现代汉语词语级歧义自动消解研究

曲维光 著

科学出版社
www.sciencep.com

现代汉语词语级歧义自动消解研究

曲维光 著

科学出版社

北京

内 容 简 介

本文提出基于词语搭配强度计算的语境计算模型RFR_SUM(SUM of Relative Frequency Ratio),用于处理各类词语级的歧义消解问题。各章节的顺序大致勾勒出该模型形成和发展的轨迹。提出广义组配理论框架,并据此建立语境计算模型RFR_SUM,用以处理语言中广泛存在的词语级歧义现象。将RFR_SUM模型应用于中文信息处理中的组合型切分歧义和交集型切分歧义的消解、兼类词的消解、多音词的消解以及词义消歧、语料库精加工、隐喻识别等多项任务中,均取得满意的结果,验证了该理论的普适性。本书可以作为从事自然语言处理和计算语言学相关研究人员的参考书。

图书在版编目(CIP)数据

现代汉语词语级歧义自动消解研究/曲维光 著.-北京:科学出版社,2008

ISBN 978-7-03-023646-3

Ⅰ.现… Ⅱ.曲… Ⅲ.汉语-词语-研究-现代 Ⅳ.H136

中国版本图书馆 CIP 数据核字(2008)第 195640 号

责任编辑:侯沈生　　责任校对:袁海滨

责任印制:李延宝　　封面设计:张祥伟

科学出版社出版

北京东黄城根北街 16 号

邮政编码:100717

http://www.sciencep.com

丹东印刷有限责任公司印刷

科学出版社发行　各地新华书店经销

*

2008 年 12 月第一版　　开本:A5(890×1240)

2008 年 12 月第一次印刷　　印张:8 1/2

印数:1—3000　　字数:280 000

定价:28.00 元

(如有印装质量问题,我社负责调换)

序 一

欣闻曲维光博士的专著《现代汉语词语级歧义自动消解研究》即将出版，我由衷地感到高兴。曲维光博士要我写个序言，实在是盛情难却。为他人的著作作序，在我的学术生涯中还是第一次。我以为，写“序言”是一件极其困难的任务，不仅要领会全书的精要，还要了解相关学科的全局以及该书对学科发展的贡献。就能力和精力而言，我确实难以胜任。然而，曲维光博士2006年初进北京大学计算机科学技术博士后工作站，两年期间与我密切合作。他不仅刻苦努力，勤于思索，出色完成了博士后研究任务，为我承担的973课题“文本内容理解的数据基础”贡献了力量；而且富有协作精神，与北京大学计算语言学研究所师生结下了深厚的友谊。同时，我知道曲维光博士的导师陈小荷教授已经为本书写了序言，相信“序言”的任务已经完成。我自觉压力不那么大了，只不过是再加上自己的读后感而已。

当前自然语言处理研究的主攻方向，是让机器能够自动地识别和消解自然语言的歧义。曲维光博士的研究重点是词语级的各种类型的歧义消解，这是自然语言处理研究的基本课题，已经研究很多年了，但没有彻底解决，甚至离彻底解决尚有很长的路要走。这种情况一方面说明，这里有创新的机会和发展的空间，另一方面也说明，创新和发展的难度很大。可以说，曲维光博士是在打攻坚战。

任何一个语言单位脱离其语境（不妨狭义地理解为该语言单位的上下文）都有可能产生歧义，消解歧义的所有方法都要利用其语境信息。不同的问题、不同的方法所利用的语境的范围各不相同。就词语级歧义而言，语境通常约束为研究对象在语句中左右相邻的若干个词语。曲维光博士提出的语境计算模型RFR_SUM利用了研究对象在整个语料库中的相关信息，取得了很好的消歧效果。这是本书最重要的创新成

果，值得向读者推荐。在这里试做一个浅显的解说。

RFR_SUM的完整表达是 SUM of Relative Frequency Ratio，SUM 就是算术“和”，而 Relative Frequency Ratio 书中解释为“相对词频比”。

设研究对象 A 有两个歧解 A_1 和 A_2。例如，字符串“学会”可以是一个名词“学会/n”，也可以是两个动词的组合“学/v 会/v”；“黄色”这个词的词义是“颜色”，另一个是“淫秽”。从包含对象 A 的语料库 C 中抽出两个分别包含 A_1 和 A_2 的子集 C_1 和 C_2。这里需要指出，语料库 C 包含的不是原始文本，而是按照需要进行了加工的带标记语料。

对于对象 A_1，将 C_1 中所有语句按 A_1 对齐，统计相对于 A_1 的每个位置 i（如左 1、左 2、右 1、右 2，等等）上的词语 w 的词频（即出现次数），称为词语 w 在位置 i 上的局部词频，记为 $LocFrq_i(w)$。称词语 w 在语料库 C 中的词频为全局词频，记为 GlobFrq(w)。局部词频与全局词频之比，即 $LocFrq_i(w)/GlobFrq(w)$，就是词语 w 关于 A_1 在位置 i 上的相对词频比 $F_i(w)$。

这里有关“词频”的各个术语有其特定的含义，与词频统计[①]中普遍使用的“频次（绝对频率）”、“频率（相对频率）”这些术语有相通之处，也有不一致，希望读者注意。

对于每个位置 i，将 $F_i(w)$ 按降序排列，选择最大的前 n 个词语。将每个位置 i 作为列，将前 n 个词语及其相应的 $F_i(w)$ 作为行，排列成表。对于对象 A_2，也有同样的一个表。这些表就是语境计算模型 RFR_SUM在训练语料 C 上所获取的最基本的参数。

曲维光博士将RFR_SUM计算模型应用于组合型切分歧义、交集型切歧义、多音词、兼类词、多义词的消歧以及隐喻的识别等任务，通过大规模实验验证了理论的可靠性，特别是对小概率的研究对象取得了令人满意的结果，这也是上面所提及的“攻坚战”的另一层含义。

我以为，曲维光博士之所以能取得这样的成绩，与其学术背景有密

① 参见国家语言资源监测与研究中心编《中国语言生活状况报告（下编）》，北京：商务印书馆，2005

切的关系。曲维光博士在大学本科和硕士阶段都是学计算机的，攻读博士学位则选择了计算语言学作为自己的研究方向，因而理科、文科都具有坚实的基础。曲维光博士不仅勤于观察分析纷繁复杂的语言现象，提出了广义组配理论；而且勇于在大规模的语言数据上进行实践，这才有语境计算模型RFR_SUM的创新。

学海无涯。无论是广义组配理论还是语境计算模型，都还有很多工作要做，需要进一步充实、完善、验证、发展。如果放眼自然语言理解的长远目标，当前的自然语言处理研究所取得的一些成就或许只能算作是一部伟大乐章的前奏曲。我用屈原的两句话作为自己的座右铭："路漫漫其修远兮，吾将上下而求索。"愿与曲维光博士共勉之。

北京大学计算语言学研究所

俞士汶

2008 年国庆前夕完稿

序　二

自然语言处理的许多问题可以归结为分类问题。例如，自动分词可以看成是将文本中的每两个字符之间的联系区分为词界和非词界，词性标注、词义标注则是将文本中的词从语法或语义角度进行分类。自动分词中的歧义（组合型歧义或交集型歧义）字串的处理也是一个分类问题，是把这些字串划分为切开和不切开、或者切在什么位置等类别。甚至隐喻识别也可以认为是分类问题，是把文本中的词划分为隐喻义和普通义两类。分类需有依据，分类需要知识。对于计算机来说，把分类所需的知识以及获取和运用这些知识的方法概括起来，就形成所谓的语言计算模型。又，如果一个语言成分或关系既可以归入这个类别，也可以归入那个类别，那么对于计算机来说，这就是一种歧义，所以在自然语言处理中分类问题经常称作歧义消解（简称消歧）问题，如分词歧义消解、词性消歧、词义消歧。

曲维光博士对歧义消解问题特别关注，在他的博士论文中提出了一种用于词语消歧的语境计算模型，后来又取了一个简洁的名字RFR_SUM，意思是相对频率比(Relative Frequency Ratio)求和。相对频率比是这个模型中的核心概念，即词语在某类样本中的相对频率除以它在整个语料库中的相对频率。这里所说的语境是指待分类的词语的简单上下文(Context)。分类也好、消歧也好，反正都是要利用上下文中透露的信息来进行计算。根据我的理解，RFR_SUM模型包括以下三个部分：

一、分类所需的知识。多分类问题可以转化为有限个二元分类问题，因此为简单起见，假设目标是将词语分为A、B两类，分别计算A类和B类的上文和下文中各个词语的相对频率比共四组数据。

二、获取知识的方法。首先对整个语料库进行词频统计，然后对其中的两个分类样本在观察窗口中的词语计算样本频率，于是可求得样本

中每个词语的相对频率比。

三、运用知识的方法。把上下文中的词语在A类和B类中的相对频率比分别相加，选择相对频率比之和较大者作为分类结果。

第一部分特别重要，因为作者就是这样来定义分类所需的语言知识的，把词的类别（包括词的意义）看成是由上下文中其他词语所共同决定的。定义中其实已经蕴含着获取和运用这些知识的方法，即语境计算。

一般来说，离待分类词语越近的词语，其分类作用越大。如果计算相对频率比的时候根据距离加权，效果是否会更好一些呢？这只是一个偶发的想法，需要通过实验来求证。

曲博士将RFR_SUM看成一种词语级的消歧模型，这个定位是准确的。它不是像隐马尔科夫模型、条件随机场那样用来解决序列标注问题（对序列中每个符号进行分类），而是用来解决单个词语的分类问题。RFR_SUM跟点式互信息（Pointwise Mutual Information）有些相似，但比后者的应用面更广、分类效果更好。从这里我们看到一种词汇主义倾向。序列标注模型着眼于整个序列的概率最大化，但对于单个词语的分类难免会有不同程度的粗糙处理。词语级消歧模型则利用各类相关样本中的颗粒度更小的语言知识，在单个词语的分类问题上取得了突出的成绩。以一阶隐马尔科夫模型为例，它计算单个标记的概率时，只考虑前一个标记到当前标记的转移概率以及当前标记产生当前词的发射概率，跟RFR_SUM模型相比，所观察的上下文范围是十分有限的，对上下文信息的利用也是比较粗糙的。

RFR_SUM模型是在作者提出的广义组配理论的基础上建立起来的。该理论的核心是三个假设：词频和分布在大规模语料库中的稳定性、在具体样本中的特殊性以及在对立样本中的区别性。其中，稳定性是语境计算的根本依据，特殊性和区别性上也体现了稳定性。唯有稳定，从训练集中得到的知识才可能成功地运用于测试集。唯有特殊并且几种特殊性能够相互区别，才可能成功地将对立样本加以分类，消除歧义。在RFR_SUM模型中，词的分布简化为两种位置：出现在待分类词语的前面、出现在待分类词语的后面。词频则聚焦为相对词频比以显示出对立

样本的区别性。作者将这个模型应用于组合型歧义消解、交集型歧义消解、词性歧义消解、词义消歧、隐喻识别等诸多任务，都取得了比别的模型更好的成绩，有的任务所获得的分类精度甚至有大幅度的提高。实验表明，广义组配理论的三个假设是符合语言实际的。

广义组配理论把词语组合关系分为固定组配、自由组配和共现组配三种类型并给予了清晰的界定。曲博士认为，“构成句子的任何词语之间都有一定的吸引关系，这是由语言系统各种规则共同作用而产生的。词语之间相互作用的程度有大有小，但都共同参与组词成句的功能。”这里一个突出的例子是虚词“的”。在一些统计模型的应用中，例如用向量空间模型来消解组合型歧义时，虚词是不予考虑的。但是事实上，“学会”后面的“的”以及“才能”前面的“的”，都明显支持“学会”、“才能”为名词，不应切开。从语言学角度来看，自动分词中许多组合型歧义实例的鉴别，既涉及语义问题也涉及语法问题，歧义消解时将上下文中的虚词排除是不合适的。广义组配理论不预设上下文中哪些词的启发性作用更大，这就为语法、语义等上下文信息的综合运用提供了广阔的空间。

在师从俞士汶教授做博士后研究时，作者对RFR_SUM模型做了进一步的完善，主要有三点：第一，将RFR_SUM模型从一元扩展为二元，增强了模型对语言的刻画能力；一元模型与二元模型相互配合，大大提高了系统的性能。第二，把二值分类器改造为多值分类器，增强了模型的实用性。第三，通过对 RFR 表剪枝（去掉上下文中的低频词语）将模型的数据量缩减为原来的 5%，速度提高 20 倍，但分类精度基本保持。

自然语言处理中大多数任务属于序列标注，能否将RFR_SUM也扩展成一种序列标注模型呢？如果现代汉语词语级歧义自动消解研究只是简单地对序列中的每个词进行RFR_SUM分类，首先一个问题就是算法的时间复杂度太高。这部书稿中提到，即使经过剪枝，对 1000 个句子的某个特定词做词义消歧也需要若干秒。更重要的问题是，各词的分类结果是否能够相容并使得整个序列的概率接近最大化？所以可能需要借鉴现有的一些序列标注模型的思想来进行扩展。

自然语言处理的历史并不长，国内外的研究基本保持同步。但是无庸讳言，我们使用的计算模型几乎都是从国外引进的，缺乏自己的创造。我们总是在不断地重复“学习—消化—移植到汉语”这样一个过程。这种过程虽然是必要的，但不应是唯一的。曲博士的RFR_SUM模型在这方面开了一个好头。我期望并相信他能把这个研究继续下去，最好是也把它应用于英语等语言的信息处理，做出来的成绩可以跟国外同行相比。

陈小荷

2008年暑期于南京白云园

目　录

序　一
序　二
绪　论 …… (1)
1　自然语言处理的根本问题 …… (1)
2　词语搭配问题的研究 …… (5)
3　本书的主要研究内容 …… (9)

第 1 章　词语组配的研究现状 …… (13)
1.1　汉语词语组配及其性质 …… (13)
1.2　国外词语搭配研究现状 …… (21)
1.3　国内词语搭配研究现状 …… (22)

第 2 章　词语搭配的自动抽取研究 …… (27)
2.1　词语搭配的抽取方法 …… (28)
2.2　搭配抽取框架的建立 …… (39)
2.3　实验及其结果 …… (41)

第 3 章　广义组配理论 …… (45)
3.1　广义组配理论的提出 …… (46)
3.2　语境的可计算性 …… (47)

第 4 章　语境计算模型RFR_SUM …… (55)
4.1　相对词频比 RFR …… (57)
4.2　基本RFR_SUM模型 …… (64)

第 5 章　RFR_SUM模型在分词消歧中的应用 …… (67)
5.1　RFR_SUM模型应用于组合型消歧 …… (67)

5.2 RFR_SUM模型应用于交集型消歧 …… (76)

第6章 兼类词与多音词的消歧 …… (87)
6.1 RFR_SUM模型在兼类词消解中的应用 …… (87)
6.2 基于RFR_SUM模型的多音词的消歧 …… (97)

第7章 词义消歧研究 …… (102)
7.1 RFR_SUM模型在词义消歧中的应用 …… (102)
7.2 无需词性标注语料的词义消歧实验 …… (111)

第8章 词义消歧的二元模型及集成研究 …… (115)
8.1 BI_RFR_SUM模型 …… (116)
8.1.1 二元搭配强度和二元相对词频比(BI_RFR) …… (116)
8.1.2 BI_RFR_SUM模型 …… (118)
8.1.3 实验及结果 …… (120)
8.2 UNI_RFR_SUM 与 BI_RFR_SUM的集成 …… (124)
8.3 多分类问题研究 …… (126)

第9章 超大规模语料精加工技术研究 …… (135)
9.1 问题的提出 …… (135)
9.2 现有标注软件的性能指标的计量研究 …… (138)
9.2.1 ICTCLAS 系统标注结果分析 …… (140)
9.2.2 系统改进探讨 …… (144)
9.3 语料精加工的方法 …… (147)
9.3.1 词表校对法 …… (147)
9.3.2 基于简单词语组合特性的方法 …… (149)
9.3.3 基于多元组比对的方法 …… (149)
9.3.4 基于RFR_SUM模型的方法 …… (152)
9.4 初步实验结果 …… (158)

第10章 隐喻识别研究 …… (159)
10.1 隐喻研究现状 …… (159)
10.2 隐喻研究的意义 …… (166)

10.3　隐喻研究的内容和方案 …… (167)
10.4　初步的研究成果 …… (171)

结　语 …… (174)
1　本研究完成的主要工作 …… (174)
2　进一步研究计划 …… (176)

主要参考文献 …… (179)

附录 1　北京大学汉语文本词性标注集 …… (189)

附录 2　组合型切分歧义强弱势比例 …… (191)

附录 3　“从小/学”训练用例句 …… (193)

附录 4　“应/用于”训练用例句 …… (197)

附录 5　“应用于”测试集 …… (203)

附录 6　“从小学”测试集 …… (211)

附录 7　“科学”词性标注开放测试中标注错误句子 …… (214)

附录 8　“黄色”词义消歧中错误句子 …… (216)

附录 9　“黄金”词义消歧中错误句子 …… (224)

附录 10　经改进后,“黄金”词义消歧中错误句子 …… (235)

附录 11　经改进后,“黄色”词义消歧中错误句子 …… (238)

附录 12　“黄色”词义开放测试错误句子 …… (240)

附录 13　“黄金”词义开放测试错误句子 …… (241)

附录 14　“分子”分类错误的句子 …… (243)

附录 15　“材料”分类错误的句子 …… (244)

附录 16　“着/u”和“着/v”校对出错误的句子 …… (246)

附录 17　“本书/r”和“本/q 书/n”校对出错误的句子 …… (251)

后　记 …… (253)

绪　论

1　自然语言处理的根本问题

《圣经·创世纪》说，人类的先民原来拥有统一的语言，交流思想非常方便，劳动效率也很高。他们曾想建立一座高达天庭的通天塔叫做“巴比塔”，以此来展示他们的伟大才能。人类建造巴比塔的壮举震惊了上帝，上帝便施展权威，让不同的人讲不同的语言，使人们难以交流思想，无法协调工作，以此来惩罚异想天开的人类。结果，巴比塔没有建成，而语言的不同却从此成为人们相互交往的最大障碍[52]。

1946年，世界上第一台电子计算机ENIAC（电子数字积分计算机的简称）在美国宾夕法尼亚大学宣告诞生。计算机与人类的共同之处在于都使用语言。与人类语言不同，计算机语言经过了良性定义，不存在歧义。但计算机的语言，从机器代码到汇编语言再到高级语言，虽然可读性大为提高，但仍然无法为大众所接受，成为可以相互交流的语言。人们喜爱的还是那些让人花费大量精力去学习、令大家绞尽脑汁去揣摩的种类繁多、充满歧义，但情感丰富、韵味无穷的自然语言，而不是严格定义、理性有余，却冷若冰霜的机器语言。

随着信息时代的到来，人们与计算机的交流日趋频繁。同时，当人们相互之间由于语言不通而难以交流时，也希望计算机的发展能够更进一步，成为人们跨越语言鸿沟的桥梁。如何能够让没有太多计算机专业

知识的人们也能顺利地同计算机进行交流，这个问题便成为摆在计算机工作者和语言学家面前的一项迫在眉睫的任务。对这个问题的深入研究，大大促进了计算语言学以及语言信息处理的发展。

张普先生曾按照表达方与理解方对象的差异，将信息时代的交际行为表示为如下四种模式[111]：

A. 人表达—— 人理解

B. 机器表达——人理解

C. 人表达—— 机器理解

D. 机器表达——机器理解

其中，研究电脑如何表达人的语言（模式 B）是“自然语言生成”，研究电脑如何理解人的语言（模式 C）就是“自然语言理解”。机器翻译需要电脑理解一种自然语言，然后转换生成另外一种语言，所以既包括自然语言理解研究，也包括自然语言生成研究（属于模式 D），还包括语言之间的转换研究。事实上，当进行 A 模式交际时，如果一方表达而对方无法理解时，人们也寄希望于计算机的帮助，去完成原来需要人类翻译所做的工作。这时，就需要 C 模式（人表达—机器理解）和 B 模式（机器表达—人理解）。此时，计算机便成为人与人进行交流的桥梁。

自然语言分析的关键就是识别与消解自然语言的歧义。人与人的交流由于有共同的知识背景，并且能领会交流的环境和过程，通常不会产生误解。但是，作为语言学研究对象的任何一个语言单位，如词、短语和句子等，如果脱离语境而孤立存在，通常都是有歧义的。当交流在人和机器之间进行时，由于机器尚不具备“背景知识”和“世界知识”，歧义现象就表现得尤为突出。

汉语信息处理很难回避的一个步骤就是把用汉字序列书写的句子切分为词的序列，或者说从句子中辨识出词。在这个最基本的步骤中，就存在大量的歧义。例如，仅“白天鹅”这 3 个汉字组成的序列就存在歧义[106]：是“白/天鹅/”还是“白天/鹅/”？如果这 3 个字的序列落在更长的汉字序列中，歧义就可能得以消解。

白天鹅飞过来了——白/天鹅/飞/过来/了/（因为鹅不会飞）

白天鹅可以看家——白天/鹅/可以/看/家/（家里通常不会养天鹅）

人如何消解歧义呢？当然是根据业已掌握的知识。人也可以把这些知识教授给计算机，存储在知识库中，计算机据此也可以消解这样的歧义。但是如果“白天鹅”落在句子“白天鹅在湖里游泳”中，仅依靠存储在人脑或电脑中的静态知识，是无法判定句中的“白天鹅”这3个字应该如何切分的，必须依赖更大的上下文语境。

动物园里，白天鹅在湖中游泳。

——动物园/里/，/白/天鹅/在/湖/中/游泳/。/

白天鹅在湖里游泳，夜晚蛙在池边鸣唱。

——白天/鹅/在/湖/里/游泳/，/夜晚/蛙/在/池/边/鸣唱/。/

我们把像“白天鹅”这样“白”、“天鹅”和“白天”、“鹅”都成词的歧义叫做交集型切分歧义。除此之外，汉语切分中还存在称为组合型的切分歧义，例如“都会”，是“都/会/”两个词，还是“都会/”一个词，也需要在确定的上下文中才能区分清楚：

我们/都/会/游泳/。/

哈尔滨/是/个/国际化/大/都会/。/

词语切分确定下来之后，还有歧义。见下例：

老子不在家——老子/不/在/家/

这里的“老子”如果读“lǎo zǐ”，是指古代的人物；如果读“lǎo zi”，则可能指“父亲”，或者指“自己”。以上句子中“子”的读音不同可以造成意义的不同，而同音词也会形成另外的歧义。下面几个例子中“连”的读音是一样的，但词性不同（当然，词义也不相同）：

一个连有三个排——“连”是名词，指军队的建制

我们兄弟心连心——“连”是动词，“连结”的意思

苹果可以连皮吃——“连”是介词，“带”的意思

当词语切分和词性标注得以正确解决之后，还会面临一词多义的问题。下面的句子：

他是个唱歌的好材料。

他们研制出新型建筑材料。

他们在整理人事材料。

这三个句子中的“材料”形态一样、读音一样、词性也一样，但是词义却各不相同：“唱歌的材料”指的是人，“建筑材料”、“人事材料”指的是物。“建筑材料”是指可以建筑高楼的东西，而“人事材料”指的则是可供参考的事实。

除了上述句子内的切词、多音词、词性、词义等都有歧义现象外，其他的语言求解问题，诸如断句（现代汉语尽管有标点符号，但是确定句法和语义相对完整、又不过长的句子仍然是个难题）、指代、省略也可归结为歧义问题。这里还没有涉及到句法结构、语义角色以及语用表达等方面存在的歧义。

让计算机理解符合规则（词法、句法、语义）的自然语言的语句和文本已经是一项十分困难的任务，不同语言单位的各种形态的歧义已经让研究者左支右绌，力不从心。进而，当自然语言处理面对语言中的各种修辞手法时，又会遭遇怎样的困难呢？

隐喻是修辞学的传统研究内容，运用隐喻是为了提高语言表达效果。作为一种修辞手段，隐喻可以归于文学语言的范畴，但从认知语言学的角度去观察，隐喻无处不在，因此它又不限于文学语言的范畴。认知语言学甚至认为“隐喻不仅仅是语言修辞手段，而且是一种思维方式——隐喻概念体系。作为人们认知、思维、经历、语言甚至行为的基础，隐喻是人类生存主要的和基本的方式[11]。”在计算语言学领域，特别是在汉语信息处理领域，中国大陆学者只是近年来才开始关注“隐喻”的识别和求解[107][113]。不过，语言信息处理要走上自然语言理解的坦途，隐喻是必须逾越的路障。

根据包含隐喻的语言单位的大小将隐喻划分为词汇级、语句级和篇章级。就隐喻的自动理解研究而言，三个级别的难度逐次增强。

汉语词汇中有许多像“山头”、“墙脚”、“垃圾”这样一些词语，它们除了本义之外，在词典中都具有隐喻义。例如，在《现代汉语词典》中“垃圾”有两个义项：

①脏土或扔掉的破烂东西。

②比喻失去价值的或有不良作用的事情。

只要词典（或机器中的词汇知识库）登录了这些词语的各种义项（包