关键的架构维度：

* 摘要（abstract）：需要使用别的词汇来描述文档的内容。
* 摘抄（extract）：最简单的摘要，从待摘要的文档中挑出短语或者句子，然后将它们连接起来。大多数文本摘要系统都是基于摘抄的，因为实现更加容易。

文本摘要系统，也叫自然语言生成系统，针对的三个问题：

* 内容选择（content selection）：要从待摘要的文档中选择信息。
  + 假设需要摘抄的粒度为句子或者从句
* 信息排序（information ordering）：对摘抄出来的单位进行排序，并且安排合理的结构
* 句子实现（sentence realization）：对摘抄出来的单位进行清理，使其在新的上下文中显得流利

## 单文档摘要

单文档摘要是以句子为摘抄单位，执行下面三个步骤：

* 内容选择：选择要从文档中摘抄的句子
  + 摘抄句子的内容选择可以看作分类任务，分类器的目标是为文档中的每个句子进行二元标注
    - 重要（摘抄）
    - 不重要（不摘抄）
* 信息排序：安排一个顺序，将这些句子安放在摘要的合适的位置
* 句子实现：清理句子，移除非必需的短语，基于连贯性原则将多个句子合并。

### 无监督的内容选择

选择包含了更显著的（salient）或者更多信息的（informative）词的句子，是依赖于表层特征（例如：词的显著性）的无监督算法。

显著性基于主题特征（topic signature）计算，即显著词（salient term）或者特征词（signature term）的集合，其中每个词的显著性得分必须大于某个阈值。

* TF-IDF
* 对数似然比（log-likelihood ratio，LLR）

基于中心的摘要（centroid-based summarization）算法族：可以把特征词的集合看成一个伪句子，这个集合是文档中所有句子的“中心”，查询就是找到最接近于中心句的句子。

* 基于LLR算法
* 基于句子所承载信息的核心程度对句子排序
* 基于图的中心度量

### 基于修辞分析的无监督摘要

使用篇章信息（例如：连贯关系）的无监督算法。

基于篇章分析器（Ref：Ch21）计算每个篇章单元之间的连贯关系。将句子分析为连贯关系或者分析树时，就可以基于核心单元对于摘要递归地抽取文本中的卫星单元

连贯分析树上每个结点的显著单元的递归定义：

* 基础情况：叶子结点的显著单元是该叶子结点本身
* 递归情况：中间结点的显著单元是直接核心的子结点的显著单元的联合

### 有监督的内容选择

给所有线索赋予权重并结合在一起的方法是有监督的机器学习方法。需要由人工创建对应的摘抄型摘要的文档集作为训练语料，将每个训练文档与其摘要对齐，找出文档中包含在摘要中的句子。

对齐（alignment）算法是找出源文档和摘要语句之间非信用词的最长公共子序列，或者计算最小编辑距离，或者使用更加复杂的短语源。

### 句子简化

句子实现（sentence realization）主要通过句子压缩（sentencce compression）或者句子简化（sentence simplification）。首先进行句法分析或者部分句法分析，通过一些代表性规则剪除或者保留某些句子。更加复杂的句子压缩模型基于有监督的机器学习方法，其中文档和人工摘要的平等语料被用来计算特定词汇或者短语结点被剪除的概率。

## 多文档摘要

多文档摘要适合基于网络的应用。

### 多文档摘要的内容选择

多文档摘要因为训练数据太少，所以主要关注于无监督算法。

多文档摘要与单文档摘要的最大区别是文档中存在大量的冗余。在一个文档集合中，除了每篇文档所表达的特定信息之外，各文档在词汇、短语和概念上都会存在明显的重叠。因此多文档摘要算法的关键是选择摘要句子时剔除冗余。

### 多文档摘要的信息排序

信息的排序或者结构化：

* 时间顺序（chronological ordering）策略：使用该报道相关的日期的排序技术
* 连贯性（coherence）策略
  + 让句子之间存在合理的连贯关系
  + 内聚性和词汇链的关系
  + 一个连贯的篇章，其中的实体提及必须相互协调，即共指。

基于共指的连贯性算法：也使用了基于中心（Centering）的思想。每个简单片段都有一个显著的实体，即焦点。焦点的句法特定实现以及实现之间的特定转移，可以创建更加连贯的对话。实体识别的结构序列可以自动抽取并且表示为一个实体网格。

根据句子或者句子序列之间的局部连贯性得分，为每个语句序列指派一个连贯性得分。然后句子之间的转移得分可以与词汇连贯性和基于实体的连贯性结合起来。句子排序的难度等同于NP完全问题。

信息排序任务是完全独立于内容选择的。

信息排序任务和内容选择任务合并在一起进行学习，得到一个对语句进行选择并且排序的模型。

### 多文档摘要的句子实现

对输出进行共指消解，抽取名称，并且运用下面的清理重写规则：

* 第一次提及时使用全名，之后的提及中使用姓；
* 第一次提及的某个修订的形式，之后的提及中移除同位语或者修饰语
* 句子合并算法连接不同句子中的短语，从而实现比摘抄语句粒度更小的实现

## 主题摘要和问答

针对查询的摘要，也叫主题摘要、基于主题的摘要、针对用户的摘要，是响应用户问题或者信息需求的一种较长的、非事实的答案。

一种针对查询的摘要就是摘录（snippet），即单个文档针对查询的摘要。

通过对多文档摘要技术简单修改就可以实现针对查询的摘要。

使用自顶向下的或者信息抽取的技术可以实现针对查询的摘要。

基于信息抽取的复杂问答系统：

* 信息检索
* 用分类器为每个句子打卡一个适合该领域的类别标签
* 通用的、多文档摘要的内容选择的音节，把还未落入特定信息抽取类别中的句子添加到答案里

## 摘要的评价

内在摘要评价进一种自动的方法，称为面向召回率的要点评估（Recall-Oriented Understudy for Gisting Evaluation，ROUGE）。根据机器生成的候选摘要和人工摘要的N-gram重叠数目，自动地为候选摘要评分。

侧重于摘要的含义的评价方法：金字塔方法（Pyramid Method），主要统计候选摘要和参考摘要共享了多少个意义单位。可以根据重要性为每个含义单位赋权；某个含义单位出现在越多的人工摘要中，那么这个含义单位的权重也就越高。

意义单位，也称为摘要内容单位（Summary Content Units，SCU），是一种子句结构的语义单位，对应命题或者连贯的命题片段。人们为每个参考摘要和候选摘要里的摘要内容单位进行标注，然后计算重叠的数量。

评价摘要的标准基准系统是随机句子和首句基准系统。当评价长度为N个句子的摘要，系统就随机选择N个句子，然后首名基准系统选择前N个句子。

## 小结

* 主导信息检索系统把文档和查询表示为词袋子
* 向量空间模型把文档和看成多维空间的向量
  + 文档与查询或者其他文档的相似度使用两个向量的夹角余弦度来表示
* 事实性问答系统的主要部分
  + 问题分类模型，用来决定答案的命名实体类型
  + 段落检索模型，用来识别相关段落
  + 答案处理模块，用来抽取和格式化最终答案
* 事实性问题可以使用平均逆排序（MRR）来评价
* 摘要可以是摘要式的或者摘抄式的，大多数算法都是摘抄式的
* 摘要算法是三个组件
  + 内容选择
  + 信息排序
  + 句子实现
* 单文档摘要算法主要关注句子的摘抄，依赖于所选择的特征
  + 句子在简单中的位置、词汇的信息量、提示短语、句子长度
* 多文档摘要
  + 多文档摘要算法在文档的句子上执行句子简化
  + 防止冗余性对于多文档摘要非常重要，实现方法是在句子摘抄时添加冗余性惩罚项
  + 多文档摘要的信息排序算法是保持整体的连贯性
* 针对查询的摘要的算法
  + 在一般摘要的算法基础上稍作修改
  + 使用信息抽取的方法

# 对话 和 会话智能代理

语言是人性和感知的标志，会话（conversation）或对话（dialogue）是语言最基本的，也是最优越的舞台。

会话智能代理（conversational agents）的基本结构和算法。会话智能代理是通过语音而不是通过文本进行交流的，因此也称为口语对话系统（spoken dialogue systems），或者口语系统（spoken language systems）。会话智能代理是一种能够使用自然语言与用户进行交流的程序。

## 人类会话的属性

会话是两个（或者多个）对话者之间的共同活动（joint activity）。基本事实有若干分支，会话由连贯的话轮（turns）组成，每个话轮由说话人和听者的共同行为（joint action）组成，听者需要做出特殊的推理，即对说话人真实意图的话语隐含（conversational implicature）的推测。

### 话轮与话轮的转换

对话的重要特征是话轮转换（turn-taking）。

会话和语言本身具有某种结构，从而能够有效地处理话轮的分配问题。

正常的人类会话中话段的时间。尽管说话人在谈论时可以与其他人重叠，但是实际平均的重叠数是非常小的。话轮之间通常有低于几百毫秒的时间间隔。

自然对话会以某种方式建立，人们能够快速领悟到下一个说话人是论证，以及确切的说话时间。这种话轮转换行为通常是会话分析（Conversation Analysis，CA）。

话轮转换的行为是由一组话轮转换规则制约的，这些规则使用在相关转换位置（Transition-Relevance Place，TRP）。

简化的话轮转换规则：

* 如果在该话轮中，当前的说话人已经选择了A为下一个说话人，那么下一个说话人就必须是A；
* 如果在该话轮中，当前的说话人没有选择下一个说话人，那么其他说话人可以在下一轮说话
* 如果没有其他人参加下一个话轮，那么当前的说话人可以接着参加下一个话轮

毗邻对（adjacency pair）或两人对话（dialogic pair）：

* 问题——回答（QUESTION-ANSWER）
* “问候”接“问候”（GREETING followed by GREETING）
* “称赞”接“自谦”（COMPLIMENT followed by DOWNPLAYER）
* “请求”接“准许”（REQUEST followed by GRANT）

在毗邻对的两个部分之间出现的沉默是有意义的沉默（significant silence）。例如：A刚刚问了B一个问题，沉默表示为拒绝，或者是不喜欢回应

说话人之间转换发生的地点不是任意的。相关转换位置（transition-relevance place）通常出现在话估（utterance）的边界。话段常常较短，更可能是单一从句，主语常常是代词而不是完整的词汇名词短语，并且充斥着停顿和修正。

### 语言作为行动：言语行为

对话中的话段是一种由说话人实施的行为（action）。行为分类：

* 以言表意（locutionary act）：也叫言内行为，发出一个带有特殊意义的句子
* 以言行事（illocutionary act）：也叫言外行为，发出一个带有询问、回答、承诺等行为的句子
* 以言生效（perlocutionary act）：也叫言后行为，发出一个对听话人的感情、信念或者行为产生一种特定效果的句子，往往是有意为之。

“言语行为”（speech act）用于描述以言行事。共分五个大类：

* 断言语（assertivees）：说话人对某事是某种情形的表态
* 指令语（directives）：说话人的目的是使听话人去做某事
* 承诺语（commissives）：说话人对将来的行为做出承诺
* 表情语（expressives）：说话人对一些事情的心理状态
* 宣告语（declarations）：说话人说出而使外在世界产生新的情景

### 语言作为共同行动：对话的共同基础

对话是说话人和听话人的共同行为，因此在对话中说话人与听话人需要不断地建立共同基础（common ground），即对话双方都认可的事物集合。需要获得共同基础意味着听话人必须依靠（ground）说话人的话段，确保听话人已经理解说话人的意思和意图。

自证的原则（principle of closure），也叫闭合原则，或者叫闭包原则。智能代理做出一个行动时，对于当前目标需要有充分的证据表明他们已经成功地完成了这个行动。

自证的过程：即每个共同的语言行动或者贡献（contribution）有两个阶段：

* 表示（presentation）阶段：说话人通过一个语段向听话人叙述，执行一种言语行动
* 接受（acceptance）阶段：听话人必须处理这个话段，即表示自己是否已经理解了说话人的意图。接受的五种方法（从弱到强）：
  + 继续关注（continued attention）：因为对A的陈述保持满意，B表明准备继续关注。
  + 相关邻接贡献（next contribution）：B开始邻接相关的贡献
  + 确认（acknowledgment）：B点头确认或者表示接续
  + 表明（demostration）：B表明她已经理解A所有的或者部分的陈述
  + 展示（display）：B逐字地展示A所有的或者部分的陈述

单词uh-huh是一个接续（continuer），也叫确认标记（acknowledgment token），或者叫反输（backchannel）。接续是一个（短的）可行的话段，它确认前面话段的某些内容，并且不需要其他人的确认。

共同基础和贡献的概念可以与言语行动结合起来，给出一个关于会话的更加复杂的共同行动模型，这些更加复杂的模型称为对话行为（dialogue acts）。

### 会话结构

会话的全局组织（overall organization）。会话包含许多其他各类的结构，包括会话结束的复杂性质，以及前序列的广泛使用。

### 会话隐含

隐含（implicature）是允许的推理中的一个特殊类别。会话隐含（conversational implicature）的4个普遍准则：

* 数量准则（maxim of quantity）：与需求正好一致的信息
  + 贡献与需求一致的信息（满足当前交流的目的）
  + 一定不要贡献多于需求的信息
* 质量准则（maxim of quality）：尽可能贡献真实的信息
  + 不要说人认为是虚假的信息
  + 不要说对你而言缺乏足够证据的信息
* 相关准则（maxim of relevance）：贡献切题的信息
* 方式准则（maxim of manner）：贡献清楚的信息
  + 避免模糊的表达
  + 避免歧义的表达
  + 简短，避免啰嗦
  + 有序，避免混乱

## 基本的对话系统

语音识别

自然语言理解

对话管理器

任务管理器

TTS组件

自然语言生成器

图24.5会话智能代理的各组件的简化结构图

### ASR（自动语音识别）组件

听写或者转录的大词汇表语音识别器，关注于转录任意话题的任意单词的任意句子。

对话系统使用基于有限状态语法的非概率语言模型，语法通常由人工编写，编写语言通常是VoiceXML。这些语法的语言模型可以自动编制，例如：使用自然语言理解中的合一语法自动编制。

会话智能代理系统中的语言模型通常是对话状态依赖（dialogue-state dependent）的，这种依赖对话状态的语言模型称为约束语法（restrictive grammar）。

为了满足说话人切换带来的语音变换，还需要使用说话人自适应技术。

对话系统还要求ASr系统返回句子的置信度（confidence）。

### NLU（自然语言理解）组件

NLU组件需要产生适合对话任务的语义表示，基于框架和槽语义学（Ref：Ch22）。

NLU生成语义表示可以基于语义附着合一语法（Ref：Ch18）。一个句法分析器产生句子的意义，然后从其中抽取出槽的填充值。其他的对话系统依赖于简单的领域特定语义分析器，例如：语义语法（semantic grammars）。

语义语法是上下文无关语法或者递归转移网络（Recursive Transition Networks，RTN）形式，规则的左部对应需要表达的语义实体。

语义语法可以使用标准的CFG语法分析算法进行分析（Ref：Ch13）。

语义语法不能处理歧义，还需要人工编写语法规则，因此代价高，建模慢。

* 使用概率论的语义语法系统来解决歧义问题，但是依然需要手工编写语法
* 使用语义HMM模型既可以解决歧义问题，还可以避免手工编写语法问题。HMM的隐藏状态是语义槽标签，观察值是槽填充。包含两个组成部分的生成模型
  + 选择表达的含义，赋值为语义槽序列的先验概率，通过概念的N-gram模型计算
  + 选择用什么词来表达含义，在给定槽的情况下，生成某一词串的依然概率经。

### 生成和TTS组件（自然语言生成器）

生成任务的两个部分：

* 内容规划（content planner）模块完成 “说什么”（what to say），即决定向用户表达什么内容
* 语言生成（language generation）模型完成“怎么说”（how to say it），即选择表达这个意思的语法结构和单词。
  + 基于模板的生成方法：对话设计者预先设定好对用户所说句子中的所有的或者大多数的单词，利用这些模板创建的句子称为提示（prompt）。
  + 自然语言生成方法。对话管理器先对需要表达的话段含义建立对应的表示，然后将这个含义的表示传递给一个完整的生成器。生成器的三个组件（图24.8）
    - 句子规则组件
    - 表层实现组件
    - 韵律赋值组件

### 对话管理器

对话管理器从ASR/NLU组件接受输入，维护一些状态，与任务管理器交互，并将输出传递给NLG/TTS模型。

对话管理器控制着对话的架构和结构。四种常用的对话管理架构：

* 有限状态对话管理器，系统完全控制着与用户的对话，这种控制对话的系统为系统主动（system initiative）或者单一主动（single initiative）系统
  + 受限单一主动有限状态对话管理器系统一起知道用户在回答哪个问题，有时系统也会提供万能（universal）指令，可以在对话的任何地方使用。
  + 完全系统主动的、有限状态的对话管理器系统的限制过于严格
* 基于框架（frame-based）或者基于表格（form-based）的对话管理器系统需要询问用户问题，并且填充框架里面的槽，同时也允许用户通过提供填充框架中的其他槽的信息来引导对话，每个槽可能会与一个问题对应，一起到有足够的信息构建数据库查询，然后返回结果给用户。基于框架的系统一般实现为生成规则（production rule）系统。
* 信息状态对话管理器
  + 基于马尔可夫决策过程的概率化的信息状态对话管理器
* 基于规划的对话管理器

主动权是指在某一时刻论证掌握着会话的控制权。

会话主动权：

* 系统主动
* 用户主动
* 混合主动（mixed initiative）
  + 系统或者用户可以多种方式，任意地接管或者放弃主动权
  + 在基于框架的系统中，存在一种受限的转换，即根据提示类型（开放类型 或者 指令类型）和ASR中使用的语法类型的结合体进行工作。
    - 开放提示（open prompt）是一个系统允许用户按其所需进行回答的提示
    - 指令提示（directive prompt）是一个系统显式要求用户如何作答的提示

严格的语法是强烈地约束ASR系统的语言模型，只识别给定提示的适当回答。

### 任务管理器的错误处理：确认和拒绝

在一个对话系统中，听错是一个特别重要的错误类别

有两种方法类确认：

* 向用户确认（confirming）理解，确认的策略：
  + 显式确认（explicit confirmation）策略：询问用户一个直接问题以确认系统的理解。
    - 让用户纠正系统的误识别更容易，因为只需要对确认问题说“不”
    - 比较笨拙，容易增加会话的长度
    - 确认的对话片段不自然、不完全人性化
  + 隐式确认（implicit confirmation）策略：使用前面所介绍的示范（demostration）或者展示（display）策略，通过复核系统对用户据说的内容的理解，向用户进行确认。
    - 隐式确认更加自然
* 拒绝（rejecting）系统可能误解的话段
  + 当一个话段被拒绝时，系统会遵循渐进提示（progressive prompting）或者逐步细化（escalating detail）的策略
  + 快速重提示（rapid reprompting）策略，系统在拒绝话段时，仅仅提示用户再说一次，只有当话段被第二次拒绝时，系统才会使用渐进提示。用户更加喜欢这个策略。

## VoiceXML

VoiceXML(Voice Extensible Markup Language)，是一种由W3C发布的基于XML的对话设计语言，也是语音标记语言（如SALT）中最常用的一种。目标是基于框架的架构中创建简单语音对话，通过使用ASR和TTS模块，并处理非常简单的混合主动。

VoiceXML文档包括一个对话集合，每个对话可以是一个表格（form）或者目录（menu）。

## 对话系统的设计与评价

### 对话系统的设计

语音用户界面（Voice User Interface，VUI）设计中以用户为中心的设计原则：

* 研究用户和任务。通过与用户面谈、调研类似系统、研究相关的人类对话、获潜在用户和任务的性质。
* 创建仿真和原型。
* 对设计进行迭代用户测试。

### 对话系统的评价

计算用户满意度，先让用户与对话系统的某一任务进行交流，然后对用户进行问卷调查。

最理想的对话系统是使用最少的代价帮助用户实现目标：

* 任务成功完成度（task completion success）。任务成功完成度可以用总体解答的正确性来衡量。
  + 在基于框架的架构中，正确性可以根据正确填充槽所占的百分比，或者完成子任务所占的百分比来计算。
  + k系数来计算任务成功完成度得分，因为这个得分已经对偶然一致进行了归一化，因此可以进行跨系统比较。
* 效率成本（efficiency cost）。是系统在帮助用户时的效率指标。可以用按秒计算的对话总时间、总话轮或者系统话轮的数量，或者总的查义数来衡量。还有系统无应答的数量和“话轮纠正率”：系统或者用户单独用来纠正错误的话轮数除以总的话轮数。
* 质量成本（quality cost）。是衡量交流时其他能够反映用户对系统看法的因素。
  + 一种量度是ASR系统没有返回任何句子的次数，或者是ASR拒绝提示的数量
  + 用户必须干扰（打断系统）的次数，或者用户回复不够迅速而引起的超时提示的次数
  + 系统对用户的理解和回复的优劣上。
    - 系统问题、答案以及错误信息的不妥（冗长或者歧义）
    - 每个问题、答案或者错误信息的正确性
  + 概率正确率（concept accuracy）或者概念错误率（concept error rate）：衡量NLU组件正确返回语义概念所占的百分比
    - 基于框架架构的系统可以通过统计用正确含义填充的槽来计算这个指标。

PARADISE算法可以采用多次回归的方法对成功完成度打分和成功指标进行融合，并且赋予权重。

## 信息状态和对话行为

信息状态架构能够提供关于言语行为和共同基础的复杂解释和生成模型，以及对话上下文的更加复杂的表示。

马尔可夫决策过程模型是信息状态方法扩展的概率化架构。

信息状态架构的组成部分：

* 信息状态（“话语上下文”或者“心智模型”）
* 对话行为解释器（“对话引擎”）
* 对话行为生成器（“生成引擎”）
* 一组随着对话行为的解释逐步进行更新的规则，包括生成对话行为的规则
* 选择应该使用哪条更新规则的控制结构

信息状态（information state）包括：话语上下文、两个说话人的共同基础、说话人的信息或者意图、用户模型等内容。

对话行为是整合了共同基础理论之后，对言语行为的扩展。

更新规则使用对话行为的信息修改信息状态。这些更新规则是基于框架的对话系统所使用的产生式规则的泛化。更新规则的子集，称为选择规则（selection rule），被用来生产对话行为。当一个问题被识别时，一条更新规则可能会指定回答该问题的需求。

行为智能代理（Behavioral Agent）是更新规则和控制结构的结合体。

信息状态包含丰富的话语模型，如话语表示理论（Discourse Representation Theory，DRT），或者关于信念、愿望和意图的复杂模型。

### 使用对话行为

会话行为（dialogue acts），也叫对话行动、会话行动，是言语行为的扩展，特别是与其他会话功能紧密相关的内部结构的扩展。

大量的对话行为标注集

“对话系统多层标记语言”（Dialogue Act Markup in Several Layer，DAMSL）体系中，每个话段被标为两种功能：

* 类似言语行为功能的向前功能（forward looking function）
* 类似共同基础和回答的向后功能（backward looking function）

行为类型的层次：

* DAMSL对话行为
  + 核心言语行为
  + 共同基础行为
* 论证关系
  + 话轮转换行为
  + 连贯关系

### 解释对话行为

采用陈述句的形式提出疑问，或者采用疑问的形式发出请求，都被称为间接的言语行为

将对话行为的解释看成一个有监督的分类任务，其中待检测的隐藏类别是对话行为的标签。训练分类器使用已经手工标注了每个话段的对话行为的语料库。用于对话行为的解释的特征可以从会话上下文以及行为的微语法。

### 检测纠正行为

用于识别用户对系统错误的纠正的检测器。纠正比一般的句子更难识别。

纠正的错误识别率是非纠正的两倍的原因：

* 说话人在纠正时使用了极为不同的韵律风格，即高清晰发音
* 说话人使用了多种策略进行纠正
  + 复述（REP）
  + 解释（PAR）
  + 添加内容（ADD）
  + 忽略内容（OMIT）

多种基于机器学习的分类器被用于检测纠正行为，常用的特征：

* 词汇信息
* 韵律和高清晰的发音
  + F0范畴的增长
  + 暂停时长
  + 单词时长
* 话段长度
* ASR特征
  + 置信度
  + 语言模型概率
* 对话特征

### 生成对话行为：确认和拒绝

Stent是一个对话行为生成模型，将一组更新规则用于内容规划：

* 如果用户刚刚释放话轮，系统可以采取话轮转换行为
* 如果系统需要为用户总结一些信息来解决问题，系统可以将信息作为语义内容来使用ASSERT会话行为，然后通过自然语言生成系统的标准技术，将内容映射到单词。

许多对话行为生成模型关注的是确认行为和拒绝行为，通常基于概率方法解决。

对话系统可以选择显式确认或者隐式确认。

分类器的特征：

* 话段的置信度：用来显式地确认低置信度的句子
  + 置信度是语音识别器赋予转换句子的一个度量，用于衡量该转换的可信度有多高（Ref：Sec10.1）
  + 置信度一般根据话段的语音对数似然计算得来（概率越高，置信度越高），韵律特征也可以用作置信度预测。
* 出错时需要付出的代价。

拒绝话段基于三层结构的置信度：

* 低于某个置信度阈值，拒绝话段
* 高于该阈值，显式确认该话段
* 置信度更高，使用隐式话段

## 马尔可夫决策过程架构

将对话建模为马尔可夫决策过程（Markov decision process，MDP），就可以基于优化某些收益和成本来改变对话系统的行为

## 基于规划的对话行为

智能规划技术是会话智能代理最早的行为模型，也是最复杂的模型。使用这种技术来生成和解释句子，需要有信念（beliefs）、愿望（desire）和意图（intentions），即BDI模型。

BDI模型是基于规划的对话模型。

### 规划推理的解释和生成

利用谓词演算对信念和愿望的形式化定义，来进一步扩展对理解和生成规划的说明。

BDI模型需要行为和规划的公理化。最简单的是基于行为方案（action schema）集，基于简单的智能规划模型。每个行为方案都有一个参数集，用于约束每个变量的类型和其他部分：

* 前提（precondition）：为实施该行为必须为真的那些条件
* 效果（effect）：为实施该行为为结果的变为真的那些条件
* 体（body）：在实施该行为中必须达到的部分有序的目标集

### 对话的意图结构

话语的片段通过：解释或者说明的连贯关系而关联，这些连贯关系描述了话语片段之间的信息关系。话段解释的BDI方法是意图方法，使得对话连贯的是意图结构，即每个话段潜在的基于规划的说话人意图。

基于每个话语具有一个称为话语目的（Discourse Purpose，DP）的潜在意图的假设，这些意图可以在模型中被实例化，并且由实例化它的人持有。话语中每个话语片段也有一个相应的目的，被称为话语片段目的（Discourse Segment Purpose，DSP）。

连贯关系：

* 支配（dominance）
* 满足——优先（satisfaction-precedence）

辅助话语片段也被称为子对话（subdialogue）。辅助对话称为信息共享、知识前提子对话，因为它们被代理人发起以帮助满足更高层目标前提条件。

对话中意图结构的推理算法类似于对话行为的推理算法

* BDI模型
* 基于线索短语的机器学习架构
* 基于韵律的机器学习架构
* 基于其他线索的机器学习架构

## 小结

会话智能代理是重要的语音和语言处理应用

* 对话系统由5个组件构成：
  + 语音识别
  + 自然语言理解
  + 对话管理
  + 自然语言生成
  + 语音合成
  + 对于某一任务领域，可能会有一个特殊的任务管理吕
* 常用的对话架构
  + 有限状态架构
  + 基于框架的架构
  + 信息状态框架
    - 马尔可夫决策过程
  + BDI（Belief-Desire-Intention）模型
* 人类对话现象
  + 话轮转换
  + 共同基础
  + 会话结构
  + 主动权
* 在对话中，说话是一种行为，称作言语行为或者对话行为。模型的作用是生成和解释这些行为。

# 机器翻译

机器翻译适合的非文学任务：

* 粗略翻译（rough translation）
  + 网络信息获取
* 采用人工译后编辑（post-editor），也叫计算机辅助翻译（computer-aided human translation，CAHT）。
  + 软件手册的本地化
* 可以实现全自动高质量翻译（Fully Automatic, High-Quality Translation，FAHQT）的子语言限定领域
  + 天气预报

## 机器翻译面临的困难

语言之间的翻译差异（translation divergences）：

* 系统性的差异，即可以以一种通用的方式进行建模的差异
* 需要单独处理的差异
  + 个体（idiosyncratic）差异
  + 词汇差异差异

### 类型学

类型学（typology）：对不同语言的相似性和差异性进行系统研究的学科

* 从形态学（morphologically）的角度，语言的两种分类方法：
  + 根据每个单词的语素数目
    - 孤立语（isolating language）：每个单词只有一个语素，如：汉语
    - 多式综合语（polysynthetic language）：每个单词可能包含多个语素
  + 根据语素是否可以被切分
    - 粘着语（agglutinative language）：语素有相对比较清晰的边界
    - 整合语（fusion language）：单个词缀可能合并了多个语素，如：俄语
* 从句法（syntactically）角度看，语言间的突出差异在于其简单陈述句的动词（verb）、主语（subject）和宾语（object）的基本顺序
  + 主语-动词-宾语（SVO）：动词放在主语和宾语之间，如：英语和汉语
    - SVO语言通常具有前置词（preposition）
  + 主语-宾语-动词（SOV）：动词放在基本句的末尾，如：日语和印地语
    - SOV语言通常具有后置词（postposition）
  + 动词-主语-宾语（VSO）：动词放在基本句的前面，如：阿拉伯语和希伯来语
* 从论元结构（argument structure）及这些论元与先行词（predicates）的链接关系
  + 中心语标记（head-marking）语言：将中心语和其附属语之间的关系标记于中心语，如：匈牙利语
  + 附属语标记（dependent-marking）语言：将中心语和其附属语之间的关系标记于非中心语，如：英语
* 将事件的概念属性映射到特定的单词
  + 动词框架（verb-framed）语言：将运动的方向标记于动词，运动的方式标记于外围词，如：日语和罗曼语
  + 外围词框架（satellite-framed）语言：将运动的方向标记于外围词，运动的方式标记于动词，如：英语和汉语
* 能够省略的事物
  + 代词省略语言（prop-drop）：允许省略代词的语言
    - 日语和汉语的省略频率远高于西班牙语
    - 指代密度（referential density）表示代词省略的频率
      * 冷语言，即指代稀疏的语言。需要语言的接收者完成更多的推理工作来恢复指代的原意
      * 热语言，显式标明指代关系的语言，语言接收者理解更加容易

### 其他的结构差异

除了类型学差异外，来自于特定语言的单一的专有特征。

### 词汇差异（翻译中词汇的问题）

翻译时的明确化（specification）问题：

* 词义排歧问题：需要将一个不确定的词在目标语言中更加确定
* 语法排歧问题：需要加入更多的语法约束来帮助词汇选择
* 伪词（spruous words）：对于不存在的词使用词汇空白（lexical gap）填充

## 经典的机器翻译方法与Vauquois三角形

三种典型的机器翻译方法：

* 直接方法（direct）：源语言文本中的词是一个接一个地进行处理，使用一部较大的双语词典，词典中的每个条目用于翻译每个词
* 转换方法（transfer）：对输入文本进行解析，利用规则将源语言的解析结果转换到目标语言的解析结构，再利用这个解析结果得到目标语言的句子
* 中间语言方法（interlingua）：对源语言文本进行分析，得到抽象的意义表示（中间语言），然后根据这个中间表示生成目标语言

Vauquois三角开表示这三种方法（图25.3）：从直接翻译方法到转换方法到中间语言方法的变化，所需要的语言分析程度不断加深，所需要的转换知识的数量不断递减。

* 直接方法需要大量的转换知识（每个词的对应条目都是转换知识）
* 转换方法的转换规则仅用于句法分析树或者语义角色
* 中间语言方法则不需要特定的转换知识

### 直接翻译

在直接翻译中，源语言需要逐词翻译。除了浅层的词形分析外，不需要使用中间结构。每个源语言的词直接映射成目标语言的词，然后基于调序规则把目标语言的词输出为目标语言的句子。

### 转换方法

对比知识（contrastive knowledge）：是指关于两种语言差异性的知识。

基于转换的模型就是使用对比知识策略的翻译系统。预先假定一个源语言的解析器，然后通过一个生成阶段以获得实际的输出句子。

模型中的三个阶段：

* 分析（analysis）
  + 词义消歧技术在源语言的分析过程中进行处理
* 转换（transfer）：
  + 句法转换（syntactic transfer）：是指从一个树结构映射到另一个树结构的操作过程。
    - 句法转换规则：将源语言句法分析树改变为目标语言句法分析树的规则
  + 词汇转换（lexical transfer）：基于一部双语词典
    - 词典本身能够处理词汇的歧义问题。
  + 语义转换（semantic transfer）：使用语义角色的方法
* 生成（generation）

### 直接翻译与转换方法的融合

商用的机器翻译系统结合了直接方法与转换方法，使用双语词典、标注器和句法分析器。

Systran翻译系统的三大组件：

* 浅层的分析阶段
  + 词型分析和词性标注
  + NP、PP以及较大短语的组块分析
  + 浅层的依存句法分析（主语、被动语态和中心词修饰成分）
* 转换阶段
  + 成语、俗语的翻译
  + 词义消歧

### 中间语言的思想与使用意义

中间语言方法，将翻译过程看成从输入句子中抽取句子意义，再将这一意义表示成目标语言。机器翻译系统不再需要对比知识，仅仅依赖于由符合该语言标准的解析器和生成器所使用的相同的句法和语义规则。所需要的知识数量将会和翻译系统能够处理的语言数成正比，而不是和语言数的平方成正比。

这种方案假定存在一种意义表示或者中间语言，这是一种与语言无关的规范形式。

中间语言使用的表示方案：

* 一阶逻辑表示脏话其变形，如：最小递归语义（minimal recursion semantics）
* 将语义分解为某些类型的原子语义基元
* 基于事件的表示。事件与论元是通过一个小规模的确定的语义角色集合进行链接的。
  + 需要指定事件的时态和语态属性
  + 表示实体之间的非事件性的关系

利用语义分析器从源语言中获得这种中间语言表示。

## 统计机器翻译

统计机器翻译通过对忠实度和流利度分别建立概率模型，再通过模型联合计算出最有可能的翻译结果。

统计机器翻译的形式化噪声信道模型和模型应用中的三个模型：

* 语言模型language model）（Ref：Ch4）
* 翻译模型（translation model）
* 解码器（decoder）

## 基于短语的翻译模型

基于短语的（phrase-based）的统计翻译：是将短语（带有词序列）以及单个的词作为翻译的基本单位。

基于短语的翻译系统的概率模型：

* 翻译概率（translation probability）
* 变形概率（distortion probability）
  + 对一个短语在两种语言（句子）中所处位置的距离度量

基于短语的翻译系统的模型：

* 解码模型：从表面的源语言字符串中得到隐藏的目标语言的字符串
* 训练模型：帮助学习解码的参数

短语对齐（phrase alignment）

## 翻译中的对齐

词语对齐（word alignment）：在平等句子集合中，源语言词和目标语言词之间的映射关系。

基于统计的对齐算法：IBM模型1和HMM模型。

注：IBM模型1 是IBM研究者论文中提到的5个模型中最简单的那个。

### IBM模型1

IBM 模型1的训练采用EM算法。

### HMM对齐模型

HMM模型是使用链规则对IBM模型1重新构造。

## 对齐模型的训练

平行语料库（parallel corpus），也叫平行文本（parallel text）或者双语文本（bitext），指的是具有两种语言版本的文本。

句子切分（sentence segmentation），也叫句子对齐（sentence alignment）。是训练的第一步，结果是将语料切分成句子。对齐方法是根据句子中的词或者字符的个数，不考虑句子中词的含义。

### 训练对齐模型的EM算法

EM训练的思想是：

* E步骤，基于对隐藏变量（对齐）的和来计算t参数的期望计数
* M步骤，利用这些计数来重新计算t概率的极大似然估计值

## 基于短语的翻译模型的对齐

基于HMM模型不能得到源语言中的多词短语与目标语言中的多词短语的对齐。因此，可以基于对称（symmetrizing）的方法对HMM模型进行扩展，使得对于句子对，能够产生短语到短语的是。

## 基于短语的统计机器翻译的解码

解码器是针对源语言的句子，根据翻译模型和语言模型来生成最佳的翻译句子。

寻找翻译模型和语言模型概率最大的句子的过程是一个搜索问题，因此解码也是一种搜索。

机器翻译中的解码器是基于最佳优先搜索（best-first search）方法的，是一种启发式（heuristic）搜索或者提示性搜索（informed search），这种搜索算法利用来自于问题领域的知识来进行指导。基于一个评价函数，在搜索空间上确定一个结点。

机器翻译的解码器是一种特殊的最佳优先搜索算法的变形，称为A\*搜索。（Ref：Sec10.2）

每个状态都伴随一个代价：

* 当前代价（current cost）：是在翻译假设中到目前为止已经翻译的短语的总的概率，即翻译概率、变形概率以及语言模型概率的乘积
* 未来代价（future cost）是对未翻译部分估计的翻译代价

搜索过程中需要使用柱搜索剪枝（beam search pruning）技术。

解码器会利用假设重组（recombining hypotheses）来减少搜索空间中的组合爆炸问题

## 机器翻译的评价

翻译可以从两个维度进行评价：

* 忠实度（fidelity）
  + 充分性（adequacy）是判断翻译结果是否包含原文中存在的信息
  + 信息性（informativeness）是判断翻译结果是否包含了足够完成任务的信息
* 流利度（fluency）
  + 可理解度
  + 清晰度（clarity）
  + 可读度
  + 自然度（naturalness）
  + 文本（style）

### 人工评价

编辑代价（edit cost），即将机器翻译结果经过后编辑修正为一个好的翻译所需要的代价。用于机器翻译输出的后编辑评价。

### 自动评价：BLEU

在BLEU值中，通过计算机器翻译结果与人工翻译结果间的N-gram重叠数目的一个事权平均值来驿机器翻译结果进行排序。

修正过的N-gram精度：首先统计一个词在不同参考译文中出现的最大次数，然后基于这个最大参考数目修正翻译中每个候选词出现的次数。

自动评价标准适用于评价单一系统的增量式改进或者评价具有相似架构的多个可比系统。

## 机器翻译的句法模型

转录语法（gransduction grammar），又称为同步语法（synchronous grammar）

* 是基于句法转换翻译系统利用现代统计原理的一个具体实现。
* 是用于描述结构上相关的语言对。
* 从语言生成的角度，转录语法是同时生成两种语言的对齐句对的一种方式。
* 转录语法是有限状态转录机的一种泛化。

反向转录语法（Inversion Transduction Grammar，ITG）。

## IBM模型3和繁衍度

IBM的模型3、模型4、模型5都使用了繁衍度（fertility）。

繁衍度是指从源语言中的一个单词对应的目标语言中单词的个数。

### IBM模型3的训练

不需要对模型3做手工对齐，使用EM算法来同步地学习对齐和概率模型。

## 机器翻译的MEMM模型

对数线性模型的常用特征：

* 语言模型
* 翻译模型
* 反向翻译模型
* 两个翻译模型的词汇化版本
* 词数惩罚
* 短语惩罚
* 未登录词惩罚

利用标准的最大互信息原则进行训练，直接在评价标准（如BLEU值）上进行优化，即最小错误率训练，又叫MERT方法。

## 小结

* 不管在结构上还是在词汇上，语言之间存在许多差异
* 类型学领域研究了翻译面临的一些困难
  + 语言可以通过带有类型维度的位置信息进行分类
* 传统机器翻译的三大范式
  + 直接方法
  + 转换方法
  + 中间语言方法
* 统计机器翻译基于噪声信道模型结合了翻译模型和语言模型
  + 基于短语的机器翻译是统计机器翻译的主要范式，基于一个双语短语表
    - 柱搜索剪枝的栈解码可以用于基于短语的机器翻译解码
  + IBM模型1，HMM模型，IBM模型3等模型对于生成句子对齐极为重要，从这些生成的对齐中可以抽取短语表
  + 对齐模型同样可以用于机器翻译解码
* 机器翻译的自动评价标准包括：BLEU、TER、METEOR、准确率和召回率
* 现代统计机器翻译系统基于对数线性模型，训练使用MERT方法。