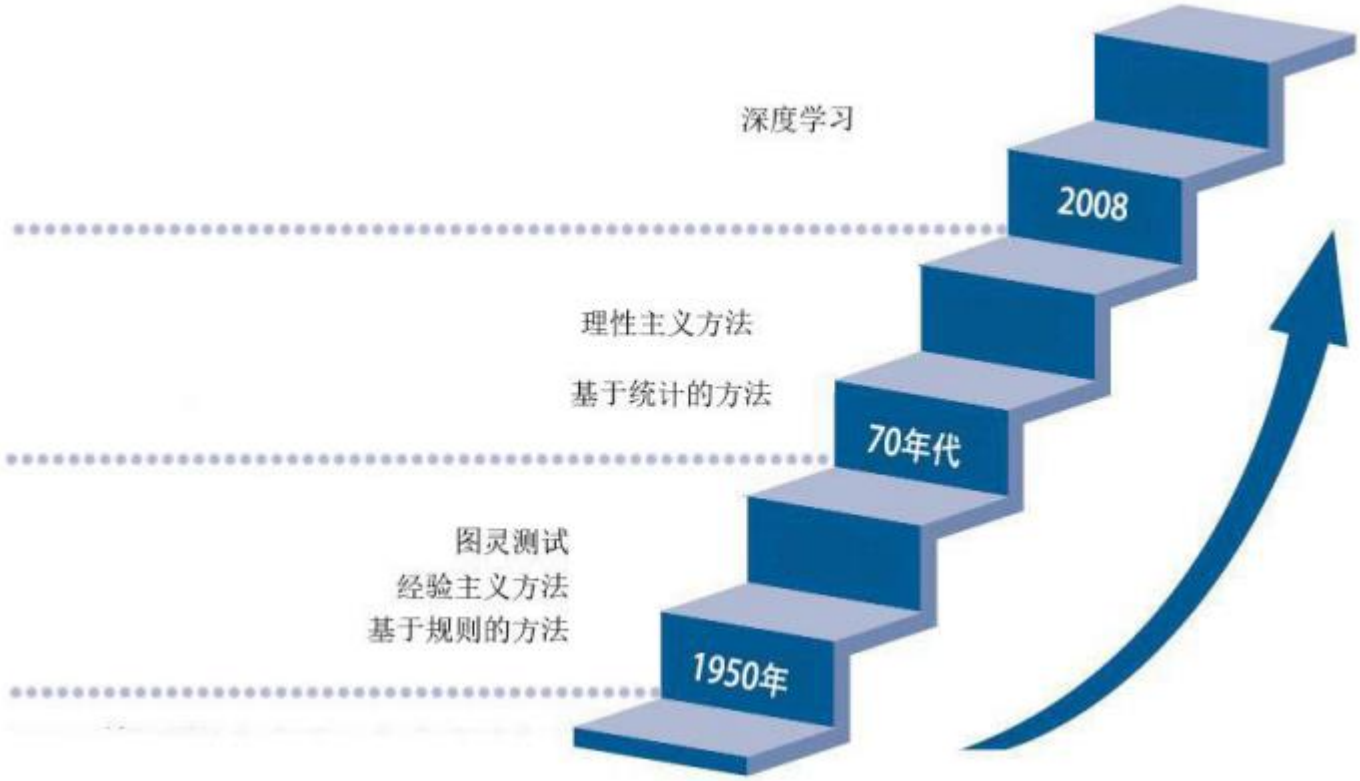
第4节 预备知识

参考宗成庆《统计自然语言》

知识回顾



自然语言发展历程

自然语言处理的两种基本方法

- 基于规则的分析方法
  - 规则库开发
  - 推导算法设计

  理论基础：形式语言与自动机理论

- 基于语料库的统计方法
  - 语料库建设
  - 统计模型建立

  理论基础：数理统计、信息论、语料库

内容概览

1. 概率论基础
2. 信息论基础
3. 形式语言与自动机基础
4. 语料库与语言知识库

1. 概率论基础

1.1 概率

概率是从随机实验中的事件到实数域的函数，用以表示事件发生的可能性。如果用 P(A) 作为事件A的概率，Ω 是实验的样本空间，则概率函数必须满足如下公理：

公理1：P(A) ≥ 0

公理2：P(Ω) = 1

公理3：如果对任意的 i 和 j (i ≠ j)，事件 Ai 和 Aj 不相交(AiⅡAj = Φ)，则有：

$$P(\bigcup_{i=0}^{\infty} A_i) = \sum_{i=0}^{\infty} P(A_i)$$

1.2 最大似然估计(Maximization likelihood estimation, MLE)

如果一个实验的样本空间是 $\{s_1, s_2, \ldots s_n\}$，在相同情况下重复实验 N 次，观察到样本 $s_k (1 \le k \le n)$ 的次数为 $n_N(s_k)$，则 s_k 的相对频率为：

$$q_N(s_k) = \frac{n_N(s_k)}{N}$$

由于 $\sum_{i-1}^{n} n_N(s_k) = N$，因此，$\sum_{i=1}^{n} q_N(s_k) = 1$

当N越来越大，相对频率 $q_N(s_k)$ 就越来越接近 s_k 的概率 $P(s_k)$，事实上，

$$\lim_{N \leftarrow \infty} q_N(s_k) = P(s_k)$$

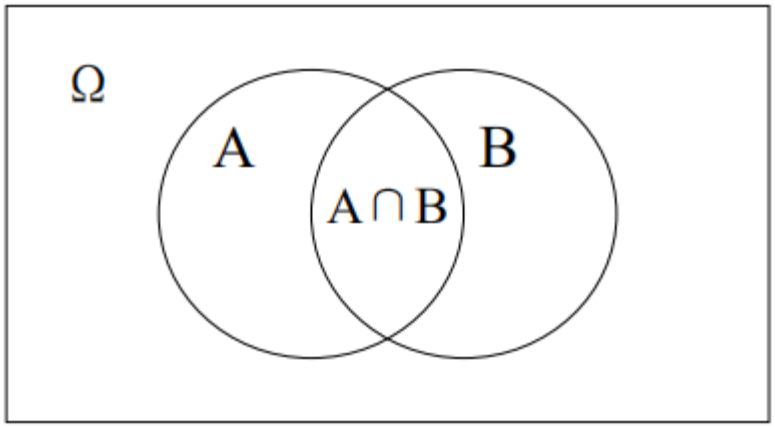
因此相对频率常被用作概率的估计值。这种概率值的估计方法称为**最大似然估计**。

1.3 条件概率 (conditional probability)

如果 A 和 B 是样本空间 Ω 上的两个事件，P(B)>0，那么在给定 B 时 A 的条件概率 P(A|B) 为

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

条件概率 P(A|B) 给出了在已知事件 B 发生的情况下，事件 A 发生的概率。一般地，P(A|B) ≠ P(A).



1.4 全概率公式

设 Ω 为实验E的样本空间， $B_1, B_2, ..., B_n$ 为Ω的一组事件，且他们两两互斥，且每次实验中至少发生一个，即：

- $B_i \cap B_j = \phi (i \neq j; i, j = 1,2,...,n)$
- $\bigcup_{i=1}^{n} B_i = \Omega$

则称 $B_1, B_2, ..., B_n$ 为样本空间为样本空间Ω的一个划分。

设A为Ω 的事件， $B_1, B_2, ..., B_n$ 为Ω的一个划分，且 P(Bi)>0 (i=1, 2, ..., n)，则全概率公式为：

$$P(A) = P(\bigcup_{i=1}^{n} AB_i) = \sum_{i=1}^{n} P(AB_i) = \sum_{i=1}^{n} P(B_i)P(A \mid B_i)$$

1.5 贝叶斯法则(Bayes’ theorem)

如果 A 为样本空间 Ω 的事件，$B_1, B_2, \ldots, B_n$ 为 Ω的一个划分，且 P(A) > 0，P(Bi) > 0 (i = 1, 2, …, n) ，那么

$$P(B_i \mid A) = \frac{P(B_i)P(A \mid B_i)}{\sum_{j=1}^{n} P(B_j)P(A \mid B_j)}$$

当n=1时，

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)}$$

1.6 贝叶斯决策理论 (Bayesian decision theory)

假设研究的分类问题有c个类别，各类别的状态用wi表示，i=1,...,c；对应于各个类别wi出现的先验概率为 $P(\mathrm{w}_i)$;在特征空间已经观察到某一向量 $\overline{\mathrm{x}} = [x_1, x_2, \ldots x_d]^T$ 是d 维特征空间上的某一点，且条件概率密度函数 $p(\bar{x} \mid w_i)$是已知的。那么，利用贝叶斯公式我们可以得到后验概率

$$P(\mathrm{w}_i \mid \bar{x}) = \frac{p(\bar{x} \mid w_i)P(w_i)}{\sum_{j=1}^{c} p(\bar{x} \mid w_j)P(w_j)}$$

基于最小错误率的贝叶斯决策规则为：

- 如果 $P(w_i \mid \bar{x}) = \max_{j=1,2\ldots c} p(w_j \mid \bar{x})$ 则 $\bar{x} \in w_i$
- 或者：如果 $p(\bar{x} \mid w_i)p(w_i) = \max_{j=1,2\ldots c} p(\bar{x} \mid w_j)p(w_j)$ 则 $\bar{x} \in w_i$
- 或者（c=2时）：如果 $l(\bar{x}) = \frac{p(\bar{x} \mid w_1)}{p(\bar{x} \mid w_2)} > \frac{p(w_2)}{p(w_1)}$ 则 $\bar{x} \in w_1$

  否则 $\bar{x} \in w_2$

贝叶斯决策理论在文本分类、词汇语义消歧(word sense disambiguation)等问题的研究中具有重要用途。

例子：假设某一种特殊的句法结构很少出现，平均大约每**100,000**个句子中才可能出现一次。我们开发了一个程序来判断某个句子中是否存在这种特殊的句法结构。如果句子中确实含有该特殊句法结构时，程序判断结果为“存在”的概率为**0.95**。如果句子中实际上不存在该句法结构时，程序错误地判断为“存在”的概率为**0.005**。那么，这个程序测得句子含有该特殊句法结构的结论是正确的概率有多大？

解：假设G表示事件 “句子确实存在该特殊句法结构 “，T 表示事件 “程序判断的结论是存在该特殊句法结构”。那么，我们有：

$$P(G)=\frac{1}{100000}=0.00001 \qquad P(\overline{G})=\frac{100000-1}{100000}=0.99999$$

$$P(T\mid G)=0.95 \qquad P(T\mid \overline{G})=0.005$$

求得：

$$P(G\mid T)=\frac{P(T\mid G)P(G)}{P(T\mid G)P(G)+P(T\mid \overline{G})P(\overline{G})}$$

1.7 二项式分布 (binomial distribution)

当重复一个只有两种输出（假定为 $\overline{A}$或 A）的实验（伯努利实验）， A 在一次实验中发生的概率为p，现把实验独立地重复 n 次。如果用X 表示A 在这n 次实验中发生的次数，那么，$X=0,1,...,\mathrm{n}$ 。考虑事件 $\{X=\mathrm{i}\}$ ，如果这个事件发生，必须在这 n 次的原始记录中有 i 个 A，n - i 个 $\overline{A}$。

$$\underbrace{A\overline{A}A...\overline{A}}_{n个}\Rightarrow \mathrm{p}^{i}(1-p)^{n-i}$$

A可以出现在n个位置中的任何一个位置，所以，结果序列有 $\binom{\mathrm{n}}{i}$ 种可能。由此，可以得出：

$$\mathrm{p_i}=\binom{\mathrm{n}}{i}p^{i}(1-p)^{n-i},\quad i=0,1,...,n$$

其中，$\binom{\mathrm{n}}{i}=\frac{n!}{(n-r)!r!},\quad 0\le r\le n$ ，有时也记作C_n^r

在自然语言处理中，我们常常以句子为处理单位。一般地，我们假设一个语句独立于它前面的其它语句，句子的概率分布近似地认为符合二项式分布。

1.8 期望 (expectation)

期望值是一个随机变量所取值的概率平均。设 X为一随机变量，其分布为 $P(X=\mathrm{x}_k)=p_k$， $k=1,2,\ldots$ 。若级数 $\sum_{\mathrm{k}=1}^{\infty} x_k p_k$ 绝对收敛，那么，随机变量X的数学期望或者概率平均值为：

$$E(X)=\sum_{\mathrm{k}=1}^{\infty} x_k p_k$$

1.9 方差 (variance)

一个随机变量的方差描述的是该随机变量的值偏离其期望值的程度。设 X 为一随机变量，其方差为：

$$\mathrm{Var}(X) = E((X - E(X))^2)$$

2. 信息论基础

2.1 熵 (entropy)

熵是信息论中重要的基本概念。熵又称为自信息（self-information），表示信源X 每发一个符号（不论发什么符号）所提供的平均信息量。熵也可以被视为描述一个随机变量的不确定性的数量。一个随机变量的熵越大，它的不确定性越大。那么，正确估计其值的可能性就越小。越不确定的随机变量越需要大的信息量用以确定其值。

如果 X 是一个离散型随机变量，其概率分布为：$\mathrm{p}(x) = P(X = x), \quad x \in X$ 。

X 的熵 $H(X)$为：

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

其中，约定 $0\log 0 = 0$。H(X) 也可以写为 H(p)。通常熵的单位为二进制位比特（bit）。

2.2 联合熵 (joint entropy)

如果 X, Y 是一对离散型随机变量 X, Y ~ p(x, y)，X, Y 的联合熵 H(X, Y) 为：

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

联合熵实际上就是描述一对随机变量平均所需要的信息量。

2.3 条件熵 (conditional entropy)

给定随机变量 X 的情况下，随机变量 Y 的条件熵定义为：

$$H(Y \mid X) = \sum_{x \in X} p(x) H(Y \mid X = x)$$

$$= \sum_{x \in X} p(x)[-\sum_{y \in Y} p(y \mid x) \log p(y \mid x)]$$

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

连锁规则

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

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

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

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

$$= H(X) + H(Y\mid X)$$

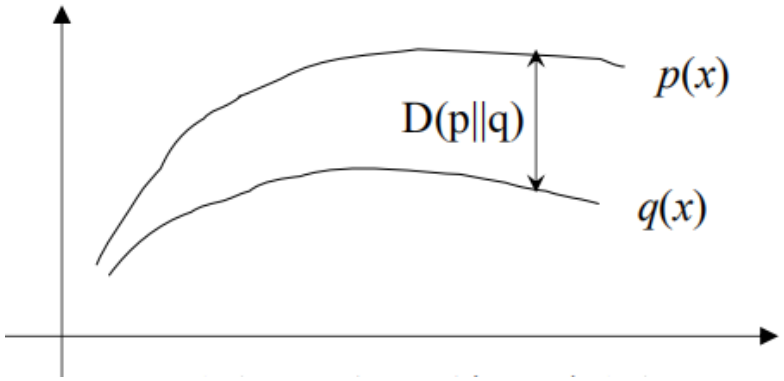
2.4 相对熵 (relative entropy or Kullback-Leibler divergence)

相对熵常被用以衡量两个随机分布的差距。当两个随机分布相同时，其相对熵为0。当两个随机分布的差别增加时，其相对熵也增加。

两个概率分布 p(x) 和 q(x) 的相对熵（或 Kullback_x0002_Leibler 距离，简称 KL 距离）定义为：

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

定义中约定 $0\log(0/q)=0,\quad p\log(p/0)=\infty$ 。



2.5 交叉熵 (cross entropy)

如果一个随机变量 X ~ p (x)，q(x)为用于近似 p(x)的概率分布，那么随机变量X和模型 q之间的交叉熵定义为：

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

交叉熵的概念是用来衡量估计模型与真实概率分布之间差异情况的。

2.6 困惑度(perplexity)

在设计语言模型时，我们通常用困惑度来代替交叉熵衡量语言模型的好坏。给定语言L的样本 $l_1^n = l_1 \ldots l_n$ ，L的困惑度 PP_q 定义为：

$$PP_q = 2^{H(L,q)} \approx 2^{-\frac{1}{n}\log q(l_1^n)} = [q(l_1^n)]^{\frac{1}{n}}$$

语言模型设计的任务就是寻找困惑度最小的模型，使其最接近真实的语言。

3. 语料库

3.1 语料库(corpus)

- **语料库(corpus)** 就是存放语言材料的仓库（语言数据库）
- 语料库一词在语言学上意指大量的文本，通常经过整理，具有既定格式与标记。其具备三个显著的**特点**：
  - 语料库中存放的是在语言的实际使用中真实出现过的语言材料
  - 语料库以电子计算机为载体承载语言知识的基础资源，但并不等于语言知识
  - 真实语料需要经过加工（分析和处理），才能成为有用的资源

生语料库是指收集之后未加工的语料库 相对而言，熟语料库经过加工的语料库

3.2 语料库语言学(corpus linguistics)

- **定义**：基于语料库进行语言学研究
- **研究的内容**：
  - 语料库的建设与编纂
  - 语料库的加工和管理技术
  - 语言研究中语料库的使用
  - 语料库语言学在计算语言学中的应用

3.3 语料库的类型

- **按内容构成和目的划分**
  - 异质的（heterogeneous）：最简单的语料收集方法，没有特定的语料收集原则，广泛收集并原样存储各种语料
  - 同质的（homogeneous）：只收集同一类内容的语料
  - 系统的（Systematic）：根据预先确定的原则和比例收集语料，使语料具有平衡性和系统性，能够代表某一范围内的语言事实
  - 专用的（specialized）：只收集用于某一特定用途的语料

- **按语言种类划分**
  - 单语的（Monolingual）
  - 双语的（Bilingual）
  - 多语的（Multilingual）

多语种语料库又可以再分为比较语料库（comparable corpora）和平行语料库（parallel corpora）。比较语料库的目的侧重于特定语言现象的对比，而平行语料库的目的侧重于获取对应的翻译实例。

平衡语料库与平行语料库

- **平衡语料库**：平衡语料库着重考虑语料的代表性与平衡性

➢ 语料采集的七项原则：语料的真实性、语料的可靠性、语料的科学性、语料的代表性、语料的权威性、语料的分布性和语料的流通性。其中，语料的分布性还要考虑语料的科学领域分布、地域分布、时间分布和语体分布等。【张普，2003】

- **平行语料库（parallel corpora）**

➢ **含义一**：在同一种语言的语料上的平行。其平行性表现为语料选取的时间、对象、比例、文本数、文本长度等几乎是一是一致的。建库的目的是进行对比研究。

➢ **含义二**：在两种或多种语言之间的平行采样和加工。例如，机器翻译中的双语对齐语料库

C: 早晨1 好2 !3

E: Good2 morning1 .3

按语料选取的时间划分

- 共时语料库（syn-chronic corpus）：为了对语言进行共时研究而建立的语料库
- 历时语料库（diachronic corpus）：为了对语言进行历时研究而建立的语料库

3.4 语料库设计需要考虑的问题

- **代表性**：一个语料库具有代表性，是指在该语料库上获得的分析结果可以概括成为这种语言整体或其指定部分的特性
- **规模性**：一般而言，在保证质量的前提下应足够大；随着语料库的增大，垃圾语料越来越多，语料达到一定规模以后，语料库功能不能随之增长
- **结构性**：目的地收集语料的集合，必须以电子形式存在，计算机可读的语料集合结构性体现在语料库中语料记录的代码、元数据项、数据类型、数据宽度、取值范围、完整性约束。
- **平衡性**：理想的情况是收入语料库的文本在题材、语体、时间跨度等方面有一个合理的平衡
- 语料库的管理与维护：错误修正或改善版本升级语料库的检索系统、分析和处理工具的维护等

语料库资源

国家语委现代汉语语料库 http://www.cncorpus.org/

LIVAC共時語料庫 http://www.livac.org/index.php

LDC中文语言资源联盟 http://www.chineseldc.org/

知网 http://www.keenage.com/html/c_index.html

搜狗实验室数据资源 http://www.sogou.com/labs/resource/list_yuliao.php

哈工大信息检索研究室对外共享语料库
http://ir.hit.edu.cn/demo/ltp/Sharing_Plan.htm

传媒大学文本语料库 http://ling.cuc.edu.cn/RawPub/

在线分词标注系统 http://ling.cuc.edu.cn/cucseg/

兰开斯特汉语语料库 http://ota.oucs.ox.ac.uk/scripts/download.php?otaid=2474

美国当代英语语料库（COCA）http://corpus.byu.edu/coca/

中英双语知识本体词网：http://bow.ling.sinica.edu.tw/

Sentiment140 http://help.sentiment140.com/for-students/

WordNet https://wordnet.princeton.edu/

维基百科语料库 http://nlp.cs.nyu.edu/wikipedia-data/

古滕堡语料库：http://www.gutenberg.org/

1. 概率论基础

- 最大似然估计
- 条件概率
- 二项式分布
- 贝叶斯公式
- 贝叶斯决策理论
- 期望和方差

2. 信息论基本概念

- 熵
- 联合熵
- 相对熵
- 交叉熵
- 困惑度

3. 语料库

- 语料库语言学的基本定义
- 语料库类型
- 语料库建设中的基本问题

小结

Thank you !