1. 概率论

语言是稳态的，可遍历的随机过程.

* 稳态的：即随机过程的概率分布只与时间长度相关.

$$f(x_1, \cdots, x_n, t_1, \cdots, t_n) = f(x_1, \cdots, x_n, t_1+\Delta t, \cdots, t_n+\Delta t).$$

语言在短时间内数学特征应保持相同.

* 可遍历的：即单个随机变量在长时间下可以遍历其所有可能取值

一个人长时间不断书写语言 $\Leftrightarrow$ 多人短时间书写语言.

2. 齐夫定律

将大规模语言数据统计词频，并按从高到低排序，设某个词 w 的词频为 f，排在第 r 位，则 $fxr \longrightarrow C$，C 为一个常数.

3. 信息论

* 熵：$H(x) = -\sum_{x \in X} p(x) \log_2 p(x)$　　$0 \log 0 = 0$

描述随机变量不确定性

* 联合熵：$H(x,y) = -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log_2 p(x,y)$.

一对随机变量平均所需要的信息量

* 条件熵：$H(Y|X) = \sum_{x \in X} p(x) H(Y|X=x)$

$$= \sum_{x \in X} p(x) \left[-\sum_{y \in Y} p(y|x) \log_2 p(y|x) \right]$$

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

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

$$= -\sum_{x \in X} \sum_{y \in Y} p(x,y) \left[\log_2 p(y|x) + \log_2 p(x) \right]$$

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

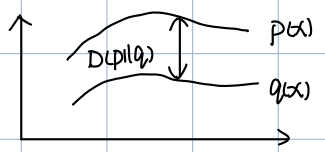
$$= H(Y|X) - \sum_{x\in X} p(x)\log_2 p(x)$$

$$= H(Y|X) + H(X).$$

* 相对熵 (K-L 散度) : $p(x)$, $q(x)$ 为两个概率分布.

$$D(p\|q) = \sum_{x\in X} p(x)\log\frac{p(x)}{q(x)} \qquad 0\log\frac{0}{q}=0 \ , \ p\log\frac{p}{0}=\infty$$

用于衡量两个随机分布的差距.



* 交叉熵 : $X\sim p(x)$. $q(x)$ 是 $p(x)$ 的近似, 则交叉熵定义为:

$$H(X,q) = D(p\|q) + H(X)$$

$$= \sum_{x\in X} p(x)\log\frac{p(x)}{q(x)} - \sum_{x\in X} p(x)\log p(x).$$

$$= -\sum_{x\in X} p(x)\log q(x).$$

交叉熵 $\Leftrightarrow$ 相对熵. [由于 $p(x)$ 为真实数据分布, 不会改变, 故两者只是数值上不同]

语言 $L=(X)\sim p(x)$. 理论模型 q. 交叉熵定义为:

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

$x_1^n = x_1,\cdots,x_n$ 为语言 L 的样本

若 L 是稳态的、可遍历的, 则

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

* 困惑度: $x_1^n = x_1\cdots x_n$ 是语言 L 的样本. 则 L 的困惑度:

$$PP_q = 2^{H(L,q)} = 2^{-\frac{1}{n}\log q(x_1^n)} = [q(x_1^n)]^{-\frac{1}{n}}$$

衡量语言模型的好坏

* 互信息 $(X,Y)\sim p(x,y)$. 则 X,Y 之间的互信息定义:

$$I(X,Y) = H(X) - H(X|Y).$$

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

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

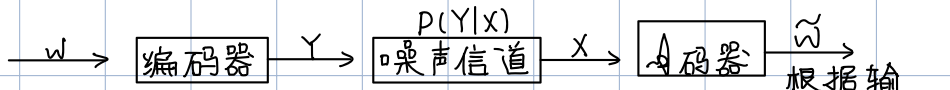
$$= \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{p(x,y)}{p(y)p(x)}$$

描述 X 在给定 Y 后不确定性减少的量.

$H(X) = H(X) - H(X|X) = I(X,X)$.

用互信息估计两个汉字结合紧密程度. 进而分词. 效果不好。但个字可以组成多种词。

4. 噪声信号模型



机器翻译: X: 源语言, Y: 目标语言. 解码 $p(Y|X)$.

4. 应用举例:

目标: 歧义消解.

1) 基于贝叶斯分类器

多义词 w, 上下文语境为 C, w 的多个语义记为 S_i $(i\geq 2)$.

则在 C 中 w 的语义为: $\arg\max_{S_i} p(S_i|C) \Leftrightarrow \arg\max_{S_i} p(C|S_i)p(S_i)$

而 $p(S_i|C) = \frac{p(C|S_i)\,p(S_i)}{p(C)}$ → 常数

$p(C|S_i) = \prod_{V_k\in C} p(V_k|S_i)$, V_k 为 w 的上下文 (如 w 前后±2词)

$p(V_k|S_i) = \frac{N(V_k,S_i)}{N(S_i)}$

$p(S_i) = \frac{N(S_i)}{N(w)}$

实际操作中. 通常将计算 $p(C|S_i)p(S_i) \Rightarrow \log p(S_i) + \sum_{V_k\in C}\log p(V_k|S_i)$

2) 基于最大熵的消歧办法.

基本原理: 在只掌握关于未知分布的部分知识的情况下, 将合已知知识的概率分布可能有多个, 取熵值最大的分布.

特征函数: $x\in X$, $y\in Y$. $f(x,y) = \begin{cases} 1, & (x,y) \text{ 满足某种条件} \\ 0, & \text{否则} \end{cases}$

模型定义：给定数据集 $T=\{(x_1,y_1),\cdots,(x_N,y_N)\}$

目标：$\min_{p\in C} -H(p)=\sum_{i=1}^{N}\tilde{p}(x_i)p(y_i|x_i)\log p(y_i|x_i)$

即寻求在给定 x 下，条件熵最大的分类器

约束：① $E_{\tilde{p}}(f_i)=E_p(f_i)$. $i=1,\cdots,K$.

其中：$E_{\tilde{p}}(f_j)=\sum_{i=1}^{N}\tilde{p}(x_i,y_i)f_j(x_i,y_i)\triangleq\tau_j$

$E_p(f_j)=\sum_{i=1}^{N}\tilde{p}(x_i)p(y_i|x_i)f_j(x_i,y_i)$

即满足特征函数期望保持不变

② $\sum_y p(y|x)=1$.

模型求解：上式模型最优解：$p^*(y|x)=\frac{1}{Z(x)}\exp\left(\sum_{j=1}^{K}\lambda_j\cdot f_j(x,y)\right)$

其中 Z 是归一化常量，λ_j 为某个固定的数

特征函数 $f(x,y)$ 确定：取 x 为上下文条件 y 为多义词词义.

上下文条件：* 表示：① 词形信息（字词）

② 词性信息（词性）

③ 词形 + 词性.

* 顺序：① 位置有关

② 位置无关.

* 窗口大小：前后五个词.

参数入（GIS 算法）：

* 要求每个实例的 K 个特征和为常量 C：$\sum_{j=1}^{K}f_j(x,y)=C$

若不满足，取 $C=\max_{x,y}\sum_{j=1}^{K}f_j(x,y)$，构造一个新特征

$f_{K+1}(x,y)=C-\sum_{j=1}^{K}f(x,y)$.

算法迭代：

① 初始化：$\lambda=0$

② 计算 $E_{\tilde{p}}f_j(x,y)=\sum_{i=1}^{N}\tilde{p}(x_i,y_i)f_j(x_i,y_i)$

③ 迭代计算 $E_p f_j(x,y)$.

$$Z(x) = \sum_y \exp\left(\sum_{j=1}^{k} \lambda_j f_j(x,y)\right)$$

$$p^*(y|x) = \frac{1}{Z(x)} \exp\left(\sum_{j=1}^{k} \lambda_j f_j(x,y)\right)$$

$$E_p(f_j) = \sum_{i=1}^{N} \tilde{p}(x_i)\, p^*(y_i|x_i)\, f_j(x_i,y_i)$$

若 超过迭代次数

$|L_{i+1} - L_i| < \varepsilon$, $L(p) = \sum_{i=1}^{N} \tilde{p}(x_i,y_i) \log p(y_i|x_i)$

则停止，否则 $\lambda^{(n+1)} = \lambda^{(n)} + \frac{1}{C} \ln\left\{\frac{E_{\tilde{p}} f}{E_{p^{(n)}} f}\right\}$

④ 确定 λ，算出 p^*.