网络用语词典的构建及问题分析*

昝红英[1]，许鸿飞[1]，张坤丽[1]，穗志方[2]

（1.郑州大学信息工程学院，河南省 郑州市 450001；2. 北京大学计算语言学研究所，北京市 100871）

摘要：随着互联网应用的快速发展，网络用语的使用越来越普遍，网络新词层出不穷。网络文本中大量的网络用语，对基于自然语言处理的情感分析、产品推荐、自助问答系统等应用带来了一定的挑战，而收集并构建网络用语词典及相关语料则是解决此类问题的突破点。本文以微博语料为出发点，综合多类网络资源，收集并整理了较为全面的网络用语词典及相关语料。同时，对网络用语词典构建中遇到的问题进行了分析和总结，并对其潜在应用进行了初步的探讨。

关键词：网络用语；词典构建；标注

中图分类号：TP391　　**文献标识码：**A

The Construction of Internet Slang Dictionary and its Analysis

Hongying Zan[1], Hongfei Xu[1], Kunli Zhang[1], Zhifang Sui[2]

(1.School of Information Engineering, Zhengzhou University, Zhengzhou, Henan 450001, China; 2. Institute of Computational Linguistics, Peking University, Beijing, 100871, China)

Abstract: With the rapid development of the Internet, the use of internet slang prevails. New words of this kind keep appearing, which has become a great challenge for natural language processing tasks like sentiment analysis, product recommendation, QA, etc. This problem can be effectively alleviated through the construction of the internet slang dictionary. This paper analyzes the problems we have encountered when collecting and annotating micro-blog texts, together with other internet resources, to build the dictionary and the related corpus. Further, the potential applications of this dictionary and the corpus will be discussed.

Key words: Internet Slang; Construction of Dictionary; Annotation

1 引言

网络用语是互联网上信息传播和交流的一种语言[1]。伴随着互联网的快速发展，网民产生了大量的数据，其中文本数据占据了很大的比重，微博、网络媒体、贴吧、博客、论坛等应用产生的数据尤其惊人。自然语言处理技术在情感分析、产品推荐、自助问答系统等领域得到了广泛应用，针对网络文本的特征，对应用在网络文本数据中的自然语言处理技术也有相应的调整。网络用语有很多与传统用法不同的地方[2]，但就目前的认知，还没有人对网络用语进行系统地收录、总结、整理和标注。词汇语料库作为文本分析及语义理解的数据基础，是信息检索中的查询扩展、机器翻译中的模块识别等方面不可或缺的资源，在句法分析、词义消歧等信息处理任务中也发挥着重要的作用[3]。

与正式文体相比，网络用语有自主化、个性化、全息化、符号化、时尚化、创新化等特性，这些特性使网络用语这种传递信息的表达方式更形象，使用也越来越广泛，逐渐从网络进入了人们的日常生活。有理论研究认为网络用语具有临时性，表现为快速出现，快速消亡，但从目前的收集工作来看，网络用语出现后有两种走向，一种往往有其鲜明的时代、社会或者市场背景，随着时间的发展，因为不符合现有情境等原因而逐渐消亡，另一种则具有普适性，随着时间的推移不但被保留下来，还被人们频繁地使用，甚至被收录到权威语言资源，进入正式文体。例如，2010年11月10日，网络用语“给力”一词登上了《人民日报》头版头条，网友惊呼“太给力了”。

虽然针对网络文本的自然语言处理技术及其相关的研究工作越来越多，但是目前仍然缺少相关的基础资源。因此，本文从词语开始，从百科、贴吧、微博、论坛等媒介收集标注网络用语，

* **收稿日期：**　　**定稿日期：**

基金项目：国家重点基础研究发展计划 973 课题（2014CB340504）；国家自然科学基金项目（61402419）；国家社会科学基金项目（14BYY096）；河南省科技厅基础研究项目（142300410231，142300410308）；河南省教育厅科学技术研究重点项目（13B520381，15A520098）

收集、整理网络词库。本文主要介绍网络用语词典构建的过程，总结网络用语标注过程中遇到的典型问题，并提出了相应的解决方案。

2 相关工作

语料库就是存放语言材料的语言数据库，语料库语言学是基于语料库进行的语言学研究。语料库在自然语言处理中占据了重要的地位，基于规则的任务需要语料库的强力支持，基于统计的计算语言学也需要从语料库中学习信息。词语是语言的基础，目前为止已经出现了大量的词典、词汇知识库或围绕词语展开的语料库，例如：布朗语料库、LLC 口语语料库、朗文语料库、现代汉语语料库等，这些语料库的构建对其它很多语言资源的构建以及相关的自然语言处理任务提供了有力的数据支持和经验参考。

北京大学俞士汶教授等构建的现代汉语语法信息词典[4]，是中文信息处理的第一块基石，在此基础上形成的现代汉语综合型语言知识库，为后续的相关工作提供了很好的中文分词和词性标注标准，也为中文分词、命名实体识别等自然语言处理任务的展开提供了有力支持。Chinese LDC 包含汉语通用词表和汉语信息词典，同时也包含分词词性标注语料。国家语委汉语语料库不仅有现代汉语语料库，同时还包含有古代汉语语料库。这些语料库对面向中文的自然语言处理任务提供了有力的支持，但是这些经典的语言资源对于随着互联网的快速发展而出现的网络用语，则缺少相应的标注和整理。

随着网络文本规模的迅速增长，目前已经有一些针对网络用语的分析和研究。张曼对网络用语的类型和特征进行了研究[5]，认为网络用语大致有词汇缩略型、图形符号型、数字谐音型、故作童言型、文符并用型、英语汉说型、旧词新说型七类，对网络用语的分类提供了有效的参考和帮助。张曼从微博产生的相关衍生词、微博使用频度较高的新词新语及同事件紧密关联的微博流行热词等三方面对微博的新词新语进行了研究阐述[6]。侯敏从语言学、社会学、传播学三个角度对 2010 年产生的 543 个新词语进行解读[7]。韩忠明等人对中文微博短文本倾向性分类算法的研究工作[8]表明，网络用语对面向网络文本的自然语言处理任务有重要影响。但是到目前为止，在自然语言处理应用的研究领域和工业界还缺少系统的网络用语词典或相关资源。如果能对常用的网络用语进行收集、整理、标注，将会为面向网络文本的信息抽取[9]、情感分析[10]、词义消歧[11]等处理系统提供有力的数据支持。

目前还有一些学者针对具体的应用任务，构建相应的语料资源，如李钰面向微博情感分析应用构建了微博情感词典，对基于词典和规则的情感分析方法进行了研究，其最显著的优点是这些规则方法即使在大型的文本语料库上也非常有效[12]，并且情感词典本身结合语料的统计信息作为特征对基于统计的方法也很有帮助；王文远等人构建了面向情感分析的微博表情词典并尝试应用，将表情情感词典反作用于对应的微博文本，重新度量其中情感词的倾向值[13]，也获得了较好的结果；陈晓东对当前已有的情感词汇资源加以总结和整理，并运用扩展的情感倾向点互信息算法对新浪微博语料进行实验，自动获得领域情感词[14]，构建了一个面向中文微博的情感词典并将其用于情感分析中。但是这些语料库都是针对具体的任务构建，具有较强的针对性，不具备普适性。鉴于此，本文旨在构建通用的网络用语词典，为面向网络文本的自然语言处理系统提供一定的数据基础。

3 网络用语的来源和特点

借鉴已有的研究成果，并通过对网络文本语料的抽取整理，本文认为网络用语主要从这几个来源产生：

1）谐音，可能是因为方言的不确定输入，也可能是因为用户觉得好玩，例如：用“虾米”代替“什么”；

2）缩略，典型代表比如：“不明觉厉”（不明白为什么，但是觉得很厉害），从语言上来讲是不规范的，但确实在互联网时代出现并且走进了人们的日常生活；

3）象形，使用文字或者符号组合成有趣的图像，来传递用户的感情（例如：“orz”看起来像一个人跪在那里，表示自己也是无奈了），也可能是对已有汉字字形的再解读，比如：“囧”；

4）转义，有些词语原本就有一定的意义，往往根据已有的词意进行扩展或延伸，比如说“马甲”本来是人们日常穿的一种衣物，网络上却往往用来表示一个人的另

一个 ID，一个新的 ID 相对于用户来说就像人又穿了一件马甲，再次包装了一下一样；

5）新词，网络中有些词语，原来并不存在，往往伴随网络中的新生应用产生，比如：“点赞”。

网络用语随着互联网的发展一起快速成长，快速、丰富多彩是互联网的特色，简单、有趣是网络用语的特色，因此得到用户的喜爱。特别地，网络用语往往随互联网或生活中的某个热点事件一起出现，在网上短期内迅速传播，甚至进入用户日常生活的交流中，并逐步融入人们的正式文体中，成为这个时代文明的特色之一。

4 网络用语词典构建标准

考虑到可能面对的应用场景，本文对网络用语从 6 个属性进行描述，分别是：类别、是否低俗、情感极性、释义、例句和词性。网络用语的某个属性可能具有多个标注，就像一词多义，在我们的标注工作中也标注了多个值。

4.1 类别

目前已有的研究工作中对网络用语的分类并不统一，借鉴汉语的语法和其它相关语言资源构建中制定的分类标准，根据网络用语的来源，本文把网络用语分为：谐音[15]、缩略[16]、象形、转义和新词五大类，其中谐音、缩略和象形又根据细节的不同分为若干子类别，具体类别如表 1 所示：

表 1 网络用语分类表

类别	子类别	解释	例子
谐音(1)	数字(1)	用数字来代替原来的词语，谐数字的读音	88（拜拜）、1314（一生一世）
	字符(2)	用英文字母或标点符号来代替原来的词语，谐符号的读音	==（等等）、3Q（thank you，谢谢你）
	汉语(3)	对汉语本身的谐音	不造（不知道）、筒子（同志）
	外语(4)	中文词语谐音英语词语	候住（Hold 住，控制住）、闹太套（not at all）
	方言(5)	谐方言的读音	大条（桂林话，指一个人目中无人，摆架子，为贬义词）、好康（好看，来自闽南方言）
	连读(6)	把词语的衔接处连起来读，可以谐音另一个字	屈原（泉）
	拆音(7)	把词的读音拆开正好可以谐音另外一句话	酱紫（这样子）、我宣你（我喜欢你，是台湾腔演变夸张了的）
缩略(2)	拼音(1)	词语的拼音取首字母组成	RMB（人民币）、RP（人品）
	汉字(2)	从一句话中抽取某些字	何弃疗（为什么放弃治疗）、累觉不爱（很累，感觉自己不会再爱了）
	外语(3)	英语词组的首字母组合，或英语中原有的缩略表示	BF（boy friend 男朋友(也可译为：best friend 最佳朋友)）、eg（举例）
象形(3)	字符(1)	利用字符拼接出的图案来传达意图	;)、(^_^)
	造字(2)	字本身像图案，传达意图	囧、亹
	拆字(3)	把一个字拆成两个或多个字来写，常用的反屏蔽、反过滤手段	弓虽（强）、女子（好）
转义(4)		原有词语，现表达另一种新的词义	杯具（悲剧之意）、盖楼（在同一个主题帖下回复）
新词(5)		原来并不存在，后来被创造出来使用的词语[17]	水帖（没有用处的帖子）
其他(0)		难以确定类别归入此类	新蚊连啵

其中括号内为标注中使用的编号，类别与子类别以“-”分隔，例如“象形”类别下的“字符”子类，标注编码为“3-1”。

4.2 是否低俗

虽然现在已经有很多净化网络环境的相关工作，但是网络上还是包含一些低俗的信息。网络已经广泛地渗透到社会生活的方方面面，成为人们工作、学习、生活密不可分的重要组成部分[18]。网络用语中不可避免的也包含了一些低俗的词语。考虑到网络净化、舆情分析等应用需求，本文对所有词语标注了这个属性，如果低俗标为1，反之标为0。

4.3 情感极性

网络文本中的情感十分丰富，网络用语中很大一部分有明显的情感倾向。情感分析是对主观性文本进行挖掘与分析，获取有用的知识和信息[19]。本文对收录到的网络用语标注了情感极性，共有正面、负面和中性三种标记，这些属性将有助于网络文本的情感分析或舆情监控等应用。

4.4 释义和例句

一般人见到的网络用语往往是有限的，面对一个给出的词语弄不清语义和用法是很正常的事情，所以对于收集到的网络用语，我们都标注了释义和例句，这样可以为查询网络用语的用法和意义提供便利和帮助，也可以用来辅助训练网络用语的词向量。

4.5 词性

对收集到的网络用语都标注了词性，词性标注标准以北京大学计算所词性标注标准[20]为基础，针对网络用语的特殊情况，对其进行了修改，删除了除“nr”、“ns”、“nt”、“nz”之外的所有非一级词性标记，并增加了两个新的标记。因为其中有些网络用语看起来像是词语，实际上是一句话，这些用语的类别多为缩略，也无法确定具体的词性，所以在语料库中增加了“dj”标记，取“短句”一词的首字母，表示这个用语用来表示一句话的意思，对难以确定词性的词语，则标记为“0”。

5 网络用语词典构建问题分析

网络用语词典的构建工作主要由网络用语的收集整理、标注属性及标注值的确定和校对三部分工作组成，每部分工作都有其特有的问题。

5.1 有效地采集网络用语及语料来源

收集数据时尽可能使用已发布的数据，加以规范化整理和筛选。本文网络用语的收集主要以微博为主，同时兼顾其他来源，初期收集的过程中也包含贴吧、论坛、聊天软件等应用中产生的网络用语。具体收集过程如下：

1) 通过搜索工具收集网上已发布的网络用语集合，对其进行去重后再由人工筛选核对；
2) 对微博语料进行分词处理，去除常规已收录词语，对去重的结果进行人工筛选抽取网络用语；
3) 从热点事件（例如：庆安枪击事件）相关的微博语料中抽取网络用语。

5.2 网络用语的判定问题

网络用语的界定并没有规范统一的标准，本文以应用为目的，认为伴随互联网出现的并且被网民大量使用并广为接受的词语或短语作为网络用语收入网络用语词典。这里，大量使用和广为接受的界限是难以确定的，实际操作中通过将收集到的、人工筛选过的网络用语送入搜索引擎，查看搜索引擎反馈的结果，如果确实大量使用并且用法、词义统一，就认为该用语已经具有特定的、被人们认可的意义，并收录到网络用语词典中。

5.3 兼类问题

有些网络用语可能同时属于多个类别（例如“白骨精”，同时属于缩略（白领、骨干、精英）和转义（区别于《西游记》中的白骨精））。对于这种情况，同时对其标注多个值，值与值之间使用“|”分割。

5.4 是否低俗的判定问题

对于什么是低俗没有严格的分界，只能通过人的感觉去判断，再加上网络用语往往比较随意，所以有些处于判定边界附近的网络用语，还有一些被用来开玩笑的词语，很难去界定是否低俗，比如“OUT”这一用语，可以说人过时，也会被朋友之间用来打趣，最终我们在讨论过程中进行举例论证，得到合理有据的一致意见，作为最终的标注。

5.5 歧义带来的多个属性值的问题

一个网络用语可能有多个语义和用法，且不同语义和用法间的属性值不同。例如“不明觉厉”一词，表示“不明白，但是感觉很厉害”，可以用来表示对于对方实力的认可和膜拜，

也可以用来调侃对方，对于这种情况，根据语义或用法（例句）词典中将对此类网络用语收录多个词条，分别标注。

5.6 词性确定的问题

大部分网络用语的词性可以根据语义和用法（例句）确定，个别存在不同语境或语义下有不同用法的情况，将其分为多个词条标注；

对于短语或句子形式的网络用语，例如：“喜大普奔”、“细思恐极”、“累觉不爱”等，直接标记为“dj”，还有一些词语难以确定词性则标记为0，表示无法确定。

5.7 校对中存在争议的分歧处理和客观性保证

标注时对每一个网络用语都采用双人标注，标注完成之后进行合并校对，对合并校对过程中产生的分歧进行集体讨论，得到一致意见之后再修改标注数据。网络用语的释义和例句尽可能地从百科、问答或论坛中抽取，再进行人工审核，选出网络上公认的解释和例句，如果确认有多个释义和例句，则分为多个条目处理，尽可能地保证标注的客观性及全面性。

6 结果分析及应用领域

按照以上规范和流程，历经 6 个月，目前共收集、标注、校对网络用语及属性 6760 个，网络用语语词典样例如表 2 所示：

表 2 网络用语词典样例

词语	类别	是否低俗	情感极性	释义	例句	词性
白骨精	2-2|4	0	1	白领+骨干+精英。	立做白骨精	n
包子	4	0	-1	某人长得难看或者笨就说他包子	包子脸是可爱的。	n
爆粗	4	0	-1	讲粗话 讲脏话	受不了了，我要爆粗	v
爆棚	4	0	1	形容客满、满座。人气旺	现在餐馆已经爆棚。	a
杯具	1-3|5	1	-1	悲剧。	杯具啦！	n
不明觉厉	2-2	0	1	又称“虽不明，但觉厉”。指“虽然不明白（对方）在说什么，但好像很厉害的样子。”用于表达菜鸟对高手的崇拜。	不明觉厉，大神就是大神。	dj
不明觉厉	2-2	0	-1	又称“虽不明，但觉厉”。指“虽然不明白（对方）在说什么，但好像很厉害的样子。”用于调侃楼主语言行为夸张和不知所云。	那个保健专家让人很是不明觉厉。	dj
PF	2-1	0	0	佩服。	佩服。	v
233	1-1	0	1	大笑。	这个笑话忍不住让人233！	v
吊炸天	5	1	0	形容非常厉害的样子。	这款游戏简直吊炸天了！	a
猪你快乐	1-3	1	1	祝你快乐。	今天生日吗？猪你快乐！	dj

在构建网络用语词典之后，对收集到的数据中的“是否低俗”、“情感极性”和“类别”三个属性进行了统计。网络用语的“是否低俗”属性可以对舆情监控起到帮助，对情感分析可能也是有一定帮助的，对它的统计信息显示：低俗网络用语所占比例约为16.59%。可以看出，网络用语中确实包含一些低俗的信息，但是占据的比例相对来说还是较少的。网络用语的“情感极性”属性可以用于情感分析，对它的统计信息显示：中性情感的网络用语所占比例约为55.57%，负面情感的网络用语所占比例约为 30.33%，正面情感的网络用语所占比例约为 14.10%。可以看出，大部分网络用语并没有分明的情感极性，表达负面情感的网络用语较表达正面情感的网络用语多，这或许也可以从一个侧面反映出：网络上人们倾向于自由地表达自己的情感，发泄在生活中遇到的不如意。

词典中网络用语类别的各个属性及子属性进行统计得到的比例如图 1-4 所示：

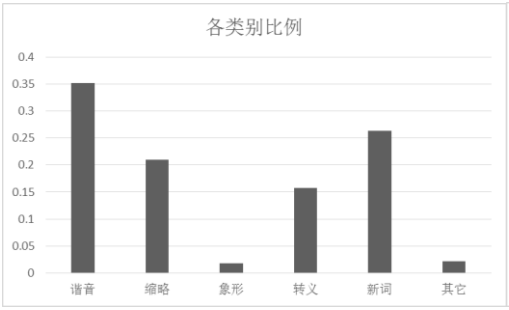

图 1 各类别比例

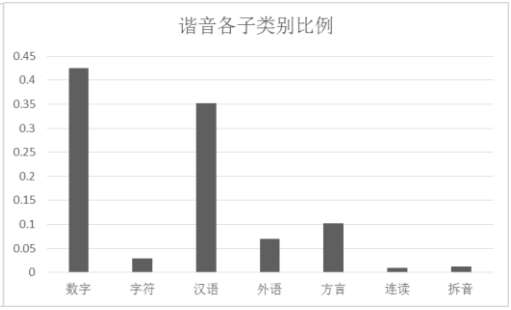

图 2 谐音子类别比例

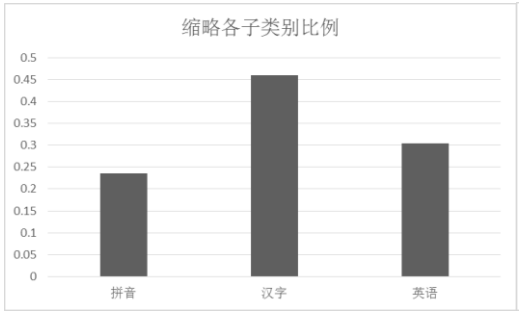

图 3 缩略子类别比例

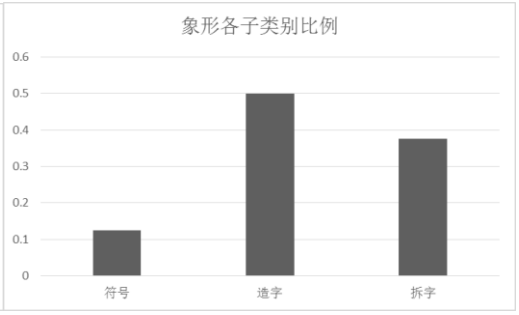

图 4 象形子类别比例

从图 1-4 中可以看出：网络用语中谐音和新词作为最主要的两部分占据了最大的比例，而象形可能因为网络中可以随意使用的大量表情和图像等各种原因，占据的比例最少，且远小于平均值；由谐音产生的网络用语中，数字和汉语的比例最高，前者应该是因为容易输入，而后者应该是因为中国的网民更熟悉；缩略类别中，汉语也占了很大的比例，和谐音很相似，但是各个比例的相差不大；象形的各个子类中，符号占据的比例远小于另外两类，这应该与我们去除了在主流输入法中涵盖的、可直接输入的表情有关。

在规范的指导下，同一个网络用语的标注工作由两人共同完成，标注的一致性用 Kappa 统计量（K）来度量[21]。双标过程中的 Kappa 值如表 3 所示：

表 3 双标过程的 Kappa 值

类别	是否低俗	情感极性	词性
0.947	0.934	0.916	0.967

对 kappa 值相对较低的情感极性属性进行分析，可以发现全部是由对中性与正面或负面的争议造成，但总体来说，一致性还是较高的。对于双标过程中不一致的部分，我们通过集体讨论来确定最终的标记。

将收集到的网络用语应用于分词系统，对处理网络文本分析将会有帮助。用未导入用户词典的分词工具对提取网络用语所使用的微博语料进行分词[22]，然后除去已有词典[23-24]中收录的词语，对比我们抽取到的网络用语，在 3 次每次约 10,000 字的测试中，只有 1 次正确地提取出了 1 个我们手工整理出的网络用语；网络用语的类别属性可以作为特征，加入到包括情感计算在内的各种网络文本的自然语言处理应用中；是否低俗属性对于网络监管和倾向分析应用很有帮助；情感极性属性对情感分析[25]有很大帮助，尤其是基于规则的方法；因为网络用语的实际用例往往在语料中出现频率很低，释义和例句对于训练网络文本的词向量有很大帮助；词性可随网络用语本身一起用来辅助分词，也可以作为特征加入到机器学习应用中；为采用机器学习方法进行自动化预构建提供帮助，也可以为其他语言资源[26-27]的建设或扩充提供支持。

7 总结与展望

网络用语对网络文本的理解及应用极其重要，本文对已有的词典资源进行了分析，根据获取到的网络资源，收集并整理了流行网络用语，构建了一定规模的网络用语词典，并对构建过程中的问题和相应的解决办法进行了总结和分析。利用已构建的网络用语词典，在微博分词和情感分析的应用进行了初探，表明网络用语词典是网络文本的分析及应用的重要资源。下一步，将尝试在包含网络文本的自然语言处理应用中加入目前已收集的网络用语；同时，以现有数据为基础，

尝试利用半监督的开放信息抽取等技术进行有关网络用语的自动收录和标注工作，减少人工收录、标注、整理的工作量，有效扩充网络用语词典的规模。

参考文献

[1] 黄晓斌, 余双双等. 网络用语对信息交流的影响[J]. 情报理论与实践, 2008, 31(1):23-25.

[2] Y Ding, F Ren. Constructing Chinese Internet Terminology Corpus[J].研究报告自然语言处理, 2009,193(4): 1-7.

[3] 石金铭, 昝红英, 韩英杰. 大规模汉语词汇语义知识库的构建[J]. 山西大学学报（自然科学版）, 2015, 38(4): 553-559.

[4] 俞士汶, 穗志方, 朱学锋. 综合型语言知识库及其前景. 中文信息学报, 2011, 25(6):12-20.

[5] 张曼. 微博新词新语探析[J]. 学理论, 2011(23):163-164.

[6] 侯敏. 2010 年度新词语解读[J]. 语言文字应用, 2011(4):64-70.

[7] 韩忠明, 张玉沙, 张慧,等. 有效的中文微博短文本倾向性分类算法[J]. 计算机应用与软件, 2012(10):89-93.

[8] 林丽. 试析框架语义标注在新闻事件抽取中的应用—以越南语军事新闻为例[J]. 山西大学学报（自然科学版）, 2013(4):510-516.

[9] 史伟, 王洪伟, 何绍义. 基于语义的中文在线评论情感分析[J]. 情报学报, 2013, 32(8):860-867.

[10] 宗成庆. 统计自然语言处理[M]. 2. 北京:清华大学出版社, 2013:297-398.

[11] 林纲. 网络用语的类型及其特征[J]. 当代修辞学, 2002(1):26-27.

[12] 李钰. 微博情感词典的构建及其在微博情感分析中的应用研究[D]. 郑州大学, 2014.

[13] 王文远, 王大玲, 冯时,等. 一种面向情感分析的微博表情情感词典构建及应用[J]. 计算机与数字工程, 2012, 40(11):6-9.

[14] 陈晓东. 基于情感词典的中文微博情感倾向分析研究[D]. 华中科技大学, 2012.

[15] 成晓杰. 谈网络语言的谐音表义[J]. 修辞学习, 2002, 3: 21-21.

[16] 殷志平. 构造缩略语的方法和原则[J]. 语言教学与研究, 1999(2): 73-82.

[17] 邹纲, 刘洋, 刘群, 孟遥, 于浩, 西野文人, 亢世勇. 面向 Internet 的中文新词语检测[J]. 中文信息学报. 2004(6): 1-9.

[18] 陈静. 网络低俗内容的监管难度与对策[J]. 网络传播. 2008(10): 58-59.

[19] 魏韡, 向阳, 陈千. 中文文本情感分析综述[J]. 计算机应用. 2011,31(12): 3321-3323.

[20] 俞士汶, 段慧明, 朱学锋, 孙斌. 北京大学现代汉语语料库基本加工规范[J]. 中文信息学报, 2002,16(6): 58-65.

[21] 贾玉祥, 黄德智, 刘武,等. 中文语音合成中的文本正则化研究[J]. 中文信息学报, 2008, 22(5):45-50.

[22] 黄昌宁, 赵海. 中文分词十年回顾[J]. 中文信息学报, 2007, 21(3): 8-19.

[23] 俞士汶, 朱学锋, 王慧, 等. 现代汉语语法信息词典详解：第 2 版[M]. 北京：清华大学出版社, 2003.

[24] 王惠, 刘群. 《现代汉语语义词典》的概要及设计[C]. 1998 中文信息处理国际会议论文集，北京：清华大学出版社, 1998: 361-367.

[25] 谢丽星, 周明, 孙茂松. 基于层次结构的多策略中文微博情感分析和特征抽取[J]. 中文信息学报. 2012, 26(1): 73-83.

[26] 张坤丽, 昝红英, 柴玉梅, 等. 现代汉语虚词用法知识库建设综述[J]. 中文信息学报, 2015,3: 1-8+15.

[27] Kunli Zhang, Hongying Zan, Yingjie Han,et.al. Preliminary Study on the Construction of Bilingual Phrase Structure Treebank.In Proceedings of CLSW2014, LNAI8922，pp403-413.

作者简介：昝红英（1966——），女，博士，教授，主要研究领域为语言资源构建、词汇语义学、信息检索、情感计算、自然语言处理等。 Email: iehyzan@zzu.edu.cn；许鸿飞（1994——），男，硕士生，主要研究领域为自然语言处理。Email: hfxunlp@foxmail.com（通讯作者）；张坤丽（1977—），女，讲师，博士生，主要研究领域为自然语言处理、语言资源构建等。 Email: ieklzhang@zzu.edu.cn 。