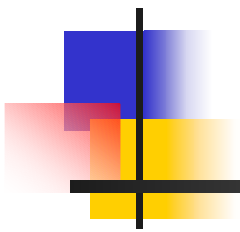


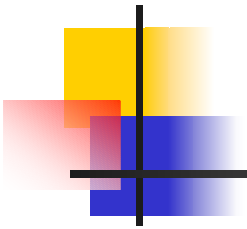
第七章 词汇分析（三）

—— 从词串到词义标记串

詹卫东

2002.5

http://icl.pku.edu.cn/doubtfire/course/CL/2001_2002_2.htm



提纲

- 1 词的多义现象
- 2 词义排歧（WSD）
  - (1) 如何确定一个词的义项？
  - (2) 如何找到能够判定词义的指示信息？
  - (3) 如何在具体语境中判定词义？
- 3 WSD方法简介
- 4 小结

1 词的多义现象（polysemous word）

打酱油　打电话　打毛衣　打手势　打哈欠　……

修门　进门　门上有把锁 | 拍子坏了　打拍子

生意很清淡　口味比较清淡

我就来　我就不来　我就记得一句话

开车　吃你的车

bank　table　title　book　eye　fly　…

所谓词的多义，就是一个“词形式”可以对应多种不同的变换形式（比如一个词对应着多个不同的翻译）

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

引自童翔1993，《汉语真实文本的语义自动标注》

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

多义词的分类

1 甲类多义词：不同词性 —— 不同意思

制服　编辑　建议　突出　秘密　特别　……

2 乙类多义词：相同词性 —— 不同义类 —— 不同意思

便衣　单位　图书馆　保管　老　红　……

3 丙类多义词：相同词性/义类 —— 不同特征 ——不同意思

表　材料　兄弟　大家　……

2 词义排歧（Word Sense Disambiguation）

WSD需要解决的三个问题:

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）对每个需要进行义项标注处理的多义词，预先得有关于它的各个不同义项的清晰的区分标准

3）对出现在具体语境中的每个多义词，给它确定一个合适的义项

解决这两个问题是提供WSD所需的基础资源

如何确定一个“word”的“sense”?

看电影 —— 看电视 —— 看病

开飞机 —— 开汽车 —— 开门 —— 开发票

炒菜 —— 炒外汇 —— 炒绯闻 …… ?

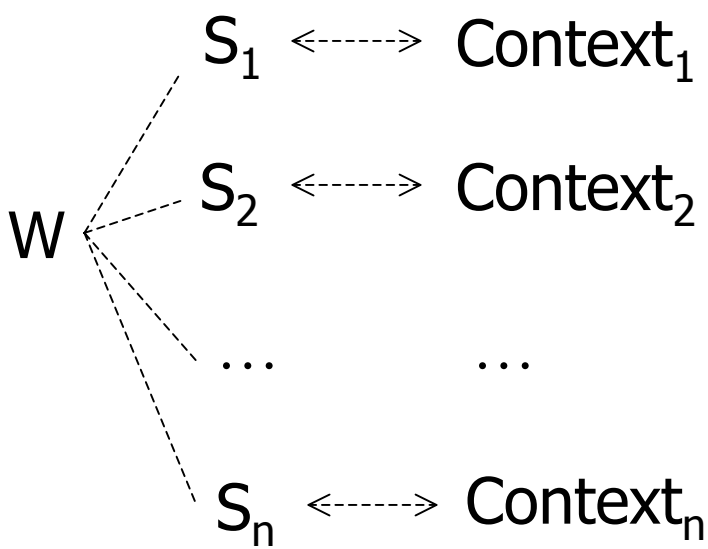
你在搞什么飞机?

你这个人真是很机车耶（台湾用法）

把多余的粮食卖给国家 —— 把多余的字句删除 ?

在言语中如何判定词义

- You shall know a word by the company it keeps
  （观其伴、知其意） —— J.R. Firth, 1957, A Synopsis of Linguistic Theory 1930-1955, In Studies in Linguistic Analysis, Philological Society, Oxford.



- 什么是Context
- 如何找Context

3 各种WSD方法简介

- 3.1 基于Bayes判别的方法
- 3.2 基于互信息的方法
- 3.3 基于词典释义的方法
- 3.4 基于义类词典的方法
- 3.5 基于实例相似度比较的方法
- 3.6 基于判定表的方法
  （词义搭配知识的自动抽取）

3.1 基于Bayesian Discrimination的方法

- Gale et al., 1992，试验了6个多义词，准确率90％

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。

Bayes decision rule

- 如果 $P(s' \mid c) > P(s_i \mid c)$ 则sense(w)=s′， $s' \neq s_i$

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

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

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

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

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

Bayes decision rule（续）

$$s' = \arg\max_{s_i} \log[P(C \mid s_i)P(s_i)] = \arg\max_{s_i}[\log P(s_i) + \log P(C \mid s_i)]$$

$$= \arg\max_{s_i}[\log P(s_i) + \log \prod_{w_j \in C} P(w_j \mid s_i)]$$

- 如果 $s' = \arg\max_{s_i}[\log P(s_i) + \sum_{w_j \in C} \log P(w_j \mid s_i)]$

  则sense(w)=s′

获取词义知识算法（Training）

for all sense s_i of w do

for all words w_j in the vocabulary do

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

end

end

for all sense s_i of w do

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

end

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 $s' = \arg\max_{s_i} score(s_i)$

词义知识库示例

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	…
…	…	…	…	…	…	…	…	…	…	…

基于Bayes判别的WSD示例

我看过由同名武侠小说改编的电影

$$score(看_1) = \log 0.3 + \log 0.1 + \log 0.27 + \log 0.01$$

$$score(看_2) = \log 0.5 + \log 0.25 + \log 0.15 + \log 0.5$$

$$score(看_3) = \log 0.2 + \log 0.03 + \log 0.05$$

显然，score(看$_2$)最大，所以当前语境下是“看”的第2个义项

“我看过由同名武侠电影改编的小说”中的“看”该是哪个义项？

3.2 基于互信息的WSD方法

- Brown, et al, 1991，应用于MT，将MT准确率从37%提高到45%

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

flip-flop算法

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

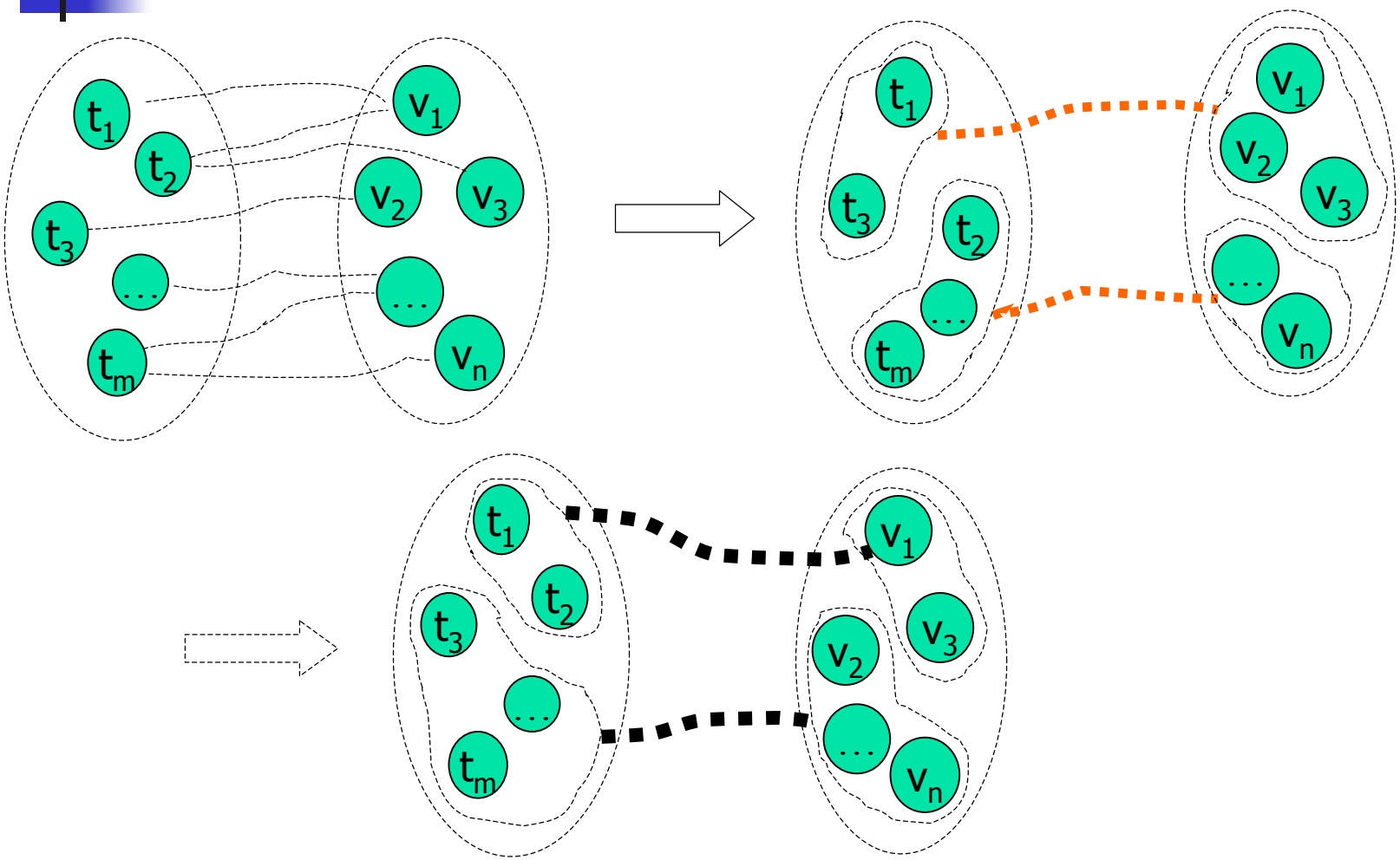
约定

- 假定一个法语词在英语中存在若干译词$t_1,t_2,\ldots,t_m$，这个法语词就是一个多义词；
- 对于一个多义词，其示意特征可能的取值为$v_1,v_2,\ldots,v_n$；

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。

基于互信息方法示意图



基于互信息的WSD方法示例

带有语义标记信息的训练语料库

看电影（观看）
看报（读）
看书（读）
看小说（读）
看电视（观看）
……

$\{t_1$=读, t_2=观看$\}$
$\{v_1$=电影,v_2=报,v_3=书,v_4=小说,v_5=电视$\}$

样本容量N=10

$Count(t_1)=3$, $Count(t_2)=2$,
$Count(v_1)\ldots=Count(v_5)=1$

$Count(t_1,v_1)=Count(t_1,v_5)=0$,
$Count(t_1,v_2)=Count(t_1,v_3)=Count(t_1,v_4)=1$

$Count(t_2,v_1)=Count(t_2,v_5)=1$
$Count(t_2,v_2)=Count(t_2,v_3)=Count(t_2,v_4)=0$

基于互信息的WSD方法示例（续）

r1:{t_1=读}　r2:{t_2=观看}

分类1

q1:{v_1=电影,v_2=报}　q2:{v_3=书,v_4=小说,v_5=电视}

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

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

分类2

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

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

q1:{v_2=报,v_3=书,v_4=小说}　q2:{v_1=电影,v_5=电视}

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

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

Bayes方法与互信息方法的比较

- 相同：
  都需要事先进行义项标注的语料库进行训练

- 不同：
  1) 对 “语境”的理解不同
  - Bayes方法：大语境——a bag of words
  - 互信息方法：小语境—— only one informative feature

  2) 对“训练语料库”的要求不同
  - Bayes方法：不需要标注结构信息
  - 互信息方法：根据训练目标，需要标注更多信息

3.3 基于词典释义的WSD方法

- 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：
  - 语境中出现tree → cone的意思将判定为“松果”
  - 语境中出现ice cream → cone的意思将判定为“冰淇淋”

基于词典释义的WSD方法的算法描述

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

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

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

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（Ew_j）实际代表了一组出现在该释义中的词$\{b_1, b_2, b_3, \ldots\}$。

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：$\text{Score}(S_i) = D_i \bigcap (\bigcup_{w_j \in C} E_{w_j})$

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

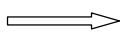
取$\text{Score}(S_i)$最大值所对应的S_i，作为该多义词的义项。

基于词典释义的WSD方法示例

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.
	…	…

He don't know exactly how a fox could have got into the sheep's pen.

Score(s_1)=0
Score(s_2)=2



Sense(pen)=s_2

基于词典释义的WSD方法小结

- 用语料库资源进行词义排歧，是利用多义词在实际使用中的实例来把握多义词的意思。
- 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，而这些描写是在语言学家观察了多义词的不同使用情况后概括归纳，抽象总结的结果。只不过跟实际语料不同的是，它是以一种概括的方式在描写词义，而语料库是以具体可感的大量重复的实例本身在描写词义。
- 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，效果不理想。

3.4 基于义类词典（thesaurus）的方法

- Yarowsky, 1992. 试验12个多义词，准确率92%
- 基本思想：
  一般而言，一个多义词在义类词典中可能分属不同的义类，这样，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。

  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WSD

解决两个问题：

1. 确定能够表示出每一个义类的特征词，以及每个特征词对于该义类的权重（weight）
2. 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

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）；

比如下面是包含“工具/仪器”类中部分词的语料(引自Yarowsky,1992)

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 hul ...
... 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 ...

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（Salient Words），并计算各个特征词的权值（Weight），计算公式为：

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

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

下面是“动物”类和“工具”类的特征词及其权重示例

“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),…

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。方法是：在该多义词的上下文中找到若干个特征词，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。

… 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

基于义类词典的WSD方法小结

- Thesaurus-based WSD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（比如名词）
- 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（比如动词和形容词）
- 对义项区别不依赖主题的多义词效果较差

3.5 基于实例的词义标注方法

- 童翔（1993）
- 基本思想

  通过将多义词所在上下文与实例库中已经标注好义项的实例进行对比，来推断该多义词的义项。

- 比如：实例库中有　打/B02　鼓/A01

  当前输入为：打锣鼓

  判断“锣鼓”与“鼓”非常相似，于是将“打锣鼓”中的“打”义项标记为B02

3.6 基于判定表（Decision list）的方法

- Yarowsky,David, 1994, *Decision list for lexical ambiguity resolution: application to accent restoration in Spanish and French*, In Proceedings of ACL 32, pp.88-95
- Yarowsky, David, 1995, *Unsupervised word sense disambiguation rivaling supervised methods*, In Proceedings of ACL 33, pp.189-196

应用背景：重音还原（accent restoration）

原始形式	重音模式	意思	语境
cote	côte [ko:t]	海滨（coast）	vivre sur notre **cote** ouest toujours verte creer sur la **cote** du labrador des travaillaient **cote** a cote, ils avaient
	côté [kote]	边（side）	du laisser de **cote** faute de temps appeler l'autre **cote** de l'atlantique passe de notre **cote** de la frontiere

Yarowsky, 1994

基于判定表的方法：基本过程

1) 确定语料中存在重音模式歧义的词；
2) 收集包含这些歧义词的语料，每行都有一个歧义词及其前后K个词
3) 对歧义词周围K个词范围内的“搭配”词（collocation）进行统计

Position	Collocation	côte	côté
-1 w	du *cote*	0	536
	la *cote*	766	1
	un *cote*	0	216
	notre *cote*	10	70
+1 w	*cote* ouest	288	1
	cote est	174	3
	cote du	55	156
+1w,+2w	*cote* du gouvernement	0	62
-2w,-1w	cote a *cote*	23	0
±k w	poisson (in ±*k* words)	20	0
±k w	ports (in ±*k* words)	22	0
±k w	opposition (in ±*k* words)	0	39

基于判定表的方法：基本过程（续）

4) 计算对数似然比（log-likelihood），构造判定表，公式为：

$$LogL = Abs\left(Log\left(\frac{P(Accent\_Pattern_m \mid Collocation_i}{P(Accent\_Pattern_n \mid Collocation_i} \right) \right)$$

LogL	Evidence/Collocation	Classification
8.28	PREPOSITION QUE **terminara**	terminara
7.24	de que **terminara**	terminara
7.14	para que **terminara**	terminara
6.87	y **terminara**	terminará
6.64	WEEKDAY (within k words)	terminará
5.82	NOUN QUE **terminara**	terminará

基于判定表的方法：基本过程（续）

5) 利用构造好的判定表对输入句子中的歧义词进行判定

多义词w + w所在的context

查表

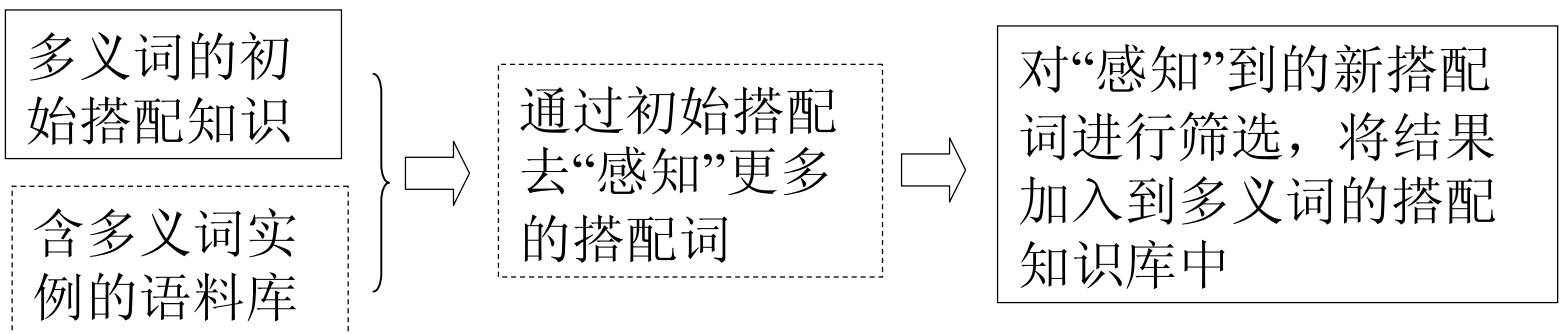
Word	Evidence/Context	LogL	Sense
terminara	…	8.28	1
terminara	…	7.24	2
…	…	7.14	…

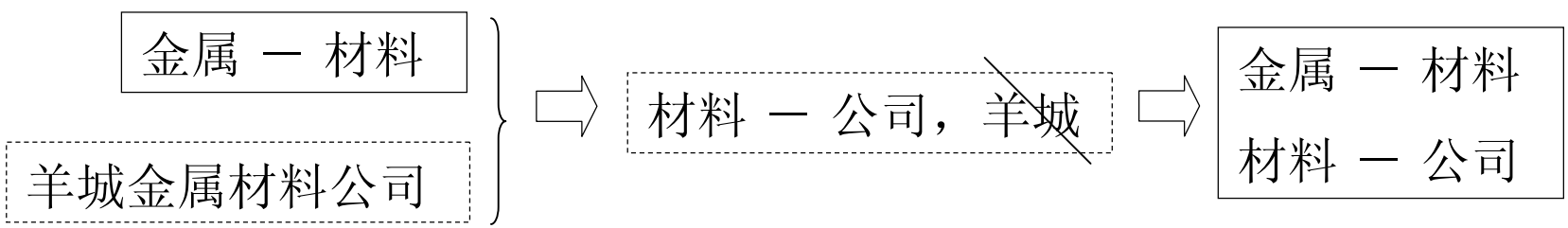
输出

匹配成功

词义搭配知识的自动获取

- 李涓子 等（1999）《一种自组织的汉语词义排歧方法》，载《中文信息学报》1999年第3期。
- Yarowsky, 1995





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
4） 对WSD系统的评价困难（pseudowords）

豆腐放坏了　豆腐放好了　豆腐放早了

打酱油 —— 打翻了酱油

打眼睛 —— 打湿了她的眼睛

WSD研究的意义

- 机器翻译：开飞机　开门　开发票 | 大雨　大小伙子　大公司
- 信息检索：计算机病毒　便衣　court
- 文本处理：简繁转换，重音复原，拼写校对，……
- 语音处理：看起来很美 —— 起来跑步
- ……

小结

纵观以往的WSD技术和方法，人们一直在努力解决两个问题：

（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？

对意义的追问，如同摸象的瞎子……

mission impossible ???

进一步阅读文献

- Lesk, Michael, 1986, Automatic sense disambiguation: How to tell a pine cone from an ice cream cone, In Proceedings of the 1986 SIGDOC Conference, pp24-26, New York, Association for Computing Machinery.
- Brown, Peter F., et al, 1991, Word-sense disambiguation using statistical methods, In Proceedings of ACL 29, pp264-270.
- Nancy Ide & Jean Veronis, 1998, Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art, Computational Linguistics, 24:1, 1-40.
- Gale, W., K. Church, and D. Yarowsky, 1992, A Method for Disambiguating Word Senses in a Large Corpus. Computers and the Humanities. 26, pp. 415-439.
- Yarowsky, David. 1992, Word-sense disambiguation using statistical models of Roget's categories trained on large corpora, In Proceedings of Coling-92.
- Christopher D. Manning & Hinrich Schutze, 1999, Foundations of Statistical Natural Language Processing, The MIT Press. Chapter 7
- 童翔1993，《汉语真实文本的语义自动标注》，载黄昌宁、夏莹编《语言信息处理专论》，清华大学出版社1996年版。
- 赵铁军等，2000，《机器翻译原理》，哈尔滨工业大学出版社。第9章。

复习思考题

1. 试归纳说明汉语中多义词的不同类型（可以有各种不同的角度，同时尽可能说明为何选取某个分类角度）
2. 分析汉语多义词示意特征的差异，比如多义量词的不同意思可能由它所修饰的名词来提示（“一段木头”——“一段历史”；“一片纸”——“一片云”），而多义动词的不同意思可能由其宾语提示，也可能由其补语提示（“打手机”——“打开手机”）
3. 分析多义词不同义项在句法功能上的差异，比如“清淡”做补语，定语，谓语等等不同功能成分，对判定它的意思有多少影响？