



教育部人文社会科学重点研究基地
黑龙江大学俄语语言文学研究中心

主办

ИССЛЕДОВАНИЯ
ПО РУССКОМУ ЯЗЫКУ И ЛИТЕРАТУРЕ

俄语语言文学 研究

第三辑 (语言学卷)
ВЫПУСК III (ЯЗЫКОЗНАНИЕ)

黑龙江人民出版社

教育部人文社会科学重点研究基地 主办
黑龙江大学俄语语言文学研究中心

俄语语言文学研究

第三辑(语言学卷)

ИССЛЕДОВАНИЯ
ПО РУССКОМУ ЯЗЫКУ И ЛИТЕРАТУРЕ

ВЫПУСК III (ЯЗЫКОЗНАНИЕ)

黑龙江人民出版社

图书在版编目(CIP)数据

俄语语言文学研究/张家骅主编. —哈尔滨:黑龙江
人民出版社, 2006.5

ISBN 7-207-06953-7

I. 俄... II. 张... III. 俄语—语言学—文集
IV. H35-53

中国版本图书馆 CIP 数据核字(2006)第 041741 号

责任编辑: 张晔明

封面设计: 张 涛

俄语语言文学研究(第三辑 语言学卷)

张家骅 主编

出版发行 黑龙江人民出版社

通讯地址 哈尔滨市南岗区宣庆小区1号楼(邮编 150008)

网 址 www.longpress.com E-mail hljmchs@yeah.net

印 刷 黑龙江省商业厅印刷厂

开 本 787×1092 毫米 1/16

印 张 22.5

字 数 420 000

版 次 2006 年 5 月第 1 版 2006 年 5 月第 1 次印刷

书 号 ISBN 7-207-06953-7/H·240

定价: 38.00 元

(如发现本书有印制质量问题, 印刷厂负责调换)

本刊(语言学卷)

主 编

张家骅

本辑执行主编

靳铭吉

本辑编辑

靳铭吉

孙淑芳

张春新

王晓阳

本刊编辑委员会

(以姓氏笔画为序)

- | | |
|-----|--------------|
| 田文琪 | 高等教育出版社 |
| 白春仁 | 北京外国语大学 |
| 任光宣 | 北京大学 |
| 刘文飞 | 中国社会科学院 |
| 余一中 | 南京大学 |
| 李 勤 | 上海外国语大学 |
| 吴国华 | 中国人民解放军外国语学院 |
| 金亚娜 | 黑龙江大学 |
| 张家骅 | 黑龙江大学 |
| 张建华 | 北京外国语大学 |
| 郝 斌 | 黑龙江大学 |

目 录

词义排歧方法研究	冯志伟(1)
名词短语的语义描写	张卫国 梁社会 刘宁静 计巍巍(14)
语义价与句法题元的对应关系	薛恩奎(20)
《用于信息技术的语义概念词典》建构策略论	易绵竹 武 斌 姚爱钢(33)
词汇函数理论在机器翻译中的应用	周清波(41)
中西语义的认知研究对比初探	王 寅(48)
俄语幽默言语含义解读的认知语用分析 ——关联理论和合成空间理论的整合性视角	蔡 晖 何英玉(58)
语言信息单位的认知功能与语用功能	彭文钊(73)
多义词 ДАЛЕКИЙ 词义发展过程中的认知范畴转换	吴 哲(87)
从《说文解字》看汉语语义研究的系统方法	戴昭铭(95)
术语学核心术语辨析	郑述谱(103)
术语词汇中的多义现象	吴丽坤(112)
论大型俄汉词典的释义	张金忠(122)
核义素与辞书的义项描写	王东海(137)
俄语宗教成语浅析	徐来娣(143)
汉语标题的语义	尹世超(152)
对语义三角理论的回顾与思考	邝友昌 孟 霞(162)
全译命题中枢转换机制	黄忠廉(176)
论俄语笑话的语义结构	王金玲(183)
社会语言学视野中的词汇意义研究	赵蓉晖(191)
“巨”类程度副词转化的语义基础和语用动因	邓 波(199)
语气词的两种隐含意义:标准与概括	杨家胜(206)

关于篇章实指信息	陈 勇(216)
知觉词的意觉语义转移	徐小波(228)
汉、英、俄语句对比分析	李 勤(237)
俄语中语义与搭配的关系分析	宁 琦(250)
论双宾构造中两个宾语的语义选择	徐德宽(262)
俄汉动词的体貌语义类别对比	张志军(273)
“中动态”语义及其与被动态的关系	彭玉海(295)
“打碎了”和“没打碎”的句法生成分析	陈喜富(305)
ДАРОМ 型否定副词的语义差异及其句法功能探索	惠秀梅(313)
论俄语语句信息结构与交际结构的关系	靳铭吉(320)
俄语小句初探	王 蓉(331)
词素性与词性的关系	张青琳(341)

Contents

Approaches for word sense disambiguation (WSD)	FENG Zhi-wei(1)
Semantic description of noun phrases	ZHANG Wei-guo, LIAN She-hui, LIU Ning-jing, JI Wei-wei(14)
Relations between semantic roles and syntactic actants	XUE En-kui(20)
On the building strategy of 《Semantic / Conceptual Dictionary for Information Technology》	YI Mian-zhu, WU Bing, YAO Ai-gang(33)
On the application of lexical functions in MT	ZHOU Qing-bo(41)
 The comparison of western and Chinese ancient scholars in terms of the cognitive semantics	WANG Yin(48)
Cognitive and pragmatic reading of Russian humor	
- From an integrative perspective of the relevance theory and blended space theory	CAI Hui, HE Ying-yu(58)
On cognitive and pragmatic functions of logopisterna	PENG Wen-zhao(73)
Cognitive categories in the development of the meaning of polysemic word "ДАЛЕКИЙ"	WU Zhe(87)
 Systematic research methods in 'Shuo Wen Jie Zhi'	DAI Zhao-ming(95)
An analysis of core words in terminology science	ZHENG Shu-pu(103)
The polysemy in terminology	WU Li-kun(112)
On several aspects of definition in big Russian-Chinese dictionaries	ZHANG Jin-zhong(122)
Kemel sememe and description of dictionary meaning	WANG Dong-hai(137)
An analysis of Russian religious idioms	XU Lai-di(143)
 The semantics of Chinese titles	YIN Shi-chao(152)
Reflections on the semantic triangle	ZHI You-chang, MENG Xia(162)
The proposition-oriented transformation mechanism of complete translation	HUANG Zhong-lian(176)
Semantic structures of Russian jokes	WANG Jin-ling(183)
Research in lexical meaning from the perspective of sociolinguistics	ZHAO Rong-hui(191)
The semantic and pragmatic explication of the functional transferring of the adverb JU	DENG Bo(199)
Two implicit meanings of particles: norm and generalization	YANG Jia-sheng(206)
On the referential information in text	CHEN Yong(216)
The meaning shifts of concept-words	XU Xiao-bo(228)

Onomasiological approach to sentence meaning

- Comparative analysis of sentences in the Chinese, English and Russian languages LI Qin(237)
- analysis of the relation between word meaning and collocation in Russian language NING Qi(250)
- The semantic selection of the two objects in double object constructions of modern Chinese XU De-kuan(262)
- The contrast study of verb aspect in Russian and in Chinese ZHANG Zhi-jun(273)
- Semantics of medium voice and its relation to the passive voice PENG Yu-hai(295)
- How "Da sui le" and "Mei da sui" are transformed in modern Chinese CHEN Xi-fu(305)
- On the semantic differences and syntactic functions of negative adverbs like *дапом* HUI Xiu-mei(313)
- On the relations between information structure and communicative structure of Russian sentences
..... JIN Ming-ji(320)
- Clauses in Russian language WANG Rong (331)
- Morpheme determines part of speech? ZHANG Qing-lin(341)

词义排歧方法研究

冯志伟 / 教育部语言文字应用研究所

提 要 本文全面总结了 40 多年来在词义排歧方面的研究成果,分别讨论和比较了各种词义排歧的方法:选择最常见义项的方法、利用词类进行词义排歧的方法、基于选择限制的方法、鲁棒的词义排歧方法、有指导的学习方法、自举的词义排歧方法、无指导的词义排歧方法、基于词典的词义排歧方法等。这些方法对于自然语言的计算机处理都是非常重要的。

关键词 自然语言处理 词义排歧 选择限制 有指导的学习方法 无指导的方法

词义排歧(Word Sense Disambiguation, 简称 WSD)是自然语言计算机处理中的一个很困难的问题。由于多义词是任何语言中都普遍存在的现象,而多义词中诸多的词义分布又很不容易找到一般的规律,多义词的排歧涉及到上下文因素、语义因素、语境因素,还甚至涉及到日常生活中的常识,而这些因素的处理,恰恰是计算机最感棘手的问题。

早在机器翻译刚刚问世的时候,美国著名数理逻辑学家 Bar-Hillel 在 1959 年就指出,全自动高质量的机器翻译(Fully Automatic High Quality MT, 简称 FAHQMT)是不可能的,他说明,FAHQMT 不仅在当时的技术水平下是不可能的,而且,在理论原则上也是不可能的。他举出了如下简单的英语片段,说明要在上下文中发现多义词 pen 的正确译文是非常困难的事情。

John was looking for his toy box. Finally he found it. The box was in the pen. John was very happy.

他的理由如下:

(i) pen 在这里只能翻译为 play-pen(“游戏的围栏”),而绝对不能翻译为书写工具“钢笔”。

(ii) 要确定 pen 的这个正确的译文是翻译好这段短文的关键所在。

(iii) 而要确定这样的正确译文依赖于计算机对于周围世界的一般知识。

(iv) 但是我们没有办法把这样的知识加到计算机中去。

可见,词义排歧问题一开始就困扰着刚刚萌芽的机器翻译研究。

从 1959 年到现在已经 40 多年了,学者们在探索多义词排歧的研究中做了大量的工作。尽管词义排歧的问题距离彻底解决还非常遥远,但是,从这 40 多

年的成就已经可以让我们看到希望的曙光。

词义排歧的方法归纳起来可以有如下几种：

1. 选择最常见义项的方法

早期的机器翻译系统没有词义排歧的功能,虽然机器词典中的多义词都列举出各种不同的义项,但实际上只是选择排列在第一位的那个最常见的义项。这样的办法虽然能够处理一些多义词,达到一定的排歧目的,但是,词义排歧的效率不高,这是早期机器翻译系统的译文质量低劣的重要原因之一。例如,在上面的例子中,由于 pen 最常见的词义是“钢笔”,因此,把 pen 翻译成“钢笔”,结果,The box was in the pen 就翻译成“盒子在钢笔中”,弄出了笑话。

2. 利用词类进行词义排歧的方法

一些多义词的词义与它们所属的词类有关。不同的词义往往属于不同的词类。因此,如果我们能够确定这些多义词的词类,词义排歧的问题也就迎刃而解了。例如:

face: 当 face 是动词时,它的词义是“面对”;当 face 是名词时,它的词义是“面孔”。在“The house faces the park”中,faces 前面为名词词组,后面也为名词词组,可判定为动词,因而它的词义是“面对”,整句的意思是“房子面对公园”。在“She pulled a long face”中,face 前面是形容词,可判定为名词,因而它的词义是“面孔”,整句的意思是“她拉长了面孔”。

May: 当 May 是助动词时,它的词义是“可以”(在句子开头,第一个字母大写,在其他情况下,第一个字母不大写),当 May 是名词并且第一个字母大写时,它的词义是“五月”。在“May I help you?”中,May 是助动词,因而它的词义是“可以”,整个句子的意思是“我可以帮助你吗?”在“May Day is first day of May”中,May 是名词,因而它的词义是“五月”,整个句子的意思是“五一劳动节是五月的第一天”。

can: 当 can 是助动词时,它的词义是“能够”,当 can 是名词时,它的意思是“罐头”。在“She can speak German”中,can 处于动词 speak 前面,人称代词 she 的后面,可判定为是助动词,因而它的词义是“能够”,整个句子的意思是“她能够说德语”。在“He opened a can of beans”中,can 前面是不定冠词,后面是介词,可判定为名词,因而它的词义是“罐头”,整个句子的意思是“他打开一瓶豆子罐头”。

will: 当 will 是助动词时,它的词义是“将要”,当 will 是名词时,它的意思是“意志”。在“It will rain tomorrow”中,will 前面是代词,后面是动词,可判定为助动词,因而它的词义是“将要”,整个句子的意思是“明天将要下雨”。在“Free will

makes us able to choose our way of life”中, will 前面是形容词, 后面是第三人称现在时动词, 可判定为名词, 因而它的词义是“意志”, 整个句子的意思是“自由的意志使得我们能够选择我们的生活方式”。

kind: 当 kind 是名词时, 它的意思是“种类”, 当 kind 是形容词时, 它的意思是“亲切”。在“ I like that kind of book”中, kind 在指示词 that 之后, 在介词 of 之前, 可判定为名词, 因而它的词义是“种类”, 整个句子的意思是“我喜欢这种书”。在“ It was very kind of you to do it”中, kind 在副词 very 的后面, 介词 of 的前面, 可判定为形容词, 因而它的词义是“亲切”, 整个句子的意思是“你做这件事显得非常亲切”。

如果我们有一个高效率的词性标注系统, 可以正确地决定兼类的多义词的词类, 那么, 我们就可以利用标注正确的词类, 来决定多义词的词义, 从而达到词义排歧的目的。可是, 当同一个词类的多义词还存在多个不同的词义的时候, 这种“以词类决定词义”的方法就显得无能为力了, 因为在判定了词类之后, 还需要对不同的词义进行选择。

例如, works 这个多义词可兼属动词和名词, 当它是动词的时候, 它的词义是“工作”, 当它是名词的时候, 它的词义可以是“工厂”, 也可以是“著作”。在句子“ My daughter works in an office”中, works 处于名词词组之后, 介词之前, 可判定为动词, 因而它的词义是“工作”, 整个句子的意思是“我女儿在一个办公室工作”。可是, 当判定 works 为名词的时候, 它的词义还没有最后决定, 这就会出现两难的尴尬局面。在“ It is a gas works”和“ I read the works of Shakespears”中, works 都可以判定为名词, 可是, 我们还决定不了前句中 works 的词义是“工厂”, 后句中的 works 的词义是“著作”。这时, 我们还需要根据上下文的选择限制来排歧。比如说, 如果我们规定, works 与表示燃料的名词连用, 可判定其词义是“工厂”, 当 works 与作家的名字连用, 可判定其词义是“著作”, 那么, 我们就可以根据这样的选择限制来进行词义排歧。

3. 基于选择限制的词义排歧方法

选择限制(selectional restriction)和语义类型的分类(type hierarchies)是词义排歧的主要的知识源。在语义分析中, 它们被用来删除不恰当的语义从而减少歧义的数量。

最早研究选择限制的是 Katz 和 Fodor(1963)。把选择限制应用于计算机处理的是 Hirst(1987)。

例如, dish 的排歧, 我们来研究下面的一段话:

“In our house, every body has a career and none of them includes washing dishes,”

he says.

In her tiny kitchen at home, Mr. Chen works efficiently, stir-frying several simple dishes, including braised pig's ears and chicken livers with green peppers. (他说道, “在我们的家里, 每一个人都有自己的事情, 但不包括洗碟子。” 在她的小厨房里, 陈先生干得很有成效, 他炒炸几个简单的菜肴, 包括炖猪耳朵和炒胡椒鸡肝。)

前句中的 dishes 是用于吃饭的物理物 (physical object), 后句中的 dishes 则是菜肴。它们的选择限制各不相同, 前者是 wash 的 PATIENT (受事), 它应该具有可洗性 (washable), 后者是 stir-fry 的 PATIENT (受事), 它应该具有可食性 (edible)。谓词选择其歧义论元的正确含义, 删除不能匹配的含义。由此可见, 使用选择限制实际上是一种“观其伴而知其意” (You shall know the word by the company it keeps) 方法。

当谓词有歧义时, 可以根据其论元 (argument) 的语义来消除歧义。例如:

Well, there was the time served green-lipped mussels from New Zealand. (好, 有时来品尝从新西兰来的绿唇蚌。)

Which airlines serve Denver? (哪一个航班到 Denver 去?)

Which ones serve breakfast? (哪一个航班提供早餐?)

前句中的 serve 要求某种食物作为其 PATIENT, 中句中的 serve 要求地名或者团体作为其 PATIENT, 后句中的 serve 要求某种饭局作为其 PATIENT。如果我们确信 mussel, Denver 和 breakfast 都是无歧义的, 那么, 就可以通过它们的语义来消除 serve 的歧义。

如果谓词和它的论元都有歧义, 则选择的可能性大大增加。例如:

I'm looking for a restaurant that serves vegetarian dishes.

Serve 有 3 个含义, dish 有 2 个含义, 则这个句子应该有 3×2 个含义。在这种情况下, 要根据谓词和论元的选择限制共同地决定其正确的选择。

谓词	论元
食物	可食性
地名或团体	可洗性
饭局	

因此, 这个句子的意思是“我正在找一个出售素食的饭馆”。

可见, 基于选择限制的词义排歧要求在语义分析中使用两方面的知识:

■ 论元的语义类型分类

■ 论元对于谓词的选择限制。

这两方面的知识都可以从词网 (WordNet) 中获取。语义类型分类的信息可以从有关词的上下位词 (hyponymy) 获得, 选择限制的信息通过把有关词的同义

词集(Synset)与谓词的论元相联系的方法获得。如果我们从词网上获得了这两方面的知识,我们就可以利用选择限制来进行词义排歧了。

然而,选择限制是有局限性的,主要表现在:

■当选择限制的一般性太强的时候,很难决定有关词的选择限制的范围。

例如,What kind of dishes do you recommend?

这里,我们难于决定 dishes 的选择限制是“可洗性”还是“可食性”。

■否定句中,否定关系明显地违反了选择限制,但是,句子的语义却是合法的。

例如,People realized you can't eat gold for lunch if you're hungry. (人们认识到,当你饥饿的时候,你不会把金子当作午餐来吃的。)

句子中的 eat gold 显然违反了 eat 的选择限制,因为 gold 不具有可食性。但是,由于有否定词 can't,这个句子却是完全合法的。

■当句子描述的事件是不寻常的事件时,尽管违反了选择限制,句子仍然是完全合法的。

例如,In his two championship trials, Mr. Kulkirni ate glass on an empty stomach, accompanied only by water and tea. (在他的两次冠军比赛中,Kulkirni 先生空腹吞食玻璃,吞食的时候只是喝点水和茶。)

句子中 glass(玻璃)是不具有可食性的,违反了 eat 的选择限制,可是,这个句子仍然是合法的。

■当句子中出现隐喻(metaphor)或借喻(metonymy)的时候,这样的隐喻或借喻是对选择限制的极大挑战。

例如,If you want to kill the Soviet Union, get it to try to eat Afghanistan. (如果你想扼杀苏联,那么,就让它去吞吃阿富汗吧。)

这时,谓词 kill 和 eat 的 PATIENT 的典型的选择限制都完全失效了,可是,这个句子的在语义上合法性却是毋庸置疑的。

1987年,Hirst指出,所有这些违反选择限制却在事实上合法的例子,都将导致词义排歧的失效。因此,他建议,与其把选择限制看成一种硬性的规定,不如把它看成是一种优选关系(preference),应该把优选的概念引入选择限制的研究中。

1997年,Resnik提出了“选择联想”(selectional association)的概念。选择联想是在谓词与该谓词所支配论元的类别之间的联想强度的一种概率测度。Resnik把 WordNet 中上下位(Hyponymy)关系与标注语料库中的动词—论元关系结合起来,从而推算选择联想的强度。Resnik用选择联想来进行词义排歧,算法选择在谓词与其论元的上位词之间具有最高选择联想的论元作为该论元的正确含义。

Risnik 选择联想方法的缺陷是,它只能用于谓词没有歧义而仅仅论元有歧义的场所。

4. 鲁棒的词义排歧方法

前面的方法是所谓“规则对规则”(rule-to-rule approach)的方法,另外,还有自立的方法(stand-alone approach)。自立的方法是一种鲁棒(robust)的词义排歧方法。鲁棒的自立的词义排歧方法依靠词类标注来工作,力求把对于信息的要求减低到最低限度,从而作到“自立”,让机器自己学习而获得信息。

这种机器学习的方法,要求对系统积极性训练,使得系统能够自行进行词义排歧。

要进行词义排歧的词叫做目标词(target word),目标词所嵌入的文本,叫做上下文(context)。输入按下面方式进行初始化的处理:

- 输入文本一般应该是经过词类标注的;
- 上下文可以看成是围绕目标词的长短不一的语言片段;
- 上下文中的单词,应该是经过词法分析的,应该把变形词还原成原形词;
- 文本最好是经过局部句法分析或者依存关系分析,能够反映出题元角色关系或者其他语法关系。

经过这样的初始化处理,输入文本要进一步提炼为包含相关信息的特征的集合。主要步骤是:

- 选择相关的语言学特征;
- 根据学习算法的要求对这些特征进行形式化描述(或者编码)。

大多数的学习系统使用简单的特征向量(feature vector),这些特征向量采用数字或者词类标记来编码。

用来训练词义排歧系统的语言学特征可以粗略地分为两类:

- 搭配特征(collocation feature);
- 共现特征(co-occurrence feature)。

搭配特征对目标词左右的上下文进行编码,要求指出特定的、能反映这些单词的语法性质的位置特征。典型的特征是单词、词根形式、词类等。这样的特征往往能把目标词特定的含义孤立起来以便处理。例如:

An electric guitar and bass player stand off to one side, not really part of the scene, just as a sort of nod to gringo expectations perhaps. (电吉他和低音乐器演奏者站在一旁,他并不是站在舞台的一部分,大概只是为了等待外国佬的到来。)

我们取特征词 bass(低音乐器)的左右两个词以及它们的词类标记为特征向量,表示如下:

[guitar, NN1, and, CJC, player, NN1, stand, VVB]

共现特征不考虑相邻词的精确的位置信息,单词本身就可以作为特征。特征的值就是单词在围绕目标词的环境中出现的次数。目标词的环境一般定义为以目标词为中心的一个固定窗口,要计算出在这个窗口中实词的出现频度。例如,对于目标词 bass,我们从语料库中选出它的 12 个共现词,然后标出它们在特定窗口中的出现频度。这 12 个共现词是: fishing, big, sound, player, fly, rod, pound, double, runs, playing, guitar, band.

在上面句子中选取 guitar and bass player stand 作为窗口,则其特征向量为:

[0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0]

根据这样的特征向量,由于第四个共现词 player 和第十一个共现词 guitar 在特征向量中的值都是 1,因此可以确定这个 bass 的词义是“低音乐器”。

在鲁棒的词义排歧系统中,一般都把搭配特征与共现特征结合起来使用。

5. 有指导的学习方法

有指导的学习方法(Supervised Learning Approaches)又可以再分为朴素的 Bayes 分类法和决策表分类法两种。

5.1 朴素 Bayes 分类法

使用朴素 Bayes 分类法(naive Bayes classifier)时,不是去寻找某个特定的特征,而是在综合考虑多个特征的基础上进行词义排歧。这种方法实际上是在给定的上下文环境下,计算一个多义词的各个义项中概率最大的义项。计算公式如下:

$$s = \operatorname{argmax}_{s \in S} P(s|V)$$

其中, S 是词义的集合, s 表示 S 中的每一个可能的义项, V 表示输入上下文中的向量。

直接根据向量的计算公式是:

$$s = \operatorname{argmax}_{s \in S} P(s) \prod_{j=1}^n P(v_j | s)$$

1992 年, Gale 等使用这个方法试验了 6 个英语的多义词(duty, drug, land, language, position, sentence)的词义排歧,正确率达到 90% 左右。

5.2 决策表分类法

决策表分类法(Decision list classifiers)根据共现词的等价类的不同制定决策表,然后将这张决策表用于输入向量,确定最佳的词义。

例如,可以制定如下的决策表来确定 bass 的词义:

规则		词义
窗口中出现 fish	◇	bass1
striped bass	◇	bass1
窗口中出现 guitar	◇	bass2
bass player	◇	bass2
窗口中出现 piano	◇	bass2
窗口中出现 tenor	◇	bass2
sea bass	◇	bass1
play/V bass	◇	bass2
窗口中出现 river	◇	bass1
窗口中出现 violin	◇	bass2
窗口中出现 salmon	◇	bass1
on bass	◇	bass2
bass are	◇	bass1

其中,bass1 表示 fish 的含义,bass2 表示 music 的含义。如果检测成功,就选择相应的词义,如果检测失败,那就进入下一个检测。这样一直检测到决策表的末尾,其缺省值就是最大可能的词义。

这个决策表可用于从 bass 的 music 含义中消除 fish 的含义。第一项检测说明,如果在输入中出现 fish,那么,就选择 bass1 为正确的答案。如果不是这样,就检测下一项一直到返回值为 True,在决策表末尾的缺省值的检测,其返回值为 True。

决策表中项目的排列根据训练语料的特征来决定。1994 年,Yarowsky 提出一种方法来计算决策表中的每个特征值偶对的对数似然比值(log-likelihood ratio),根据计算所得的比值调整 Sense1 和 Sense2 在决策表的顺序,从而确定整个决策表中特征值的排列顺序。计算公式如下:

$$\text{Abs} \left(\text{Log} \left\{ \frac{P(\text{Sense}_1 | f_i = v_j)}{P(\text{Sense}_2 | f_i = v_j)} \right\} \right)$$

其中,v 表示 Sense 的特征向量,f 表示该 Sense 的绝对频度。

根据这个公式来比较每个特征值偶对,便可以获得一个排列最佳的决策表。Yarowsky(1996)采用这样的方法,得到了 95% 的正确率。

6. 自举的词义排歧方法

有指导的学习方法的问题是需要训练大量的标注语料。Hearst 在 1991 年,