一种基于组合语义的文本情绪分析模型

乌达巴拉 [1, 2] 汪增福 [1, 2]

摘 要 文本情绪分析属于细颗粒度文本情感分析范畴. 传统的基于监督学习的方法, 大多注重从表面词形提取特征, 对语言的结构化特征考虑较少, 无法应对特征稀疏问题, 也无法挖掘文本中隐含的深层语言信息 (包括词语搭配和语义韵). 上述问题的存在导致现有系统的分类性能不高, 尤其对隐性文本情绪分类问题表现出较大的局限性. 本文尝试将基于依存句法的词语搭配特征和基于组合语义的深度特征应用于文本情绪分类, 提出了一种以短语为主要线索的半马尔科夫条件随机场文本情绪分析模型. 为了验证模型的有效性, 利用实际构建的相关实验语料, 开展了相关实验研究. 实验结果表明, 本文方法不仅可以显著提高文本情绪分类的准确率, 而且对解决隐性情感分析问题也具有重要作用.

关键词 文本情绪分析, 隐性情绪分类, 组合语义, 半马尔科夫条件随机场

引用格式 乌达巴拉, 汪增福. 一种基于组合语义的文本情绪分析模型. 自动化学报, 2015, **41**(12): 2125−2137

DOI 10.16383/j.aas.2015.c150064

Emotion Analysis Model Using Compositional Semantics

Odbal[1, 2] WANG Zeng-Fu[1, 2]

Abstract Emotion analysis is one of the challenging tasks in the fine-grained sentiment analysis field. Conventional supervised learning approaches reply on features based on surface word forms and neglect linguistic structure of text. Hence, they usually suffer from sparse features and are unable to exploit implicit information such as collocation and semantic prosody, leading to unsatisfactory performance, especially for identifying implicit emotion expression. This paper first explores dependency parsing features and compositional semantics information in emotion. Then it proposes a phrase level emotion detection model based on semi-CRFs (semi-Markov conditional random fields) and finally creates several corpora and carries out corresponding experiment. The experimental results indicate that the model and features are remarkably effective to classify sentence or short text by fine-grained emotion tags and play an important role in implicit emotion detection.

Key words Emotion analysis, implicit emotion recognition, compositional semantics, semi-CRFs (semi-Markov conditional random fields)

Citation Odbal, Wang Zeng-Fu. Emotion analysis model using compositional semantics. *Acta Automatica Sinica*, 2015, **41**(12): 2125−2137

情绪是对一系列主观认知经验的通称[1]. 现代社会中情绪无处不在. 特别地, 随着计算机和网络技术的快速发展, 情绪的表达方式也日益多样化. 它广泛存在于博客、微博和微信等互联网社交媒体中. 所谓情绪分析 (Emotion analysis), 就是对各种信息资源中的情绪以 “喜悦、愤怒、悲伤、恐惧、惊慌、厌恶” 等进行分类的研究. 由于现在大部分信息资源都是基于文本的, 又可称为文本情绪分析. 文本情绪分析的研究内容呈现跨学科的特点, 具有极其广泛的研究和应用价值.

情绪的表达有显性和隐性之分. 显性情绪表达是直观的, 主要呈现在语言形式的表层, 我们可以在绝大多数情况下仅根据表达情绪的词汇本身就能比较准确地分析出文本的情绪. 褒义词、贬义词、情态语、强化语等都是表达显性评价或态度的可利用手段. 而隐性情绪表达则是暗藏的, 它存在于语言表述的深层涵义中, 有时甚至是说话者不经意地 “言不由衷” 表述下掩盖的真实情绪, 在不同的语境中所体现的语义色彩是不同的. 例如,“时髦” 这个词往往表现出消极的语义倾向. 但是, 在句子 “这小镇就形成了古朴而又时髦, 偏僻而又繁华的特色” 中,“时髦” 一词则显示了积极的语义特征. 对于隐性或暗藏的态度意义, 我们需要借助于特定的词语搭配和语义韵 (Semantic prosody) 来渲染. 语义韵[2] 是词语搭配

收稿日期 2015-02-13 录用日期 2015-09-23
Manuscript received February 13, 2015; accepted September 23, 2015
国家自然科学基金 (61472393) 资助
Supported by National Natural Science Foundation of China (61472393)
本文责任编委 赵铁军
Recommended by Associate Editor ZHAO Tie-Jun
1. 中国科学院合肥智能机械研究所 合肥 230031 2. 中国科学技术大学自动化系 合肥 230027
1. Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei 230031 2. Department of Automation, University of Science and Technology of China, Hefei 230027

的延伸和发展, 它揭示了词汇极具特征的表现方式和功能指向. 语义韵具有短语属性, 因此对隐性情绪表达的分析需要从短语层面进行处理.

在以往的工作[3−5] 中, 研究者们通常将文本情绪分析看成是分类问题进行处理: 首先, 基于人工标注的方式构建情绪标注语料; 然后将标注好的语料分成训练集和测试集两个部分, 从训练集中提取各种词汇、语法、符号等特征, 生成相应的文本情绪分类训练集, 并利用朴素贝叶斯 (Naive Bayes, NB)、支持向量机 (Support vector machine, SVM)、最大熵 (Maximum entropy, ME) 等模型得到分类器模型; 最后利用分类器模型对测试集进行预测, 给出测试集可能表现出的情绪类别. 然而, 已有方法主要存在以下两个方面的问题: 1) 仅引入词语级别的特征, 比如词语的 unigram、bigram 以及词性信息等, 未考虑短语层面的特征, 比如依存句法信息, 而这类特征对文本情绪分析有重要的作用, 特别是对隐性情绪分析; 2) 仅根据表面词形提取特征, 存在特征稀疏问题; 此外, 因无法挖掘隐含的语言信息 (包括词语搭配和语义韵), 导致系统分析能力有限, 尤其对文本的隐性情绪分类问题表现出较大的局限性. 下面通过例子来说明利用表面词形特征的词语级文本情绪分析方法存在的问题.

例句 1. 不过在教堂里, 站在讲台上的牧师却是大叫大嚷, 非常生气. [Anger]

例句 2. 迷信使她的血一会儿变冷, 一会儿变热. [Fear]

例句 3. 中式内衣展现东方诱惑. 服装注重剪裁, 完美展现年轻少女气息. [Joy]

由于显性情绪表达存在显式的情绪词, 我们可以在句子的层面以直截了当的方式使用明显的情绪词语就可以比较准确地分析出文本的情绪类别. 例如, 例句 1 中存在显式的情绪词 “大叫”、“大嚷” 和 “生气”. 这些词语通常表示生气或愤怒的情绪. 显而易见, 例句 1 的情绪是 [愤怒 (Anger)]. 相反, 由于隐性情绪表达不存在显式情绪词, 我们只能根据一些词语的搭配和意义等语言学特征对其进行分类. 比如, 例句 2 中 “变冷” 具有消极语义韵, 更多的时候这个词是用来表达人的心情变失落, 感受不到温暖, 传递一种伤感的情感. 而 “变热” 常和表示褒义的词语组合, 表达一种热烈、积极、主动、友好的情感或态度. 但在这里 “变冷” 和 “变热” 的组合隐含了一种 [害怕 (Fear)] 的情绪. 再如, 例句 3 中词语 “诱惑” 倾向于表达某些贬义的目的和不光彩的方式方法, 但在这里与 “展现” 相结合, 表达一种赞美的情感, 显示了 [喜悦 (Joy)] 的情绪. 从例句 2 和例句 3 可以看出, 由词语搭配形成的语义信息对情绪分类的影响, 它具有词语级别特征所不具备的优势.

然而, 基于表面词形的词语搭配特征缺少对词语组合之间语义规则的利用, 并且有可能导致一部分语义信息的丢失. 例如, 例句 2 中的词语 “变冷” 和 “变热” 的组合属于低频共现, 利用这种稀疏特征无法获取到准确的情绪类别信息. 组合语义 (Compositional semantics) 方法的引入为上述问题的解决提供了有效的技术手段. 组合语义采用概率统计的方法来描述词语搭配的趋势, 并希望据此发现文本上下文间的语义关系和潜在的概念结构 (如词汇间的同义关系). 例如, 例句 2 中的 “迷信” 一词, 其指示意义包括 [鬼]、[神] 等语义成分, 其内涵意义却可能有 “盲目”、“不科学”、“害怕” 等含义. 组合语义方法使从大规模语境中学习词语的内涵概念成为可能, 这为隐含情绪分析问题的解决奠定了良好的基础.

此外, 将词语搭配作为一种特征引入到现有判别式模型时存在一些问题. 传统的判别式模型 (比如 Conditional random fields, CRFs) 通常遵循马尔科夫独立性假设. 在这样的情况下, 虽然能够引入丰富的特征, 但是模型本身的适用性是建立在词条 (Token) 基础之上. 每个词条的标记与其他词条无关的假设限制了将其应用于更长单位 (比如短语或块) 的可能性. 因此, 当词语组合被视为观察序列时, 上述马尔科夫性假设使得传统的判别式模型表现出较大的局限性. 相比而言, 半马尔科夫条件随机场 (Semi-Markov condition random fields, Semi-CRFs) 是传统 CRFs 的扩展, 其所具有的半马尔科夫性允许其处理短语级别的特征, 从而通过使标记序列附着于短语级别的单位之上可以形成对基于词语搭配的短语层面特征的表达.

针对上述问题, 本文探索了将词语搭配特征和组合语义特征应用于文本情绪分析的可能性; 同时, 为了弥补现阶段判别式模型对词语搭配等短语级特征应用上的缺陷, 提出了一种以短语为主要线索的半马尔科夫条件随机场文本情绪分析模型. 最后, 为了验证本文所提出特征和模型的有效性, 实际构建了相关的实验语料, 并利用这些语料开展了相应的实验研究. 实验结果表明, 本文方法不仅可以显著提高文本情绪分类的准确率, 而且对解决隐性情绪分析问题也提供了一种很好的解决思路.

本文以下部分组织如下: 第 1 节对相关研究进行简要介绍; 第 2 节重点描述本文所采用的情绪分类特征; 第 3 节提出基于短语的文本情绪分析模型, 并讨论所涉及的参数估计和解码问题; 第 4 节介绍实验语料构建和实验方案, 并对实验结果进行分析. 最后是结论及下一步的工作.

1 相关工作

1.1 文本情绪分类

目前, 文本情绪分析研究主要采用监督的学习方法. Alm 等[3] 在 SNoW 的学习框架下研究了基于监督学习的文本情绪分类问题. 首先对包括喜悦、悲伤、恐惧、愤怒 – 厌恶、惊奇等基本情绪在内的情绪语料进行了人工标注. 然后采用朴素贝叶斯分类器对文本情绪进行分类. 采用的特征包括: 文本中的第一个句子、WordNet 情感词汇、特殊标点符号、正极和负极词数、相互影响的词汇等等. 他们对来自儿童寓言故事领域的语料 (185 个故事) 做了相应的测试, SVM 得到的最高准确率 (Accuracy) 达到了 69.37 %. 然而, 该方法只涉及基于词汇级别的特征. Aman 等[4] 基于 Ekman 定义的六种基本情绪, 即喜悦、悲伤、恐惧、愤怒、厌恶、惊奇等, 开展了对句子的自动情绪分类研究. 他们设计了基于语料库的 Unigram 特征和情感词汇特征, 并采用朴素贝叶斯和支持向量机对博文进行情绪分类. 使用的情感词典是基于 Roget 的词典和基于 WordNet 构建的词典. 采用了 PMI-IR 的语义相似度计算和概率情感评分的方法构建词典. 他们的方法取得的精确率 (Precision) 分别为: 高兴, 81.3 %; 悲伤, 60.5 %; 愤怒, 65 %; 厌恶, 67.2 %; 惊讶, 72.3 %; 害怕, 86.8 %. Das 等[5] 以博客为研究对象研究了基于句子的孟加拉文情绪分类方法. 他们的实现过程是: 首先利用英文 SentiWordNet 词典, 将英文情感词汇翻译成孟加拉文, 构建孟加拉文的 SentiWordNet 词典, 并根据 Ekman 的六种基本情绪进行标注. 然后采用条件随机场模型实现情绪分类任务. 使用的特征包括: 词性信息、论坛的第一个句子、特殊标点符号、否定词、表情符号 (Emoticons) 等. 近年来 CRFs 模型受到一些研究者们的青睐, 通过引入句法信息被应用于文本情感倾向性分析, 比如文献 [6 − 7] 等. Neviarouskaya 等[8−9] 利用博客领域的语料研究了基于规则的文本情绪标注方法. 他们首先将采集的博客内容按照句子进行划分, 然后依次进行符号性提示 (Symbolic cue) 分析、句法结构分析、基于词的文本情绪分类和基于短语的文本情绪分类, 最后是基于句子的文本情绪分类. 然而, 他们所提出的上述基于规则的方法, 强烈依赖于语言学约束, 缺乏对未登录词或隐含情绪的处理. Chaffar 等[10] 采用监督的学习方法, 利用多个领域混杂的语料库 (包括新闻标题、故事集以及博文等) 来识别六种基本情绪. 他们的实验结果表明, 在新闻领域句子 (1 250 条) 上 SVM 得到的准确率 (Accuracy) 是 39.6 %, 在 Alm 的语料[3] 上得到的结果是 61.88 %, 混合语料上得到的最高准确率是 71.69 %. 上述研究已证实 SVM 和 CRFs 分类器的效果要优于 NB 的结果. 然而, 上述研究仍停留在基于表面词形的不同词汇级别特征上. Mohammad 等[11] 研究了词语级情感词典 (Word level affect lexicons) 对句子级文本情绪分类的作用. 他们在基于 WordNet 的词语级情感词典 (WordNet affect lexicon) 和 NRC-10 的词语级情感词典的基础上, 利用 Logistic Regression 和 SVM 分类器对句子的文本情绪进行了预测, 取得了不错的成绩. 他们的研究表明语义关系对情绪分类的影响.

除了词汇相关的特征之外, 也有一些研究者注意到影响情感情绪的其他一些特征. 比如颜色、表情符号 (Emoticons)、话题标签 (Hashtags) 等. 相关研究举例如下: Volkova 等[12] 利用 Crowd sourcing 构建了一个颜色 – 情绪 – 概念词典, 并利用该词典开展了文本情绪分类方面的研究. Purver 等[13] 研究了面向微博的六种基本情绪的分类问题. 他们的研究证实了符号特征 (如 Emoticons 和 Related conventional markers) 的优势. Qadir 等[14] 采用 Bootstrapping 学习算法生成情感话题标签 (Hashtags), 并利用该特征对微博可能表现出的情绪进行预测, 相比 N-gram 分类器取得了较好的预测结果. 然而, 上述方法大部分是在表面词形的基础上进行特征提取, 存在词语特征稀疏问题, 且无法挖掘情感表达中隐含的语言现象.

与英文的情绪分析研究相比, 中文的情绪分析研究主要集中于文本情绪资源的建设. Quan 等[15] 构建了面向博客的中文情绪语料, 人工标注了八种基本情绪, 包括期待、喜悦、爱、惊讶、焦虑、悲伤、愤怒和仇恨等. 除此之外还标记了情绪的强度、情绪的施事者/受事者以及情绪词汇/短语、程度词汇、否定词、连接词、修辞 (Rhetoric)、标点符号等, 并利用该语料对博文进行情绪预测. Xu 等[16] 研究基于图的中文情绪词典构建方法. 该方法可以根据一些种子情绪词汇对词汇进行排序. 排序算法利用词与词之间的相似性和多种相似矩阵. 这些矩阵可以通过词典, 未标注语料库以及启发式规则获取. Wang 等[17] 研究了基于习语 (Idiom) 的情绪分析库的构建问题. Yang 等[18] 提出了一种基于情绪意识的 LDA (Emotion-aware LDA) 模型, 并据此建立了一个特定领域的词典. 预定义的情绪包括愤怒、厌恶、恐惧、快乐、悲伤、惊讶等. 该模型使用特定领域的小规模领域无关的种子词作为先验知识发现领域特定的词汇.

1.2 组合语义

根据 Montague (1974) 的阐述, 组合语义 (Compositional semantics) 的意义在于一个复合

表达式是由构成它的每个成分及其将这些成分组合在一起的句法规则所描述. 从表层意义来看, 组合语义是句子里词语搭配关系的体现. 近年来, 相关的研究在自然语言处理领域受到了一定的关注. 组合语义主要采用的表示方法有符号表示方法和分布式学习方法. 符号表示方法主要利用一些逻辑符号, 加上人工设计的规则进行语义组合 (Choi 等[19]). 而分布式学习方法是利用大规模的语料库进行规则学习. 首先将词语表示为向量集, 然后通过向量积、向量加、线性或非线性计算等方法实现组合语义 (Mitchell 等[20]; Baroni 等[21]; Socher 等[22−23]). 相关研究成果被应用于识别短语及计算短语相似度[24−27].

到目前为止, 国内外将组合语义方法应用于文本情感分析的研究还很少, 尤其是涉及文本情绪类别的分类问题更是如此. Choi 等[19] 提出了一种基于组合语义的子句情感倾向性分析方法. 他们首先评估子句表达式成分的极性, 然后递归运用一套相对简单的启发式推理规则 (根据组合语义学定义的规则, 组合表达式) 对文本的倾向性进行分析; Socher 等[22] 提出了一种基于半监督递归自动编码 (Autoencoders) 算法的句子级情感倾向性分布预测的新型机器学习框架. Autoencoders 是一种神经网络算法. 他们在没有使用任何情感词典以及情感转移规则的前提下运用该算法从层次结构和组合语义学角度分析情感, 并最终得到句子的情感极性; Moilanen 等[24] 基于组合语义规则, 建立了情感传递、情感转变、情感冲突的组合模型; Yessenalina 等[25] 利用组合语义方法研究了短语情感倾向性分类问题. 他们采用组合矩阵空间模型 (Compositional matrix-space model) 来解决语义组合关系表达问题, 采用有序逻辑回归模型 (Ordered logistic regression) 实现短语情感极性的多级分类. 通过对 MPQA 英文影评语料进行实验, 证实了该模型优于 Bag-of-Words 模型.

上述方法都是将组合语义应用于文本情感倾向性分析问题, 一般只考虑 “正极”、“负极” 和 “中性” 三个方面. 因此其组合结果也只涉及三种结果: “正极”、“负极” 和 “中性”. 然而, 文本情绪分类相比文本倾向性分析更加细致, 可以包括更多的种类. 比如 “喜悦、愤怒、悲伤、恐惧、惊慌、厌恶” 等, 甚至可以扩展到 “鼓励、讽刺、调侃” 等. 因此, 文本情绪分类问题相比文本倾向性分析更为细致复杂.

2 特征描述

本文引入了基于表面词形的基本特征和基于组合语义的深度特征. 基本特征包括功能实词及其词性信息、情绪词汇以及基于依存关系的词语搭配等特征. 下面分别加以阐述.

2.1 基于表面词形的基本特征

1) 功能实词及其词性信息. 我们通过对真实语料的分析, 发现表达情感的词汇往往是形容词. 除此之外, 动词和名词也有类似的作用. 但是数词、量词、介词等词汇的作用并无太大. 因此, 本文考虑将形容词、动词、名词等功能实词以及其词性作为特征. 中文分词和词性标注利用 Stanford 的分词和词性标注工具包 (http://nlp.stanford.edu/index.shtml) 实现.

2) 情绪词汇及其词性信息. 情绪词汇对于判断一条句子或一段文本的情绪类别起着非常重要的作用, 特别是对显性情感分析提供直接有效的信息. 我们可以仅根据情绪词汇本身就可以比较准确地识别出 “显性” 句子的情绪类别. 例如, 句子 1 “当她想着这事情的时候, 恐惧刺激着她的脚, 使她加快了步子.” 只要利用显式的情绪词 “恐惧” 就可以很容易判断出该句子的情绪类别是 [恐惧 (Fear)]. 句子 2 “ ‘这世界是多么美丽啊!’ 毛毛虫说.” 我们可以根据情感词 “美丽” 和附加语 “啊” 来判断该句子的情绪类别是 [喜悦 (Joy)]. 情绪词汇库的建立在第 4.1 节中有详细描述.

3) 基于依存关系的词语搭配特征. 词语搭配是一种具有一定结构的, 词语共现达到统计意义上的显著程度的词语组合. 本文考虑基于依存关系的词语搭配作为特征. 我们认为发生情绪变化的词语之间存在某种依存关系, 而这类依存关系对于情绪类别信息的标注具有重要的作用, 尤其是隐性情绪类别的判别. 比如否定词 “不”、“没有”、“无” 等使得

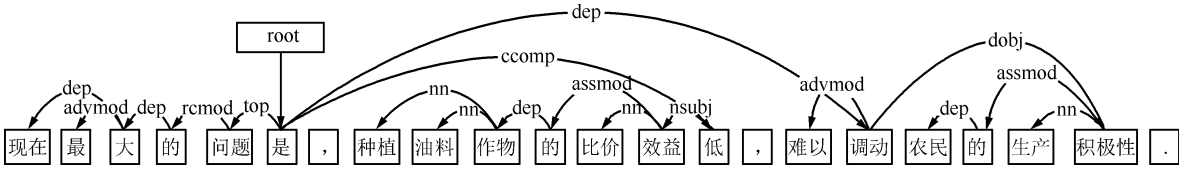

图 1 一条例句的依存句法分析结果之图示

Fig. 1 The dependency parsing result of an example in our experimental corpora

与它们有依存关系的情绪词汇的情绪类别发生改变. 当引入这层依存关系的词语搭配作为特征时, 就无需考虑专门人工设置这类规则的问题. 但是, 这类结构比较宠大, 需要更多的训练时间, 而且过多的单词会引入相应的噪声. 鉴于此, 本文探索一种骨架依存关系[28]. 这种骨架依存关系能够保存原有依存句法树中的谓词和与谓词有直接相关联的单词信息, 以及情绪词汇和与情绪词汇有直接相关联的单词信息, 这部分信息对于分类器具有更大的区分作用. 表 1 给出了图 1 所示例子中存在的所有依存关系以及提取的骨架依存关系的示例.

表 1 图 1 所示例子中所有依存关系和骨架依存关系的词语搭配

Table 1 All dependency relationship and skeleton dependency relationship of the example in Fig. 1

依存关系	骨架依存关系
dep (大-4, 现在-2)	root (ROOT-0, 是-7)
advmod (大-4, 最-3)	top (是-7, 问题-6)
dep (的-5, 大-4)	ccomp (是-7, 低-15)
rcmod (问题-6, 的-5)	nsubj (低-15, 效益-14)
top (是-7, 问题-6)	dep (是-7, 调动-18)
root (ROOT-0, 是-7)	advmod (调动-18, 难以-17)
nn (作物-11, 种植-9)	dobj (调动-18, 积极性-22)
nn (作物-11, 油料-10)	
dep (的-12, 作物-11)	
assmod (效益-14, 的-12)	
nn (效益-14, 比价-13)	
nsubj (低-15, 效益-14)	
ccomp (是-7, 低-15)	
advmod (调动-18, 难以-17)	
dep (是-7, 调动-18)	
dep (的-20, 农民-19)	
assmod (积极性-22, 的-20)	
nn (积极性-22, 生产-21)	
dobj (调动-18, 积极性-22)	

2.2 基于组合语义的深度特征

根据 Montague (1974) 的阐述, 组合语义 (Compositional semantics) 的意义在于一个复合表达式是由构成它的每个成分及其将这些成分组合在一起的句法规则所描述. 我们认为通过组合语义不仅可以扩充同义词 (或近义词) 而且也可以扩充同义 (或近义) 短语. 基于 "语义相似的语言单位往往具有相同的情绪趋向" 的认知, 同义 (或近义) 词或者短语具有相同的语义倾向, 表达相同的情绪. 因此, 基于组合语义的深度特征可以有效表达词语搭配和语义韵特征, 对隐含情绪分析有一定的作用.

本节给出一种基于语义空间模型 (Distributional semantic models, DSM) 的词语表示和组合语义计算方法. 具体计算步骤概括如下:

步骤 1. 首先我们对语料库进行预处理, 包括分词、词性标注以及依存句法分析. 汉语的分词和词性标注利用 Stanford 的分词和词性标注工具包实现, 依存句法分析采用 Stanford Dependency Parsing 句法分析器实现.

步骤 2. 选择目标词语集 V. 这是构建语义空间模型的第一步. 本文从大量的语料库中选择目标词语. 利用已预处理的汉语语料进行词语统计, 去掉停用词, 并根据情绪词典给出其初始的情绪类别信息. 例如 (消沉 |Sad; 享受 |Joy; 仁爱 |Joy) 等.

步骤 3. 选择上下文信息集 C. 构建语义空间模型的下一步, 也是比较关键的一步是选择一个适当的上下文信息. 上下文信息可以是一个文档、一条句子、围绕目标词语一定窗口范围内 (比如$-2, +2$, $-5, +5$, $-10, 10$ 等) 的词语集或某个句法路径上的词语集. 不同的上下文信息会产生不同的 DSM. 本文利用大量已经过依存句法分析的汉语语料库选择 2-词或 3-词窗口 (2-word or 3-word window) 的词语集作为上下文信息, 且满足: ①上下文信息必须是实词 (i.e., 名词、动词、形容词等); ②上下文信息不能是停用词. 我们认为语料库的高频词都可能显现杂合的情绪趋向. 以单个词作为研究单位观察其搭配词数据和词语索引很难发现语义特征规律. 因此, 本文以与这些高频词搭配的短语和复合词作为研究单位, 进一步进行词语搭配统计. 比如 "生活" 这个词与 "条件" 产生一定的搭配关系. 那么, 我们继续以 "生活条件" 为单位, 进一步统计会发现与 "生活条件" 产生主要搭配的词语有 "不满"、"差"、"提高" 等.

步骤 3. 建立词–上下文信息的关联矩阵 $F : V \times C$. 我们采用均值为零的高斯分布对该矩阵进行初始化, 即 $x \sim \mathrm{N}(0, \sigma^2)$. 再利用神经概率语言模型 (Neural language model, NLM) 的学习方法在大规模语料下进行进一步更新. 具体的实现过程可参照 Turian 等[29] 的研究工作.

步骤 5. 根据构建的关联矩阵, 建立词语向量查询表 LT_v, 将每个词语向量 $\phi_w(w_i) \in \mathbf{R}^n$ 投射到 d 维空间: $LT_v(i) = \phi_w(w_i)$. 其中 w_i 是第 i 列对应的元素.

步骤 6. 组合语义计算. 目前, 绝大多数的方法是以 Mitchell 等[20] 提出的方程式 $p = f(u, v, R, K)$ 为基础, 采用向量加、点积、张量积或是非线性的计算方法. 文献 [21−22] 等的研究工作已经验证了非线性函数的计算方法对短语相似度计算和短语识别非常有效. 鉴于此, 本文同样借鉴非线性函数的计算方法 (计算公式如式 (1)). 但不同于他们的工作之处在于, 我们并不被限制在考虑指定的某些特定的短

语, 比如形–名短语, 而是不特定长度的某种依存关系的短语 (Segment). 比如图 1 所示例子中, 我们选择不同长度的短语集 “现在最大的问题是”、“种植油料作物的比价效益低”、“难以调动农民的生产积极性”、“调动农民的生产积极性”、“是效益底” 等.

$$\phi(s) = f(u, v, R) = g(W(u||v) + b) \tag{1}$$

式中 s 是短语切分, $\phi(s)$ 表示 s 的组合语义表示; u, v 是词语向量; R 表示某种依存句法关系, 比如 mod; W 是权值矩阵; b 是偏移量. 函数 g 的表达形式采用双曲正切函数 (tanh).

本文采用 Recursive autoencoders 最小化目标函数, 使得输入向量 u_i, v_i 和它们的重构 u'_i, v'_i 之间的差距最小, $i \in (0, \cdots, N)$, 从而优化组合语义计算过程.

$$e_i = f^{enc}(u_i, v_i, R) = W^{enc}(u_i||v_i) + b^{enc} \tag{2}$$

$$e'_i = f^{rec}(e_i) = W^{rec}e_i + b^{rec} \tag{3}$$

$$E = \frac{1}{2}\sum_{i}^{N} \|e'_i - e_i\|^2 \tag{4}$$

式中 $u, v \in \mathbf{R}^n$, $W^{enc} \in \mathbf{R}^{m\times n}$, $W^{rec} \in \mathbf{R}^{n\times m}$, $b^{enc} \in \mathbf{R}^m$, $b^{rec} \in \mathbf{R}^n$. 图 2 显示了一个简单的计算例子.

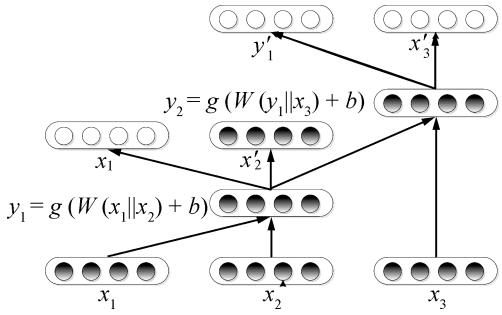

图 2 通过 Recursive autoencoders 的组合语义计算过程示意[27]

Fig. 2 Implementation of compositional semantics process using recursive autoencoders[27]

3 基于短语的文本情绪分类模型

本文提出一种基于短语 (Segment) 的文本情绪分析模型. 该模型主要由以下 3 个部分组成: 首先对输入文本进行依存句法分析, 并根据该结果将文本 (短文本或句子) 切分为若干短语; 然后对短语的情绪类别信息进行标注; 最后通过将它们的情绪类别信息进行组合来最终确定整个文本 (句子或短文本) 的情绪类别. 图 3 显示了文本情绪分类模型示意.

本节首先介绍如何获取短语切分, 接下来给出基于短语的文本情绪分类模型的数学模型、参数学习和解码过程.

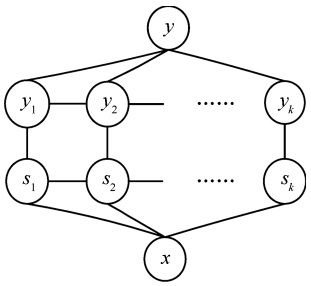

图 3 文本情绪分类模型示意

Fig. 3 Illustration of emotion detection model

3.1 短语切分依据

本文首先需要将已分词的一条句子或一段文本切分为几个短语. 短语是由两个或两个以上的词语组合构成的. 在计算语言学中, 短语可以是具有一定句法结构的词语组, 比如 “伤心的孩子” 是修饰关系的短语, 前边的形容词修饰后边的名词, 也可以是不具有任何关系的连续词语组, 比如图 1 例子中的 “底, 难以”.

短语识别及获取研究多是针对具有一定句法关系的词语组展开. 本文依据依存句法将输入文本切分为短语. 之所以选择依存句法, 其原因在于依存句法结构能够直接发现句子中词语之间的情感传递和情感组合关系. 例如, 图 1 给出了本文实验语料中出现的一条例句 “现在最大的问题是, 种植油料作物的比价效益低, 难以调动农民的生产积极性.” 的依存句法分析结果之图示. 易见, 通过 “效益” 和 “低”、“难以”、“调动” 和 “积极性” 这样的依存短语是可以判断该例句所包含的情绪类别信息.“低” 与 “效益” 共现时往往产生一种消极的、悲伤的情绪; “积极性” 体现积极的语义特征, 表现出 “喜悦” 的情绪, 但是与 “难以” 共现时将改变其情感趋向.

基于上述分析, 本文的短语切分过程包括: 首先利用 Stanford 句法分析器 (http://nlp.stanford.edu/software/ lex-parse.shtml) 对每个输入文本进行依存句法分析, 对生成的依存句法树结构抽取每个单词间的上下层链接结构, 生成一种单词依存树结构; 然后根据 Root 节点的谓词将两个上下层链接结构分别对应的单词对依存树结构进行组合, 生成短语切分. 比如, 图 1 的例子中, 我们得到的短语切分是 “现在最大的问题是”, “是, 种植油料作物的比价效益低”, 以及 “是 $\cdots$ 难以调动农民的生产积极性” 等.

3.2 基于短语的文本情绪分类模型

本文将文本情绪分析问题转化为以短语为主要线索的序列标注问题. 这样, 文本情绪分类系统的基本任务是对一个输入的文本 (比如句子或短文本) 进行 "愤怒、厌恶、喜悦、恐惧、悲伤、惊讶" 等情绪类别的标注.

假设 $x=\{x_1,x_2,\cdots,x_n\}$ 表示输入文本, 由 n 个词语组成; y 表示 x 的情绪标注类别, 且 $y\in\sum$, 其中 $\sum=$ {愤怒、厌恶、喜悦、恐惧、悲伤、惊讶}; 引入一个隐含变量 $s=\{s_1,\cdots,s_K\}$ 表示 x 的短语切分, 代表由 K 个短语切分组成; 令 $y_s=\{y_{s_1},\cdots,y_{s_K}\}$ 分别对应 s_1 到 s_K 个短语切分的情绪标注信息, 且 $y_{s_i}\in\sum, 1\le i\le K$. 那么, 文本情绪分析建模的基本任务就是要确定在接受文本为 x 的条件下对应的情绪类别标签为 y 的后验概率:

$$p(y|x)=\sum_{s,y_s}p(s|x)\cdot p(y_s|s,x)\cdot p(y|y_s,s,x)=\sum_{s,y_s}\prod_{k=1}^{K}p(y_{s_k},s_k|x)\cdot p(y|y_{s_k}) \tag{5}$$

显而易见, 式 (5) 可以分解为 2 个因式: $p(y_s,s|x)$ 和 $p(y|y_s)$. 我们定义每个因式对应一个子模型. 第一个子模型 $p(y_s,s|x)$ 定义为短语情绪分类模型, 它分析每个短语切分 (e.g, 从生成的句法子树中得到) 的情绪类别. 第二个子模型 $p(y|y_s)$ 对应短语切分的情绪类别到整个文本的情绪类别的转换过程. 接下来我们将讨论如何对这两个子模型进行建模.

本文采用半马尔科夫条件随机场 (Semi-CRFs) 作为短语情绪分类的基本框架, 对第一个子模型进行建模. 采用 Semi-CRFs 模型来判断短语的情绪类别信息的优点在于可以适用于引入表达基于词语搭配的短语特征, 且不需要考虑这些特征之间是否独立.

Semi-CRFs 是在半马尔科夫链 (Semi-Markov chain) 的基础上定义的. 半马尔科夫链是马尔科夫与非马尔科夫过程相结合的一种特殊的离散随机过程. 假设第 i 时刻的状态为 x_i, 该状态在一段时间 d_i 之内保持不变, 系统的下一个状态为 x', 该状态完全由当前状态决定. 而在 i 到 $i+d_i$ 的时间内系统的行为是非马尔科夫的.

根据 Sarawagi 等[30] 的定义, 半马尔科夫条件随机场描述如下. 假设 $X=x_t$, $Y=y_t$ $(t=1,\cdots,T)$ 分别表示需要标记的观察序列和它相应的标记序列的分布随机变量. x 由一系列切分集 $s=\{s_1,s_2,\cdots,s_p\}$ 构成, x 的某一个切分 s_i $(1\le i\le p)$ 被定义为一个三元组 $s_i=\langle t_i,u_i,y_i\rangle$: t_i 为起始位置, u_i 是终点位置, y_i 是 s_i 的标记. 例如, 图 1 例子中的有效切分包括 $s=\langle$(2, 7, 悲伤), (7, 15, 悲伤), (17, 22, 悲伤)$\rangle$ 等. 那么, Semi-CRFs(x,y) 就是一个以观察序列 x 为条件的无向图模型. 在给定观察序列 x 的条件下, x 的切分序列 s 和标记序列 y 的概率分布为

$$p(y,s|x)=\prod_{c\in S}\psi_c(x,y_c)=\frac{1}{Z(x)}\exp(-E(x,s,y)) \tag{6}$$

式中势函数 $\psi_c(x,y_c)$ 和能量函数 $E(x,s,y)$ 的计算方程式为

$$\psi_c(x,y_c)=\exp\left\{\sum_{k=1}^{K}\lambda_k f_k(x_c,y_c)\right\} \tag{7}$$

$$E(x,s,y)=-\sum_{c\in S}\sum_{k=1}^{K}\lambda_k f_k(x_c,y_c) \tag{8}$$

其中, $Z(x)=\sum_{(s',y')}\prod_{c\in s'}\psi_c(x,y'_c)$ 是归一化因子, f 是特征函数, λ_k 是 f_k 对应的权重.

第二个子模型 $p(y|y_s)$ 的目的是通过对短语的情绪类别信息进行组合来最终确定整个文本 (句子或短文本) 的情绪类别. 为此, 本文设置了几个情感传递规则: 如果两个短语情绪类别相同, 则它们的组合情绪类别也与它们相同, 例如 ⟨ 喜悦 ⟩+⟨ 喜悦 ⟩→⟨ 喜悦 ⟩; 如果两个短语情绪类别中某一个短语的情绪类别信息为中性, 即它没有任何情绪, 则两个短语的组合情绪类别为另一个短语的情绪类别, 例如 ⟨ 中性 ⟩+⟨ 喜悦 ⟩ →⟨ 喜悦 ⟩; 如果多种情绪类别信息的组合, 例如 ⟨ 喜悦, 喜悦, 悲伤 ⟩ 的情况下, 本文将采用类似 N-gram 语言模型的计算方式来获取.

3.3 参数估计与解码过程

参数估计问题即在给定训练集 $D=\{x^{(i)},y^{(i)}\}_{i=1}^{N}$ 的条件下, 计算特征权重 $\theta=\{\lambda_k\}$, 使得 $p(y|x,\theta)$ 最大. 本文采用对数似然估计计算:

$$L(\theta)=\sum_{i=1}^{n}\log p(y_i|x_i,\theta)-\frac{1}{2\sigma^2}||\theta||^2 \tag{9}$$

利用 L-BFGS (Limited memory BFGS) 算法可对参数训练过程进行进一步的优化.

需要说明的一点是基于组合语义的模型需要对切分后的结果进行形式化表示. 因此需要考虑两个参数, 即 $\theta=\{w_k,\delta_k\}$. 如图 4 所示. 每个输出单元 $-E_c(x,s_c,y_c)$ 在最后的隐含层对应一个特征函数 $f_c(x_c,y_c,w)$, 该特征函数对应的权重为 δ. 公式表示

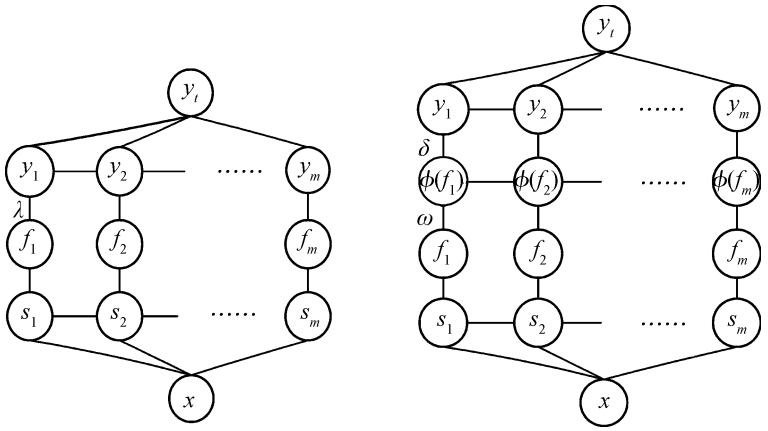

图 4 左图和右图分别表示基于短语的文本情绪分析模型和基于组合语义的文本情绪分析模型

Fig. 4 The graphs show the segment-based and the compositional semantics based emotion analysis model respectively

如下:

$$p(y,s|x) \propto \prod_{c\in S} \exp(-E_c(x,s_c,y_c)) = \prod_{c\in S} \exp(\langle \delta, f_c(x_c,y_c,w)\rangle) \tag{10}$$

解码 (Decoding) 过程, 即根据已知训练模型, 对未知变量的解释或推理. 在本文中, 即根据已知训练模型, 获取文本的情绪类别. 本文采用 Viterbi 算法实现整个解码过程. 那么, 给定一个具体模型, 最佳的标记是:

$$\widehat{y} = \arg\max_{y,s}\{p(y|y_s)\cdot p(y_s,s|x)\} \tag{11}$$

4 实验设计及结果分析

4.1 实验语料准备

目前尚未发现公开的中文情绪标注语料库. 因此, 我们首先需要构建训练和测试用的情绪标注语料库. 本文共构建了四种语料库, 包括情绪词汇词典、新闻领域情绪标注语料库、儿童故事集标注语料库以及未标注的语料库. 见表 2: 本文使用的语料库.

表 2 本文使用的语料库

Table 2 Emotion corpus used in this paper

	中文情绪词典	儿童故事集
词条/句条	1 810	1 223
	新闻领域语料	未标注语料
词条/句条	1 135	115 M

1) 中文情绪词汇库. 本文利用 WordNet Affect Lexicon (http://wndomain.fbk.eu/wnaffect.html) 中收集的英文情绪词汇, 首先对英文词典中的词汇进行大小写转换、词汇符号化 (Tokenization)、去重等操作; 然后利用中英双语词典进行翻译, 选取候选的中文情绪词汇; 之后, 基于 "语义相似的词汇往往具有相同的情绪趋向" 的认知, 对选取的候选中文情绪词汇进行语义相似度计算; 最后对候选的中文情绪词汇进行情绪标注. WordNet 英文情感词典标注了 "喜悦、愤怒、厌恶、悲伤、恐惧、惊慌" 等六种情绪类别. 同样, 我们也标注了这六种基本情绪.

2) 新闻领域情绪标注语料. 该情绪语料库的构建参照了文献 [3] 和文献 [6] 的工作. 其标注流程为: 首先由两名标注者分别对抽取的文本进行独立标注. 每个标注者将每条句子标注为 "喜悦、愤怒、悲哀、恐惧、惊讶、厌恶" 等六个值的一种或两种, 且每条句子最多可以被标注为两种情绪; 然后通过 Kappa 值来计算互标注者的一致性 (The inter-annotator agreement). 不同情绪类别的平均成对一致性 (The average pair-wise agreement) 的范围约在 0.51 ∼ 0.8 之间. 在实验语料的选择上, 我们选择了一致性大于等于 0.7 的句子. 依照上述过程, 本文共构建了 1 135 条句子, 平均句长为 27.09 词.

3) 儿童故事集. 该情绪语料库的构建基于 Alm 的英文情绪标注语料库 (http://people.rc.rit.edu/∼coagla/). Alm 的情绪标注语料来源于儿童故事集, 取自安徒生童话故事和格林童话故事等. 我们构建该语料库的流程为: 首先从 Alm 的情绪标注语料中选取情感高一致性 (Affective high agreement) 的句子; 然后搜集并整理对齐的英文 – 中文故事集; 之后利用句子对齐技术对其进行对齐, 选取每条英文句子对应的中文句子; 每条中文句子的情绪标注与英文的保持一致. Alm 的情绪标注语料只对 "喜悦、愤怒或厌恶、悲伤、恐惧、惊讶" 等五种情绪进行了标注. 为了与之前构建的情绪词典以及新闻领域语料中情绪标注信息保持一致, 本文将 "愤怒" 和 "厌恶" 进行分开标

注. 该部分的标注由人工完成. 其余的参照 Alm 的标注结果. 目前共整理了 1 223 条句子, 平均句子长度为 34.76 词.

4) 未标注语料. 该语料主要用于计算词语分布及组合语义计算, 使用的单语语料约 115 M, 57 万词语级, 领域包括新闻、儿童故事、博客等.

4.2 实验设计及分析

为了验证本文提出特征和模型的有效性, 在第 4.1 节构造的两组语料: 儿童故事集和新闻领域语料上进行情绪标注的对比实验. 采用的评价指标为准确率 ACC (Accuracy).

4.2.1 对本文提出特征和模型的评价

首先验证本文提出模型的有效性以及不同特征对模型产生的影响. 实验比较了 4 种统计计算模型: SVM (http://www.csie.ntu.edu.tw/∼cjlin/libsvm/)、ME (http://mallet.cs.umass.edu/)、CRFs (http://crf.sourceforge.net/) 以及本文提出的一种基于 Semi-CRFs 的模型 (Yet another semi-CRFs, 简记为 YAsemiCRFs). SVM 是目前文本情绪分类采用的主要分类器[6,10], 作为基准 (Baseline) 系统. 而实验中采用的其他三个判别式模型 (ME, CRFs, Semi-CRFs) 能够对序列输入上下文相关标注的特点. 同时为了验证不同特征对统计模型产生的影响, 在实验中针对本文设计的几类特征: 基本词汇 (BOW)、功能词汇 (ContentBOW)、情绪词汇 (Emotion)、词性标注信息 (POS) 以及依存词语搭配特征 (Dependency) 等分别作了相应的实验. 本组对比实验采用 10-fold 交叉验证的方法进行. 即将数据集分成十份, 轮流将其中 9 份作为训练集 1 份作为测试集, 10 次结果的均值作为对算法的估计. 对比实验结果如表 3 所示.

从以下三点对表 3 的实验结果进行分析: 1) 不同特征的对比. 从实验结果可以看出, ContentBOW 特征的结果高于 BOW 特征. 以 ContentBOW 特征的引入作为基准 (Baseline) 系统, 分别加入词性特征 (POS)、情绪词汇特征 (Emotion)、基于依存关系的词语搭配特征 (Dependency). 从结果可以看出这几类特征的加入从不同角度提升了系统效果. 比如在儿童故事集上, Emotion 特征的加入使得 SVM 分类器的准确率相比 Baseline 系统提高了 5.88 %, ME 的准确率提高了 1.67 %, CRFs 提高了 3.57 %, YAsemi-CRFs 的效果提高了 4.38 %, 从而验证了我们的预期, 即情绪词汇对于判断一条句子或一段文本的情绪类别起着重要的作用. 当加入 Dependency 特征时, 各分类器的准确率相比基准系统均提高不少, SVM 提高了 8.25 %, ME 提高了 4.46 %, CRFs 提高了 9.81 %, YAsemiCRFs 的效果提高了 8.7 %. 说明基于依存关系的词语搭配特征对判断文本的情绪类别信息非常有效. 但是, POS 特征的加入产生了一些不同的效果. 比如, SVM、CRFs 和 YAsemiCRFs 的准确率相比基准系统均有小幅度提高, 但是 ME 的准确率却下降了 1.77 %. 同样, 在 ContentBOW + Emotion 的组合特征上加入 POS 特征时, ME 分类器的效果也下降了 1.28 %. POS 特征产生这样的结果, 这一方面可能是因为 ME 存在标记偏置的问题. 另外一方面也可能是因为 ContentBOW 特征本身已经考虑了形容词、动词、名词等功能实词, 使得 POS 特征的加入未产生准确率提高的效果.

组合语义的深度特征, 其实际意义在于对基于表面词形特征的深度表示. 组合语义特征的提取是在 100 维的语义空间模型下进行的. 基于依存句法

表 3 基于基本特征的各种分类器对比实验结果 (%)

Table 3 Comparison of experimental results of various classification based on surface word form features (%)

	儿童故事集				新闻领域语料			
	SVM	ME	CRFs	YAsemiCRFs	SVM	ME	CRFs	YAsemiCRFs
BOW	39.59	40.30	35.79	40.09	46.84	46.1	53.29	53.33
ContentBOW	39.98	40.59	35.87	42.11	48.59	47.8	55	56.74
ContentBOW + POS	40.19	38.82	36.05	43.95	48.64	47.67	56.78	57.69
ContentBOW + Emotion	45.86	42.26	39.44	46.49	51.46	50.7	57.41	59.55
ContentBOW + Emotion+ POS	46.15	40.98	41.89	48.95	50.2	47.1	58.53	61.53
ContentBOW + Emotion + POS + Dependency	48.23	45.05	45.68	50.81	54.45	54.32	59.06	65.12
Compositional semantics	48.56	–	–	52.19	55.37	–	–	65.3

的短语被切分后, 通过词语向量的非线性组合将其转化为组合语义表达, 并作为 SVM 和 YAsemiCRFs 的输入. 相比于基于表面词形的组合特征 Content BOW + Emotion + POS + Dependency, 组合语义的深度特征得到了最高的准确率. 但是, 提高的幅度不是很明显. 比如, 在儿童故事集上, 基于组合语义的结果相比基于依存短语的组合特征提高了 1.31 %, 而在新闻领域的数据集上, 基于组合语义的结果相比基于依存短语的结果仅有 0.14 % 的提高. 造成这个结果的原因, 可能与所使用语料的领域特征有关. 例如, 从受众易于接受的角度考虑, 新闻领域的语料通常比较简洁, 情绪表达也比较直接. 统计结果表明, 本文采用的新闻领域语料的平均句长仅为 27.09 词. 此外, 复杂的情绪表达手法也鲜有利用. 因此, 组合语义的方法在这样的语料上体现不出其优势.

2) 不同分类器效果对比. 本组实验以 SVM 作为基准系统. 综合两组实验语料的效果, 本文提出的 YAsemiCRFs 模型在两种风格差异较大的数据集上都能取得稳定的性能优势, 准确率均优于基准系统. 而其他模型, 比如 ME 的综合表现均明显劣于 SVM. 而 CRFs 模型在儿童故事集上的综合表现不如 SVM 的效果好, 但是在新闻领域语料上的效果又比 SVM 的效果好. 说明 CRFs 模型对于短文本的处理存在一定的不确定因素, 另外也说明分类模型也是受语料库和领域的影响. 在不同特征的引入上所表现的能力也有所不同. 比如 SVM 在采用 ContentBOW + Emotion + POS 的组合特征时其准确率要高于 ME 和 CRFs, 但是应用 ContentBOW + Emotion + POS + Dependency 的组合特征时其效果又低于 ME 和 CRFs. 可能的原因是 SVM 对短语级特征融合能力有一定的局限性. 另外, 我们增加了统计显著性实验. 假设样本服从正态分布, 我们利用来自同一数据集上用于交叉验证的 10 组实验数据做了 T-test, 对 YAsemiCRFs 模型与其他三种模型 (SVM, ME, CRFs) 分别进行了对比, 每种对比进行了 7 次实验. 在大多数情况下, 我们得到的 P 值范围在 8.24E−4∼0.0345 之间. 由此说明我们的实验结果具有统计显著性. 基于 YAsemiCRFs 模型的结果优于其他模型. 综上, YAsemiCRFs 模型有效提高了系统的准确率, 在一定程度上弥补了现阶段判别式模型对词语搭配等短语特征应用上的缺陷, 比较适合文本情绪分析的研究.

3) 不同语料的对比. 儿童故事集的总体效果稍逊于新闻领域语料的效果. 产生这个结果的原因可能在于儿童故事集很多是短文本 (Short text), 相比新闻领域语料, 其句子长度要长一些, 且句子结构或语句比较复杂, 情绪表达不明显.

接着进一步验证本文提出组合语义特征的有效性. 为此, 我们依据第 2.2 节中的组合语义深度特征的生成算法步骤, 生成 100 维 (100D) 的语义空间模型, 并采用基于线性的组合计算方法和基于非线性的组合计算方法提取出组合语义特征, 然后将其加入到 YAsemiCRFs 模型中, 以 ContentBOW 特征 (T1) 的引入作为基准系统, 以引入组合特征 ContentBOW + Emotion + POS + Dependency (T2) 的系统为对比系统进行对比实验. 线性组合方法采用 $f(x,y)=x+y$ (T3) 和 $f(x,y)=xy$ (T4); 非线性组合计算采用 $f(x,y)=g(W(x||y)+b)$. 其中, $x+y$ 表示两个词汇向量的加; xy 表示两个词汇向量的积; $g(W(x||y)+b)$ 表示基于依存句法树切分的非连续词汇向量之间的非线性计算. 针对非线性组合计算, 我们分别训练了两组: $g(W(x||y)+b)-I$ (T5) 和 $g(W(x||y)+b)-II$ (T6). 其中, $g(W(x||y)+b)-I$ 完全遵循第 2.2 节中的组合语义特征的生成算法步骤, 而 $g(W(x||y)+b)-II$ 的训练过程中有一定的人工干预. 在参数学习过程中以提高准确率为目的, 改变其参数值和迭代次数. 本组对比实验仍采用 10-fold 交叉验证的方法进行. 实验结果如图 5 所示.

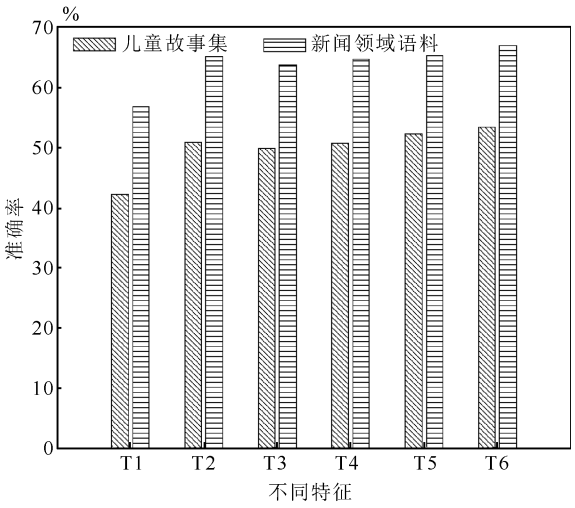

图 5 基于组合语义的特征对比实验

Fig. 5 Comparison of experimental results of various different features based on compositional semantics

从图 5 的实验结果可以观察到, 引入依存词语搭配特征 (T2) 的系统和基于组合语义深度特征 (T3∼T6) 的系统所得到的准确率均高于基准 (T1) 系统的准确率. 基于线性组合特征 (T3∼T4) 的系统准确率和基于特征 T2 的系统准确率相差不大, 而基于非线性的组合语义特征 (T5∼T6) 的系统准确率不仅高于引入特征 T2 的系统准确率, 也高于引

入基于线性组合特征 (T3 ∼ T4) 的系统准确率. 具体而言, 在本文提出的 YAsemiCRFs 模型中引入基于非线性计算方法 ($g(W(x||y)+b)-II$) 得到的组合语义特征在新闻领域测试集上取得了最好的成绩 66.89 %, 相比基准系统提高了 10.15 %. 相比引入依存词语搭配特征的系统提高了 7.7 %. 在儿童故事集上取得了最好的成绩 53.38 %, 相比 Baseline 系统提高了 17.51 %. 相比引入依存词语搭配特征的系统提高了 1.77 %. 上述实验结果说明基于依存关系的词语搭配特征和组合语义特征对文本情绪分类系统准确率的提高都有一定的效果. 基于非线性的组合语义表示方法相比基于线性的组合方式更有效, 而有针对性的训练 ($g(W(x||y)+b)-II$) 能取得更好的效果. 但是, 基于非线性的组合语义计算过程比线性组合计算复杂, 且模型参数训练必须使用神经网络 (Neural network) 进行, 其复杂度较高. 这将是我们下一步工作中考虑的主要问题. 上述实验结果同时也说明本文提出的 YAsemiCRFs 模型可以有效利用组合语义特征.

4.2.2 对隐性文本情绪分析的评价

我们认为凡是不包含任何情绪词汇的句子属于 "隐性" 情绪表达的句子. 进一步, 如果某条句子中包含了某个情绪词汇, 但是该情绪词汇的初始情绪类别发生了改变, 那么这条句子也属于 "隐性" 情绪表达的句子. 我们通过对本文的两组实验语料进行统计, 发现不包含任何情绪词典中关键词的句子或短文本约占 35 %; 包含情绪词典中关键词, 但是其情绪类别与其初始的情绪类别不一致 (或发生改变) 的句子或短文本约占 40.9 %. 这说明本文构建的情绪标注语料中 "隐性" 句子所占比例是不容忽视的, 而这些 "隐性" 句子的分类准确率对整个系统性能有一定影响.

为了验证本文提出模型和特征对隐性情绪分类的有效性, 分别从儿童故事集和新闻领域语料中选取了 105 条隐含情绪的句子或短文本作为测试集, 其余作为训练集进行测试. 对比模型采用 SVM 和 YAsemiCRFs, 语义空间模型是 100 维的. 实验结果如表 4 所示.

从表 4 的实验结果可以看出, 我们得到的准确率虽然不算高, 但是组合语义的结果在众多特征中取得了最优的效果, 说明组合语义的表示方式完全适用于隐性情绪分析任务, 可以有效提高隐性情绪分类的准确率, 从而验证了我们方案的可行性. 同时也注意到基于依存句法的词语搭配特征的加入也取得了不错的成绩. 说明基于依存句法的词语搭配特征对隐性情绪分析也是有重要作用的. 组合语义特征和基于依存句法的词语搭配特征之所以能取得不错的成绩, 其原因可能是因为它们和情绪之间存在较强的关联性. 尤其是对于隐含情绪的分析问题而言, 基于组合语义的特征表现出明显的优势, 其原因可能正在于其所包含的语义层面的信息和情绪之间存在的强关联性.

5 结论

针对以往文本情绪分类研究多着眼于表面词形特征的问题, 本文对词语搭配或语义韵等特征在文本情绪表达中的作用进行了探索, 提出了一种有效的基于组合语义的深度特征. 并在此基础上提出了一种基于半马尔科夫条件随机场的文本情绪分析模型, 有效提高了系统的分类性能. 为了验证方法的有效性, 构建了包括情绪词汇词典、新闻领域情绪标注语料库、儿童故事集标注语料库以及用于构建词语表示向量的未标注语料库等四种语料库, 开展了相

表 4 隐性情绪表达文本测试集对比实验 (%)

Table 4 Comparison of experimental results of implicit emotion detection (%)

	儿童故事集		新闻领域语料	
	SVM	YAsemiCRFs	SVM	YAsemiCRFs
BOW	32.38	38.33	33.33	40.09
ContentBOW	32.69	38.69	34.33	42.43
ContentBOW + POS	34.88	39.33	36.14	41.51
ContentBOW + Emotion	36.89	41.19	37.14	46.48
ContentBOW + Emotion + POS	36.98	42.14	38.01	45.11
ContentBOW + Emotion+ POS + Dependency	37.09	43.12	40.56	47.77
$x+y$	37.69	43.20	41.63	47.43
xy	37.02	43.86	41.66	47.86
$g(W(x\|\|y)+b)-I$	37.55	44.98	42.27	48.04
$g(W(x\|\|y)+b)-II$	–	47.33	–	50.95

关的实验研究. 实验结果表明, 本文提出的组合语义特征和半马尔科夫条件随机场文本情绪分析模型对于完成文本情绪分类任务具有很好的适用性.

在下一步的研究工作中, 一方面, 我们将进一步研究组合语义方法对于隐性情绪分类的影响; 另一方面, 我们将探索更多语言学理论对于提高文本情绪分类性能的作用. 此外, 扩建情绪标注语料库是另外一项重要的工作.

References

1 Lutz C, White G M. The anthropology of emotions. *Annual Review of Anthropology*, 1986, **15**: 405−436

2 Wei Nai-Xing. *Fundamentals of Lexis*. Shanghai: Shanghai Foreign Language Education Press, 2011.
(卫乃兴. 词语学要义. 上海: 上海外语教育出版社, 2011.)

3 Alm C O, Roth D, Sproat R. Emotions from text: machine learning for text-based emotion prediction. In: Proceedings of the 2005 Conference on Human Language Technology and Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005. 579−586

4 Aman S, Szpakowicz S. Using roget′s thesaurus for fine-grained emotion recognition. In: Proceedings of the 3rd International Joint Conference on Natural Language Processing. Hyderabad, India: Researchgate, 2008. 312−318

5 Das D, Bandyopadhyay S. Word to sentence level emotion tagging for Bengali blogs. In: Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009. 149−152

6 Nakagawa T, Inui K, Kurohashi S. Dependency tree-based sentiment classification using CRFs with hidden variables. In: Proceedings of Human Language Technologies: the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010. 786−794

7 Xu Bing, Zhao Tie-Jun, Wang Shan-Yu, Zheng De-Quan. Extraction of opinion targets based on shallow parsing features. *Acta Automatica Sinica*, 2011, **37**(10): 1241−1247
(徐冰, 赵铁军, 王山雨, 郑德权. 基于浅层句法特征的评价对象抽取研究. 自动化学报, 2011, **37**(10): 1241−1247)

8 Neviarouskaya A, Prendinger H, Ishizuka M. Compositionality principle in recognition of fine-grained emotions from text. In: Proceedings of International Conference on Weblogs and Social Media. San Jose, California: The AAAI Press, 2009. 6−24

9 Neviarouskaya A, Prendinger H, Ishizuka M. Recognition of affect, judgment, and appreciation in text. In: Proceedings of the 23rd International Conference on Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010. 806−814

10 Chaffar S, Inkpen D. Using a heterogeneous dataset for emotion analysis in text. In: Proceedings of the Canadian AI′11 Proceedings of the 24th Canadian conference on Advances in Artificial Intelligence. St. John′s, Canada: Springer, 2011. 62−67

11 Mohammad S, Turney P D. Emotions evoked by common words and phrases: using mechanical Turk to create an emotion lexicon. In: Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010. 26−34

12 Volkova S, Dolan W B, Wilson T. CLex: a lexicon for exploring color, concept and emotion associations in language. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012. 306−314

13 Purver M, Battersby S. Experimenting with distant supervision for emotion classification. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012. 482−491

14 Qadir A, Riloff E. Bootstrapped learning of emotion hashtags #hashtags4you. In: Proceeding of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. Stroudsburg, PA, USA: Association for Computational Linguistics, 2013. 2−11

15 Quan C Q, Ren F J. Construction of a blog emotion corpus for Chinese emotional expression analysis. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009. 1446−1454

16 Xu G, Meng X F, Wang H F. Build Chinese emotion lexicons using a graph-based algorithm and multiple resources. In: Proceedings of the 23rd International Conference on Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010. 1209−1217

17 Wang L, Yu S W. Construction of a Chinese idiom knowledge base and its applications. In: Proceedings of the Multiword Expressions: From Theory to Applications (MWE 2010). Beijing, China: Association for Computational Linguistics, 2010. 11−18

18 Yang M, Peng B L, Chen Z, Zhu D J, Chow K P. A topic model for building fine-grained domain-specific emotion lexicon. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2014. 421−426

19 Choi Y, Cardie C. Learning with compositional semantics as structural inference for subsentential sentiment analysis. In: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics, 2008. 793−801

20 Mitchell J, Lapata M. Vector-based models of semantic composition. In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2008. 236−244

21 Baroni M, Zamparelli R. Nouns are vectors, adjectives are matrices: representing adjective-noun constructions in semantic space. In: Proceedings of the 2010 Conference

on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010. 1183−1193

22 Socher R, Pennington J, Huang E H, Ng A Y, Manning C D. Semi-supervised recursive autoencoders for predicting sentiment distributions. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011. 151−161

23 Socher R, Huval B, Manning C D, Ng A Y. Semantic compositionality through recursive matrix-vector spaces. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012. 1201−1211

24 Moilanen K, Pulman S. Sentiment composition. In: Proceedings of Recent Advances in Natural Language Processing. Borovets, Bulgaria: University of Oxford, 2007. 378−382

25 Yessenalina A, Cardie C. Compositional matrix-space models for sentiment analysis. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011. 172−182

26 Mohammad S. Portable features for classifying emotional text. In: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012. 587−591

27 Hermann K M, Blunsom P. The role of syntax in vector space models of compositional semantics. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2013. 894−904

28 Sui Zhi-Fang, Yu Shi-Wen. The research on recognizing the predicate head of a Chinese simple sentence in EBMT. *Journal of Chinese Information Processing*, 1998, **12**(4): 39−46 (穗志方, 俞士汶. 面向 EBMT 的汉语单句谓语中心词识别研究. 中文信息学报, 1998, **12**(4): 39−46)

29 Turian J, Ratinov L, Bengio Y. Word representations: a simple and general method for semi-supervised learning. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010. 384−394

30 Sarawagi S, Cohen W W. Semi-Markov conditional random fields for information extraction. In: Proceedings of the Eighteenth Annual Conference on Neural Information Processing Systems. Montreal, Quebec, Canada, 2004.

乌达巴拉 中国科学技术大学自动化系博士研究生. 中国科学院合肥智能机械研究所助理研究员. 主要研究方向为自然语言处理、情感计算、人工智能与模式识别. 本文通信作者.
E-mail: hwdbl@126.com
(**Odbal** Ph. D. candidate at the Department of Automation, University of Science and Technology of China and assistant researcher at the Institute of Intelligent Machines, Chinese Academy of Sciences. Her research interest covers natural language processing, affective computing, artificial intelligence and pattern recognition. Corresponding author of this paper.)

汪增福 中国科学院合肥智能机械研究所研究员. 中国科学技术大学语音及语言信息处理国家工程实验室教授. 主要研究方向为立体视觉、生物特征识别、情感计算以及智能机器人.
E-mail: zfwang@ustc.edu.cn
(**WANG Zeng-Fu** Professor at the Institute of Intelligent Machines, Chinese Academy of Sciences and the National Engineering Laboratory for Speech and Language Information Processing, University of Science and Technology of China. His research interest covers stereo vision, biometrics, affective computing and intelligent robot.)