文章编号：1003-0077（2011）00-0000-00

基于图排序的词汇情感消歧研究*

杨亮，张绍武，林鸿飞，宋艳雪

（大连理工大学，辽宁省大连市，116024）

摘要：词汇情感消歧是文本情感倾向性分析的关键技术之一。本文在分析比较了词汇情感消歧和词义消歧异同后，从情感分析角度出发，提出了基于图排序的词汇情感消歧方法。该方法通过自动获取和人工校正相结合的方式获得多情感词汇，然后根据语义关系构建词义关系图，进而在词义关系图上迭代计算直至收敛，最后选择多情感词汇的词义中权值最大的词义作为结果输出，从而实现情感消歧。本文分别在新浪微博语料库和情感语料库上验证了该方法的有效性。

关键词：多情感词汇；图排序；情感消歧

中图分类号：TP391　　**文献标识码：**A

Word Emotion Disambiguation Based on Graph Ranking

Liang Yang, Shaowu Zhang, Hongfei Lin, Yanxue Song

(Dalian University of Technology, Dalian, Liaoning 116024, China)

Abstract: Word emotion disambiguation is vital to sentiment analysis, so we analyzed the differences between word emotion disambiguation and word disambiguation, then selected the multi-emotional word automatically and artificially. From the aspect of sentiment analysis, we promoted the method named word emotion disambiguation based on graph ranking which builds directed meaning graphs according to semantic relations, and then iteratively computed on the graphs, selected the largest iterative value meaning among others of multi-affect word as the right output. We compared our method on MicroBlog corpus and emotional corpus with two others that one was based on part of speech and emotional frequencies, the other was based on Bayesian model, and the results proved that our method was effective than the two others.

Key words: Multi-Affect Words; Graph Ranking; Word Emotion Disambiguation

1 引言

文本情感倾向性分析逐渐成为一个研究热点[1][2]，词语级倾向性分析是文本情感分析的基础。但是，同一个词语在不同的语境下可能表达出不同的情感倾向性。例如下面两个句子：

(1) 这种幼稚的做法最终会让你后悔莫及。

(2) 我那幼稚的弟弟今年才两岁就已经能数到一百了。

在《现代汉语词典》中，“幼稚”有 2 个词义：(1)年纪小；(2)形容头脑简单或缺乏经验。生活中，词义(1)经常被用来形容小朋友在思想上的天真无邪，纯真可爱；词义(2)则常常会被人们用来形容成人思想不成熟，眼界狭隘，目光短浅，看问题难以洞悉实质。由上述例句可以看出，在不同的语境中，“幼稚”表达了不同的词义及情感倾向性：在句(1)中的“幼稚”表达的词义是负向的情感倾向性，而在句(2)中表达的词义却是正向的情感倾向性。由上可见，单纯通过情感词典判断类似“幼稚”这样的含有多词义且多情感倾向性的词语有一定局限性，因此需要结合其处所的上下文环境进行词义及倾向性的判断。

目前在词义消歧上，国内外已有不少成熟的方法。其中，何径舟等[3]在分析了特征模板。

* **收稿日期：**　　**定稿日期：**

基金项目：国家自然科学基金资助项目（60973068，61277370）；辽宁省自然科学基金（201202031）

作者简介：作者一杨亮（1986——），男，博士研究生，主要研究领域为情感计算与观点挖掘。Email: yangliang@mail.dlut.edu.cn；作者二张绍武（1967——），男，副教授，主要研究领域为情感计算与观点挖掘。 Email: zhangsw@dlut.edu.cn；作者三林鸿飞（1962——）,男，教授博导，主要研究领域为搜索引擎、文本挖掘、情感计算和自然语言处理。Email: hflin@dlut.edu.cn。作者四宋艳雪（1986——），女，硕士研究生，主要研究领域为情感计算与观点挖掘。

对消歧结果影响的基础上，提出一套基于最大熵分类模型的自动特征选择方法来实现词义消歧。张仰森等[4]针对最大熵原理只能利用上下文中的显性统计特征构建语言模型的缺点，提出了隐最大熵原理构建词义消歧模型；通过构建面向词义消歧的条件随机场模型库，车玲等[5]通过实验证明，低频义项可以取得较好的消歧效果。与此同时，Mihalcea[6]提出了基于Wikipedia 进行词义消歧的方法。Navigli 等[7]提出了一种多语联合词义消歧方法。该方法通过利用多语知识库和不同语言的译文作为补充，进行了基于图的词义消歧。另外，通过从Web 上自动地抽取不同领域的术语并将这些术语作为语义知识，Stefano[8]提出了一种无监督的领域词义消歧方法。然而，目前鲜有研究者从情感倾向性角度进行词义消歧。以情感消歧为出发点，陈建美等[9]通过贝叶斯方法取得了较好的效果。然而有指导的监督学习方法跨领域性适用性差，针对不同领域需要重新标注部分信息，因此需要耗费大量的人力物力，鉴于此，本文提出了基于图排序的无监督词汇情感消歧算法，以此解决上述类似问题。

本文在解决词汇情感消歧时，充分考虑情感词所处的上下文语境。在对语料进行预处理后，利用《现代汉语词典》构建词义关系图，并通过 PageRank 算法进行迭代计算直至其收敛。然后，选取多情感词所含词义中具有最大权值的词义作为该情感词的最终词义，从而实现词汇的情感消歧。最后，在新浪微博数据集和大连理工大学信息检索实验室情感语料库[10]（下文简称情感语料库）两个语料集上验证了本文方法的有效性。

2 理论基础

2．1 情感词汇本体

本文使用的情感词典资源为大连理工大学信息检索实验室的情感词汇本体[11]（下文简称情感词汇本体），该情感词汇本体将情感分为 7 大类 20 小类，目前收录情感词 17000 余条。对于每个情感词，通过一个三元组来描述：

$$Lexicon = (B, R, E) \quad (1)$$

其中 B 表示词汇的基本信息，主要包括编号、词条、对应英文、词性等信息。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。E 代表词汇的情感信息，包括情感类别、情感强度、情感极性，是情感词汇描述框架中比较重要的一部分。图 1 表示“美丽”一词在情感词汇本体中的存储状态以及各个变量所存储的值。其中<num>表示“美丽”的编号，<lex>表示本词条所存储的词汇，<ccat>表示词性，<eng>表示英文表达方式，<emotion>域表示该词包含的大类情感，其中的“PA”、“PH”、“PB”分别代表大类情感中的“快乐”、“赞扬”、“喜欢”。<intensity>域采用 20 维向量形式表示，每一维代表 20 小类相应情感的强度。其中 0 表示不含该类情感，强度 1、3、5、7、9 表示强度由小到大。<polarity>表明词汇极性，有褒义、贬义、中性、褒贬兼有四类。<emotion_ class>表明词汇包含的主要情感是消极、积极还是中性。

```
<num>APA00108</num>
<lex>美丽</lex>
<ccat>a</ccat>
<eng> beautiful</eng>
<emotion>PA,PH,PB</emotion>
<intensity>3,0,0,0,0,5,7,0,0,0,0,0,0,0,0,0,0,0,0,0</intensity>
<polarity>1</polarity>
<syn>漂亮、好看</syn>
<emotion_class>A</emotion_class>
<standard>0</standard>
```

图 1　情感本体存储示例

Fig.1　A Example of Affective Lexicon Ontology

由于大量网络流行用语经常出现在社交媒体的文本中，而且常常带有明显的情感倾向性。为了使情感词典涵盖范围更广，本文在情感词汇本体的基础上整合了如“给力”、“顶”等当前网络流行词汇，其主要来自中文倾向性评测任务，共 153 个网络常用流行词汇，以此辅助本文情感消歧任务。

2.2 PageRank 算法

PageRank[12]用于衡量特定网页相对于搜索引擎索引中其他网页的重要程度。它充分利用了互联网资源中浩瀚复杂的链接结构。一个页面的“得票数”，即重要性，由所有链向它的页面的重要性来决定。所以，到一个页面的超链接相当于对该页面的投票。一个页面的 PageRank 值是由所有链向它的页面（“链入页面”）的重要性经过递归计算得到的。一个有较多链入的页面会有较高的等级，相反如果一个页面没有任何链入页面，那么它没有等级。PageRank 算法目前已经被广泛的应用到了网页链接分析、社交网络、引文分析等领域中。它通过公式(2)计算每个网页的 PageRank 值，其中 c 设定为 0.85[15]。

$$PR(X)=(1-c)+c\left[\frac{PR(T_1)}{N_{T_1}}+\ldots+\frac{PR(T_n)}{N_{T_n}}\right] \tag{2}$$

PageRank 之所以成功，归咎于它考虑到了以下三个要点：首先，web 页反向链接的数目，即该 web 页受欢迎的程度。其次，web 页反向链接是否来源于权威性网页，即要考虑反向链接网页的重要性。最后，web 页反向链接页面的链接数，即要考虑该 web 页被选中的概率。

3 基于图排序的词汇情感消歧模型

3.1 多情感词汇的获取

多情感词汇是指具有不同情感倾向性的词汇，其表达的情感倾向性依赖于所处的语境，。如“骄傲”一词在下面两个句子中所要表达的情感倾向性：

a. 莉莉考上了名牌大学，爸爸妈妈都感到非常的骄傲。

b. 公主般的莉莉总是那么骄傲，从来不把别人放在眼里。

在《现代汉语词典》中，“骄傲”有三个词义：(1)自以为了不起，看不起别人；(2)自豪；(3)值得自豪的人或事物。显然，在句 a 中“骄傲”表达的是词义(2)。而在句 b 中，其所要表达的却是“自以为了不起，看不起别人”的意思，即词义(1)。从情感倾向性来看，“骄傲”一词在句 a 中表达的是正向情感倾向性，而在句 b 中表达的是负向情感倾向性。类似于“骄傲”这样在不同语境中表达不同情感色彩的词汇本文称之为多情感词。一个词汇有多种情感的问题可以看作是词汇多义问题造成的。那么解决词汇情感消歧问题相对应的看作解决词义消歧问题的延续，因而它们之间存在共性。多情感词汇的挖掘和其情感的确定可以依赖词义消歧方法，但是二者之间又有所差异，需要根据多情感词汇本身的特性进行相应改进及处理。

多义词的确定可以根据《现代汉语多义词词典》、《常用多义词词典》等词典实现。然而，目前没有权威的准则或词典来确认一个情感词是否为多情感词汇，更不可能确定多情感词汇到底包含哪几种情感。因此，为从情感词汇本体中挖掘出多情感词汇，本文提出了机器过滤与人工校对相结合的方法，具体过程如下所述：

(1) 机器过滤

该阶段主要通过两层过滤手段实现。根据语言习惯及观察实验语料，本文发现一个能表达多种情感的词也往往含有多个词义，且每个词义可能表现出不同的情感，故多情感词汇很可能是多义词。为了挖掘多情感词，首先要筛选出多义词。为此，本文通过参照《同义词词林》[13]筛选出包含在情感词汇本体中且存在多个词义的词汇，将其作为候选。在《同义词词林》中，如果一个词存在于多个组中，本文认为此类词是多义词，例如“骄傲”在《同义词词林》中存在于下面的两个组中：

Da13A01 = 荣誉 荣耀 荣幸 光荣 光彩 光荣 骄傲 桂冠 殊荣。
Ee34D01 = 骄傲 骄矜 矜夸 傲慢 骄慢 神气 高傲 傲视 傲岸。

图 2　多义词示例

Fig.2　A Sample of Polysemy

依据上述分析，第一层过滤首先提取在《同义词词林》中有两个及以上词义且被情感词汇本体收录的词汇，如“骄傲”等。经统计，首次过滤出来的词集合 M 包含 901 个词汇。

第二层过滤是通过情感词汇本体描述框架中的 20 维向量<intensity>进行的。这 20 维的向量代表该情感词在 20 小类情感上的相应情感强度。集合 M 中的词汇，并不一定都是多情感词汇。所以，对于集合 M 中的每个词汇，若其在<intensity>向量上只有一个分量大于 0，则表明其只有一种情感，故不属于多情感词，应过滤掉。若在情感词汇本体描述框架中在<intensity>向量上含有两个及两个以上分量大于 0 的，类似“骄傲”一词的<intensity>向量形式为：<intensity>5，0，0，7，0，0，0，0，0，0，0，0，0，0，0，0，7，0，0，0</intensity>，这表明“骄傲”分别有“快乐”、“赞扬”和“贬责”三种情感，即表明“骄傲”为多情感词汇，应该保留下来。本文将第二次过滤后保留下来的词汇集合表示为 N。

(2) 人工校对

为进一步保证多情感词汇的选取质量，本文接下来进行人工校对。对于词汇集合 N，我们根据《现代汉语词典》提取出精准的多情感词汇。为避免个人主观性影响，校验过程中，本文采取 3 人独立校验，然后取 3 人校验结果的交集部分，最后得到确定多情感词 236 个。

3.2 基于 PageRank 排序的词汇情感消歧

针对消歧原理，本文对 PageRank 进行改进，并将其应用在词语情感消歧问题中。下面是一个 PageRank 的计算例子。图 3 表示的是一个 web 页面的链接结构图。其中节点 A、B、C 代表三个 Web 页面，有向边代表页面的链接结构。PR(a)、PR(b)、PR(c)分别表示节点 A、B、C 的 PageRank 值，在图的右侧定义了各个节点 PageRank 值的计算公式。图下方给出了各个节点前三次迭代值和最终迭代值的详细计算过程。

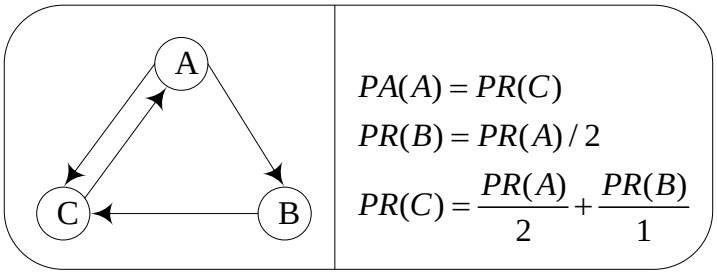

图 3　网页链接示例

Fig.3　A Sample of Web Link

第一次：PR(C)=0.33/2+0.33=0.5　PR(A)=0.33　PR(B)=0.17
第二次：PR(C)=0.33/2+0.17=0.33　PR(A)=0.5　PR(B)=0.17
第三次：PR(C)=0.42　PR(A)=0.33　PR(B)=0.25
最终值：PR(C)=0.4　PR(A)=0.4　PR(B)=0.2

由上述例子可以看出，迭代结束后，图中每个顶点的 PageRank 值代表了该顶点在图中的重要程度，即在随机游走过程中找到该顶点的可能性。PageRank 算法的“投票”思想同

样适用于词汇的情感消歧。本文将词汇的多个词义视为图上的节点，链接到某一个词义顶点的链接数目越多说明该顶点与上下文语境的相关性越大，即该词义越有可能是符合该语境下的词义。在进行情感消歧时，本模型通过在词义关系图上游走，最终的稳定分布概率值可以被用来决定所给定序列最可能的词义集合。

此部分将介绍关系图的构造。对于一个给定的词序列 $W=\{w_1, w_2, \ldots, w_n\}$，《现代汉语词典》中，每一个词 w_i 的词义表示成如下形式：

$$L_{w_i} = \{l_{w_i}^1, l_{w_i}^2, ..., l_{w_i}^m\} \tag{3}$$

其中 m 表示词 w_i 的词义数。n 表示词序列 W 中词语的个数。

本文将基于词义的有向图表示为：G = (V ， E)，其中 V、E 分别表示顶点和边集合。顶点集合 V 由词序列 W 中所有的词的全部词义构成，每个词义映射为图 G 中的一个顶点。$wg(w_i^k, w_j^h)$表示词 w_i 的第 k 个词义与 w_j 的第 h 个词义之间的词义关系，通过公式(4)并结合同义词词林中的各词义的概念计算得到。

$$wg(w_i^k, w_j^h) = \frac{SAME(w_i^k, w_j^h)}{N_{w_i^k} + M_{w_j^h}} \tag{4}$$

其中 $SAME(w_i^k, w_j^h)$表示词 w_i 的第 k 个词义和词 w_j 的第 h 个词义之间相同词的总个数，分母为 w_i 的第 k 个词义和 w_j 的第 h 个词义中词数总数。

鉴于句子中下位词对上位词的语义选择有很大的影响，本文用公式(5)计算下位词 w_j 的第 k 个词义与上位词 w_i 的第 h 个词义之间的语义关系 $wg(w_j^k, w_i^h)$。

$$wg(w_j^k, w_i^h) = \log_2 \frac{P(w_j^k, w_i^h)}{P(w_j^k)P(w_i^h)} \tag{5}$$

其中 $j > i$，w_i^k 表示词 w_i 的第 k 个词义。

在词义关系图中，词义与词义间的依赖关系可通过有向边权重的大小表示。通过权重大小来衡量依赖关系的强弱，当边的权重为 0 时表示两个词义之间没有依赖关系。图 4 展示了 4 个序列词构成的词义关系图，表示了 4 个词序列词义间的依赖关系。对于一个给定的词义关系图，可以通过图排序算法得到每个词中各个词义被选中的权值。即在词义关系图上随机游走后得到的稳定权值，其决定了该顶点的重要性。图中每个顶点旁边方括号中的数字表示最终的稳定权值分布。迭代开始时，每个顶点的初始值都为 1，待收敛后，所有词义中概率最大的词义即为该情感词的最终词义。如图所示，由于在 w_1 的所有词义中，词义 1 的最终迭代权值 1.39，在三个词义中最大，故选取词义 1 作为最终词义。

图排序算法的全局性是解决词汇情感消歧问题的关键，其不仅仅依赖于本地的特殊顶点或者单个顶点信息，而是从全体性出发挖掘词义之间的依赖关系。设已给定的顶点 b 和 a 间有向边的权重是 w_{ba}，则顶点 a 的迭代计算公式（6）如下所示。

$$WP(V_a) = (1-c) + c \sum_{V_b \in In(V_a)} \frac{w_{ba}}{\sum_{V_c \in Out(V_b)} w_{bc}} WP(V_b) \tag{6}$$

基于图排序的词汇情感消歧算法主要包含以下三个步骤：(1) 构造词序列 W 的词义关系图；(2) 计算图中每个顶点的 WP 值；(3) 利用 WP 值实现词汇的情感消歧。具体过程为：对于所有词，将其在《现代汉语词典》中的每个词义作为顶点加入图中。通过公式(4)、(5)计算任意两个顶点之间的权重，并将其作为有向边的权重加入图中。构建图时，本文通过最

大距离 MaxDist 来约束权重的计算，即在寻找与词义 i 有关系的词义 j 时，允许跨越最多 MaxDist 的距离。对于 MaxDist 的设定主要考虑语言及思维习惯，即一个词在句中的词义受其前后词影响。若选择过大的间距则会引入较大的噪音，若选择间距过小则可能丢失词与词之间的语义信息，综合二者考虑，本文将 MaxDist 设为 3，即最大允许跨越的距离为 3。在词义关系图构建完成后，通过公式(6)迭代计算直至收敛，最后得到每个顶点的 *WP* 值。对于每个多情感词，选取其所有词义中 *WP* 值最大的词义作为当前语境下的词义。

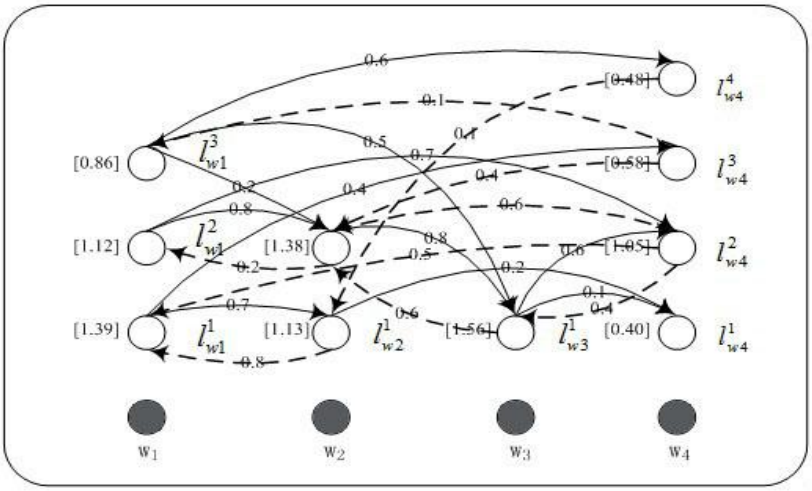

图 4　词义关系图示例

Fig. 4　Sample Graph

4 实验结果与分析

4.1 实验设置

为了说明本文算法的可移植性和鲁棒性，本文分别在微博语料和情感语料库上与两种方法作了对比，二者为基于词性和情感频率的方法和基于贝叶斯模型的词汇情感消歧方法。情感语料库中包含 250021 个句子，句子覆盖小学教材、电影剧本、童话故事、文学期刊，内容表达比较规范，从时间、空间、学科、风格和构成上看覆盖面大。而微博内容的主题多样，表达随意，并且每条微博所包含的信息量少，文字简短，事件核心突出。所以，采用两种风格不同的语料更能验证本文所提出的方法的可移植性和鲁棒性。

本文首先爬取新浪微博文本内容作为备选语料集，然后筛选出带有多情感词的句子。另外，为了构建相对完整的词义关系图，本文不考虑特别短小的句子（存在信息丢失等问题）和广告等噪音数据。对筛选过后的微博句子进行分句，根据每个多情感词汇，选取包含它的 30 个句子作为后续实验语料。之所以选择 30 条作为标准，是由于根据观察，超出 30 条后所获取的重复句子明显增加，很少能再获得新的实例。最后，对筛选出来的语料中每个句子所出现的多情感词汇进行词义及情感标注，标准参照为 3.1 节中的部分。

为检验方法的有效性，本文设置如下两个对比实验：(1)词性 ccat 与情感频率 fs 相结合的词汇情感消歧，此方法是一种基于统计的方法；(2)基于贝叶斯模型的词汇情感消歧，该方法在已标注语料上提取多情感词汇的属性及特征，然后训练模型进而对测试语料进行词汇情感消歧。在进行基于贝叶斯模型的词汇情感消歧时，将语料按照 2:1 的比例分为训练语料和测试语料进行实验。下面介绍上述两种对比方法的实现过程：

1.基于词性和情感频率的词汇情感消歧：(1) 使用分词软件 NLPIR[14]对句子进行分词、词性标注，去停用词，并设句子中的多情感词汇为 w，词性为 p。(2) 在情感词汇本体中查

找<lex>域为 w、<ccat>域为 p 的词条，查看其<emotion>域中的值是否唯一。当<emotion>中的值多于 1 个时，统计其各个情感在标注语料中的出现频率，并将该词在语料库中情感频率最高的作为该词在此句中的情感。

2.基于贝叶斯模型的词汇情感消歧：该方法首先在已标注语料中统计多情感词的词义和其上下文语境的关系，进而得到一个知识库。然后计算多情感词 w 在特定的语用环境 C 下表现各种情感的后验概率值，最后根据后验概率大小决定其所述类别。其中，*count* 表示所获得的相关句子在语料库中所出现的总数。

$$\begin{aligned} sense(w) = s_i &= \arg\max P(s_i|C) = \arg\max \frac{P(C|s_i)P(s_i)}{P(C)} \\ &= \arg\max P(C|s_i)P(s_i) = \arg\max \prod_{w_k \in C} \frac{Count(w_k, s_i)}{Count(s_i)Count(w)} \end{aligned} \tag{7}$$

下面介绍本文所提出方法的基本流程：对于每个筛选出的微博语句进行分词、词性标注、去停用词后，将剩余词在《现代汉语词典》中的所有词义映射为词义关系图中的顶点。对于图中的任意两个顶点，根据公式(4)和公式(5)计算其之间的有向边权重。另外，在计算 SAME 值时应用《同义词词林》进行扩展。通过对 $wg(w_j^k, w_i^h)$的计算可以进一步充分利用语义信息。

通过实验，本文发现随着 MaxDist 的增大，词义间的依赖性逐渐衰退，且当 MaxDist=3 时所得到的信息最大。待词义关系图建成后，初始每个词义顶点的 *WP* 值为 1，按照公式(6)对图中的顶点迭代计算。实验中发现经过 20 次的迭代计算后，每一个词义顶点的 WP 值基本趋于稳定。最后，选取情感词的所有词义中 *WP* 值最高的作为该情感词的情感倾向性，实现词汇的情感消歧。

4.2 结果及分析

本文用准确率作为实验结果评价指标，此处指的是情感倾向性判断正确的多情感词数量占待预测词汇总量的比例。表 1 展示了三种方法在微博语料上的实验结果：

表 1　微博语料上的对比实验

Tab.1　The Comparison Based On MicroBlog Corpus

实验名称	正确率
基于词性和情感频率的词汇情感消歧	68.22%
基于贝叶斯模型的词汇情感消歧	71.46%
基于图排序模型的词汇情感消歧	73.51%

分析实验结果可以发现，基于词性和情感频率的情感消歧方法的正确率为 68.22%。虽然多情感词有多个词义，但在生活中，人们通常只会常用其某一个词义，表达某一种情感。即最常用的词义，最多见的情感会应用在日常表达交流中。所以，基于词性和情感频率的词汇情感消歧方法能获得 68.22%准确率。伴随着网络文化的发展，许多网络流行用语日益涌现。微博作为当下比较流行的社交媒体，其文本形式受限于时间、空间等诸多因素，即某时段的微博语料主要和该时间段内所发生的热门话题有关。鉴于此，该方法的正确率有待提高。

相比基于词性和情感频率的词汇情感消歧方法，基于贝叶斯模型的词汇情感消歧方法大约提高了 3.24%，但是其效果却低于基于图排序模型的词汇情感消歧方法约 2%。本文认为主要由以下原因导致：

(1) 训练集的规模、领域都会都对贝叶斯分类模型有一定的影响。另外，特征选取的质量直接影响到分类结果。

(2) 由于微博更新速度较快、内容短小，主题多样，所以当测试集和训练集主题有所差异时，对测试集语料来说，分类模型可能无法获得部分先验知识作为参考，故导致分类结果

不理想。这也就是其针对跨领域问题上没有图排序模型效果好的最主要原因。

相比前两种情感消歧方法，本文在微博语料上相比前两种消歧方法体现了其优越性。基于图排序模型方法在准确率上分别有 2.04%和 5.29%的提高。这是由于该方法是基于词义依赖关系，从整体出发充分考虑了上下文的语义环境。在进行情感消歧时，不依赖于训练集的规模和特征的选取质量，同时也不受限于文本内容的领域和主题。综上所述，该方法取得了更好的效果，但仍有提高的余地。分析实验结果我们发现存在以下问题：

(1) 在词典中描述词汇词义的句子一般较为短小，包含的词语比较少，导致在计算词汇间相似度时受到影响。

(2) 在微博语料中，表达相对随意，且新组合词、网络流行用语以及新生僻词较多。而通常这些比较流行的网络用语及组合词却没有被《现代汉语词典》所收录，在一定程度上影响了实验精确度。同时微博句子比较短小，表达形式随意，相对不规范，甚至经常出现只言片语的情况。所以导致词义关系图构建相对比较困难，进而影响词义相似度的计算，也是影响实验精度的重要因素之一。

(3) 在现代汉语词典中，示例信息往往可以更好的反应该词义所要表达的情感信息，因为相比词义的定义，示例内容更接近人们表达的实际情况。所以，充分利用示例信息是我们下一步的工作之一。

为了验证本文所提出的方法在标准语料集上的有效性，本文将上述提到三种方法在情感语料库上进行了实验，并和在微博语料上取得的精度进行了对照，结果如图 5 所示：

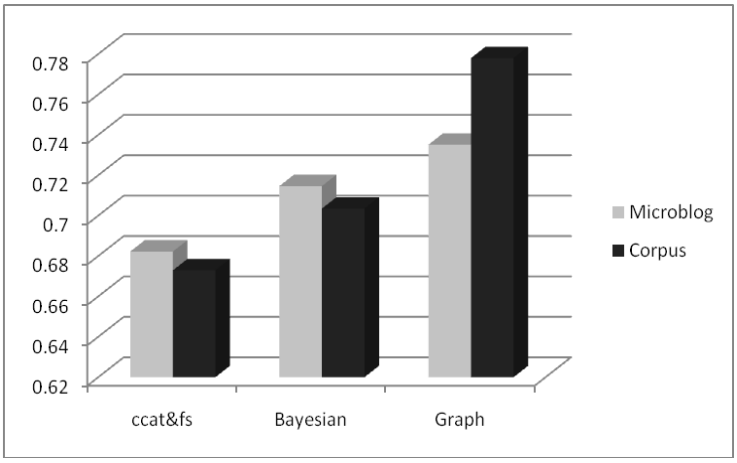

图 5 语料库和微博的实验对比结果

Fig.5 The Comparison of Methods Between MicroBlog and Corpus

从上图可以看出，在情感语料库上，基于图模型的情感消歧方法仍优于其它两种方法。这主要是由于该方法基于语义分析，不受限于特征的提取精度和语料自身特性，所以在情感消歧准确率上表现相对较好。

分析基于词性和情感频率的词汇情感消歧法在情感语料库和微博语料上的结果可以看出，在微博语料上取得的精度相对较高。这主要是由于两种语料在行文风格、知识背景、描述主题等方面的差异所导致。情感语料库中表达比较规范，相对较为书面化，通常采用比较含蓄的方法抒发感情。而在微博中表达比较随意，较为口语化，情感抒发方式相对直接。相比情感语料库上，基于贝叶斯模型消歧方法在微博语料取得的结果也相对较好。这主要是由于情感语料库覆盖范围较广，包括小学教材、电影剧本、童话故事、文学期刊等。所以分类模型很可能无法获得某些领域或主题的先验知识，进而影响了分类精度。这也验证了监督学习在跨领域问题处理上的欠缺。

从图5我们可以发现，不同于前两种方法，基于图模型的消歧方法在情感语料库上表现相对较好。这主要是由于微博的内容相对短小，构建完整的词义关系图比较困难，进而影响了实验准确率。而情感语料中的表达方式比较规范，能够较为准确地构建词义关系图，因此实验结果相对微博数据较好。

综上所述，通过在两种表达方式不同的语料集上进行测试，验证了本文提出的基于图排序模型的词汇情感消歧方法都优于其它两种对比方法。这充分说明了该方法的有效性，也体现了本文方法在跨领域性适用性和鲁棒性方面的优势。

5 结束语

本文详细介绍了基于图模型的词汇情感消歧的方法，并在微博语料库和情感语料库上验证了该方法的有效性。下一步的工作是充分利用《现代汉语词典》中的示例信息，因为示例比词义定义更接近人们的用语习惯，将示例和上下文的互信息性也考虑到词义的相似度计算中。另外，由于在特定领域内语义与情感关联性很强，因此将词义的领域信息融入词汇情感消歧中也是未来重要的工作之一。

参考文献

[1] Pang B, Lee L. Opinion mining and sentiment analysis[J]. Foundations and trends in information retrieval, 2008, 2(1-2): 1-135.

[2] Liu B, Zhang L. A survey of opinion mining and sentiment analysis[M]. Mining Text Data. Springer US, 2012: 415-463.

[3] 何径舟, 王厚峰. 基于特征选择和最大熵模型的汉语词义消歧[J]. 软件学报, 2010, 21(6): 1287-1295.

[4] 张仰森, 黄改娟, 苏文杰. 基于隐最大熵原理的汉语词义消歧方法[J]. 中文信息学报, 2012, 26(3): 72-78.

[5] 车玲, 张仰森. 面向词义消歧的条件随机场模型库构建[J]. 计算机工程, 2012, 38(20).

[6] Mihalcea R. Using wikipedia for automatic word sense disambiguation[C]//Proceedings of Human Language Technology conference and conference on Empirical Methods in Natural Language Processing, Rochester, 2007, 196–203.

[7] Navigli R, Ponzetto S P. Joining forces pays off: Multilingual joint word sense disambiguation[C]//Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning. Association for Computational Linguistics, 2012: 1399-1410.

[8] Faralli S, Navigli R. A new minimally-supervised framework for domain Word Sense Disambiguation[C]//Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Association for Computational Linguistics, 2012: 1411-1422.

[9] 陈建美,林鸿飞.基于贝叶斯模型的词汇情感消歧[C]//第九届全国计算语言学学术会议论文集,大连, 2007:594-599.

[10] Yang L, Lin H. Construction and application of Chinese emotional corpus[M]. Chinese Lexical Semantics. Springer Berlin Heidelberg, 2013: 122-133.

[11] 徐琳宏,林鸿飞,潘宇,等.情感词汇本体的构造[J]. 情报学报, 2008, 27(2): 180-185.

[12] 曹军. Google的PageRank技术剖析[J].情报学报, 2002,10:15-18.

[13] 哈尔滨工业大学《同义词词林》扩展版[EB/OL]. http://ir.hit.edu.cn/phpwebsite/index.php?module=pagemaster&PAGE_user_op=view_page&PAGE_id=162.

[14] NLPIR分词系统[EB/OL]. http://ictclas.nlpir.org/.

[15] PageRank[EB/OL]. http://zh.wikipedia.org/wiki/PageRank.

作者简介：

作者一杨亮（1986——），男，博士研究生，主要研究领域为情感计算与观点挖掘。Email: yangliang@mail.dlut.edu.cn；

作者二张绍武（1967——），男，副教授，主要研究领域为情感计算与观点挖掘。 Email: zhangsw@dlut.edu.cn；

作者三林鸿飞（1962——）,男，教授博导，主要研究领域为搜索引擎、文本挖掘、情感计算和自然语言处理。Email: hflin@dlut.edu.cn;