自然语言理解

第二章 数学基础

宗成庆

中科院自动化研究所

模式识别国家重点实验室

cqzong@nlpr.ia.ac.cn

http://www.nlpr.ia.ac.cn/English/cip/cqzong.htm

第二章　数学基础

2.1 内容回顾

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：Chomsky 文法理论

- 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

2.2 概率论基础

❑ 概率（probability）

概率是从随机试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： $P(A) \geq 0$

公理2： $P(\Omega) = 1$

公理3：如果对任意的 i **和** j $(i \neq j)$**，事件** A_i **和** A_j **不相交** $(A_i \cap A_j = \Phi)$**，则有：**

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

2.2 概率论基础

❑ 最大似然估计(likelihood estimation)

如果一个试验的样本空间是$\{s_1, s_2, \ldots, s_n\}$，在相同情况下重复实验 N 次，观察到样本 s_k 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越来越大时，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，

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

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。

2.2 概率论基础

❑ 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 **的概率。一般地，**P(A|B) ≠ P(A).

2.2 概率论基础

❑ 全概率公式

设 Ω 为试验E的样本空间，$B_1, B_2, \ldots B_n$ 为Ω的一组事件，且他们两两互斥，且每次试验中至少发生一个。即：

(1) $B_i \cap B_j = \Phi \quad (i \neq j, i, j = 1, 2, \ldots, n)$

(2) $\bigcup_{i=1}^{n} B_i = \Omega$

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

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

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

2.2 概率论基础

❑ 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** $P(A) > 0$，$P(B_i) > 0$ ($i = 1, 2, \ldots, n$) **，那么**

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

2.2 概率论基础

例2-1：给定语音信号 A，找出对应的语句S，使得 P(S|A) 最大，那么，

$$\hat{s} = \arg\max_{s} P(S \mid A)$$

根据贝叶斯公式，

$$\hat{s} = \arg\max_{s} \frac{P(S)P(A \mid S)}{P(A)}$$

由于 P(A) 在 A 给定时是归一化常数，因而，

$$\hat{s} = \arg\max_{s} P(A \mid S)P(S)$$

语音识别问题

语音模型

语言模型

2.2 概率论基础

例2-2：假设某一种特殊的句法结构很少出现，平均大约每100000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？

2.2 概率论基础

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：P(G|T)= ？

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

$$=\frac{0.95\times 0.00001}{0.95\times 0.00001\ +0.005\times 0.99999}\approx 0.002$$

2.2 概率论基础

❑ 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）的试验（伯努利试验），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，…，n。

考虑事件$\{X=i\}$，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$\overline{A}$。

$$\underbrace{A\,\overline{A}\,AA\,\cdots\,\overline{A}}_{n\text{个}} \Longrightarrow p^{i}(1-p)^{n-i}$$

2.2 概率论基础

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$\binom{n}{i}$种可能。由此，可以得出：

$$p_i = \binom{n}{i} p^i (1-p)^{n-i}, \qquad i = 0,1,\cdots,n$$

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

X所遵从的概率分布（2-13）称为二项式分布并记为：

$X \sim B(n, p)$

2.2 概率论基础

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。

2.2 概率论基础

❑ 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w_i表示，$i = 1, 2, \cdots, c$；对应于各个类别w_i出现的先验概率为$P(w_i)$；在特征空间已经观察到某一向量 $\bar{x} = [x_1, x_2, \cdots, x_d]^T$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

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

2.2 概率论基础

基于最小错误率的贝叶斯决策规则为：

(1) **如果** $P(w_i \mid \bar{x}) = \max_{j=1,2,\cdots,c} P(w_j \mid \bar{x})$ **则** $x \in w_i$

(2) **或者：如果** $p(\bar{x} \mid w_i)P(w_i) = \max_{j=1,2,\cdots,c} p(\bar{x} \mid w_j)P(w_j)$ **则** $x \in w_i$

(3) **或者**(c=2**时**) **：如果** $l(\bar{x}) = \dfrac{p(\bar{x} \mid w_1)}{p(\bar{x} \mid w_2)} > \dfrac{P(w_2)}{P(w_1)}$ **则** $x \in w_1$

否则 $x \in w_2$ 。

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

2.2 概率论基础

❑ 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 $\mathrm{P}(X = \mathrm{x}_k) = \mathrm{p}_k$，$k = 1,2, \ldots$，若级数 $\sum_{k=1}^{\infty} x_k p_k$ 绝对收敛，那么，随机变量 X 的数学期望或概率平均值为：

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

2.2 概率论基础

❑ 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：

$$Var(X) = E((X - E(X))^2)$$

$$= E(X^2) - E^2(X)$$

2.3 信息论基础

❑ 熵（entropy）

香农（Claude Elwood Shannon）于1940年获得麻省理工学院数学博士学位和电子工程硕士学位后，于1941年加入了贝尔实验室数学部，并在那里工作了15年。1948年6月和10月，由贝尔实验室出版的《贝尔系统技术》杂志连载了香农博士的文章《通讯的数学原理》，该文奠定了香农信息论的基础。

熵是信息论中重要的基本概念。

2.3 信息论基础

如果 X 是一个离散型随机变量，其概率分布为：

$p(x) = P(X = x)$，$x \in X$。X 的熵 $H(X)$ 为：

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

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

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。

2.3 信息论基础

例2-3 计算英文（26个字母和空格，共27个字符）信息源的熵：

（1）假设27个字符等概率出现；

（2）英文字母的概率分布如下：

字母	空格	E	T	O	A	N	I	R	S
概率	0.1956	0.105	0.072	0.0654	0.063	0.059	0.055	0.054	0.052

字母	H	D	L	C	F	U	M	P	Y
概率	0.047	0.035	0.029	0.023	0.0225	0.0225	0.021	0.0175	0.012

字母	W	G	B	V	K	X	J	Q	Z
概率	0.012	0.011	0.0105	0.008	0.003	0.002	0.001	0.001	0.001

2.3 信息论基础

解：（1）等概率出现情况：

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

$$= 27 \times \{-\frac{1}{27} \log \frac{1}{27}\} = \log 27 = 4.75 \quad \text{(bits/letter)}$$

（2）实际情况：

$$H(X) = -\sum_{i=1}^{27} p(x_i) \log p(x_i) = 4.02 \quad \text{(bits/letter)}$$

说明：考虑了英文字母和空格实际出现的概率后，英文信源的平均不确定性，比把字母和空格看作等概率出现时英文信源的平均不确定性要小。

2.3 信息论基础

❑ 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：

$$H(X,Y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log_2 p(x,y) \qquad (1)$$

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 信息论基础

❑ 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：

$$H(Y \mid X) = \sum_{x \in X} p(x) H(Y \mid X = x) \qquad (2)$$

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

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

2.3 信息论基础

将 (1)式 中的 $\log_2 p(x,y)$ 根据条件概率公式展开，有：

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

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

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

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

$$= H(X) + H(Y\mid X) \qquad (3)$$

(**连锁规则**)

2.3 信息论基础

❑ 互信息（mutual information）

如果 (X, Y) ~ p(x, y)，X, Y **之间的互信息** I(X; Y) **为：**

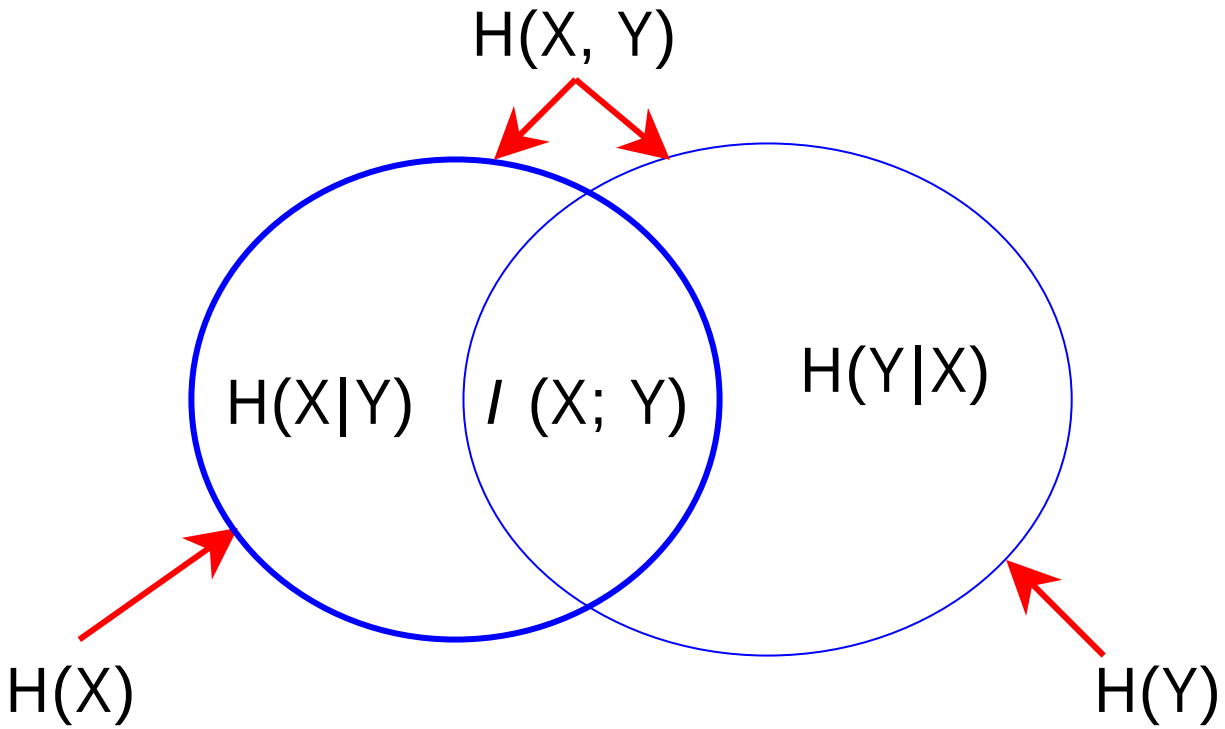
$$I(X;Y) = H(X) - H(X \mid Y)$$

根据定义，展开 H(X) **和** H(X|Y) **容易得到：**

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

互信息I (X; Y) **是在知道了** Y **的值后** X **的不确定性的减少量。即，** Y **的值透露了多少关于** X **的信息量。**

2.3 信息论基础



2.3 信息论基础

由于 $H(X, X) = 0$, **所以，**

$$H(X) = H(X) - H(X|X) = I(X; X)$$

这一方面说明了为什么熵又称自信息，另一方面说明了两个完全相互依赖的变量之间的互信息并不是一个常量，而是取决于它们的熵。

2.3 信息论基础

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

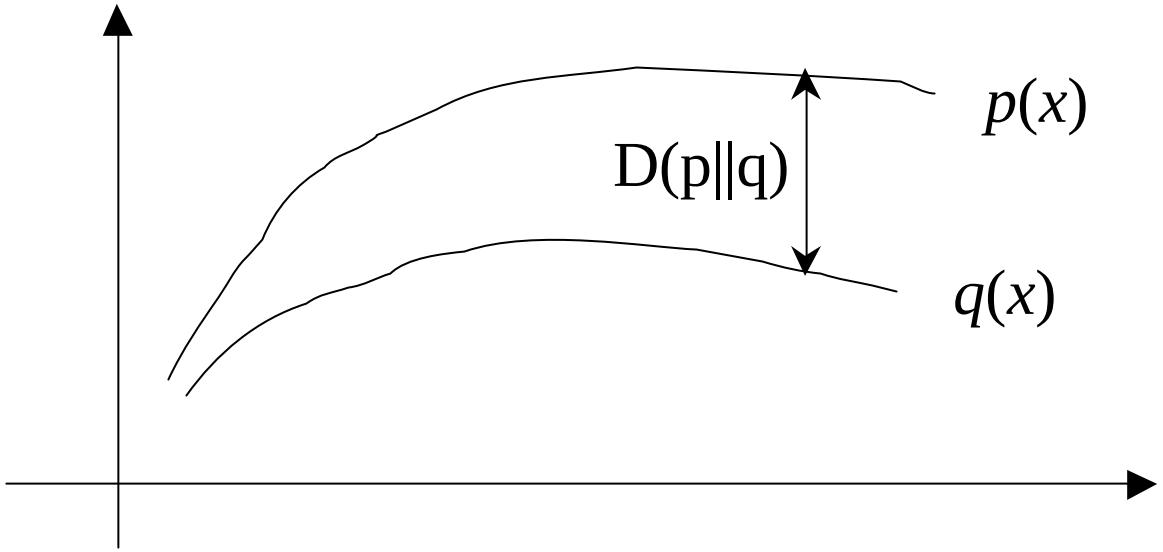
两个概率分布 p(x) **和** q(x) **的相对熵（或** Kullback-Leibler **距离，简称** KL **距离）定义为：**

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

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

2.3 信息论基础

相对熵常被用以衡量两个相对随机分布的差距。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。



2.3 信息论基础

❑ 交叉熵（cross entropy）

如果一个随机变量 $X \sim p(x)$，$q(x)$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：

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

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。

2.3 信息论基础

对于语言 $L = (X_i) \sim p(x)$ 与其模型q的交叉熵定义为 ：

$$H(L,q) = -\lim_{n\to\infty}\frac{1}{n}\sum_{x_1^n} p(x_1^n)\log q(x_1^n)$$

其中，$x_1^n = x_1,\cdots,x_n$ 为语言L的语句，$p(x_1^n)$为L中语句的概率，$q(x_1^n)$为模型q对 x_1^n 的概率估计。

我们可以假设这种语言是“理想”的，即n趋于无穷大时，其全部“单词”的概率和为1。也就是说，根据信息论的定理：假定语言L是稳态(stationary) ergodic随机过程，L与其模型q的交叉熵计算公式就变为：

2.3 信息论基础

$$H(L,q) = -\lim_{n\to\infty}\frac{1}{n}\log q(x_1^n)$$

由此，我们可以根据模型q和一个含有大量数据的L的样本来计算交叉熵。在设计模型q时，我们的目的是使交叉熵最小，从而使模型最接近真实的概率分布$p(x)$。

2.3 信息论基础

❑ 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 $l_1^n = l_1 \cdots l_n$ ，

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。

2.3 信息论基础

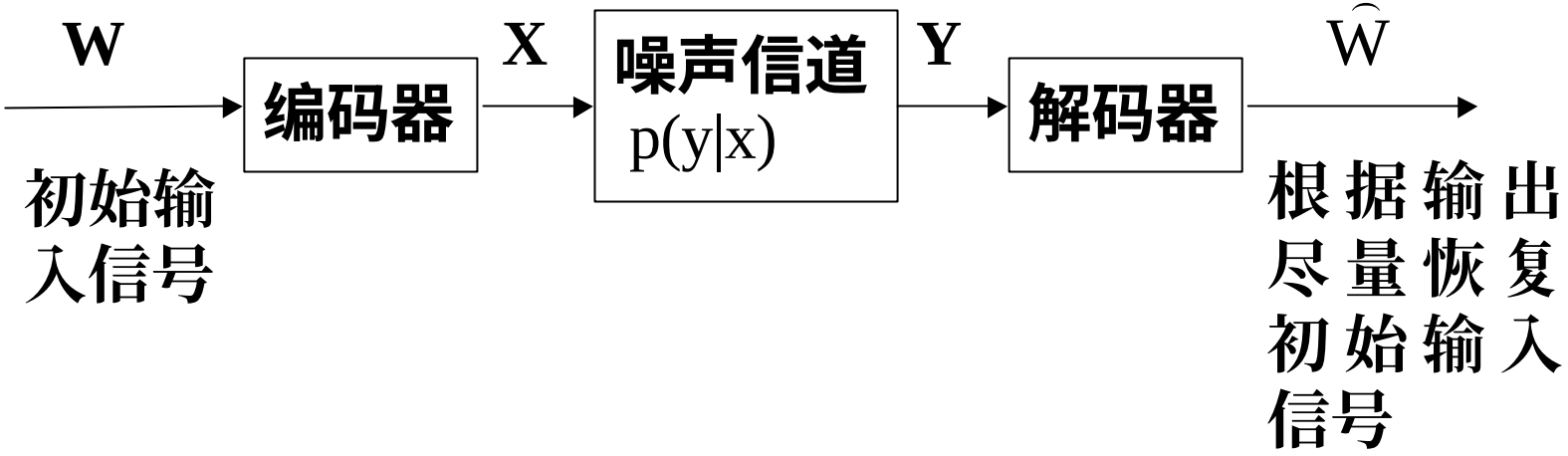
❑ 噪声信道模型（noisy channel model）

在信号传输的过程中都要进行双重性处理：一方面要通过压缩消除所有的冗余，另一方面又要通过增加一定的可控冗余以保障输入信号经过噪声信道后可以很好的恢复原状。这样的话，信息编码时要尽量占有少量的空间，但又必须保持足够的冗余以便能够检测和校验错误。而接收到的信号需要被解码使其尽量恢复到原始的输入信号。

噪声信道模型的目标就是优化噪声信道中信号传输的吞吐量和准确率，其基本假设是一个信道的输出以一定的概率依赖于输入。

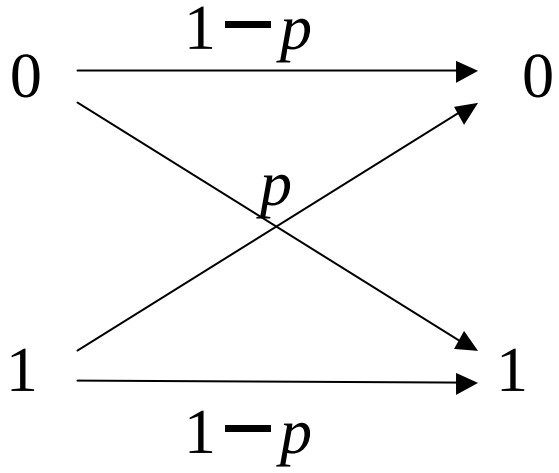
2.3 信息论基础

过程示意图：



2.3 信息论基础

一个二进制的对称信道(binary symmetric channel, BSC)的输入符号集X:{0, 1}，输出符号集Y:{0, 1}。在传输过程中如果输入符号被误传的概率为 p，那么，被正确传输的概率就是 $1-p$。这个过程我们可以用一个对称的图型表示如下：



2.3 信息论基础

信息论中很重要的一个概念就是信道容量（capacity），其基本思想是用降低传输速率来换取高保真通讯的可能性。其定义可以根据互信息给出：

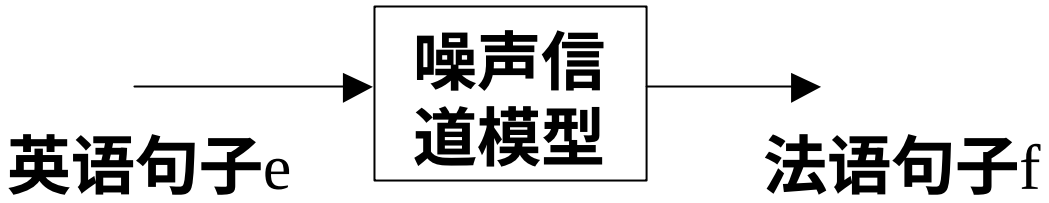
$$C = \max_{p(X)} I(X;Y)$$

根据这个定义，如果我们能够设计一个输入编码X，其概率分布为p(X)，使其输入与输出之间的互信息达到最大值，那么，我们的设计就达到了信道的最大传输容量。

2.3 信息论基础

在自然语言处理中，我们不需要进行编码，只需要进行解码，使系统的输出更接近于输入。

例如，法语翻译成英语：



根据贝叶斯公式：

$$P(e|f)=\frac{P(e)P(f|e)}{P(f)}$$

2.3 信息论基础

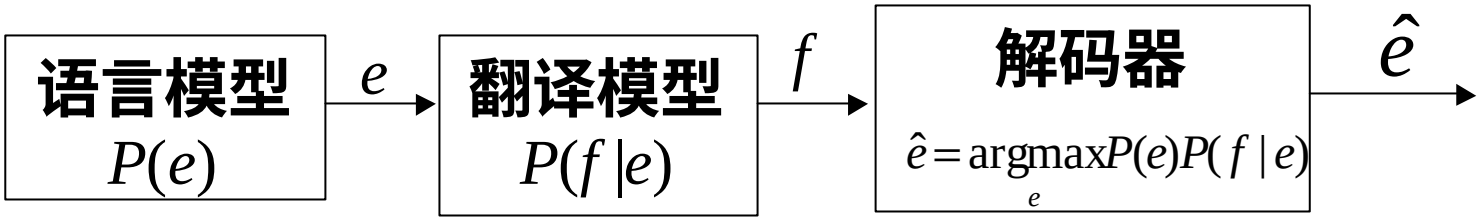
由于右边分母$P(f)$与e无关，因此，求该式的最大值相当于寻找一个使得右边分子的两项乘积 $P(e) \times P(f|e)$ 最大，即：

$$\hat{e} = \arg\max_{e} P(e)P(f \mid e)$$

- $P(e)$称为目标语言的语言模型（language model, LM）
- $P(f|e)$为给定e的情况下f的翻译概率，也称翻译模型(translation model)

2.3 信息论基础

统计翻译的噪声信道模型：



也就是说，如果我们要建立一个源语言f到目标语言e的统计翻译系统，我们必须解决三个关键的问题：

（1）估计语言模型概率$P(e)$；

（2）估计翻译概率$P(f|e)$；

（3）设计有效快速的搜索算法求解 $\hat{e}$ 使得$P(e)\times P(f|e)$最大。

本章小结

❑概率论基础

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

❑信息论基本概念

- ➢ 熵
- ➢ 互信息
- ➢ 交叉熵
- ➢ 噪声信道模型
- ➢ 联合熵
- ➢ 相对熵
- ➢ 困惑度

习题

2-1. 任意摘录一段文字，统计这段文字中所有字符的相对频率。假设这些相对频率就是这些字符的概率，请计算其分布的熵。

2-2. 任意取另外一段文字，按上述同样的方法计算字符分布的概率，然后计算两段文字中字符分布的 KL 距离。

2-3. 举例说明（任意找两个分布 p 和 q），KL 距离是不对称的，即 $D(p \| q) \neq D(q \| p)$。

2-4. 设 $X \sim p(x)$，$q(x)$ 为用于近似 $p(x)$ 的一个概率分布，则 $p(x)$ 与 $q(x)$ 的交叉熵定义为 $H(p, q) = H(p) + D(p \| q)$。请证明：

$$H(p,q) = -\sum_{x} p(x) \log q(x)$$

Thanks

谢谢!