文章编号：1673-3193(2024)01-0105-09

基于多任务学习的同行评审细粒度情感分析模型

朱金秋，檀　健，韩斌彬，殷秀秀

（南京邮电大学 理学院，江苏 南京 210023）

摘　要：学术论文同行评审能够直接反映审稿人对论文的主观评价，对审稿文本进行情感分析有利于挖掘审稿人对论文多维度的评价信息。现有的情感分析模型仅能挖掘专家单一的评审维度和相应的情感倾向，本文提出了一种基于多任务学习的同行评审细粒度情感分析模型。该模型在多任务学习框架下，通过在BERT-LCF模型的基础上增加BiLSTM-CRF模块，使其具备了同时完成属性词抽取和细粒度情感分析任务的能力。与传统的基于Pipeline模式的单任务细粒度情感分析模型相比，本模型在保证精度的情况下可以同时完成评审属性提取和情感分析任务。在这两项任务中，所提出模型的$F1$分数分别达到了89.01%和90.71%。对比实验证明，在多任务场景下，引入BiLSTM-CRF模块对评审文本属性词提取任务有一定的提升作用。

关键词：同行评审；多任务学习；属性词抽取；细粒度情感分析；BiLSTM-CRF

中图分类号：TP391.1　　**文献标识码**：A　　**doi:** 10.3969/j.issn.1673-3193.2024.01.014

引用格式：朱金秋，檀健，韩斌彬，等. 基于多任务学习的同行评审细粒度情感分析模型[J]. 中北大学学报（自然科学版），2024，45(1)：105-113.

ZHU Jinqiu，TAN Jian，HAN Binbin，et al. Fine-grained sentiment analysis model of peer review based on multi-task learning[J]. Journal of North University of China(Natural Science Edition)，2024，45(1)：105-113.

Fine-Grained Sentiment Analysis Model of Peer Review Based on Multi-Task Learning

ZHU Jinqiu，TAN Jian，HAN Binbin，YIN Xiuxiu

（School of Science，Nanjing University of Posts and Telecommunications，Nanjing 210003，China）

Abstract：The peer review of academic papers can directly reflect the subjective evaluation of reviewers on the papers，and the extraction of sentiment information from the review text is beneficial to mining rich information of reviewers'evaluation on each dimension of the papers. The existing sentiment analysis task could only extract the single review dimension and sentiment of experts. A fine-grained sentiment analysis model for peer review is proposed based on multi-task learning. The model is equipped with the ability to accomplish both attribute word extraction and fine-grained sentiment analysis tasks by adding the BiLSTM-CRF module to the BERT-LCF model in a multi-task learning framework. Compared with the traditional single-task fine-grained sentiment analysis model based on the Pipeline model，the proposed model can complete the review attribute extraction and sentiment analysis tasks simultaneously while ensuring the accuracy of the model. In the two tasks，$F1$-score of the proposed model reaches 89.01% and 90.71%，respectively. In addition，the introduction of BiLSTM-CRF module has a certain enhancement

收稿日期：2023-02-27

作者简介：朱金秋(1995－)，男，硕士生，主要从事自然语言处理的研究。

通信作者：檀　健(1989－)，男，副教授，博士，主要从事分析数学、机器学习中的数学理论的研究。E-mail：tj@njupt.edu.cn。

(C)1994-2024 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

effect on the review text attribute word extraction task in a multi-task scenario, as demonstrated by comparison experiments.

Key words: peer review; multi-task learning; attribute word extraction; fine-grained sentiment analysis; BiLSTM-CRF

0 引 言

随着科学技术的飞速发展，科技领域的论文产出呈井喷式的增长，尤其在计算机和人工智能领域，这给论文的评审带来了巨大挑战。同行评审(Peer review)作为学界主流的评审方式，为论文评审和裁定的公平性提供了保障[1]。尽管如此，同行评审过程中的随机性、不一致性造成的评分偏差在学术界引起了广泛的讨论[2]。Ragone等[3]发现，评审人员水平的参差不齐会造成同行评审过程中无法发现论文的主要缺陷。正是这些同行评审制度问题的存在，使得目前学界越来越重视对同行评议数据领域的研究。以OpenReview为代表的线上论文公开评审平台的出现，促使同行评议日趋公开透明化，为学术论文评价领域研究提供了大量的开源数据支撑。

情感分析作为自然语言处理领域的成熟技术，对于分析和挖掘专家在论文评审文本数据的主观评价信息方面有着巨大的优势。然而，专家评审文本往往包含对论文多个方面的评价，传统的情感分析模型无法深入挖掘句子中多维度情感信息。如图1所示，该句子反映了评审对论文评价的两个维度情感倾向：writing(论文写作)和novelty(创新性)。相比于细粒度情感分析模型，传统粗粒度情感分析模型会因为只能识别句子的整体情感而造成情感判别的偏差。

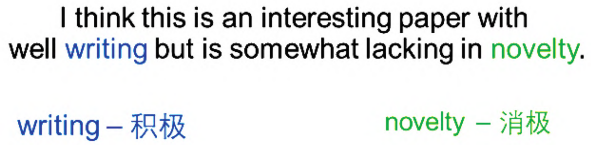

图1 细粒度情感语句示例

Fig. 1 Example of fine-grained sentiment statements

为了自动化地精确提取专家评审维度和情感，本文引入细粒度情感分析技术，提出一种同行评审细粒度情感分析模型BLBC(BERT-LCF-BILSTM-CRF-Joint model)。该模型利用多任务学习框架，在BERT-LCF的基础上融入BiLSTM-CRF模块，实现了同时提取专家评审维度和相应情感倾向的功能。

1 相关工作

常见的细粒度情感分析模型主要采用深度学习方法，如TD-LSTM[4](Target-dependent long short term memory)、ATAE-LSTM[5](Attention-based LSTM with aspect embedding)和IAN[6](Interactive attention networks)。这些模型都利用了长短期记忆网络(Long short term memory)和注意力机制(Attention)，在文本序列建模的基础上，通过注意力机制将情感极性的预测聚焦于与之对应的属性词上。随着大规模预训练模型的兴起，在细粒度情感分析领域，以BERT为输入层的模型取得较高的性能提升。基于BERT开发的BERT-PT[7](BERT post-training)、BERT-SPC[8]和BERT-LCF[9](BERT local context focus)，都是在传统的BERT输入序列末尾通过拼接属性词的方式，使得BERT模型更关注属性词序列的相关信息，以此提高该属性情感倾向的预测效果。然而，由于此类数据的标注是一个既费时又费力的工作，大多数研究往往不会提供语句里属性词的标记。因此，现有的细粒度情感分析任务很少会去考虑属性词的抽取任务，造成其在专有领域的应用性研究较少。

在同行评审文本研究方面，关于专家评审情感信息的提取主要还是使用粗粒度情感分析模型。Wang等[10]引入多实例学习和注意力机制，分别对评审文本中每个短句的情感得分进行建模以提升论文接受预测的效果。Ghosal等[11]利用文档编码器和情感编码器分别提取论文内容信息和评审文本中的情感信息，同时完成论文接收预测和评分预测。林原等[12]提出了一种观点注意力机制，通过增加包含情感信息的观点句权重的方式来提高模型预测论文接收的效果。

现有的细粒度情感分析在同行评审领域的应用主要通过人工划分好的论文评价维度分别对文本进行建模，但无法直接提取单个句子中内部属性词的情感分布。Yuan等[13]利用序列标注模型，实现了可自动化标注论文评审单个句子中的维度

(C)1994-2024 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

词及情感。Chakraborty 等[14]从同行评审文本拆分的短句中标注了论文创新性、清晰性等 9 个论文评价维度以及其相应的情感极性，并以此分别对维度和情感进行建模。张明阳等[15]和 Chakraborty 等[9]类似，通过关键词匹配的方式，从论文的创新性、动机性等维度进行标注，并对单个句子中的单属性情感进行识别。类似地，这些研究虽然形式上加上了对文本评价维度的识别，但并没有考虑一个句子包含多个属性词或多个评价维度的情况，其本质上还是一个粗粒度的多标签分类和标注问题。不仅如此，如果要同时识别评审维度和情感倾向，需要构建两个 Pipeline 模型，这极大地增加了模型占用空间和复杂度[16]。

综上所述，尽管情感信息的引入对同行评审文本接收预测有正向影响，但现有的研究均为基于粗粒度的情感分析建模，且现有的同行评审细粒度情感分析研究本质上也是通过文本分类或序列标注模型，以实现判断评审文本中单一的情感倾向的功能。本文提出了一种端到端的多任务细粒度情感分析模型 BLBC，该模型可以同时提取句子中论文评审属性词以及其相应的情感倾向。

2　模型构建

BLBC 模型任务为给定一个包含属性的同行评审句子序列 $S=\{w_1,\cdots,w_m,w_{m+1},\cdots,w_{m+k},\cdots,w_n\}$，其中，$w_i$ 为句子中的单词，$w_m\sim w_{m+k}$ 为属性词序列，n 为句子的长度，输出为{属性词，情感极性}，模型结构如图 2 所示。

2.1　BERT 共享输入层

BERT（Bidirectional encoder representation from transformers）是一种基于大规模预训练的语言表征模型，能表征文本丰富的语义信息[17]。输入层采用基于 BERT 共享层的词嵌入方式，使用 BERT-BASE（如图 3）和预训练模型 BERT-SPC（如图 4），分别构建属性抽取输入表征模型和情感分析输入表征模型，用以提取全局和局部上下文信息。这两个模型的输入都由词向量、块向量和位置向量组成，区别在于输入形式不同。在 BERT 输入层中，以两个独立的 BERT 进行编码。全局输入序列为 $\boldsymbol{Input}_g$，局部输入序列为 $\boldsymbol{Input}_l$，其中，$w_1,\cdots,w_m$ 为输入句子序列，$q_1,\cdots,q_n$ 为句子中的属性词序列。

$$\boldsymbol{Input}_g=([CLS],w_1,\cdots,w_m,[SEP],q_1,\cdots,q_n,[SEP]),\quad(1)$$

$$\boldsymbol{Input}_l=([CLS],w_1,\cdots,w_m,[SEP])。\quad(2)$$

输入序列经过 BERT 编码后得到全局和局部上下文特征，分别为 $\boldsymbol{O}_g$, $\boldsymbol{O}_l$。

$$\boldsymbol{O}_g=BERT^1(\boldsymbol{Input}_g),\quad(3)$$

$$\boldsymbol{O}_l=BERT^2(\boldsymbol{Input}_l)。\quad(4)$$

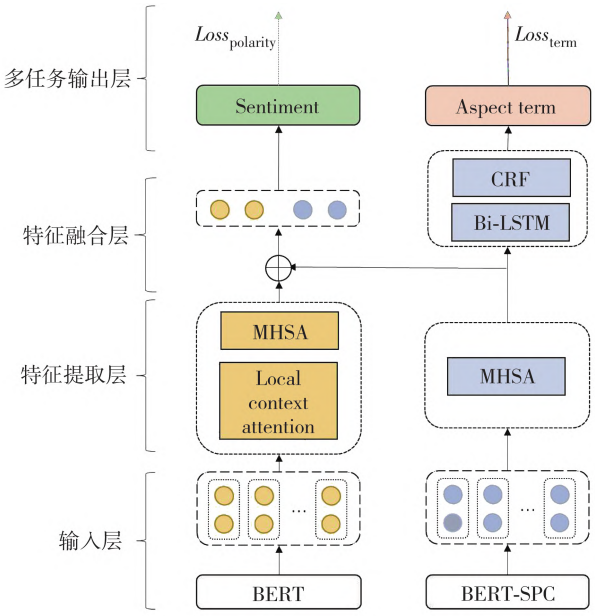

图 2　BERT-LCF-BILSTM-CRF-Joint 多任务细粒度情感分析模型

Fig. 2　BERT-LCF-BILSTM-CRF-Joint multi task fine grained emotional analysis model

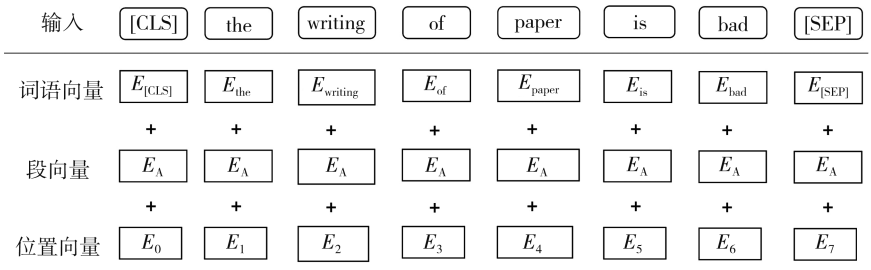

图 3　BERT-BASE 输入特征

Fig. 3　BERT-BASE input feature

(C)1994-2024 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

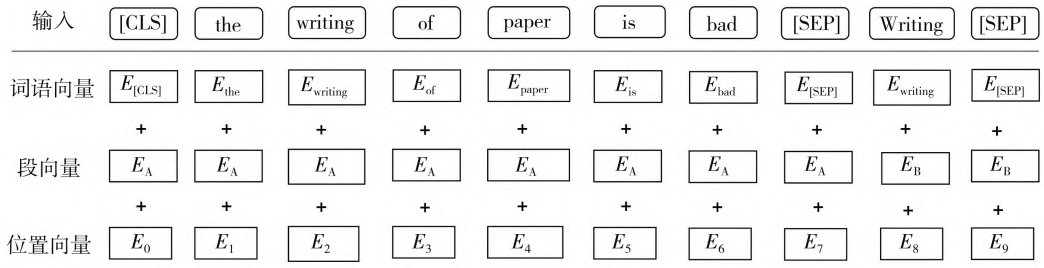

图4 BERT-SPC输入特征

Fig. 4 BERT-SPC input feature

2.2 局部语义注意力层

全局上下文往往包含过多的信息，提取评审句子中的局部关键信息比使用全局信息对于训练模型更有效。局部语义注意力层(Local context attention)通过局部下文聚焦机制LCF(Local context focus)，以加权的方式增加属性词附近的单词权重，降低距离属性词较远的单词权重。LCF以属性词为中心，分别计算各个非属性词与属性词的相对语义距离SRD(Semantic-relative distance)，最后，采取上下文特征动态加权CDW(Context dynamic weighted)的方式将更多的注意力聚焦在设置的属性词相对语义范围内。其中，属性词的相对语义距离d_i为

$$d_i=|i-p^{\mathrm{t}}|-\left|\frac{m}{2}\right|, \tag{5}$$

式中：$i(1\leqslant i\leqslant n)$表示词在句子中的位置；$p^{\mathrm{t}}$表示属性词的中心位置；$m$表示属性词的长度。若句子序列中各非属性词与属性词的相对语义距离d_i超过设定的阈值a，则对其进行权重衰减。若在设定的阈值内，则权重为1，权重计算方式为

$$w_i=\begin{cases}E, & d_i\leqslant\alpha,\\ \dfrac{n-(d_i-\alpha)}{n}\cdot\boldsymbol{E}, & d_i>\alpha,\end{cases} \tag{6}$$

$$\boldsymbol{W}=[w_1,w_2,\cdots,w_n], \tag{7}$$

$$\boldsymbol{O}_l^{\mathrm{CDW}}=\boldsymbol{O}_l\odot\boldsymbol{W}, \tag{8}$$

式中：$\boldsymbol{E}$为元素全为1的向量；n为句子序列的长度；$\odot$表示Hadamard积；$\boldsymbol{W}$为动态权重矩阵。对BERT编码后特征$\boldsymbol{O}_l$进行局部注意力加权，得到上下文动态加权特征$\boldsymbol{O}_l^{\mathrm{CDW}}$。

2.3 多头自注意力机制

多头自注意力机制(Multi-head self-attention，MHSA)是在缩放点积注意力的基础上构建的，能避免评审文本中上下文的长距离依赖对学习语义特征造成的负面影响[18]。设$\boldsymbol{X}$为输入特征，单头注意力(缩放点积注意力)$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})$计算公式为

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=Softmax\left(\frac{\boldsymbol{Q},\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\cdot\boldsymbol{V}, \tag{9}$$

$$\boldsymbol{Q}=\boldsymbol{X}\cdot\boldsymbol{W}^q,\ \boldsymbol{K}=\boldsymbol{X}\cdot\boldsymbol{W}^k,\ \boldsymbol{V}=\boldsymbol{X}\cdot\boldsymbol{W}^v, \tag{10}$$

$$MHSA(\boldsymbol{X})=\tanh\left(concat\{\boldsymbol{H}_1,\cdots,\boldsymbol{H}_h\}\cdot\boldsymbol{W}^{\mathrm{MH}}\right), \tag{11}$$

式中：$\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$由输入矩阵线性变换得到；$\boldsymbol{W}^q\in\mathbf{R}^{d_{\mathrm{h}}\times d_q}$，$\boldsymbol{W}^k\in\mathbf{R}^{d_{\mathrm{h}}\times d_k}$，$\boldsymbol{W}^v\in\mathbf{R}^{d_{\mathrm{h}}\times d_v}$为可学习的权重参数；$d_{\mathrm{h}}$为隐藏层大小，$d_q=d_k=d_v=\sqrt{d}_{\mathrm{h}}$。多头自注意力得分如式(11)，由多个并行的单头注意力H_i拼接，并经过tanh函数激活得到。其中，h为单头注意力的个数，$\boldsymbol{W}^{\mathrm{MH}}$为可学习权重矩阵。上一层的浅层局部语义特征和全局语义特征经过多头注意力MHSA编码后得到深层语义特征，即

$$\boldsymbol{O}_{\mathrm{l}}^{\mathrm{MHSA}}=MHSA(\boldsymbol{O}_{\mathrm{l}}^{\mathrm{CDW}}), \tag{12}$$

$$\boldsymbol{O}_{\mathrm{g}}^{\mathrm{MHSA}}=MHSA(\boldsymbol{O}_{\mathrm{g}})。 \tag{13}$$

2.4 特征融合层

特征融合层融合了局部上下文信息和全局上下文信息，以增强特定属性词的文本表征能力，输出特征$\boldsymbol{X}_{\mathrm{polarity}}$计算公式为

$$\boldsymbol{X}_{\mathrm{polarity}}=\boldsymbol{W}_f\left(concat\left[\boldsymbol{O}_{\mathrm{l}}^{\mathrm{CDW}};\boldsymbol{O}_{\mathrm{g}}^{\mathrm{MHSA}}\right]\right)+\boldsymbol{b}_f, \tag{14}$$

式中：$\boldsymbol{W}_f\in\mathbf{R}^{d_{\mathrm{h}}\times 2d_{\mathrm{h}}}$与$\boldsymbol{b}_f\in\mathbf{R}^{d_{\mathrm{h}}}$为可学习的权重矩阵与偏置项。

2.5 BiLSTM-CRF层

BiLSTM-CRF为属性词序列标记模型，如图5所示。在对输入特征进行双向LSTM编码后，条件随机场CRF(Conditional random fields)可以学习前后文属性词标签之间的约束关系，以降低属性词提取的错误率，得到具有最大概率的合理序列[19]。在CRF中，设全局特征经过

(C)1994-2024 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

BiLSTM（Bidirectional long short-term memory）编码后为$\boldsymbol{X}_{\text{term}}$，其对应的属性词标签预测结果$Y_{\text{t}}$的得分$Score(\boldsymbol{X}_{\text{term}}, Y_{\text{t}})$计算方式为

$$\boldsymbol{X}_{\text{term}} = BiLSTM\left(\boldsymbol{O}_{\text{g}}^{\text{MHSA}}\right), \tag{15}$$

$$Score(\boldsymbol{X}_{\text{term}}, Y_{\text{t}}) = \sum_{i=0}^{n} \boldsymbol{A}_{y_i, y_{i+1}} + \sum_{i=1}^{n} \boldsymbol{P}_{i, y_i}, \tag{16}$$

式中：$\boldsymbol{A}$为转移概率矩阵；$\boldsymbol{A}_{y_i, y_{i+1}}$表示标签$y_i$转移到$y_{i+1}$的概率；$\boldsymbol{P}_{i, y_i}$表示第$i$个词标记为标签$y_i$的概率。

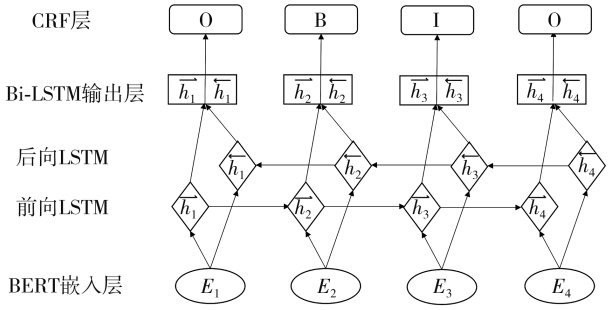

图 5　BiLSTM-CRF 模型

Fig. 5　BiLSTM-CRF model

2.6　情感分析输出层

本文提出的多任务细粒度情感分析有两个输出层，二者分别用 Softmax 激活函数进行指数归一化来预测属性词情感极性分布概率$\widehat{Y}_{\text{p}}$和属性词标签的分布概率$\widehat{Y}_{\text{t}}$。

$$\widehat{Y}_{\text{p}} = Softmax\left(\boldsymbol{X}_{\text{polarity}}\right), \tag{17}$$

$$\widehat{Y}_{\text{t}} = Softmax\left(Score(\boldsymbol{X}_{\text{term}}, Y_{\text{t}})\right)。 \tag{18}$$

使用交叉熵损失函数作为情感极性预测任务和属性词抽取任务的损失函数。此时，模型训练的联合损失为

$$L = -(1-\sigma) L_{\text{polarity}} - \sigma L_{\text{term}} + \lambda \sum_{\theta \in \Theta} \theta^2, \tag{19}$$

$$L_{\text{polarity}} = -\sum_{c=1}^{C} \hat{y}_p \log y_{\text{p}}, \tag{20}$$

$$L_{\text{term}} = -\sum_{n=1}^{N} \sum_{k=1}^{K} \hat{y}_t \log y_{\text{t}}, \tag{21}$$

式中：L_{polarity}为情感极性预测任务的损失；L_{term}为属性词抽取任务的损失；σ为联合损失的调整系数；λ为L_2正则项系数；Θ表示模型的参数集；C表示情感标签类型的数量；N表示上下文单词的数量；K表示属性词标签类型的数量。

3　实验过程及结果

3.1　论文评审维度和显式关键词设置

对论文评审文本进行细粒度情感分析，需划定论文的评价维度。本文主要采取对评价短句中“显式词语”进行人为归纳的方式，以反映论文的各个评价属性。本文综合以往研究，汇总了一套显式关键词设置，部分词表如表 1 所示，将论文评审的总体维度划分为清晰性、对比、创新性、动机和实验设计。

表 1　论文评审维度和部分显式关键词设置

Tab. 1　Paper review dimensions and part of explicit keyword settings

评价维度	关键词
Clarity（清晰性）	Presentation, Writing, Written, Clarity, Details, Structure, Organization, …
Comparison（对比）	Comparison, Compared, Baselines, Related work, …
Originality（创新性）	Originality, Novel, Novelty, Originallity, Ideal, Contribution, …
Motivation（动机）	Motivation, Motivated, Motivate, Meaning, …
Experiment design（实验设计）	Analysis, Evaluation, Results, Experiments, Dataset, Performance, …

3.2　数据准备

本文所使用的数据集为 OpenReview 平台爬取的数据，并采用人工标注的方式标注了包含 3 488 个属性词及相应情感倾向的同行评审文本句子，其中积极情感为 1 384 个，消极情感为 2 104 个。通过随机划分训练集与测试集的方式，最终得到训练样本 2 740 个，测试样本 748 个。图 6 展示了实验样本的句子长度，发现本文标注句子的长度集中分布在 15 个～30 个单词。选取数量最多的属性词类别，绘制它们在不同情感倾向的分布情况，如图 7 所示。评价论文创新维度和论文动机性维度的词数量最多，如 Idea，Novelty，Motivation，Motivated 等，且它们在情感倾向的分布上有着不平衡的特点。

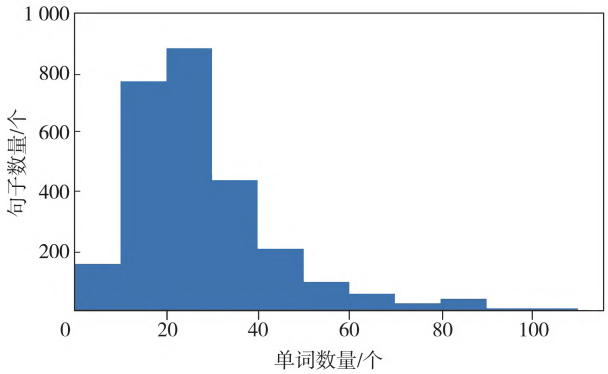

图 6　句子长度分布图

Fig. 6　Sentence length distribution

(C)1994-2024 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

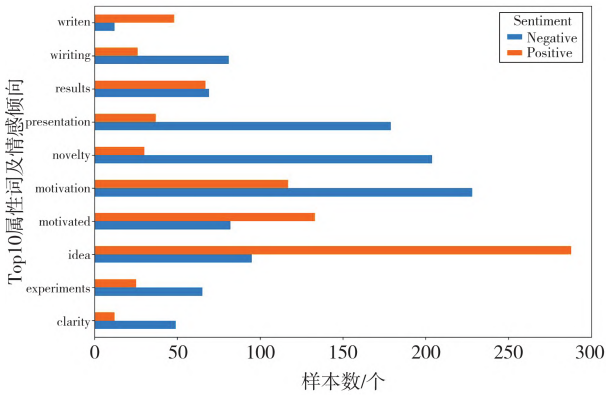

图7 Top10属性词分布图

Fig. 7 Top10 aspect word distribution

本文实验数据的标注形式以表2为例，其中，属性词标签为$y_t \in \{B,I,O\}$，B表示属性词/短语的开始词，I表示属性词/短语的中间词，O表示非属性词。情感标签为$y_p \in \{NEG,POS\}$，分别表示审稿人评审文本中对属性词的情感倾向，正面为POS，负面为NEG。表2中Novelty与Results为属性词，其对应的情感均为负面情感，而其他单词用占位符—1代替。

3.3 实验设计

为了验证本文提出的多任务模型在科技论文同行评审数据集上属性抽取和细粒度情感预测的效果，进行了对比实验，如表3所示。实验主要分为3组，两组单任务对比模型，一组多任务对比模型。单任务对比模型主要基于Pipeline模式，即先进行单独的属性词抽取，再根据给定的属性词进行情感分析；多任务对比模型为同时提取属性词和情感的多任务对比实验。

表2 同行评审文本细粒度情感标签标注示例

Tab. 2 Examples of fine-grained sentiment labels for peer-reviewed text

标签	输入									
	there	is	a	lack	of	novelty	and	no	significant	results
y_t	O	O	O	O	O	B	O	O	O	B
y_p	—1	—1	—1	—1	—1	NEG	—1	—1	—1	NEG

表3 对比实验设计

Tab. 3 Comparison experiment design

实验组	对比模型
实验一(单任务) 属性词抽取任务	BiLSTM-CRF:GLOVE词向量作为输入,使用BiLSTM编码,最后接条件随机场CRF。 BERT-linear:BERT词向量作为输入,最后接*Softmax*层。 BERT-CRF:BERT词向量作为输入,最后接条件随机场CRF。 BERT-BiLSTM-CRF:BERT词向量作为输入,BiLSTM进行上下文编码,最后接条件随机场CRF。
实验二(单任务) 细粒度情感分析任务	ATAE-LSTM:GLOVE词向量作为输入,将属性词和上下文单词拼接送入LSTM和注意力机制进行学习属性词情感。 BERT-BASE:输入序列以BERT词向量作为输入,最后接入*Softmax*激活函数输出层。 BERT-SPC:在BERT-BASE的基础上,将输入序列换为属性词拼接后的BERT输入序列。 BERT-AEN:在BERT的基础上融入注意力编码网络,建立上下文和属性词的关系。 BERT-LCF:在BERT的基础上利用局部上下文机制学习局部上下文信息,同时融合BERT-SPC的全局上下文信息进行情感预测。
实验三(多任务) 基于多任务学习的属性词抽取与细粒度情感分析	BERT-BASE-Joint(模型1):输入序列以BERT编码,共享其BERT词嵌入向量,用两个*Softmax*激活函数输出层同时完成属性抽取与细粒度情感分析。 BERT-LCF-linear-Joint(模型2):在BERT-LCF的基础上,利用BERT全局上下文信息通道加入一个属性抽取的*Softmax*激活函数输出层。 BERT-LCF-CRF-Joint(模型3):在模型2的基础上,属性抽取的*Softmax*输出层前加入一个条件随机场。 ATEPC-LCF[20](模型4):基于属性抽取和情感分析的联合学习模型。 BLBC(our):在模型3的基础上,属性抽取的*Softmax*输出层前加入一个BiLSTM-CRF层。

3.4 实验参数设置

本实验参数设置如表4所示，其中，BERT和GLOVE分别表示所使用的预训练词嵌入模型，它们的词嵌入维度大小(Embedding size)分别为768与300。Learning rate为模型初始学习率。考虑到模型训练过程中的鲁棒性以及减少参数调整的多余操作，选择带有自适应功能的Adam作为本实验的优化器。为了防止模型过拟合，Dropout和L2分别设置为0.1与10^{-5}。

(C)1994-2024 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表 4　实验参数设置

Tab. 4　Experimental parameter settings

参数	设定值
BERT	Bert-base-uncased
BERT embedding size	768
GLOVE	glove.840B.300d
GLOVE embedding size	300
Learning rate	10^{-5}
优化器	Adam
Dropout	0.1
Epochs	10
Batch size	32
L2	10^{-5}

3.5　实验结果分析

3.5.1　基于pipeline单任务模型的对比实验

为了验证实验结果的有效性，本文选取了细粒度情感分析领域两个常见的评价指标*Acc*（Accuracy）和*F1*得分（*F1*-Score），其中，*Acc*表示所有样本中正确分类的指标，*F1*综合考虑了模型的精确率和召回率，能有效评价不平衡样本数据集模型的效果。表 5 中两组对比模型分别统计了它们在各项单任务中的预测效果，其中，“—”表示单任务模型无法完成一项任务的占位符。在属性抽取模型中，基于GLOVE静态词向量的BiLSTM-CRF虽然利用了BiLSTM提取句子上下文语义特征，但其性能明显弱于BERT类的模型。同样是引入条件随机场，BERT-CRF比BiLSTM在*F1*上提高了7.44%。在BERT类模型中，各模型的预测性能差距不大，BERT-BiLSTM-CRF仅比BERT-linear在*F1*上提高了0.54%。

表 5　单任务模型实验结果

Tab. 5　Single task model experiment results

单任务模型	属性词提取任务	细粒度情感分析任务	
	F1/%	*Acc*/%	*F1*/%
BiLSTM-CRF	80.62	—	—
BERT-linear	88.06	—	—
BERT-CRF	88.53	—	—
BERT-BiLSTM-CRF	89.07	—	—
ATAE-STM	—	70.39	70.18
BERT-BASE	—	83.69	82.59
BERT-SPC	—	91.18	90.89
BERT-AEN	—	91.31	90.96
BERT-LCF	—	92.36	92.00

在细粒度情感分析模型方面，基于GLOVE的ATAE-LSTM预测性能最差。在BERT类模型中，BERT-BASE由于没有考虑属性词信息，*F1*仅为82.59%。在输入序列中增加属性词信息的BERT-SPC在*F1*上比它提高了8.3%。BERT-AEN基于注意力编码网络，在BERT的基础上建立了上下文和评审属性词之间的联系，使得其在*F1*得分上高于BERT-SPC。在BERT-SPC的基础上集成局部注意机制的BERT-LCF表现最佳，*F*1达到92.0%。

3.5.2　基于多任务模型的对比实验

表 6 为本文提出的多任务属性抽取和情感分析联合模型BLBC及其他多任务模型的对比。结果表明，多任务模型能同时完成属性词提取和细粒度情感分析任务。由于在单任务对比实验中，基于GLOVE的模型均表现不佳，本文多任务对比模型均以BERT为基础。BERT-BASE-Joint模型由于没有属性词信息的拼接，在细粒度情感分析任务中表现较差，*F*1仅为82.41%。基于BERT-LCF网络框架的多任务模型在该任务里表现优异，*F*1达到90.71%。在属性词提取方面，多任务场景下BLBC通过引入BiLSTM-CRF模块，有助于提升多任务中的属性词提取，*F*1达到了89.01%。ATEPC-LCF相比BERT-LCF-CRF-Joint加强了其在细粒度情感分析上特征提取的能力，但在属性抽取能力上与其他基于BERT的非CRF基准模型一样差，*F*1仅为87.25%。这与单属性提取任务对比实验中得出的结论一致。

表 6　多任务模型实验结果

Tab. 6　Multi task model experiment results

多任务模型	属性词提取任务	细粒度情感分析任务	
	*F*1/%	*Acc*/%	*F*1/%
BERT-BASE-Joint	87.19	83.06	82.41
BERT-LCF-linear-Joint	87.30	89.65	89.32
BERT-LCF-CRF-Joint	88.38	90.46	90.08
ATEPC-LCF	87.25	90.90	90.57
BLBC	89.01	90.99	90.71

4　案例研究

为了验证所提出的多任务细粒度情感分析模型在同行评审上的预测效果，选取了一篇2019年ICLR公开的论文评审文本，旨在通过可视化的方式展示模型预测结果。评审文本示例及其细粒度情感分析结果如表 7 和图 8 所示。表 7 列出了属

(C)1994-2024 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

性词及其情感极性，其中，正向情感为“+”，负向情感为“—”。

图8为评审文本原文，评审文本最后一行为属性词所属的细粒度评审维度，即综合情感倾向。在所示案例中，多任务模型中的属性抽取任务对每个单词的所属属性标签进行了预测，得到Idea，Presentation，Methodology，Explanation等属性词。根据表1指定的评审维度和属性词设置(用颜色区分)，最终得到评审对该论文的创新性评价为正向，对论文内容的呈现(清晰性)以及实验设计的评价为负向。

表7 同行评审文本细粒度情感分析示例

Tab. 7 Example of fine-grained sentiment analysis of peer-reviewed text

属性词	情感极性
Idea	+
Written	+
Presentation	—
Methodology	—
Explanation	—
Results	—
Discussions	—
Performance	—

The work suggests reshuffle images blocks of adversarial examples during adversarial training, in order to improve the generalization performance on benign and adversarial test samples. The main method is based on the hypothesis in [Zhang et al 2019], [Ilyas et al 2019]. The assumption claims that robust models rely on global structural features, and non-robust models rely on local features. Thus, the work tries to learn local robust features, by cutting and reshuffling the image blocks. Overall the idea is interesting and the paper is well written. However, there are some concerns about the presentation and the main methodology: Can the paper give more explanation on the purpose of inserting the feature transfer term in the objective function? What is the difference of the proposed one with directly minimizing the loss on both original PGD image and reshuffled image? For CIFAR10, TRADEs and PGDAT's performance in the result is not as good as the ones shown in their original works, which is comparable to the performance of the proposed RLFAT method. More discussions are needed, otherwise the experimental results are not convincing. More intuitions are needed on what local and global features are, and why training on the reshuffled images can help learn generalizable robust local features. Overall the paper is easy to understand, but we suggest that more insight should be given on the success of the proposed method.

Novelty + Clarity — Experiment Design —

图8 同行评审文本示例及其细粒度情感分析结果

Fig. 8 Example of peer-reviewed text and its fine-grained sentiment results

5 结 论

对科技论文评审文本的情感分析任务研究主要集中于粗粒度的建模，如粗粒度的论文评审情感分析以及根据审稿文本预测论文接收情况等任务。在细粒度情感分析方面，传统的Pipeline模式会造成多任务的误差传递和参数冗余。本文以OpenReview同行评审文本数据集为研究对象，提出了一种基于多任务学习的细粒度情感分析方法。相比于以往对同行评审文本粗粒度的建模，本模型可以更细粒度地发掘审稿人对论文各评价维度的情感倾向，为论文智能化评审提供决策辅助。与此同时，该模型真正实现了端到端的评审文本属性词提取和情感极性预测任务，并取得了优异的效果，其属性词提取任务的$F1$达到89.01%，细粒度情感极性预测任务的$F1$达到了90.71%。

参考文献：

[1] BENOS D J，BASHARI E，CHAVES J M，et al. The ups and downs of peer review[J]. Advances in Physiology Education，2007，31(2)：145-152.

[2] BORNMANN L，DANIEL H D. Reliability of reviewers' ratings when using public peer review：a case study[J]. Learned Publishing，2010，23(2)：124-131.

[3] RAGONE A，MIRYLENKA K，CASATI F，et al. On peer review in computer science：Analysis of its effectiveness and suggestions for improvement[J]. Scientometrics，2013，97(2)：317-356.

[4] TANG D，QIN B，FENG X，et al. Effective LSTMs for target-dependent sentiment classification[J]. 2015，

(C)1994-2024 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

arXiv：1512.01100.

[5] WANG Y，HUANG M，ZHU X，et al. Attention-based LSTM for aspect-level sentiment classification[C]//2016 Conference on Empirical Methods in Natural Language Processing. 2016：606-615.

[6] MA D，LI S，ZHANG X，et al. Interactive attention networks for aspect-level sentiment classification[C]//26th International Joint Conference on Artificial Intelligence. 2017：4068-4074.

[7] XU H，LIU B，SHU L，et al. BERT post-training for review reading comprehension and aspect-based sentiment analysis[C]//NAACL-HLT. 2019：2324-2335.

[8] SONG Y，WANG J，JIANG T，et al. Attentional encoder network for targeted sentiment classification[J]. arXiv：1902.09314，2019.

[9] ZENG B，YANG H，XU R，et al. LCF：A local context focus mechanism for aspect-based sentiment classification[J]. Applied Sciences，2019，9(16)：3389.

[10] WANG K，WAN X. Sentiment analysis of peer review texts for scholarly papers[C]//The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. 2018：175-184.

[11] GHOSAL T，VERMA R，EKBAL A，et al. Deepsentipeer：Harnessing sentiment in review texts to recommend peer review decisions[C]//The 57th Annual Meeting of the Association for Computational Linguistics. 2019：1120-1130.

[12] 林原，王凯巧，杨亮，等. 基于pu-learning的同行评议文本情感分析[J]. 计算机工程与应用，2023，59(3)：143-149.
LIN Yuan，WANG Kaiqiao，YANG Liang，et al. Sentiment analysis of peer review texts based on pu-Learning[J]. Computer Engineering and Applications，2023，59(3)：143-149.（in Chinese）

[13] YUAN W，LIU P，NEUBIG G. Can we automate scientific reviewing?[J]. Journal of Artificial Intelligence Research，2022，75：171-212.

[14] CHAKRABORTY S，GOYAL P，MUKHERJEE A. Aspect-based sentiment analysis of scientific reviews[C]//ACM/IEEE Joint Conference on Digital Libraries in 2020. 2020：207-216.

[15] 张明阳，王刚，彭起，等. 学术论文公开评审平台数据分析[J]. 计算机科学，2021，48(6)：63-70.
ZHANG Mingyang，WANG Gang，PENG Qi，et al. Data analysis of open review[J]. Compute Science，2021，48(6)：63-70.（in Chinese）

[16] VANDENHENDE S，GEORGOULIS S，VAN GANSBEKE W，et al. Multi-task learning for dense prediction tasks：a survey[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence，2022，44(7)：3614-3633.

[17] KENTON J D M W C，TOUTANOVA L K. Bert：Pre-training of deep bidirectional transformers for language understanding[C]//NAACL-HLT. 2019：4171-4186.

[18] VASWANI A，SHAZEER N，PARMAR N，et al. Attention is all you need//31st Advances in Neural Information Processing Systems. 2017：1-11.

[19] 丁晟春，方振，王楠. 基于Bi-LSTM-CRF的商业领域命名实体识别[J]. 现代情报，2020，40(3)：103-110.
DING Shengchun，FANG Zhen，WANG Nan. Business domain named entity recognition based on Bi-LSTM-CRF[J]. Journal of Modern Information，2020，40(3)：103-110.（in Chinese）

[20] YANG H，ZENG B，YANG J H，et al. A multi-task learning model for chinese-oriented aspect polarity classification and aspect term extraction[J]. Neurocomputing，2021，419：344-356.

(C)1994-2024 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net