

基于多视角建模的汉语议论文写作质量评估方法

贺亚琼, 蒋峰, 褚晓敏, 李培峰

引用本文

贺亚琼, 蒋峰, 褚晓敏, 李培峰. [基于多视角建模的汉语议论文写作质量评估方法](#) [J]. 计算机科学, 2023, 50(3): 315-322.

HE Yaqiong, JIANG Feng, CHU Xiaomin, LI Peifeng. [Chinese Argumentative Writing Quality Evaluation Based on Multi-perspective Modeling](#) [J]. Computer Science, 2023, 50(3): 315-322.

相似文章推荐 (请使用火狐或 IE 浏览器查看文章)

Similar articles recommended (Please use Firefox or IE to view the article)

[基于联合模型的端到端事件可信度识别](#)

End-to-End Event Factuality Identification with Joint Model

计算机科学, 2023, 50(2): 292-299. <https://doi.org/10.11896/jsjcx.211200108>

[基于多模态注意力的噪声事件分类模型](#)

Noise Event Classification Model Based on Multimodal Attention

计算机科学, 2022, 49(11A): 211000161-7. <https://doi.org/10.11896/jsjcx.211000161>

[基于相似度矩阵学习和矩阵校正的无监督多视角特征选择](#)

Unsupervised Multi-view Feature Selection Based on Similarity Matrix Learning and Matrix Alignment

计算机科学, 2022, 49(8): 86-96. <https://doi.org/10.11896/jsjcx.210700124>

[三维城市场景中的小物体检测](#)

Small Object Detection in 3D Urban Scenes

计算机科学, 2022, 49(6): 238-244. <https://doi.org/10.11896/jsjcx.210400174>

[基于隐式视角转换的视频异常检测](#)

Video Anomaly Detection Based on Implicit View Transformation

计算机科学, 2022, 49(2): 142-148. <https://doi.org/10.11896/jsjcx.210900266>

基于多视角建模的汉语议论文写作质量评估方法

贺亚琼 蒋 峰 褚晓敏 李培峰

苏州大学计算机科学与技术学院 江苏 苏州 215006

(20204227070@stu.suda.edu.cn)

摘 要 自动作文评分是一项代替人工为学生作文进行等级评分的任务,其中丰富的语义、严密的组织和合理的逻辑是重要的考虑因素。已有的研究大多数只从语义或组织等视角出发评估作文的质量,未考虑如逻辑等更高层次的因素。因此,文中提出了一个多视角评价框架(Multi-perspective Evaluation Framework, MPE),从语义表达、组织结构和整体逻辑 3 个方面对学生议论文进行了客观、可靠的评价。具体来说,多视角评价框架首先利用预训练模型编码句子并获得由低到高 3 个层次的语义信息,来评估文章的语义表达;其次,框架将句子功能识别与段落功能识别相结合,用于评估文章的组织结构;然后,通过计算段落之间的连贯性来评估文章的整体逻辑;最后,该框架综合这 3 个方面的评估特征,对作文评分。实验结果表明,所提出的多视角评价框架能够有效地对不同质量的作文进行评分,优于所有基准系统。

关键词: 多视角;作文评分;议论文;XLNet;全局连贯性

中图法分类号 TP391

Chinese Argumentative Writing Quality Evaluation Based on Multi-perspective Modeling

HE Yaqiong, JIANG Feng, CHU Xiaomin and LI Peifeng

School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu 215006, China

Abstract Automated essay scoring is a task that replaces manual grading for students' essays, where rich semantics, rigorous organization, and reasonable logic are important considering factors. Most previous studies only consider the semantics or organization of the essay from a single perspective, lacking considering higher-level factors such as logic. Therefore, this paper proposes a multi-perspective evaluation framework (MPE) to more objective and reliable evaluate the essay from semantics, organization, and logic. MPE first utilizes the pre-trained model to encode sentence and obtain three levels semantic information to evaluate the essay's semantic expression. Then, it combines sentence function identification and paragraph function identification to evaluate the essay's organization. Moreover, MPE evaluates the essay's logic by calculating the coherence between paragraphs. Finally, the framework scores the essay by integrating these three evaluation perspectives. Experimental results show that the proposed multi-perspective evaluation framework can effectively score the essays at various qualities, outperforming all the baselines.

Keywords Multi-perspective, Essay score, Argumentation, XLNet, Global coherence

1 引言

自动作文评分 (Automated Essay Scoring, AES) 旨在利用语言学、统计学和自然语言处理技术对书面作文进行自动评估和评分。自动作文评分能够避免主观因素对作文评分的影响,大大减轻阅卷老师的工作量。目前,自动作文评分在英文作文评分上已经有了广泛的研究,如 Mesgar 等^[1]、Liu 等^[2]、Yang 等^[3] 均在英文作文数据集 ASAP¹⁾ 上进行了研究,而汉语作文评分的研究还处在尝试阶段。

在自动作文评分的相关研究中,传统方法依赖于手工

特征^[4-7]对作文进行评分,近年来深度学习模型取得了更好的性能。在使用深度学习模型对作文评分的相关工作中,一些工作从文本自身语义出发对作文进行整体评分^[8-11],也有一些工作对文本的特定特征如叙事质量^[12]、写作立场^[13]、写作风格^[14]、主题清晰度^[15]、组织结构^[16]等进行等级评价,从而获得了更丰富的反馈信息,但很少有工作将上述作文特征结合起来对汉语作文进行整体评分。作文特征指决定作文质量的不同方面,如英文数据集 ASAP 给出的决定英文议论文质量特征包含内容、组织、单词选择、句子流畅度和写作习惯等,而对于中文议论文,高考作文等级评价从 3 个角度

¹⁾ <https://www.kaggle.com/c/asap-aes>

到稿日期:2022-01-14 返修日期:2022-08-14

基金项目:国家自然科学基金(61836007,62006167);江苏省高校优势学科建设工程资助项目

This work was supported by the National Natural Science Foundation of China(61836007,62006167) and Priority Academic Program Development of Jiangsu Higher Education Institutions(PAPD).

通信作者:李培峰(pfli@suda.edu.cn)

评估作文质量,分别是内容、表达和发展等级。表 1 列出了 高考作文的评判标准¹⁾。

表 1 高考作文评判标准
Table 1 College entrance examination composition evaluation criteria

	一等	二等	三等	四等
内容	符合题意	符合题意	基本符合题意	偏离题意
	中心突出	中心明确	中心基本明确	中心不明确
	内容充实	内容较充实	内容单薄	内容不当
	思想健康	思想健康	思想基本健康	思想不健康
	感情真挚	感情真实	感情基本真实	感情虚假
表达	符合文体要求	符合文体要求	基本符合文体要求	不符合文体要求
	结构严谨,文脉畅通,段落划分恰当,衔接紧凑,过渡自然	结构完整,有头有尾,分段合理	结构基本完整,段落划分不够合理	结构混乱,段落划分有严重缺陷,杂乱无章
	语言流畅	语言流畅	语言基本通顺	语言不通顺
	字迹工整	字迹清楚	字迹基本清楚	字迹潦草难辨
	深刻;思想上的深刻	较有文采	略显文采	个别语句有深意
发展	丰富;内容丰富	较有创意	略有创意	个别例子较好
	有文采;语言文采有创意			个别语句较精彩
	有创意;见解新颖;材料新鲜			个别地方有新意

为了验证结合作文特征对汉语作文进行等级评价是否可以取得更好的效果,本文提出结合作文特征对作文进行整体评分。相比之前的工作仅从单一视角出发对作文进行等级评价,本文综合高考作文评判标准,从文本语义、组织结构和整体逻辑 3 个视角对 Song 等^[17]发布的 1 220 篇高中汉语议论文进行了等级评价,通过综合该 3 个视角的信息,更加全面地评价议论文,获得了更为优秀的性能表现。具体而言,语义表达从多个层次出发考虑文本内容,组织结构考虑文本的组织结构和每个部分的功能,整体逻辑考虑篇章整体的连贯性。由于汉语作文评分工作中尚未有标注好的作文特定特征评分语料库,因此本文不对作文特征进行等级评价,而是采用不同的方法获取不同视角特征的信息,将 3 种信息相结合对作文整体进行评分。

汉语作文文本语义表达所面临的问题主要有两个,一个是中文文本在输入到英文自动作文评分的常用模型如 CNN^[18]或 LSTM^[19-20]中时,句子会被分割为词组。对于中文文本,不恰当的分词会曲解文章含义,从而导致中文自动作文评分任务表现不佳。另一个是英文作文中单词的使用是决定其质量的一个重要因素^[21],英文文本也主要通过结合单词之间的相似性来形成文本语义表示^[22-24],而根据高考作文评分标准,中文则更注重句子级别或段落级别的文本表达,如文本的连贯性和语言的流畅等。为了解决上述问题,本文使用预训练模型 XLNet^[25]获得句子表示。首先,与传统的词嵌入相比,XLNet 可以将单个汉字作为输入,无须考虑分词问题;其次,相比早期研究通过长短时记忆网络(LSTM)和卷积神经网络(CNN)建模文本语义表示,XLNet 更能获得深层次语义。

作文组织结构的评估所面临的问题主要是如何有效表示组织结构以及如何将其与作文质量联系起来。一篇好的议论文应具有合理的组织结构,论点、论据、论证等成分缺一不可,在已有的研究中,Attali 等^[26]、Persing 等^[16]、Song 等^[17]提出使用篇章功能(句子功能和段落功能)来表现作文的组织形式,他们认为篇章功能能够清晰地表现出议论文的组织结构,可以提高组织结构质量评估的可解释性。因此,本文使用

篇章功能表现作文整体的组织。

作文的整体逻辑所面临的问题是如何表示文本的逻辑关系。Liang^[27]的研究表明,文本连贯性是一种逻辑关系,高质量的文章更多的依赖文章的全局连贯性;Louis 等^[28]、Mitsakaki 等^[29]和 Liao 等^[30]的研究工作表明,语篇连贯性特征对作文评分任务有显著提升。因此,本文使用篇章连贯性来表示作文逻辑,从而为作文质量等级评估提供文本逻辑信息。

基于以上分析,本文提出了一个多视角评价框架(MPE),相比以往仅考虑单一因素如作文语义表达来预测作文等级,该框架从多个视角出发,不仅考虑了作文的语义表达和组织结构,还提出使用篇章连贯性来表示作文的整体逻辑。

本文的主要贡献如下:

(1)提出将作文的整体逻辑作为评估作文质量的一个因素,并使用篇章连贯性来表示作文的整体逻辑。

(2)提出了一种针对议论文质量评估的多视角评价框架(MPE),结合了文章的语义表达、组织结构和整体逻辑 3 个决定作文质量的因素。

(3)实验结果证明,MPE 框架能有效提高对不同等级作文的识别能力,其性能明显优于基准系统。

2 相关工作

对于一篇议论文优劣的评价需要从多个方面进行考虑,如语义表达、组织结构和整体逻辑等。已有的工作大多分别从这 3 个方面对作文进行评价,如仅从篇章语义信息出发对作文进行等级评价,或是评价作文的组织质量和连贯性。本节归纳了关于作文评分相关的语义表达、组织结构以及文本连贯性的表示等方法的已有研究。

2.1 语义表达

语义表达是决定文本质量的基本因素,长短时记忆网络(LSTM)^[19-20]和卷积神经网络(CNN)^[18]被广泛应用于作文的语义表达中。例如,Alikaniotis 等^[8]、Taghipour 等^[9]使用单层 LSTM 对一篇文章的单词序列编码得到篇章级的文本语义表示,并将其输入线性层预测作文等级,这种方法只考虑了单词级别的信息。Dong 等^[10]使用两层 CNN 分别对句子

¹⁾ <http://www.xuefen.com.cn/144691.html>

和篇章进行建模,从而加强对文本语义的理解。Dong 等^[11]则使用 CNN 获得句子语义表示,再将句子语义表示输入到 LSTM 获得篇章语义表示,从而对作文进行评分。然而,上述方法仅考虑了单词或句子对作文质量的影响,且使用 LSTM 和 CNN 仅能捕获浅层语义,难以获得深层次语义信息。

2.2 组织结构评估

对于议论文,文本的组织是决定其质量的重要因素,对于一篇好的议论文,论点、论据和论证等成分缺一不可。Attali 等^[26]、Persing 等^[16]和 Song 等^[17]提出使用篇章功能对作文组织结构进行评估,篇章功能指句子功能或段落功能。Attali 等^[26]将一篇议论文包含的段落功能与标准的五段落式结构作对比来评估议论文的组织结构质量,即假设一篇议论文包含 1 个引言段落、3 个正文段落(每个正文段落包含论点和论据)、1 个总结段落,缺少某个篇章功能会对作文的组织结构质量造成影响。Persing 等^[16]构建了第一个语料库,对 1 003 篇英文议论文的组织质量等级、句子功能和段落功能进行标注,并使用句子功能序列和段落功能序列的 ngrams 作为特征探讨其与组织质量之间的相关性。Song 等^[17]新标注了一个包含 1 220 篇汉语议论文的语料库,标注内容包括组织质量等级、句子功能和段落功能,并提出了一种分层的多任务学习模型,该模型由局部到整体划分了 3 个层次,分别为句子功能识别、段落功能识别与组织评估。

2.3 文本连贯性表示

文本连贯性是一种逻辑关系,这种逻辑关系将文本中的命题或思想有效地连接起来,使得文本在意义上具有连续性。Louis 等^[28]、Miltakaki 等^[29]和 Liao 等^[30]的研究工作表明,语篇连贯性特征对作文评分任务有显著提升。对于作文连贯性, Xu^[31]使用衔接手段如连词的使用来衡量语篇连贯性;然而 Ma^[32]的研究表明,中国学生作文中连词使用过多,甚至有滥用连词的现象,因此把连词这种衔接手段的出现频率作为评判学习者作文连贯性的标准可能是不合适的;而 Zhu^[33]提出连贯是一种语义概念。在语篇连贯性分析工具 Coh-Metrix^[34]中,段落间 LSA 这一变量计算同一文本内所有段落之间的语义相似性,该变量测量文本的潜在特征,可以解释文本中所有段落之间的连贯性。Liang^[27]表明,低质量的文章更多地依赖文章的局部连贯性,高质量的文章更多地依赖文章的整体连贯性。因此,本文使用篇章连贯性表示作文的整体逻辑。

2.4 小结

综上,本文认为文本的语义表达、组织质量和整体逻辑都是决定作文质量的重要因素,相比仅从单一角度出发对作文质量进行等级评价,多个视角使得作文质量评估更具可解释性。

3 多视角评价框架

基于对相关工作的研究和对作文评分任务的分析,以往的工作多数仅从语义表达或组织结构等视角评估文章质量,没有综合考虑各类因素,也没有考虑更高层次的诸如整体逻辑性这类因素。本文综合考虑各类因素对作文质量进行评估,从而更加客观地对作文进行等级评价。因此,本文提出了多视角评价框架(MPE),其结构如图 1 所示。具体而言,本文期望通过语义表达评估、组织结构评估和整体逻辑评估

这 3 个视角来获取不同的信息,以评价作文的整体质量。3 个视角的内容如下。

(1)语义表达评估:从多个层次出发由下往上依次获得句子级、段落级、篇章级的语义信息。综合考虑单词、句子和段落对作文质量的影响。

(2)组织结构评估:采用分层多任务学习的方法获得句子功能和段落功能,并将其用于篇章组织结构信息的表示。

(3)整体逻辑评估:获取篇章高层次的连贯性信息表示作文逻辑。

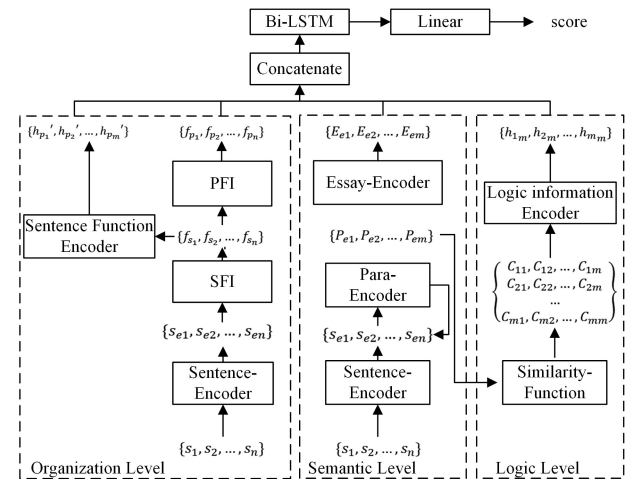


图 1 多视角评价(MPE)框架结构

Fig. 1 Structure of multi-perspective evaluation(MPE) framework

3.1 语义表达评估

对于作文语义表示, Alikaniotis 等^[8]、Tagh-ipour 等^[9]使用单层 LSTM 编码文章的单词序列得到篇章语义表示,仅捕获了单词级别的信息; Dong 等^[10-11]采用双层结构获得句子表示和篇章表示,考虑了单词和句子表示作文语义。与之前的模型不同,我们采用 3 层结构分别获得句子表示、段落表示和篇章表示,综合考虑单词、句子和段落对作文质量的影响。

3.1.1 句子语义表示

本文使用 XLNet 作为句子内容编码器从而获得更完整的句子语义信息。给定一个句子的单词序列 $s = \{w_1, w_2, \dots, w_n\}$, 使用 XLNet 对句子 s 进行编码, 得到句子级的语义信息表示 s_{ei} , 如式(1)所示:

$$s_{ei} = XLNet(s_i), i = 1, 2, \dots, n \quad (1)$$

3.1.2 段落级语义表示

在获得句子语义信息表示后, 我们使用段落内容编码器获得段落语义表示, 首先使用 Bi-LSTM 对句子序列为 $\{s_{e1}, s_{e2}, \dots, s_{en}\}$ 的段落进行编码, 得到隐藏状态 $h_p = \{h_1, h_2, \dots, h_n\}$ 。第 i 步的隐藏状态如式(2)所示:

$$h_i = BiLSTM(s_{ei}, h_{i-1}) \quad (2)$$

在 Bi-LSTM 的基础上, 本文叠加注意力层以获得段落级的语义信息表示 P_e , 叠加注意力层有助于获得不同句子对段落内容表示的重要程度。

3.1.3 篇章级语义表示

在获得段落语义表示后, 我们进一步使用篇章内容编码器获得篇章级的语义表示。具体来说, 我们使用 Bi-LSTM 对段落序列进行编码, 如式(3)所示。将隐藏层 $\{H_1, H_2, \dots, H_m\}$ 作为篇章级语义表示 $I_{sem} = \{E_{e1}, E_{e2}, \dots, E_{em}\}$ 。

$$H_i = Bi-LSTM(P_i, H_{i-1}) \tag{3}$$

3.2 组织结构评估

议论文的组织结构是评估议论文质量的一个重要因素，一篇好的议论文应具有合理的组织和清晰的结构，论点、论据、论证等成分缺一不可。作文组织结构评估所面临的问题是如何有效地表示组织结构以及如何将其与作文质量联系起来。

表2 篇章功能(句子功能、段落功能)的定义

Table 2 Definition of discourse function(sentence function, paragraph function)

句子功能	定义	段落功能	定义
Introduction	引言,文章背景介绍	Introduction-Para	包含引言但不包含中心论点和分论点的段落
Thesis	中心论点,表达作者对文章主题的主要观点	Thesis-Para	包含中心论点的段落
Main Idea	分论点,支持中心论点	Idea-Para	包含分论点但不包含中心论点句的段落
Evidence	论据,举例论证中心论点或分论点	Support-Para	包含论据或解说但不包含分论点和总结句的段落
Elaboration	解说,进一步解释中心论点或分论点	Conclusion-Para	包含总结句不包含中心论点句
Conclusion	总结,总结全文呼应中心		

3.2.1 句子功能识别和段落功能识别

(1)句子功能识别模块(SFI)

句子功能识别模块整合句子语义表示和句子位置编码来识别句子功能。其中句子语义表示使用3.1.1节中的句子级语义表示 $S = \{s_{e1}, s_{e2}, \cdots, s_{en}\}$ 。

1)位置编码。本文使用3种句子位置信息对句子的位置进行编码。其中,句子在一篇文章所有句子组成的句子序列中的位置为全局位置信息;句子所在段落在一篇文章的段落序列中的位置为段落位置信息;句子在其所在段落的所有句子组成的句子序列中的位置为局部位置信息。

这3种句子位置信息使用transformer模型中的正弦位置进行编码,分别表示为 $pos_{global}(i)$, $pos_{para}(i)$ 和 $pos_{local}(i)$,位置编码维度为 d 。最后使用线性层将3种位置信息整合为最终的位置信息表示,如式(4)所示:

$$pos(i) = \sum_{r \in \{global, local, para\}} \beta_r pos_r(i) \tag{4}$$

其中, $\{\beta_r\}$ 是训练期间需要学习的参数。

2)句子功能识别。本文将句子语义表示和位置编码相加输入Bi-LSTM得到隐藏层表示,取隐藏层的平均值输入非线性层,从而获得结合上下文的句子语义表示 $f_s = \{f_{s_1}, f_{s_2}, \cdots, f_{s_n}\} \in R^{d \times n}$,如式(5)、式(6)所示:

$$H = Bi-LSTM(S + pos) \tag{5}$$

$$f_s = \tanh(\text{mean}(H)) \tag{6}$$

最后本文将结合上下文的句子语义表示依次输入到线性层和Softmax层,从而获得每个句子功能标签的概率分布 P_{s_i} ,如式(7)所示:

$$P_{s_i} = \text{Softmax}(\text{Linear}(f_s)) \tag{7}$$

该模块的损失函数采用负对数似然损失函数,记为 L_{SFI} ,如式(8)所示:

$$L_{\text{SFI}}(\theta) = -\frac{1}{n} \sum_{i=1}^n \log(P_{s_i}) \tag{8}$$

其中, θ 表示模型的可训练参数。

(2)段落功能识别模块(PFI)

假设第 i 段包含的句子索引为 j 到 k ,将 f_s 作为句子功能表示输入到Bi-LSTM,并使用注意力机制^[35]筛选出重要的句子功能,最后得到段落功能表示,维度为 d ,如式(9)所示:

$$H_p = \text{Bi-LSTM}(f_{s_{[j:k]}}) \tag{9}$$

同样,使用线性层和Softmax层来获得段落功能标签的

遵循Song等^[17]的思想,本文使用句子功能和段落功能来评估议论文的组织结构。具体而言,首先识别出句子功能和段落功能,然后将句子功能表示和段落功能表示整合为组织评估信息。

表2列出了Song等^[17]定义的句子级别和段落级别的篇章功能。

概率分布 P_{p_i} 。损失函数采用负对数似然损失函数,记为 L_{PFI} ,如式(10)所示:

$$L_{\text{PFI}}(\theta) = -\frac{1}{m} \sum_{i=1}^m \log(P_{p_i}) \tag{10}$$

其中, θ 表示模型的可训练参数。

3.2.2 组织评估部分

本文将每个段落所包含句子对应的句子功能表示序列输入到句子功能编码器Bi-LSTM,取隐藏状态最后一层的输出作为每个段落内的上下文句子功能表示,最后得到序列 $\{h'_{p_1}, h'_{p_2}, \cdots, h'_{p_m}\}$, h'_{p_i} 表示第 i 段对应的上下文句子功能表示,维度为 d ,计算方法如式(11)所示:

$$H_s = \text{Bi-LSTM}(f_{s_{[j:k]}}) \tag{11}$$

最后,本文将每个段落的段落功能表示 f_p 和 h'_{p_i} 拼接,作为组织信息评估作文的组织结构, f_p 为段落功能表示,如式(12)所示:

$$I_{\text{ORG}} = \text{Concatenate}(h_p'; f_p) \tag{12}$$

3.3 整体逻辑评估

篇章连贯性是一种逻辑关系,这种逻辑关系将文本中的命题或思想有效地连接起来,使得文本在意义上具有连续性。研究表明高质量的文章更多地依赖文章的全局连贯性,因此本文使用全局连贯性对文章的整体逻辑进行评估。

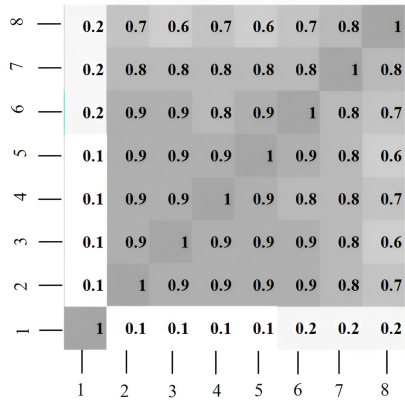
本文使用相似性函数计算各个段落之间的语义相似性,从而获得评估文章整体逻辑的连贯性矩阵。具体来说,本文使用余弦相似度(Cosine Similarity)计算段落间语义的相似性,段落语义表示使用3.1.2节中的段落语义表示序列 $\{P_{e_1}, P_{e_2}, \cdots, P_{e_m}\}$,最终获得连贯性矩阵(Essay Coherence Matrix, ECM),用 C 表示, C_{ij} 表示第 i 段与第 j 段的相似性,范围为 $0 \sim 1$,如式(13)所示:

$$C_{ij} = \begin{cases} \text{cosine}(P_i, P_j), & i \neq j \\ 1, & i = j \end{cases} \tag{13}$$

如图2所示,矩阵每一行对应一个段落的相似性序列。如式(15)所示,本文将第 i 段的相似性序列 $(C_{i_1}, C_{i_2}, \cdots, C_{i_m})$ 编码后输入Bi-LSTM,获取隐藏层最后一层作为该段落与全局的联系 h_{i_m} ,维度为 d 。

$$(h_{i_1}, h_{i_2}, \cdots, h_{i_m}) = \text{Bi-LSTM}(C_{i_1}, C_{i_2}, \cdots, C_{i_m}) \tag{14}$$

最后,将 $I_{\text{LOG}} = (h_{1_m}, h_{2_m}, \cdots, h_{m_m})$ 作为逻辑信息来评估文章的整体逻辑。



注: X 轴和 Y 轴表示段落位置, 矩阵中数字为段落间的语义相似性, 颜色由浅至深表示段落间相似性由低到高

图 2 连贯性矩阵(ECM)

Fig. 2 Coherence matrix(ECM)

本文将评估文章语义表达的语义信息、文章组织结构的组织信息和文章整体逻辑的逻辑信息, 将其线性拼接后依次输入到 Bi-LSTM、线性层和 Softmax 层, 得到作文等级的概率分布 P_E , 如式(15)、式(16)所示:

$$H = \text{Bi-LSTM}(\text{Concatenate}(I_{\text{SEM}}; I_{\text{ORG}}; I_{\text{LOG}})) \tag{15}$$

$$P_E = \text{softmax}(\text{Linear}(H)) \tag{16}$$

损失函数采用负对数似然函数, 记为 L_E , 如式(17)所示:

$$L_E(\theta) = -\frac{1}{n} \sum_{i=1}^n \log(P_E) \tag{17}$$

最终得到损失函数, 如式(18)所示:

$$L = L_{\text{SFI}} + L_{\text{PFI}} + \gamma \cdot L_E \tag{18}$$

参照 Song 等^[17]的做法, 本文假设 SFI 和 PFI 同等重要, 使用超参数 γ 来控制综合评估的相对重要性, 本文采用式(19)的方式动态更新 γ 。

$$\gamma = \max\left(\min\left(\frac{L_E}{L_{\text{SFI}}} \cdot \gamma, 1\right), 0.01\right) \tag{19}$$

其中, γ 初始值设为 0.1, 因此该模型会优先优化句子功能识别模块和段落功能识别模块, 当 L_{SFI} 小于 L_E 时, γ 值会变得更

4 实验

为了验证本文所提出的多视角评价框架(MPE), 本文使用 Song 等^[17]发布的数据集进行实验, 本节首先介绍实验数据、实验设置和基准系统, 然后给出实验结果及实验分析。

4.1 数据

本文采用 Song 等^[17]发布的数据集进行实验。该数据集是从网站乐乐课堂¹⁾上选取的 1 220 篇高中议论文, 所有文章的平均段落数为 8, 平均句子数为 28, 每个句子平均词语数为 21。数据集中包含已标注的篇章功能(句子功能和段落功能)。作文等级则使用乐乐课堂网站上的等级评分, 并由两位标注者对存有异议的作文进行讨论以确认作文的最终等级, 最终分为 3 个等级: 差(Bad)、中(Medium)、优(Great)。其中, Bad 等级的文章有 245 篇, Medium 等级的文章有 670 篇, Great 等级的文章有 305 篇。

4.2 实验设置

本文按照 Song 等^[17]的方法对数据进行预处理, 每篇文章包含的句子数、段落数以及每个段落包含的句子数(n, m, n_p)的阈值被设为 50, 20, 20。短于或长于阈值的句子和段落被填充或截断。

本文按照 Song 等^[17]的数据处理方式使用 5 折交叉验证进行实验, 每一折在作文等级上具有相似的分

布。训练过程中随机选取 10% 的训练数据作为验证集。梯度下降优化器使用随机梯度下降(Stochastic Gradient Descent, SGD)。

本文使用的评价指标有 4 个, 分别是 QWK(Quadratic weighted Kappa)、均方误差(Mean Squared Error, MSE)、平均绝对误差(Mean Absolute Error, MAE)和宏 F1。作文的 3 个等级表示为 0, 1, 2。其中, MSE 和 MAE 的表达式如式(20)、式(21)所示, 它们的值越小, 表明实验结果具有更高的精确度。

$$\text{MSE} = \frac{1}{m} \sum_{i=1}^m (y - \hat{y})^2 \tag{20}$$

$$\text{MAE} = \frac{1}{m} \sum_{i=1}^m |y - \hat{y}| \tag{21}$$

其中, y 表示作文实际等级, $\hat{y}$ 表示作文预测等级。

4.3 基准系统

本文与下列方法进行比较。

(1) Dong 等提出的模型^[11]。该模型使用 CNN 获取句子语义信息, 使用 LSTM 获取篇章语义信息, 并使用注意力机制捕捉关键的单词信息和句子信息, 只考虑文章的语义表达来学习预测作文等级。

(2) Dong 等提出的模型^[16]。该模型使用段落功能序列和句子功能序列的 $n\text{grams}$ 作为特征, 并使用 SVM 学习预测作文组织等级。

(3) Dong 等提出的模型^[17]。该模型使用多任务学习模型(SFI+PFI)来获得句子和段落功能表示, 并利用网格将句子功能表示和段落功能表示相结合来评估议论文的组织结构, 最后使用 2dCNN 来学习预测作文组织等级。

(4) XLNet 文本分类模型^[25]。本文使用 XLNet 模型获得段落语义表示, 将其拼接后通过线性层来预测作文等级。

4.4 实验结果

实验结果如表 3 所列, 其中 MPE(-LOG-ORG)表示模型仅考虑文章的语义表达, MPE(-LOG)表示模型仅考虑文章的语义表达和组织结构, MPE(-ORG)表示模型仅考虑文章的语义表达和整体逻辑。

表 3 作文评估结果
Table 3 Essay evaluation results

Systems	F1	QWK	MAE↓	MSE↓
Dong 等的模型 ^[11]	0.474	0.446	0.539	0.480
Persing 等的模型 ^[16]	0.478	0.449	0.386	0.383
Song 等的模型 ^[17]	0.591	0.561	0.391	0.373
XLNet	0.584	0.556	0.338	0.411
MPE	0.664	0.608	0.315	0.333
MPE(-LOG)	0.641	0.592	0.357	0.342
MPE(-ORG)	0.637	0.588	0.343	0.325
MPE(-LOG-ORG)	0.625	0.578	0.352	0.338

¹⁾ <http://www.leleketang.com/zuowen/>

实验结果表明,本文模型在整体上明显优于所有基准系统。相比文献[11]仅考虑句子和篇章语义表示,MPE使用预训练模型 XLNet 获得句子语义表达,对于中文文本的表示无需分词且获得了更深层次的语义信息,并且 MPE 考虑了句子、段落和篇章 3 个层次的语义信息。相比文献[16-17],MPE 进一步考虑了文本的语义和逻辑因素,使得评分结果更具可解释性。

相比预训练模型 XLNet,MPE 的 QWK 提升了 9%,宏 F1 提升了 14%,MAE 降低了 7%,MSE 降低了 19%,这说明本文考虑的文章的语义表达、组织结构以及整体逻辑可以有效提升 MPE 对不同质量的作文进行评分的性能。

相比目前效果最好的基线文献^[17],MPE 的 QWK 提升了 8%,宏 F1 提升 12%,MAE 降低了 19%,MSE 降低了 11%。这是因为文献[17]仅考虑了文章的组织结构,而 MPE 不仅考虑了文章的组织结构,还考虑了决定议论文质量的重要因素,即语义表达和整体逻辑。如在使用文献[17]的模型对文献[36]进行评判时其等级为 Great,因为该文章内容完整,论点、论证和论据分布合理,但该文章的论述语言无感染力,论证缺乏层次条理,因此本文模型在考虑了语义表达和整体逻辑后,对其评分为 Bad。

因此,文章的最终等级是由多个角度综合考虑所决定的,仅从单一角度出发会对一些文章进行错误的评估。

4.5 实验分析

为了比较文章的语义表达(Semantic,SEM)、组织结构(Organization,ORG)以及整体逻辑(Logic,LOG)3个因素对 MPE 结果的影响,本文进行了一系列的消融实验。

4.5.1 语义表达评估分析

如表 3 所列,相比使用预训练模型 XLNet 的基准系统,MPE(-LOG-ORG)方法仅考虑语义表达,其 QWK 提升了 4%,其宏 F1 的结果提升了 7%,MSE 下降了 18%。而在表 4 对各类别消融实验的结果中,相比 XLNet,MPE(-LOG-ORG)使得 3 个作文等级的 F1 值均有明显提升,这表明 MPE 中语义表达评估模块所获取的由低到高 3 个层次的语义信息能够提高对作文全局语义信息的理解,从而提高每个等级作文的识别性能。

表 4 各类别消融实验结果

Table 4 Results of ablation experiments by category

Essay grade	Systems	F1
Bad	XLNet	0.4817
	MPE(-LOG-ORG)	0.5184
	MPE(-ORG)	0.5362
	MPE(-LOG)	0.5689
	MPE(ALL)	0.5919
Medium	XLNet	0.6927
	MPE(-LOG-ORG)	0.7258
	MPE(-ORG)	0.7467
	MPE(-LOG)	0.7033
	MPE(ALL)	0.7400
Great	XLNet	0.5774
	MPE(-LOG-ORG)	0.6318
	MPE(-ORG)	0.6291
	MPE(-LOG)	0.6499
	MPE(ALL)	0.6584

相比文献[17]仅考虑文本的组织结构,MPE 仅考虑文本

语义表达,其 QWK 提升了 3%,宏 F1 的结果提升了 5.8%,MAE 和 MSE 分别下降了 10%和 9.4%,这是因为相比作文的组织结构,作文的语义表达是决定作文质量好坏的基本因素。

与仅考虑文章的语义信息来学习预测作文等级的基线文献[11]相比,MPE(-LOG-ORG)的 QWK 提升了 30%,宏 F1 提升了 31.8%,MAE 和 MSE 分别下降了 34.7%和 29.6%。这是因为文献[11]更关注单词和句子对作文语义表达的影响,而英文作文的单词使用是决定其作文质量的一个重要因素,中文作文在语义表达上则更注重句子和段落的流畅性以及句式的灵活性。因此,本文提出由低到高分别获得句子级、段落级和篇章级的文本语义表达,从而在整体上提升作文的识别性能。

4.5.2 组织结构评估分析

如表 3 所列,在语义表达的基础上进一步考虑文章的组织结构,即 MPE(-LOG)相比 MPE(-LOG-ORG)在 QWK 和宏 F1 上有所提升,在 MAE 和 MSE 上基本不变。这表明不同等级的作文在组织结构上也有不同,组织结构信息的增加能进一步帮助 MPE 区分作文等级。

从表 4 对各类别消融实验结果的分析来看,在语义表达的基础上进一步考虑文章的组织结构 MPE(-LOG),等级为 Bad 的作文宏 F1 值提升了近 10%,等级为 Great 的作文宏 F1 值提升了 3%,而等级为 Medium 的作文宏 F1 值下降了 3%。这说明文本的语义表达在根本上不能决定文章好坏,根据实验结果分析,当仅考虑文本语义表达时,一些等级为 Great 和等级为 Bad 的文章被模型判定为 Medium 等级,而加入结构信息后,原本识别错误的文章又被识别正确。如表 5 所列,该文章在语义表达上语义流畅,善用修辞,但其内容空泛,仅考虑语义表达对其评分得到的等级为 Medium,而这篇文章在结构上缺少论据,因此在考虑了文章的组织结构后,即使用 MPE(-ORG)对其评分得到的等级为 Bad,这说明文章的组织结构也是决定其质量的重要因素。

表 5 作文样例分析

Table 5 Sample analysis of essay

小议宽容
(Introduction)宽容别人是对别人的理解,是一种放得下的大度,是一种与人为善的观念释然。而宽容自己则是一种豁达、冷静与理智,宽容自己并不是放纵自己。
(Main idea)人应该学会宽容。(Elaboration)多一些宽容就少一些心灵的隔膜;多一份宽容,就多一份理解,多一份信任,多一份友爱。
(Main idea)我觉得宽容就是在心里上接纳别人,理解别人的处事方法,尊重别人的处事原则。(Main idea)……
(Elaboration)宽容是一种非凡的气度、宽广的胸怀,是对人对事的包容和接纳。
(Elaboration)宽容是一种高贵的品质、崇高的境界,是精神的成熟、心灵的丰盛。
(Elaboration)宽容是一种仁爱的光芒、无上的福分,是对别人的释怀,也即是对自己善待。
(Elaboration)宽容是一种生存的智慧、生活的艺术,是看透了社会人生以后所获得的那份从容、自信和超然(Elaboration)……
(Conclusion)宽容有 3 种境界,可以养鱼为喻:最初级的境界是……;中等境界是……;最高境界则是……

4.5.3 整体逻辑评估分析

如表 3 所列,在语义表达的基础上进一步考虑文章的

整体逻辑,即相比 MPE(-LOG-ORG),MPE(-LOG)的 QWK 和 宏 F1 有所提升,在 MAE 和 MSE 上有所下降。这表明,文本段落间的语义相似性所捕捉的文本连贯性信息能够进一步提升 MPE 对不同等级作文的识别能力。

从表 4 所列的对各类别消融实验结果的分析来看,在语义表达的基础上进一步考虑文章的整体逻辑,等级为 Bad 的作文宏 F1 提升了 3.4%,等级为 Medium 的作文宏 F1 提升了 2.9%,而等级为 Great 的作文宏 F1 基本不变。根据实验结果分析,一些原本为 Bad 等级的文章被识别为 Medium,而一些原本为 Medium 等级的文章被识别为 Great,而高质量的文章更多地依赖文章的全局连贯性,即整体逻辑,因此,在考虑逻辑信息后,一些被识别为更高等级的文章识别正确。

4.5.4 综合分析

本文提出综合语义表达、组织结构和整体逻辑 3 个因素对作文进行等级评价,这是因为仅考虑其中某一个因素并不能决定文章质量的好坏。

如表 3 所列,MPE(-ORG)相比 MPE(-LOG-ORG)其 QWK 提升了 1.7%,而 MPE 相比 MPE(-LOG)其 QWK 提升了 2.7%;MPE(-LOG)相比 MPE(-LOG-ORG)其 QWK 提升了 2.4%,而 MPE 相比 MPE(-ORG)其 QWK 提升了 3.4%,这说明组织结构和整体逻辑对模型预测作文等级起到了促进的作用。

4.6 错误分析

在对实验结果进行分析的过程中发现,识别错误的少部分为 Medium 等级的文章被识别为 Bad 等级或 Great 等级,而更多的则是 Bad 等级和 Great 等级的文章被识别为 Medium。通过分析,这是因为 Medium 等级的文章在各个角度的等级不同,如表 5 中的文章所示,这篇文章在语义表达上接近 Medium 等级,但其在结构和逻辑上却接近它原有的等级 Bad。我们的实验数据中还有许多类似的文章,其中有的 Great 等级的文章在整体结构上接近 Medium 等级,就被识别为 Medium,有的 Bad 等级的文章在语义表达上接近 Medium 等级,就被识别为 Medium。这也说明,各个角度对作文等级高低的决定程度是不一样的。而目前汉语作文缺少各个角度的评分数据,因此这也是我们以后的工作期望改进的一点。

结束语 本文提出了一种对议论文质量进行评估的多视角评价框架(MPE)。该框架从作文的语义表达、组织结构和整体逻辑 3 个视角出发,以篇章功能表示组织结构,从句子级、段落级、篇章级语义表示出发评估文章的语义表达,以文章的篇章连贯性表示作文整体逻辑,最后汇总这 3 个方面的评估特征对作文整体进行评分。实验结果表明,这种方法对议论文质量评估是有效的。另外,实验分析显示从多个视角出发使得实验结果更具可解释性,根据实验结果可以看出每个视角都能提升作文评分效果。在未来,我们将使用多任务学习对作文的不同特征以及作文的整体进行评分,从而为读者提供更有价值的见解。

参 考 文 献

[1] MESGAR M,STRUBE M. A neural local coherence model for text quality assessment[C]// Proceedings of the 2018 Confe-

rence on Empirical Methods in Natural Language Processing. 2018:4328-4339.

[2] LIU J,XU Y,ZHU Y. Automated essay scoring based on two-stage learning[J]. arXiv:1901.07744,2019.

[3] YANG Y,ZHONG J. Automated essay scoring via example-based learning[C]// International Conference on Web Engineering. Cham:Springer,2021:201-208.

[4] CHEN H,HE B. Automated essay scoring by maximizing human-machine agreement [C]// Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013:1741-1752.

[5] SOMASUNDARAN S,BURSTEIN J,CHODOROW M. Lexical chaining for measuring discourse coherence quality in test-taker essays[C]// The 25th International Conference on Computational Linguistics; Technical papers(COLING 2014). 2014:950-961.

[6] YANNAOUDAKIS H,BRISCOE T,MEDLOCK B. A new dataset and method for automatically grading ESOL texts [C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics; Human Language Technologies. 2011:180-189.

[7] PHANDI P,CHAI K M A,NG H T. Flexible domain adaptation for automated essay scoring using correlated linear regression [C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015:431-439.

[8] ALIKANIOTIS D,YANNAOUDAKIS H,REI M. Automatic text scoring using neural networks[J]. arXiv:1606.04289, 2016.

[9] TAGHIPOUR K,NG H T. A neural approach to automated essay scoring[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 2016:1882-1891.

[10] DONG F,ZHANG Y. Automatic features for essay scoring—an empirical study[C]// Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 2016:1072-1077.

[11] DONG F,ZHANG Y,YANG J. Attention-based recurrent convolutional neural network for automatic essay scoring[C]// Proceedings of the 21st Conference on Computational Natural Language Learning(CoNLL 2017). 2017:153-162.

[12] SOMASUNDARAN S,FLOR M,CHODOROW M,et al. Towards evaluating narrative quality in student writing[J]/OL. Transactions of the Association for Computational Linguistics, 2018, 6: 91-106. [https://direct. mit. edu/tac/article/doi/10.1162/tac_a_00007/43428/Towards-Evaluating-Narrative-Quality-In-Student](https://direct.mit.edu/tac/article/doi/10.1162/tac/article/doi/10.1162/tac_a_00007/43428/Towards-Evaluating-Narrative-Quality-In-Student).

[13] PERSING I,NG V. Modeling stance in student essays [C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016: 2174-2184.

[14] MATHIAS S,BHATTACHARYYA P. Thank “Goodness”! A Way to Measure Style in Student Essays[C]// Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications. 2018:35-41.

[15] KE Z,INAMDAR H,LIN H,et al. Give me more feedback II:

Annotating thesis strength and related attributes in student essays[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019:3994-4004.

[16] PERSING I, DAVIS A, NG V. Modeling organization in student essays[C]// Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 2010:229-239.

[17] SONG W, SONG Z, LIU L, et al. Hierarchical Multi-task Learning for Organization Evaluation of Argumentative Student Essays[C]//IJCAI. 2020:3875-3881.

[18] CHEN Y. Convolutional neural network for sentence classification[D]. Canadian:University of Waterloo, 2015.

[19] SHI X J, CHEN Z, WANG H, et al. Convolutional LSTM network: A machine learning approach for precipitation nowcasting [C] // Advances in Neural Information Processing Systems, 2015:802-810.

[20] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11):2278-2324.

[21] WON Y. The Prediction of Writing Scores Using Vocabulary Features in ESL University Students' Essays[J]. Modern English Education Society, 2019, 20(4):31-40.

[22] LE Q, MIKOLOV T. Distributed representations of sentences and documents [C] // International Conference on Machine Learning. PMLR, 2014:1188-1196.

[23] LIU C, ZHAO S, VOLKOV S M. Unsupervised document embedding with cnns[J]. arXiv:1711.04168, 2017.

[24] WU L, YEN I E H, XU K, et al. Word mover's embedding: From word2vec to document embedding[J]. arXiv:1811.01713, 2018.

[25] YANG Z, DAI Z, YANG Y, et al. XLNet: Generalized autoregressive pretraining for language understanding [J/OL]. Advances in neural information processing systems, 2019, 32. <https://proceedings.neurips.cc/paper/2019/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html>.

[26] ATTALI Y, BURSTEIN J. Automated essay scoring with e-rater © V. 2 [J/OL]. The Journal of Technology, Learning and Assessment, 2006, 4(3). <https://ejournals.bc.edu/index.php/jtla/article/view/1650>.

[27] LIANG M C. A study of coherence in EFL learners' written production [J]. Modern Foreign Languages, 2006, 29(3):284-292.

[28] LOUIS A, NENKOVA A. A coherence model based on syntactic patterns[C]//Proceedings of the 2012 Joint Conference on Em-

pirical Methods in Natural Language Processing and Computational Natural Language Learning, 2012:1157-1168.

[29] MILTSAKAKI E, KUKICH K. Evaluation of text coherence for electronic essay scoring systems[J]. Natural Language Engineering, 2004, 10(1):25-55.

[30] LIAO D, XU J, LI G, et al. Hierarchical Coherence Modeling for Document Quality Assessment[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(15):13353-13361.

[31] XU W C. Cohesion, coherence and quality in English composition[J]. Journal of GuangZhou University, 2000(5):71-75.

[32] MA G G. A comparative analysis of the linguistic features of English composition between Chinese and American College Students [J]. Foreign Language Teaching Research, 2002, 34(5):345-350.

[33] ZHU Y S. Halliday's standard of discourse coherence is misunderstood by the outside world and its own shortcomings[J]. Foreign Language Teaching and Research, 1997(1):23-27.

[34] MCNAMARA D S, LOUWERSE M M, GRAESSER A C. CohMetrix: Automated cohesion and coherence scores to predict text readability and facilitate comprehension[R]. Technical report, Institute for Intelligent Systems, University of Memphis, Memphis, TN, 2002.

[35] VASWANIA, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems. 2017:5998-6008.

[36] 乐乐, 问题 [OL]. <http://www.leleketang.com/zuowen/287886.shtml>.



HE Yaqiong, born in 1997, postgraduate, is a member of China Computer Federation. His main research interests include natural language processing and so on.



LI Peifeng, born in 1971, Ph.D, professor, Ph.D supervisor, is a member of China Computer Federation. His main research interests include natural language processing and machine learning.

(责任编辑:喻藜)