Constructing the affective lexicon ontology

Article · January 2008

CITATIONS
105

READS
924

5 authors, including:

Hongfei Lin
Dalian University of Technology
304 PUBLICATIONS **1,845** CITATIONS

SEE PROFILE

情 报 学 报 ISSN 1000 – 0135
第 27 卷 第 2 期 180 – 185，2008 年 4 月

JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION ISSN 1000 – 0135
Vol.27 No.2，180 – 185 April 2008

情感词汇本体的构造[1)]

徐琳宏 林鸿飞 潘 宇 任 惠 陈建美
（大连理工大学计算机科学与工程系，大连 116024）

摘要 情感计算是目前人工智能领域的热门课题，而大规模的情感词汇本体的构造是准确完成文本情感识别的基础。本文首先根据目前情感分类发展的现状，确定情感分类体系，在此基础上综合现有的各种情感词汇资源构造情感词汇本体。在本体的知识获取过程中采用手工分类和自动获取相结合的方法填充词汇本体的框架。详细描述了词汇的情感类别、强度和极性等，并进一步统计了情感词汇的分布情况。

关键词 情感计算 情感分类 互信息 本体

Constructing the Affective Lexicon Ontology

Xu Linhong，Lin Hongfei，Pan Yu，Ren Hui and Chen Jianmei
（*Department of Computer Science and Engineering*，*Dalian University of Technology*，*Dalian* 116024）

Abstract Affective computing has received more and more interests in the field of artificial intelligence. However，constructing the affective lexicon Ontology is the basis of textual affective computing. Firstly，the paper analyzes the status of the emotional classification，and then classification system is determined. Finally，affective lexicon ontology which synthesizes various resources is constructed. In the process of acquiring the knowledge，the framework of ontology is filled by the combination of manual classification and acquiring the intensity automatically. The paper describes emotional classification and lexical intensity etc，and the distribution of affective lexicons.

Keywords affective computing，emotional classification，mutual information，ontology

1 引 言

情感计算是人工智能方面一个热门的研究领域，它的目标是使计算机拥有情感，这主要包含两个方面，即计算机能识别人类的情感和表达自身的情感。目前国内外对情感计算的研究主要集中在图像、声音、生理信号和文本几方面。随着 Internet 的发展，以文本形式出现的信息越来越多，逐渐成为我们最容易获取也是最为丰富的一种交互资源。从大量的文本中提取其中包含的情感信息在许多方面都有广阔的应用前景，如邮件系统、个性化文本、解析文章情感结构、网页评价等。现在主要的研究成果有 EmpathyBuddy 邮件系统[1]，颜色条表示文章的情感结构[2]，动画文本[3]、名人的评价[4]等。

一篇文章是由许多句子组成的，而每个句子又由若干的词汇构成，因此对词汇的情感色彩的理解是分析整篇文章的情感色彩的基础。目前关于情感词汇方面的资源较少，国外主要有 WordNet。可以利

收稿日期：2007 年 3 月 5 日

作者简介：徐琳宏，硕士生，研究方向：情感计算。林鸿飞，博士，教授，博士生导师，研究方向：搜索引擎、文本挖掘和自然语言理解。E-mail：hflin@dlut.edu.cn。潘宇，硕士生，研究方向：信息推荐和情感计算。任惠，硕士生，研究方向：情感计算。陈建美，硕士生，研究方向：情感资源建设和情感标注。

1) 基金项目：国家自然科学基金资助项目（编号：60373095，60673039）和国家 863 高科技计划资助项目（编号：2006AA01Z151）。

用 WordNet Domain 对其进行分类,划分出 WordNet-Affect 类。国内在这方面的资源还比较少。本文在综合现有多种词典和语义资源的基础上,构建了一个情感词汇的本体,旨在为段落和篇章级的情感分类提供基础和依据。

2 情感分类

到目前为止,心理学界对情感的划分还没有一个公认的标准。情感的分类有 4、6、8、10 乃至 20 余类不等,这主要是因为人类的情感复杂多变,并且人们对情感的认识还不够深入和全面导致的。但是对情感划分的研究仍在不断的进步和发展中,主要的情感分类方法有以下几种:

6 类:高兴,悲伤,愤怒,恐惧,厌恶和惊奇[5]。

Plutchik 在 1960 年提出的 8 种纯情感有:快乐,悲伤,愤怒,恐惧,期望,惊奇,憎恨,接受。并认为其他复杂情感都是这些情感混合而成。

12 类:高兴,悲哀,恐惧,厌恶,愤怒,惊奇,喜爱,期待,焦虑,内疚,赞扬,羞[6]。

Plutchik 等提出的 8 大类情感:狂喜,警惕,悲痛,惊奇,狂怒,恐惧,接受,憎恨。

中国传统的"七情"大致分为:好,恶,乐,怒,哀,惧,欲。

心理学家林传鼎将情绪划分为 18 类:安静,喜悦,恨怒,悲痛,哀怜,忧愁,忿急,烦闷,恐惧,惊骇,恭敬,抚爱,憎恶,贪欲,嫉妒,骄慢,惭愧,耻辱[7]。

许小颖等人将情感词汇划分为基于心理感受和基于表现力的两大类,其中将基于心理感受的词汇又细化为 24 类:喜,乐;爱;愁,闷;悲;慌;敬;激动;羞,疚;烦;急;燃;吃惊;怒;失望;安心;恨;嫉;蔑视;悔;委屈;谅;信;疑;其他。将基于表现力的词汇细化为态度词、品性词、声音词和其他[8]。

仇德辉等人提出人的情感可分为对物情感、对人情感、对己情感以及对特殊事物的情感四大类,其中对特殊事物的情感又细分为对他人评价的情感、对交往活动的情感、对不确定事物的情感和对自身状态的情感。

上面所列出的情感分类方法是目前国内外比较有影响的情感分类方法,在参照这些分类体系的基础上,综合现有的情感词汇资源,本文将情感分为 7 大类、20 小类。具体划分如表 1。

表 1 情感分类

编号	情感大类	情感类	例词
1	乐	快乐	喜悦、欢喜、笑咪咪、欢天喜地
2		安心	踏实、宽心、定心丸、问心无愧
3	好	尊敬	恭敬、敬爱、毕恭毕敬、肃然起敬
4		赞扬	英俊、优秀、通情达理、实事求是
5		相信	信任、信赖、可靠、毋庸置疑、
6		喜爱	倾慕、宝贝、一见钟情、爱不释手
7	怒	愤怒	气愤、恼火、大发雷霆、七窍生烟
8	哀	悲伤	忧伤、悲苦、心如刀割,悲痛欲绝
9		失望	憾事、绝望、灰心丧气、心灰意冷
10		疚	内疚、忏悔、过意不去、问心有愧
11		思	相思、思念、牵肠挂肚、朝思暮想
12	惧	慌	慌张、心慌、不知所措、手忙脚乱
13		恐惧	胆怯、害怕、担惊受怕、胆颤心惊
14		羞	害羞、害臊、面红耳赤、无地自容
15	恶	烦闷	憋闷、烦躁、心烦意乱、自寻烦恼
16		憎恶	反感、可耻、恨之入骨、深恶痛绝
17		贬责	呆板、虚荣、杂乱无章、心狠手辣
18		妒忌	眼红、吃醋、醋坛子、嫉贤妒能
19		怀疑	多心、生疑、将信将疑、疑神疑鬼
20	惊	惊奇	奇怪、奇迹、大吃一惊、瞠目结舌

首先将情感分为 7 大类,这是在国外比较有影响的 Ekman 的 6 大类情感的基础上划分的。因为 6 大类情感中的积极情感只有"高兴"一类,刻画得不够细致,所以本文又在"乐"的基础上增加了"好"一类来描述喜好、喜欢类型的情感。同时这 7 大类情感也基本上与中国传统的"七情"说法一致,只是少了"欲"一类。这是因为在本文目前的语义资源中描述"欲"类的情感词汇较少,所以没有单独划分出来作为一大类。

确定了 7 大类情感后,在每个大类内按照情感强度和复杂度的区别细化情感大类,最终分为 20 小类。如"安心"和"快乐"同属于"乐"类,但是"安心"类的词汇快乐的强度大部分要弱于"快乐"类。20 个情感小类的划分主要是参考林传鼎和许小颖等的情感分类方法,将一些词汇量较少、意思比较接近的情感类别合并。将情感词汇分为 20 小类主要是为了详尽地划分大部分的情感词汇,并为以后情感类

别的增加、减少和细化等提供方便。在情感类别的划分过程中,预先确定了一个初步的情感划分,然后将部分词汇按照初步的划分方法分类,当某些词汇的情感不在现有的划分中时,则根据该类词汇的性质和数量考虑是否增加该类情感。本文预先录入约500个情感词汇,采用上述类似基于转换的错误驱动学习方法,修正初步的情感划分,最终确定了表1中的20个情感分类。由于情感的发展具有"绝对的变化性"和"相对的稳定性",所以随着社会的不断发展,情感的分类也不是一成不变的。文中的情感分类体系也可以随着时间的推移,不断的修正,如增加一些分类来更准确地描述词汇的情感信息,具有一定的可扩展性。

3 词汇本体的建设

本文给出的情感词汇本体利用20类基本情感描述词汇的情感信息,并对每个词汇从极性和强度等多个方面进行描述。

3.1 词汇本体的描述框架

情感词汇本体通过一个三元组来描述:

Lexicon = (B, R, E)

其中,B 表示词汇的基本信息,主要包括编号、词条、对应英文、词性、录入者和版本信息。

R 代表词汇之间的同义关系,即表示该词汇与哪些词汇有同义的关系。该部分主要参考哈尔滨工业大学的同义词词林,从同义词中人工挑选具有情感色彩的词汇录入,然后修改具有同义关系的一组词的"syn"域,以记录情感词汇之间的同义和近义关系。

E 代表词汇的情感信息,是情感词汇描述框架中比较重要的一部分。

情感认知中有基本情绪论和维度论两种不同的研究途径。基本情绪论认为情绪有几种原型,其他情感是在基本情感的基础上演化和综合而来[9]。而维度理论是用几个维度空间来描述人类的情感。人们普遍认同的是"大二"(Big Two)模式,快乐度和唤醒度。但是两者并不是完全矛盾的,例如可以把"快乐"和"不快乐"看作两种基本情感。本文在 E(情感信息)中就综合利用上述两种途径描述词汇的情感信息,分别通过情感分类、强度和极性3个维度描述。每个情感词汇可以同时拥有多种情感分类,并且对每个分类都有一个强度的等级。

例如,词汇"惊喜"的描述如下:

〈num〉APA00032〈/num〉
〈lex〉惊喜〈/lex〉
〈ccat〉a〈/ccat〉
〈eng〉 pleasantly surprised〈/eng〉
〈emotion〉PA〈/emotion〉
〈intensity〉7,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,5〈/intensity〉
〈polarity〉1〈/polarity〉
〈syn〉〈/syn〉
〈emotion_class〉A〈/emotion_class〉
〈standard〉0〈/standard〉

其中〈emotion〉域中的"PA"是快乐类情感的编号。〈intensity〉域是采用向量的形式表明词汇都包含哪些情感以及强度等级,该域由20个0~9之间的数字组成,每个分量分别代表一个情感分类。0表示不含该类情感,1~9表示包含该类情感。因为"惊喜"包含"快乐"和"惊奇"两种情感,所以分别在两个情感的相应分量上用7和5表示。强度分为1、3、5、7、9五个等级。对于"惊喜"来说,在"快乐"上的等级为7,在"惊奇"上的等级为5,表明主要情感是快乐。〈polarity〉表明词汇极性有褒义、贬义、中性、褒贬兼有4类。〈emotion_class〉表明词汇包含的主要情感是消极、积极还是中性。

上述描述框架包含了词汇的静态和动态两方面的属性,并从定量和定性两方面表示词汇的情感信息,为以后段落乃至篇章级的情感分析和褒贬义识别提供更多的参考和利用的信息。

3.2 本体的知识获取

3.2.1 知识的来源

情感本体的基本知识主要来源于现有的一些词典、语义网络等。其中词典包括《现代汉语分类词典》[10]、《汉语褒贬义词语用法词典》[11]、《汉语形容词用法词典》[12]、《中华成语大词典》[13]、《汉语熟语词典》[14]、《新世纪汉语新词词典》[15]。语义知识网络有《知网》和WordNet。另外还加入了《汉语情感系统中情感划分的研究》中的部分词汇。

3.2.2 获取的方法

以大量情感语料为基础,采用手工情感分类和自动获取强度两种方法,从资源中获取情感信息。

情感词汇通过两级筛选得到。一级筛选是从各类资源中初步挑选可能与情感相关的词汇。各类资

源的一级筛选方法如下:词典资源主要是利用词典中与情感相关的子类,如心理、感觉、情感、性格、态度等类的词汇,但是也有一些词典没有相应的子类划分,需要整本都进行人工过滤。《知网》中情感词汇的获取是先选择包含情感色彩的义原,主要有'情绪'、'态度'等几大类,然后从《知网》中选取包含这些义原的词汇。WordNet 是利用 WordNet-Domain 抽取初级的情感词汇。WordNet-Domain 划分 WordNet 中的 Synet,选择与 psychology 类型相关的 Synet,然后根据它们与词汇的对应关系选择相应的词汇。

二级筛选是采用手工分类的方法,人工从一级筛选的词汇中选择有情感色彩的词汇,并划分情感类别,即指出词汇包含哪几种情感,并分别为每种情感类别的每个强度等级确定一定量的标准词汇。

自动获取强度主要是指在上述手工分类的基础上,自动获取词汇的情感强度。具体思想是在大规模的语料库中查找待定词汇和标准词汇的互信息,从而将待定词汇的强度确定为与之互信息最高的标准词汇的强度。这样情感强度的计算首先需要两类资源,一个是标准词,另一个是大规模的语料库。

标准词就是为 20 类情感的 5 个情感等级分别确定的一定量的标准词汇。通过计算词汇与每个等级的标准词汇在语料中的互信息(即共现概率)来初步确定情感强度,然后再对不合理的进行人工调整。本文采用的是点互信息(pointwise mutual information),它的计算公式如下:

$$I(W_u, S_{ui}) = \log_2 \frac{P(W_u S_{ui})}{P(W_u)P(S_{ui})} \qquad (1)$$

其中,W_u 表示包含 u 类情感的词汇,S_{ui} 表示 u 类情感的第 i 个标准词。计算 W 与所有 u 类情感的标准词汇之间的互信息,选择互信息最大的那个标准词的强度作为词汇 W 在 u 类情感上的强度。

语料库是计算点互信息的基础,如果语料数量较少或涉及词汇的范围较窄,那么在计算点互信息时会出现数据稀疏的问题,不能正确反映词汇的强度。我们的语料是从网上下载的与情感相关的文章,从风格上看包括寓言、散文、戏剧、小说、杂文、新闻等多种文体;从时间和空间上涉及国内和国外不同时代的多个作家。去噪后的语料大约 150M,涉词范围较广,保证了互信息计算的有效性。图 1 统计了自动获取强度的准确率(与人工校正过的对比),可见自动获取的强度与实际强度完全相等的不是很多,但是与人工校正的强度相差一个等级的却比较多,其中'妒忌'类强度符合程度最高,达到 84%。所以自动获取的强度有较大的参考价值,为人工核查提供了依据。

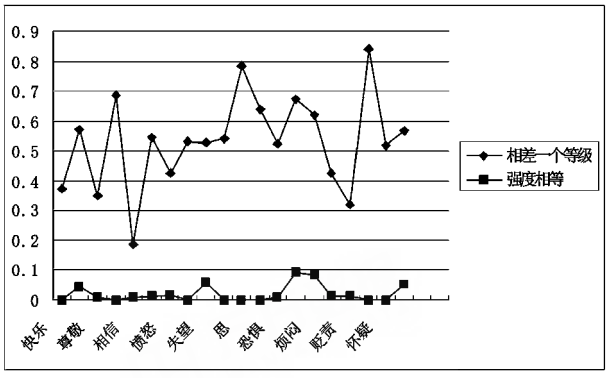

图 1 自动获取强度的准确率

3.3 词汇本体的质量保证措施

质量是词汇本体的生命,是应用和更新的基础。但是由于汉语词汇的信息量庞大和人类情感的复杂多变,质量的保证更成为本体建设过程中的一个关键问题。为此,我们设计了一套规范化的操作方法,严格控制词汇情感信息的更新,并对词条信息采用多重的人工检查流程。为了减少建设本体过程中的误操作,设计了一个方便快捷的录入界面,如图 2,以保证词条语义属性的正确性与一致性。

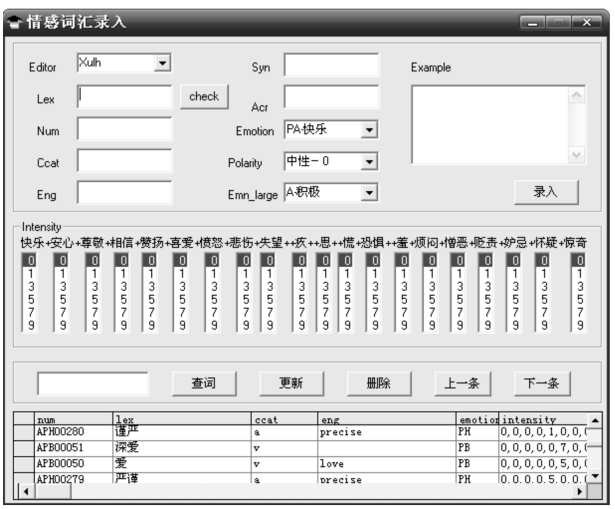

图 2 情感本体建设录入界面

汉语词汇中不断有新词出现,还有一些词汇产生了新的释义,所以情感词汇的本体也在不断的更新和维护中。通过不同时期版本的比较可以输出词汇信息的更新和修订情况。

4　统计数据

情感词汇本体第 1 期收录词汇 10 259 条，各类情感包含的词汇数量及使用频率等信息如表 2 所示。

表 2　各类情感词汇的数量

情感类	词汇数	情感类	词汇数	情感类	词汇数	情感类	词汇数
快乐	609	喜爱	562	惊奇	47	尊敬	327
安心	151	思	83	相信	92	赞扬	3 046
愤怒	187	悲伤	362	恐惧	182	憎恶	845
烦闷	456	羞	59	疚	50	慌	144
失望	74	妒忌	29	怀疑	37	贬责	2 917

由表 2 可以看出，妒忌、怀疑、相信和惊奇类的词汇数量较少，在 50 个以下，而贬责和赞扬类的词汇数量较多。这表明很多词汇并不具有快乐、惊奇、愤怒、悲伤等强烈的情感，而对某事物表达一种褒或贬的倾向词汇比较多。

各类情感词汇的强度分布如图 3 所示。

由图 3 可以看出，强度等级为 5 和 3 的词汇数量在各个情感类别中都普遍高于其他强度等级的词汇。其他强度等级的词汇在数量上分布比较平均。总体来说，贬责和赞扬类的词汇数量较多，思、疚、慌等类的词汇数量较少。

情感词汇的在语料库中的平均使用频率如图 4 所示。“妒忌”类的平均使用频率最低，为 56.17 次，而“惊奇”类词汇的使用频率最高，达到 568.96 次。所有词汇的总平均使用频率为 143.37 次，一定程度上保证了词汇间互信息的获取，从而保证了自动获取强度的可行性和有效性。

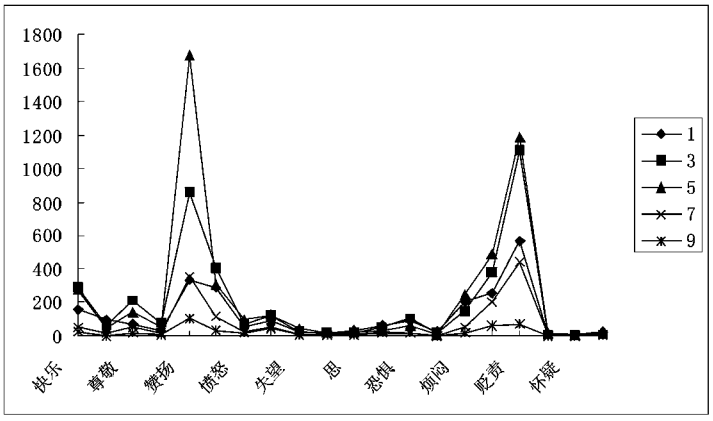

图 3　情感词汇的强度分布

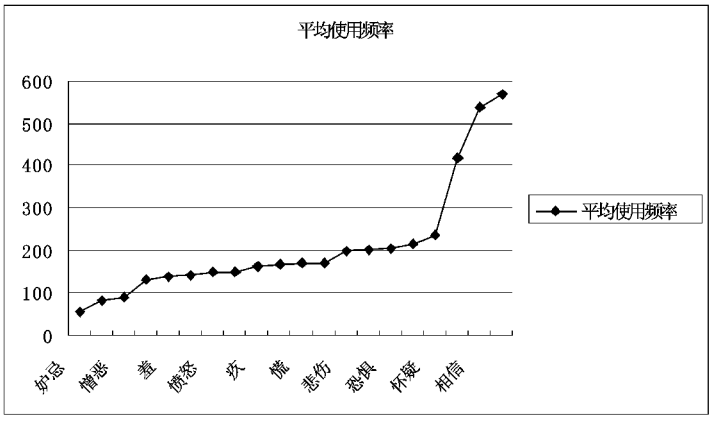

图 4　各类情感词汇在语料中的平均使用频率

5 进一步的工作

目前词汇本体的建设工作还在继续进行中，我们计划加入更多的语义资源来丰富词汇本体，在综合多种语义资源后第2期的词汇总量预计将达到3万词以上。同时以大规模的语料库建设为基础，统计本身没有情感倾向的词汇，通过大规模的实际场景的训练得到一个词汇以多大的概率出现在哪方面的情感中[1]。这种扩展的情感词汇将极大地丰富我们的词汇本体。通过带标语料库的建设，还可以进一步验证和修订目前词汇的情感分类。

本文所介绍的情感词汇本体是篇章情感分析的基础，可以应用在多种情感识别系统中。由于人类对情感认识的局限性和汉语词汇的复杂多变性，情感词汇本体的建设是一个长期而繁杂的工程。词汇本体还有许多需要完善地方，如增强情感词汇复杂度的描述，录入大量例句等。今后还应根据实际的需要增加情感词汇的数量，不断修正词汇的描述信息，使情感词汇的描述体系更加完善。同时还应从大规模的标注语料中抽取更多情感信息，校验现有的情感分类。

参 考 文 献

[1] Hugo Liu，Henry Lieberman，Ted Selker. A model of textual affect sensing using real-world knowledge [C] // Proceedings of the 8th International Conference on Intelligent User Interfaces. 2003：125-132.

[2] Hugo Liu，Ted Selker，Henry Lieberman. Visualizing the affective structure of a text document [C] // Proceedings of Conference on Human Factors in Computing Systems. 2003：740-741.

[3] Hua Wang，Helmut Prendinger，Takeo Igarashi. Communicating emotions in online chat using physiological sensors and animated text [C] // Proceedings of Conference on Human Factors in Computing Systems. 2004：1171-1174.

[4] Tsou Benjamin K Y，Kwong O Y，Wong W L. Sentiment and content analysis of Chinese news coverage [J]. International Journal of Computer Processing of Oriental Languages，2005，18(2)：171-183.

[5] Ekman P. Facial expression and emotion [J]. Americam Psychologist，1993，48：384-392.

[6] Yu Zhang，zhuoming Li，Fuji Ren，Shingo Kuroiwa. Semi-automatic emotion recognition from textual input based on the constructed emotion thesaurus [C] // Proceedings of 2005 IEEE International Conference on Natural Language Processing and Knowledge Engineering（IEEE NLP-KE′05）. 2005：571-576.

[7] 林传鼎. 社会主义心理学中的情绪问题 [J]. 社会心理科学，2006，21(83)：37-62.

[8] 许小颖，陶建华. 汉语情感系统中情感划分的研究 [C] // 第一届中国情感计算及智能交互学术会议论文集. 2003：199-205.

[9] Ekman P. An argument for basic emotions [J]. Cognition and Emotion，1992，6：169-200.

[10] 董大年. 现代汉语分类词典 [M]. 上海：汉语大词典出版社，1998.

[11] 王国璋. 汉语褒贬义词语用法词典 [M]. 北京：华语教学出版社，2001.

[12] 郑怀德，孟庆海. 汉语形容词用法词典 [M]. 北京：商务印书馆，2004.

[13] 程志强. 中华成语大词典 [M]. 北京：中国大百科全书出版社，2003.

[14] 杨兴发. 汉语熟语词典 [M]. 成都：四川辞书出版社，2005.

[15] 王均熙. 新世纪汉语新词词典 [M]. 上海：汉语大词典出版社，2006.

（责任编辑 许增棋）