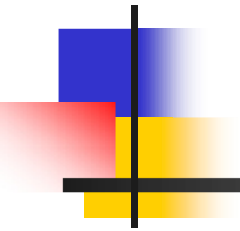
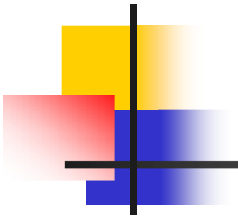


第9章 语义分析

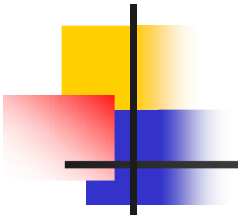
主要内容

- 定义
- 有监督消歧
- 基于词典的消歧



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
- 语义消歧：确定一个歧义词的哪一种语义在一个特定的环境中被调用

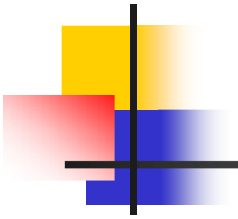


例：bank

- the rising ground bordering a lake, river, or sea …
- an establishment for the custody, loan exchange, or issue of money, for the extension of credit, and for facilitating the transmission of funds

例：title

- Name/heading of a book, statute, work of art or music, etc. 标题
- Material at the start of a film 字幕
- The right of legal ownership (of land) 权利
- The document that is evidence of this right
- An appellation of respect attached to a person's name 头衔
- A written work [by *synecdoche*, i.e., putting a part for the whole]



应用

- 英德翻译
  - bank的第一个语义→Ufer
  - bank 的第二个 语义→Bank (financial institution)
- 信息检索
  - 查询：financial bank，
  - 返回：使用第二个语义的文档

有监督词义消歧

- 贝叶斯分类

  将上下文看作一个无结构词集，整合上下文中众多的词汇信息。

- 基于互信息的方法

  仅仅考虑上下文中的一个信息特征，其可以灵敏地反映上下文结构，但需要谨慎地选取此特征

符号约定

Symbol	Meaning
w	an ambiguous word
$s_1, \ldots, s_k, \ldots, s_K$	senses of the ambiguous word w
$c_1, \ldots, c_i, \ldots, c_I$	contexts of w in a corpus
$v_1, \ldots, v_j, \ldots, v_J$	words used as contextual features for disambiguation

贝叶斯分类

- 原理：在一个大的上下文窗口中考虑歧义词周围词的信息。
  - 每个实词都含有潜在的有用信息
  - 前提：语料库中的歧义词事先被正确地进行语义标注。
- 贝叶斯决策规则：最小化错误概率

$$\text{Decide } s' \text{ if } P(s'|c) > P(s_k|c) \text{ for } s_k \neq s'$$

为w指定语义s’

$$
\begin{aligned}
s' &= \arg\max_{s_k} P(s_k|c) \\
&= \arg\max_{s_k} \frac{P(c|s_k)}{P(c)} P(s_k) \\
&= \arg\max_{s_k} P(c|s_k) P(s_k) \\
&= \arg\max_{s_k} [\log P(c|s_k) + \log P(s_k)]
\end{aligned}
$$

朴素贝叶斯分类器

Naïve Bayes Classifier

- 朴素贝叶斯的两个假设
  - 上下文c中的词语顺序可忽略；
  - 每个词互相独立

$$P(c|s_k) = P(\{v_j|v_j \text{ in } c\}|s_k) = \prod_{v_j \text{ in } c} P(v_j|s_k)$$

I walk along the river bank.

修正的分类决策规则

Decision rule for Naive Bayes

Decide s' if $s' = \arg\max_{s_k}[\log P(s_k) + \sum_{v_j \text{ in } c} \log P(v_j|s_k)]$

- 其中

$$P(v_j|s_k) = \frac{C(v_j, s_k)}{C(s_k)}$$

语料库中语义Sk出现的总次数

$$P(s_k) = \frac{C(s_k)}{C(w)}$$

多义词w出现的总次数

贝叶斯消歧算法

1 **comment:** Training
2 **for** all senses s_k of w **do**
3 **for** all words v_j in the vocabulary **do**
4 $P(v_j|s_k) = \frac{C(v_j,s_k)}{C(v_j)}$
5 **end**
6 **end**
7 **for** all senses s_k of w **do**
8 $P(s_k) = \frac{C(s_k)}{C(w)}$
9 **end**

贝叶斯消歧算法(续)

10 **comment:** Disambiguation
11 **for** all senses s_k of w **do**
12 $\quad$ score$(s_k) = \log P(s_k)$
13 $\quad$ **for** all words v_j in the context window c **do**
14 $\quad\quad$ score(s_k) = score$(s_k) + \log P(v_j|s_k)$
15 $\quad$ **end**
16 **end**
17 choose $s' = \arg\max_{s_k}$ score(s_k)

贝叶斯模型消歧结果

- 对Hansard语料库中6个歧义名词(duty, drug, land, language, position, sentence)的消歧正确率可达90%
- Drug中两个语义的“线索词”

Sense	Clues for sense
medication	*prices, prescription, patent, increase, consumer, pharmaceutical*
illegal substance	*abuse, paraphernalia, illict, alcohol, cocaine, traffickers*

贝叶斯消歧实例

阅读

使视线接触人或物

观察并加以判断

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	…
…	…	…	…	…	…	…	…	…	…	…

我看过由同名武侠小说改编的电影

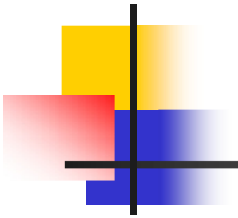
$\log P(s_k)$

$\log P(v_j|s_k)$

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



基于词典的消歧

- 基于语义定义的消歧
- 基于义类辞典的消歧

基于语义定义的消歧

- 词典中词条本身的定义可以作为判断其语义的一个很好的依据。
- 设cone在词典中有两个定义：
  - a mass of ovule-bearing or pollen-bearing scales of 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 圆锥体

- 如果cone的上下文中出现了tree，说明cone的语义是语义1。
- 如果cone的上下文中出现了ice，说明cone的语义是语义2

实现算法

1) 一个多义词有若干义项$(S_1, S_2, ..., S_m)$ ；

2) 多义词的每个义项(S_i)在词典中分别有一个释义$(D_1, D_2, ..., D_m)$，每个释义(D_i)实际上代表了一组出现在该释义中的词$\{a_1, a_2, a_3, ...\}$；

义项在词典中的定义(释义)

3) 多义词在一个具体的上下文（C）中出现时，前后有一些词$(v_1, v_2, ...)$，这些词将作为判定该多义词意思的上下文特征词(v_j)；

每个上下文词条的释义

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

5) 当要判断一个多义词在具体语境中的义项时，就对该多义词的每个义项(S_i)，计算：$Score(S_i) = D_i \cap (\bigcup_{v_j \in C} E_{v_j})$

即 $\{a_1, a_2, a_3, ...\} \cap (\{b_1, b_2, ...\} \cup ... \cup \{b_{1'}, ..., b_{k'}\})$，取$Score(S_i)$最大值所对应的$S_i$，作为该多义词的义项。

原理：比较 每个义项的释义 与 多义词的每个上下文词条的释义

基于语义定义消歧实例

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

$Score(s_1)=0$
$Score(s_2)=2$
} → 取S_2

实现算法

- 实际应用中，可以将百度百科或维基百科作为语义词典使用。
- 如：对乔丹进行消歧，可以发现其在百度百科有多个义项和释义。

实现算法

- 迈克尔·乔丹的释义

迈克尔·乔丹（Michael Jordan），全名迈克尔·杰弗里·乔丹（Michael Jeffrey Jordan），1963年2月17日生于美国纽约州布鲁克林，前美国职业篮球运动员，司职得分后卫/小前锋，现为夏洛特黄蜂队老板。 [22] [48]

乔丹在1984年NBA选秀中于第1轮第3位被芝加哥公牛队选中，职业生涯曾效力于芝加哥公牛队以及华盛顿奇才队，新秀赛季当选NBA年度最佳新秀。1986-87赛季，乔丹场均得到37.1分，首次获得NBA得分王称号。1991-93赛季，乔丹连续2次荣膺常规赛MVP（1991、1992）和3次总决赛MVP（FMVP） [2-3] ，率领芝加哥公牛队3夺NBA总冠军。1993年10月6日因父亲被害而宣布退役，两年后宣布复出。1996年入选NBA50大巨星。1996-98赛季，乔丹荣膺个人职业生涯第10次（共10次）NBA得分王以及第5次（共5次）常规赛MVP，并再次率领公牛队3夺（共6次）NBA总冠军，自己当选共第6次总决赛MVP。1999年1月13日在劳资谈判失败后再次宣布退役，两年后在华盛顿奇才队再次宣布复出。乔丹的职业生涯年年入选NBA全明星阵容（共14次）并3次当选NBA全明星MVP，10次入选NBA最佳阵容一阵，1985年入选NBA最佳阵容二阵，1988年荣膺NBA年度最佳防守球员，9次入选NBA最佳防守阵容一阵，3次荣膺NBA抢断王，两次夺得NBA全明星扣篮大赛冠军，1984年以及1992年夺得奥运会金牌。

- 中国体育用品品牌的释义

乔丹篮球鞋是乔丹体育用品有限公司设计、生产与经营的运动类服饰之一。按照不同乔丹篮球鞋的运动类别，设计制作了运动鞋、篮球鞋、休闲鞋、板鞋、跑步鞋、乒乓球鞋、网球鞋等多个种类。

乔丹篮球鞋为满足专业运动员以及运动爱好者进行篮球运动的需要，设计制作出能有减震和助弹跳双重功能和保护脚踝的鞋体设计的比赛用鞋。

基于义类词典的消歧

- 又称类义词典

词典内容按照义类(每个义项所属的类别层次结构)进行组织。如现代汉语分类词典。

如：义类词典中mouse有两个义项：老鼠 和 鼠标，其分属的义类为：哺乳动物 和 电子器件。

读者需要查找某个词的时候，先根据概念逐层搜索义项，再根据义项找到确切表达的词语。

基于义类词典的消歧

消歧原理：多义词的义类应该与多义词上下文的义类相同或相似。

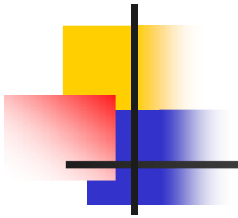
比如英语词“crane”有两个意思，一是指“吊车”，一是指“鹤”。前者属于“工具/机械”这个义类；后者属于“动物”这个义类。如果能够确定“crane”出现在具体语境中时属于哪个义类，实际上也就知道了“crane”的义项。

基于义类词典的消歧算法

对多义词的每个可能语义Sk，假定t(Sk)为义类词典中Sk的某个义类，那么通过统计将t(Sk)也作为义类的上下文单词的个数，即可对该词进行消歧。

1 **comment**: Given: context c
2 **for** all senses s_k of w **do**
3 $\quad \text{score}(s_k) = \sum_{v_j \text{ in } c} \delta(t(s_k), v_j)$
4 **end**
5 choose s' s.t. $s' = \arg\max_{s_k} \text{score}(s_k)$

$\delta(t(s_k), v_j) = 1$ 表示上下文词V_j与Sk具有相同的义类 $\delta(t(s_k)$



谢谢！！