文章编号:1671-9352(2025)09-0071-16 DOI:10.6040/j.issn.1671-9352.0.2024.039

基于情节描述的中文长篇小说高潮章节识别方法

王文晶[1], 刘忠宝[2*], 万广文[2], 胡迦南[3]
(1.山西工程科技职业大学信息工程学院, 山西 太原 030619; 2.北京语言大学信息科学学院, 北京 100083; 3.中北大学软件学院, 山西 太原 030051)

摘要:在精准刻画中文长篇小说情节的基础上,探讨中文长篇小说高潮章节识别方法。该方法由关键要素抽取和高潮章节识别 2 部分组成,其中前者包括观点段落、非观点段落、章节关键词、主要角色等关键要素抽取,后者在建立章节情节描述矩阵的基础上,引入 BiGRU 模型与多头注意力机制,实现中文长篇小说高潮章节识别。金庸小说语料集上的比较实验表明,与朴素贝叶斯(naive Bayesian, NB)、支持向量机(support vector machine, SVM)、预训练模型 Roberta-large、双向长短时记忆网络(bi-directional long short-term memory, BiLSTM)等模型相比,本文所提方法具有更优的识别性能。消融实验验证所提方法主要组成部分的有效性。

关键词:中文长篇小说;关键要素抽取;章节情节描述矩阵;高潮章节识别
中图分类号:TP391 **文献标志码**:A
引用格式:王文晶, 刘忠宝,万广文,等. 基于情节描述的中文长篇小说高潮章节识别方法[J]. 山东大学学报(理学版),2025,60(9):71-86.

Climaxchapter recognition method of chinese long novel based on plot description

WANG Wenjing[1], LIU Zhongbao[2*], WAN Guangwen[2], HU Jianan[3]
(1. College of Information Engineering, Shanxi Vocational University of Engineering Science and Technology, Taiyuan 030619, Shanxi, China; 2. School of Information Science, Beijing Language and Culture University, Beijing 100083, China; 3. School of Software, North University of China, Taiyuan 030051, Shanxi, China)

Abstract: How to quickly and accurately identify the climax chapter has become a common problem faced by the majority of readers in their reading choices. In view of this, the method of identifying the climax chapters of Chinese long novel on the basis of accurately portraying the plot of Chinese long novel is explored, which consists of two parts, namely, key element extraction and climax chapter recognition, where the former includes the extraction of key elements such as viewpoint and non-viewpoint passages, keywords of the chapter, main characters, etc., and the latter, based on the establishment of the chapter plot description matrix, introduces the BiGRU model and the multi-head attention mechanism to realize the climax chapter recognition of Chinese long novel. Comparative experiments on Jin Yong's novel corpus show that the proposed method in this paper has better recognition performance compared with models such as Naive Bayesian (NB), Support Vector Machine (SVM), pre-trained model named Roberta-large, and Bi-directional Long Short-Term Memory (BiLSTM). Ablation experiments validate the effectiveness of the main components of the proposed method.

Key words: Chinese novel; main component extraction; chapter plot description matrix; climax chapter recognition

0 引言

中文长篇小说由于人物丰富多彩、情节跌宕起伏,一直深受读者的喜爱。读者通过故事情节感受小说表

收稿日期:2024-01-30; **网络出版时间**:2024-06-13 12:25:46
基金项目:国家社科基金重点项目"大数据时代古籍活化赋能文化自信自强的理论、方法与路径研究"(23AZD047)
第一作者:王文晶(1981—),女,副教授,硕士,研究方向为智能计算. E-mail:806214106@qq.com
*通信作者:刘忠宝(1981—),男,教授,博士,研究方向为数字人文、文化数字化. E-mail:liuzb@nuc.edu.cn

达的主题思想以及主要角色的性格特征。作为故事情节的重要组成部分,高潮章节是小说基本矛盾冲突发展到最紧张、最尖锐的部分,是决定主要角色命运、发展前景以及事物成败的关键内容。准确找到小说的高潮章节,对于吸引读者阅读兴趣至关重要。高潮章节具有以下特点:一是高潮章节往往分散在小说文本之中;二是高潮章节一般围绕少量主题展开故事情节;三是高潮章节通常包含强烈的情感关系。中文长篇小说具有篇幅庞大、人物数量丰富、人物关系复杂、情节时空跨度大等特点,进一步加大了小说高潮章节识别的难度。

随着大数据时代的到来,大数据技术与方法成为重要的技术样态。大量研究表明,大数据技术与方法适用于处理规模庞大、关系多样、结构复杂的数据。因此,如何充分发挥大数据技术与方法的优势,进行小说高潮章节自动识别,值得深入探讨。鉴于此,本文在抽取观点段落、非观点段落、章节关键词、主要角色等关键要素的基础上,建立由章节情节矩阵和章节情节差异矩阵组成的章节情节描述矩阵,利用 BiGRU 模型与多头注意力机制对章节情节描述矩阵进行特征提取与融合,进而实现中文长篇小说的高潮章节识别。

1 相关研究

中文长篇小说典型研究成果有:肖天久和刘颖[1]利用主成分分析和文本分类方法,从句子破碎度、从众性等方面对金庸与古龙的小说进行对比研究。研究表明,金庸小说从众性高于古龙,而且较多使用俚语方言,口语性更强,两人在语法结构、短语结构、文本节奏、文本可读性、语言变化程度上具有较大差异。姚睿琦等[2]以《射雕英雄传》和《神雕侠侣》为研究对象,利用中文信息处理技术和社会网络分析方法探究了小说角色关系,该研究为大数据时代小说定量研究提供了新的思路。张旋等[3]提出一种基于复杂网络的小说角色关系识别模型,该模型对文本长度、角色关系复杂度和情节的时间序列特征均具有较高的鲁棒性。郜沁清等[4]将金庸小说作为研究对象,在抽取角色、环境、情节等元素的基础上,引入文化词典匹配金庸小说中的道德与文化元素,探讨了金庸文化思想的多元性和包容性,并绘制了电子地图与门派分布图。Liu 和 Xiao[5]将古龙最具代表性的 16 部小说作为研究对象,并将其分为前期、中期和末期 3 个时期,选取平均段落长度、词语长度、句子长度、词长离散度等特征,研究了古龙的写作风格,并利用层次聚类算法对小说进行聚类。研究表明,古龙小说的风格在不同时期存在明显差异。Xia 等[6]将金庸小说作为研究对象,通过统计词性、句长及其分布,探究了金庸小说的用词、句法、地理特征。研究表明:在用词方面,金庸小说含有大量的名词、动词、副词和代词;在句法方面,金庸小说的语言风格较为口语化;在地理方面,金庸小说具有广泛的文学地理分布。

时至今日,小说高潮章节识别相关研究还不多。小说高潮章节识别与篇章级情感分析有关,两者均将小说文本作为研究对象,在融合上下文语义、领域知识和语句信息的基础上进行情感极性判断。不同之处在于:篇章级情感分析考虑整个篇章的情感极性[7],高潮章节识别考虑章节情感极性的波动变化情况。鉴于此,本文梳理了中文情感分析相关文献,为小说高潮章节识别提供借鉴参考。小说情感分析大多采用情感词典嵌入、情感词抽取、机器学习算法等。Kim 和 Klinger[8]将 20 篇短篇小说作为研究对象,探讨了表情、手势、姿势、声音等特征在情感分析中的作用。Zehe 等[9]将 1750 年至 1920 年的 212 部德国小说分为圆满结局和不圆满结局 2 类,将每部小说分成 n 个等长的片段,根据 NRC 情感词典[10]计算每个片段的情感值,利用支持向量机(support vector machine, SVM)进行情感分类。Horton 等[11]将《汤姆叔叔的小屋》等 19 世纪美国小说作为研究对象,为小说各章节标记相应的情感强度,利用朴素贝叶斯(naive Bayesian, NB)算法得到不同情感强度的词语及其使用章节,为小说情感分析做出了有益探索。Yu[12]分别利用 NB 和 SVM 算法,探究了 19 世纪美国情感小说以及艾米莉·迪金森书信中情色诗的情感分析问题。

对上述研究梳理可以看出,目前鲜有小说高潮章节识别研究,但与之相关的小说情感分析取得了一些进展。随着研究的深入,这些研究面临一些重要挑战:首先,基于情感词典嵌入、情感词抽取、统计理论的研究依赖于领域专家知识、人工标注数据和识别规则,耗时耗力;其次,机器学习算法难以利用丰富的文本语义信息及其上下文依赖关系,导致该方法无法取得令人满意的情感识别效果;最后,小说情感分析大多基于英文文本,目前鲜有中文长篇小说情感分析研究。因此,本文在借鉴小说情感分析研究成果的基础上,对中文长篇小说高潮章节识别问题展开深入研究,以期丰富中文长篇小说研究的技术体系和方法体系,拓展新一代信

息技术背景下文学作品的研究思路。

2　研究框架

金庸是著名的中文长篇小说家,擅长撰写武侠小说,其代表作有《射雕英雄传》《神雕侠侣》《倚天屠龙记》和《笑傲江湖》等。本文将《射雕英雄传》《神雕侠侣》《倚天屠龙记》和《笑傲江湖》4 部小说组成的金庸小说语料集作为研究对象,给出如图 1 所示的研究框架。该框架包括 2 部分:① 关键要素抽取。首先,搜集并整理金庸小说语料集小说文本,并利用 Hanlp 模型对其段落进行情感值打分及中文分词;然后,利用 *K*-means++算法将金庸小说语料集的每个段落聚类为正向情感段落、负向情感段落、平淡段落 3 类,并将正向情感段落和负向情感段落组成观点段落 O,平淡段落组成非观点段落 $\overline{O}$;接着,将 TF-IDF 算法和隐含狄利克雷分布(latent Dirichlet allocation, LDA)算法抽取的关键词和主题词做交集,得到金庸小说语料集的章节关键词;最后,在金庸人物集合的指导下,统计小说文本中所有角色出现的频次,按角色频次排列并拟合为曲线,将曲线二阶导为 0 之前的角色作为主要角色[13]。② 高潮章节识别。首先,根据前面得到的关键要素,建立章节情节矩阵和章节情节差异矩阵,两者共同组成章节情节描述矩阵;然后,利用双向门控循环单元(bidirectional gate recurrent unit, BiGRU)模型与多头注意力机制对章节情节描述矩阵进行特征提取与融合,得到小说文本的深层语义特征;最后,将特征送入全连接神经网络,并利用 softmax(·)函数进行归一化处理,得到输出概率,最终选择概率最大的值作为高潮章节识别结果。

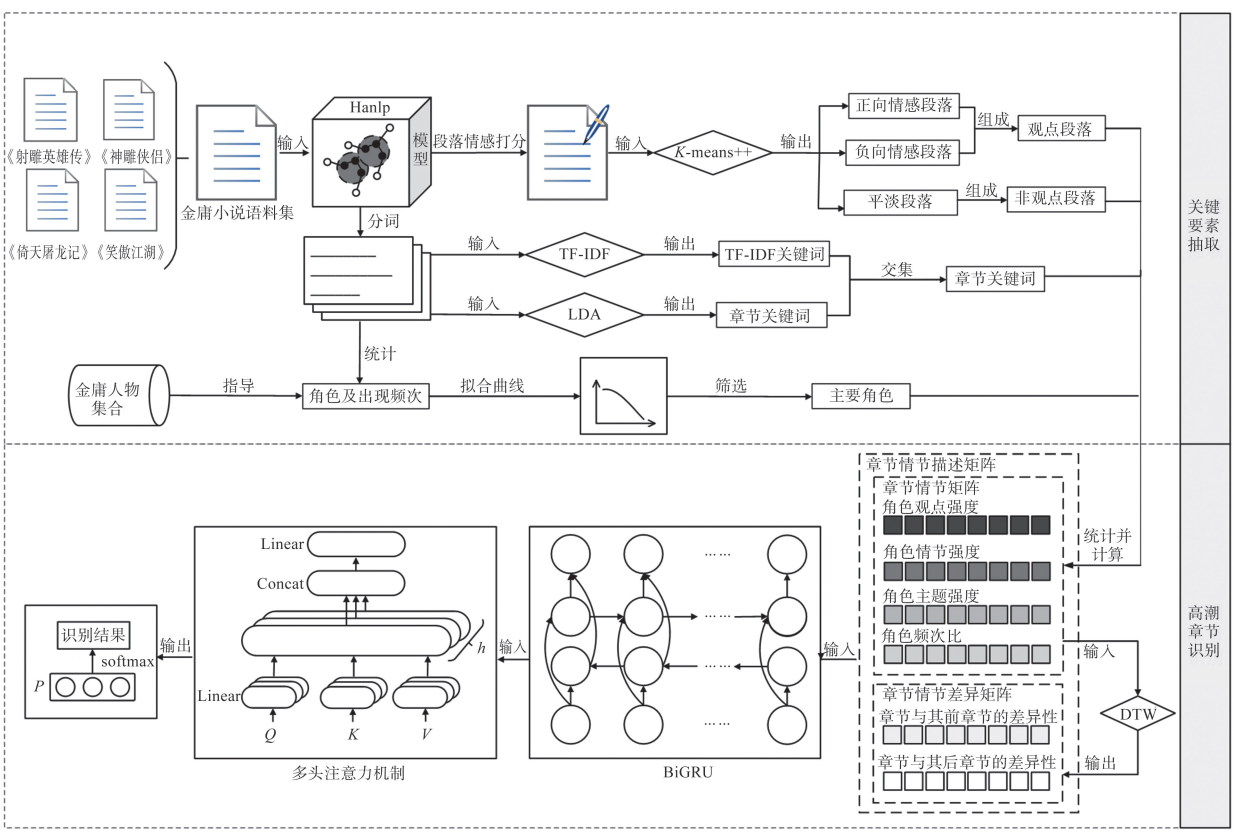

图 1　研究框架

Fig.1　Research frame

3　关键要素抽取

小说文本包含多个描述章节内容的关键词,章节关键词的多少决定了情节描述的丰富程度[14]。一般而言,一个段落情感越强烈,情节描述词越多,包含该段落的章节就越有可能是高潮章节。因此,分别引入角色情节强度和角色观点强度来描述主要角色与章节内容的关联程度以及主要角色观点的丰富程度(即主要角

色与章节关键词的共现次数。

小说文本的高潮章节往往出现众多角色[15]。描写主要角色的笔墨越多,与章节主题词的关系越密切,主要角色与章节主题词的关联程度就越高,章节的描述也就越精彩,越有可能是高潮章节。因此,分别引入角色频次比和角色主题强度来表征主要角色出现的频次比和主要角色与章节主题的契合度。

小说文本的情节发展基本遵循"开头铺垫、中间高潮、最后收尾"的逻辑顺序[3,13]。通过对比每一章节与其前、后章节精彩程度的变化,可以判断其是否为高潮情节。因此,引入章节情节差异特征,构造章节情节差异矩阵,用以表征每一章节与其前、后章节精彩程度的变化。

3.1 主要角色集合

小说文本的故事情节往往围绕主要角色展开,相较于其他角色,主要角色对于情节发展具有重要作用。本文首先利用 Hanlp 模型[16]进行中文分词,并在金庸人物集合的指导下统计小说文本中所有角色出现的频次;然后,按角色出现频次排列并拟合为曲线;最后,将曲线二阶导为 0 之前的角色作为主要角色。一系列主要角色组成主要角色集合 R。

3.2 观点段落集合

小说文本段落往往包含积极、消极、中性等情感极性。情感极性高,则表明该段落包含观点。观点段落通常给读者提供更多的指向性信息,也为小说高潮章节识别提供重要线索。

情感极性识别常用的方法是,利用情感词典并设定相应的语义规则来计算情感极性。但该方法存在如下不足:首先,通用领域情感词典在小说文本上的效果欠佳;其次,面向小说文本的情感词典构建依赖于人工挑选种子词,该方式耗时耗力;最后,小说文本语言简洁、情感表达隐晦,构建语义规则较为困难。因此,本文借助 Hanlp 情感分析模型计算小说段落的情感极性,根据 Hanlp 模型对段落的打分,利用 K-means++算法将金庸小说语料集的每个段落聚类为正向情感段落、负向情感段落、平淡段落,将正向情感段落和负向情感段落组合为观点段落 O,平淡段落组合为非观点段落 $\overline{O}$。

3.3 章节关键词集合

章节关键词集合 W 与章节主题紧密联系,用以描述章节的主要内容。章节关键词集合获取过程是:首先,利用 TF-IDF 算法抽取 TF-IDF 关键词;其次,将 TF-IDF 关键词与 LDA 算法抽取的主题词做交集,得到金庸小说语料集的章节关键词。

4 章节情节描述矩阵

4.1 角色情节强度

角色情节强度(character plot intensity, CPI)描述的是主要角色与章节内容的关联程度。CPI 越大,表明主要角色与章节内容的关联度越高。$\mathrm{CPI}=\{\mathrm{cpi}_{1,1},\mathrm{cpi}_{i,j},\cdots,\mathrm{cpi}_{c,r}\}$,其中,$\mathrm{cpi}_{i,j}$表示章节 i 中主要角色 j 的情节强度。在观点段落和非观点段落中,分别计算主要角色出现的频次、章节关键词出现的频次、主要角色与章节关键词共现频次,得到角色情节强度。角色情节强度计算公式如下:

$$\mathrm{cpi}_{c,r}=\frac{\sum_{i=1}^{n}\ln\dfrac{O(r,w_i)}{O(w_i)*O(r)}}{\sum_{i=1}^{m}\ln\dfrac{\overline{O}(r,w_i)}{\overline{O}(w_i)*\overline{O}(r)}},\tag{1}$$

$$O(\eta)=\mathrm{Occurrence}(\eta\mid O),\tag{2}$$

$$\overline{O}(\eta)=\mathrm{Occurrence}(\eta\mid\overline{O}),\tag{3}$$

其中,r 为主要角色,w_i 为章节关键词,O 为观点段落,$\overline{O}$ 为非观点段落,$\ln(\cdot)$为以 e 为底的指数函数,$\mathrm{Occurrence}(\cdot)$表示频次计算函数,$O(\eta)$为 η 在 O 中出现的频次,$\overline{O}(\eta)$为 η 在 $\overline{O}$ 中出现的频次。

4.2 角色观点强度

角色观点强度(character opinion intensity, COI)描述的是主要角色观点的丰富程度。COI 越大,表明主要角色在小说文本中表达的观点越丰富。通过统计主要角色在观点段落、非观点段落共现频次之比来计算 COI。$\mathrm{COI}=\{\mathrm{coi}_{1,1},\mathrm{coi}_{i,j},\cdots,\mathrm{coi}_{c,r}\}$,其中 $\mathrm{coi}_{i,j}$表示章节 i 中主要角色 j 的观点强度,可表示为

$$\mathrm{coi}_{c,r}=\frac{\mathrm{Occurrence}(r,O)}{\mathrm{Occurrence}(r,\overline{O})}。\tag{4}$$

4.3　角色主题强度

角色主题强度(character theme intensity, CTI)描述的是主要角色与章节主题的契合度。通过统计主要角色与章节主题词的共现频次之和来计算 CTI。$\mathrm{CTI}=\{\mathrm{cti}_{1,1},\mathrm{cti}_{i,j},\cdots,\mathrm{cti}_{c,r}\}$,其中 $\mathrm{cti}_{i,j}$表示章节 i 中主要角色 j 的主题强度,可表示为

$$\mathrm{cti}_{c,r}=\sum_{i=1}^{k}\sum_{j=1}^{n}\mathrm{Occurrence}(r,w_j)。\tag{5}$$

4.4　章节情节矩阵

如式(6)—(7)所示,主要角色的 CPI、COI、CTI、CR 构成章节情节矩阵 $\boldsymbol{M}$。该矩阵将章节情节信息表示为向量矩阵,有助于实现章节之间相似性比较,可表示为

$$\boldsymbol{M}=(\boldsymbol{m}_1,\cdots,\boldsymbol{m}_i,\cdots,\boldsymbol{m}_c),\tag{6}$$

$$\boldsymbol{m}_i=\begin{pmatrix}\mathrm{cpi}_{i,1} & \mathrm{coi}_{i,1} & \mathrm{cti}_{i,1} & \mathrm{cr}_{i,1}\\ \mathrm{cpi}_{i,2} & \mathrm{coi}_{i,2} & \mathrm{cti}_{i,2} & \mathrm{cr}_{i,2}\\ \vdots & \vdots & \vdots & \vdots\\ \mathrm{cpi}_{i,r} & \mathrm{coi}_{i,r} & \mathrm{cti}_{i,r} & \mathrm{cr}_{i,r}\end{pmatrix},\tag{7}$$

其中,$\boldsymbol{m}_c$ 表示章节 c 的情节矩阵,CR(character frequency ratio)为角色频次比,$\mathrm{cr}_{i,j}$表示主要角色在每一章节出现频次与所有主要角色出现频次之比。$\mathrm{cr}_{i,j}$越大,表明主要角色在当前章节的重要性越大,其关联的情节信息越多。

4.5　章节情节差异矩阵

章节情节差异矩阵 $\boldsymbol{D}$ 描述的是每一章节与其前、后章节的差异性。$\boldsymbol{D}$ 越大,说明该章节与其前、后章节变化程度越大,越有可能是高潮章节。本文利用动态时间弯曲(dynamic time warping, DTW)算法计算每一章节与其前、后章节的差异性,构建章节情节差异矩阵。由于第一章没有前章节,最后一章没有后章节,故采用均值化后的章节情节矩阵 $\boldsymbol{M}$ 作为第一章的前章节、最后一章的后章节。$\boldsymbol{D}$ 计算公式如式下:

$$\boldsymbol{D}=\begin{pmatrix}\mathrm{pre}_1 & \mathrm{next}_1\\ \mathrm{pre}_2 & \mathrm{next}_2\\ \vdots & \vdots\\ \mathrm{pre}_c & \mathrm{next}_c\end{pmatrix},\tag{8}$$

$$\mathrm{pre}_i=\mathrm{DTW}(\boldsymbol{m}_{i-1},\boldsymbol{m}_i),\tag{9}$$

$$\mathrm{next}_i=\mathrm{DTW}(\boldsymbol{m}_i,\boldsymbol{m}_{i+1}),\tag{10}$$

其中,pre_i 和 next_i 分别表示章节 i 的与其前、后章节的差异性。

4.6　章节情节描述矩阵

对章节情节差异矩阵 $\boldsymbol{D}$ 归一化处理后,将其中每一章节与其前、后章节的差异性 $\boldsymbol{d}_i$ 延展到与 $\boldsymbol{m}_i$ 相同的维度,拼接 $\boldsymbol{d}_i$ 与 $\boldsymbol{m}_i$,得到章节情节描述矩阵 $\boldsymbol{T}$。$\boldsymbol{T}$ 计算公式如下:

$$\boldsymbol{t}_i=\mathrm{concat}(\boldsymbol{m}_i,\boldsymbol{d}_i),\tag{11}$$

$$\boldsymbol{T}=(\boldsymbol{t}_1,\boldsymbol{t}_2,\cdots,\boldsymbol{t}_i,\cdots,\boldsymbol{t}_c),\tag{12}$$

其中,concat(·)表示拼接操作,$\boldsymbol{t}_i$ 表示章节 i 的情节描述矩阵。

5　高潮章节识别

图 2 给出基于情节描述的中文长篇小说高潮章节识别方法(plot description-based model for climax chapter recognition, CCRPCM),该模型由输入层、融合层、输出层组成。输入层由章节情节描述矩阵组成;融合层引入 BiGRU 模型和多头注意力机制对章节情节描述矩阵进行特征提取与融合,得到小说文本的深层语义特征;输出层将上述语义特征送入全连接神经网络,利用 softmax(·)函数进行归一化处理,得到输出概率,选择概率最大的值作为高潮章节识别结果。CCRPCM 计算过程如下:

$$\boldsymbol{Y}=\mathrm{BiGRU}(\boldsymbol{T}), \tag{13}$$

$$\boldsymbol{V}=\mathrm{MultiHead}(\boldsymbol{Y},\boldsymbol{Y},\boldsymbol{Y}), \tag{14}$$

$$\boldsymbol{P}=\mathrm{softmax}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{V}+\boldsymbol{b}), \tag{15}$$

$$y=\mathrm{argmax}(\boldsymbol{P}), \tag{16}$$

其中,$\boldsymbol{W}$ 和 $\boldsymbol{b}$ 表示权重矩阵和偏置,BiGRU(·)表示 BiGRU 模型函数,MultiHead(·)表示多头注意力机制函数,softmax(·)表示归一化函数,argmax(·)表示概率最大函数。

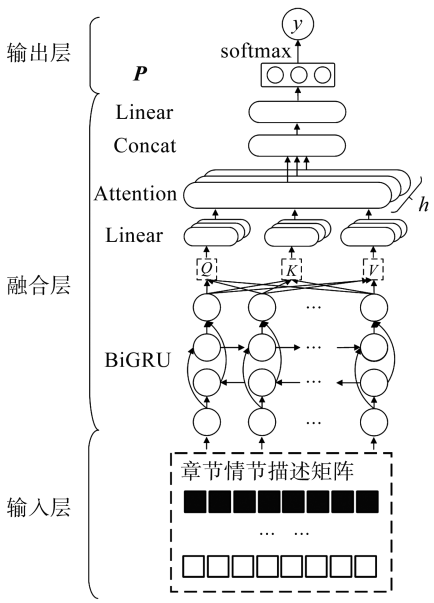

图 2 高潮章节识别模型

Fig.2 Climax chapter recognition model

5.1 BiGRU

鉴于小说文本具有显著的序列特征,故采用 BiGRU 模型[17]作为高潮章节识别的基础模型。BiGRU 模型通过拼接具有正向和反向的 GRU 模型的特征向量,实现了小说文本上下文语义特征的有效利用,计算公式如下:

$$\vec{h}_t=\mathrm{GRU}(\vec{h}_{t-1},\boldsymbol{x}_t), \tag{17}$$

$$\overleftarrow{h}_t=\mathrm{GRU}(\vec{h}_{t+1},\boldsymbol{x}_t), \tag{18}$$

$$\boldsymbol{y}_t=[\vec{h}_t,\overleftarrow{h}_t], \tag{19}$$

其中,$\boldsymbol{x}_t$ 为 t 时刻的输入向量,$\vec{h}$、$\overleftarrow{h}$ 分别表示正向和反向 GRU 模型得到的特征向量,$\boldsymbol{y}_t$ 为当前时刻 BiGRU 模型得到的特征向量。

5.2 多头注意力机制

多头注意力(multi-head attention, MultiHead)机制[18]将多个注意力机制进行横向拼接,用以表征不同位置、不同方面的语义信息,有助于进一步提高模型的学习能力,可表示为

$$\mathrm{head}_i=\mathrm{Attention}(\boldsymbol{Q}\boldsymbol{W}_i^{\boldsymbol{Q}},\boldsymbol{K}\boldsymbol{W}_i^{\boldsymbol{K}},\boldsymbol{V}\boldsymbol{W}_i^{\boldsymbol{V}}), \tag{20}$$

$$\mathrm{MultiHead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{Concat}(\mathrm{head}_1,\mathrm{head}_2,\cdots,\mathrm{head}_h)\boldsymbol{W}^o, \tag{21}$$

其中,head 表示注意力的头,h 表示头的个数,$\boldsymbol{Q}$、$\boldsymbol{K}$ 和 $\boldsymbol{V}$ 为 Query 向量、Key 向量和 Value 向量,$\boldsymbol{W}_i^{\boldsymbol{Q}}$、$\boldsymbol{W}_i^{\boldsymbol{K}}$、$\boldsymbol{W}_i^{\boldsymbol{V}}$ 为第 i 个 head 的 Query、Key 和 Value 的权重矩阵,$\boldsymbol{W}^o$ 为权重矩阵,$\boldsymbol{Q}\boldsymbol{W}\boldsymbol{Q}_i$、$\boldsymbol{K}\boldsymbol{W}\boldsymbol{K}_i$、$\boldsymbol{V}\boldsymbol{W}\boldsymbol{V}_i$ 和Concat(·)$\boldsymbol{W}^o$表示 Linear 层运算。

6 实验设计与分析

6.1 实验语料集

在网站 https://github.com/weiyinfu/JinYong 上下载《射雕英雄传》《倚天屠龙记》和《笑傲江湖》小说文

本,组成金庸小说语料集并将其作为实验语料集对这些进行数据清洗、全角字符转为半角字符、标点符号归一化、繁简转化、去重等预处理后得到各部小说 40 章,共计 160 章。由于目前没有标注好的高潮章节语料集,因此,在分析小说社区评论数据的基础上,结合领域专家标注结果,得到标注好的高潮章节语料集。

表 1 统计出金庸小说语料集部分主要角色及其出现频次;表 2 给出"金庸小说语料集的章节总数、高潮章节数以及高潮章节示例;表 3 给出金庸小说语料集的段落总数、观点段落数、非观点段落数。

表 1　部分主要角色及其频次

Table 1　Part of the main characters and their frequencies

《射雕英雄传》		《神雕侠侣》		《倚天屠龙记》		《笑傲江湖》	
主要角色	频次/次	主要角色	频次/次	主要角色	频次/次	主要角色	频次/次
郭靖	5 084	杨过	6 233	张无忌	4 784	令狐冲	5 947
黄蓉	3 722	小龙女	2 371	周芷若	1 737	岳不群	1 192
欧阳锋	904	郭靖	1 476	赵敏	1 311	林平之	932
周伯通	688	黄蓉	1 466	谢逊	1 919	岳灵珊	921
欧阳克	624	李莫愁	1 051	张翠山	1 156	仪琳	736

表 2　金庸小说语料集的章节数、高潮章节数、高潮章节示例

Table 2　The number of chapters, climactic chapters and the examples of climactic chapters in Jin Yong's novel corpus

小说名称	章节数/节	高潮章节数/节	高潮章节示例
《射雕英雄传》	40	8	大闹禁宫、岛上巨变
《神雕侠侣》	40	7	生辰大礼、生死茫茫
《倚天屠龙记》	40	10	屠狮有会孰为殃、百尺高塔任回翔

表 3　金庸小说语料集的段落总数、观点段落数、非观点段落数

Table 3　The number of paragraphs, viewpoint paragraphs and non-viewpoint paragraphs in Jin Yong's novel corpus

单位:个

小说名称	段落总数	观点段落数	非观点段落数
《射雕英雄传》	6 796	4 271	2 525
《神雕侠侣》	6 908	4 421	2 487
《倚天屠龙记》	7 859	3 495	4 364

6.2　参数设置

利用网格搜索法来确定高潮识别模型的最优参数。max_epoch 最大为 512,并设置连续 10 轮训练,模型性能没有提升就停止训练;batch_size 在网格[4,8,16]中搜索选取;lr 在网格[0.000 1, 0.000 2, 0.000 4]中搜索选取;GRU_dropout 在网格[0.1, 0.2, 0.4]中搜索选取;num_heads 在网格[4, 8, 16]中搜索选取。表 4 给出模型的主要参数含义及取值。

表 4　参数设置

Table 4　Parameter setting

参数	含义	取值
max_epoch	最大迭代次数	128
batch_size	批量大小	4
lr	初始学习率	0.000 1
GRU_dropout	BiGRU 模型丢失率	0.1
linear_dropout	线性层丢失率	0.1
num_heads	注意力机制头数	8

6.3　评价指标

采用召回率(Recall, R)、准确率(Precision, P)以及调和平均值($F1$-score, $F1$)来衡量小说高潮章节识别性能,其计算如下:

$$R=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}\times 100\%, \tag{22}$$

$$P=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}\times 100\%, \tag{23}$$

$$F1=\frac{2\times \mathrm{P}\times \mathrm{R}}{\mathrm{P}+\mathrm{R}}\times 100\%, \tag{24}$$

其中,真正例(true positive, TP)表示被正确分类的正例样本,假正例(false positive, FP)表示被错误分类的正例样本,假负例(false negative, FN)表示被错误分类的负例样本。而 P 表示模型预测正确的正例样本占预测为正例的样本的比例,R 表示模型预测正确的正例样本中占实际为正例的样本的比例。

6.4 实验结果与分析

为了验证所提模型 CCRPCM 的有效性,设计对比实验和消融实验。对比实验用到的模型有朴素贝叶斯[10]、支持向量机[9]、双向长短时记忆网络(bidirectional long short-term memory, BiLSTM)[19]、预训练模型 Roberta-large[20]。消融实验用来验证 CCRPCM 模型主要组成部分的有效性。主要消融模型有:不含情节描述矩阵的 CCRPCM 模型(CCRPCM without plot description matrix, CCRPCM-PCM)、不含章节情节差异矩阵的 CCRPCM 模型(CCRPCM without chapter plot difference matrix, CCRPCM-PDM)、不含 BiGRU 的 CCRPCM 模型(CCRPCM without BiGRU, CCRPCM-BiGRU)、不含多头注意力机制的 CCRPCM 模型(CCRPCM without multi-attention, CCRPCM-MultiAtt)。实验结果如表 5 所示。

表 5 实验结果

Table 5 The experimental results 单位:%

模型	《射雕英雄传》			《神雕侠侣》			《倚天屠龙记》			《笑傲江湖》		
	R	*P*	*F*1	*R*	*P*	*F*1	*R*	*P*	*F*1	*R*	*P*	*F*1
NB	53.96	52.80	53.37	55.68	55.49	55.58	57.07	60.56	58.76	54.92	59.07	56.92
SVM	57.97	60.84	59.37	56.45	54.98	55.71	59.87	61.62	60.73	59.61	56.95	58.25
Roberta-large	67.93	70.38	69.13	63.01	66.72	64.82	65.78	64.61	65.18	63.53	67.90	65.64
BiLSTM	66.25	70.87	68.48	67.89	69.87	68.87	69.92	71.17	70.53	68.73	69.48	69.10
CCRPCM-PCM	59.07	62.73	60.85	64.82	66.26	65.53	67.99	70.08	69.02	65.00	64.11	64.56
CCRPCM-PDM	53.76	53.64	53.70	53.68	53.05	53.36	59.56	60.93	60.24	58.33	57.84	58.08
CCRPCM-BiGRU	51.43	51.08	51.25	48.44	49.04	48.24	48.57	49.10	48.83	47.55	48.67	48.10
CCRPCM-MultiAtt	73.20	72.02	72.61	76.88	75.00	75.93	69.51	71.47	70.48	72.10	75.27	73.65
CCRPCM	**82.43**	**78.37**	**80.35**	**80.11**	**82.81**	**81.43**	**73.36**	**74.40**	**73.88**	**78.16**	**79.70**	**78.92**

由表 5 可以看出:在《射雕英雄传》语料集上,CCRPCM 性能最优,Roberta-large 次之,NB 最差。CCRPCM 的 *F*1 值分别比 NB、SVM、Roberta-large、BiLSTM 大 26.98%、20.98%、11.22%、11.87%。在《神雕侠侣》语料集上,CCRPCM 性能最优,BiLSTM 次之,NB 最差。CCRPCM 的 *F*1 值分别比 NB、SVM、Roberta-large、BiLSTM 大 25.85%、25.72%、16.61%、12.56%。在《倚天屠龙记》语料集上,CCRPCM 性能最优,BiLSTM 次之,NB 最差。CCRPCM 的 *F*1 值分别比 NB、SVM、Roberta-large、BiLSTM 大 15.12%、13.15%、8.70%、3.35%。在《笑傲江湖》语料集上,CCRPCM 性能最优,BiLSTM 次之,NB 最差。CCRPCM 的 *F*1 值分别比 NB、SVM、Roberta-large、BiLSTM 大 22.00%、20.67%、13.28%、9.82%。由此可见,与 NB、SVM、Roberta-large、BiLSTM 等模型相比,CCRPCM 具有更优的小说高潮识别效果,BiLSTM 与 Roberta-large 较为接近,NB 和 SVM 均不高。其主要原因是:

(1)NB 假设特征之间相互独立,但小说章节特征之间存在较大的相关性,导致其分类效果欠佳。而且,该模型需要事先给定先验概率,且先验概率一般取决于假设,这进一步增加了错分的风险。

(2)SVM 适用于处理两类规模均衡的分类问题,但由表 3 可以看出,实验语料集中的观点段落与非观点段落的规模明显失衡,故其分类性能收到较大影响。但较之 NB,其性能更优,主要原因有二:一是采用结构风险最小化原则进行模型选择,这种原则能够有效地避免“过拟合”现象的发生,使得该模型具有较强的泛化能力;二是该模型对异常点的鲁棒性较好,可以有效地避免异常点对分类结果的影响。

(3)BiLSTM 是一种典型的深度学习模型,其具有强大的特征学习能力,因此,该模型较之传统机器学习模型,如 NB 和 SVM,具有更优的识别性能。该模型通过 2 层 LSTM 的堆叠,使得模型摆脱了只能依据之前时刻的时序信息来预测下一时刻输出的限制,能更好地结合上下文信息进行特征提取。因此,该模型适用于处理序列化的小说文本,故而在本研究中取得了较好的识别效果。但该模型较之 CCRPCM 性能稍差,其主要原因是该模型致力于提取小说文本的语义特征,没有针对小说高潮章节识别的具体任务提取有效特征。

(4)Roberta-large 是基于 Transformer 结构的预训练语言模型,它一方面考虑小说文本的上下文信息,另一方面引入自注意力机制捕捉长距离依赖关系,因而具有更优的小说文本理解能力。因此,该模型在《射雕英雄传》语料集上较之 BiLSTM 具有更优的识别性能。该模型在《神雕侠侣》、《倚天屠龙记》和《笑傲江湖》语料集上的识别性能不如 BiLSTM。主要原因是该模型规模大、参数多,模型结构不够灵活,网络结构不易

改变,影响了其在不同小说文本上的表现。

消融实验的目的是通过移除某一组件之后的性能,来判断该组件对整个模型的作用。由表 5 可以看出:在《射雕英雄传》语料集上,CCRPCM 的 $F1$ 值分别比 CCRPCM-PCM、CCRPCM-PDM、CCRPCM-BiGRU、CCRPCM-MultiAtt 高 19.50%、26.65%、29.10%、7.74%。在《神雕侠侣》语料集上,CCRPCM 的 $F1$ 值分别比 CCRPCM-PCM、CCRPCM-PDM、CCRPCM-BiGRU、CCRPCM-MultiAtt 高 15.90%、28.07%、33.19%、5.50%。在《倚天屠龙记》语料集上,CCRPCM 的 $F1$ 值分别比 CCRPCM-PCM、CCRPCM-PDM、CCRPCM-BiGRU、CCRPCM-MultiAtt 高 4.86%、13.64%、25.05%、3.40%。在《笑傲江湖》语料集上,CCRPCM 的 $F1$ 值分别比 CCRPCM-PCM、CCRPCM-PDM、CCRPCM-BiGRU、CCRPCM-MultiAtt 高 14.36%、20.84%、30.82%、5.27%。由此可见,CCRPCM-MultiAtt 的识别性能最优,之后是 CCRPCM-PCM 和 CCRPCM-PDM,CCRPCM-BiGRU 最差。出现上述实验结果的原因是:

(1)BiGRU 集成了不同方向的 GRU 模型,该模型参数量较少,训练速度较快,较之 BiLSTM 具有更高的处理效率。该模型擅长捕捉小说文本的上下文语义信息,能够有效地抑制“梯度消失”或“梯度爆炸”等问题,因而,在缺省情节描述矩阵的情况下,CCRPCM-PCM 依然具有较高的识别性能。

(2)由 BiGRU 模型的工作机理不难看出,该模型往往将长距离信息或特征弱化,这不利于小说高潮章节识别。多头注意力机制对不同特征差异化对待,它自动对小说高潮章节识别具有重要作用的特征给予更多关注,有助于提高模型的识别性能。因此,在不含多头注意力机制的情况下,CCRPCM-MultiAtt 的 $F1$ 值低于 CCRPCM。

(3)小说高潮章节往往与其前、后章节变化程度较大,故而引入情节差异矩阵来描述每一章节与其前、后章节的差异性。由于情节描述矩阵是由章节情节矩阵和情节差异矩阵组成,在缺省情节差异矩阵的情况下,情节描述矩阵仅由章节情节矩阵组成。尽管章节情节矩阵对小说的角色和情节进行了刻画,但章节之间的差异性没有表征,进而影响到 CCRPCM-PDM 的识别性能。

(4)情节描述矩阵对小说章节及章节之间的关系进行了表征,该矩阵是在关键要素抽取的基础上形成的显式特征表达。BiGRU 作为深度学习模型的重要代表之一,其具有优良的特征学习能力。该模型能够对情节描述矩阵中的上下文语义进行进一步分析,有助于发现一系列隐式特征,这些特征对于小说高潮章节识别具有重要作用。因此,在缺省 BiGRU 的情况下,CCRPCM-BiGRU 的识别性能较差。

利用 CCRPCM 识别的高潮章节如表 6 所示,其中《射雕英雄传》的第 18 回为“三道试题”,第 23 回为“大闹禁宫”,第 24 回为“密室疗伤”,第 27 回为“轩辕台前”,第 34 回为“岛上巨变”,第 35 回为“铁枪庙中”,第 40 回为“华山论剑”;《神雕侠侣》的第 13 回为“武林盟主”,第 27 回为“斗智斗力”,第 33 回为“风陵夜话”,第 36 回为“生辰大礼”,第 38 回为“生死茫茫”,第 39 回为“大战襄阳”;《倚天屠龙记》的第 20 回为“与子共穴相扶将”,第 21 回为“排难解纷当六强”,第 24 回为“太极初传柔克刚”,第 27 回为“百尺高塔任回翔”,第 29 回为“四女同舟何所望”,第 35 回为“屠狮有会孰为殃”,第 36 回为“夭矫三松郁青苍”,第 38 回为“君子可欺之以方”;《笑傲江湖》的第 3 回为“救难”,第 14 回为“论杯”,第 20 回为“探狱”,第 25 回为“闻讯”,第 28 回为“积雪”,第 36 回为“伤逝”,第 38 回为“夭聚歼”,第 40 回为“曲谐”。

读者在阅读小说中,通常会沉浸在跌宕起伏的故事情节和主人公细腻而生动的形象塑造中。以《射雕英雄传》为例,结合豆瓣书目评分网、金融江湖网、金庸吧等论坛评论,读者普遍认为第 18 回“三道试题”、第 23 回“大闹禁宫”、第 24 回“密室疗伤”、第 27 回“轩辕台前”、第 34 回“岛上巨变”、第 35 回“铁枪庙中”以及第 40 回“华山论剑”这 7 个章节,不仅准确地捕捉并展现了小说的核心情节,还展现了主人公郭靖的成长历程。这些高潮章节情节紧凑、张力十足,成功推动了整个故事的发展。“三道试题”突显了他的智慧和毅力,“大闹禁宫”展现了他的勇气和正义感,“华山论剑”则是全书的巅峰。郭靖在这些关键情节中从青涩少年蜕变为成熟英雄,读者清晰地见证了他的成长与心智发展,同时也使得郭靖的形象更加立体丰满。此外读者普遍认为上述高潮章节不仅富有戏剧性和观赏性,还具有深刻的情感共鸣。比如,“密室疗伤”中的郭靖与黄蓉之间的感情描写让人动容,而“大闹禁宫”和“轩辕台前”等章节的紧张刺激情节更是让读者欲罢不能。总体而言,CCRPCM 识别的高潮章节与读者的反馈一致,因此其识别结果具有较高的可接受程度。

表 6 已识别出的高潮章节

Table 6 The identified climactic chapters

小说名称	章节 1	章节 2	章节 3	章节 4	章节 5	章节 6	章节 7	章节 8
《射雕英雄传》	第 18 回	第 23 回	第 24 回	第 27 回	第 34 回	第 35 回	第 40 回	—
《神雕侠侣》	第 13 回	第 27 回	第 33 回	第 36 回	第 38 回	第 39 回	—	—
《倚天屠龙记》	第 20 回	第 21 回	第 24 回	第 27 回	第 29 回	第 35 回	第 36 回	第 38 回
《笑傲江湖》	第 3 回	第 14 回	第 20 回	第 25 回	第 28 回	第 36 回	第 38 回	第 40 回

由表 6 可以看出,《射雕英雄传》、《神雕侠侣》、《倚天屠龙记》、《笑傲江湖》已识别高潮章节具有矛盾突出、情节跌宕的特点。由表 4 可知,CCRPCM 在《倚天屠龙记》语料集上的识别性能(73.88%)明显低于《射雕英雄传》(80.35%)、《神雕侠侣》(81.43%)和《笑傲江湖》(78.92%)。结合表 5、6 的实验结果可以看出,出现上述显著性能差异与小说文本行文特点紧密联系。《倚天屠龙记》未能识别的高潮章节——第 32 回“冤蒙不白愁欲狂”和第 39 回“秘笈兵书此中藏”,并非基本矛盾、冲突发展到紧张、尖锐的部分,情节较为平淡,角色情节强度、角色观点强度、角色主题强度等指标较低,进而导致高潮章节的识别效果较差。

各部小说已识别高潮章节的观点段落与平淡段落数分别如表 7—10 所示。

表 7 《射雕英雄传》已识别高潮章节的观点段落数与非观点段落数

Table 7 The number ofviewpoint and non-viewpoint passages in the identified climactic chapters of *The Legend of the Condor Heroes*

《射雕英雄传》	观点段落数	非观点段落数	《射雕英雄传》	观点段落数	非观点段落数	《射雕英雄传》
第 18 回	151	51	第 34 回	117	62	第 18 回
第 23 回	96	53	第 35 回	100	48	第 23 回

表 8 《神雕侠侣》已识别高潮章节的观点段落数与非观点段落数

Table 8 The number ofviewpoint and non-viewpoint passages in the identified climactic chapters of *Divine Eagle Heroes*

《神雕侠侣》	观点段落数	非观点段落数	《神雕侠侣》	观点段落数	非观点段落数
第 13 回	106	90	第 36 回	93	51
第 27 回	110	71	第 38 回	109	47

表 9 《倚天屠龙记》已识别高潮章节的观点段落数与非观点段落数

Table 9 The number ofviewpoint and non-viewpoint passages in the identified climactic chapters of *Heaven Sword and Dragon Sabre*

《倚天屠龙记》	观点段落数	非观点段落数	《倚天屠龙记》	观点段落数	非观点段落数	《倚天屠龙记》	观点段落数
第 20 回	121	91	第 29 回	28	80	第 20 回	121
第 21 回	75	122	第 35 回	104	76	第 21 回	75

表 10 《笑傲江湖》已识别高潮章节的观点段落数与非观点段落数

Table 10 The number ofviewpoint and non-viewpoint passages in the identified climactic chapters of *The Swordsman*

《笑傲江湖》	观点段落数	非观点段落数	《笑傲江湖》	观点段落数	非观点段落数	观点段落数
第 3 回	157	71	第 28 回	101	44	157
第 14 回	124	79	第 36 回	151	71	124

各部小说已识别高潮章节的主要角色与频次分别如表 11—14 所示。

表 11 《射雕英雄传》高潮章节出现的角色与频次

Table 11 The characters and their appearances in the identified climactic chapters of *The Legend of the Condor Heroes*

序号	角色	第 18 回	第 23 回	第 24 回	第 27 回	第 34 回	第 35 回	第 40 回
1	郭靖	140	145	64	92	201	36	175
2	黄蓉	96	190	79	73	98	136	115
3	洪七公	74	32	1	26	30	15	89
4	黄药师	134	6	26	8	136	42	56
5	周伯通	15	61	1	1	2	24	3
6	欧阳克	88	0	6	2	0	3	4
7	梅超风	7	1	2	1	3	1	0

表 11(续)

序号	角色	第 18 回	第 23 回	第 24 回	第 27 回	第 34 回	第 35 回	第 40 回
8	杨康	1	23	13	89	4	26	4
9	柯镇恶	0	0	0	0	34	96	1
10	裘千仞	0	0	0	44	0	17	0
11	朱聪	0	0	0	2	12	8	0
12	蓉儿	16	18	3	4	20	1	12
13	完颜洪烈	0	35	32	0	0	23	1
14	铁木真	0	0	0	0	0	0	0
15	穆念慈	1	0	0	0	0	2	21
16	彭连虎	2	9	28	1	2	15	1
17	陆冠英	0	0	93	0	0	0	0
18	瑛姑	0	0	0	0	0	0	0
19	杨铁心	0	0	0	0	0	0	0
20	拖雷	1	0	0	0	0	0	28
21	梁子翁	2	9	21	0	0	7	0
22	成吉思汗	0	0	1	0	1	0	34
23	华筝	0	0	0	2	2	0	7
24	王处一	0	0	1	2	13	8	0
25	包惜弱	0	0	6	0	0	0	0
26	傻姑	0	69	27	1	0	48	0
27	韩小莹	0	0	0	0	11	4	0
28	老顽童	0	14	2	0	1	10	0
29	韩宝驹	0	0	0	1	12	6	0
30	侯通海	0	6	55	0	0	0	0
31	沙通天	2	10	15	0	2	12	1
32	黄老邪	4	0	0	0	17	5	6
33	鲁有脚	0	0	0	45	0	0	1
34	程瑶迦	0	0	78	2	0	0	0
35	尹志平	0	0	71	1	13	2	0
36	小王爷	0	0	0	0	0	17	0
37	段天德	0	0	0	0	0	0	0
38	全金发	0	0	0	0	8	3	0

表 12　《神雕侠侣》高潮章节出现的角色与频次

Table 12　The characters and their appearances in the identified climactic chapters of *Divine Eagle Heroes*

序号	角色	第 13 回	第 27 回	第 33 回	第 36 回	第 38 回	第 39 回
1	杨过	261	124	3	41	57	174
2	小龙女	95	63	1	9	30	59
3	郭靖	44	41	7	48	14	99
4	黄蓉	49	114	5	59	102	54
5	李莫愁	1	78	0	0	3	0
6	郭芙	14	22	16	34	1	16
7	郭襄	0	31	69	43	128	91
8	周伯通	0	0	0	4	58	30
9	陆无双	0	1	0	0	19	6
10	赵志敬	1	13	0	0	0	0
11	姑姑	14	11	0	0	0	0
12	霍都	109	13	0	9	4	1
13	裘千尺	0	0	0	0	1	0

表 12(续)

序号	角色	第 13 回	第 27 回	第 33 回	第 36 回	第 38 回	第 39 回
14	金轮法王	37	11	0	0	20	17
15	过儿	14	10	0	5	1	7
16	公孙止	0	0	0	0	1	0
17	尹志平	1	7	0	0	0	0
18	耶律齐	0	0	1	42	0	28
19	黄药师	1	4	0	3	16	47
20	潇湘子	0	11	0	0	0	0
21	武三通	0	0	0	0	0	5
22	武修文	0	0	0	12	0	4
23	尼摩星	0	17	0	3	0	0
24	朱子柳	31	0	0	1	0	9
25	尹克西	0	10	0	0	0	0
26	孙婆婆	4	1	0	0	0	0
27	洪七公	8	0	0	0	0	0
28	达尔巴	79	10	0	0	1	0
29	忽必烈	0	1	1	0	3	10
30	樊一翁	0	0	0	9	0	0
31	完颜萍	0	0	0	4	0	1
32	柯镇恶	0	0	0	0	2	0
33	马光佐	0	8	0	0	0	0
34	洪凌波	0	0	0	0	0	0
35	老顽童	0	0	0	1	19	8
36	一灯大师	2	0	0	1	12	10
37	武敦儒	0	0	0	3	0	3
38	郝大通	5	2	0	0	0	0

表 13 《倚天屠龙记》高潮章节出现的角色与频次

Table 13 The characters and their appearances in the identified climactic chapters of *Heaven Sword and Dragon Sabre*

序号	角色	第 20 回	第 21 回	第 24 回	第 27 回	第 29 回	第 35 回	第 36 回	第 38 回
1	张无忌	154	227	140	93	121	135	183	151
2	赵敏	0	0	48	64	41	157	49	25
3	谢逊	7	7	0	0	69	23	39	48
4	张翠山	5	0	3	0	0	0	0	1
5	周芷若	2	1	1	29	15	3	0	155
6	殷素素	3	1	2	0	0	0	0	0
7	杨逍	13	5	13	5	5	1	52	16
8	张三丰	5	0	105	2	3	0	2	4
9	灭绝师太	0	0	0	50	1	0	0	6
10	小昭	68	6	4	0	17	7	0	0
11	金花婆婆	0	1	0	0	32	0	0	0
12	殷梨亭	3	0	2	2	3	0	0	36
13	周颠	1	1	12	0	0	0	18	25
14	韦一笑	5	2	38	12	0	1	11	6
15	郭襄	0	0	1	0	0	0	0	1
16	胡青牛	0	13	1	0	0	0	0	0
17	范遥	0	0	0	61	1	1	26	24
18	宋青书	1	0	0	3	0	0	0	65
19	俞岱岩	0	0	44	1	0	0	0	2
20	何太冲	0	28	0	4	0	0	13	0
21	成昆	13	7	0	1	1	7	18	0
22	纪晓芙	0	0	0	0	1	0	0	2

表 13(续)

序号	角色	第 20 回	第 21 回	第 24 回	第 27 回	第 29 回	第 35 回	第 36 回	第 38 回
23	宋远桥	20	0	4	6	0	0	0	1
24	鹿杖客	0	0	0	39	0	1	1	0
25	殷天正	69	4	15	0	0	1	34	1
26	陈友谅	0	0	0	0	1	5	1	1
27	杨不悔	2	3	1	0	0	0	0	1
28	空智	17	4	7	4	0	7	26	6
29	朱长龄	0	0	0	0	0	0	0	0
30	常遇春	0	0	1	0	0	0	0	0
31	空闻	3	1	6	10	0	11	21	1
32	张松溪	11	0	0	5	0	0	0	0
33	张教主	0	0	11	3	1	3	27	14
34	殷离	0	0	0	0	28	1	0	0
35	说不得	4	1	16	0	0	0	2	2
36	鹤笔翁	0	0	0	51	0	0	1	0

表 14　《笑傲江湖》高潮章节出现的角色与频次

Table 14　The characters and their appearances in the identified climactic chapters of *The Swordsman*

序号	角色	第 3 回	第 14 回	第 20 回	第 25 回	第 28 回	第 36 回	第 38 回	第 40 回
1	令狐冲	43	79	158	147	134	211	261	168
2	岳不群	0	55	2	0	38	148	21	0
3	林平之	14	7	1	0	24	77	28	2
4	岳灵珊	0	30	1	1	31	59	4	0
5	仪琳	74	0	0	11	0	0	4	8
6	田伯光	89	1	0	9	0	0	2	1
7	任我行	0	0	0	0	31	5	4	29
8	向问天	0	0	43	0	9	1	0	18
9	左冷禅	0	0	0	4	20	15	45	2
10	余沧海	40	0	0	0	0	5	0	0
11	刘正风	16	0	0	1	0	0	0	1
12	冲虚	0	0	0	0	3	0	0	90
13	林震南	14	0	0	0	0	0	0	0
14	劳德诺	73	3	0	1	0	30	0	7
15	仪和	0	0	0	24	0	0	0	5
16	陆大有	11	0	0	1	0	1	0	1
17	祖千秋	0	47	0	1	1	0	0	3
18	莫大	0	0	0	42	0	0	6	3
19	木高峰	0	0	0	0	0	5	0	0
20	岳先生	0	4	0	1	1	3	5	0
21	桃根仙	0	35	0	0	0	0	6	1
22	桃枝仙	0	15	0	0	0	0	3	2
23	莫大先生	0	0	0	33	0	0	6	1
24	桃花仙	0	16	0	0	0	0	4	2
25	任教主	0	0	0	0	6	1	3	36
26	上官	0	0	0	0	0	1	0	0
27	桃实仙	0	23	0	0	0	0	2	1
28	桃干仙	0	28	0	0	0	0	3	2
29	曲非烟	0	0	0	0	0	0	0	0
30	仪清	0	0	0	9	0	1	0	8

表 14(续)

序号	角色	第 3 回	第 14 回	第 20 回	第 25 回	第 28 回	第 36 回	第 38 回	第 40 回
31	风清扬	0	0	11	0	0	0	0	4
32	上官云	0	0	0	0	0	1	0	0
33	圣姑	0	0	0	1	0	1	13	0
34	平一指	0	49	5	0	0	0	0	0
35	杨莲亭	0	0	0	0	0	0	0	0
36	秃笔翁	0	0	38	0	0	0	0	0
37	蓝凤凰	0	0	0	0	0	0	18	0
38	游迅	0	0	0	0	0	0	37	0
39	郑萼	0	0	0	12	0	0	0	1
40	绿竹翁	0	2	0	0	1	0	0	3

一般而言,小说高潮章节的观点段落数较多、表达的情感较为强烈。由表 7—10 可以看出,《射雕英雄传》已识别高潮章节的观点段落数明显大于非观点段落数;《神雕侠侣》除第 33 回外,已识别高潮章节的观点段落数明显大于非观点段落数;《倚天屠龙记》仅有第 21 回和第 29 回外,已识别高潮章节的观点段落数明显大于非观点段落数;《笑傲江湖》已识别高潮章节的观点段落数明显大于非观点段落数。《倚天屠龙记》有 2 个章节出现观点段落数小于非观点段落数,明显多于《射雕英雄传》的 0 个、《神雕侠侣》的 1 个和《笑傲江湖》的 0 个,鉴于观点段落对于小说高潮识别的重要作用,不难理解表 4 所示的《倚天屠龙记》$F1$ 值低于其他 3 部小说语料集。

小说的高潮章节往往出现更多角色[15,21]。由表 11—14 可以看出,《射雕英雄传》已识别高潮章节第 18、23、24、27、34、35、40 回出现的角色数分别为 15、15、22、19、21、26、18 位,出现频次最高的角色及频次分别为郭靖(140)、黄蓉(190)、陆冠英(93)、郭靖(92)、郭靖(201)、黄蓉(136)、郭靖(175),其中括号内的数值为角色出现频次。《神雕侠侣》已识别高潮章节第 13、27、33、36、38、39 回出现的角色数分别为 19、23、8、19、20、21 位,出现频次最高的角色及频次分别为杨过(261)、杨过(124)、郭襄(69)、黄蓉(59)、郭襄(128)、杨过(174)。《倚天屠龙记》已识别高潮章节第 21、24、27、29、35、36、39 回出现的角色数分别为 17、22、20、16、16、18、24 位,出现频次最高的角色及频次分别为张无忌(227)、张无忌(140)、张无忌(93)、张无忌(121)、赵敏(157)、张无忌(183)、周芷若(155)。《笑傲江湖》已识别高潮章节第 3、14、20、25、28、36、38、40 回出现的角色数分别为 9、15、8、15、12、17、20 位、25 位,出现频次最高的角色及频次分别为田伯光(89)、令狐冲(79)、令狐冲(158)、令狐冲(147)、令狐冲(134)、令狐冲(211)、令狐冲(261)、令狐冲(168)。结合章节关键词集合可以看出,出现频次较高的主要角色是高潮章节的参与者和推动者,在高潮情节中扮演重要角色。

《射雕英雄传》第 17、18、19 回相关指标 cpi、coi、cti、cr、pre、next 如表 15—17 所示。

表 15 《射雕英雄传》第 17 回相关指标 cpi、coi、cti、cr、pre、next

Table 15 The indicators cpi, coi, cti, cr, pre, next in the 17th chapter of *The Legend of the Condor Heroes*

角色名称	cpi	coi	cti	cr	pre	next
郭靖	0.52	0.07	1167.00	0.33		
黄蓉	0.10	0.03	696.00	0.15		
洪七公	0.00	0.00	44.00	0.01	0.09	0.30
黄药师	0.00	0.00	108.00	0.01		
欧阳克	0.00	0.00	88.00	0.01		

表 16 《射雕英雄传》第 18 回相关指标 cpi、coi、cti、cr、pre、next

Table 16 The indicators cpi, coi, cti, cr, pre, next in the 18th chapter of *The Legend of the Condor Heroess*

角色名称	cpi	coi	cti	cr	pre	next
郭靖	2.90	0.26	1 712.00	0.35		
黄蓉	4.25	0.25	1 309.00	0.34		
洪七公	1.80	0.15	714.00	0.15	0.30	0.10
黄药师	6.48	0.36	956.00	0.21		
欧阳克	1.95	0.20	658.00	0.14		

表 17　《射雕英雄传》第 19 回相关指标 cpi、coi、cti、cr、pre、next

Table 17　The indicators cpi, coi, cti, cr, pre, next in the 19th chapter of *The Legend of the Condor Heroes*

角色名称	cpi	coi	cti	cr	pre	next
郭靖	2.32	0.16	1 469.00	0.23		
黄蓉	9.96	0.11	1 066.00	0.16		
洪七公	1.81	0.21	756.00	0.12	0.10	0.25
黄药师	3.14	0.08	1 516.00	0.23		
欧阳克	2.28	0.06	982.00	0.14		

当小说主要角色之间的矛盾冲突随着情节的发展达到最尖锐,主要角色出现次数越多、主要角色的人物形象塑造得最为丰满,章节主题在主要角色上表达得最为清晰时,小说情节会在这一章迅速发展,从而达到一个情节上的高潮。例如,结合表 11 和表 15—17,《射雕英雄传》第 18 回“三道试题”郭靖、黄蓉、黄药师、欧阳克和洪七公是故事中出现频次最多的角色,相比第 17 回“双手互博”,他们的角色情节强度、角色频次比较高,分别为 2.90、4.25、1.80、6.48、1.95 和 0.35、0.34、0.15、0.21、0.14,表明他们在推动情节发展中发挥着关键作用。此外,他们的角色观点强度、角色主题强度同样显著,分别为 0.26、0.25、0.15、0.36、0.20 和1 712.00、1 309.00、714.00、956.00、658.00,这表明章节所要传达的主题思想都通过他们的行为和对话得到了生动的展现。通过前面章节的铺垫,本章节中的冲突更为尖锐,人物形象的塑造更为丰满,与前一章节相比,差异值为 0.30,差异较为显著。因此,本章节达到了情节上的高潮。CCRPCM 模型识别的《射雕英雄传》其他高潮章节第 23 回为“大闹禁宫”,第 24 回为“密室疗伤”,第 27 回为“轩辕台前”,第 34 回为“岛上巨变”,第 35 回为“铁枪庙中”,第 40 回为“华山论剑”也呈现出类似的特征。而第 19 回“洪涛群鲨”虽然出现频次最多的角色具有较高的角色情节强度、角色频次比和角色主题强度,分别为 2.32、9.96、1.81、3.14、2.28, 0.23、0.6、0.12、0.23、0.14 和 1 469.00、1 066.00、756.00、1 516.00、982.00,但是其角色观点表达不够明显,黄药师、欧阳克角色观点强度仅为 0.08 和 0.06。此外,尽管第 19 回“洪涛群鲨”也是高潮章节,与后一章节差异值有 0.25,但是与前一章节的差异值不高,仅为 0.10,因此 CCRPCM 模型未能识别。总的来说,章节情节描述矩阵是对章节多维度特征的全面反映,揭示了每个角色如何通过自己的行为和观点,共同推动情节的发展,体现章节的主题,因此 CCRPCM 模型在表 5 中四本小说语料集上表现出最优的高潮章节识别效果。

7　总结

本文在精准刻画小说情节的基础上,提出中文长篇小说高潮章节识别方法。金庸“射雕三部曲”语料集上的比较实验表明本文所提方法较之已有方法具有一定优势。此外,本文还设计了消融实验,用以验证所提方法主要组成部分的有效性。尽管本文研究取得了一定进展,但由于标注语料规模相对有限,导致深度学习模型的性能没有充分发挥,高潮章节识别性能有待于进一步提高。

参考文献:

[1] 肖天久,刘颖. 基于聚类和分类的金庸与古龙小说风格分析[J]. 中文信息学报,2015,29(5):167-177.
XIAO Tianjiu, LIU Ying. A styistic analysis of Jin Yong's and Gu Long's fictions based on text clustering and classification [J]. Journal of Chinese Information Processing, 2015, 29(5):167-177.

[2] 姚睿琦,张辉,姚云洪. 社会网络分析方法在金庸小说人物关系中的应用研究[J]. 文献与数据学报,2021,3(3):68-80.
YAO Ruiqi, ZHANG Hui, YAO Yunhong. Research on application of social network analysis on character relationships in Jin Yong's novels[J]. Journal of Library and Data, 2021, 3(3):68-80.

[3] 张旋,梁循,李志宇,等. 金庸小说中主角复杂爱情模式的识别与分析[J]. 中文信息学报,2019,33(4):109-119.
ZHANG Xuan, LIANG Xun, LI Zhiyu, et al. Identification and analysis of love relationships of protagonists in Jin Yong's fictions[J]. Journal of Chinese Information Processing, 2019, 33(4):109-119.

[4] 邰沁清,夏恩赏,饶高琦,等. 数字人文视角下的金庸文本挖掘研究[J]. 数字人文,2020,4:115-136.
TAI Qinqing, XIA Enshang, RAO Gaoqi, et al. Research on Jin Yong with text mining from the perspective of digital humanities[J]. Digital Humanities, 2020, 4:115-136.

[5] LIU Ying, XIAO Tianjin. A stylistic analysis for Gu Long's Kung Fu novels[J]. Journal of Quantitative Linguistics, 2020, 27

(1):32-61.

[6] XIA Enshan, TAI Qingqing, LI Qi, et al. Digital humanities research of Jin Yong's works based on quantitative linguistics[J]. International Journal of Knowledge and Language, 2021, 12(1):1-10.

[7] ZHANG Le, WANG Shuai, LIU Bing. Deep learning for sentiment analysis: a survey[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2018, 8(4):e1253.

[8] KIM E, KLINGER R. An analysis of emotion communication channels in fan-fiction: towards emotional storytelling[C]// Proceedings of the Second Workshop on Storytelling. Florence:ACL, 2019:56-64.

[9] ZEHE A, BECKER M, HETTINGER L, et al. Prediction of happy endings in German novels based on sentiment information [C]//Proceedings of the 3rd Workshop on Interactions between Data Mining and Natural Language. Riva del Garda:[s.n.], 2016:9-16.

[10] MOHAMMAD S M, TURNEY P. NRC emotion lexicon[EB/OL]. (2011-07-10)[2024-01-30]. http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm.

[11] HORTON T, TAYLOR K, YU B, et al. "Quite right, dear and interesting": seeking the sentimental in nineteenth century American fiction[C]// Digital Humanities Conferences. Paris:[s.n.], 2006:81-82.

[12] YU Bei. An evaluation of text classification methods for literary study[J]. Literary and Linguistic Computing, 2008, 23(3): 327-343.

[13] 梁循. 基于深度学习的社会信息挖掘应用实例分析[M]. 北京:科学出版社,2020.
LIANG Xun. Application instance analysis of social information mining based on deep learning[M]. Beijing: Science Press, 2020.

[14] 宋琦. 武侠小说从"民国旧派"到"港台新派"叙事模式的变迁[D]. 济南:山东大学,2010.
SONG Qi. The narrative model changes of martial arts novels from "old school during the republican period" to "new breed of Hong Kong and Taiwan"[D]. Jinan:Shandong University, 2010.

[15] 曹正文.中国侠文化史[M].上海:上海书店出版社, 2014.
CAO Zhengwen. History of Chinese chivalrous culture[M]. Shanghai: Shanghai Bookstore Publishing House, 2014.

[16] HAN H, CHOI J D. The stem cell hypothesis: dilemma behind multi-task learning with transformer encoders[C]// Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Punta Cana: ACL, 2021:5555-5577.

[17] KUMAR A, VEPA J. Gated mechanism for attention based multi modal sentiment analysis[C]// Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. Washington, D.C.: IEEE, 2020:4477-448.

[18] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//31st Conference on Neural Information Processing Systems (NIPS 2017). Long Beach: ACM, 2017:5998-6008.

[19] 过临朋. 基于NLP的小说人物属性抽取系统[D]. 北京:北京邮电大学,2021.
GUO Linpeng. A NLP-based novel character attribute extraction system[D]. Beijing: Beijing University of Posts and Telecommunications, 2021.

[20] XU Liang, HU Hai, ZHANG Xuanwei, et al. CLUE: a Chinese language understanding evaluation benchmark[C]// Proceedings of the 28th International Conference on Computational Linguistics. Barcelona: ACL, 2020:4762-4772.

[21] BAL M. Narratology: introduction to the theory of narrative[M]. Toronto: University of Toronto Press, 2009.

(编辑:李艺)