寻找巧妙的共现

# 语义计算初步
# ——词义消歧

杨沐昀

哈工大教育部-微软语言语音重点实验室

MOE-MS Joint Key Lab of NLP and Speech (HIT)

# 1.概述

◆**语义计算的任务**：解释自然语言句子或篇章各部分(词、词组、句子、段落、篇章)的意义。

◆**面临的困难**：

- 自然语言句子中存在大量的歧义，涉及指代、同义/多义、量词的辖域、隐喻等；
- 同一句子对于不同的人来说可能有不同的理解；
- 语义计算的理论、方法、模型尚不成熟。

# 1.概述

- 语义的定义
  - 符号学：词的指称(signified)
  - 心理图像：image
  - 说话者的意图：speech act
  - 情景语义：
- 语义计算的经典框架
  - 格语法(Fillmore,1966)：施事、受事、工具….
  - 语义网络(Quilian, 1968)：is-a, part-of, is
  - 概念依存(Schank, 1970s)：动作基元、剧本、计划

# 2.多义词

## 多义词

- 多义词是自然语言中普遍存在的现象

  生意淡　　口味淡

  我就来　　我就不来　　我就记得一句话

  bank　　time　　fly

- 在NLP的许多应用领域，都需要识别出多义词在具体语境中的意思。

# 2.多义词

## 定义

- 语义歧义：很多词语具有几个意思或语义，如果将这样的词从上下文中独立出来，就会产生语义歧义

打酱油　打电话　打毛衣　打手势　打哈欠

生意很清淡　口味比较清淡

拍子坏了　打拍子

我就来　我就不来　我就记得一句话

# 2. 多义词

## 常用词（字）的多义情况

| Marrian-Webster袖珍词典 | | 《现代汉语通用字典》 | |
|---|---|---|---|
| 词形 | 义项数 | 词形 | 义项数 |
| go | 63 | 打 | 26 |
| fall | 35 | 上 | 20 |
| run | 35 | 下 | 19 |
| turn | 31 | 干 | 19 |
| way | 31 | 子 | 18 |
| work | 31 | 着 | 18 |
| do | 30 | 生 | 18 |
| draw | 30 | 和 | 18 |
| play | 29 | 点 | 18 |
| get | 26 | 折 | 17 |

# 2. 多义词

## 同义词词林

《同义词词林》，梅家驹 等，1983，上海辞书出版社

| | 单字词 | | 多字词 | | |
|---|---|---|---|---|---|
| | 词条数 | 百分比 | 词条数 | 百分比 | |
| 单义词 | 1973 | 52.3% | 40751 | 87.9% | 42724 |
| 多义词 | 1801 | 47.7% | 5629 | 12.1% | 7430(14.8%) |
| 总计 | 3774 | 100% | 46380 | 100% | 50154 |

引自黄昌宁 等《词义排歧的一种语言模型》，载《语言文字应用》2000年第3期

# 2. 多义词

## 多义词

多义词的分类

- 甲类多义词：不同词性——不同义项

  如“编辑”：N 和 V

  - 兼类词：一定是多义词
    - 动名兼类：制服　建议　book……
    - 形名兼类：　秘密　经济　现实……
    - 形动兼类：充实　负责　饿 ……
  - 甲类多义词的义项标注实际上即词性标注

# 2. 多义词

## 多义词

多义词的分类

- 乙类多义词：相同词性——不同意思

  如“材料”：他是做外交工作的好材料； 装饰材料

  “打”：打电话；打酱油；打毛衣；打架

  “淡”：颜色淡；生意淡；味道淡

→ 一个多义词的多个义项之间的差别体现在哪里？

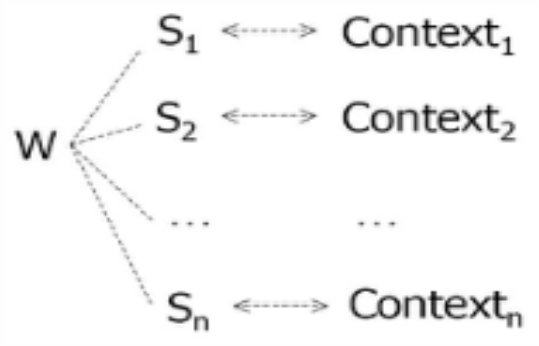
# 2. 多义词

## 多义词

- You shall know a word by the company it keeps.

  ——J.R. Firth, 1957, A Synopsis of Linguistic Theory 1930-1955, In Studies in Linguistic Analysis, Philological Society, Oxford.

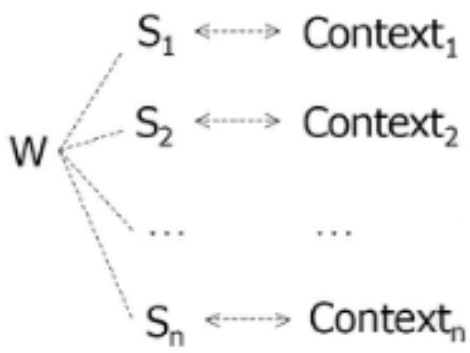
- 通过一个词周围的搭配词（即上下文语境）来了解其意义




- 什么是Context
- 如何找Context

# 2. 多义词

## 多义词




- 什么是Context
- 如何找Context

- 如何确定具体语境中多义词的确切意义？——词义标注/消歧
- 对乙类多义词的判别：寻找具有区别意义的$C_1, C_2, \ldots, C_n$
- 对$C_i$的认识不同，寻找$C_i$的途径也不同 → 不同的WSD方法

# 3. 词义消歧

## 词义标注 / 排歧（WSD）

- 词义标注/消歧：确定一个多义词在具体语境中的义项
- WSD需要解决三个问题：

（1）如何判断一个词是不是多义词？

如何表示一个多义词的不同意思？

（2）对每个多义词，预先要有关于它的各个不同义项的清晰的区分标准

WSD所需的基础资源

（3）对出现在具体语境中的每个多义词，为它确定一个合适的义项

# 3.词义消歧

## 词义标注 / 排歧（WSD）

WSD的基础资源：

- 传统的语文词典：通常列举每个多义词的不同义项
- 义类词典/同义词词典：每个词语按照意义归入不同的语义类
- 标注好义项的语料库：每个多义词的义项都跟一些确定的语境相关联

# 3.词义消歧

## 词义标注/排歧（WSD）

WSD技术策略：

- 将多义词与已经标注好义项的语料进行对比，确定多义词的义项；
- 事先制定语义消歧规则，根据规则来确定具体语境下多义词的义项；
- 利用词典、百科辞典等知识库中的信息来确定多义词的义项。

# 3.词义消歧

## 词义标注/排歧（WSD）

不同的WSD系统在实现两步骤的具体策略不同：

- 基于机器词典的WSD
- 基于义类词典的WSD
- 基于语料库的WSD
- 基于统计方法的WSD
- 基于规则的WSD
- ……

由名字可知WSD使用哪种资源，采用什么策略进行词义排歧

# 4.词义消歧——基于知识库的方法

## 基于词典释义的W

- 词典是语言学家对词语知识归纳总结的结果；
- 词典中对多义词的各个义项的描写是对多义词的不同使用情况的总结

从 cóng 325

craft

【从实】 cóngshí 按真实情况；如实 in the light of the fact (that ...)；based on the fact；～回答 answer honestly (frankly)

【从事】 cóngshì ❶ 投身到（事业中去） pursue；go in for；devote oneself to；throw oneself into；work on；occupy oneself with；take part in；go about；take up；be engaged in；be bound up in；～革命 devote oneself to revolutionary work | ～文艺创作 engage in literary and artistic creation ❷ （按某种办法）处理 (in certain way) deal with；军法～ deal with according to military law；court-martial sb.

【从属】 cóngshǔ 依从；附属 subordinate；dependent；～关系 relationship of subordination；affiliation

【从俗】 cóngsú ❶ 按照风俗习惯；遵循通常做法 follow local custom；follow tradition；conform to convention；～办理 proceed according to local customs | ～就简 conform to conventions while adhering to the principle of simplicity ❷ 指顺从时俗 follow what the majority are doing；～浮沉 experience ups and downs like most people do；live an ordinary life without much of a struggle to better one's situation

【从速】 cóngsù 赶快；赶紧 as soon as possible；without delay；～处理 deal with the matter as soon as possible；settle the matter quickly | 存货不多，欲购～。 Buy now, while they last.

【从头】 cóngtóu (～儿 cóngtóur) ❶ 从最初（做） from the beginning；from scratch；～儿做起 start from the very beginning ❷ 重新（做） afresh；anew；once again；～儿再来 start afresh；start all over again

【从先】 cóngxiān 〈方 *dial.*〉 same as 从前 cóngqián；他身体比～结实多了。 He's much stronger than before.

# 4.词义消歧——基于知识库的方法

## 基于词典释义的WSD方法

- 基于词典释义的WSD方法：利用词典中的释义文本进行WSD

  Lesk，1986，准确率50% -70%之间
- 词典释义： cone
  - a mass of ovule-bearing or pollen-bearing scales or bracts in trees of the pine family or in cycads that are arranged usually on a somewhat elongated axis. 松果
  - something that resembles a cone in shape : as ...a crisp cone-shaped wafer for holding ice cream. 蛋卷冰淇淋
- 语境消歧：
  - 上下文中出现了tree → 第1个义项
  - 上下文中出现了ice → 第2个义项

# 4.词义消歧——基于知识库的方法

## 基于词典释义的WSD方法

- 已知：

1)一个多义词W有若干义项($S_1, S_2, \ldots, S_m$)；

2)多义词W的每个义项($S_i$)在词典中分别有一个释义($D_i$)，每个释义($D_i$)实际上代表了一组出现在该释义中的词{$a_1, a_2, a_3, \ldots$}；

3)多义词W在一个具体的上下文(C)中出现时，前后有一些词($W_1, W_2, \ldots$)，这些词将作为判定多义词W意思的上下文特征词；

4)每个特征词($W_j$)在词典中也分别有释义($E_1, E_2, \ldots$)，每个释义($E_{wj}$)实际代表了一组出现在该释义中的词{$b_1, b_2, b_3, \ldots$}。

- 判断多义词在语境中的义项：对每个义项$S_i$计算 $\mathrm{Score}(S_i) = D_i \cap (\bigcup_{w_j \in C} E_{w_j})$

即 $\{a_1, a_2, a_3, \ldots\} \cap (\{b_1, b_2, \ldots\} \cup \ldots \{b_1', \ldots, b_k'\})$

取最大值所对应的$S_i$，即为该多义词的义项。

# 4.词义消歧——基于知识库的方法

## 基于词典释义的WSD方法

| Word | Sense | Definition (from Collins COBUILD) |
|---|---|---|
| pen | $S_1$:笔 | A pen is a long thin object which you use to write in ink. |
| | $S_2$:围栏 | A pen is a small area with a fence round it in which farm animals are kept for a short time. |
| sheep | $S_1$:羊 | A sheep is a farm animal with a thick woolly coat. |
| | … | … |

- 多义词pen： The sheep has been penned for three days.

在pen的上下文中只有sheep这个词的释义跟pen的一个释义有交集词

$$\left.\begin{array}{l}\text{Score}(s_1)=0\\ \text{Score}(s_2)=2\end{array}\right\} \rightarrow \text{取}S_2$$

# 4.词义消歧——基于知识库的方法

## 基于词典释义的WSD方法

总结：

- 用词典资源进行词义排歧，是利用词典中对多义词的各个义项的描写，求多义词的释义跟其上下文环境词的释义之间的交集，判断词义的亲和程度，来确定词义；
- 由于词典释义的概括性，这种方法应用于实际语料中多义词的排歧，效果不一定理想。

# 4.词义消歧——基于知识库的方法

## 基于义类词典的WSD方法

- Yarowsky, 1992. 试验12个多义词，准确率92%
- 基本思想：一个多义词在义类词典中可能分属不同的义类，在具体语境中，确定了一个多义词的义类实际上就刻画了它的一个义项。
  - 如：“crane”有两个意思，一是指“吊车”，一是指“鹤”。前者属于“工具/机械”这个义类；后者属于“动物”这个义类。如果能够确定“crane”出现在具体语境中时属于哪个义类，实际上也就知道了“crane”的义项。

# 4.词义消歧——基于知识库的方法

## 基于义类词典的WSD方法

基于义类词典进行WSD需要解决两个问题：

- 表示每一个义类的特征词，以及每个特征词对于该义类的权重；
- 对于一个具体语境中的多义词，根据其周围词隶属于某个义类的可能性大小，选择其中可能性最大的那个义类作为该多义词对应的义项标记。

# 4.词义消歧——基于知识库的方法

## 基于义类词典的WSD方法

基于义类词典的WSD的过程（第一步）：

- 对Roget词典中每个义类（共1041个类）中所有的词，收集包含这些词的上下文C（每个词的上下文长度为前后100个词）作为训练数据
- Yarowsky收集的训练语料来自Grolier百科全书1991年的电子版，1000万词规模。 如包含“工具/仪器”类中部分词的语料：

| Training Data (Words in Context) |
| --- |
| ... CARVING .SB The gutter **adz** has a concave blade for form ... |
| ... uipment such as a hydraulic **shovel** capable of lifting 26 cubic ... |
| ... on .SB Resembling a power **shovel** mounted on a floating bul ... |
| ... uipment , valves for nuclear **generators** , oil-refinery turbines ... |
| ... 00 BC , flint-edged wooden **sickles** were used to gather wild ... |
| ... l-penetrating carbide-tipped **drills** forced manufacturers to fi ... |
| ... ent heightens the colors .SB **Drills** live in the forests of equa ... |
| ... traditional ABC method and **drill** were unchanged , and dissa ... |
| ... nter of rotation .PP A tower **crane** is an assembly of fabricat ... |
| ... rshy areas .SB The crowned **crane** , however , occasionally ... |

# 4.词义消歧——基于知识库的方法

## 基于义类词典的WSD方法

基于义类词典的WSD的过程（第二步）

- 对C进行统计，找出能够有效地标示每个义类的特征词，并计算各个特征词的权值：

$$Weight(w) = \log(\frac{P(w \mid RCat)}{P(w)})$$

P(w|Rcat)表示w出现在Rcat类中的概率，P(w)表示w出现在训练语料库中的总概率

- 如：

| “动物”类特征词 | “工具”类特征词 |
| --- | --- |
| species(2.3), family(1.7), bird(2.6), fish(2.4), breed(2.2), animal(1.7), tail(2.7), … | tool(3.7), machine(2.7), engine(2.6), blade(3.8), cut(2.6), saw(5.1), lever(4.1),… |

# 4.词义消歧——基于知识库的方法

## 基于义类词典的WSD方法

基于义类词典的WSD的过程（第三步）

- 判断在某个具体的语境中出现的多义词所属的义类：
  - 如果在该多义词的上下文中能够且只能找到一个义类的特征词，则该多义词即属于这个义类；
  - 如果在该多义词的上下文中找到若干个特征词，且分别对应着不同的义类，根据Bayes法则，分别求这些特征词所对应的不同义类的权值之和，哪个义类的特征词权值之和最大，该多义词就属于哪个义类。

# 4.词义消歧——基于知识库的方法

## 基于义类词典的WSD方法

…lift water and to grind grain .PP Treadmills attached to cranes were used to lift heavy objects from Roman times , …

| TOOLS/MACHINE | Weight | ANIMAL/INSECT | Weight |
|---|---|---|---|
| lift | 2.44 | water | 0.76 |
| lift | 2.44 | | |
| grain | 1.68 | | |
| used | 1.32 | | |
| heavy | 1.28 | | |
| Treadmills | 1.16 | | |
| attached | 0.58 | | |
| grind | 0.29 | | |
| water | 0.11 | | |
| TOTAL | 11.30 | TOTAL | 0.76 |

# 4.词义消歧——基于知识库的方法

## 基于义类词典的WSD方法

总结：

- 可以理解为是对一个多义词所处语境的“主题领域”的猜测，假定如果当前主题领域猜对了，该多义词的义项也能判定正确；
- 对训练语料库不需要事先标注；
- 对义项区别依赖大语境的多义词效果较好（如名词）；
- 对义项区别对应着义类区别的多义词效果较好；
- 对那些不依靠大语境提示词义的多义词效果较差（如动词和形容词）；
- 对义项区别不依赖主题的多义词效果较差。

——随着电子义类词典资源的丰富（如Wordnet），Thesaurus-based WSD 可以在更多资源基础上应用，效果会有一定程度的提高。

# 5.词义消歧——基于统计的方法

## 基于互信息的WSD方法

- 基于互信息的WSD方法：Brown, et al, 1991
- 思路：要判断多义词在具体语境下的意义，关键是找到能够指示该多义词意义的示意特征（indicator）

| 多义词（法语） | 译词（英语） | 示意特征 | 示意特征的具体取值 |
|---|---|---|---|
| Prendre [prã:dr] | take | 当前词的宾语 | 当prendre的宾语是mesure时 |
| | make | 当前词的宾语 | 当prendre的宾语是décision时 |
| vouloir [vulwa:r] | want | 当前词的时态 | 当vouloir为现在时形式时 |
| | like | 当前词的时态 | 当vouloir为条件时态形式时 |
| cent [sã] | percent | 当前词的左边一个词 | 当cent左边词语为per时 |
| | c. | 当前词的左边一个词 | 当cent左边是数字时 |

# 5.词义消歧——基于统计的方法

## 基于互信息的WSD方法

如何得到多义词的示意特征？其取值是什么？— Flip-Flop算法

- 假设：
  - 一个法语多义词在英语中存在若干译词$t_1,t_2,\ldots,t_m$
  - 对于一个多义词，其示意特征可能的取值为$v_1,v_2,\ldots,v_n$

$$I(R,Q)=\sum_{r_i\in R}\sum_{q_j\in Q}P(r_i,q_j)\log\frac{P(r_i,q_j)}{P(r_i)P(q_j)}$$

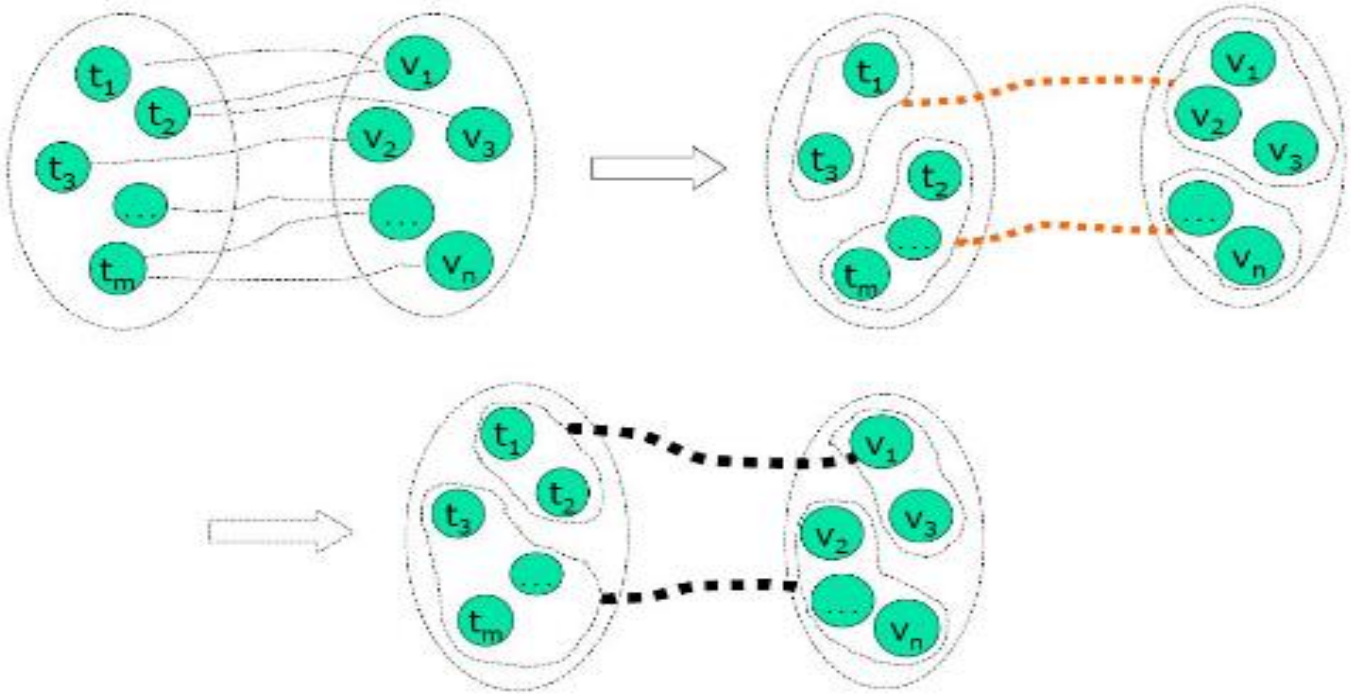
- 算法：

（1）随机地将$t_1,t_2,\ldots,t_m$分为两类，可记作$R=\{r_1, r_2\}$；

（2）寻找$v_1,v_2,\ldots,v_n$的一个分类$Q=\{q_1,q_2\}$，使得Q与R的互信息值I(R,Q)最大。根据Q，再调整R的分类，反复进行这个过程，直到I(R,Q)的值不能再提高（或变化甚微）为止。

# 5.词义消歧——基于统计的方法

## 基于互信息的WSD方法



# 5.词义消歧——基于统计的方法

## 基于互信息的WSD方法

示例：

- 看：{t1=读, t2=观看}
- {v1=电影,v2=报,v3=书,v4=小说,v5=电视}

- Count(t1)=3, Count(t2)=2,
  Count(v1)…=Count(v5)=1
  Count(t1,v1)=Count(t1,v5)=0,
  Count(t1,v2)=Count(t1,v3)=Count(t1,v4)=1
  Count(t2,v1)=Count(t2,v5)=1
  Count(t2,v2)=Count(t2,v3)=Count(t2,v4)=0

带有语义标记的训练语料库

看电影（观看）
看报（读）
看书（读）
看小说（读）
看电视（观看）
……

样本容量N=10

# 5.词义消歧——基于统计的方法

## 基于互信息的WSD方法

r1:{t1=读}　　r2:{t2=观看}

分类1：　q1 {v1=电影,v2=报}　　q2 {v3=书,v4=小说,v5=电视}

$$I_1(R,Q)= p(t_1,q_1)\log\frac{p(t_1,q_1)}{p(t_1)p(q_1)}+\ldots+p(t_2,q_2)\log\frac{p(t_2,q_2)}{p(t_2)p(q_2)}$$

$$=\frac{1}{10}\log\frac{10\times1}{3\times2}+\frac{2}{10}\log\frac{10\times2}{3\times3}+\frac{1}{10}\log\frac{10\times1}{2\times2}+\frac{1}{10}\log\frac{10\times1}{2\times3}$$

$$=\frac{5}{10}\log10-\frac{1}{10}\log2430$$

$I_2(R,Q)> I_1(R,Q)$

分类2：　q1 {v2=报,v3=书,v4=小说}　　q2 {v1=电影,v5=电视}

$$I_2(R,Q)=\frac{3}{10}\log\frac{10\times3}{3\times3}+\frac{0}{10}\log\frac{10\times0}{3\times2}+\frac{0}{10}\log\frac{10\times0}{2\times3}+\frac{2}{10}\log\frac{10\times2}{2\times2}$$

$$=\frac{5}{10}\log10-\frac{1}{10}\log108$$

# 5.词义消歧——基于统计的方法

## 基于互信息的WSD方法

多义词的示意特征以及特征值都确定下来后，判定多义词的义项：

1. 扫描该多义词所在的上下文环境，取得该多义词示意特征的当前值$V_i$；
2. 如果$V_i$属于$q_1$，则多义词义项为$r_1$；如果$V_i$属于$q_2$，则多义词义项为$r_2$。

# 5.词义消歧——基于统计的方法

## 基于Bayes判别的WSD方法

- 基于Bayes判别的WSD方法：

  Gale et al., 1992，试验了6个多义词，准确率90%

- 基本思想：

  计算多义词W出现在给定上下文语境C（包括多个词$w_1,w_2,\ldots,w_n$）中，标注为各个义项$S_i$概率大小$P(s_i|C)$，使$P(s_i|C)$最大的义项即为该多义词W的义项标注。

# 5.词义消歧——基于统计的方法

## 基于Bayes判别的WSD方法

- 多义词w 有多个义项 $s_1, s_2, \ldots, s_i, \ldots$
- 上下位语境C = $w_1, w_2, \ldots, w_n$

→

$$P(s_i \mid C) = \frac{P(C \mid s_i)P(s_i)}{P(C)}$$

$$s' = \arg\max_{s_i} P(s_i \mid C) = \arg\max_{s_i} \frac{P(C \mid s_i)P(s_i)}{P(C)}$$

$$= \arg\max_{s_i} P(C \mid s_i)P(s_i)$$

$$P(s_i) = \frac{Count(s_i)}{Count(w)}$$

$$P(C \mid s_i) = P(\{w_j \mid w_j \in C\} \mid s_i) = \prod_{w_j \in C} \frac{Count(w_j, s_i)}{Count(s_i)}$$

# 5.词义消歧——基于统计的方法

## 基于Bayes判别的WSD方法

词义排歧算法（Disambiguation）：

```
for all sense s_i of w        do
        score(s_i) = logP(s_i)
        for all words w_j in the context of w  do
                score(s_i)=score(s_i) + logP(w_j|s_i)
        end
end
choose s' = argmax score(s_i)
```

# 5.词义消歧——基于统计的方法

## 基于Bayes判别的WSD方法

- 例： 词义知识库示例

| $s_i$ \ $w_j$ | $P(s_i)$ | … | 书 | 武侠 | 电影 | 股市 | 行情 | 桌子 | 小说 | … |
|---|---|---|---|---|---|---|---|---|---|---|
| 看$_1$ | 0.3 | … | 0.40 | 0.10 | 0.01 | 0.01 | 0 | 0.20 | 0.27 | … |
| 看$_2$ | 0.5 | … | 0 | 0.25 | 0.5 | 0.01 | 0 | 0 | 0.15 | … |
| 看$_3$ | 0.2 | … | 0.01 | 0.03 | 0.05 | 0.45 | 0.45 | 0 | 0 | … |
| … | … | … | … | … | … | … | … | … | … | … |

# 5.词义消歧——基于统计的方法

| $w_j$ / $s_i$ | $P(s_i)$ | … | 书 | 武侠 | 电影 | 股市 | 行情 | 桌子 | 小说 | … |
|---|---|---|---|---|---|---|---|---|---|---|
| 看$_1$ | 0.3 | … | 0.40 | 0.10 | 0.01 | 0.01 | 0 | 0.20 | 0.27 | … |
| 看$_2$ | 0.5 | … | 0 | 0.25 | 0.5 | 0.01 | 0 | 0 | 0.15 | … |
| 看$_3$ | 0.2 | … | 0.01 | 0.03 | 0.05 | 0.45 | 0.45 | 0 | 0 | … |
| … | … | … | … | … | … | … | … | … | … | … |

- 我看过由同名武侠小说改编的电影

score(看$_1$) = log0.3 + log0.1 + log0.27 + log0.01

score(看$_2$) = log0.5 + log0.25 + log0.15 + log0.5

score(看$_3$) = log0.2 + log0.03 + log0.05

→ score(看$_2$)最大，所以当前语境下是“看”的第2个义项

# 5.词义消歧——基于统计的方法

## 基于Bayes判别的WSD方法

总结：

(1) 标注好词义的语料库（training corpus）；

(2) 从标注语料库训练“语境”与词义之间的依赖关系，得到“词义知识库”；

(3) 对于一个输入句子中的多义词，根据“词义知识库”中的知识，计算它在当前“语境”下，取哪一个义项的可能性最高，就将该义项判定为这个多义词在当前语境下的意思。

# WSD小结

各种WSD技术和方法解决两个问题：

（1）如何确定用于词义排歧的可靠知识？

语境中的某个特定的提示特征？大语境？普通语文词典？义类词典？带语义标记的语料库？

（2）如何低代价，高效地，大规模地获得这样的知识？

人工？统计—机器自动获取？

# WSD小结

WSD研究的困难：

1）词义缺乏明确清晰的定义

2）搭配并不能完全确定一个词的意义

“有的是钱”——“有的是医生”

3）词义是相互依赖的

豆腐放坏了　豆腐放早了

打酱油　打翻了酱油

打眼睛　打湿了她的眼睛

4）对WSD系统的评价困难