DOI:10.7544/issn1000-1239.20210537
59(11):2467-2496,2022

文本情感原因自动提取综述

邱祥庆[1,2,4] 刘德喜[1,3] 万常选[1,3] 李 静[1,2,4] 刘喜平[1,3] 廖国琼[1,3]
[1](江西财经大学信息管理学院 南昌 330013)
[2](福建江夏学院电子信息科学学院 福州 350108)
[3](数据与知识工程江西省高校重点实验室(江西财经大学) 南昌 330013)
[4](福建江夏学院数据分析与智能决策研究所 福州 350108)
(qiuxq_2020@163.com)

Survey on Automatic Emotion Cause Extraction from Texts

Qiu Xiangqing[1,2,4], Liu Dexi[1,3], Wan Changxuan[1,3], Li Jing[1,2,4], Liu Xiping[1,3], and Liao Guoqiong[1,3]
[1](*School of Information Management*, *Jiangxi University of Finance and Economics*, *Nanchang* 330013)
[2](*School of Electronic Information Science*, *Fujian Jiangxia University*, *Fuzhou* 350108)
[3](*Jiangxi Key Laboratory of Data and Knowledge Engineering* (*Jiangxi University of Finance and Economics*), *Nanchang* 330013)
[4](*Research Institute for Data Analysis and Intelligent Decision Making*, *Fujian Jiangxia University*, *Fuzhou* 350108)

Abstract Emotion cause extraction(ECE) is a new research direction of affective computing. As a kind of fine-grained sentiment analysis, its purpose is to find out the part of the given document text that triggers the emotion, which is also called tracing to the source of an emotion. ECE has a high academic research value and a wide range of application scenarios in reality, because of its involvement in the fields of linguistics, psychology and other related domains. Recognizing the emotion cause in a document is more useful than just only identifying the emotion. For example, it can help people understand the source of stress and better control their emotions. Although most of the research in affective computing focuses on emotion recognition, emotion prediction and emotional information extraction, many scholars have turned to analyze the causes behind emotions, and have produced rich results in recent years. We make a comprehensive review and analysis of automatic ECE from texts in multiple perspectives, starting from the problem definition and the classification of ECE. Then we review the main methods for ECE especially based on deep learning. After that, the benchmark datasets and the evaluation metrics for ECE task are detailed summarized. Finally, we discuss the deficiency of the existing works on ECE and forecast the future challenges.

收稿日期:2021-06-03;**修回日期**:2021-11-09

基金项目:国家自然科学基金项目(62272206,61972184,62076112);教育部人文社会科学研究项目(20YJC630229);福建省中青年教师教育科研项目(JAT170623);福建省科技厅引导性项目(2020H0029);福建省财政厅专项课题(闽财指[2020]822号);江西财经大学研究生创新专项项目;江西省主要学科学术和技术带头人培养计划领军人才项目(20213BCJL22041);江西省自然科学基金项目(20212ACB202002);福建江夏学院青年科研人才培育基金项目(JXS2016010)

This work was supported by the National Natural Science Foundation of China (62272206, 61972184, 62076112), the Humanities and Social Sciences Foundation of Ministry of Education of China (20YJC630229), the Education Scientific Research Project of Young Teachers of Fujian Province (JAT170623), the Pilot Project of Fujian Provincial Department of Science and Technology (2020H0029), the Specific Project of Fujian Provincial Department of Finance (mincaizhi[2020]822), the Graduate Innovation Project of Jiangxi University of Finance and Economics, the Leading Talent Project of Jiangxi Provincial Major Disciplines Academic and Technical Leaders Training Program (20213BCJL22041), the Natural Science Foundation of Jiangxi Province (20212ACB202002), and the Youth Scientific Research Talent Cultivation Project of Fujian Jiangxia University (JXS2016010).

通信作者:刘德喜(dexi.liu@163.com)

Key words sentiment analysis; emotion cause extraction; emotion cause pair; machine learning; deep learning

摘 要 情感原因提取是情感计算领域研究的一个新方向,是一种细粒度的情感分析,其目的是要找出给定文档中触发情感的那部分文本,是对情感的一种追根溯源.情感原因提取涉及到语言学、心理学等相关的领域知识,具有较高的学术研究价值和广泛的应用场景.尽管情感计算的相关研究大多集中在情感识别、情感预测、情感信息抽取等方面,但近些年不少学者已开始深入到情感背后的原因分析与提取上,并产生了较为丰富的成果.从问题定义、任务类别、研究方法、主流数据集、评测指标等多个角度对基于文本的情感原因自动提取的研究成果进行全面回顾和分析,重点对情感原因提取的方法特别是基于深度学习的方法进行了梳理,最后总结了现有情感原因提取工作的不足及其未来所面临的挑战.

关键词 情感分析;情感原因提取;情感原因对;机器学习;深度学习

中图法分类号 TP391

随着 WEB2.0 时代的到来,通过在线社交媒体来搜集用户观点变得越来越普遍,推特、脸书、微博等平台已成为与读者分享想法和观点的有效渠道.这些信息获取的便利性大大吸引了那些观点挖掘者,并被用于商务、社会、教育和娱乐等特定目的.在此背景下,情感分析作为自然语言处理中的一个研究领域越来越受到学术界和工业界的关注[1],该领域中的情感分类、情感检测、情感预测等也成为近年来研究的热点.

然而对文本进行情感分类、检测或预测只是一种浅层的分析,对政策制定者、社会管理或服务者、商业组织或企业来说,他们有时更关心主观文本所表达的某种情感背后更深层次的原因.下面通过 3 个示例分别从社会治理、餐饮住宿、心理健康这 3 个领域来说明情感原因提取任务的应用场景及其意义.

示例 1. "突然被告知公司裁员了……很不幸我中招了虽然没了工作但是有几万块的赔偿拿……也不知道是开心还是不开心……本命年果然有些不顺……希望疫情快快过去让我这个失业人士找到个满意的工作遇到可爱的同事暖心的领导求求惹?①"

示例 1 中的"不幸""不顺"等词表达的是负面的情感,对文本进行更深入的挖掘,可以得出该负面情感的原因是"公司裁员了",也就是说,对于该民众来讲失业是他最担心最害怕的事情.对此,政府可以从如何复工复产、扩大就业等角度入手来缓解民众的情绪.

新冠疫情给社会带来了巨大的冲击,很多人会通过网络来发表自己的感受,从这些在线社交文本中可以了解到民众的真实情感以及其产生的原因.情感原因提取(emotion cause extraction, ECE)可以为相关部门进行灾后民众心理健康辅导以及灾后重建政策的制定提供辅助.同理,在网络舆情方面,如果能迅速提取人们情感变化的原因,则可以在第一时间针对性地进行疏导,减少负面影响.此外,传统的舆情监控一般只在事件已经发生或者已造成不良影响后才可能监控到,但文本情感原因研究则可以基于对历史事件的分析,对其中触发民众异常情感的原因进行归纳推导,从而提前判断有可能产生异常的舆论事件.

示例 2. "房间的隔音太差!我定的房朝内街,对讲机声和空调声极大!建议加装双层玻璃窗!②."

示例 2 中,传统的情感分析可以识别出文档中表达的负面情感,通过方面级情感分析可进一步知道用户不满的对象是酒店的设施,但这还是不够的,如果能再进一步分析出是因为"隔音太差"和"对讲机声和空调声极大"造成对设施的不满,那么酒店的管理人员就能更清楚地知道为什么顾客不喜欢他们的酒店,而不是简单的喜好分类,从而有明确的改进方向.

示例 3. "为什么我的父母都不理解我,看到我手腕上的伤只会骂我打我,让我更绝望,我也好想死,但哪一次是成功的呢?③"

① 数据来源自 CCIR 2020 疫情期间网民情绪识别数据集

② 数据来源自谭松波等的酒店评论语料 ChnSentiCorp

③ 数据来源自中科院心理研究所微博自杀意念数据集

示例 3 中,文本反映出用户存在心理问题,有自杀的倾向,如果能把用户为什么有这种负面情绪或者说为什么想自杀的原因"父母都不理解我"提取出来,在进行心理辅导时就更有针对性.

民众的心理健康问题也日益突出,很多患者由于某些主观原因并不愿意通过传统途径来获取帮助,他们反而愿意通过在线社交媒体来寻求支持.因此,在通过帖子发现心理健康异常或情绪异常的用户后,提取帖子中反映出的"情感原因",可以提高疏导的针对性,改善心理辅导效果.

以上示例显示,情感原因的自动提取具有广泛的应用场景.同时,情感原因的自动提取需要能够对文本语义和情感表达有更深刻的理解,需要探索更有效的自然语言处理技术,甚至需要利用心理和认知领域的知识,因此具有较高的学术研究价值.

近年来已有不少学者关注情感原因提取这一重要方向.目前从事情感原因提取研究的团队主要集中在国内[2-32],其中 Lee 等人[2]在 2009 年开始从事情感原因提取的相关研究,Xu 等人[14]做了大量基础性工作并发布了目前唯一公开的中文情感原因数据集,Xia 等人[25]于 2019 年提出情感原因对提取任务,为情感原因提取研究打开了一个新的方向.近 3 年,该领域的研究更是处于一个明显的上升势头.

虽然在情感原因提取这一领域目前产生了不少很好的成果,但并没有相关的综述报道,因此,本文对基于文本的情感原因自动提取的相关成果进行全面回顾和分析,梳理情感原因自动提取的主要方法和模型,廓清该领域的发展状况与趋势,展望未来发展方向,旨在为情感分析工作的深入研究提供参考.

1 问题定义与分类

情感原因简单理解就是导致或诱发某种情感产生的直接或间接原因,而文本情感原因提取则是从蕴含情感的文本中提取出描述情感产生原因的事件、子句、短语或词.不同学者从不同的学科和角度出发,对情感原因的理解也不尽相同.考虑到一些文献中将"情感"和"情绪"视为 2 种完全不同的概念,本文中所指的"情感"是更为广义的概念,它既包括普通的正向、负向和中性的情感含义,也包括高兴、生气、害怕等具体层面的情绪含义.

Lee 等人[2]于 2009 年开始开展文本情感原因提取的研究.他们基于心理学领域关于情感的理论研究成果[33],将情感原因看成是一种明确表达的、引发相应情感出现的对象或事件,通过识别"原因事件"来完成情感原因的提取.这里的"原因事件"可以是实际的情感触发事件,也可以是对情感触发事件的一种感知.在该任务中原因事件是以短语级的粒度进行表示的.

由于触发情感的原因事件可能是名词短语、动词短语,也可能是由若干个短语组成的短句,这其中涉及许多复杂的语言学知识,造成传统的短语级情感原因提取任务复杂度大、提取准确率不高.因此,2016 年 Gui 等人[15]提出了子句级情感原因提取任务,即从给定的包含情感的文档中提取出触发该情感的原因,并以子句的粒度进行提取.

由于需要事先对情感子句中的情感关键词及其情感类别进行标注,这限制了其在现实场景中的应用.2019 年,Xia 等人[25]又在此基础上提出了"情感原因对"提取任务,该任务将情感子句和情感原因子句进行组合形成"情感原因对",提取任务的目标是成对地提取文档中潜在的情感子句和相应的情感原因子句,其优势在于提取原因子句时不需要提前知道情感子句的具体位置和情感类别.

情感原因提取任务和情感原因对提取任务的形式化定义如下:

给定一篇包含情感关键词和情感原因的文档 d,将该文档按子句的粒度进行划分 $d=\{c_1,c_2,\cdots,c_n\}$,n 为文档中子句的数量.每个子句 c_i 包含若干个单词,即 $c_i=\{w_1,w_2,\cdots,w_k\}$,$k$ 为子句中单词的数量,其中包含了情感关键词 E 的子句称为情感子句记为 c^{emo},包含情感原因的子句称为情感原因子句记为 c^{cau}.

子句级情感原因提取的目的是提取出文档中所有能够触发情感关键词 E 的原因子句 c^{cau},即由 $E\rightarrow c^{\text{cau}}$.值得注意的是,某个情感可能由多个原因触发,因此,情感原因子句 c^{cau} 的数量可能不止一个.

相对于这种子句级的粗粒度情感原因提取,细粒度情感原因提取则是以词或者短语块的粒度来对情感原因的边界进行限定,即在情感原因子句 c^{cau} 中提取出子字符串 $W^{\text{cau}}=\{w_i,w_{i+1},\cdots,w_j\}$,$0\leqslant i\leqslant j\leqslant k$.

对于"情感原因对"提取任务,其输入为文档 d 中的所有子句,输出则是以 Pair 对的形式进行组织,即给出二元组的集合 $P=\{(c^{\text{emo}},c^{\text{cau}})_1,\cdots,(c^{\text{emo}},c^{\text{cau}})_i,\cdots,(c^{\text{emo}},c^{\text{cau}})_m\}$,任务的目标是提取出文档中所有的情感原因对 $(c^{\text{emo}},c^{\text{cau}})_i$,其中 c^{cau} 子句是 c^{emo} 子句所对应的原因(注:同一个情感可能由

不同的原因触发,同一个原因也可能触发不同的情感).该任务与传统情感原因提取任务的最大区别在于其并不需要事先对情感子句 c^{emo} 进行标注,也就是任务本身并不依赖于是否给定情感关键词 E.

下面给出 2 个任务的 1 个实例:

给定文档 d =“昨天上午,一名警察带着丢失的钱拜访了那个老人,并告诉他小偷已经抓住了.老人十分高兴,并把钱存进了银行.”,将其划分为 5 个子句.

c_1 =“昨天上午”;

c_2 =“一名警察带着丢失的钱拜访了那个老人”;

c_3 =“并告诉他小偷已经抓住了”;

c_4 =“老人十分高兴”;

c_5 =“并把钱存进了银行”.

子句级情感原因提取任务是在给定子句 c_4 中表达的情感关键词“高兴”的基础上提取出触发“高兴”这一情感的原因是子句 c_2 和 c_3,即“高兴”→c_2,“高兴”→c_3.而对于短语级的情感原因提取则需要输出:“高兴”→“警察带着丢失的钱”,“高兴”→“小偷已经抓住了”.情感原因对提取任务则无需对情感关键词进行标注,直接输出文档中的情感和对应原因的子句对,即(c_4,c_2)和(c_4,c_3).

相对于传统的情感分析研究任务,情感原因提取的研究仍处于起步阶段,该领域公开发表的研究文献相对较少,本综述选取了近 10 年公开发表的情感原因研究文献作为研究对象,从提取粒度、研究方法、研究对象等多个角度对情感原因的研究工作进行分类归纳和总结,其统计结果如表 1 所示:

Table 1 Statistical Data of the Literature Classification for Emotion Cause Extraction

表 1 情感原因提取研究的文献分类统计表

分类依据	类别	代表文献
原因粒度	短语级	文献[2-7,9,12-13,15-16,22,26,30-32,34-43]
	子句级	文献[8,10-11,14,17-21,23-25,27-29,44-75]
研究对象	短文本	文献[9-10,12-13,22,30-32,34,37,40,55]
	长文本	文献[2-8,11,14-21,23-29,35-36,38-39,41-54,56-75]
研究方法	基于规则	文献[2-3,5-6,12-13,30-31,34-36,44]
	基于统计	文献[4,7,12-16,18-19,26,32,37-38,41-43,45,54]
	基于深度学习	文献[8-11,17,20-29,39-40,46-53,55-75]
语种	中文	文献[2-34,37,45-51,53-60,62-75]
	英文	文献[35,38-43,52,61]
	其他语言	文献[36,44]

情感原因提取任务从原因粒度来看主要有短语级和子句级两大类,近几年的研究大都是以子句为粒度,将情感原因的提取作为分类问题来进行处理.从研究对象来看主要分为微博类的短文本和新闻类的长文本两大类,现有的公开语料为基于新浪新闻的长文本.从研究方法来看,主要有基于规则、基于统计和基于深度学习这三大类,其中近几年的研究大都是以深度学习模型为基础.在语料的语言类型上,虽然曾有不同的学者从中文、英文、日文、意大利文这 4 种语言上开展研究,但目前研究最多的还是中文语料.

2 情感原因提取的主要方法

2.1 基于规则的方法

基于规则的情感原因提取方法主要是通过分析语料库,找出与文本情感原因相关的语言学线索并构建相关规则,之后利用规则提取导致情感变化的原因.

Lee 等人[2-5]首先针对“高兴”和“惊讶”这 2 种最基本的情感设计了若干语言学规则,对其进行情感原因的提取和分析.他们从情感认知理论出发,认为情感原因主要有动词类原因和名词类原因 2 种,设计了一套标注模式对包含文本情感原因的语料库进行标注,该标注模式对样本的情感类别列表、包含情感关键词的焦点句、焦点句中情感词的类别、焦点句的前一子句和后一子句等几个方面进行了标注.然后,利用使役动词、感官动词、连词、介词以及其他线索词提取出如表 2 所示的 6 组文本情感原因语言学线索词.他们随后又基于表 2 中定义的线索词

Table 2 Linguistic Cue Words for Emotion Cause Events[4]

表 2 情感原因事件语言学线索词[4]

类号	类别	线索词
Ⅰ	介词	为,为了,对,对于,以
Ⅱ	连词	因,因为,由于,于是,所以,因而,可是
Ⅲ	使役动词	让,令,使
Ⅳ	报导动词	想到,想起,一想,想来;说到,说起,一说,讲到,讲起,一讲,谈到,谈起,一谈,提到,提起,一提
Ⅴ	认知标志	听,听到,听说; 看,看到,看见,见到,见,眼看,瞧见; 知道,得知,得悉,获知,获悉,发现,发觉; 有;
Ⅵ	其他	的是,的说,于,能

生成了如表3所示的14条语言学规则.最后,根据这些规则开发了一套基于规则的情感原因提取系统.

Table 3 Linguistic Rules for Emotion Cause Extraction 1[4]

表3 情感原因提取语言学规则集1[4]

编号	规则
1	C(B/F)+Ⅲ(F)+K(F):C=F/B中Ⅲ组之前最近的名词/动词
2	Ⅳ/Ⅴ/Ⅰ/Ⅱ(B/F)+C(B/F)+K(F):C=F中K之前最近的名词/动词
3	Ⅰ/Ⅱ/Ⅳ/Ⅴ(B)+C(B)+K(F):C=B中Ⅰ/Ⅱ/Ⅳ/Ⅴ组之后最近的名词/动词
4	K(F)+Ⅴ/Ⅵ(F)+C(F/A):C=F/A中Ⅴ/Ⅵ组之后最近的名词/动词
5	K(F)+Ⅱ(A)+C(A):C=A中Ⅱ之后最近的名词/动词
6	Ⅲ(F)+K(F)+C(F/A):C=F或A中K之后最近的名词/动词
7	越C越K(F):C=F中2个"越"之间的动词
8	K(F)+C(F):C=F中K之后最近的名词/动词
9	Ⅴ(F)+K(F):C=F中的动词体
10	K(F)+的(F)+C(F):C=F中"的"之后最近的名词/动词+"的"+名词
11	C(F)+K(F):C=F中K之前最近的名词/动词
12	K(B)+Ⅳ(B)+C(F):C=F中Ⅳ组之后最近的名词/动词
13	Ⅳ(B)+C(B)+K(F):C=B中Ⅳ组之后最近的名词/动词
14	C(B)+K(F):C=B中K之前最近的名词/动词

注:C为情感原因,K为情感关键词,B为位于焦点句之前的子句,F为焦点句(包含情感关键词的子句),A为位于焦点句之后的子句,Ⅰ/Ⅱ/Ⅲ/Ⅳ/Ⅴ/Ⅵ为对应于线索词的组号.

以上规则的构建主要是基于语言学线索词、情感关键词、情感原因以及三者相对位置的组合.以规则1"C(B/F)+Ⅲ(F)+K(F):C=F/B中Ⅲ组之前最近的名词/动词"为例,该规则表明原因位于第Ⅲ组线索词(使役动词:让,令,使)的前面,因此为了提取原因C,就要在包含情感动词的焦点句F或者F之前的子句B中,找到第Ⅲ组线索词之前并且离之最近的名词或动词,并将包含该名词或动词的子句提取为原因子句.如对于文本"伊拉克细菌武器的曝光,使联合国大为震惊"可以提取成满足规则1的形式,即"[C伊拉克细菌武器的曝光],[Ⅲ使]联合国大为[K震惊]",其中"使"为线索词,"震惊"为情感关键词,"伊拉克细菌武器的曝光"为原因子句.在Lee等人[3-5]研究的基础上,Gui等人[12]又针对微博短文本的特点,增加了如表4所示的4条匹配规则.

Li等人[34]于2013年构建了一个中文微博情感语料库,基于触发情感原因的事件是情感的重要组成部分这一理论,提出一种利用情感原因作为特征之一来进行情感分类的方法,其情感原因事件的提取仍采用基于规则的方法.Neviarouskaya等人[35]则根据Ortony等人[76]于1988年提出的22类情感认知模型OCC构建了一个情感标注语料库,并通过依存句法分析、语义分析提取出了8种类型的情感原因提取规则.Gao等人[30-31]也是基于OCC模型设计了一个层次结构的情感原因分类体系模型ECOCC(emotion cause OCC),随后利用情感触发条件机制从事件结果、主体行为和实体对象这3类评价对象出发,将与模型框架中的情感规则相匹配的文本成份分为了6类,并分别设计了对应的评价成分和评价标准,其中在规则产生上引入了基础情感产生规则、复合情感产生规则以及延伸情感产生规则,最后通过建立子事件集的模型,从外部事件和内部事件2个角度来实现对情感原因的提取.

Table 4 Linguistic Rules for Emotion Cause Extraction 2[12]

表4 情感原因提取语言学规则集2[12]

编号	规则
1	动宾式(C,E),E为动词:C=包含原因候选词和情感词且二者之间是动宾关系的句子
2	E+"的"+C(F):C=包含"的"和原因候选词的焦点句
3	C(包含"的")+E:C=包含"的"和原因候选词的焦点句
4	C(B/F)+E:E=微博中特殊的情感表达,C=包含特殊情感表达词和原因候选词的句子或包含原因候选词的焦点句之前的句子

注:C为情感原因,E为情感关键词,F为包含情感词的焦点句,B为F左边的子句.

除了手工设计规则外,也有一些学者尝试借助一些外部知识来进行规则的自动构建.Russo等人[44]借助情感原因知识库,利用相关语言模式之间的相互作用关系,提出了一种自动提取意大利报纸文章中可能引起情感或情感状态变化原因的方法.该方法主要采用了最大期望算法(expectation maximization,EM)聚类模型和分类器的数据挖掘技术,自动归纳事件原因短语表达的规则.Yada等人[36]则采用了自举(bootstrapping)方法来自动获取情感原因的提取规则.该方法认为当某一情感的原因事件出现在另一个具有相同情感的句子中时,2个句子中位于情感原因和情感词之间的线索短语应该具备相同的连接功能.例如"过生日令我十分开心"和"过生日让我十分开心"这2句话具备相同的情感"开心"和原因事件"过生日",那么它们之间的线索短语"令"和"让"就应该具备相同的功能.因此,他们先通过人工给定的线索短语来收集情感原因;然后,从包含与先前收集的情感原因相似的情感短语中获得新的连接线索

短语,通过迭代不断地提取出新的线索短语.

2.2 基于统计的机器学习方法

传统基于统计的机器学习方法主要通过设计情感原因特征,然后将情感原因提取问题看作是一个文本分类或序列标注问题,进行有监督的文本情感原因提取.此类方法一般先假定触发情感的原因是一个或者一系列的事件,情感原因就在情感词附近.因此,先找出一段话中有意义的实词,然后确定分类的特征,比如事件特征、语言学特征、距离特征、词法特征等,最后利用这些特征完成情感原因的分类或序列标注.已有文献中,用于情感原因提取的特征大致可分为 6 类.

1) 事件特征

Talmy[77]从认知学角度出发,认为情感常常是由 1 个或多个事件触发产生的.Balahur 等人[78]也将情感的产生看成是动态的过程,这一动态过程主要是由一系列引发情感的事件所触发,他们通过构建"情感-事件"常识库来建立情感与其引发事件之间的关系,并在此基础上进行情感分类.因此,早期基于规则和基于统计机器学习的情感原因提取研究大都是将情感原因当成是一种特殊事件来进行提取,如 Lee 等人[3,5]是从规则中提取事件,Chen 等人[4]则是用机器学习的方法来提取事件特征.

Gui 等人[14-15]使用遵循万维网联盟(World Wide Web Consortium, W3C)标准的情感标记语言方案,建立了新浪新闻情感原因标注数据集,并提出了一种事件驱动的情感原因提取方法,该方法通过对包含情感的文本上下文进行句法分析来提取事件.同时他们对情感原因事件进行了正式定义,并通过七元组的方式来进行事件结构的表示.事件七元组的形式化定义为

$$e=(Att_{O_1},O_1,Adv,P,Cpl,Att_{O_2},O_2). \quad (1)$$

该定义基于中文是一种典型的主谓宾(SVO)结构,七元组中的 Att_{O_1} 和 Att_{O_2} 分别表示主语对象 O_1 和宾语对象 O_2 的属性;P 是谓语,表示一种动作或者行为;Adv 是用于修饰谓语 P 的状语;Cpl 则是谓语 P 的补语.由于一个事件中不一定会包含所有的 7 个元素,因此元组中某些元素的值可以为空.在通过依存句法对句子进行解析后,使用事件树进行表达和存储,最后再利用基于卷积核(树核)的支持向量机(support vector machine, SVM)算法进行情感原因事件的提取.同时,考虑到实际处理的需要,文献[14-15]的作者还设计了不同形式的核函数来进行分类.

王九硕[32]也从事件角度出发提出了一种用于中文微博文本的情感原因提取方法.该方法抽取博文中包含的子事件并标记,然后通过情感原因成分比例来挖掘情感与原因成分之间的对应关系,并以此提取出博文中包含的情感原因成分,找出与情感对应的原因事件.

2) 语言学特征

Chen 等人[4]在前期研究的基础上,提出一种多标签分类方法来提取情感原因,该方法不仅可以检测多个原因子句的问题,还可以捕获远距离的信息.他们将语言学线索词和语言学规则作为特征,同时考虑到手工设计特征的复杂性以及覆盖率低的问题,他们还设计了泛化性更好的特征来提取局部功能词结构、长距离连词结构以及中文所特有的一些泛化动词和认知动词结构,这些泛化特征的设计可以有效地保证特征的完备性.

Gui 等人[12]构建了一个包含 1 333 条语料的微博情感原因标注文本,并从中构造了 25 条情感原因匹配规则,随后从规则、距离、词性等角度进行特征的设计,最后采用 SVM 算法和条件随机场(conditional random field, CRF)算法进行情感原因分类和序列标注.其中,规则特征的使用方法是将规则转换为二元逻辑特征,即如果某子句符合某条规则,那么其对应的特征就是 1,否则为 0.除了基本的规则特征,情感原因出现的位置和其上下文之间也存在着一定的语言学规则.

Gao 等人[30-31]以 22 种细粒度的情感类型为基础,设计相关的提取规则,构建情感词汇,用于分析不同的情感原因触发不同情感的比例情况,在此基础上设计了多种语言学相关特征,用于中文微博的情感原因提取.这些特征包括各种表情符、程度副词(如"极其""很""欠""较""稍"等)、否定词、标点符号(如"!!!!"等)、连词(如"但是")等.袁丽[13]也是在构建微博文本情感原因数据集的基础上,利用统计模型提取了微博文本的情感原因提取规则,并结合句子距离、词语距离、候选词词性、表情符号、情感关键词及其词性等特征进行文本情感原因的提取.王赵煜[16]则基于中国知网情感词典(HowNet)和同义词词林的常识库扩展方法构造了一个认知常识库,并结合语言学特点,将常识库中的知识作为特征,用于情感原因的提取.

3) 距离特征

距离特征主要包括子句间的距离特征和词语间的距离特征,其中子句间的距离特征是指情感原因

子句和情感表达子句之间的相对距离,词语间的距离特征则是指情感原因子句中触发情感的词语和情感表达子句中情感关键词之间的相对距离.

针对子句的距离特征,文献[12]对中文微博情感原因数据集的分析表明,有近60%的情感原因和情感表达是在同一子句,有近30%的情感原因子句是在情感表达子句的前一子句或后一子句,在这30%的情感原因子句中有近80%是位于情感表达子句的前一子句.文献[15]对新浪新闻情感原因数据集的分析表明,有23.6%的情感原因和情感表达位于同一子句,有54.45%的情感原因子句位于情感表达子句的前一子句,因此可将子句的距离特征设置为-2,-1,0,1,2等,其中-2或-1分别表示位于情感表达子句前面的第2句或前一句,0表示和情感表达子句位于同一子句,即该子句就是情感表达子句,以此类推.

词语间的距离特征则是考虑到词语上下文的语境以及语用的特点,距离情感表达关键词越近的实词,其成为触发情感产生的关键词的可能性就越大.因此,可以将某实词的距离特征值设置为"1"或"-1"表示它位于情感表达关键词右边或左边且是距离其最近的第1个实词.

基于以上2种距离特征,文献[13]利用线性链条件随机场的特征,将文本情感原因提取问题看作是一个序列标注问题,在语言学特征和微博语义特征的基础上,添加词语距离特征和子句距离特征,提高标注的准确性.

4) 词法特征

考虑到情感原因通常包括名词性原因和动词性原因,因而词语的词法特征,如词性(part-of-speech, POS)等,也作为一种特征被用于情感原因提取任务.词法特征可分为情感原因候选词词法特征和情感关键词词法特征.其中,情感原因候选词词法特征主要考虑该词的词性是否是名词、动词、代词、限定词等,它主要用于对候选的情感原因子句中的词语进行判别;而情感关键词的词法特征则是指情感关键词的词性,主要有动词、名词、形容词、语气词等.情感关键词的词性和情感原因之间存在着一定的关联,例如文献[13]发现,名词性的情感原因其情感关键词一般为动词或形容词.除了基本的词性特征外,李逸薇等人[7]也将子句中名词个数、动词个数作为特征.

5) 上下文特征

文本中的子句以及词语并不是独立的,子句之间和词语之间都存在着上下文的语义关联以及一些常识性关联.文献[7]设计了上一子句中的动词、名词、标签以及下一子句中的动词、名词这5个特征作为上下文特征,用于情感原因的提取.

6) 主题特征

情感的产生和文本的主题存在较大的相关性,相同或相似的主题会触发相同或相似的情感.因此在利用主题模型方面,Song等人[37]提出了一个概念层面的情感原因模型CECM(concept-level emotion cause model),用来发现微博用户在特定热点事件中多样化情感的原因.CECM使用改进的二元词主题监督模型来检测某事件相关的推文中的情感主题,然后使用PageRank来检测有意义的多词表达作为情感原因.同时,该模型还能够检测出情感表情符和情感之间的关系.文献[13]也利用了主题模型来提取情感认知知识和情感的语义知识.

除以上6类特征外,Ho等人[38]结合心理学相关知识,提出了一种利用高阶隐马尔可夫(hidden Markov model, HMM)模型来模拟心理状态序列引发情感的过程,其核心思想在于:先将输入文本转换为导致心理状态的一系列事件,然后使用HMM对导致情感变化的状态序列进行建模.在构造HMM状态和将输入文本与这些状态的匹配过程中,将向量空间模型(vector space model, VSM)和潜在语义分析(latent semantic analysis, LSA)作为语义相似度比较机制,该机制可以检测出一些通用术语所表达的情感,并最终在数据集上取得了较好的效果.该方法既考虑了作为情感唤起过程的情感心理特征,又考虑了作为输入文本语法关系的语言信息.

与传统的机器学习方法不同,Xu等人[45]从信息检索的角度出发,基于文档排序的思想,提出了一种基于学习排序的方法来提取情感原因.该方法以文档中被触发的情感词作为查询,以情感段落中的各候选子句作为候选文档,设计了一套原因导向的子句级排序方法,用于对候选子句进行排序.该方法的重点在于将候选子句表示为包含情感独立特征和情感依赖特征的特征向量,学习有效的子句排序模型.其中情感独立的特征(子句长度、POS、线索词等)用于捕捉候选原因子句触发情感的可能性;而情感依赖特征(相对位置、词向量相似度、主题相似度)则用于捕捉候选原因子句与情感词之间的相关性.在排序方法上,文献[45]的作者分别从pointwise, pairwise, listwise这3个级别出发,采用了多种经典的信息检索排序算法来进行学习排序.

2.3 基于深度学习的方法

随着深度学习在自然语言处理领域中的广泛应用,基于神经网络的方法从2017年开始被应用于文本情感原因提取.其一般过程为:首先将词映射到向量空间中;其次通过神经网络模型来对文本特征进行自动提取;最后使用softmax函数将结果映射到概率空间来完成情感原因的提取.

从深度学习技术发展的脉络来看,神经网络模型经历了卷积神经网络(convolutional neural network, CNN)、循环神经网络(recurrent neural network, RNN)、长短时记忆网络(long short-term memory, LSTM)、门控循环网络(gate recurrent unit, GRU)、Transformer、图卷积神经网络(graph convolutional neural network, GCN)等基础模型的演变,现有的基于深度学习的情感原因提取模型也是在这些模型的基础上,通过组合、变形、融合注意力机制等方式来构造更为复杂的模型,提升提取效果.由于大多数模型都涉及好几项技术的交叉,特别是注意力机制基本在每个模型上都有一定程度的应用,因此本节对用于情感原因提取任务的深度学习模型进行介绍时行文的组织为:首先介绍采用基础神经网络的几种典型模型,随后介绍涉及多种基础模型的混合模型,最后分析在神经网络模型的基础上借助特定技术(如多任务、知识蒸馏等)来融入外部信息进行辅助的几个代表性模型.本节的总体组织结构如图1所示.

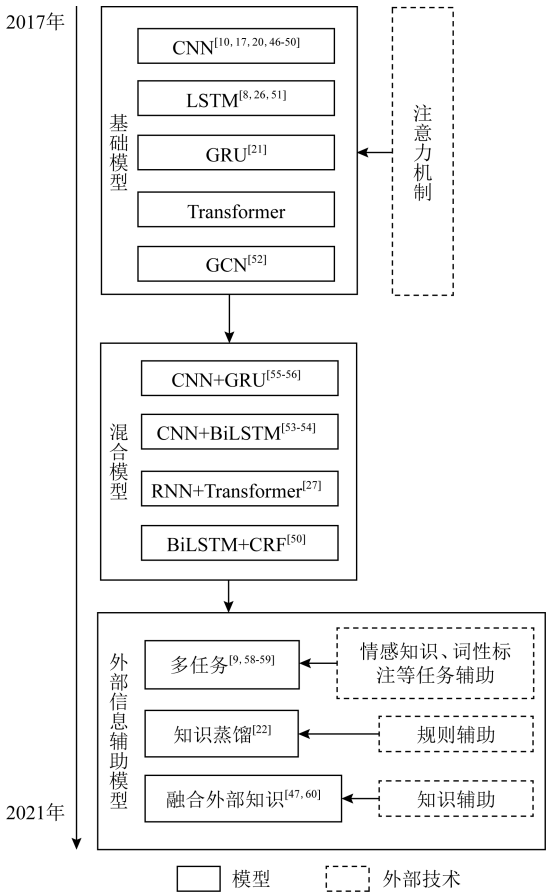


Fig. 1 Emotion cause extraction based on deep learning

图1 基于深度学习的情感原因提取技术

2.3.1 基础神经网络结合注意力机制模型

1) CNN

Gui等人[17]受问答领域的启发,将情感关键词作为查询词,将其上下文作为查询文本,通过问答的方式来判断当前子句是否为情感原因.他们基于该思想设计了名为ConvMS-Memnet(convolutional multiple-slot memory network)的模型,该模型利用CNN的卷积机制,并通过多槽记忆网络来实现对远距离上下文信息的建模,达到同时提取词级序列特征和词汇特征的目的.为了验证网络深度的作用,他们分别设计了单层的网络模型,如图2所示,以及多层的网络模型.傅科达[46]也在句子级别上分别设计了基于端到端、基于词向量、基于注意力机制、基于关键词-值网络等多种记忆网络模型来提取情感原因.

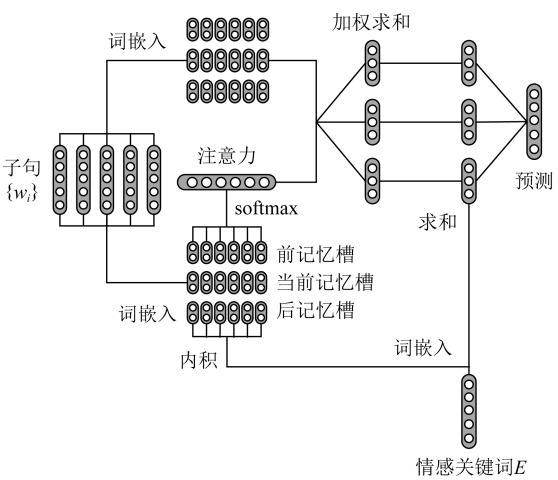


Fig. 2 A CNN and memory network based model for ECE[17]

图2 基于CNN和记忆网络的情感原因提取模型[17]

Chen等人[10]针对中文微博中的情感原因提取问题,提出了一种分层卷积神经网络模型来提取微博中的事件特征,该模型设计了子句级编码器和子推文级编码器来分别处理特征稀疏问题和事件信息不足问题.首先,底层的子句级编码器结合多个神经网络提取每个子句的局部特征;然后,上层的子推文级编码器将这些局部特征作为一个序列,通过卷积

神经网络为每个子句提取序列特征.此外,考虑到由于小规模实验数据中的特征稀疏问题,其子句级编码器还提取了 2 种相互补充的局部特征,即基于卷积神经网络的显著特征和基于注意力网络的加权特征.

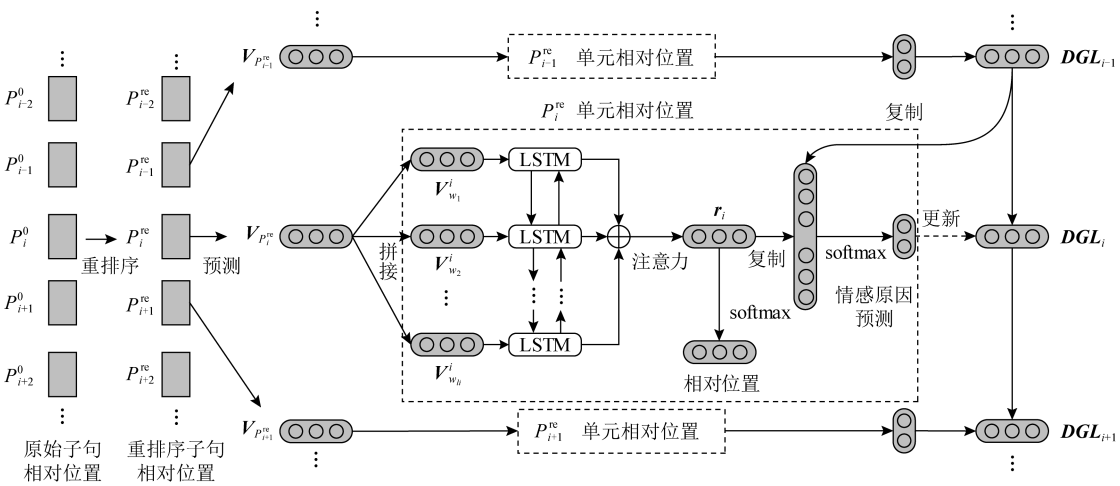
Diao 等人[47]也提出了一种基于注意力的上下文卷积网络模型 EACN(enhanced-representation attention convolutional-context network),该模型采用了一种新的处理机制,即在情感词信息背后引入分层上下文,并将这种上下文作为输入,通过卷积运算提取情感原因,充分捕捉子句之间的层次语义关系,从而构建复杂句子结构中情感词及其情感成因之间的关系,以便更好地理解情感词及其上下文语境.

慕永利等人[48-49]则是从解决数据集不平衡性的角度出发,提出了一种基于集成 CNN 的情感原因提取方法.该方法充分利用了 CNN 中的卷积和池化等相关技术来对句子的语义信息进行融合.此外,考虑到情感原因数据集中非原因子句和原因子句的不平衡性,他们将非原因子句集合分解为多个子集并与原因子句集组合构成多个较平衡的数据集,然后分别通过一个单独的 CNN 模型进行训练,最后将它们进行集成并用于最终的提取.

郑胜协[50]设计了名为 CAES(compare aggregate model with embedding selector)的网络模型,该模型在 CNN 的基础上,结合互注意力机制和自注意力机制,并利用 K-max 方式对注意力权重进行剪枝.其中互注意力机制用于捕获情感词和子句的关系,自注意力机制用于捕获每个子句自身内部词语的重要性,而剪枝则可以去除不相关的文本片段在注意力归一化时造成的噪音.此外,他还把情感原因发现问题当成一个排序问题,利用 pairwise-rank 方式进行建模.

2) LSTM

为了进一步挖掘子句间的因果关系,Ding 等人[26]创新性地提出,除了文本本身的内容外,子句的相对位置信息和全局标签信息对于情感原因的提取也至关重要.其中相对位置信息主要是表示候选子句和情感句之间的相对距离,而全局标签则是为了表示除当前子句外的其他所有子句的当前预测结果.为了整合这些信息,他们提出基于双向长短时记忆(bidirectional long short term memory, BiLSTM)的 PAE-DGL(relative position augmented embedding learning- reordered prediction with dynamic global labels)模型,如图 3 所示,以统一的端到端的方式来编码这 3 个要素(文本内容、相对位置和全局标签),模型中采用了一种相对位置增广的嵌入学习算法,将任务从一个独立的预测问题转化为一个包含动态全局标签信息的重排序预测问题.该方法最大的创新在于在预测过程中能够随着已有子句的预测结果动态调整当前子句的预测结果,也就是说如果前一个子句被预测出有较高的概率为原因子句,则其后的子句被预测为原因子句的概率则自动降低,反之亦然.



注:P 代表相对位置,V 代表位置向量,r 代表子句的向量表示,DGL 代表动态全局标签向量.

Fig. 3 A relative position and global label based model for ECE[26]

图 3 基于相对位置和全局标签的情感原因提取模型[26]

夏林旭等人[51]同样采用注意力机制和 BiLSTM 神经网络模型来进行情感原因的提取,但他们采用字符向量来表示文本的语义信息,并且在提取文本特征时还结合了人工提取的子句特征.

与现有大多数研究仅针对单用户的单条微博内容进行情感原因提取不同的是,Cheng 等人[8] 于 2017 年提出了一种基于多用户结构(某一微博下多个用户的交互,其中最原始发布的微博称为原推文,其回复称为子推文)的中文微博情感原因提取方法.为此,他们首先专门设计了一种情感原因标注方案,用来处理在多用户结构中某个用户的情感原因可能来自于其他用户这一复杂情况,并基于该标注方案构建情感原因标注语料库;然后,通过对该语料库的分析,提出了基于子推文和基于原推文的 2 种情感原因提取任务;最后基于 LSTM 模型来实现情感原因的提取.

3) GRU

Fan 等人[21] 通过对语篇的上下文信息进行建模,并引入外部情感知识库来进一步辅助情感原因的发现,在此基础上提出了一种正则化的层次神经网络(regularized hierarchical neural network, RHNN)模型,如图 4 所示.

该模型通过 GRU 并结合层次化注意力网络来对词级和子句级的语篇结构信息进行建模,并最终为子句表征生成有用信息.考虑到情感原因子句中存在一些蕴含情感极性的关键词,及情感原因子句和情感子句中情感词的相对位置关系,他们还设计了基于情感字典和相对位置的正则化机制来对训练模型中损失函数进行约束.

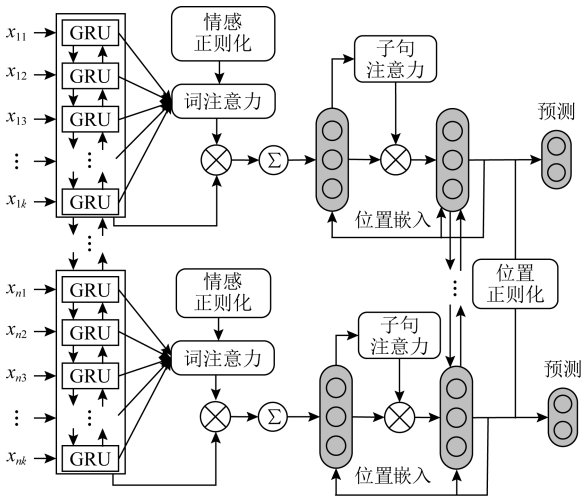
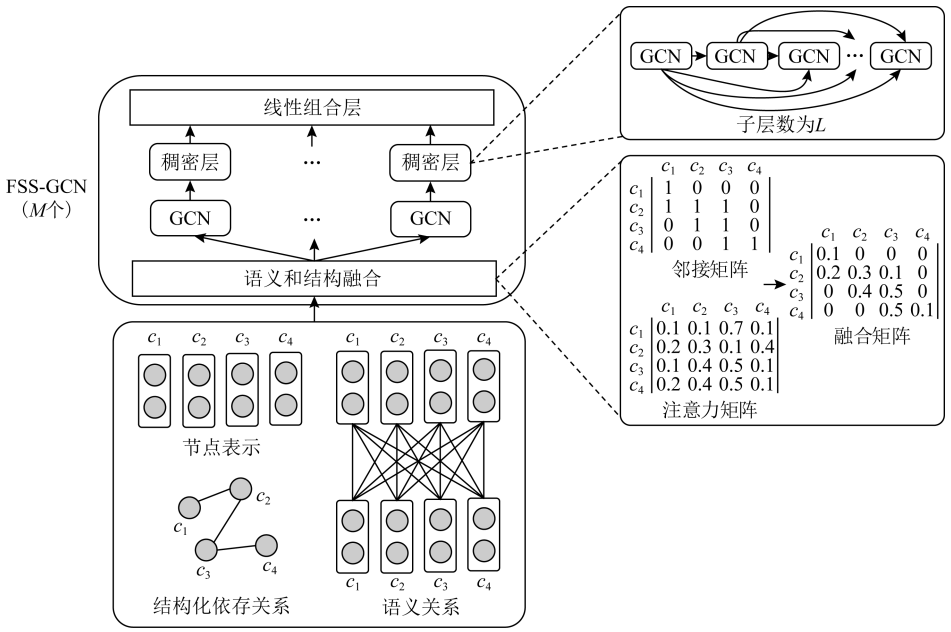


Fig. 4 A knowledge regularization based model for ECE[21]

图 4 基于知识正则化的情感原因提取模型[21]

4) GCN

随着图卷积神经网络技术的发展,该技术也被广泛应用于链接预测、事件检测以及推荐系统等领域,许多自然语言处理(natural language processing, NLP)任务中的问题也通过图卷积神经网络得到了成功解决.现有的情感原因提取方法大多通过注意力机制或联合学习来获取语义信息,其子句编码器大都是以 LSTM 或 GRU 为基础序列模型.这类模型难以刻画子句之间的长距离或全局依赖,从而忽略子句间的深层依赖关系.因此,Hu 等人[52] 提出了



注:M 为 FSS-GCN 模块的堆叠个数,L 为每个 FSS-GCN 模块中 GCN 的层数.

Fig. 5 A GCN based model with fusion of semantics and structural constrict for ECE[52]

图 5 融合语义和结构约束的图卷积网络情感原因提取模型[52]

一种基于子句依存关系的融合语义和结构约束的图卷积网络结构(graph convolutional structure with fusion of semantics and structural constrict, FSS-GCN)模型,如图 5 所示,该模型通过将 GCN 基础模型和基于注意力引导的图卷积神经网络(attention guided graph convolutional network, AGGCN)作为子句级编码器,利用子句之间的依赖关系来加深对文本语义的理解.模型还通过不断向网络中注入结构约束,将焦点从全局结构缩小到局部结构,使得该模型能够选择性地注意到有助于情感原因分析的相关子句.

2.3.2 基础神经网络混合模型

1) CNN+LSTM

Li 等人[53-54]认为,前期的研究方法忽略了可能为情感原因提供线索的上下文,然而上下文中的子句在激发某种特定情感方面发挥着不同的作用.借助于注意力机制的特点,他们提出基于情感上下文感知的共注意力机制神经网络(co-attention neural network,CANN)模型.该方法首先通过 BiLSTM 模型对原因候选子句和情感子句进行编码,然后送入 CNN 的卷积层进行情感原因提取.此后,他们又提出基于多注意力机制的神经网络(multi-attention-based neural network, MANN)模型,如图 6 所示.该模型通过 BiLSTM 整合词语的上下文信息,并利用多注意力机制捕获情感子句和候选子句之间的相互影响,生成情感子句和候选原因子句的向量表示.其中的多注意力机制主要分为情感词引导的注意力和候选子句引导的注意力.

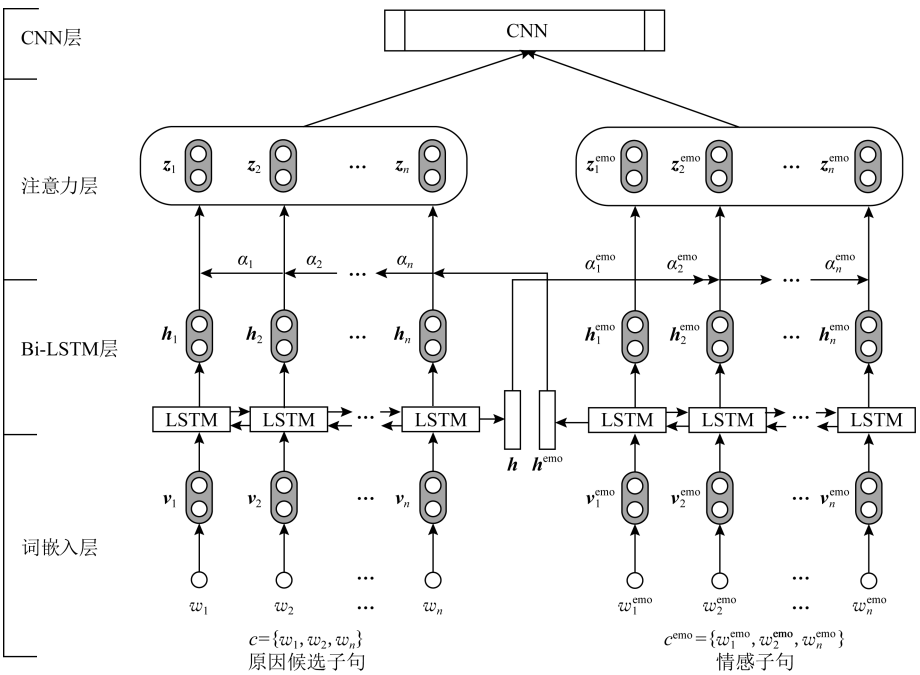


Fig. 6 A multi-attention-based model for ECE[54]

图 6 基于多注意力机制的情感原因提取模型[54]

2) CNN+GRU

在利用基本的深度学习模型配合注意力机制的基础上,也有一些学者从情感分析问题本身出发,从文本粒度的角度开展研究.在文本分析粒度方面,Yu 等人[55]基于层次化网络的多级别子句选择框架来提取情感原因,框架由低到高依次由词级、短语级和子句级 3 个网络构成,如图 7 所示.具体地,通过基于内容和位置注意力矩阵的单词级网络建模单词级信息,通过 CNN 建模短语级信息,通过双向门循环单元(bidirectional gated recurrent unit, Bi-GRU)建模子句级信息.这种多级别的建模方式的优点在于综合考虑了文档特征的多个因素,如词的位置、不同的语义级别(单词和短语)、子句的交互等.

Diao 等人[56]将情感原因提取作为一个机器阅读理解问题,设计了一个名为 MBiAS(multi-granularity bidirectional attention stream)的多粒度双向注意力流网络模型,模型中的双向注意流层能够捕获情感查询感知上下文表示中的深层次交互,从而学习和理解其中的语义关联.模型在字符级、词级、类别级、句子级和位置级等多个层面对上下文

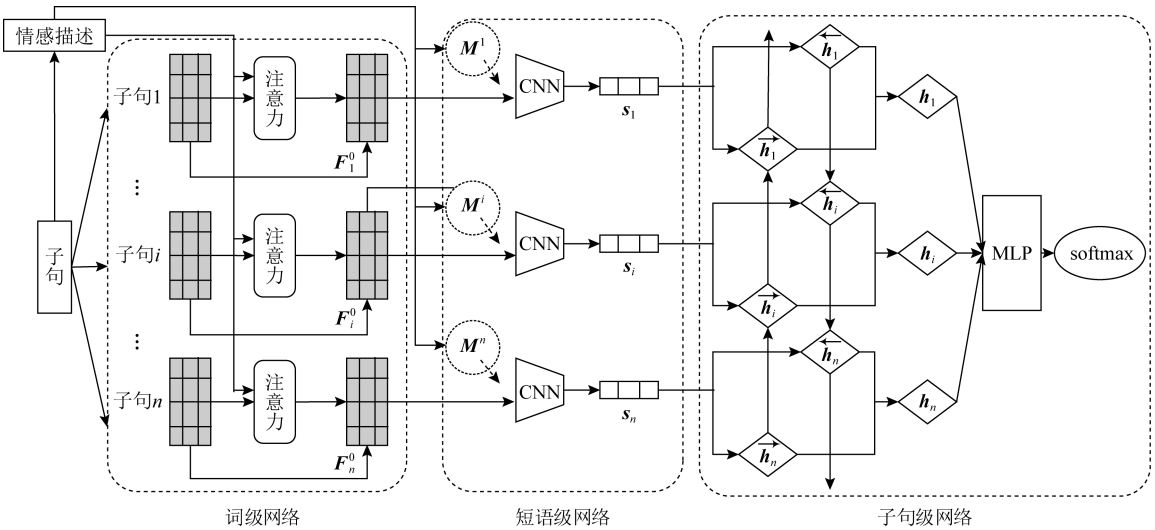


Fig. 7 A multiple level hierarchical network-based model for ECE[55]

图 7 基于多级别层次网络的情感原因提取模型[55]

段落和查询进行建模,随后基于双向注意流机制,从查询-上下文和上下文-查询 2 个方向获取情感查询感知的上下文表示.

3) RNN+Transformer

2018 年以来,基于自注意力机制的 Transformer 模型在深度学习领域大放异彩.Xia 等人[27]在前期研究的基础上利用 Transformer 模型设计了名为 RTHN(RNN-transformer hierarchical network)的联合情感原因提取框架,同步地对多个子句进行编码和分类.该框架由一个基于 RNN 的低层词级编码器和一个基于 Transformer 的高层子句级编码器组成,前者用于在每个子句中编码多个单词,后者用于学习文档中多个子句之间的相关性.此外,该模型还将相对位置信息和全局预测信息编码到转换器中,以便更好地捕获子句之间的因果关系.

4) BiLSTM+CRF

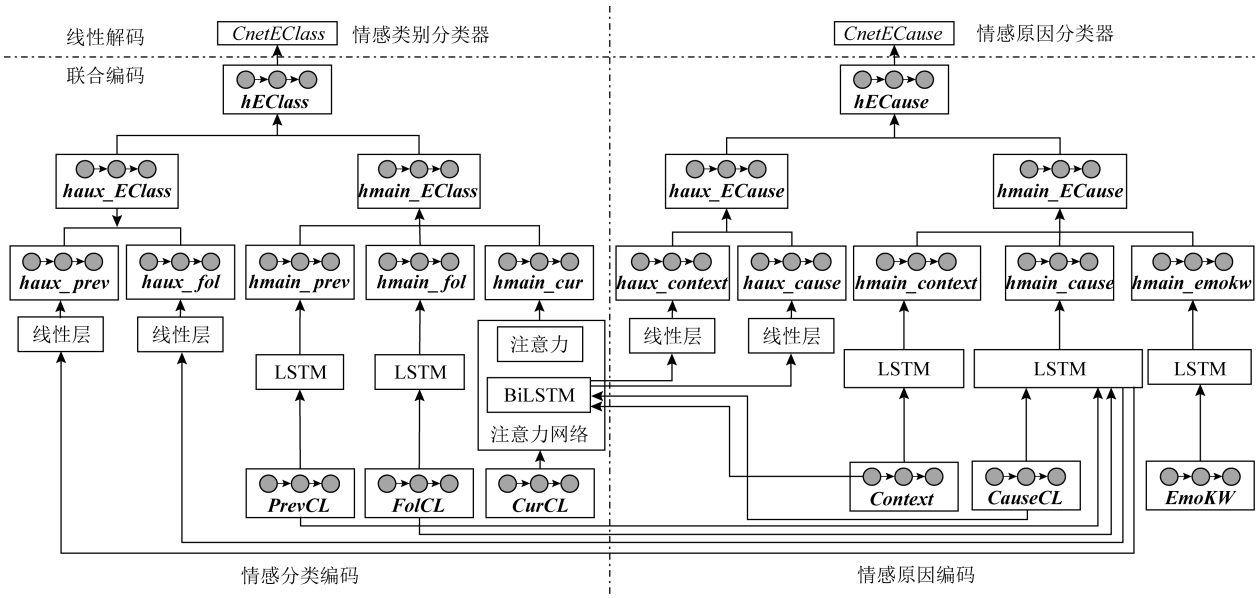
在用传统的 CRF 进行情感原因提取时,由于特征的提取效果与词之间的距离有很大关系,而原因子句和情感关键词之间经常相隔较长的距离,这就限制了 CRF 的抽取效果.而传统的 LSTM 模型虽然有强大的序列建模能力,能够处理较长的文本数据,但对输出标签的约束能力较弱,无法很好地反映当前时间步的标签是否受其他时刻标签的影响.因此,张晨等人[57]从序列标注的角度出发,将条件随机场与神经网络相结合,提出了一种基于 BiLSTM+CRF 的情感原因提取和表情符情感识别的联合模型.模型将情感原因提取以及情感识别形式化为统一的序列标注问题,充分利用了情感与其原因之间的全局关联性.

2.3.3 外部信息辅助模型

1) 任务辅助——多任务联合学习模型

随着多任务学习技术在深度学习中的广泛应用,有一些学者从多任务学习的角度对情感原因提取开展研究.Chen 等人[9]提出了一种基于神经网络的情感分类和情感原因提取的联合方法,如图 8 所示,针对情感分类和情感原因提取需要不同类型的特征(分别基于情感和事件),提出了一种联合编码器,使用统一的框架来提取 2 个子任务的特征,并用多任务模型同时学习 2 个分类器.此外,由于实验数据存在特征稀疏问题,注意力网络无法有效提取到能够表达事件的特征,因此在进行联合编码时只用 LSTM 提取事件特征,而增加注意力机制以提取情感特征.该方法本质上是一种多任务学习,它试图借助情感分析的 2 个子任务之间的内在关联来提升效果.

Hu 等人[58]认为每一个子句都可以从情感和原因 2 个角度来理解,并以此提出了一种情感原因联合检测(emotion cause joint detectoin, ECJD)模型,将情感识别和原因提取作为 2 个子任务,与情感原因联合提取这一主任务统一到同一个框架中,以同步和联合的方式来增强子任务之间的交互.他们将问题形式化为一个四分类问题(普通子句、情感子句、原因子句、既是情感也是原因子句).子句的特征表示从情感和原因的双重视角来评估,即从情感的角度关注子句对情感的贡献,同时也从原因的角度关注子句对原因的贡献.余传明等人[59]也提出了一个基于 LSTM 和多任务的情感原因提取模型,该模型利用词性标注这一辅助任务来帮助情感原因提取主任务.



注:*haux* 代表辅助向量表示,*hmain* 代表主向量表示,*prev* 代表前一子句,*fol* 代表后一子句,*cur* 代表当前子句,*emokw* 代表情感关键词,*context* 代表上下文,*cause* 代表原因,*class* 代表情感类别(首字母不论大小写).

Fig. 8 A multi-task model for ECE[9]

图 8 多任务情感原因提取模型[9]

2) 规则辅助——基于知识蒸馏的模型

深度神经网络的训练往往依赖于大量高质量的标注数据,但缺乏对人工构造的语言表达规则的有效利用,同时也存在解释性和可控性不强等问题.恰当地利用规则或学习规则可以提高模型的可解释性,减少训练样本的数量.

巫继鹏等人[22]将情感原因发现的语言学规则通过知识蒸馏技术引入到模型训练中,从而实现传统基于规则的方法和深度学习方法的有机融合.他们提出了一种结合规则蒸馏的情感原因发现模型RD-HAN,该模型由四大组件构成:教师编码器 E_T、学生编码器 E_S、分类器 H、辨别器 D,如图 9 所示.其中教师编码器和学生编码器均为结合位置信息和残差结构并基于 Bi-GRU 的层次注意力网络.这一层次网络结构用于捕获词级和子句级的序列特征,而注意力机制用于捕获子句与情感表达之间的潜层语义表示.

模型中教师编码器是一种融入了语言学规则的复杂编码器,为了有效利用语言学规则并将其嵌入到深度神经网络中,文献[22]的作者将原始的输入文本根据情感词、原因线索词、情感持有者以及情感原因这 4 种角色进行了标注和编码,并通过规则约束

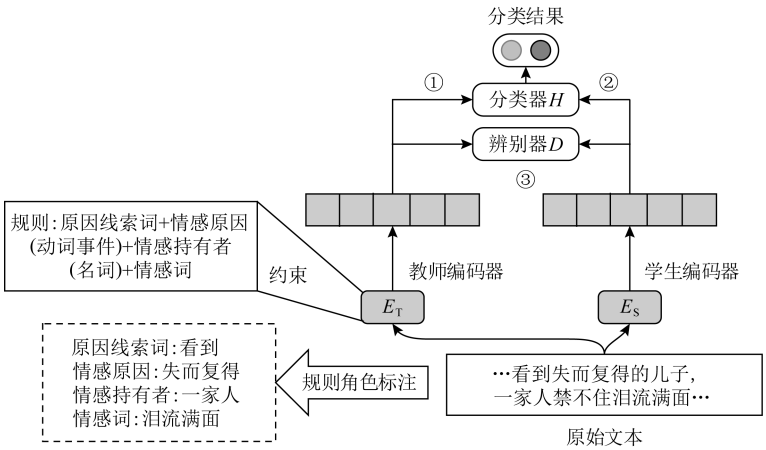


Fig. 9 A rule distillation based model for ECE[22]

图 9 基于规则蒸馏的情感原因提取模型[22]

训练得到教师编码器.而学生编码器的输入不需要包括额外的规则信息,但通过知识蒸馏的方式它可以从教师编码器中隐式地学习到规则的相关信息.

分类器的作用是判断某子句是否为原因子句,辨别器的作用则是用于判断输出到底是来自教师编码器还是学生编码器.模型首先对教师编码器和分类器进行训练,训练的输入既包括文本语义的编码信息也包括规则角色的编码信息;接着固定分类器来对学生编码器进行训练,此时的输入就只包括文本语义的编码信息;最后再通过基于对抗学习的知识蒸馏技术来交替训练辨别器和学生编码器,其目的在于让学生编码器的输出能够尽量接近教师编码器的输出.

3) 知识辅助——融合外部知识的模型

考虑已有的情感原因提取模型性能均受到情感层面的语义信息融合不足以及语料库规模有限的影响,也有一些学者尝试将外部知识引入到模型中.Hu 等人[60]提出了一种融合外部情感知识的情感原因识别(external sentiment knowledge for emotion cause detection, ExSenti-ECD)模型,该模型采用了一种情感特定的嵌入方法,将情感文本中包含的外部情感知识编码成词向量,以此来提高词向量中对于情感知识的表示能力.他们首先将多个公开的语料库合并成一个新的包含情感极性的语料库,然后采用 BERT 模型对其进行预训练,以此来赋予模型更高的融合文本情感信息的能力.

Diao 等人[47]也从增强情感语义表示这一角度出发,提出了一种利用情感词及其同义词的语义增强表示方法.该方法对普通词向量、情感词向量以及基于同义词词林的情感词同义词向量这 3 种向量进行融合,获得增强的向量表示,随后将其添加到基于注意机制的词级和子句级向量表示中,捕获其中与情感相关的重要信息.

2.4 各类方法的特点分析

基于规则文本情感原因提取中,规则的构建过程相当于情感的语义理解过程,所以规则清晰易懂,并且准确率比较高,计算复杂度也相对较低.但是基于规则的方法局限性也是明显的.首先,规则通常依赖于语言学线索词,但在情感文本语料库中,含有语言学线索词的情感句子比例较低,制定的规则并不能完全覆盖所有的语言现象,造成覆盖度低、泛化能力差的问题.其次,同一个子句可能同时匹配多个规则,容易造成规则冲突.再者,基于规则的方法通常无法应对包含多个原因子句的情况.最后,不同领域语料的语言结构有一定的区别,针对特定领域的文本制定的规则并不能很好地适用于其他领域.

基于统计机器学习的方法主要依赖于特征工程.对于情感原因提取任务来说,除了考虑传统的语法、语义、词性、上下文等特征外,相对位置、情感、语言学等特征也起到十分关键的作用.该类方法的优点在于通过概率来描述模型的不确定性,从而进行不确定性推理,具有较强的泛化能力,它们能够根据特征工程最大限度地从原始数据中提取特征,供算法和模型使用,并且在数据的驱动下不断地进行参数优化.然而,特征工程是一件十分繁琐的任务,需要有较强的业务背景和很高的人力成本.另外,情感和情感原因之间存在不同程度的因果语义关联,如何设计和提取有效特征来反映这种深层次的因果语义关系,仍然面临许多挑战.

基于深度学习的方法优势在于:它抛开或简化了繁琐的特征工程设计,能够自动从数据中学习到有效的特征表示.在情感原因提取任务中,通常的做法是在充分理解语义的基础上,采用深度神经网络模型并结合注意力来捕获原因子句和情感子句之间的关联.由于多任务学习可以通过多个相关任务之间的联合训练来捕获任务间的一些内在关联,因此结合多任务机制也是当前许多主流模型所采用的一种解决方案.同时,由于文本情感原因提取涉及较为复杂的语言学和情感认知领域的知识,因此目前取得较好效果的模型则是通过知识蒸馏或者知识正则化的方式来利用这些知识.然而,深度神经网络模型本质上是一种数据驱动的模型,对样本的数量和标注质量有较高的要求,在情感原因提取任务上,数据资源的缺乏在一定程度上限制了该类模型的效果和应用场景.

总的来说,情感原因提取方法是随着机器学习技术的发展而不断更新,在其发展的早期一般是以基于规则的方法为主,随后是基于统计的机器学习方法,而近几年则是以深度学习方法为主流.但由于情感原因提取与传统情感分析相比是一种更深层次的文本挖掘任务,因此与传统的文本分类相比,如果仅仅用时下主流的深度模型进行文本分类难以取得理想的效果,所以现有的一些效果较好的模型均是将深度学习技术与传统方法相结合,通过引入外部知识来提升效果,例如融入规则知识或引入额外设计的特征等.

除模型本身的特点外,语种也是影响模型选择的重要因素.这主要是由于不同语种本身的语言特

点以及数据资源的情况引起的.

首先,不同语种在语素识别、词性标注、词义消歧、词汇粒度抽取、句法结构分析、指代消解等方面均存在差异,这些差异会影响情感原因提取方法的选择和提取效果.

以一词多义现象的影响为例,文献[21]用同样的模型分别在英文和中文数据集上开展了实验,实验结果表明 SVM 和 Word2vec 方法在中文数据集上性能相差不大,但在英文数据集中 SVM 方法比 Word2vec 方法的 $F1$ 值高了 11 个点,一个主要原因就是一词多义限制了 Word2vec 的效果.而文献[21]的作者提出的深度神经网络模型在中文数据集上的 $F1$ 值达到了 79.14%;而在英文数据集上却只有 59.75%.此外,Oberländer 等人[61]对比了序列标注方法和基于子句的分类方法在英文情感原因数据集上的效果,实验结果表明目前在中文数据集上广泛采用的子句级粒度的分类方法并不适合于英文语料.

其次,许多方法和模型中都会利用如情感词典、预训练语言模型等外部资源来丰富语义的表示,但不同语种可利用的外部资源在种类、质量上存在差别,这也会影响模型的选择和最终的效果,使得一些小语种上的情感原因提取研究聚焦在知识库或者数据集的构建上.例如文献[44]通过众包方式来获取意大利文的语料,并构建相应的情感常识库来辅助情感原因的提取,文献[36]则是通过自举的方式对日文的情感原因标注集进行自动扩充.最后,标注数据的缺乏限制了在某些语种上研究方法的选择.现有的深度学习技术通常是需要有大量的标注数据,如果数据量太少,基于规则或者统计的方法或许也是不错的选择.

3 情感原因对提取

传统的情感原因提取任务需要对情感句先进行标注,这大大限制了其应用场景.Xia 等人[25]在前期研究的基础上,提出了情感和情感原因联合提取任务,即情感原因对提取(emotion cause pair extraction, ECPE).文献[25]的作者提出了一种"2 阶段"的方法进行情感原因对的联合提取:阶段 1:独立的情感子句提取和原因子句提取;阶段 2:情感原因的配对和过滤.阶段 1 中主要设计了 2 种多任务网络模型来进行子句的提取,一种为独立的多任务学习模型,另一种为交互的多任务学习模型(如图 10 所示),其中后者是对前者的一个增强版本,它能够捕获情感和原因之间的内在关系.阶段 2 则是通过笛卡儿积来对阶段 1 提取出的情感子句和原因子句配对,再通过因果分析对配对结果进行过滤.

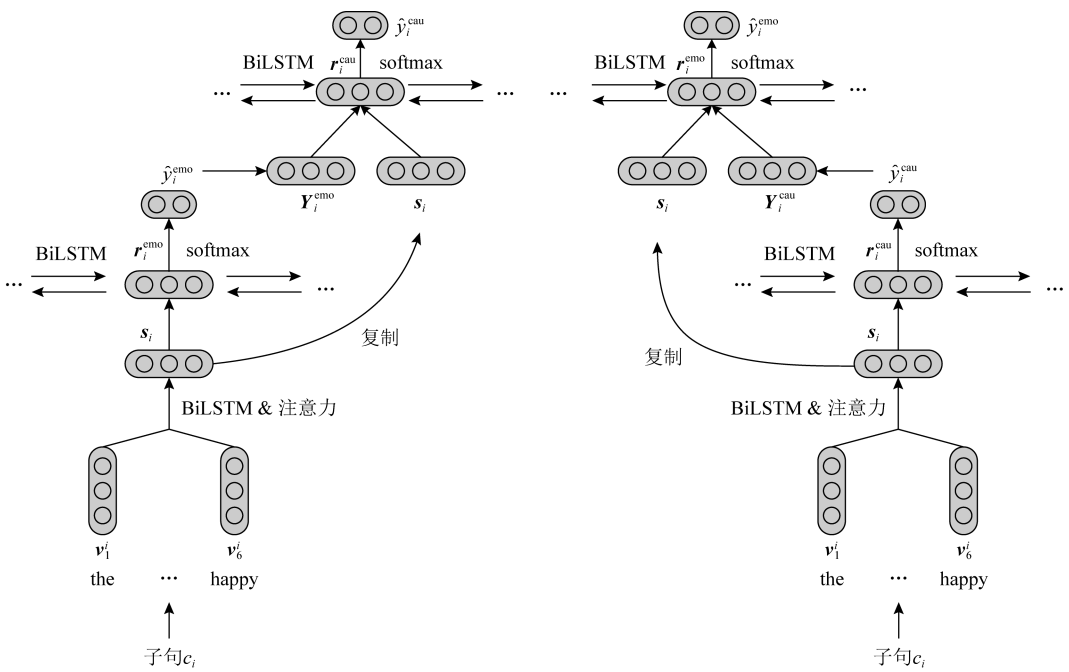


Fig. 10 A multi task based model for emotion casue pair extraction[25]

图 10 基于多任务的情感原因对提取模型[25]

以图 10(a)的 Inter-EC 模型为例,模型底层首先以分词后的子句 c_i 作为输入,通过 BiLSTM 及注意力机制获得每个子句的向量表示 $\boldsymbol{s}_i$;然后将 $\boldsymbol{s}_i$ 进行复制,一份作为情感子句提取的输入,另一份则与情感子句提取任务的输出 $\boldsymbol{Y}_i^{\text{emo}}$ 进行拼接,并作为情感原因提取的输入;最后再通过 BiLSTM 获得情感原因提取的输出 $\boldsymbol{Y}_i^{\text{cau}}$.将情感提取任务的输出与子句的表示进行拼接体现了多任务学习中的信息共享,同时也在某种程度上将情感和情感原因之间关联融入到模型之中.图 10(b)的 Inter-CE 模型结构本质上和图 10(a)的 Inter-EC 是对称的,Inter-EC 是利用情感提取来改进原因的提取效果,而 Inter-CE 则是利用原因的提取来改进情感提取的效果.该模型的最大特点在于利用了多任务联合学习的优点来进行原因对的联合提取.

Shan 等人[62]认为图 10 的 Inter-EC 模型并没有充分考虑情感子句和原因子句的相关性,并且对于上下文的利用也十分有限,因此他们在该模型基础上提出了一种基于 Transformer 的组件来对子句进行编码.该组件通过自注意力机制将更远距离的上下文信息编码进子句的向量表示,从而优化不同任务间的交互.

Yu 等人[63]则认为"2 阶段"方法中情感子句提取任务与原因子句提取任务之间仅仅通过单向关联并不能充分利用二者之间的相关性,而且第 1 阶段的误差将会直接降低第 2 阶段的性能.针对这 2 个不足,他们提出了一个相互辅助的多任务模型,通过添加 2 个与原任务相同的辅助任务来促进情感子句和原因子句的提取.该模型将 2 个辅助任务产生的预测结果作为彼此主要任务的额外特征,从而建立情感与原因提取的双向关联.其次,为了减少误差传播对第 2 阶段的影响,Yu 等人还针对这种"任务对"形式的模型通过"自蒸馏"的知识蒸馏技术进行训练,进一步提升第 1 阶段任务提取的准确率.

虽然文献[62-63]提出的方法在一定程度上提高了情感原因对提取任务的效果,但这种"2 阶段"的 Pipeline 方法不可避免地存在误差传递的问题.为了从根本上解决该问题,近 2 年有越来越多的学者尝试通过构建"端到端"(end-to-end)的统一模型来一次性地完成情感原因对的提取.由于情感原因对的提取涉及到情感子句的提取、原因子句的提取这 2 个基本任务,因此这些端到端的模型基本上还是基于多任务学习的思想而设计的,模型间的区别主要体现在情感原因子句对的构建方式及处理上.

相对于情感原因提取任务,情感原因对提取任务中,情感子句与原因子句的配对是关键.传统的"2 阶段"方法是通过先筛选出可能的情感子句集合和原因子句集合,然后再将 2 个集合以笛卡儿积的方式来构造候选的子句对.该方法存在的问题主要是计算代价较大,同时由于情感子句或原因子句在阶段 1 未被正确提取导致子句对缺失.因此,很多学者尝试了不同的子句对构建方式.

1) 基于关系分类的子句配对

Wu 等人[64]专门设计了一个"子句对关系分类"子任务来处理子句的配对问题,并将该任务与情感识别和原因提取任务一起进行多任务联合学习.在该模型中子句对关系的识别并不依赖于情感识别或者原因提取的结果.在进行子句对关系分类这一任务时,如果将所有的子句都进行两两配对,那么真正具有因果关系的子句对样本将极其不平衡.因此,在训练阶段对于子句的选取是以数据集中的真实标签为依据,只有子句对$\langle c_1, c_2\rangle$中的 c_1 为包含情感子句,或者 c_2 为原因子句,该子句对才会被用作训练样本.

2) 基于矩阵变换的子句配对

Ding 等人[28]提出了 ECPE-2D(emotion cause pair extraction two dimensional)模型,该模型先视文档中所有子句既是情感子句也是原因子句,在此基础上两两配对,构造一个 2 维方阵进行子句对的表示.由于方阵中真正有因果交互的子句对只占很小的一部分,基于 Transformer 设计了基于窗口大小限制、基于行列十字交叉等变换方法对 2 维方阵中的子句交互进行建模,再通过一个标准的二分类预测来完成情感原因对的提取.

文献[28]的方法虽然在一定程度上减少了子句对匹配时的计算复杂度,但在对筛选出的子句对进行预测时本质上还是只依赖于当前的子句对,而借助于双仿射矩阵及其计分函数则可以从全局的角度计算每一候选子句对中是否为因果关系的可能性.因此通过双仿射机制来处理子句对这种以"对"形式存在的目标时就有其一定的优势.Tang 等人[65]构建了一个基于多注意力和双仿射机制的多任务模型 LAE-MANN(latent variable enhanced multi-level attentional neural network).该模型首先利用自注意力机制和互注意力机制来识别情感词和情感子句的关系,以及情感子句和原因子句间的关系;随后,通过构建一个双仿射矩阵来表示所有的候选子句

对;最后基于自注意力和互注意力的结果,通过多级注意力模块计算每一个子句对的特征表示,用于最终的预测.Song 等人[66]则将情感原因对的提取看成是从情感子句到原因子句的有向链接学习过程,并为此设计了一个端到端情感原因对提取模型(end to end emotion cause pair extraction, E2EECPE).在进行关系预测时,由于传统的方法是针对无向的情形而设计的,因此,文献[66]的作者通过双仿射注意力机制来为每一个节点生成"指向该节点"和"从该节点发出"2 种独立的表示,最后通过双仿射变换来构建一个非对称且方向依赖的子句对表示矩阵.

3) 基于解析式转移系统的子句配对

Fan 等人[23]提出 TransECPE(transition-based emotion cause pair extraction)模型,将情感原因对抽取任务转换成一个通过动作序列来构造有向图的过程,图中边的方向和标签表明子句之间的触发关系,有向图的构建依赖于一种新型的基于转移系统的解析器.该文定义了包含 6 个动作以及对应状态转换的动作集,然后利用栈、缓冲器来进行动作的选择和归约,其目标是寻找一个最优的动作序列.此外,针对某个子句可能既是情感子句也是原因子句的情形,该文还专门设计了一个二分类器进行判断,并定义了特定的动作来对其进行处理.此外,由于该模型是以从左到右的顺序来处理输入序列,从而减少了需要解析的潜在配对数量,从时间效率上来讲是一种线性的时间复杂度,大大低于将所有子句以笛卡儿积形式来处理时的情形.

4) 基于局部邻域搜索的子句配对

当人类在处理情感原因对提取任务时,情感子句和原因子句的提取及匹配是同时进行的,这一过程主要是通过局部搜索来完成,即当一个子句被标记为情感子句时,人们就会自然而然地在它的局部上下文中去寻找它对应的原因子句.局部搜索的好处在于可以避免一些在局部范围外的错误配对.此外,人们不仅会判断局部范围内的这一子句是否是原因子句,还会判断它是否与情感子句相匹配,这就可以避免在局部上下文范围内的错误匹配.通过对现有主流情感原因数据集的分析表明,情感原因子句大都位于情感子句的附近.

Cheng 等人[67]提出了一种对称式局部搜索网络(symmetric local search network, SLSN)模型,通过局部搜索同时进行情感原因子句检测和匹配.该模型由对称的情感子网络和原因子网络构成,每一个子网络由子句表示学习器和局部对搜索器(local pair search, LPS)组成,其中局部对搜索器是一种专门设计的跨子网组件,它能够对局部搜索的上下文范围进行限制.在局部搜索过程中,LPS 首先判断目标子句是否为情感子句;然后再判断本地上下文窗口内的每个子句是否是相应的原因;最后,模型为文档中的每个子句输出"对标签"(目标子句是否为情感/原因子句,其本地上下文窗口中的子句是否为对应的原因/情感子句),再基于该标签来提取最终的情感原因对.

Ding 等人[29]提出了一种基于多标签联合学习的情感原因对提取模型(emotion cause pair extraction based on multi-label learning, ECPE-MLL).该方法先假设文档中的所有子句都是情感子句,并以每个情感子句为支点引入一个面向情感的滑动窗口;然后在每个滑动窗口内使用一个多标签学习框架提取一个或多个可以与当前情感子句配对的原因子句,其中滑动窗口的设置本质上也是一种基于局部搜索的策略.

此外,也有不少研究利用图神经网络模型的特性来处理子句间的局部搜索问题.Wei 等人[68]提出 RANKCP(rank clause pair)模型,将情感原因对的提取看成是一种排序问题,他们将文档看成是全连接的子句图,利用图注意力网络模型来学习子句的表示,模型利用多个图注意层来加强对子句间相互作用的建模,并通过自适应地融合其他子句的信息来生成每个子句的表示,随后采用基于核的子句相对位置嵌入方案来进一步增强子句对的表示.在生成候选子句对时,文献[68]也采用添加约束的方式,将子句对中 2 子句的相对位置限制在一定的范围内,最终通过对候选子句对的预测分数进行排序来实现情感原因对的提取.Fan 等人[69]也是根据情感和原因之间的位置相关性,设计了一个范围控制器来缩小情感原因对的预测分布和真实分布之间的差异,进而将情感原因对的预测限制在一个高概率区域内.

Chen 等人[11]则考虑了另一种形式的局部邻域搜索问题,他们在对情感-原因的共现属性分析的基础上指出,在一个局部邻域中,如果一个候选对被检测为情感-原因对,其他候选对通常是非情感-原因对.因此,在建模上下文信息时,这种"对级别"的依赖关系也应该考虑进去.这里的"局部邻域"是指一个候选子句对的集合,这些子句对中情感候选子句

都是相同的,而原因候选子句彼此间的距离是比较近的.通过构造对图(pair graph)和对图卷积网络(PairGCN)来建模局部邻域候选对之间的依赖关系,图中的节点是候选的情感原因对,而节点间边则设计了3种类型的依存关系,即自循环边、原因候选子句间距离为1的边、原因候选子句间距离为2的边,每一种依赖关系都有其各自的传播上下文信息的方式.

5) 基于序列标注的子句配对

与大部分研究将情感原因对提取任务当成是子句级的二分类问题不同,Chen等人[70]和Yuan等人[24]将情感原因对提取任务转换成子句级的序列标注问题,并分别设计了不同的标注模式来对文档中的所有子句进行整体标注.其中文献[70]设计了因果标签集和情感标签集来对文档中的每一子句进行标注.因果标签集为四分类(O表示非情感原因句,E表示情感句,C表示原因句,B表示既是情感句也是原因句),情感标签集为传统的7类情感标签(O表示不含情感,H表示高兴,Sa表示伤心,A表示生气,D表示厌恶,Su表示惊讶,F表示害怕),这种标注方式更易于区分不同情感类型的情感原因对.例如标签(B-A)表示该子句既是情感子句也是原因子句,同时该子句对应的情感为生气.在此基础上设计了一个端到端的统一序列标注模型来进行情感原因对的提取,该模型包含1个卷积神经网络、2个BiLSTM网络和1个CRF,卷积神经网络用于编码邻域信息,2个BiLSTM网络分别用于预测因果标签和情感标签,CRF用于实现子句级的序列标注.

文献[24]则将子句之间关系直接以距离的方式编码到标签中.每一个子句的标签由类型标签和距离标签组成,类型标签只分为2种:"C"表示原因子句和"O"表示非原因子句;距离标签集为$\{-(n-1),\cdots,-1,0,1,\cdots,n-1,\perp\}$,距离标签的值代表了该子句与对应情感句之间的相对距离.例如,标签(C,2)表示该子句为原因子句,而情感子句位于其右边第2个子句.如果当前子句为非原因子句,那么距离标签就为特殊符号"$\perp$".基于该标注模式,先通过BERT来对子句进行编码,再通过BiLSTM来进一步对子句级的上下文进行建模,最后通过softmax对每一子句进行标签预测.该模型的优点在于它采用的是一种端到端的模式来自左向右地处理输入文本,其时间复杂度总是线性的,从而大大提高了模型的训练和推理速度.实验表明,该模型比当时的SOTA模型(文献[65]中提出的LAE-MANN)在训练阶段快了36%、在推理阶段快了44%,并且$F1$值也高了2.26个百分点.

4 其他情感原因相关研究

针对文本的情感原因研究除了第2节和第3节介绍的主流研究任务外,近2年也有一些学者从语义角色、条件因果、文本对话、社会情感、提取粒度、子句级序列标注等角度开展相关研究,为情感原因的研究提供了新的视角.

1) 语义角色.Oberländer等人[39]从语义角色的角度出发,分析了情感体验者、情感原因、情感目标这3种不同的情感语义角色是如何使得机器学习能够进行情感推理的.他们在5个数据集上进行情感分类的训练,训练时至少标注了其中的一种语义角色,同时以一种可控的方式来隐藏其他角色,以验证不同角色的作用.实验结果表明情感原因和情感目标携带了许多情感信息,而情感体验者则是一个干扰因子.同时发现,如果将情感角色的位置信息提供给模型会有更好的分类效果.

2) 条件因果.Chen等人[71]认为有些情感子句和原因子句之间的因果关系只有在特定语境条件下才会成立,因此他们定义了一个新任务,用于判断给定的文本对(情感子句,原因子句)在不同的上下文语境中是否存在有效的因果关系.同时,针对该新任务,在情感原因对数据集基础上通过人工标注和负采样的方式构建了新的数据集.

3) 文本对话.现有的情感原因研究大多以新闻或微博类文本为研究对象,而Poria等人[40]构建了一个名为RECCON(recognizing emotion cause in conversations)的数据集,用于提取对话中的情感原因.该文作者定义了原因块提取和情感原因推理2个子任务,并设计了相关的模型和评价指标来开展实验.该文作者还分析了文本块的数量、情感动力学、常识、复杂的共指关系等给情感推理带来的挑战.

4) 社会情感.Xiao等人[72]提出了社会情感原因提取(social emotion cause extraction, SECE)任务.社会情感是指读者在阅读某些文本类的文档时产生的情感,该任务考察的是读者层面的情感而非作者层面的情感.该文作者提出了一种词汇增强的记忆网络模型来应对这一新的任务,模型主要通过构建的情感诱发词典和情感记忆的动态机制来实现情感原因的提取,该机制可以在每个子句中迭代地

学习特定的情感相关信息,并在训练过程中动态更新.

5) Span 块提取.Li 等人[73]认为现有的情感原因研究大多局限于子句层面的二分类,但并非子句中的所有词都能表达有用的情感原因信息,因此他们提出了更细粒度的情感原因块(span)提取任务,并结合情感感知注意力、上下文感知注意力和位置感知注意力等机制,构造了原因块提取和原因块分类两大模块来进行情感原因的提取.

6) 子句级序列标注.Xiao 等人[74]认为现有的研究大都只针对子句的情感依赖性语言表征进行了建模,而忽略了包括因果指示符在内的子句的情感独立性特征.因此,该文作者提出一种上下文多视图注意力网络(context-aware multi-view, COMV)用于情感原因提取,并将任务转换成子句级的序列标注问题,即将文档中的所有子句视为一个整体,共同预测这些子句的标签.模型主要通过注意力机制来分别学习以情感导向为视角的情感依赖特征表示和以子句导向为视角的情感独立特征表示.Liang 等人[75]同样将情感原因提取看成是子句级的序列标注问题,提出一种基于注意力的 BiLSTM-CRF 模型.模型首先通过 BiLSTM 来分别捕获上下文信息以及情感表达和候选子句的潜在语义关系;随后设计 2 种注意力机制来分别编码情感表达与候选子句、相对位置和候选子句之间的相互影响;最后,将获得的子句表示送入条件随机场进行子句标注.

5 情感原因数据集

5.1 中文数据集

由于情感原因数据集在标注过程中需要耗费大量的人力,因此公开的数据集并不多.Lee 等人[5]于 2010 年构造了第 1 个用于情感原因分析的中文数据集,该数据集中包含了 6 058 个句子条目,这些条目是基于高兴、伤心、恐惧、生气、吃惊这五大类情绪而提取的,其中 72%的条目中有表达明确的情感,在这些表达情感的条目中 80%都包含了情感原因.

Gui 等人[15]在 2016 年发布了一个基于新浪新闻的情感原因数据集,这个数据集包括 2 105 篇文档,共 11 799 个子句,2 167 个情感原因子句,其中,包含一个原因子句的文档 2046 篇,包含 2 个原因子句的文档 56 篇,包含 3 个原因子句的文档 3 篇.该数据集是目前唯一被公开发布的中文情感原因数据集,近年来的许多研究工作都基于这个数据集进行.该数据集遵循 W3C 的情感标记语言(emotion markup language)格式进行标注,主要标签及含义如表 5 所示:

Table 5 Label Interpretation for Emotion Cause Corpus

表 5 情感原因语料中标签及含义说明表

标签名	含义
emotion *id*	样本的编号
category *name*	情绪类别,如高兴、伤心、害怕、生气、吃惊等
value	情绪类别的编号
clause *id*	子句的编号
cause	子句是否为原因子句,是则值为 Y,否则为 N
keywords	子句是否为情感子句,是则值为 Y,否则为 N
text	子句的文本内容
cause *id*	原因子句的编号
type	原因的类别,是动词类则为 v,名词类则为 n
begin	原因字符串在原因子句中的开始位置
length	原因字符串的长度
key-words-begin	情感关键词在情感子句中的开始位置
keywords-length	情感关键词的长度
emotionml	一个样本标注的结束标记

以文本“劝说过程中,消防官兵了解到,该女子是由于对方拖欠工程款,家中又急需用钱,生活压力大,无奈才选择跳楼轻生.”标注示例如下:

```
〈emotion id="4"〉
  〈category name="sadness" value="4"/〉
  〈clause id="1" cause="N" keywords="N"〉
    〈text〉
      劝说过程中,
    〈/text〉
  〈/clause〉
  〈clause id="2" cause="N" keywords="N"〉
    〈text〉
      消防官兵了解到,
    〈/text〉
  〈/clause〉
  〈clause id="3" cause="Y" keywords="N"〉
    〈text〉
      该女子是由于对方拖欠工程款,
    〈/text〉
    〈cause id="1" type="v" begin="6"
     length="7"〉
```

```
        对方拖欠工程款
      〈/cause〉
    〈/clause〉
    〈clause id="4" cause="Y" keywords="N"〉
      〈text〉
        家中急需用钱,
      〈/text〉
      〈cause id="2" type="v" begin="0" length="7"〉
        家中又急需用钱
      〈/cause〉
    〈/clause〉
    〈clause id="5" cause="Y" keywords="N"〉
      〈text〉
        生活压力大,
      〈/text〉
      〈cause id="3" type="n" begin="0" length="5"〉
        生活压力大
      〈/cause〉
    〈/clause〉
    〈clause id="6" cause="N" keywords="Y"〉
      〈text〉
        无奈才选择跳楼轻生
      〈/text〉
      〈keywords key-words-begin="0" keywords-length="2"〉
        无奈
      〈/keywords〉
    〈/clause〉
  〈/emotion〉
  〈/emotionml〉
```

对该数据集的其他统计信息如表 6、表 7 所示，其中表 6 表示情感类型的分布情况，表 7 表示情感原因子句和情感子句的相对位置关系.

此外，针对“情感原因对提取”这一新任务，Xia 等人[25]对上述数据集进行了整合，并形成了适合于该新任务的情感原因对数据集.Chen 等人[71]从为原因子句构造不同上下文语境的角度出发，在情感原因对数据集基础上通过人工标注和负采样的方式构建了条件因果情感原因数据集.

Table 6 Distribution of Emotions

表 6 情感类型的分布情况

情感类型	数量	占总数的百分比/%
高兴	544	25.83
伤心	567	26.94
害怕	379	18.00
生气	302	14.35
厌恶	225	10.69
吃惊	88	4.18

Table 7 Relative Position of Emotion Cause Clause and Emotion Clause

表 7 情感原因子句和情感子句的相对位置情况

位置	数量	占总数的百分比/%
前 3 个子句	37	1.71
前 2 个子句	167	7.71
前 1 个子句	1 180	54.45
在同一个子句	511	23.58
后 1 个子句	162	7.47
后 2 个子句	48	2.22
后 3 个子句	11	0.51
其他位置	42	1.94

5.2 英文数据集

虽然在情感计算领域有许多对情感类别进行标注的英文数据集，但专门针对情感原因任务而设计的英文数据集并不多.Gao 等人[79]在 2017 年 NTCIR13 会议上专门为情感原因提取子任务发布的数据集包括中文数据集和英文数据集，其中中文数据集即为 5.1 节中提到的新浪新闻数据集，英文数据集的语料则取材自英文小说.该英文数据集包括 2 156 篇文档的 16 259 条子句，其中原因子句 2 421 条，包含 1 个、2 个和 3 个原因子句的文档分别为 1 949 篇、164 篇和 32 篇，其标注方法与表 5 一致.

Ghazi 等人[41]利用 FrameNet 的情感导向框架自动建立一个包含情感和情感原因标注的英文数据集，其中包括 820 条包含情感原因的情感句子和 1 594 条未包含情感原因的情感句子.该数据集的标注相对比较简单，每个句子的开头标注具体的情感类别，如果句子中包含有情感原因，则用〈cause〉〈\cause〉的形式标注对应的文本块，示例如下.

1) 含情感原因句的标注示例

〈happy〉

These days he is quite happy 〈cause〉 travelling by trolley〈\cause〉.

〈\happy〉

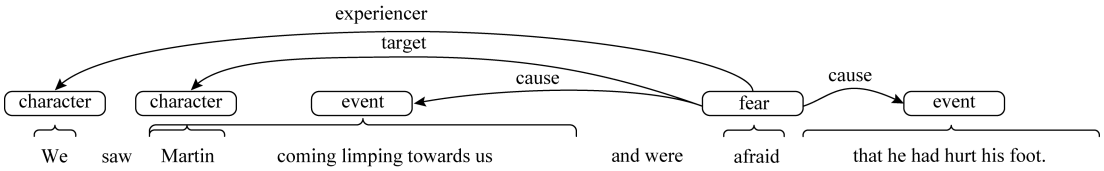
2) 不含情感原因句的标注示例

〈anger〉

Bernice was so angry she could hardly speak.

〈\anger〉

考虑到传统的情感分析任务仅仅检测文本中所表达的情感,这其实是一种简化,会导致对分析结果的一种过度概括的解释,因为它并没有考虑到谁经历了一种情感以及为什么有这种情感.在心理学看来,情感在角色和他们所参与的事件之间的互动中起着至关重要的作用,因此针对没有专门的语料库来捕捉这种互动的情况.Kim 等人[42]于 2018 年提供了一个来自于 Gutenber 项目的基于关系情感标注(relational emotion annotation, REMAN)的公共可用语料库.该语料库是针对小说文本中情感和实体之间的关系进行语义角色标注,由 1 720 条句子三元组构成,其中 1 115 条是包含情感的.三元组的中间元素为包含情感的句子,第 1 个元素为该情感句的上一句,第 3 个元素为该情感句的下一句.数据集只对三元组中的中间句子进行标注,其前后句子只是用于辅助标注者更好地理解情感句的上下文.标注者从情感的具体类别、情感体验者(experiencer)、情感目标(target)以及情感原因(cause)等方面进行标注,如图 11 所示.其中对情感原因的标注主要是将句子中的情感触发短语、实体/事件标注出来.REMAN 语料中的标签及含义如表 8 所示.



注:character 为角色,event 为事件,fear 为情感标签,target 为目标对象,cause 为原因,experiencer 为情感体验者.

Fig. 11 An annotation example for REMAN

图 11 REMAN 标注示例

Table 8 Label Interpretation for REMAN

表 8 REMAN 语料中标签及含义说明表

标签名	含义
span	一个文本块(可能是情感类型、角色或事件)
annotatorId	标注者的编号
cbegin	文件块的开始位置
cend	文本块的结束位置
modifier	表示情感的修饰级别的强弱
type	文本块的类型(如某种具体情感、角色、事件等)
relation	表示某种关系(如体验者关系,目标关系,原因关系等)
relation_id	关系编号
left	关系左边的实体的文本块起始位置
right	关系右边的实体的文本块结束位置
source_annotation_id	关系源标注编号
target_annotation_id	关系目标标注编号
type	关系的具体类型(如体验者、原因、目标)

图 11 所示文本"We saw Martin coming limping towards us and were afraid that he had hurt his foot."中涉及情感原因相关的部分标注示例如下:

〈spans〉

〈span *annotation_id*="10001" *annotatorId*="A" *cbegin*="48" *cend*="54" *type*="fear"〉afraid〈/span〉

〈span *annotation_id*="10002" *annotatorId*="A" *cbegin*="7" *cend*="39" *type*="event"〉Martin coming limping towards us〈/span〉

〈span *annotation_id*="10003" *annotatorId*="A" *cbegin*="55" *cend*="80" *type*="event"〉that he had hurt his foot〈/span〉

〈/spans〉

〈relations〉

〈relation *annotatorId*="A" *left*="7" *relation_id*="20001" *right*="54" *source_annotation_id*="10001" *target_annotation_id*="10002" *type*="cause"〉Martin coming limping towards us [EVENT]… afraid [FEAR] 〈/relation〉

〈relation *annotatorId*="A" *left*="48" *relation_id*="20002" *right*="80" *source_annotation_id*="10001" *target_annotation_id*="10003" *type*="cause"〉afraid [FEAR] … that he had hurt his foot [EVENT]

〈/relation〉

〈/relations〉

(注:为了便于读者理解,此处对各 ID 编号进行了简化,在涉及字符相对位置时未加入前一子句的长度,只涉及在当前句子中的相对位置.)

其中涉及到的情感类型有"fear"、事件有"Martin coming limping towards us"和"that he had hurt his foot",2 个事件均被标注为情感"fear"的原因.此外,由于该语料库是面向情感分析的整个体系进行的标注,并非只针对情感原因这一个任务,因而目前并未见有基于该语料库开展情感原因提取研究的文献.

Bostan 等人[80]于 2020 年发布了一个包含 5 000 条新闻标题的"GoodNewsEveryone"情感角色英文数据集,该数据集对每个新闻标题中的情感体验者、线索词、情感目标和情感原因等与情感有关的语义角色进行了标注.这是目前为止与情感原因有关的最大规模的英文数据集.由于其语料来源是新闻标题,因此每条文本的长度并不大,平均只有 13 个单词.

此外,Balahur 等人[43]构建了一个 EmotiNet 知识库,用于表示和存储对现实生活环境的情感反应,为情感原因的研究提供了很好的语义资源.该知识库主要基于 ISEAR(international survey on emotion antecedents and reactions)语料库中一系列关于情感状况的自我报告,通过对示例样本进行聚类,并使用语义角色提取三元组信息.

5.3 其他语种数据集

除了英文数据集外,文献[44]针对意大利语料开展情感原因提取的研究.该文作者首先在基于众包的方式获取的新闻语料关键词上进行情感状态和情感原因事件提取;随后构建了包含 32 525 个单词的情感上下文语料库,并分别对情感关键词、情感原因短语、情感原因词元进行了标注,其对应的标签分别为〈emotionWord〉〈causePhrase〉〈causeEmotion〉.语料中共包含 356 个情感词、104 个情感原因短语、84 个情感原因词元.

文献[36]则采用了自举的技术对日文语料进行情感原因数据集的构建.其语料主要来自报刊文章、门户网站的新闻、问答网站的帖子.每种语料均采样了约 100 000 条包含情感词的句子,然后在此基础上,以少量人工标注的情感原因线索短语为种子,采用自举技术通过迭代方式不断提取新的线索短语,并最终达到样本增广效果.

6 情感原因提取的评测

6.1 评测指标

情感原因提取任务的评测指标最早是由 Lee 等人[3]提出的,评估的主要指标为传统的准确率(precision, P)、召回率(recall, R)和 $F1$ 值,计算公式分别为:

$$P=\frac{\sum\limits_{S_i\in GF}\sum\limits_{em_j\in S_i}Score(SCList_j,GCList_j)}{\sum\limits_{S_i\in SF}\sum\limits_{em_j\in S_i}1},\tag{2}$$

$$R=\frac{\sum\limits_{S_i\in GF}\sum\limits_{em_j\in S_i}Score(SCList_j,GCList_j)}{\sum\limits_{S_i\in GF}\sum\limits_{em_j\in S_i}1},\tag{3}$$

$$F1=\frac{2\times P\times R}{P+R},\tag{4}$$

其中,GF 表示人工标注的原因数据集,SF 表示系统提取出来的原因数据集,S_i 表示某个句子,em_j 表示情感关键词或情感子句,$GCList_j$ 和 $SCList_j$ 分别是对应于 em_j 的人工标注情感原因列表和系统提取出来的情感原因列表.$Score$ 是记分函数,用于对系统结果和人工结果进行对比.由于一种情感可能由多个原因触发,记分函数计算的是 2 个原因文本列表 $GCList_j$ 和 $SCList_j$ 中原因子句之间的重叠程度,即原因子句之间的字符串匹配程度.考虑到原因提取的复杂性,文献[3]中使用了 2 种匹配方案:方案 1 中,只要系统结果和人工结果中的 2 个原因存在交集,这 2 个原因子句就算匹配成功;方案 2 则更为严格,它考虑了系统结果和人工结果中原因子句之间重叠文本的长度.

以上评测指标是针对提取粒度为词的情况,近年来的情感原因提取任务大都是以子句为提取粒度,此时评测指标中对于准确率和召回率的计算则依据文献[15]简化为:

$$P=\frac{\sum\limits_{correct\_causes}1}{\sum\limits_{proposed\_causes}1},\tag{5}$$

$$R=\frac{\sum\limits_{correct\_causes}1}{\sum\limits_{annotated\_causes}1},\tag{6}$$

其中,$correct\_causes$ 是系统提取正确的原因子句,$proposed\_causes$ 则是系统提取出来的原因子句,$annotated\_causes$ 表示人工标注的原因子句.

针对"情感原因对"提取这一新任务,其测评

对象由单个原因子句扩充成了情感及情感原因组合而成的一对子句.句子中有 n 个原因子句,该任务需要提取出 n 个子句对.目前,该任务上的评测是以子句对为粒度的,只有子句对提取正确才算是匹配成功,并没有对子句对中单独的某一项的提取结果进行评测.因此其评测指标与子句级情感原因提取的评测指标是一样的,即在形式与式(5)、式(6)一致,只是将公式中相关参数的下标由 $correct\_causes$, $proposed\_causes$, $annotated\_causes$ 分别替换成 $correct\_pairs$, $proposed\_pairs$, $annotated\_pairs$,分别表示系统提取正确的情感原因对、系统提取出来的情感原因对和人工标注的情感原因对.

6.2 评测结果分析

表 9 给出了在同一数据集(Gui 等人[15]在 2016 年发布的情感原因数据集)上不同的情感原因提取方法的实验结果.表 9 显示,在情感原因提取任务上,基于深度学习模型的各项指标总体上均高于传统的基于规则和基于统计机器学习的方法.

Table 9 Comparison of Performance with Existing Models for ECE

表 9 情感原因提取各模型性能对比

模型名称	文献	P/%	R/%	$F1$/%	主要方法
RB	[4,14]	67.47	42.87	52.43	规则
CB	[14,44]	26.72	71.30	38.87	知识库
RB+CB	[14]	54.35	53.07	53.70	规则+知识库
RB+CB+ML	[4,14]	59.21	53.07	55.97	规则+知识库+机器学习
K_{new*M}	[15]	66.73	68.41	67.56	SVM
B_{*multi}	[18]	75.63	73.52	74.56	SVM+装袋
ConvMS-Mement	[17]	70.76	68.38	69.55	卷积多槽记忆网络
LambdaMART	[45]	77.20	74.99	76.08	学习排序算法
CANN	[53]	77.21	68.91	72.56	CNN+注意力
HCS	[55]	73.88	71.54	72.69	CNN+注意力
MANN	[54]	78.43	75.87	77.06	CNN+注意力
RTHN	[27]	76.97	76.62	76.77	RNN+Transformer
PAE-DGL	[26]	76.19	69.08	72.42	BiLSTM+注意力
RD-HAN	[22]	77.06	72.03	74.46	规则知识蒸馏
RHNN	[21]	81.12	77.25	79.14	注意力+正则化
ExSenti-ECD	[60]	87.69	70.62	78.23	外部知识
MBiAS	[56]	73.6	72.59	73.09	双向注意力流
Span2Clause	[73]	76.25	79.73	77.95	多任务
EACN	[47]	72.15	73.76	72.94	外部知识+注意力
FSS-GCN	[52]	78.61	75.72	77.14	GCN
COMV	[74]	78.23	76.32	77.21	BiLSTM+注意力
* ECPE	[25]	70.41	60.83	65.07	多任务
* ECPE-2D	[28]	73.36	69.34	71.23	2D 建模+Bert
* TransECPE	[23]	75.62	64.71	69.74	转移系统+Bert
* RANKCP	[68]	69.27	67.43	68.24	图注意力+Bert
* PairGCN-BERT	[11]	79.07	69.28	73.75	局部搜索+GCN

注:标"*"的模型是情感原因对提取模型在情感原因提取这一任务上的表现,和其他情感原因提取模型相比,该任务采用的数据集并未对情感子句中的具体情感进行标注,因而任务的难度更高.

基于规则的方法由于规则设计的复杂性及覆盖率低的影响,其效果并不理想,即使融入了知识库以及传统的机器学习方法,其最好的 $F1$ 效果也未达到 60%;基于统计机器学习的方法中目前效果最好的模型主要采用了 SVM 算法,这其中核函数的设计是影响效果的关键.

虽然总体上看基于深度学习的方法性能最优,但情感原因提取任务并非普通的文本分类任务,如果仅仅依靠时下主流的基础深度学习模型,则优势并不明显.例如文献[18]采用传统的 SVM 算法并结合装袋(bagging)的分类器技术同样能取得与一些主流深度学习模型相当的效果.从各模型的实验

效果看,大部分的深度学习模型在实验效果中差距并不大,如文献[26,53,55]等,其 $F1$ 值均在 72%左右,而文献[22]由于在深度学习模型中融合了传统的规则知识,因此在 $F1$ 值上有近 2 个百分点的提升.目前情感原因提取任务取得 SOTA 效果的是文献[21]提出的模型,其 $F1$ 值达到了 79.14%.该模型最大的特点在于其引入了情感词典知识以及子句的相对位置这 2 项外部知识来对深度学习模型的损失函数进行约束.这表明,要将深度学习模型与传统的方法相结合才能在情感原因提取问题上取得较好的提升.

由表 9 可以看出,目前情感原因提取任务的最好模型在 $F1$ 值上仍然没有能够超过 80%.这说明,现在的研究方法在对性能的提升上已经遇到了瓶颈,未来必须要有更具创新的思路才可能有新的突破.

表 10 是"情感原因对"提取任务上不同模型的实验结果,使用的数据集依然是 Gui 等人[15]发布的数据集,只是按情感原因对抽取任务进行了重新整理.

Table 10 Comparison of Performance with Existing Models for ECPE

表 10 情感原因对提取各模型性能对比

模型名称	文献	P/%	R/%	$F1$/%	子句对的构建方式
ECPE-Inter-EC	[25]	67.21	57.05	61.28	2 阶段笛卡儿积
Inter-ECNC	[62]	66.01	57.34	61.38	2 阶段笛卡儿积
MAM-SD	[63]	69.63	57.99	63.20	2 阶段笛卡儿积
TransECPE	[23]	73.74	63.07	67.99	解析式转移系统
ECPE-2D	[28]	72.92	65.44	68.89	矩阵变换
MTNECP	[64]	69.44	60.17	64.40	矩阵变换
LAE-MANN	[65]	71.1	60.7	65.5	矩阵变换
E2EECPE	[66]	64.91	61.95	63.15	矩阵变换
SLSN-U	[67]	68.36	62.91	65.45	局部搜索
ECPE-MLL-ISML	[29]	70.9	64.41	67.40	局部搜索
RANKCP	[68]	69.10	62.54	65.62	局部搜索
RHNSC	[69]	69.56	58.71	63.57	局部搜索
PairGCN-BERT	[11]	76.92	67.91	72.02	局部搜索
IE-CNN+CRF	[70]	71.49	62.79	66.86	序列标注
ECPE-SL	[24]	72.43	63.66	67.76	序列标注

虽然情感原因对提取任务是 2019 年提出,但近 2 年来已涌现出 10 余个针对该任务的模型.表 10 中除了前 3 个模型是采用"2 阶段"的训练方式外,其余模型均为统一的端到端的训练方式.从模型效果来看后者明显高于前者,这也再次体现了端到端训练方法的优点.这些端到端的模型大部分都采用了多任务联合学习的思想,模型间的主要区别在于子句对的构建及处理方式上.

从表 10 可以看出,虽然各模型在最终的实验效果上各有千秋,但局部搜索的方式总体上还是优于其他方式,这主要是由于实验数据集中原因子句和情感子句间的相对位置存在很明显的特点,即大部分原因子句都离情感子句比较近,而局部搜索的配对方式能更好地建模这 2 种子句间的这种特殊位置关系.

文献[11]利用图卷积神经网络来建模子句对之间的关联,并取得了该任务上的 SOTA 效果,这也进一步反映出图神经网络在建模邻域关系中的优势.此外,文献[24,70]中基于序列标注的模型也体现出了一定的竞争力,这主要是由于文档中各子句之间的标签本身存在一种相互制约的关系,传统的子句级二分类方法大多是对每一个子句进行独立的预测,而序列标注是从整体上对所有子句进行标注,因而能够更好地捕获这种子句标签之间的依赖关系.

当然,从表 10 的模型效果也可以看出,情感原因对提取任务目前的 SOTA 效果仅为 72%左右,与传统的情感原因提取任务上近 80%的 SOTA 效果相比仍存在不小差距,这主要是由于前者在提取时没有给定明确的情感子句信息.尽管情感原因对提取任务的应用场景更广泛,但其提取的难度高于情感原因提取,存在更大的挑战.

7 展 望

文本情感原因提取是情感计算的一个新兴方向,得益于近年来自然语言处理和深度学习技术的飞速发展,该领域也越来越受到学者的关注,并产生了较为丰富的成果,特别是近 2 年在如 ACL 等自然语言处理国际顶会上均有不少关于情感原因提取的文章.

早期学者提出的基于规则的方法充分利用了语言学机制,规则清晰易懂,准确率较高,同时也为后期基于统计的机器学习方法和基于深度学习的方法提供了很好的理论基础.此后学者利用特征工程从统计机器学习的角度出发,设计了大量有效的情感原因提取特征,提高了该任务的准确率和覆盖率.然而,基于规则的方法和基于统计的方法均需要消耗大量的人力成本,较难适应新的领域或不同的数据.

深度学习技术在文本情感分析问题上获得了成功应用,启发研究者通过构造端到端的深度神经网络模型来解决情感原因提取问题,像注意力机制、

多任务联合学习、知识蒸馏等技术均在该领域得到了有效的应用,这些深度学习技术的应用在很大程度上降低了传统方法所带来的人力消耗,促进了情感原因提取技术的发展.然而,从文献中的实验结果看,情感原因提取的研究仍然还有提升空间,一些最新的深度学习成果仍未能广泛地应用于该领域.情感原因提取工作所面临的挑战主要体现在7个方面:

1) 情感原因语料库较少,涵盖领域不够丰富

近年来使用较多的情感原因数据集只有文献[15]中发布的中文数据集,但该数据集中标注语料仍相对较少,只有2000余条,若采用一些复杂的深度网络,算法较容易产生过拟合,并且由于样本的不均衡或者样本数太少,导致个别情感原因无法被很好地学习和表达,影响提取的准确率.此外,该数据集中的样本都是新闻类的长文本,其文本特点并不适用于时下流行的在线社交短文本类的情感原因提取工作,也较难迁移到其他语种上.

情感原因本质上是触发情感的某种事件或者场景,而这些事件或场景的种类是五花八门的,在生活中各个领域发生的事件都有可能触发人的某类情感,例如娱乐事件、体育事件、社会事件、政治事件等.此外,某类事物本身也可能触发人的情感,就像平时我们说的"触景生情"一样.而现有数据集中采集的数据只是来自于社会新闻领域,在没有足够丰富的语料的情况下,仅仅依靠现有数据来训练是很难达到理想效果的,哪怕在当前数据集中表现良好,也难以泛化到其他领域.标注语料无论在种类还是数量上的匮乏都在很大程度上给方法设计带来很大的限制.

因此,情感原因标注数据集的扩展和新建仍然是未来开展情感原因提取任务中一项十分重要的基础工作.

2) 情感语义特征的挖掘仍不够充分

从文献看,通过深度学习的方法自动抽取文本的特征已成为主流,其中词向量的表示和预训练语言模型便是深度学习在自然语言处理中的一大研究成果.词向量是词的一种分布式表示,向量间的相对相似度和语义相似度是相关的.然而,传统的词向量是根据上下文词语学习获得的,只包含语义和语法信息,而词语的情感信息对于情感分析任务至关重要,现有大多数词向量学习方法忽略了词语的情感信息,不能很好地解决情感分类以及情感原因提取等任务.

同样地,目前一些主流的预训练语言模型在获得句子的向量表示时并不能很好地反映出该句子中所蕴含的情感倾向.并且,对于情感原因子句来说,它虽然本身并不包含情感词,但该子句能够触发人类情感,因此,理想的预训练模型应该将这种虽非情感表达但又能触发某种情感的子句进行学习并表示出来,例如从"他中彩票了"这一句子中学习出其包含的触发正向情感的语义,从"他参加比赛输了"学习出其包含的触发负向情感的语义.

除此之外,不同领域的情感及情感原因表达也存在差异,因此如何将特定领域中文本的情感信息融入到词向量和句子向量中,从而提供更深层的语义表征,这值得进一步探索.

3) 情感和原因的内在因果分析不够深入

对现有情感原因语料库的统计发现,每篇文档大都包含多个子句,但情感原因子句则相对较少,通常只有一句,这要求模型能够很好地建模情感和情感原因之间的内在关系.虽然现有的基于深度学习的方法在提取深层语义特征方面有一定的优势,但情感与情感原因之间的因果关联和普通事件及其原因之间的因果关联存在一定差别,它和情感本身有很大的关系.现有的一些情感原因提取模型经常错误地把普通事件的原因识别成情感的原因,或者根本就无法找出情感原因,这其中除了训练数据不充分外,还有一大原因就在于对情感及其原因之间的因果分析还不够深入,二者之间的关联还无法准确提取.

普通的因果关系可以通过一些显式的因果连词来发现,例如"因为""所以""由于"等,但情感和情感原因之间有时候并不存在这种显式的因果连接.人类在判断触发情感的事件时经常依据的是一种常识信息.就像打比赛输了就会不开心,赢了就会高兴,被人打了就会伤心,这些都是很自然的一种情感常识.除此之外,对于情感原因提取、情感分类、情感角色提取等目的不同但存在相关性的任务,它们彼此之间能否通过构建多任务模型来挖掘内在的情感因果关联,这些都有待学者进一步的研究.

4) 隐式情感语境下情感原因提取有待研究

目前的情感原因提取研究中,其数据集中每个样本都是包含情感子句的,各类模型都有借助情感子句特别是情感子句中的情感词来辅助情感原因子句的提取.然而,情感原因提取的另一挑战在于很多语句中可能并没有用明显的情感词来表达情感,此时传统的借助情感词来提取情感原因的方法就行不通了.例如"这家店的装修风格让人有一种活在诗里

的感觉.",这其中"让人有一种活在诗里的感觉"表达的其实是一种褒义的情感,而造成这种情感的原因就是"装修风格",很明显这种表达缺少显式情感词作为情感引导,且表达更为含蓄和隐晦.如何有效地挖掘隐式情感和情感原因之间的关联,进而把情感原因"装修风格"提取出来,这一任务难度显然更具挑战性.

虽然已有的情感原因对抽取方法不需要对情感子句中的情感进行显式标注,但该数据集中的情感子句里面还是包含了显式的情感关键词,因此已有的方法能否很好地应对这种无显式情感关键词的情形还有待进一步验证.虽然可以将现有的隐式情感分析手段结合到情感原因提取的任务上,以流水线的方式先进行隐式情感的提取,然后再以提取出的情感为基础进行情感原因的提取,但这种 2 步式的方式会带来误差的传递以及较大的计算代价.

此外,现有的隐式情感分析研究一般是指用户在文本中有表达出情感只是未使用明显的情感词,但情感原因并非是用户针对某件事物所表达的观点,而是它可能就是事物本身,也就是说情感原因子句很可能只是一种对客观发生的事件的一种描述,它本身是不带有任何主观色彩的.例如"我走在路上滑了一跤"这句话本身并不包含任何主观情绪在里面,但是我们通过常识可以判断"滑了一跤"这种事件会触发人的某种不愉快的情感.这也就是上文提到的对文本情感语义的深层次挖掘问题.如何更有效地处理该问题也是未来很有挑战性的一项工作.

5) 自然语言处理技术的发展对情感原因研究带来的机遇

现有的情感原因模型在技术方面主要以 CNN,LSTM,GRU,Transformer 等深度学习模型为基础,同时配合注意力机制、知识蒸馏技术等.近几年自然语言处理技术的飞速发展也给情感原因的研究带来了许多新的机遇,像图神经网络、知识图谱、对抗学习、少样本学习等新技术在自然语言处理方面的广泛应用都为情感原因的研究提供了新的解决思路.首先,情感原因的提取是需要借助外部领域知识的,例如情感方面的常识以及语言学的知识.而知识图谱能够将网络上的信息和数据资源关联为语义知识,使得网络的智能化水平更高,更加接近于人类的认知思维.如果在现有的通用预训练语言模型基础上再融合常识及领域知识图谱,则可以更有效地对文本语义进行表示.

其次,情感和情感原因之间本质上是一种特殊的因果关系,现有的模型也是期望挖掘出子句和子句间的这种因果关联.从传统的基于规则的方法可知,文档语篇信息的句法结构和语篇关系在情感原因提取中是十分重要的.通过图的方式对关系进行挖掘和建模是一种朴素的想法,因此时下较为流行的图神经网络也是一种很值得尝试的方案,例如通过将子句建模成节点,然后构建图神经网络来识别节点间的关系.

最后,针对现有情感原因标注数据太少的问题,也可以尝试利用少样本学习、数据增强或者伪标签技术等方式来解决.

6) 情感原因提取任务的新挑战

随着情感分析研究的不断深入,也会对情感原因的研究提出新的需求,例如由现有子句级的"粗粒度"分析转向"短语级"或者"文本块"级的"细粒度"分析.同时,情感原因研究也只是情感分析的一部分,从情感认知的角度来看,一个完整的情感表达是涉及情感、情感主体、情感目标、情感原因、情感结果等多种语义角色的,因此,未来对情感中各种语义角色的研究也会给情感原因及情感分析的研究带来新的机遇和挑战.

此外,现有的情感原因提取都是针对个体的,但对决策者来说,群体的情感及其原因才更具参考价值,因此群体情感原因提取也是未来情感原因研究的一个新方向.

7) 情感原因提取的应用

情感原因提取作为一种更深层次的情感挖掘,不仅能够丰富情感计算领域的研究成果,为情感分析提供新的研究方向,而且也能为人工智能和自然语言处理的一些分支提供有益帮助.例如,在商品推荐领域,如果在进行商品推荐算法设计时,能准确地定位用户对某商品喜恶的具体原因,就能更有针对性结合这些原因来进行商品的推荐.在人机对话领域,现有的一些人机对话技术能够识别出用户在对话过程中的情感变化,并通过该情感来引导文本的生成,如果能够进一步提取出用户表现该情感的原因,就能在生成回复时结合具体的原因事件提供更具方向性的文本回复.

最后,心理学、语言认知学、社会学领域的研究成果能够为情感原因研究提供更为丰富的理论基础,而情感原因的研究也可以反过来促进这些领域的研究和发展.例如,利用模型自动提取大规模文本中的情感原因,为探索心理学、语言认知学和社会学规律提供大规模样本.

作者贡献声明:邱祥庆负责资料收集、研究方案的构思和设计、论文撰写及修订;刘德喜提供研究思路、论文组织结构的设计、论文审阅及修订、全过程监督;万常选、刘喜平、廖国琼负责论文审阅及修订;李静负责论文图表及参考文献的核实及修订.

参 考 文 献

[1] Li Ran, Lin Zheng, Lin Hailun, et al. Text emotion analysis: A survey [J]. Journal of Computer Research and Development, 2018, 55(1): 30-52 (in Chinese)
(李然, 林政, 林海伦, 等. 文本情绪分析综述[J]. 计算机研究与发展, 2018, 55(1): 30-52)

[2] Lee S Y M, Chen Ying, Huang Churen. Cause event representations for happiness and surprise [C] //Proc of the 23rd Pacific Asia Conf on Language, Information and Computation. Hong Kong: City University of Hong Kong Press, 2009: 297-306

[3] Lee S Y M, Chen Ying, Huang Churen. A text-driven rule-based system for emotion cause detection [C] //Proc of the 11th NAACL HLT Workshop on Computational Approaches to Analysis and Generation of Emotion in Text. Stroudsburg, PA: ACL, 2010: 45-53

[4] Chen Ying, Lee S Y M, Li Shoushan, et al. Emotion cause detection with linguistic constructions [C] //Proc of the 23rd Int Conf on Computational Linguistics. Beijing: Tsinghua University Press, 2010: 179-187

[5] Lee S Y M, Chen Ying, Li Shoushan, et al. Emotion cause events: Corpus construction and analysis [C] //Proc of the 7th Int Conf on Language Resources and Evaluation. Paris: ELRA, 2010: 1121-1128

[6] Lee S Y M, Chen Ying, Huang Churen, et al. Detecting emotion causes with a linguistic rule-based approach [J]. Computational Intelligence, 2013, 29(3): 390-416

[7] Lee S Y M, Li Shoushan, Huang Churen, et al. Detecting emotion cause with sequence labeling model [J]. Journal of Chinese Information Processing, 2013, 27(5): 93-99 (in Chinese)
(李逸薇, 李寿山, 黄居仁, 等. 基于序列标注模型的情绪原因识别方法[J]. 中文信息学报, 2013, 27(5): 93-99)

[8] Cheng Xiyao, Chen Ying, Cheng Bixiao, et al. An emotion cause corpus for chinese microblogs with multiple-user structures [J]. ACM Transactions on Asian and Low-Resource Language Information Processing, 2017, 17(1): 1-19

[9] Chen Ying, Hou Wenjun, Cheng Xiyao, et al. Joint learning for emotion classification and emotion cause detection [C] //Proc of the 2018 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2018: 646-651

[10] Chen Ying, Hou Wenjun, Cheng Xiyao. Hierarchical convolution neural network for emotion cause detection on microblogs [C] //Proc of the 27th Int Conf on Artificial Neural Networks. Berlin: Springer, 2018: 115-122

[11] Chen Ying, Hou Wenjun, Li Shoushan, et al. End-to-end emotion-cause pair extraction with graph convolutional network [C] //Proc of the 28th Int Conf on Computational Linguistics. New York: ICCL, 2020: 198-207

[12] Gui Lin, Yuan Li, Xu Ruifeng, et al. Emotion cause detection with linguistic construction in Chinese Weibo text [C] //Proc of the 3rd CCF Int Conf on Natural Language Processing and Chinese Computing. Berlin: Springer, 2014: 457-464

[13] Yuan Li. Research on emotion cause detection towards text [D]. Harbin: Harbin Institute of Technology, 2014 (in Chinese)
(袁丽. 基于文本的情绪自动归因方法研究[D]. 哈尔滨: 哈尔滨工业大学, 2014)

[14] Gui Lin, Xu Ruifeng, Lu Qin, et al. Emotion cause extraction, a challenging task with corpus construction [C] //Proc of the 5th Chinese National Conf on Social Media Processing. Berlin: Springer, 2016: 98-109

[15] Gui Lin, Wu Dongyin, Xu Ruifeng, et al. Event-driven emotion cause extraction with corpus construction [C] //Proc of the 2016 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2016: 1639-1649

[16] Wang Zhaoyu. Text emotion cause detection based on emotional common sense knowledge-base [D]. Harbin: Harbin Institute of Technology, 2016 (in Chinese)
(王赵煜. 基于情绪认知常识库的文本情绪原因发现[D]. 哈尔滨: 哈尔滨工业大学, 2016)

[17] Gui Lin, Hu Jiannan, He Yulan, et al. A question answering approach to emotion cause extraction [C] //Proc of the 2017 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2017: 1593-1602

[18] Xu Ruifeng, Hu Jiannan, Lu Qin, et al. An ensemble approach for emotion cause detection with event extraction and multi-kernel SVMs [J]. Tsinghua Science and Technology, 2017, 22(6): 646-659

[19] Wu Dongyin. Event-driven emotion cause detection from text [D]. Harbin: Harbin Institute of Technology, 2017 (in Chinese)
(吴冬茵. 事件驱动的文本情绪原因发现研究[D]. 哈尔滨: 哈尔滨工业大学, 2017)

[20] Hu Jiannan. Study on emotion cause detection from Chinese text [D]. Harbin: Harbin Institute of Technology, 2018 (in Chinese)
(胡健楠. 中文文本情绪原因发现研究[D]. 哈尔滨: 哈尔滨工业大学, 2018)

[21] Fan Chuang, Yan Hongyu, Du Jiachen, et al. A knowledge regularized hierarchical approach for emotion cause analysis [C] //Proc of the 2019 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2019: 5614-5624

[22] Wu Jipeng, Bao Jianzhu, Lan Gongqiang, et al. Emotion cause extraction using rule distillation [J]. Journal of Tsinghua University: Science and Technology, 2020, 60(5): 422-429 (in Chinese)
(巫继鹏, 鲍建竹, 蓝恭强, 等. 结合规则蒸馏的情感原因发现[J]. 清华大学学报: 自然科学版, 2020, 60(5): 422-429)

[23] Fan Chuan, Yuan Chaofa, Du Jiachen, et al. Transition-based directed graph construction for emotion-cause pair extraction [C] //Proc of the 58th Annual Meeting of the ACL. Stroudsburg, PA: ACL, 2020: 3707-3717

[24] Yuan Chaofa, Fan Chuan, Bao Jianzhu, et al. Emotion-cause pair extraction as sequence labeling based on a novel tagging scheme [C] //Proc of the 2020 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2020: 3568-3573

[25] Xia Rui, Ding Zixiang. Emotion-cause pair extraction: A new task to emotion analysis in texts [C] //Proc of the 57th Annual Meeting of the ACL. Stroudsburg, PA: ACL, 2019: 1003-1012

[26] Ding Zixiang, He Huihui, Zhang Mengran, et al. From independent prediction to reordered prediction: Integrating relative position and global label information to emotion cause identification [C] //Proc of the 33rd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2019: 6343-6350

[27] Xia Rui, Zhang Mengran, Ding Zixiang. RTHN: A RNN-transformer hierarchical network for emotion cause extraction [C] //Proc of the 28th Int Joint Conf on Artificial Intelligence. Berlin: Springer, 2019: 5285-5291

[28] Ding Zixiang, Xia Rui, Yu Jianfei. Ecpe-2d: Emotion-cause pair extraction based on joint two-dimensional representation, interaction and prediction [C] //Proc of the 58th Annual Meeting of the ACL. Stroudsburg, PA: ACL, 2020: 3161-3170

[29] Ding Zixiang, Xia Rui, Yu Jianfei. End-to-end emotion-cause pair extraction based on sliding window multi-Label learning [C] //Proc of the 2020 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2020: 3574-3583

[30] Gao Kai, Xu Hua, Wang Jiushuo. A rule-based approach to emotion cause detection for Chinese micro-blogs [J]. Expert Systems with Applications, 2015, 42(9): 4517-4528

[31] Gao Kai, Xu Hua, Wang Jiushuo. Emotion cause detection for chinese micro-blogs based on ecocc model [C] //Proc of the 19th Pacific-Asia Conf on Knowledge Discovery and Data Mining. Berlin: Springer, 2015: 3-14

[32] Wang Jiushuo. Research on emotion cause analyzing method based on microblog posts [D]. Shijiazhuang: Hebei University of Science and Technology, 2015 (in Chinese)
(王九硕. 基于微博文本的情绪诱因分析方法研究[D]. 石家庄: 河北科技大学, 2015)

[33] James W. What is an emotion [J]. Mind, 1884, 9(34): 188-205

[34] Li Weiyuan, Xu Hua. Text-based emotion classification using emotion cause extraction [J]. Expert Systems with Applications, 2014, 41(4): 1742-1749

[35] Neviarouskaya A, Aono M. Extracting causes of emotions from text [C] //Proc of the 6th Int Joint Conf on Natural Language Processing. Nagoya: Asian Federation of Natural Language Processing, 2013: 932-936

[36] Yada S, Ikeda K, Hoashi K, et al. A bootstrap method for automatic rule acquisition on emotion cause extraction [C] //Proc of the 17th IEEE Int Conf on Data Mining Workshops (ICDMW). Piscataway, NJ: IEEE, 2017: 414-421

[37] Song Shuangyong, Meng Yao. Detecting concept-level emotion cause in microblogging [C] //Proc of the 24th Int Conf on World Wide Web. New York: ACM, 2015: 119-120

[38] Ho D T, Cao T H. A high-order hidden Markov model for emotion detection from textual data [C] //Proc of the 12th Pacific Rim Conf on Knowledge Management and Acquisition for Intelligent Systems. Berlin: Springer, 2012: 94-105

[39] Oberländer L A M, Reich K, Klinger R. Experiencers, stimuli, or targets: Which semantic roles enable machine learning to infer the emotions [C] //Proc of the 3rd Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media. Stroudsburg, PA: ACL, 2020: 119-128

[40] Poria S, Majumder N, Hazarika D, et al. Recognizing emotion cause in conversations [J]. arXiv preprint, arXiv: 2012.11820, 2020

[41] Ghazi D, Inkpen D, Szpakowicz S. Detecting emotion stimuli in emotion-bearing sentences [C] //Proc of the 16th Int Conf on Intelligent Text Processing and Computational Linguistics. Berlin: Springer, 2015: 152-165

[42] Kim E, Klinger R. Who feels what and why? Annotation of a literature corpus with semantic roles of emotions [C] //Proc of the 27th Int Conf on Computational Linguistics. Stroudsburg, PA: ACL, 2018: 1345-1359

[43] Balahur A, Hermida J M, Montoyo A. Building and exploiting emotinet, a knowledge base for emotion detection based on the appraisal theory model [J]. IEEE Transactions on Affective Computing, 2011, 3(1): 88-101

[44] Russo I, Caselli T, Rubino F, et al. Emocause: An easy-adaptable approach to emotion cause contexts [C] //Proc of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis. Stroudsburg, PA: ACL, 2011: 153-160

[45] Xu Bo, Lin Hongfei, Lin Yuan, et al. Extracting emotion causes using learning to rank methods from an information retrieval perspective [J]. IEEE Access, 2019, 7: 15573-15583

[46] Fu Keda. Research on emotion cause extraction for news [D]. Harbin: Harbin Institute of Technology, 2018 (in Chinese)
(傅科达. 面向新闻文本的情感原因抽取算法研究[D]. 哈尔滨:哈尔滨工业大学, 2018)

[47] Diao Yufeng, Lin Hongfei, Yang Liang, et al. Emotion cause detection with enhanced-representation attention convolutional-context network [J]. Soft Computing, 2021, 25(2): 1297-1307

[48] Mu Yongli. Research on emotional cause detection based on deep learning [D]. Taiyuan: Shanxi University, 2018 (in Chinese)
(慕永利. 基于深度学习的情绪原因识别方法研究[D]. 太原:山西大学, 2018)

[49] Mu Yongli, Li Yang, Wang Suge. Emotion cause detection based on ensembled convolution neural networks [J]. Journal of Chinese Information Processing, 2018, 32(2): 120-128 (in Chinese)
(慕永利, 李旸, 王素格. 基于 E-CNN 的情绪原因识别方法[J]. 中文信息学报, 2018, 32(2): 120-128)

[50] Zheng Shengxie. Research and implementation of emotion cause extraction method based on deep learning [D]. Beijing: Beijing University of Posts and Telecommunications, 2019 (in Chinese)
(郑胜协. 基于深度学习的文本情绪原因发现方法的研究与实现[D]. 北京:北京邮电大学, 2019)

[51] Xia Linxu, Liu Maofu, Hu Huijun. Emotion cause recognition based on AM-BiLSTM model [J]. Journal of Wuhan University: Natural Science Edition, 2019, 65(3): 276-282 (in Chinese)
(夏林旭, 刘茂福, 胡慧君. 基于 AM-BiLSTM 模型的情绪原因识别[J]. 武汉大学学报:理学版, 2019, 65(3): 276-282)

[52] Hu Guimin, Lu Guangming, Zhao Yi. FSS-GCN: A graph convolutional networks with fusion of semantic and structure for emotion cause analysis [J]. Knowledge-Based Systems, 2021, 212: 106584

[53] Li Xiangju, Song Kaisong, Feng Shi, et al. A co-attention neural network model for emotion cause analysis with emotional context awareness [C] //Proc of the 2018 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2018: 4752-4757

[54] Li Xiangju, Feng Shi, Wang Daling, et al. Context-aware emotion cause analysis with multi-attention-based neural network [J]. Knowledge-Based Systems, 2019, 174: 205-218

[55] Yu Xinyi, Rong Wenge, Zhang Zhuo, et al. Multiple level hierarchical network-based clause selection for emotion cause extraction [J]. IEEE Access, 2019, 7: 9071-9079

[56] Diao Yufeng, Lin Hongfei, Yang Liang, et al. Multi-granularity bidirectional attention stream machine comprehension method for emotion cause extraction [J]. Neural Computing and Applications, 2020, 32(12): 8401-8413

[57] Zhang Chen, Qian Tao, Ji Donghong. Joint model of microblog emotion recognition of cmoticons and emotion cause detection based on neural network [J]. Journal of Computer Applications, 2018, 38(9): 2464-2468 (in Chinese)
(张晨, 钱涛, 姬东鸿. 基于神经网络的微博情绪识别与诱因抽取联合模型[J]. 计算机应用, 2018, 38(9): 2464-2468)

[58] Hu Guimin, Lu Guangming M, Zhao Yi. Emotion-cause joint detection: A unified network with dual interaction for emotion cause analysis [C] //Proc of the 9th CCF Int Conf on Natural Language Processing and Chinese Computing. Berlin: Springer, 2020: 568-579

[59] Yu Chuanming, Li Haonan, An Lu. An analysis of text emotion cause based on multi-task deep learning [J]. Journal of Guangxi Normal University: Natural Science Edition, 2019, 37(1): 50-61 (in Chinese)
(余传明, 李浩男, 安璐. 基于多任务深度学习的文本情感原因分析[J]. 广西师范大学学报:自然科学版, 2019, 37(1): 50-61)

[60] Hu Jiaxing, Shi Shumin, Huang Heyan. Combining external sentiment knowledge for emotion cause detection [C] //Proc of the 8th CCF Int Conf on Natural Language Processing and Chinese Computing. Berlin: Springer, 2019: 711-722

[61] Oberländer L A M, Klinger R. Token sequence labeling vs clause classification for english emotion stimulus eetection [C] //Proc of the 9th Joint Conf on Lexical and Computational Semantics. Stroudsburg, PA: ACL, 2020: 58-70

[62] Shan Jingzhe, Zhu Min. A new component of interactive multi-task network model for emotion-cause pair extraction [J]. Journal of Physics: Conference Series, 2020, 1693: 012022

[63] Yu Jiaxin, Liu Wenyuan, He Yongjun, et al. A mutually auxiliary multitask model with self-distillation for emotion-cause pair extraction [J]. IEEE Access, 2021, 9: 26811-26821

[64] Wu Sixing, Chen Fang, Wu Fangzhao, et al. A multi-task learning neural network for emotion-cause pair extraction [C] //Proc of the 24th European Conf on Artificial Intelligence. Clifton, VA: IOS Press, 2020: 2212-2219

[65] Tang Hao, Ji Donghong, Zhou Qiji. Joint multi-level attentional model for emotion detection and emotion-cause pair extraction [J]. Neurocomputing, 2020, 409: 329-340

[66] Song Haolin, Zhang Chen, Li Qiuchi, et al. An end-to-end multi-task learning to link framework for emotion-cause pair extraction [J]. arXiv preprint, arXiv: 2002.10710, 2020

[67] Cheng Zifeng, Jiang Zhiwei, Yin Yafeng, et al. A symmetric local search network for emotion-cause pair extraction [C] //Proc of the 28th Int Conf on Computational Linguistics. New York: ICCL, 2020: 139-149

[68] Wei Penghui, Zhao Jiahao, Mao Wenji. Effective inter-clause modeling for end-to-end emotion-cause pair extraction [C] //Proc of the 58th Annual Meeting of the ACL. Stroudsburg, PA: ACL, 2020: 3171-3181

[69] Fan Rui, Wang Yufan, He Tingting. An end-to-end multi-task learning network with scope controller for emotion-cause pair extraction [C] //Proc of the 9th CCF Int Conf on Natural Language Processing and Chinese Computing. Berlin: Springer, 2020: 764-776

[70] Chen Xinhong, Li Qing, Wang Jianping. A unified sequence labeling model for emotion cause pair extraction [C] //Proc of the 28th Int Conf on Computational Linguistics. New York: ICCL, 2020: 208-218

[71] Chen Xinhong, Li Qing, Wang Jianping. Conditional causal relationships between emotions and causes in texts [C] //Proc of the 2020 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2020: 3111-3121

[72] Xiao Xinglin, Wang Lei, Kong Qingchao, et al. Social emotion cause extraction from online texts [C/OL] //Proc of the 17th IEEE Int Conf on Intelligence and Security Informatics. Piscataway, NJ: IEEE, 2020 [2021-10-08]. https://ieeexplore.ieee.org/document/9280532

[73] Li Min, Zhao Hui, Su Hao, et al. Emotion-cause span extraction: A new task to emotion cause identification in texts [J]. Applied Intelligence, 2021, 51: 7109-7121

[74] Xiao Xinglin, Wei Penghui, Mao Wenji, et al. Context-aware multi-view attention networks for emotion cause extraction [C] //Proc of the 16th IEEE Int Conf on Intelligence and Security Informatics. Piscataway, NJ: IEEE, 2019: 128-133

[75] Liang Liyuan, Ji Xiaodong, Ren Fuji. Attention-based BiLSTM-CRF network for emotion cause extraction in texts [C] //Proc of the 17th IEEE Int Conf on Mechatronics and Automation. Piscataway, NJ: IEEE, 2020: 1670-1675

[76] Ortony A, Clore G L, Collins A. The Cognitive Structure of Emotions [M]. London: Cambridge University Press, 1988

[77] Talmy L. Toward a Cognitive Semantics. Volume Ⅰ: Concept Structuring Systems [M]. Cambridge, MA: MIT Press, 2000

[78] Balahur A, Hermida J M, Montoyo A. Detecting implicit expressions of sentiment in text based on commonsense knowledge [C] //Proc of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis. Stroudsburg, PA: ACL, 2011: 53-60

[79] Gao Qinghong, Hu Jiannan, Xu Ruifeng, et al. Overview of NTCIR-13 ECA task [C] //Proc of the 13th NTCIR Conf on Evaluation of Information Access Technologies. Tokyo: National Institute of Informatics, 2017: 361-366

[80] Bostan L A M, Kim E, Klinger R. Goodnewseveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception [C] //Proc of the 12th Int Conf on Language Resources and Evaluation Conf. Paris: ELRA, 2020: 1554-1566

Qiu Xiangqing, born in 1981. PhD candidate. His main research interests include natural language processing and affective computing.

邱祥庆,1981 年生.博士研究生.主要研究方向为自然语言处理和情感计算.

Liu Dexi, born in 1975. PhD, professor, PhD supervisor. Senior member of CCF. His main research interests include social media processing, information retrieval and natural language processing.

刘德喜,1975 年生.博士,教授,博士生导师.CCF 高级会员.主要研究方向为社会媒体处理、信息检索和自然语言处理.

Wan Changxuan, born in 1962. PhD, professor, PhD supervisor. Distinguished member of CCF. His main research interests include Web data management, data mining, information retrieval and sentiment analysis.

万常选,1962 年生.博士,教授,博士生导师.CCF 杰出会员.主要研究方向为 Web 数据管理、数据挖掘、信息检索和情感分析.

Li Jing, born in 1983. PhD candidate. Her main research interest is social network user behavior analysis.

李 静,1983 年生.博士研究生.主要研究方向为社会网络用户行为分析.

Liu Xiping, born in 1981. PhD, professor, PhD supervisor. Member of CCF. His main research interests include big data analysis and text mining.

刘喜平,1981 年生.博士,教授,博士生导师.CCF 会员.主要研究方向为大数据分析和文本挖掘.

Liao Guoqiong, born in 1969. PhD, professor, PhD supervisor. Senior member of CCF. His main research interests include social network mining and Internet of things data management.

廖国琼,1969 年生.博士,教授,博士生导师.CCF 高级会员.主要研究方向为社会网络挖掘、物联网数据管理.