第1章 引言

计算机自然语言处理：用计算机通过可计算的方法对自然语言的各级语言单位（字、词、语句、篇章等等）所进行的转换、传输、存储、分析等加工处理。

中文语言处理：以计算机为工具，采用可计算的方法对中文信息进行自动加工处理。

从技术路线来区分：

- 基于统计的语言处理技术：从大规模真实语料库中获得各级语言单位上的统计信息，并且依据低级语言单位上的统计信息，用相关的统计推理技术计算机高级语言单位上的统计信息
- 基于语言学规则的语言处理技术：通过对语言学知识的形式化，形式化规则的算法化，以及算法实现等步骤将语言学知识转化为计算机可以处理的形式。

从语言处理对象来区分：

- 字处理技术
  - 信息处理用的汉字机内码，定义了汉字在计算机内部的存储方式
  - 汉字输入码（或称汉字外码）提供了汉字输入的途径。
  - 汉字字形库（或称汉字字形码、汉字发生器编码）存储汉字的各种字体的点阵或者曲线矢量字形信息，通过专门的处理程序把要输出的汉字转换成对应的汉字字形后在显示器或者打印机上输出。
- 词处理技术
  - 词是自然语言中最小的有意义的构成单位，是自然语言处理中最基本的研究对象，是其他研究的先行和基础
  - 分词
    - 分词规范：GB13715，判别标准就是“使用频繁，结合紧密”
    - 常用方法
      - 正向最大匹配
      - 反向最大匹配
      - 基于词网格的统计方法
    - 主要难点

- 歧义消解
- 新词识别

- 词性标注
  - 常用方法
    - 基于隐马尔可夫模型
    - 基于词典知识库
    - 基于统计分类
      - 朴素贝叶斯
      - 最大熵模型
- 词义消歧

- 语句处理技术
  - 语句处理是建立在汉字编码、汉语词语切分、汉语词法分析等基础之上
  - 语句处理是汉语篇章理解的基础
  - 汉语短语处理技术
    - 基于语法规则的方法
    - 基于数学统计的方法
    - 规则与统计结合的方法
    - 概念层次网络模型（HNC）
- 篇章处理技术
  - 话语结构分析：研究文章的话语结构，跨越语句本身的多个语句、段落之间在结构或者语义上的相互关系的分析
    - 基于语法结构的衔接性（cohesion）分析：结构与形式上的衔接

- 文本呈现出来的表面结构如何彼此串联，文本中间是如何运用适当的连接词或者副词来串联句子；
- 语法层次上，句子和句子之间是如何依赖同样的主题词以及类似的句法结构来串联

◆ 基于语义之间的连贯性（coherence）分析：语义上的连贯，语义层面上更加抽象的一致性，即文本实体之间是否基于相同的主题进行讨论。

◆ 语法与语义结合考虑

- 人类在判断文本实体间的语义连贯性会参照实体之间的语法衔接关系
- 人类在表达内容时借助形式上的语法衔接关系来反映语义上的连贯性

◆ 话语结构分析应用于自动文摘领域

从语言处理的应用领域来区分：

- 应用基础技术
- 应用技术
  - 汉字处理的应用技术
    - ◆ 汉字排版
    - ◆ 印刷体识别
    - ◆ 联机手写汉字识别

第2章 数学基础

语言的模型化：通过将语言信息形式化，使之能够按照严密规整的数学形式表现出来。

2.1 初等概率理论

统计语言处理：以自然语言为处理对象进行统计指导。

统计语言处理的两个步骤：

- 收集自然语言词汇的分布情况，即统计语言单位出现的频率
- 根据这些分布情况进行统计推导

基本概念

概率论是研究随机现象的数学分支。

随机试验：随机现象的实现和对随机现象的观察。

基本事件：随机试验的每一个可能的结果。

随机事件：简称事件，是一个或者一组基本事件。

事件的概率：是衡量事件发生的可能性的度量。

样本空间：随机试验的所有可能结果或者全体基本事件构成的集合。

频率：描述了事件出现的频繁程度。概率的统计定义，由此确定的概率称为统计概率。

概率的公理化定义：

- 非负性：$0 \le P(A) \le 1$
- 规范性：$P(\Omega)=1$，$P(\varphi)=0$
- 完全可加性，也称为概率的加法定理。
  - $P(A_1 \cup A_2 \cup \dots \cup A_n)=P(A_1)+P(A_2)+\dots+P(A_n)$
- $P(A \cup B) = P(A)+P(B)-P(A \cap B)$

事件分类：

- 差事件：B-A

- 积事件：A∩B
- 子事件：A⊆B
- 逆事件：$\bar{A}$ =Ω -A
- 和事件：A∪B

条件概率与独立性

条件概率：P(A|B)。事件 B 出现条件下事件 A 出现的概率。

先验概率：不考虑先决条件而得到的发生这个事件的概率。

后验概率：在具备该事件出现的信息或者知识的条件下得到的发生这个事件概率。

概率的乘法定理：P(AB)=P(B)P(A|B)=P(A)P(B|A)

概率的链规则：P(A_1∩ A_2∩ …∩ A_n)=P(A_1)P(A_2|A_1)…P(A_n|∩^(n-1)_(i=1) A_i)

事件的独立性：P(A∩ B)=P(A) P(B)，P(A∩ B∩ C)=P(A|C)P(B|C)

全概率公式与贝叶斯公式

划分：满足如下条件的一组事件 B_1,B_2,…,B_n，称为样本空间 Ω 的一组划分。

- B_i B_j = φ, i ≠ j, j = 1, 2, …, n
- B_1∪ B_2∪ … ∪ B_n = Ω

全概率公式（breakdown law）：

P(A)=P(B_1)P(A|B_1)+P(B_2)P(A|B_2)+…+P(B_n)P(A|B_n)=∑^(n)_(i=1) P(B_i)P(A|B_i)

贝叶斯定理：

随机变量

随机变量：E 是随机实验，样本空间是 Ω = {e}，对于每个 e∈Ω，都有一个实数ξ(e)与之对应，

就得到一个定义在 Ω 上的单值实值函数 ξ = ξ (e)。

随机变量通常或者。

随机变量根据取值不同分类：

- 离散型随机变量
  - 当且仅当 Ω 是有限集合或者可数集合
  - 使用大写英文字母 X, Y, Z 表示
  - f(x) = P(X=x)，f(x)是频率函数
  - $P(X \leq x) = \sum_{(x_i \leq x)} f(x_i)$
  - $P(\Omega) = \sum_{(x \in \Omega)} f(x) = 1$
- 连续型随机变量
  - 当且仅当其分布函数 F 连贯并且除了有限个点之外处处连续可导
  - 使用希腊字母 ξ, η, ζ 表示
  - f(x) = F'(x)=d/dx F(x)，f(x)是概率密度函数
  - $P(\Omega) = \int^{(+\infty)}_{(-\infty)} f(x)\, dx = 1$

多维随机变量

二维随机变量，联合分布函数，边缘分布函数：

- 离散型二维随机变量
  - 联合分布律
  - 边缘分布律
- 连续型二维随机变量
  - 联合分布密度

- 边缘分布密度

数学期望和方差

随机变量的两个重要的数值特征：

- 数学期望：反映了随机变量的平均取值
- 方差：反映了随机变量的分散程度。
  - 标准差：方差的平方根

常用分布

离散型随机变量

- 二项分布
  - 只有两个可能结果的随机试验称为 Bernouli 试验
  - n 重 Bernouli 试验服从的分布是二项分布，记为 X~B(n,p)
  - 期望是 np，方差是 np(1-p)
- 泊松分布
  - 在某一固定大小的范围(或者时间段)内 ,某种特定类型事件的分布服从泊松分布，记为 X~P(λ)
  - 期望是 λ，方差是 λ。
  - 泊松分布与二项分布之间的关系由泊松定理给出。

连续型随机变量

- 正态分布，又称高斯分布。
  - 服从参数 μ 和 σ 的正态分布，记为 ξ~N(0,1)

- 期望是 μ，方差是 σ^2。

2.2 信息论基础

熵是不确定函数在统计力学中的应用。

熵是热力学下物质系统状态的一个函数，表示微观粒子之间无规则的排列程度，即表示系统的混乱程度。

信息熵

对于信息接收者，通信的目的是获取信息，即通过信息消除某种不确定性。因此，不确定的程度是信息量的一个量度。不确定性越大，则信息量也越大。

在信息处理领域的计量单位

- 对数的底为 2，计量单位为比特（bit）
- 对数的底为 10，计量单位为哈特利（hartley）
- 对数的底为 e，计量单位为奈特（nat）

汉字是当今世界上信息量最大的文字符号系统。

两个随机变量：

- 联合熵
- 条件熵
- 熵的链规则
- 互信息：知道一个随机变量的取值后对另一个随机变量的不确定性的减少量，或者一个随机变量包含另一个随机变量的信息量
    - 互信息是非负的、对称的量度

- 互信息可以用于衡量两个随机变量的依赖程度（或者独立性）
- 当两个随机变量相互独立时，它们的互信息为 0
- 互信息可以描述两个单词之间关联程度大小

- 相关熵：以称为 KL 距离，衡量在同一事件空间的两个概率分布的差异。
  - 相关熵是非负的，但不是对称的量度。

语言与熵

- 语言的熵
  - 平均每个字符或者词汇的熵称为熵率。
- 语言的交叉熵
  - 衡量一个语言统计模型的优劣的定量评价方法
  - 迷惑度（perplexity）衡量统计模型的质量，是与交叉熵等价的量度

语言模型：描述了任一语句属于某一种语言的可能性的大小。

声学模型：描述了一个语句对应于声学信号的可能性。

语言识别器的任务：找到声学信号对应的可能性最大的语言文本。

语言模型就是自然语言的概率模型。

2.3 粗糙集

粗糙集：处理模糊和不确定性知识的数学工具。主要以不可分辨关系为基础，通过概念的上近似、下近似和属性约简来解决带有模糊的或者不确定性的结构化数据的分类以及规则获取等问题。

粗糙集通过成员函数提供了对模糊概念的描述能力。

信息系统

论域内的数据是采用信息表或者信息系统的形式来存储的

信息系统可以使用四元组定义：

- U 是所有个体的非空有限集合
- A 是属性的非空有限集合
  - 条件属性子集 C
  - 决策属性子集 D
  - C∩D=∅, C∪D=A
  - 把区分了条件子集和决策子集的信息表称为决策表
- 对于任意一个属性 a∈A，有
  - 一个属性值的集合 V_a
  - 一个信息函数 f_a：U→V_a

不可分辨关系是一种等价关系。

集合近似：上近似和下近似

约简的方式：

- 基于不可分辨关系将信息系统进行划分后，可以用每个个体子集中的任意一个个体来代替该子集进行存储，因此实现了对原有信息系统的约简
- 基于属性的约简，只保留能够保证原有的不可分辨关系（以及集合近似）不被改变的属性，并且将其他属性从信息系统中移除。

一个信息系统的约简可能不止一个，信息系统的所有约简的交集称为信息系统的核，核中所有的属性都是信息系统保持原有概念表述能力不变时所必需的特征。

计算一个信息系统的约简是一个 NP-Hard 问题。可以使用基于辨识函数与分辨矩阵的约简算

法来解决这个问题。

属性依从关系：依从度。

决策规则合成：条件公式，决策规则的前趋项和后继项。决策规则的支持集。

第3章 汉语自动分词技术

词是自然语言中最小的有意义的构成单位。汉语文本是基于单字的文本，以汉字作为最小单位，词与词之间没有明显的界限标志，因此分词是汉语文本分析处理中的首要问题。

本章的重点：

- 汉语分词的规范
- 歧义的分类和识别
- 新词的识别
- 分词方法
  - 正向最大匹配分词方法
  - 反向最大匹配分词方法
  - 全切分词网络分词方法
- Zipf 规律

3.1 引言

分词：识别句子中隐含的词语边界，即为句子添加明显词语边界标志，使得所形成的词串反映句子的本意。

汉语分词系统面临的困难：

- 未登录词的识别。

- 常见的未登录词
  - 专有名词
  - 重叠词
  - 派生词
  - 与领域相关的术语

- 获取分词规则
  - 汉语分词语料
  - 分词算法面临的参数空间非常大
  - 没有好的无监督机器学习算法

- 词语边界歧义
  - 交叉歧义
    - 真歧义：存在两种或者两种以上的可实现的切分形式，需要依赖于所处的上下文环境才能正确处理真歧义字段。
    - 伪歧义：伪歧义字段切分的结果与上下文无关，仅依据字段内部的信息（例如：词频或者字间互信息）就可以正确地切分伪歧义字段。
  - 组合歧义

- 分词的实时性

3.2 分词规范

表音文字中的词是由历史确定的，不存在分词规范问题。

汉字是表意文字，以汉字为单位，缺少严格意义的形态变化，没有明显的形态界限可以作为分词标志，存在特有的分词问题。

《信息处理用现代汉语分词规范及自动分词方法》（刘源，1994）中描述的分词标准：

- 结合紧密，使用稳定的二字或者三字看作分词单位
- 四字成语看作分词单位
- 五字或者五字以上的谚语、格言等，分词后不违背原有组合的意义则应该切分
- 结合紧密，使用稳定的词组看作分词单位
- 惯用语和有转义的词或者词组，在转义的语言环境下，看作分词单位
- 略语看作分词单位
- 分词单位加形成儿化音的“儿”，还是看作分词单位
- 阿拉伯数字等，仍保留原有形式，看作分词单位
- 现代汉语中其他语言的汉字音译外来词，看作分词单位
- 不同的语言环境中的同形异构现象，按照具体语言环境的语义进行切分

动词分词规范：

- 动词前的否定副词需要切分
- 用肯定加否定的形式表示疑问的动词词组需要切分，不完整的看作分词单位
- 动宾结构的词看作分词单位
- 结合不紧密的或者有众多与之相同结构的词组的动宾词组需要切分
- 动宾结构的词或者词组，如果中间插入其他成分，那么需要切分
- 动补结构的二字词或者结合紧密、使用稳定的二字动补词组，看作分词单位
- “1+1”或者“1+2”或者“2+1”或者“2+2”结构的动补词组需要切分
- 偏正结构的词以及结合紧密的词看作分词单位
- 复合趋向动词看作分词单位
- 动词与趋向动词结合的词组需要切分

- 多字动词无连词并列需要切分

3.3 常用的分词方法

最大匹配分词算法

正向最大匹配分词（Forward Maximum Matching，FMM）

- 根据自动分词词典中的最长词条所含汉字个数 L，取处理材料中当前位置起始 L 个汉字作为查找字符串，搜索分词词典是否存在匹配词条。如果成功匹配就完成搜索；如果失败，就取消查找字符串中最后一个汉字，然后继续搜索，直到搜索成功则完成这轮匹配任务。然后重复上面的步骤，直到切分出所有的词为止。
- 错误切分率较高。
- 不能处理交叉歧义和组合歧义。

反向最大匹配分词（Backward maximum Matching，BMM）：分词过程与 FMM 方法相同，只是从句子（或者文章）的末尾开始处理。

- 错误切分率比 FMM 低。
- 可以处理交叉歧义，不能处理组合歧义

基于统计的词网格分词

词网格分词的步骤：

- 选择网格构造。
  - 利用词典匹配，列举输入句子中所有可能的切分词语，并以词网格的形式保存
    - 词网格是一个有向无环图（DAG），蕴含了输入句子中所有可能的切分，其中的每条路径代表一种切分。

- 选择计算词网格中的每条路径的权值。
  - 权值通过计算图中每个节点(词)的一元统计概率和节点之间的二元统计概率的相关信息获得
  - 根据图搜索算法找出图中权值最优的路径，路径对应的就是最优的分词结果。

3.4 歧义的分类与识别

歧义的分类：

- 交叉歧义
- 组合歧义

从分词的结果上看，歧义切分字段的分类：

- 具有确定分法的歧义切分字段：在需要分词的短句中就可以实现正确切分
- 具有不确定分法的歧义切分字段:在需要分词的短句中无法实现正确切分，需要增加对上下文语义信息的处理，即增加语义理解的处理。
  - 这类分词的歧义问题可以放在短语理解、句子理解和篇章理解阶段去处理，并且这类歧义所占比重很小，可以不在分词阶段解决。

歧义的识别：

- 对交叉歧义的抽取
  - 交叉歧义抽取算法和处理策略
- 对组合歧义的抽取
  - 组合歧义抽取算法和处理策略：采用 WSD 的向量空间法

3.5 新词的识别

新词，又叫未登录词，是指系统词典中没有收录的词。

Yao Yuan, Statistics Based Approaches towards Chinese Language Processing, Ph.D. thesis, 1997.

提出了基于构词力的新词识别方法：

- 基于 head-middle-tail 结构的汉字构词模式
  - 词首（head）：多字词的首字
  - 词尾（tail）：多字词的尾字
  - 词中（middle）：多字词的中部
  - 多字词的构词模式概率
- 基于构词能力的未登录词识别算法的核心是对分词算法输出中的某一单字词串重新进行词语边界划分，从中找到满足代价函数的新的词串。
  - 新词识别是词语边界识别的后处理过程，应用简单叠加原理，将下列特征信息采用线性插值方法整合在一起
    - 汉字构词能力（Word Formation Power，WFP）
      - 功能字（function character）：的，了等字，出现频度高，常常单独存在，与其他字组合在一起形成词的构词能力较弱。
      - 汉字构成词语的概率表示
    - 字结合点
    - N-gram 模型

3.6 关于分词的若干统计结果

- 最简单的最大匹配方法的分词精度的下界在 8%

- 交叉歧义是影响汉语分词精度的重要因素
- 组合歧义的消解虽然困难，但是其出现的概率很小
- 未登录词是影响汉语词语边界划分准确率的重要因素
    - 人名和地名是未登录词的重要来源

3.7 语言单位的统计分布规律（Zipf 定律）

Zipf 定律，实质是一个多项式衰减函数。说明了单纯依赖增大语料库是无法解决数据稀疏性问题的。

《现代汉语计算语言模型中语言单位的频度——词序关系》（关毅，王晓龙，张凯，中文信息学报，1993，13（2）：8~15）介绍了现代汉语的字和词二元对等层次结构上同样存在 Zipf 形式的词频——词序关系。因为人们使用文字时遵循两个原则：

- 关联原则：使用先前使用过的词汇
- 模仿原则：模仿和借鉴自己或者其他人已经存在的作品

第4章 基于数学统计的语言模型

语言模型（Language Model）是描述自然语言内在规律的数学模型，构造语言模型是计算语言学的核心。

语言模型的分类：

- 基于语法的语言模型：人工编制语言学语法规则，语法规则来源于语言学家提供的语言学知识，不太适合处理大规模真实文本。
- 基于统计的语言模型：也叫概率模型，借助于模型的概率参数，估计出自然语言中每个句子出现的可能性。

- N 元语法模型
- 隐马尔可夫模型
- 最大熵模型

4.1 统计语言模型概述

统计语言模型以概率分布的形式描述了任意语句（字符串）属于某种语言集合的可能性。

概率的估算采用最大似然度估计。

训练数据：用于估算基于统计的计算语言模型中的概率分布的训练语料库文本。

训练：根据训练数据估算概率分布的过程。

数据平滑技术可以解决数据稀疏性问题。

4.2 现有的主要统计语言模型

根据上下文空间不同的划分方法，将现有的语言模型分为 5 个大类。

上下文无关模型

只考虑当前词本身的概率，即训练文本中词出现的频度估算词的概率，不考虑该词所对应的上下文环境。

优点：模型简单，训练数据少，易于实现。

缺点：没有考虑上下文信息，统计信息不充分，实用性不强。

N 元语法模型

依赖于上下文环境的词频统计概率。主流使用：三元语法模型（Trigram）。

优点：包含了前 N-1 个词所能提供的全部信息，对于当前词的出现具有很强的约束力。

缺点：需要更大规模的训练文本来确定模型的参数。

N 元词性模型

依赖于上下文环境的词性统计概率。

将词按照其语法功能进行分类，由这些词类决定下一个词出现的概率。这样的词称为词性，而相应的语言模型称为 N 元词性模型。

模型要求词类各不相交，但是现实中存在着一词多类的问题。

优点：需要的训练数据比 N 元语法模型少，模型的参数空间也水上。

缺点：词的概率分布依赖于词性而并非词本身，得到的概率分布相对不够精细。

基于决策树的语言模型

统计决策树包括所有的概率分布以及根据当前上下文查询其分布的机制。

统计决策树中包括两种类型的结点：

- 中间节点：包括关于上下文的一个提问。
- 叶子节点：包括惟一的概率分布。

对于当前词的上下文 ,查询从根节点开始 ,由对根节点提问的不同回答进入到不同的子节点，直到到叶子节点。从而得到当前词上下文的分布信息。

前述所有的基于统计的语言模型（上下文无关模型、N 元语法模型、N 元词性模型）都可以使用统计决策树的形式表示出来。因此，统计决策树模型是一种更加通用的语言模型。

优点：分布树不是预先固定好的，而是根据在训练语料库中的实际情况确定的

缺点：构造统计决策树的时空消耗非常大

动态、自适应、基于缓存的语言模型

静态语言模型：概率分布是预先从训练语料库中估算好的，不会在应用过程中发生改变

动态语言模型：也叫自适应模型，或者基于缓存的语言模型。能够根据词在局部文本中的出现情况，动态地调整语言模型中的概率分布数据。

例如：基于缓存的一元语法模型可以利用窗口将那些与当前词相关的词框在一起，如果当前词在窗口中出现不止一次就可以加强这个词的概率，从而动态调整模型中的概率分布。

这种混合模型可以有效地避免数据稀疏性问题，同时可以提高原静态模型的表现能力。

4.3 数据平滑技术

解决数据稀疏问题（即零概率问题）。

数据平滑算法的评价标准

数据平滑的度量方法：

- 测试文本上的交叉熵：语言模型的熵值越小，平滑算法的性能越好
- 测试文本上的困惑度（perplexity），也叫迷惑度。语言模型的困惑度越小，平滑算法的性能越好
- 测试文本的生成概率：语言模型的生成概率越大，说明模型和测试集的符合程度越高，对应的数据平滑算法的性能越好。

三种数据平滑的度量方法是相互等价的。

常见的平滑算法

- 加法平滑：将 N 元语法模型中每个 N 元对的出现次数加上一个常数

- Good-Turing 平滑：基于绝对折扣平滑方法提出的，是许多数据平滑技术的基础，将 N 元语法模型中出现 r 次的 N 元对的出现次数依据公式进行调整
- 线性插值平滑：也称为 Jelinek-Mercer 平滑。利用低元的 N 元语法模型对高元的 N 元语法模型进行线性插值。
- 回退式平滑 :将 N 元语法模型中出现次数低于阈值的 N 元对使用 Good-Turing 估计对其进行平滑，将其部分概率折扣给未曾出现的 N 元对。将未曾出现的 N 元对的概率回退到低元的 N 元模型，按照比例来分配折扣得到的概率。
  - 相比线性插值算法，参数较少，计算较快，不需要通过某种迭代重估算法训练，实现也更加方便。
- Kneser-Ney 平滑：基于绝对折扣平滑方法提出的。一元概率应该与其不同的前向邻接词的数量成正比，而不是与其频度成正比。
- Witten-Bell 平滑 :是线性插值平滑算法的特例。提出了设置线性插值中插值参数的方法。

4.4 隐马尔可夫模型

随机过程

在同一样本空间下的一个随机变量序列 ,这些变量可能出现的结果称为这个随机过程的状态。

- 离散随机过程的例子：不断掷出的骰子所形成的序列
- 连续随机过程的例子：在不同时间点某个商场内的人数

马尔可夫链与马尔可夫性

马尔可夫链是一种特殊的离散随机过程 ,这种随机过程的下一时刻的状态完全由当前时刻的状态决定，而与历史的状态无关，即“未来”与“过去”无关，只与“现在”有关，具有以

上性质的随机过程为马尔可夫过程，具有马尔可夫性。

如果随机过程的马尔可夫性与时间无关，则称这个随机过程是齐次马尔可夫过程。

概率转移矩阵是描述马尔可夫链的重要工具。

马尔可夫模型

马尔可夫模型是一个双重随机过程

- 一重随机过程是描述基本的状态转移
- 一重随机过程是描述状态与观察值之间的对应关系

马尔可夫模型的数学描述：M={Ω, Σ, P, A, Θ}

- Ω 是所有状态的集合
- Σ 是所有观察序列的集合
- P 是转移概率矩阵
- A 是发射概率矩阵
- Θ 是初始概率向量，π 是状态的初始概率

隐马尔可夫模型

马尔可夫模型与隐马尔可夫模型的数学模型是相同的，当马尔可夫模型中的“状态”不可见时，就转换成了隐马尔可夫模型，两种模型面对的现实问题的情况不同、对模型的限定条件不同以及需要解决的问题不同。

隐马尔可夫模型的 3 个基本问题：

- 对于给定的观察序列，根据现有的隐马尔可夫模型，序列出现的概率是多大
- 对于给定的观察序列，在状态序列未知的情况下，根据现有的隐马尔可夫模型，求得最

有可能的隐含状态序列是什么

- 对于给定的观察序列，通过学习可以得到最有可能的隐马尔可夫模型的参数是什么，这个参数使得观察序列的概率最大，即最好地“解释”了这个观察序列。这个问题就是隐马尔可夫模型的训练问题。

隐马尔可夫模型的 5 种算法：

- 前向算法
- 后向算法
- 前向——后向算法
- Viterbi 算法
- Baumm-Welch 算法

这 5 种算法都是动态规划算法。

4.5 最大熵模型

模型介绍

基本思想：在满足系统当前提供的所有条件下寻找分布最均匀的模型，即熵最大的模型。

基本算法：迭代规整算法（Generalized Iterative Scaling，GIS）

模型评价

优点：

- 简单并且直观，只是通过添加主要的约束来融合多种信息，没有其他的前提假设条件
- 通用性强，对任何事件空间的任何子集的概率估计都可以使用最大熵原理
- 任何现在的语言模型的知识都可以添加到模型中去

- GIS 算法是一个逐步适应的过程，新的约束条件可以被随时添加到模型当中

缺点：

- 计算复杂度太高
- 迭代算法一定能够收敛，得到模型的解，但是无法确定算法迭代次数的理论上限

模型建立

- 与 N 元语法模型的融合
- 与触发器的融合

第5章 基于语言理解的处理方法

5.1 引言

为了使计算机能够真正地理解语言，必须使用某种语言模型来描述自然语言的规律。

常用的语言模型：

- 以基于知识的方法为代表的理性主义方法，这个方法以语言学理论为基础，强调语言学家对语言现象的认识，采用非歧义的规则形式描述或者解释歧义行为或者歧义特性。
- 以基于语料库的统计分析为基础的经验语义方法，这个方法以数学为基础，从能够代表自然语言规律的大规模真实文本中发现知识，抽取语言现象或者统计规律。

5.2 常用的基于语言理解的分类标注体系

词性分类体系

划分词类的依据有：

- 形态标准：中文没有形态变化。
- 意义标准
- 分布标准：中文只能根据词在句法结构里所担当的语法功能，即分布进行分类。

标准集的确定原则：

- 标准性 :是指尽量采纳当前已经成为各种语言的词性标准或者正在成为词性标准的分类体系和标记符号
- 兼容性：是指尽量保证标注集的表示与已经存在的标注集的表示能够相互转化
- 扩展性 :是指对未解决的遗留问题或者是未来可能的技术发展方向提供扩充和修改的能力，并且使得扩充和修改对系统的整体影响代价最小。

北京大学计算语言学研究所的《现代汉语语法信息词典》提供了 39 个词类汉语词性标注集。

词义分类体系

汉语缺乏语法形态，因此词义知识非常重要。

词义：在一定的语言环境中所阐明的内容。

描述一种词语所表述的意义的常见方式：

- 同义词分类的方法：出现的最早。
  - 一个词的词义完全可以利用属于同一个集合中的其他词的词义来表示
- 即基于语义成分(义素分析法)的词汇语义学 :把一个词的意义分解为概念原子的组合。
  - 定义一套概念原子的难度很大。
  - 董振东先生的《知网》对 6000 个汉字抽取了 1000 多个义原
    - 义原 : 是知网中最基本的，意义不能再分割的最小单位，是解释知识词典的基本要素，其他的词条全部通过这些义原来定义。

- 基于关系的词汇语义学：基于“网”的形式描述词语意义，当前的发展主流。
  - 词汇所表示的概念相互之间存在着联系，彼此构成一个知识网络，因此词义词典在整体上就是一个词义网络或者是知识网络。
  - WordNet，在线义类词典，基础是同义词集合。名词和动词都是分层级组织词语之间的语义关系
    - 在名词中，有上下位关系（hyper-hyponymy）
    - 在名词中，有整体部分关系（meronymy）
    - 在动词中，有下位关系（troponymy）
    - 在动词中，有继承关系（entailmment）
  - 在“知网”中也定义了相应的关系，但是这种关系只定义在义原中，没有定义词与词的关系，这些关系通过 KDML 来描述

5.3 常用的基于语言理解的语法理论

语法，也叫文法，或者句法。是用来精确并且无歧义地描述语言构成方式的。语法描述语言的时候不考虑语言的含义。

形式化是指用一定的范畴和作用在这些范畴上的规则来描述它们。

使用有限的范畴和规则来描述和生成无限的语言，并且这些范畴和规则都是可以形式化的，即可以通过数理符号来表示和推理。

基于语言知识认知的功能主义语法：

- TG（Chomsky 转换生成语法）
- DG（特尼埃尔的依存语法，也叫配价语法）
- CaseG（菲尔格的格语法）

- SFG（韩礼德的系统功能语法）
- Langacker 的认知语法

基于语言知识表达的形式语义语法：

- PSG（短语结构语法）
- ATN（扩充转移网络语法）
- GB（支配约束理论）
- FUG（功能合一语法）
- LFG（词汇功能语法）
- HPSG（中心词驱动的短语结构语法）
- GPSG（广义短语结构语法）
- CG（范畴语法）
- LG（链接语法）
- TAG（树邻接语法）

从语法所基于的基础来分类：

- 基于范畴
  - TG（Chomsky 转换生成语法）
  - PSG（短语结构语法）
  - TAG（树邻接语法）
  - 基于词的合一运算
    - FUG（功能合一语法）
    - HPSG（中心词驱动的短语结构语法）
    - GPSG（广义短语结构语法）

- 基于词
  - DG（依存语法，也叫配价语法）
  - LG（链接语法）
  - CG（范畴语法）

常用的语法理论

1. 短语结构语法（Phrase Structure Grammar，PSG）：第一次提出的关于语言和语法的数学模型

   PSG 可以表示成（T，V，P，S）的四元组：

   - T 是终结符号集，是基本符号，不需要做出进一步的定义
   - V 是非终结符号集，是需要定义的语法范畴，专门用于描述语法，不能出现在最终生成的句子中
   - S∈V 为语法的开始符号
   - P 为产生式规则集，规则集中的所有规则都是非空的有限集

   用 PSG 处理自然语言时，它区分歧义的能力不足，为了对短语结构的形式体系增加某些约束，Chomsky 提出了语法分类体系：

   （1） 无约束短语结构语法（0 型语法）

   P 中每条产生式规则集的形式为：α→β，并且 α∈(TYV)+，β∈(TYU)*

   0 型语法是生成能力最强的一种形式，但是仍然不足以描述自然语言，然后其对程序设计语言的描述又过于一般化，需要一定的约束。

   （2） 上下文有关语法（1 型语法）

   P 中每条产生式规则集的形式为：aAb→aβb，并且 a,b∈(TYU)*，A∈V，β∈(TYU)+

当 A 的前面符号是 a，后面符号是 b 时，可以将 A 重写为 β，因此重写规则依赖于上下文。

（3） 上下文无关语法（CFG，2 型语法）

P 中每条产生式规则集的形式为：A→β，并且 A∈V，β∈(TYU)+，每条产生式规则的左侧必须是一个单独的非终结符，规则应用时不依赖于 A 所在的上下文

（4） 正则语法（3 型语法），可以用有限状态转移图（FST）来表示，所以也叫有限状态语法。所表述的语言可以使用有限自动机（FSA）来识别，一般用于描述程序设计语言的单词结构，不适宜于描述自然语言。

- 右线性语法：P 中每条产生式规则集的形式为：A→aB，A→a，并且 A，B∈V，a∈T
- 左线性语法：P 中每条产生式规则集的形式为：A→Ba，A→a，并且 A，B∈V，a∈T

语法的型号越高，对规则附加的限制就越多，语法的生成能力就越弱。

四种语法之间的关系：0 型语法⊃1 型语法⊃2 型语法⊃3 型语法

短语结构语法的优点：非常方便地对一个句子进行结构的描述

短语结构语法的缺点：因为没有引入语义的成分，在生成语法正确句子的同时，可能是语义错误的句子。

短语结构语法的扩展：

（1） GPSG（广义短语结构语法）

a) 只有一个句法对象，即短语结构

b) 只有一个句法描写平面，即表层结构

c) 语法和语义并重，在建立语法的同时，试图提示出语法与语义之间的相互关系，并

且把这种关系当作一种语法理论的中心目标

d) 更加致力于语法普遍性的探索

e) 是一种高度严格的形式化方法，总体结构图：

词汇直接支配规则→元规则→扩展后的直接支配规则的集合→投射函数→树

非词汇直接支配规则→→→→→→↑ ↑

特征共现限制，特征标示默认值，头特征规约，足特征规约、控制一致原则，线性顺序说明

（2） FUG（功能合一语法）：功能描述和合一运算相结合的方法。

a) 复杂特征集就是对事物从多方面进行描述

b) 复杂特征集在功能合一语法中叫做功能描述

c) 一个功能描述是由一组描述元组成

d) 描述元是一些带值的属性，叫做“属性——值”对。

e) 合一运算是对复杂特征集进行运算的方法 ,即把若干个功能合并成一个单独的功能描述。合一运算不同于求并运算，如果两个功能描述有相同的属性，不相等的值，那么这两个功能是不相容的，合一运算失败。

f) FUG 把复杂特征集全面地、系统地应用到语言描写中，在词条定义、语法规则、语义规则和句子描述中都能够使用复杂特征集。

（3） LFG（词汇功能语法）

a) 与转换生成语法相同的地方，都有两个语法层次

i. 成分结构

ii. 功能结构

b) 与转换生成语法不同的地方，在实现方法上：

i. 在转换生成语法中，句子的深层结构和表层结构都表示为短语结构语法，主题角色（也叫论旨角色，或者题旨角色）和表层结构之间的转换是建立在深层的短语结构语法之上的。

ii. 在词汇功能语法中，以无序的语法功能作为语法理论的基础，在语法——语义的分析过程中，所依据的语言知识主要存储在机器词典的词汇项中，整个过程是词汇驱动的。

c) LFG 采用复杂特征集作为信息表示的基本手段

d) LFG 采用合一运算作为语法——语义分析的基本算法，由成分结构和功能结构的映射均使用合一运算完成。

e) LFG 的分析步骤：

i. 由语法规则和词法规则生成句子的成分结构

ii. 由成分结构求出功能描述

iii. 由功能描述式构造功能结构

f) 功能合格条件：

i. 惟一性：在任何功能结构中，一个属性至多只能有一个值。惟一性检查体现在合一运算中。合一成功，则满足惟一性；合一失败，则不满足惟一性

ii. 完备性：当且仅当一个功能结构它的谓词应该管辖的所有语法功能时，这个功能结构是局部完备的；当且仅当一个功能结构内所有的子功能结构都是局部完备是这个功能描述是完备的。

iii. 一致性：当且仅当一个功能结构所包含的可被管辖的语法功能都被一个局部谓词所管辖时，这个功能结构是局部一致的；当且仅当一个功能结构内的所有子功能结构都是局部一致的，那么这个功能结构是一致的。

（4） HPSG（中心语驱动短语结构语法）：

a) 继承了广义短语结构语法的原则，是一种基于约束的语法理论。

b) 同时吸收了词汇功能语法和范畴语法的优点，强调了词汇在语言构成中的重要地位，强记中心语在语法分析中的作用，使得整个语法系统由中心语驱动。

c) 把符号看作含有音系、语法、语义、话语以及短语结构信息的结构性复合体，并且采用分类特征结构来描述

d) HPSG 的特性

i. 特征结构分类性：即每个特征结构都有一个分类，利用这个分类来表明这个特征结构所描写的语言客体的类型

ii. 特征合适性：即一个特征结构里面出现什么梳妆打扮特征取决于这个特征结构的分类，所以特征必须适合特征结构的分类

iii. 特征完整性：适合一个特征结构的所有特征都必须出现在这个特征结构中

iv. 分类空心性：所有的特征的值都必须是最具体的值分类。

e) HPSG 中的规则和语法的普遍性原则都是对特征结构的制约，这些制约包括：

i. 词汇规定

ii. 语法普遍原则

iii. 直接支配格式

2. 基于语言知识认知的语法：侧重于从语言事实中发现范畴，建立规则。

“价”：反映动词对名词性成分的支配能力

- DG 把“价”放在非常核心的位置
- CaseG 和 SFG 把“价”放在比较重要的位置
- TG（转换生成语法）等形式语法则把“价”当作一种记载在词库的动词词条之下

的词汇、语法特征，认为一个动词所必有的论元构成了动词的论元结构，基础句式（原子句）是动词的论元结构的一个投影。

- 下面 4 种语法是从不同的角度来认识语言的本质
  - 基于词汇来认识语言，没有过多地考虑语言的结构，而是首先考虑了语义，都是从语言的内部来表示语言的
    - DG：用词与词之间的关系来描述的
    - CaseG：用词本身的格来描述的
  - 从语言的外部来认识语言，有太多哲学理论，不太适合直接在计算机上使用
    - SFG：从语言的外部功能来认识语言
    - CG：从人的认知能力来认识语言

a) DG（依存语法，也叫配价语法）：是表述词与词之间的一种最基本的联系。

i. 依存关系的 4 大公理：

1. 一个句子中只有一个成分是独立的，不受任何成分支配的
2. 其他成分直接依赖依存于另外的某一个成分
3. 任何一个成分都不能直接依存于两个或者两个以上的其他成分
4. 如果 A 成分直接依存于 B 成分，而 C 成分在句中位于 A 成分和 B 成分之间，则 C 成分或者直接依存于 A 成分，或者直接依存于 B 成分，或者直接依存于 A 成分和 B 成分之间的某一成分。

ii. 依存语法的主要特点：表示方法简洁、易懂。

b) CaseG（格语法）

i. “格”不是表层的语法格，而是深层的语义格，是一切语言的普遍现象。

ii. 英语的 6 个格：

1. 施事
2. 工具
3. 与格
4. 使成
5. 处所
6. 客体

iii. 格语法今生语义，而忽略了词、短语之间的语法关系。

iv. 鲁川、林杏光根据汉语的特点把格语法改成了格关系，并且提出了格关系的层次性，编写出了《动词大辞典》。

c) SFG（系统功能语法）：韩礼德提出的 SFG 把语言看作一个与人类社会密切相关的职能体系，对这个体系的描述和解释都依赖于这个体系之外的语言的社会职能

i. SFG 的主要特点是外部性，而不是从语言体系本身去解释语言。

ii. SFG 的组成部分：

1. 系统语法：说明语言系统表现为选择关系
2. 功能语法：阐明语言的 3 大功能：
   a) 概念功能
   b) 人际功能
   c) 语篇功能

iii. SFG 的 6 个核心思想：

1. 纯理功能的思想
2. 系统的思想
3. 层次的思想

4. 功能的思想

5. 语境的思想

6. 近似的思想

iv. 语言本身是个意义系统，语言的功能首先体现在语义功能上。

d) CF（Cognitive Grammar，认识语法）：以所示概括性为首要目标，力图找出一些基本的认知原则，对语言不同层次、不同方面存在的并行现象做出统一的解释，并且收到化繁为简的效果

i. CF 不是以数理符号的形式表示的语言，而是一种具有描述机制的语言，也是一种具有生成机制的语言，但是真实的语言却不是靠这些形式化的描述形成的。

ii. CF 从认知的的角度研究语法的理论，因为语法不是生成的，而是与人的认知能力有密切联系的。

3. 其他语法（与短语结构语法无关的语法）

a) ATN（扩充转移网络语法）：成功应用于有限领域的问答系统中。

i. 有限状态转移图（FSTD）由若干有限的状态，以及从一个状态转移到另一个状态的弧组成，但是只能识别正则语言

ii. 递归转移网络（RTN）是对 FSTD 赋予了一种递归机制，具有了识别上下文无关语言的能力

iii. RTN 的局限性：

1. 只是一个识别器，而不是一个分析器，即只能指出输入句子是否合法，而不能进一步产生对句子的结构分析

2. 不能充分提示句子成分之间的某种依赖性

iv. ATN 对 RTN 的扩充

1. 寄存器组。用来存放分析过程中产生的关于句子或者成分的结构信息
2. 测试。弧上除了用词类、语法结构等作为标记之外，还允许附加任意的测试，只能弧上测试满足条件后才能通过
3. 动作。弧上还可以附加某些动作，当弧被通过时，动作便被执行

v. ATN 的特点：

1. 具有比较强的生成能力（相当于 0 型语法），
2. ATN 是过程性的，不是描述性的
3. ATN 对语法的过分依赖，限制了它处理某些合乎语义但是不完全合乎语法的话语的能力

b) TAG（树邻接语法）

i. 处于上下文无关语法和上下文有关语法之间的一种语法表示形式

1. 上下文无关语法过于简单，无法限制一些不合法的语言现象
2. 上下文有关语法的分析算法过于复杂，不适合实际应用

ii. TAG 的句子结构是用树形式来表示，基本操作是剪插。

iii. TAG 的形式化体系包括两种原子树

1. 初始树：根节点都是 S，而叶节点都是终结符或带有替换标志的非终结符。
2. 辅助树：包括一个特殊的叶节点，这个特殊的叶节点与其根节点具有同样的符号，这个叶节点称为脚节点。

iv. TAG 通过附加上词汇项的语法结构集合引入词汇信息，并且使用一个或者一组用于组合结构的操作。操作通过在初始树和附加树中添加词汇锚点，并且引入合一运算来实现。

c) LG（链接语法）

i. 一部链接语法就是一个单词的集合，其中每个单词后面记录着各自的链接要求

ii. 链接要求通过一系列链接子表达式指定

iii. 单词之间的链接要求的元规则：

1. 平面性：链之间互相不交叉

2. 连通性：所有的链会把所有的单词联系在一起

3. 顺序性：公式中较左边的链接子必须与距离单词比较近的单词连接；公式中较右边的链接子必须与距离单词比较远的单词连接

iv. 基于动态规划的分析算法（识别算法）：以一种自顶向下的方式建立句子的链接集，先建立较大的链，再在这些链的范围内建立范围比较小的链

v. LG 主要特点：链接语法是词汇语义的语法系统，对于任何一个单词，词典中都详尽地描述了这个单词在句子中的使用方式。

所有词汇语义的语法系统的优点是：构造大型语法系统比较容易，一个单词的描述只对包含这个单词的句子的分析产生影响，允许逐步建立一个大型语法系统，方便描述特殊的不规则的词语，这些词语描述在词典中单独占据一个条目。词汇语义铁语法系统易于构造统计模型，单词间的关系描述方便收集词汇之间的统计信息。

语言中的任何现象都可以分为典型的和非典型的：

- 典型规律的规则描述很容易
- 非典型规则的规则描述则比较困难

浅层语法分析技术

浅层语法分析，也叫组块分析，或者叫局部语法分析。

自然语言处理层次：词、短语、句子。缺少短语处理层次，会出现大量歧义问题，因此将组

块分析和语法分析分开处理，可以通过提高组块分析的正确率，从而为语法分析打下基础。

浅层语法分析技术分类：

- 基于统计的
  - 基于 HMM 的方法
  - 基于互信息的方法
  - 基于χ^2 检验的方法
  - 基于中心词依存概率的方法
- 基于规则的 :根据人工书写或者半自动获取的语法规则标注出短语的边界和短语的类型
  - 根据标注策略的方法分类：规则的使用比较简单，规则的获取比较困难，规则多是人工书写
    - 增加语法标记法
    - 删除语法标记法
  - 从语料库中自动获取语法规则的方法
    - 基于转换的错误驱动的学习方法
      - 初始标注。把训练语料中所有的基本名词短语的标记去掉，用一个简单的初始标注程序标注出训练集中可能的基本名词短语，并把这个结果作为系统的底线
      - 形成候选规则集。在每个初始标注错误的地方，规则模板用来生成候选规则，规则的条件就是词的上下文环境，动作就是改正错误标记所要做的动作。
      - 获取规则：把候选规则集中的每条规则分别运用于初始标注的结果，选出得分最高的规则。把这条规则运用于初始标注的结果作为下一轮循环的基

础，并把这条规则作为规则序列中的第一条规则输出。重复以上过程，直到所有规则的得分都低于某个阈值。

- 基于实例的方法：把标注好的短语信息的语料库分成两个部分。一部分用于训练，一部分用于剪枝。
  - 从训练的语料中得到一组名词短语的组成模式规则
  - 把得到的这些规则应用到剪枝的语料中，并对这些规则打分
  - 根据每条规则的总得分情况删除那些得分少的规则

5.4 语料库多级加工

语料库（corpus）是按照一定的原则组织在一起的真实的自然语言数据集合，主要用于研究自然语言的规律，特别是统计语言学模型的训练以及相关系统的评价和测试。

国外常用语料库：

- Brown 语料库
- London-Oslo-Bergen 语料库
- Penn 树库

汉语常用语料库：

- 清华大学的原始语料库
- 中国台湾中央研究院的汉语平衡语料库

研究汉语树库的组织：

- 中国台湾中央研究院的中文信息处理组
- 马里兰大学 CLIP 实验室的汉字森林
- 美国宾夕法尼亚大学的 XTAG 计划

语料库的多级加工

语料库的标注或者加工就是对电子语料进行不同层次的语言学分析 ,并且添加相应的“显性”解释性语言学信息的过程。

生语料（raw corpora）：也叫原始语料，是未经加工的语料，对语料库进行一层层标注后，得到熟语料，即语言学知识“显性”化，方可使用。

生语料库：也叫原始语料库，没有任何标注的语料库

语料库的加工内容：

- 词性标注
- 语法标注
- 语义标注
- 言语标注
- 语用标注
- 分词（汉语语料库）

语料库的加工方式：

- 人工：非常昂贵，需要大量的人力资源
- 自动：容易给语料库标注带来一些错误
- 人机结合的方式
    - 由计算机自动选择语料库中自动处理后需要人工干预的标注
    - 由计算机先对语料加工，然后人工校对

歧义消解技术是语料库自动处理的基础：

- 语料库的多级加工是一个面向真实文本的自然语言多级歧义消解过程
- 经过加工的语料库也为歧义消解提供了资源支持

分词

汉语自动分词就是把没有分割标记的汉语字符串转换到符合语言实际的词串。

汉语自动分词的两个困难：

- 歧义消解
- 未登录词的识别问题

词性标注

词性标注就是根据一个词在某个特定句子中的上下文，为这个词标注正确的词性。其实质是研究词语所表现的语法功能的聚合关系，它要解决的主要问题是词性歧义（词性兼类）和未登录词词性的确定问题。

词性歧义，也叫词性兼类，即词语中的词性多于一个的歧义现象。

汉语词性兼类现象有几十种，其分布特征：

- 兼类词的数量不多
- 兼类词的实际使用频率很高，即越是常用的词，其词性兼类现象越严重
- 兼类现象分布不均匀。动名兼类和形副兼类占了三分之二。

词性标注方法主要有以下 3 种：

- 基于规则的方法
  - TAGGIT 采用上下文框架规则
  - Brill 提出基于转换规则
  - Voutilainen 采用约束语法，这个准确率达到 99.3%
- 基于统计的方法
  - 基于频度的方法

- 基于 N 元模型的方法
- 基于 HMM 的方法，结合 Viterbi 算法最常见
- 基于最大熵模型，融合不同阶的 N-gram 信息、长距离 N-gram 和其他有关词法的统计信息

- 混合方法：统计与规则相结合的方法
  - Lancaster 大学的 CLAWS 系统

汉语词性标注的困难：

- 没有形态信息
- 词序相对自由
- 词性与语法成分之间没有明确的对应关系

词性标注的 HMM 模型

隐马尔可夫模型引入了独立性假设：

- 词性标记的出现只依赖于有限的前 N-1 个词性标注，即 N-POS 模型
- 一个词语的出现不依赖于其前面的任何词语，只依赖于前面的词性标记，并进一步假设词语的出现只依赖于词性标记

经过上述假设，可以得到一个 N-1 阶词性标注的隐马尔可夫模型

Viterbi 词性标注算法

主要任务：为给定的观察值序列（即输入词串），找到一个最佳的状态序列（词性标记串）。

语法分析

语法分析研究的问题：

- 句子中的组合关系：语言单位组合的方式
- 句子中的聚合关系：语言单位间的相似关系或者相似度

语法分析的层次：

- 词法结构的自然语言分析：词在形态上（词形）的组合关系以及各个词在语法功能和意义上的聚合关系（即词法分析）
- 语法结构的自然语言分析：各个句子成分在结构上的组合关系和在语法功能上的聚合关系（即语法标注）

因此，语法分析是一个自益问题（bootstrapping problem），也叫自举问题，自助问题。

常用的语法分析方法：

- 上下文无关语法分析
- 依存语法分析

常用的语法分析算法：

- 从上到下的语法分析算法：包含一个可能状态的线性表，表中的第一个元素是当前状态，其他状态叫做备份状态，每一个状态都是一个“（符号列表，词位置）”对。
  - 分析过程是一个搜索问题
    - 深度搜索：状态线性表是一个堆栈，按照先入后出原则
    - 宽度搜索：状态线性表是一个队列，按照先入先出原则
- 从下到上的语法分析算法：对句子从左到右进行扫描，将扫描得到的一系列词性范畴与语法规则右边进行匹配。如果想同时进行规约，则可以规约成语法规则的左边。
  - 使用线图的数据结构保存局部的分析结构，避免了相同的结构被反复规约
  - 线图是算法的关键数据结构
    - 保存了分析过程中推导出来的成分

- 保存了部分匹配的规则，这个部分匹配的规则叫做活动弧

- 用来保存新的已经规约成分的线性表叫做代理表

概率上下文无关语法（PCFG）

语法分析中使用概率的方法：

- 利用概率加速语法分析
- 利用概率在语法分析的不同结果中选择正确的结果
- 利用概率进行句子确定

PCFG 既可以定量描述语法歧义的能力，而且可以描述不符合语法的句子。

语料库的应用

语料库的应用：

- 经过标注和预处理的语料库可以为所有的基于统计的自然语言理解提供统计的数据资源
- 语料库的自动加工技术是具体应用的基础

第6章 音字转换技术

音字转换是汉字智能拼音键盘输入和汉语连续语音识别中的关键问题。

6.1 引言

文字输入方法依据输入设备分类：

- 语音输入

- 手写输入
- 键盘输入：使用最广泛

汉语的发音以音节为单位，音节是语音识别的基本单位，也是拼音键盘输入的基本单位。

每个音节对应多个汉字，在以单字或者词为单位进行语音输入或者拼音键盘输入时，给定输入音节，对应多个同音字或者同音词，需要用户进一步确定，影响输入速度。

6.2 声音语句输入

从声音到文字的输入过程：

- 语音识别阶段：把自然的声音信号转换成机器可以处理的数字表达的音节形式（或者拼音形式）
  - 单音节识别（字识别）
    - 汉语是单音节语言
    - 汉语的音节数量不大
    - 单音节输入比多音节输入变化范围小，系统资源要求低
  - 多音节识别（词识别）
  - 采用模块匹配法或者隐马尔可夫模型将听觉信号转换为数字表达的音节形式。
- 语音理解阶段或者音字转换阶段，把音节转换为汉字形式。
  - 字处理：把语音识别器给出的和输入音节相近的几个音所包含的近音字提供给用户选择
  - 词处理：根据用户读入音节停顿时间的不同来确定词的长短，再将词与系统词库进行近音匹配，然后将近音房屋中提供比值用户选择
  - 语句处理：明显优于字和词处理形式

- 从操作心理学的角度来看，人倾向于按照有一定意义的短语或者句子为单位进行短时记忆
- 从信息论的角度来看，汉字的多维熵要低于一维熵，语句输入法相比字和词输入需要更少的输入信息

- 从音节候选量中选择出正确的音节，再从音节中选出正确的汉字

声音语句输入系统的主要模块：

- 语音识别模块：把声音输入转化为音节候选二维向量
- 音词自动切分模块：系统自动分词
- 语义和语法推理模块：根据分词后的近音语句查找知识库，根据语义和语法规则进行自底向上的归约推理
- 概率推理模块：进行歧义判断选出最有希望的结果作为输出
- 输出验证和机器学习模块：
  - 如果转换结果正确，则系统对转换机制及所用的知识加以肯定，即对转换该语句所使用的字、词、规则等的可能性或者优先度进行小幅度增值处理，也可对其他知识进行小幅度减值处理，这称为自然记忆。
  - 如果转换结果错误，则系统为用户提供一个实时修正转换错误的方法，即句内编辑。系统将对用户改正后的汉字语句进行词法、语法、语义等方面的分析，并且进行词和规则的自动生成、对相关字词的可能性及优先度大幅度增（减）值处理等操作，对知识库进行较大的修改和更新，以便保证今后对该类语句声音输入的正确性，称为强化记忆。

6.3 汉字智能拼音键盘输入

汉字键盘输入方法分类：

- 基于字形的输入法：重码率低，平均码长较短，输入速度快，学习困难
- 基于拼音的输入法：易学易用，重码率高，输入速度慢。

6.4 拼音输入的多种表达形式

音节是语音中最小的结构单位，也是人们可以感知的最小语音单位。

音素是从音质角度划分的最小语音单位，包括：元音和辅音。

普通话的音节由 3 部分构成：

- 声母
- 韵母
- 声调

汉语拼音输入法：

- 基本形式是采用汉语拼音的原型，即“全拼”。
- 拼音的压缩表达（简拼、双拼、三拼）
    - 快速语句输入
- 用户自定义简拼
- 模糊拼音输入
- 面向数字键盘的数字拼音输入

拼音提示输入：

- 声母提示
- 拼音提示

6.5 拼音预处理

拼音流的切分

拼音流的切分就是确定拼音串中的各个音节。

拼音音节的自动断开可通过事先在大规模拼音语料基础上建立统计模型 ,即通过拼音语音模型来实现。这种方法通过整句的上下文信息来确定切分结果 ,利用统计信息来确定多个切分结果中的最大概率结果，从而获得很高的准确率，但是在用户修改切分错误时会有不便。

基于拼音规则的拼音切分纠错算法 利用几条拼音边界切分的规则即可实现拼音音节的断开，少数切分歧义通过用户手工断开即可。

算法需要两个数据结构支持：

- 汉语拼音的声母表：保存汉语拼音中的所有声母，用于判断输入字符是否为声母
- 汉语拼音的有效拼音表：保存汉语拼音中的所有有效拼音以及它们的 ID，主要用于判断拼音串是否为有效拼音或者有效拼音的一部分
    - 在判断拼音串时 ,利用二分雾里看花方法查找有效拼音表 ,给出当前拼音串的判断结果

拼音输入的纠错

用户在输入拼音串的过程中引起输入错误的原因：

- 人们使用语言文字的习惯
- 中国不同地域人的发音差异
- 用户对输入拼音的正确性不太关注，只关注转换后的汉字是否正确
- 用户对键盘的熟悉程度

拼音纠错技术采用了“词组匹配纠正法”：先找到错误拼音所在的位置，然后用“可信度”来衡量音字转换结果中汉字的正确程度。

“可信度”计算考虑的因素：

- 转换结果中单字词容易出错，而多字词不容易出错
- 某词与其前后相邻词的关系越小，其可信度越小

6.6 音字转换的实现方法

- 基于理解的方法：利用汉语语法知识来消化同音字、词，以及化解歧义分词
  - 表述为计算机识别和处理的一系列固定描述、公式和自定义规则
    - 根据自动分词得到的同音字、词的候选集，查找知识库得到相关的规则，再经过归约推理，得出转换结果。
    - 利用句内编辑实时修正转换错误和批量学习可以使得系统知识不断完善和充实，这是自学习功能。
  - 这个方法属于自然语言理解领域
    - 优点是采用了自行构造的“语法体系”，正确率比较稳定
    - 缺点是覆盖面较小，对于语法不规范时，无法有效处理；在建立知识库时，知识的表达和获取都非常困难
- 基于语用统计的方法：利用语用统计的数据来消化同音字、词，以及化解歧义分词
  - 主要通过汉语中字与字或者词与词之间的同时出现概率来完成汉语语用统计库的构造。
  - 这个方法属于统计学和运筹学范畴
    - 优点：适应于大规模真实文本的应用，对于已经进行过语用统计或者具有相同

类型的领域，系统的转换正确率比较高

- 缺点：方法具有一定的偏向性
- 对于某个用户，语用统计库会在使用过程中从通用模型转变为用户专用模型

- 基于模板匹配的方法：认为汉语语法知识存在于巨量的短语串中，利用这些短语串来消化同音字、词，以及化解歧义分词
  - 这些短语串称之为“模板词”
  - 这个方法基于计算机的存储和搜索技术实现
    - 优点：对于已经搜索过模板词的或者具有相同类型的领域，系统的转换率较高
    - 缺点：模板词的数量巨大，因此对计算机的存储空间要求较高
- 基于上下文关联的音字转换：利用上下文关联的语用环境来智能选择重码字词的方法
  - 这个方法属于自动控制分支非线性控制范畴

第7章 自动文摘技术（Auto-Summarization）

7.1 引言

文本摘要是指通过对全文信息进行处理，从中提取出最重要的内容，经过重组后生成比原谅更加简短、更加精练的文本（原文摘要）的过程。

人工摘要的缺点：

- 每个人对文本内容的理解受到其自身背景知识的影响，使得人工撰写的摘要存在主观性
- 对于面向特定任务或者基于用户请求的文摘任务，无法满足用户对信息获取的实时性要求

自动文摘系统：

- 文本内部表示：将用户输入的文档进行词语、语句、段落、章节等的划分，由于文档的结构特点，使用结构树的形式化方法表示划分后的结果
- 文档分析：对文档进行不同层次的分析，基于一个定量的标准，衡量每个文档基本单元（语句、段落或者章节）的重要程度，并且基于分析将度量的结果赋值给每个基本单元
  - 浅层分析：对文档中蕴含的浅层的特征进行统计和分析，然后将其中的某些特征按照特定的量化模型结合起来作为文档信息的量化度量，并且根据度量选择预见文档的核心内容。分析结果只需要划分出文摘提取时的基本单元，不需要复杂的文档内部结构表示
    - 主题特征(thematic feature)：也叫题旨特征，是指文档中出现的在统计意义上信息含量较高的词语，主要根据对短语的频度统计来进行分析
    - 位置特征：指在文本、段落中的位置
    - 背景特征：指在标题、子标题、文档首段或者用户查询中出现的词或者短语
    - 指示性（cue）词语和短语：指在文档内的一些特定的词语，或者起头强调作用的词语，以及一些特定领域的专有词汇，因为这些词汇往往对文档的中心主题有关很好的指示作用
  - 实体层分析：将文档转化成内部表示的形式，然后分化出文档的各个实体，并且建立起文档实体间的相互关系。
    - 实体是指组成文档的各个基本单元，每个语句或者每个段落都可以称为一个实体。
    - 对文档实体及其相互关系将有助于确定各个实体表述文档内容的作用。实体间的相互关系：
      - 相似关系

- 相近关系：即文档各个单元之间的距离
- 同现关系：指两个词语在同一个上下文中出现。
  - 词语同现
  - 词性同现
  - 词义同现
- 基于辞典的关系
  - 上下位关系
  - 同义关系
  - 反义关系
- 逻辑关系
- 语法关系：语法分析树
- 语义表示关系：断言——论点关系

- 话语层分析：对全文的宏观结构进行建模
  - 文档格式
  - 区分文本主题的线索
  - 文本的修辞结构

- 文摘提取：关键在于提取哪个层次的文档基本单元用来生成摘要
  - 低层次语言生成技术：提取词语来生成语句，形成摘要，技术简单，效果较差
  - 高层次语言生成技术：提取单句或者复句来形成摘要，技术复杂，效果较好
- 摘要生成：将文摘基本单元进行合成，并做进一步的加工。
  - 低层次语言生成技术：需要处理语句生成问题
  - 高层次语言生成技术：需要解决指代消解（anaphora resolving）

因为当前技术下文摘生成只能做浅层工作，因此文档分析是当前文摘系统的核心部分。

文摘方法分类：

- 机械式文摘方法：基于浅层分析的文摘方法
- 理解式文摘方法：基于实体层及话语层分析技术的文摘方法
- 复合文摘方法：综合机械式文摘方法和理解式文摘方法

7.2 文本的内部表示方法——文档结构树

文档结构树：章节⊃段落⊃复句⊃分句⊃词语

7.3 基于浅层分析的文摘技术

基于浅层分析的文摘系统的处理过程：

- 脱机处理：手工建立自动文档系统的训练语料库，并且通过对训练语料库的统计来建立一个通用（与特定文档无关）的特征库。
- 联机处理：
  - 根据输入文档以及通用特征库来抽取输入文档中所蕴含的特征信息，即文档特征库
  - 根据文档特征库以及相应的可能会计算方法来计算每个语句的权值
  - 根据语句权值的大小，以及文摘系统的连贯性来抽取文摘，再生成摘要。

建立特征库

Edmundson 的算法中用到的特征信息：

- 线索词词典：
  - 奖励词典：如果一个语句中包含有该词典中的词或者短语，则该语句作为文摘句的可能性就会增加，即可以获得正向加权

- 惩罚词典：如果一个语句中包含有该词典中的词或者短语，则该语句作为文摘句的可能性就会减小，即可以获得反向加权
- 无关词典：这个词典中的词条与语句的加权无关
- 基于训练语料库，统计每个词语的相关信息（频度、分布、选择率）来决定词语属于哪个词典。
  - 奖励词选择：当词语的选择率大于 λ_奖
  - 惩罚词选择：当词语的选择率小于 λ_惩
  - 无关词选择：当词语的分布大于给定的域值，并且选择率在两个给定的阈值之间（λ_奖<λ<λ_惩）
  - 后备词选择：当词语的分布小于给定的域值，并且选择率在两个给定的阈值之间（λ_奖<λ<λ_惩），这个词列入候选列表，这个列表用于调整其他词典

- 关键词词典：采用词频统计完成。即高频度的实词对于一篇文档所表述的内容起着重要作用。
- 标题词库：在标题以及各级子标题中出现的，并且没有在无关词典中出现的实词。
- 位置特征：处于关键位置的语句中的词语选择率高过处于文章其他位置的语句中的词语选择率。
  - 标题词加权
  - 位置加权：文章的开头和结尾

文摘句抽取

基于浅层分析文摘抽取方法

- 关键工作：

- 建立特征辞典
- 对词典中的元素进行加权

- 优点：系统容易实现，处理效率高，适合于大规模文档的在线处理
- 缺点：缺乏对文本内容的深层分析与理解
  - 无法保证文摘的逻辑连贯性
  - 无法准确地判断文本的中心主题
  - 无法根据用户提交的关键词或者语句来抽取用户特别关心的内容

7.4 基于实体分析的文摘技术

特征提取

TF-IDF：

- 词频（Term Frequency，TF）：术语在文章中出现的次数；
- 文档频率的倒数（Inverse Document Frequency，IDF）：出现术语的文章数目的倒数

主要特征：

- 词语搭配库：存储词语搭配关系，为每个词语提供相应的语境，排除语义模糊性。
  - 语法上的搭配
  - 习惯用法
  - 事理上的搭配
- 命名实体特征库：命名实体识别
  - 人名与文档所涉及的领域之间并没有明显的关系
  - 机构与文档所涉及的领域之间存在明显的关系
- 语义词典：提供语义所属的语义类别，以及各个词语之间或者各个义素之间的语义关系

- 同义词关系的作用
  - 在同一文档中，不同的词语表达相同的意义
  - 在面向任务的自动文摘系统中 ,根据用户要求提取与用户所用语义相同的内容

文摘抽取 :通过建立好的特征库对每个语句进行加权 ,最后根据给定的文摘比率来抽取权值最大的多个语句，构成文摘语句集。

7.5 基于话语结构的文摘技术

文摘中抽取的语句必须能够重建文章中的内在结构。

话语结构的类型：

- 衔接性（结构衔接）: 结构与形式上的衔接，考虑的是文本呈现的表面结构如何彼此串联
  - 衔接的类别
    - 文本之间运用适当的连接词或者副词来串连形成句子
    - 句子之间依赖同样的主题词以及类似的语法结构来串连形成段落
  - 衔接性的处理
    - 指代
    - 省略
    - 关联关系
    - 词汇层关系
      - 反复
      - 同义
      - 上位

- 连贯性（语义衔接）：语义上的连贯性，表示文本实体之间是否基于相同的主题进行讨论
    - 连贯性的处理偏重于段落、复句、分句之间的宏观关系，这种关系可以通过明显的线索词反映出来，也可以通过统计方法来获得文本实体内部的语义连贯性

常用的基于话语结构的文摘方法：

- 基于词汇衔接的文摘方法：是利用统计方法来获取文章的衔接结构，并且在此基础上生成文摘。
- 基于话语树的文摘方法：利用话语树分析器来描述文章的连贯性，并且通过话语结构分析来生成文摘

基于词汇衔接的文摘方法

衔接性的分类：词汇衔接和连接关系在文摘抽取中具有比较重要的作用

- 词汇衔接
- 连接关系
- 指涉
- 代替
- 省略

基于词汇链的文摘系统中，词汇衔接是通过相应的词汇链来描述的，词汇链是从文档中抽取出来的，是具有某种相似或者相关特征的词语链表。

词汇链的生成过程：

- 选择一个候选词汇集（特征词集）。
- 对于每个候选词，根据相应的判别准则来寻找相应的词汇链

- 如果词汇链存在，则将候选词插入到词汇链中；否则，创建一个或者多个新的词汇链。

基于话语树的文摘方法

话语结构：是采用话语树的形式来描述的

话语树中节点的关系：

- 对称关系：涉及两个或者多个节点，这些节点称为核节点（nucleus），在文章中所有节点表达的信息具有同等重要程度
- 非对称关系：只涉及两个节点【核节点和辅助节点（satellite）】，在文章中核节点表达的信息比辅助节点的更为重要，辅助节点依赖于核节点，并且在不同的依赖关系中以不同的方式改变着核节点所表述的内容

话语结构是一个层次结构，因为一个关系中的辅助节点可能是另一个关系中的核节点，同样一个关系中的核节点也可能是另一个关系中辅助节点。

基于话语树的文摘方法的过程：

- 节点关系分类：11 种类型【解释，并列，递进，对立，选择，充分，必要，让步，无条件，因果，转折等】。
  - 根据语言学知道判断 11 种类型是对称关系还是非对称关系
  - 通过机器学习的方法来确定依赖关系的权值，权值用来度量节点之间的关联程度
- 话语结构分析：确定各个文本单元所对应的节点之间的相互关系。
  - Marcu 给出的完整的分析方法存在的问题
    - 不含关联词的分句的分析，可以基于启发式规则来解决
    - 关联词本身的歧义，也叫词语兼类现象，可以通过词性标注算法来判别
- 文摘建立：话语结构树给出了各个文本单元之间修辞关系，还提供了建立文摘的基础

7.6 文摘系统评测方法

人工主观评测：语言学家阅读系统生成的文摘，并且根据文摘的连贯性、流畅性、概括性等性能给每篇文摘一个主观评价，再根据所有文摘的平均性能评价文摘系统的整体性能

人工对比评测：语言学家和系统对相同的语料自动生成文摘，由第三方评测人员对每篇文摘打分，再对所有文摘进行排序，依据这个结果判断系统性能。

自动评测体系：语言学家和系统对相同的语料自动生成文摘，基于字符串比较方法来对系统生成文摘进行打分，再根据所有文摘的得分来评价文摘系统的整体性能。

- 优点：相对客观，效率较高
- 缺点：只能评测抽取式文摘，无法判断文摘的连贯性、流畅性等无法量化的指标。

7.7 关键词自动抽取

主题词抽取：从领域的主题词表中选择能够代表文章内容的词

关键词抽取：允许出现自由词，即抽取的词可以不在主题词表中。

候选关键词短语的抽取方法：

- 从文中抽出的 N-gram 词对作为候选关键词短语
- 抽取名词短语 chunk 作为候选关键词短语
- 根据词性标注的匹配模式抽取词对

文章关键词的重要性评价：是一个有监督的机器学习问题，可以实现从原文中抽取关键词短语。所依赖的特征：

- 词的全文频率
- 词的首次出现位置
- 词性

- 词汇集聚现象：词汇链计算对文章进行词汇集聚分析，使每个链上比较重要的词的权重都得以增加

冗余消除：同义词或者近义词会造成冗余；存在简略关系的词也会造成冗余。可以构造一个按照重要性程度排序的候选关键词集合，再对其进行冗余处理输出最终的关键词

7.8 小结

自动文摘技术的分析方法：

- 浅层分析
- 实体分析
- 话语结构分析
- 基于特定框架或者模板的文摘技术
  - 优点：引入领域相关的模板，方便生成摘要
  - 缺点：每个模板都只能满足特定领域的摘要生成

第8章 信息检索技术

8.1 信息检索综述

信息检索的定义

信息检索（Information Retrieval）是指从非结构化的数据记录，特别是包含自由格式的自然语言文本的数据记录中获取与用户的信息需求相关的数据记录的系统、方法与过程。

文档（Document）：自然语言文本数据记录。

数据全集（Collection）：将大量的非结构化的数据记录按照方式组织和存储起来，构成的数

据记录的集合称为信息检索中的数据全集。

检索分类：

- 信息检索：从非结构化的数据记录中获取用户信息需求
  - 受到自然语言的丰富性和二义性的限制，信息检索在精确性和召回率等方面都比数据库数据检索低，其难度也更大。
  - 信息检索系统的实现需要借助于数据库管理系统。
- 数据库数据检索：从结构化的数据记录中获取用户的信息需求

信息检索过程：是根据用户特定的信息需求，在数据全集中获取所有和仅有的与用户信息需求相关的文档，并将这些文档按照相关性的大小由大到小地排列。

查询：就是用户特定的信息需求，是反映用户信息需求的字符串。

- 由关键字序列描述的字符串
- 由布尔表达式描述的字符串
- 使用自然语言表达的问句

相关性：是信息检索结果符合用户信息需求的程度。

信息检索系统

信息检索系统：是一个能够对数据全集的数据记录进行存储、组织与维护的系统，还可以根据用户的查询获取相关的信息。

信息检索系统的组成：

- 8 个基本处理模块
  - 用户接口模块：与信息检索系统的用户交互信息，接受用户的查询，根据用户对信息检索结果的反馈调整信息检索系统的有关参数，显示用户查询的结果等

- 用户查询文本操作模块：对用户的查询字串进行停用词过滤、词干抽取等处理，并且转换为机器内部的用户查询表示格式
- 文档文本操作模块：对文档数据库中的文档进行停用词过滤、词干抽取等处理，并且将文档转换为机器内部的文档表示格式，是建立索引模块处理的基础。
- 用户查询处理模块：对用户查询的词汇进行同义词扩充，或者根据用户对信息检索的倾向性对查询的词汇进行转换处理。
- 索引构建模块：建立从词汇到这个词汇出现的文档（编号）的倒排索引表，从而对用户查询中的词汇进行快速定位。
- 数据库管理模块：将文档以数据库的格式存储、管理、编辑和访问。
- 搜索模块：根据用户查询，借助倒排索引表和数据库管理模块从数据库中抽取出包含用户查询关键字的文档。
- 相关度排序模块：逐一计算用户查询与搜索模块返回文档的相关度，最后将这些文档的相关度按照从大到小的顺序排列。

- 2 大系统资源
  - 整个系统公用的语义词典
    - 系统词汇及其语法语义信息
    - 停用词表
    - 词形转换表
  - 以数据库形式存放的数据全集
- 2 个子系统
  - 检索子系统：
    - 接受用户查询，对用户查询词汇进行停用词过滤和词干抽取等处理，然后进行

同义词扩充等转换处理，转换为用户查询的机器内部格式（通常是经过同义词扩充的关键词序列）。

- 通过由信息存储管理子系统预先建立的倒排索引表，找到包含这些关键字的所有文档，并将它们作为候选文档，然后逐一计算用户查询与候选文档之间的相关度，将候选文档的相关度按照从大到小的顺序排列，并返回给用户。

- 信息存储管理子系统：
  - 将数据全集文档存储于数据库中，并提供对这些文档的编辑、增删等操作
  - 对文档进行停用词过滤、词干抽取等处理后生成的副本也存储于数据库中，并对文档的副本建立倒排文件供检索使用

信息检索系统的评价

信息检索系统的目标：使满足系统用户信息需求的开销最小。

开销：是指从用户向系统输入了一个查询开始，到他讲到了包含他的信息需求的文档为止的全部时间。

一个系统在实际应用中的时间和空间消耗是衡量一个系统优劣的重要指标。

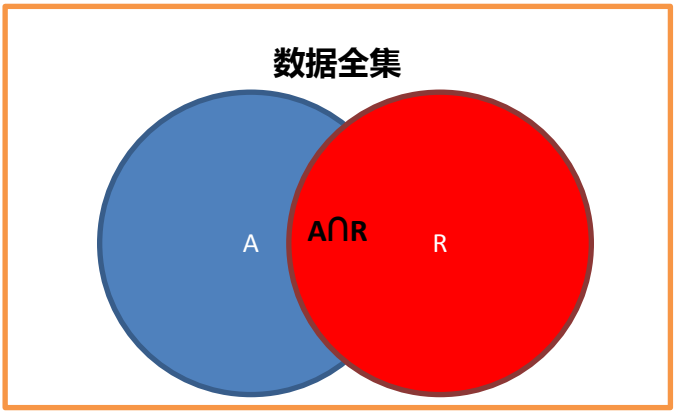
一个信息检索系统在实际应用中“相关性”才是评价这个系统的重要指标。

基于相关性的系统评价指标：

- 精确度（Precision）=|A∩R|/|A|
  - 检索获取的相关的数据记录个数与检索获得的所有数据记录个数的比值
  - 给定一个检索获取的数据记录，与用户查询相关的概率
  - 反映了系统能够返回与用户查询相关数据记录的能力
- 召回率（Recall）=|A∩R|/|R|

- 检索获取的与用户查询相关的数据记录的个数与数据全集中所有与用户查询相关的数据记录个数的比值
- 一个相关数据记录能够被检索系统获取的概率
- 反映了系统能够找到全部相关数据记录的能力

- 精确度——召回率曲线来定量分析一个信息检索系统的改进情况或者比较几个信息检索系统的优劣。
- 系统有效性度量：E≈1 - (1/ (α(1/P) + (1-α)(1/R)))



8.2 基于统计方法的信息检索模型

基于统计的方法对用户查询和数据全集中的数据的统计量度计算相关性。

信息检索模型是对实际信息检索过程加以抽象而构造的信息检索的数学模型，是关于信息检索的各个主要处理阶段的形式化框架，

信息检索模型使用三元组 IRM=（D,Q,R）表示：

- D 是文档的集合
- Q 是用户需求的集合
- R：DxQ→R 是集合 D 与 Q 的笛卡尔乘积到实数集 R 的一个映射，对每个用户查询 q∈Q，每个文档 d∈D，映射 R 将(d,q)映射为一个实数，称为用户查询 q 与文档 d 的相关度。

基于统计的信息检索模型：

- 布尔模型：文档中索引词的权重只有 0 和 1 两种取值，分别表示文档中包含或者不包含这个索引词
  - 优点：机制简单，检索效率高
  - 缺点：分类能力有限，不能给出相关性度量值导致相关性高的文档排序靠后
- 扩展布尔模型
- 向量空间模型：
  - 模型的组成
    - N 表示整个信息检索系统中的关键词的总数
    - 文档表示为由文档中索引词的权重构成的 N 维向量
    - 用户查询表示为由用户查询中索引词的权重构成的 N 维向量
  - 模型的关键问题
    - 索引词权重的计算
      - 词频与逆文档频度（TF.IDF）：将一个索引词在单个文档中的重要性和在整个数据全集中的重要性结合起来，成为一个统一的量度
        - 一个索引词在文档中出现的频度描述了这个词的重要性
        - 一个索引词的重要性与所在文档的总数成反比或者近似反比的关系，反映了包含这个索引词的文档区别于其他文档的程度，是一个索引词在整个数据命令中重要性的全局性统计特征
      - TF.IDF 的扩展：“正规化”的改进方法增加对文档中索引词总数的考虑
        - 最大正规化法
        - 增量正规化词频

- 对数词频法
- 余弦正规化法 :解决了文档中少数高频词对其他词权值扰动过大问题

◆ 用户查询与文档的相似度计算

- 相似度是用户查询与文档相关性的度量
- 最简单的相似度计算：两个 N 维向量的内积
  - L1 范数计算内积：马氏距离
  - L2 范数计算内积：欧氏距离
  - 无穷范数计算内积：最大方向距离
- 余弦相似度：当两个 N 维向量都经过了余弦正规化处理后，它们的内积恰好就是两个向量夹角的余弦，称为余弦相似度
  - 缺点 :对于一个用户查询，包含索引词个数较多的长文档往往计算结果偏低

- 概率模型：贝叶斯推理网络模型，可以将不同来源的证据结合起来，以确定给定文档满足用户查询或者信息需求的概率要求的方法。
  - 贝叶斯网络是一个描述随机变量之间因果关系的有向无环图（ DAG ）

在模型中，文档被表示为关键词的集合；这一表示方式又称为文档的平面结构；关键词又称为索引词，指除停用词之外的代表文档内容的词汇，大多数是名词。

基于统计的信息检索模型的假设：

- 出现在文档中的词汇彼此独立
- 词汇在文档中的出现没有二义性，即一个词的词义由其词形惟一的确定

以上的假设虽然并不完全正确，但是简化了处理过程，提高了信息检索系统的实时性，且对系统的精度影响不大。

8.3 基于语义方法的信息检索模型

基于语义的方法对用户查询和数据全集中的数据进行语法语义分析 ,即基于用户查询和数据全集内容理解的基础上进行两者的相关性计算 ,将信息处理的层次深入到文档中文本的内容。

语义相似度：是将词汇间的直接或者间接的语义关系映射为表示词汇间语义相关性的数值。

语义相关度计算方法：

- 基于按照概念间结构层次关系组织的语义词典的方法 :这种方法基于一个假设 , 即两个词汇在概念的结构层次网络图中存在一条通路 (主要是上下位关系) 时 , 它们具有一定的语义相关性。
- 基于上下位关系网络中两个词的公共祖先节点的最大信息量来衡量两个词的语义相关度。

基于统计的方法 , 是将词汇的上下文信息的概率分布作为词汇间语义相关度计算的参照 , 这类方法建立在两个词汇具有某种程度的语义相关当且仅当它们出现在相同的上下文中这一假设的基础上。

基于语义词典的方法：依赖于比较完备的按照概念间结构层次关系组织的大型语义词典。

8.4 文本自动分类技术

文本自动分类技术（Text Automatic Classification）：是对一篇文档，根据其内容，从预先定义好的标记集中，找出一个或者多个最适合于这个文档的标记。

文档分类决策矩阵：

- 类别标记是一组符号
- 文档对类别的隶属度是基于文档的内容
- 条件概率表示文档属于类别的可能性

文档分类的特征选择：

- 基于文档频率的特征选择
- 基于信息增益度的特征选择
- 基于χ^2 分布的特征选择
- 基于矩阵分解的特征选择
  - 两个特征的相似度使用“正向”乘法计算
  - 两个文档的相似度使用“反向”乘法计算
  - 文档和特征的相似度词语和文档组成的矩阵表示

文档分类的常用算法：

- Rocchio 算法：经典算法。为每个类别建立原型向量，然后根据文档向量和类别原型向量的距离来确定文档的类别
  - 类别的原型向量是通过计算属于这个类别的所有文档向量的平均值得到
  - 特点：计算速度快，计算精度低
- Naïve Bayes 算法：概率方法。假设文档中词语的出现是相互独立的。
- 决策树算法：用文档向量去匹配一个建立在训练集上的决策树，以决定与类别主题是否相关。算法的关键就是建立决策树。
  - 决策树是一个树型结构，在内部节点上选用一个属性进行分割，每个分叉都是分割的一个部分，叶子节点表示一个分布。
  - 决策树的建立过程：
    - 决策树生成
    - 决策树剪枝
- K 近邻算法：找出训练集中与待分类文档最相似的 K 篇文档，然后根据这 K 个文档的类

别确定这个文档的类别

- 相似度度量标准是欧氏距离或者余弦夹角。
- 是一种基于实例的“惰性”学习算法，不需要独立的模型训练过程。

- 最大熵分类算法：可以集成不同种类信息的分类模型，把在训练集中的、与分类有关的数据描述为一系列的特征
  - 对数线性模型整合不同的特征
  - 使用 GIS 算法，迭代循环过程来求解

第9章 文字识别技术

9.1 引言

汉字输入方式：

- 非智能的输入方式：编码输入法
- 智能拼音输入法：汉字识别和语音输入

汉字识别是指计算机自动辨识印刷或者手写体汉字的技术，依据识别对象以及输入设备的不同，汉字识别可以分为：

- 印刷体汉字识别：由计算机识别通过光电扫描仪输入的印刷体汉字的技术
- 脱机手写体汉字识别：由计算机识别人写在纸上的手写体汉字的技术
- 联机手写体汉字识别：由计算机识别人写在数字化仪器上的手写体汉字的技术，组成部件：
  - 前端单字识别器（Single Character Recognizer，SCR）：一般可以根据输入的汉字图像序列生成一个由候选汉字组成的矩阵

- 汉字识别后处理器，也叫语言解码器。基于字的汉语 N-gram 模型计算每个候选语句的出现概率
- 语言解码器 :自动处理句子中的上下文信息 ,从而对文字图像进行某种程度的理解，以便自动修正识别错误，选择最优的候选识别字，提高手写汉字识别的正确率。

9.2 联机手写体汉字识别的国内外研究概况

国外研究概况：

- 美国 Buffalo 大学的 CEDAR
- 加拿大 Concordia 大学的 CENPARMI
- 荷兰的 NICI 研究所

国内研究概况：

- 清华大学
- 北京大学
- 中科院自动化研究所
- 哈尔滨工业大学

9.3 联机手写体汉字识别方法综述

基于统计的识别方法

以统计为特征基础，以汉字的结构信息为辅助特征的识别方法。采用特征向量与模板匹配的方法。缺点是分辨相似字的能力较弱，在识别字集增大或者畸变幅度增大时，这类方法只适合作为组分类方法使用。

常用的整字的统计特征：

- 续波形谱分析特征 :笔尖的运动轨迹是两个坐标采样的序列 ,对此序列做 Fourier 变换，把展开式的低阶系数作为汉字识别的特征。
  - 低阶系数既能够有效反映曲线的大致变化，又能够克服高频噪声，适合描述汉字
- 字曲线描述特征 :将输入汉字听前一笔划的结束点与下一笔划的起始点连接起来 ,将汉字看作一条曲线，并且把这条曲线表示为具有固定分段数的 Freeman 链码序列，并且以此序列描述汉字。
  - 匹配时 ,将链码序列中的元素与特征模板中的元素按顺序一一对应 ,产生相应的误差向量 ,再将误差向量与加权向量相乘 ,得到匹配误差值 ,以匹配误差值最小的模板作为识别的结果。
  - 为了提高匹配的速度，可以采用在整字曲线描述基础上的动态规划算法。

基于整字的联机汉字识别特征及其匹配方法的共同特点：

- 允许输入汉字的笔划之间有连笔，但是要求笔顺正确
- 忽略了汉字的结构信息，当识别字域增大时，识别效果会变差

基于结构的识别方法

汉字是有结构的线段图形，汉字笔划之间有一定的相关性，无论如何变化其相互位置不变。基于结构的识别方法需要先识别笔划，然后以这些笔划作为主要特征来识别汉字。

笔划的识别方法：

- 按笔划的书写方向及其变化识别笔划
  - 常见的笔划的基本类型：横、竖、撇、捺、点。这些基本笔划可以组成复杂笔划。
  - 基本笔划都是直线型笔划
  - 根据书写笔划时方向的变化，检测出笔划中的折点，根据折点的方向和个数，就可

以进行复合笔划的判断

- 动态规划识别笔划：直接根据笔划上的线段方向，形成笔划方向码序列，然后用动态规划法来识别笔划。
  - 缺点：迭代速度慢，影响识别速度
- 模糊属性自动机识别笔划：利用笔划的方向和长度信息，借助模糊信息处理方法，以不变嵌入原理为基础，提出一种模糊属性语法及其相应的模糊属性自动机。
  - 这种语法形式是有限状态语法，语义规则中上下文信息。
  - 输入笔划的方向码序列和方向码长度输入自动机后，得到的结果是对笔划类的隶属度，隶属度值最大的类就是识别出的笔划类型

经过笔划识别之后，输入文字就被描述成为一个笔划序列或者笔段序列，因此汉字识别问题也就转化为笔划序列或者笔段序列的识别问题。

汉字的组成原则是：笔段→笔划→部件→整字，所以根据对笔划序列或者笔段序列的识别完成部件的识别，最后完成整字的识别。

基于神经元网络的识别方法

- Fukushima 模型：专门用于视觉模式识别的模型
- Neocognition 模型：具有对位置、大小变化的容忍能力
- 清华大学，吴佑寿，自组织聚类网络，用于汉字的自组织聚类
- 中国科学院，刘迎建，组块神经元网络模型：由大量基本单元——组块神经元通过一定的组织关系构成的复杂系统

基于机器学习的识别方法

智能识别的难点在于识别知识的获取，采用概念函数来指导优选属性。

9.4 典型联机手写体汉字识别系统

- 汉王中文手写体汉字识别系统：结构模式识别方法
  - 借助模糊信息处理方法，以不变嵌入原理为着眼，提出了一种模糊属性语法及模糊属性自动机，把在线手写汉字分为笔划、笔段、字根、整字等几个层次，最后进行词组校对。
  - 优点：对笔划变化的容忍度大
  - 缺点：运算比较复杂，自动机的设计比较困难
- 豪文中文手写体汉字识别系统：统计模式识别与结构模式识别相结合的方法，建立了识别系统的信息传递模型。
  - 识别过程分为：笔段识别、字根识别、整字识别
    - 在笔段排序和部件分析过程中采用的是专家知识，表现形式为一条条规则
    - 在整字识别中采用的模板也是基于专家知识
  - 优点：知识表达精练，内存开销小，识别效果好，识别速度快
  - 缺点：知识库维护困难

9.5 联机手写体汉字识别后处理系统

手写体汉字识别：基于词的汉语 N-gram 模型能够利用汉语词典中的构词信息。

联机手写体汉字识别系统具有实时性要求，因此主要采用基于词的汉语 Bigram 模型和简单的 Cache 自学习机制作为语言解码器的统计知识库。

联机手写体汉字识别后处理系统的核心算法模块：

- 词网格生成模块：根据前端识别器生成的候选汉字失量序列生成词网格，构成语言解码器的搜索空间
  - 遍历所有的候选字，生成全部的字节点
  - 遍历所有的字节点，把相邻列的可组词的字节点全部捆绑与词节点
  - 所有的节点按照其结束时刻进行排序，最后生成的词网格构成语言解码器的搜索空间
- 最优候选语句搜索模块：语言解码器使用 Viterbi 算法在词网格中搜索具有最大路径评价值的候选语句，并将搜索结果作为语言解码器的搜索结果，再经过用户的联机校正成为正确的句子，输出到 Cache 自学习模块
  - 根据候选汉字的可信度 汉语中基于词的 Bigram 模型以及 Cache 模型提供的 Bigram 和 Unigram 概率计算出每个候选语句的评价函数值，以便选择最优的候选语句。
- 基于 Cache 的自学习模块：Cache 自学习模块根据用户校正的句子修改相应的 Cache 学习库，调整系统去适应用户的习惯
  - Cache 是一种短期记忆机制。是一组用于学习的统计数据组成，数据来源于用户联机校正的句子，数据存储格式与相应的 Unigram 和 Bigram 相同
  - 语言解码器工作时，Cache 中的概率通过加权可以与系统语言模型的概率结合起来，Cache 的加权系数与系统的学习率成正比

基于词网络的手写体汉字识别的语言学解码方法

相邻列的候选汉字能否组成一个词是正确选择一个最优的候选语句的关键。对于所有的相邻汉字节点，如果它们可以组成一个词，就把它们捆成一个词节点，并把这个词节点挂在这个

词的尾字所在的列中。相邻列的所有字词节点均用边连接起来，它们就构成了一个有向图，称之为词网格。