自然语言处理

Natural Language Processing

权小军 教授

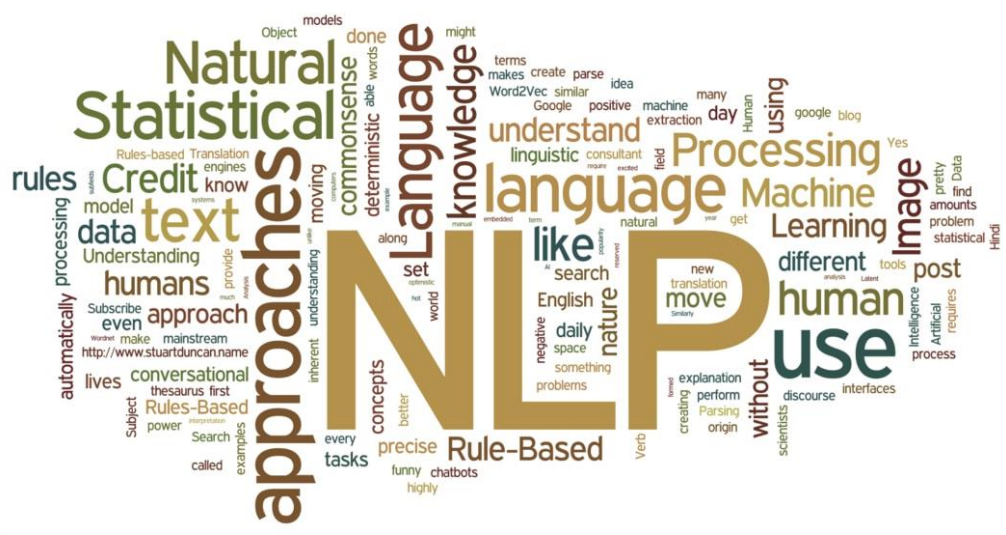
中山大学数据科学与计算机学院

quanxj3@mail.sysu.edu.cn

课程回顾

自然语言处理(NLP)

自然语言处理（自然语言理解），是计算机科学与人工智能领域中的一个重要方向。它研究能实现人与计算机之间通过自然语言进行交互的各种理论和方法.



自然语言处理(NLP)

自然语言处理的研究方向包括:

- 中文自动分词
- 句法分析
- 信息抽取
- 情感计算
- 机器翻译
- 对话系统
- 信息检索
- 自动摘要

Why NLP?

- 语言是思维的载体，是人类交流思想、表达情感最自然、最直接、最方便的工具
- 人类历史上以语言文字形式记载和流传的知识占知识总量的80％以上

基本概念

关于“理解”的标准

- 如何判断计算机系统的智能?

  计算机系统的表现(act)如何?

  反应(react)如何?

  相互作用(interact)如何?

  ⇩

  与有意识的个体（人）比较如何?

图灵设计的“模仿游戏” – 图灵实验(Turing test)

基本问题

- 基本研究问题之一：形态学 (Morphology)
  - 研究词(word) 由有意义的基本单位－词素的构成问题。
  - 单词的识别/ 汉语的分词问题。
- 基本研究问题之二：句法 (Syntax) 问题
  - 研究句子结构成分之间的相互关系和组成句子序列的规则
  - 为什么一句话可以这么说也可以那么说？如何建立快速有效的句子结构分析方法？

基本问题

- **基本研究问题之三：语义(Semantics) 问题**
  - 研究如何从一个语句中推导出词的意义，以及这些词在该语句句法结构中的作用来推导出该语句的意义。

- **基本研究问题之四：语用学(Pragmatics) 问题**
  - 研究在不同上下文中语句的应用，以及上下文对语句理解所产生的影响。

主要困难

- 困难之一：大量歧义(ambiguity)现象
  - I. 词法歧义
  - II. 词性歧义
  - III. 结构歧义
  - IV. 语义歧义
  - V. 语音歧义

主要困难

- **困难之二：大量未知语言现象**
  - 新词、人名、地名、术语等
  - 新含义
  - 新用法和新句型等，尤其在口语中或部分网络语言中，不断出现一些“非规范的”新的语句结构

2.1 概率论基础

概率论基础

◆基本概念

- 概率(probability)
- 极大似然估计(maximum likelihood estimation)
- 条件概率(conditional probability)
- 全概率公式(full probability)
- 贝叶斯法则(Bayes' theorem)
- 二项式分布(binomial distribution)
- 期望(expectation)
- 方差(variance)

极大似然估计(MLE)

一个试验的样本空间是$\{s_1, s_2, \ldots, s_n\}$,在相同情况下重复试验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：

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

$$\because \sum_{k=1}^{n} n_N(s_k) = N, \quad \therefore \sum_{k=1}^{n} q_N(s_k) = 1$$

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$P(s_k)$：$\lim_{N\to\infty} q_N(s_k) = P(s_k)$，因此相对频率常被用作概率的估计值，这种估计方法称为最大似然估计。

现代汉语字频统计结果：前20个最高频汉字及其频率

汉字	频率	汉字	频率	汉字	频率	汉字	频率
的	0.040855	了	0.008470	中	0.006012	国	0.005406
一	0.013994	有	0.008356	大	0.005857	我	0.005172
是	0.011758	和	0.007297	为	0.005720	以	0.005117
在	0.010175	人	0.006821	上	0.005705	要	0.004824
不	0.009034	这	0.006557	个	0.005488	他	0.004685

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：

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

一般地，$P(A|B)\neq P(A)$。

例子

- 当预测“大学”一词出现的概率时，如果已经知道出现在它前面的两个词是“上海”和“交通”，“大学”一词出现的概率会大大增加。

全概率公式

设Ω为实验的样本空间，$B_1, B_2, \ldots, B_n$为Ω的一组两两互斥的事件，且每次试验中至少发生一个，则称$B_1, B_2, \ldots, B_n$为样本空间Ω的一个划分。

如果A为样本空间Ω的事件，$B_1, B_2, \ldots, B_n$为样本空间Ω的一个划分，且$P(B_i) > 0$（$i = 1, 2, \ldots, n$），则**全概率公式**为：

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

贝叶斯定理

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

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

当 $n = 1$ 时，$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)}$

先验概率、后验概率

- 先验概率(Prior probability):不考虑先决条件（信息或者知识）而得到的该事件的概率：一般在试验前已知，常常是以往经验的总结
- 后验概率(Posterior probability):在具备该事件出现的信息或者知识的条件下得到的该事件的概率：反映了试验之后对各种原因发生的可能性大小的新知识

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：

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

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

求解： $P(G\mid T)=?$

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

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

二项式分布(binomial distribution)

当重复一个只有两种输出（假定为 $\overline{A}$ 和 A ）的试验（伯努利试验），A 在一次实验中发生的概率为 p，现将实验独立地重复 n 次，如果用 X 表示 A 在这 n 次实验中发生的次数，那么，$X = 0,1,\cdots,n$。则 n 次独立实验中成功的次数为 r 的概率为：$p_r = C_n^r p^r (1-p)^{n-r}$，其中，$C_n^r = \dfrac{n!}{(n-r)!r!}$，$0 \le r \le n$。此时 X 所遵从的概率分布称为二项式分布，并记为：$X \sim B(n,p)$。

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 $P(X=x_k)=p_k,\quad k=1,2,\ldots,$

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：$E(X)=\sum_{k=1}^{\infty} x_k p_k$ 。

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：

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

2.2 信息论基础

信息论基础

❑ 熵(entropy)

香农(Claude Elwood Shannon)于1940年获得MIT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

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：

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

其中，约定 $0 \log 0 = 0$

通常熵的单位为二进制位比特 (bit)

信息论基础

熵又称为自信息(self-information)，表示信源X每发一个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。

信息论基础

例2-1： 计算下列两种情况下英文(26个字母和1个空格，共27个字符)信息源的熵：

1) 假设27个字符等概率出现；

2) 假设英文字母的概率分布如下：

信息论基础

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

信息论基础

解:（1）等概率出现情况:

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

$$=27\times\{-\frac{1}{27}\log_2\frac{1}{27}\}=\log_2 27=4.75 \quad (\text{bits})$$

（2）实际情况:

$$H(X)=-\sum_{i=1}^{27}p(x_i)\log_2\ p(x_i)=4.02 \quad (\text{bits})$$

信息论基础

解：（1）等概率出现情况：

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

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。

$$H(X) = -\sum_{i=1}^{27} p(x_i) \log_2 p(x_i) = 4.02 \quad (\text{bits/letter})$$

信息论基础

- 法语、意大利语、西班牙语、英语、俄语字母的熵 (冯志伟，1989)

语言	熵 (bits)
法语	3.98
意大利语	4.00
西班牙语	4.01
英语	4.03
俄语	4.35

信息论基础

- 法语、意大利语、西班牙语、英语、俄语字母的熵 (冯志伟，1989)

语言	熵 (bits)
法语	3.98
意大利语	4.00
西班牙语	4.01
英语	
俄语	4.35

英语词的熵约为10 bits

信息论基础

- 中文?

1970年代末期冯志伟教授首先开展了对汉字信息熵的研究，经过几年的文本收集和手工统计，在当时艰苦的条件下测定了汉字的信息熵为9.65比特(bit)。1980年代末期，刘源等测定了汉字的信息熵为9.71 比特，而汉语词的熵为11.46比特。

信息论基础

- 联合熵(joint entropy)

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

$$H(X,Y) = -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log_2 p(x,y) \qquad (2)$$

信息论基础

❑ 联合熵(joint entropy)

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

$$H(X,Y) = -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log_2 p(x,y) \qquad (2)$$

联合熵是描述一对随机变量平均所需要的信息量

信息论基础

❑ 条件熵(conditional entropy)

给定随机变量X，随机变量Y的条件熵定义为:

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

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

$$= -\sum_{x \in X} \sum_{y \in Y} p(x, y) \log_2 p(y \mid x) \quad (3)$$

信息论基础

例2-2：假设(X, Y)服从如下联合概率分布：

Y \ X	1	2	3	4
1	1/8	1/16	1/32	1/32
2	1/16	1/8	1/32	1/32
3	1/16	1/16	1/16	1/16
4	1/4	0	0	0

请计算 $H(X)$、$H(Y)$ 、$H(X|Y)$、 $H(Y|X)$ 和 $H(X, Y)$

信息论基础

- 相对熵(relative entropy, 或称 Kullback-Leibler divergence, KL 距离)

  两个概率分布 $p(x)$ 和 $q(x)$ 的相对熵定义为:

$$D(p \| q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)} \tag{4}$$

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

信息论基础

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。

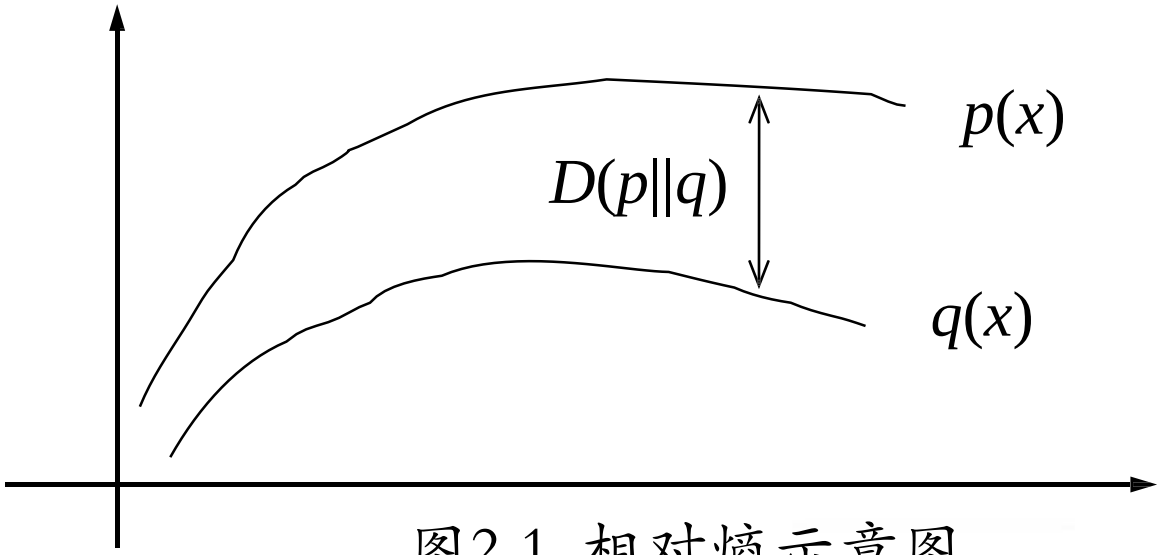


图2-1. 相对熵示意图

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：

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

$$= -\sum_{x} p(x) \log q(x) \tag{5}$$

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：

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

$$= -\sum_{x} p(x)\log q(x) \tag{5}$$

交叉熵用以衡量估计模型与真实概率分布之间的差异

信息论基础

❑ 互信息(mutual information)

如果 $(X, Y) \sim p(x, y)$，X, Y 之间的互信息 $I(X; Y)$ 定义为：

$$I(X; Y) = H(X) - H(X|Y) \qquad (6)$$

根据 $H(X)$ 和 $H(X|Y)$ 的定义：

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

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

信息论基础

$$
\begin{aligned}
I(X;Y) &= H(X) - H(X \mid Y) \\
&= -\sum_{x \in X} p(x) \log_2 p(x) + \sum_{x \in X} \sum_{y \in Y} p(x,y) \log_2 p(x \mid y) \\
&= \sum_{x \in X} \sum_{y \in Y} p(x,y) \left(\log_2 p(x \mid y) - \log_2 p(x) \right) \\
&= \sum_{x \in X} \sum_{y \in Y} p(x,y) \left(\log_2 \frac{p(x \mid y)}{p(x)} \right)
\end{aligned}
$$

$$
I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)} \tag{7}
$$

信息论基础

$$
\begin{aligned}
I(X;Y) &= H(X) - H(X \mid Y) \\
&= -\sum_{x \in X} p(x) \log_2 p(x) + \sum_{x \in X} \sum_{y \in Y} p(x,y) \log_2 p(x \mid y) \\
&= \sum_{x \in X} \sum_{y \in Y} p(x,y) \left(\log_2 p(x \mid y) - \log_2 p(x) \right) \\
&= \sum_{x \in X} \sum_{y \in Y} p(x,y) \left(\log_2 \frac{p(x \mid y)}{p(x)} \right)
\end{aligned}
$$

$$
I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)} \tag{7}
$$

信息论基础

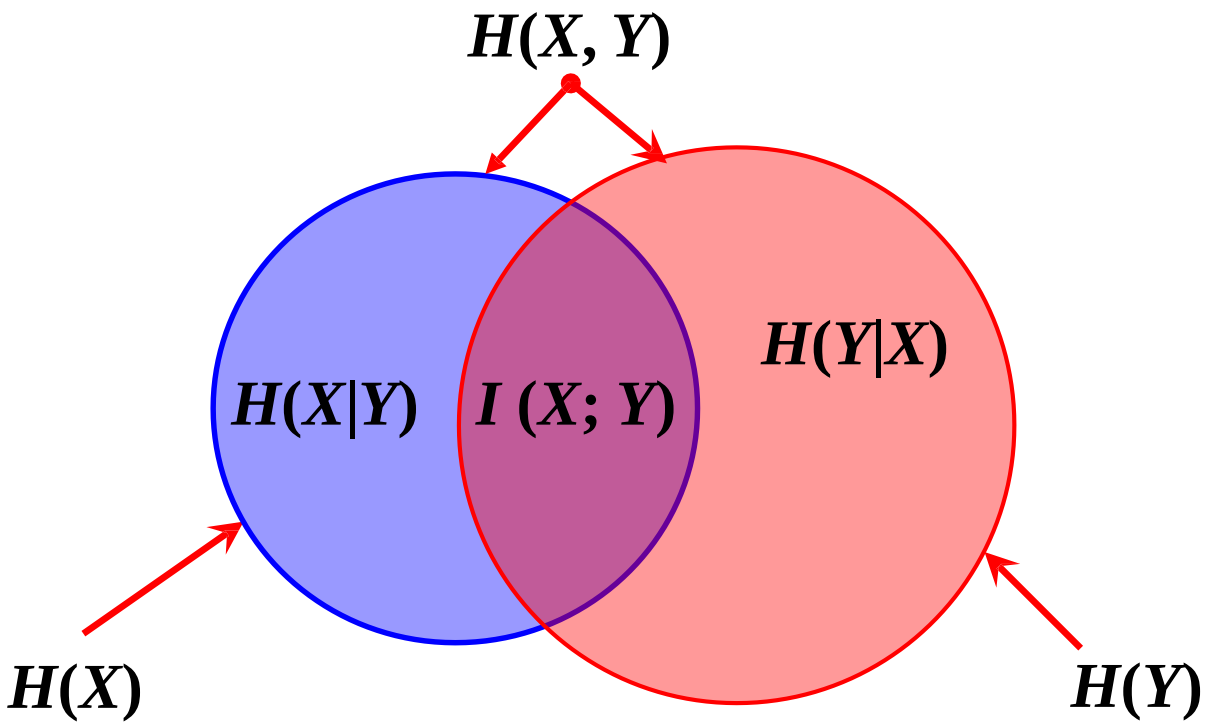


图. 互信息、条件熵与联合熵

信息论基础

例如：汉语分词问题

中文分词：为人民服务

为人// 民//服务 或者 为// 人民// 服务

信息论基础

例如：汉语分词问题

中文分词：为人民服务

为人// 民//服务 或者 为// 人民// 服务

- 利用互信息值估计两个汉字结合的强度：

$$I(x;y)=\log_2\frac{p(x,y)}{p(x)p(y)}=\log_2\frac{p(y\mid x)}{p(y)}$$

信息论基础

例如：汉语分词问题

中文分词：为人民服务

为人// 民//服务 或者 为// 人民// 服务

- 利用互信息值估计两个汉字结合的强度：

互信息值越大，表示两个汉字之间的结合越紧密，越可能成词。反之，断开的可能性越大。

2.3 应用实例

应用实例

例1： 词汇歧义消解

❖ 基本思路

任何一种自然语言中，一词多义（歧义）现象是普遍存在的。如何区分不同上下文中的词汇语义，就是词汇歧义消解问题，或称词义消歧(word sense disambiguation)。

词义消歧是自然语言处理中的基本问题之一.

应用实例

例如：

1) 他打鼓很在行。
2) 他会打家具。
3) 他把碗打碎了。
4) 他在学校打架了。
5) 他很会与人打交道。
6) 他用土打了一堵墙。
7) 用面打浆糊贴对联。
8) 他打铺盖卷儿走人了。
9) 她会用毛线打毛衣。
10) 他用尺子打个格。
11) 他打开了箱子盖。
12) 她打着伞走了。
13) 他打来了电话。
14) 他打了两瓶水。
15) 他想打车票回家。
16) 他以打鱼为生。

应用实例

❖ 解决思路

每个词表达不同的含意时其上下文（语境）往往不同，也就是说，不同的词义对应不同的上下文，因此，如果能够将多义词的上下文区别开，其词义自然就明确了。

应用实例

❖ 解决思路

每个词表达不同的含意时其上下文（语境）往往不同，也就是说，不同的词义对应不同的上下文，因此，如果能够将多义词的上下文区别开，其词义自然就明确了。

他/P 很/D 会/V 与/C 人/N 打/V 交道/N 。/PU

-2 -1 0 +1 +2

基本的上下文信息：词、词性、位置

应用实例

□ 基于上下文分类的消歧方法

(1) 基于贝叶斯分类器 (Gale *et al.*, 1992)

数学描述:

假设某个多义词 w 所处的上下文语境为 C,如果w 的多个语义记作 s_i, 那么, 可通过计算 $\arg\max p(s_i|C)$ 确定w的词义.

应用实例

根据贝叶斯公式：$p(s_i \mid C) = \dfrac{p(s_i) \times p(C \mid s_i)}{p(C)}$

考虑分母的不变性，并运用如下独立性假设：

$$p(C \mid s_i) = \prod_{v_k \in C} p(v_k \mid s_i)$$

出现在上下文中的词

应用实例

根据贝叶斯公式：$p(s_i \mid C) = \dfrac{p(s_i) \times p(C \mid s_i)}{p(C)}$

考虑分母的不变性，并运用如下独立性假设：

$$p(C \mid s_i) = \prod_{v_k \in C} p(v_k \mid s_i)$$

因此，
$$\hat{s_i} = \arg\max_{s_i} \left[p(s_i) \prod_{v_k \in C} p(v_k \mid s_i) \right] \qquad (1)$$

概率$p(v_k|s_i)$和 $p(s_i)$ 都可用极大似然估计求得：

应用实例

$$p(v_k \mid s_i) = \frac{N(v_k, s_i)}{N(s_i)} \tag{2}$$

其中，$N(s_i)$ 是在训练数据中词w用于语义s_i时的次数，而 $N(v_k, s_i)$ 为w用于语义s_i时词 v_k 出现在w 的上下文中的次数.

$$p(s_i) = \frac{N(s_i)}{N(w)} \tag{3}$$

$N(w)$ 为多义词 w 在训练数据中出现的总次数.

应用实例

举例说明：$$p(v_k \mid s_i) = \frac{N(v_k, s_i)}{N(s_i)}$$

对于打字而言，假设做实词用的25个语义分别标记为：$s_1 \sim s_{25}$. 假设 s_1 的语义为敲击(beat)。那么，$N(s_1)$表示打字的意思为敲击(beat)时在所有统计样本中出现的次数；$N(v_k, s_1)$ 表示某个词 v_k 出现在 s_1 的上下文中时出现的次数。例如，句子：

他	对	打	鼓	很	在	行	(取上下文： ±2)
-2	-1	↑	+1	+2			

应用实例

他	对	打	鼓	很	在	行	(取上下文： ±2)
-2	-1	↑	+1	+2			

那么，上下文 C=(他，对，鼓，很)。如果 v_k=他，

N(他, s_1) =5, $N(s_1)$=100，那么，

$$p(v_k \mid s_i) = p(\text{他} \mid s_1) = \frac{N(\text{他}, s_1)}{N(s_1)} = \frac{5}{100} = 0.05$$

假若打在所有样本中总共出现了800次，那么，

$$p(s_i) = \frac{N(s_i)}{N(w)} = \frac{N(s_1)}{N(\text{打})} = \frac{100}{800} = 0.125$$

应用实例

➢ 消歧算法描述：

a) 对于多义词 w 的每个语义 s_i 执行如下循环：对于词典中所有的词 v_k 利用训练语料计算

$$p(v_k \mid s_i) = \frac{N(v_k, s_i)}{N(s_i)}$$

b) 对于 w 的每个语义 s_i 计算：

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

应用实例

对于 w 的每个语义 s_i 计算 $p(s_i)$, 并根据上下文中的每个词 v_k 计算 $p(w|s_i)$, 选择:

$$\hat{s}_i = \arg\max_{s_i} \left[p(s_i) \prod_{v_k \in C} p(v_k \mid s_i) \right]$$

标注过程或称测试过程

应用实例

对于 w 的每个语义 s_i 计算 $p(s_i)$, 并根据上下文中的每个词 v_k 计算 $p(C|s_i)$，选择:

$$\hat{s}_i = \arg\max_{s_i} \left[p(s_i) \prod_{v_k \in C} p(v_k \mid s_i) \right]$$

标注过程或称测试过程

说明：在实际算法实现中，通常将概率 $p(v_k|s_i)$ 和 $p(s_i)$ 的乘积运算转换为对数加法运算:

$$\hat{s_i} = \arg\max_{s_i} \left[\log p(s_i) + \sum_{v_k \in C} \log p(v_k \mid s_i) \right]$$

Thank you!

权小军 中山大学数据科学与计算机学院