基于 BERT-BiGRU-ATT 的社交媒体用户身份识别研究

张翼翔, 芦天亮, 李 默
(中国人民公安大学信息网络安全学院, 北京 100038)

摘 要 随着互联网井喷式发展,社交媒体发展迅猛,但是伴随网络匿名特性出现的失范现象时有发生,如何准确判定社交媒体用户从属问题亟待解决。目前社交媒体信息载体多以短文本为主,语法语义过于灵活,难以准确获得文本特征向量。传统短文本作者识别多采用人工建模的方式对文本特征加以提取,设计纷繁复杂。结合深度学习的方法,提出 BERT-BiGRU-ATT 短文本作者身份识别模型。该模型对中文短文本使用 BERT 中文预训练模型生成字符向量,利用双向门控循环单元(BiGRU)结合注意力机制高效捕获序列上下文特征,最终通过 A-softmax 分类器实现文本作者的识别。在制作的中文微博短文本数据集上的实验结果表明,BERT-BiGRU-ATT 模型与其他模型相比,在中文短文本作者识别的准确率上取得较好的成绩,其 F1 值达到 93.6% 的精度。

关键词 BERT 预训练模型; 双向门控循环单元; 作者识别; 注意力机制; 短文本
中图分类号 D035.39 **文献标志码** A

Research on Social Media User Identity Recognition Based on BERT-BiGRU-ATT

ZHANG Yixiang, LU Tianliang, LI Mo
(School of Information Cyber Security, People's Public Security University of China, Beijing 100038, China)

Abstract: With the blowout development of the Internet, social media has developed rapidly, but the anomie phenomenon that accompanies the anonymity of the network has occurred from time to time. How to accurately determine the affiliation of social media users needs to be solved urgently. At present, social media information carriers are mostly short texts, and their syntax and semantics are too flexible, and it is difficult to accurately obtain text feature vectors. Traditional short text author recognition uses manual modeling to extract text features, and the design is complicated. Combined with deep learning methods, a BERT-BIGRU-ATT short text author identification model is proposed. This model uses BERT Chinese pre-training model to generate character vectors for Chinese short texts, and uses Bi-Gated Recurrent Unit (BiGRU) combined with the attention mechanism to efficiently capture sequence context features, and finally realizes the recognition of the text author through the A-softmax classifier. The experimental results on the produced Chinese Weibo short text data set show that compared with other models, the BERT-BIGRU-ATT model has achieved better results in the accuracy of Chinese short text author recognition, with an F1 value of 93.6%.

Key words: BERT pre-training model; BiGRU; identity recognition; attention mechanism; short text

收稿日期 2020-10-29
基金项目 中国人民公安大学研究生科研项目(2020ssky008)。
作者简介 张翼翔(1996—),男,江苏南京人,在读硕士研究生。研究方向为网络空间安全。
通信作者 芦天亮(1985—),男,博士,副教授,E-mail: ltl135@126.com

0 引言

近年来，伴随着互联网的兴起与飞速发展，人们的生活维度随之拓宽，社会活动轨迹不仅限于自然社会，网络空间逐渐演变成为人类社会的“第二类生存空间”。层见叠出的信息互联技术，譬如电子邮件、即时通讯工具、论坛、社交媒体、博客等，带给人们更快、更加有效的方式进行信息交流交换。据 Global Web Index（GWI）发布的 2020 年第一季度《社交媒体趋势报告》显示，全球网民中有 63% 的用户社交媒体持续在线，高于 2019 年的 56%，这一趋势还将继续上升。由此可见，社交媒体已经成为人们日常生活中越来越重要的工具。

由于网络传播具有匿名性与高效性等特点，近些年互联网犯罪数量呈指数级别增长，网络空间失范现象时有发生。同时，社交媒体目前正在成为网民表达诉求、反映民意的重要节点，但也逐渐成为宣泄情绪、散播负面信息的场所，更有甚者成为各政治力量同台竞技的新舞台。

通过识别社交媒体用户，关联同一自然人的不同虚拟身份，有助于降低网络匿名性带来的风险，协助网络监管者的监管活动，保护公民合法权益。目前如微博，推特等主流社交媒体平台由于其及时性，碎片性特点，使得平台信息载体多为短文本。短文本的内容简短、主题多元化、语法表达具有随意性等特点导致已有长文本识别模式无法应用于短文本上，也同时导致短文本的特征提取更加困难。本文的方法将使用 BERT[1] 预训练模型，结合门控循环单元网络（GRU）并引入注意力机制，实现对短文本作者的分类。加以实验证明该方法在短文本作者识别问题上具有较高准确率。

1 相关工作

关于文本作者身份分析的研究最初起源于语言学研究领域关于文体即文学风格[2] 的归纳分析。由于社交网络的增长，越来越多关于作者识别的研究集中在网络文本的分析上。Mohtasseb 等人[3] 结合了心理学的工具语言探索与字词技术（LIWC）首次对博客作者进行识别。Pillay 等人[4] 针对网络论坛文本采用无监督与有监督学习相结合的方法训练出分类器实现了论坛发文作者的识。Cristani 等人[5] 基于二元聊天对话语料库，采用由会话提炼的特征值进行分析，从而识别出文本作者。Inches 等人[6] 首次使用统计数据研究会话文档完成在线即时聊天的作者归属问题研究。Hollingsworth[7] 提出以一种基于最相邻词频排名的作者识别方法，使用 DepWords 代替原文标记单词以发掘单词间依赖关系对于作者识别的帮助，并在小说作者识别上得到较好反馈。但上述研究主要针对英文语料，不适用于处理中文文本。

国内学者针对中文文本开展了大量的研究，起初关于文本作者识别研究的主要对象偏向于长文本，如长篇文章或书籍作者的判定。王少康等人[8] 以文章语句节奏控制角度为切入点，构建节奏特征矩阵，采用 KL 距离算法于点积法的结合衡量矩阵差异，提出最优区拟合的中轴线提取算法。李晓军等人[9] 将复杂网络理论引入利用文本特征的作者识别研究领域，选取新闻报道文章做数据集，构造复杂网络模型提取文本特征，利用文本风格相似度识别作者身份。Tang 等人[10] 选择押韵，体裁，叠词等特征采取监督机器学习，实现小说作者与多位诗歌作家的同一作者认定。

但是，上述方法也存在弊端，在面对篇幅小、表述方式灵活的短文本时，以文体风格为主的研究便显得捉襟见肘。由于社交媒体的蓬勃发展，其信息载体大多为短文本，于是针对短文本作者身份识别的研究应运而生。祈瑞华等人[11] 面向短文本博客，抽取字符、词汇、句法等特征建立多层面文体风格特征的模型并验证了该方法的准确性。Yang 等人[12] 提出一种对时间信息及单词顺序较为敏感的主题漂移模型（TDM）从写作风格和主题方向入手来完成作者识别任务。Zhang 等人[13] 将文本中语句的语法解析树编码得到分布式表示，即为每个词构造与之唯一对应的嵌入向量，将路径编码置于该单词对应的语法树中，并将获得的向量输入 CNN 模型中完成文本作者的认定并取得较好成效。徐晓霖等人[14] 采用深度学习的方法提出了 CABLSTM 模型，可高质量完成中文微博作者识别任务。冯勇等人[15] 提出了融合中文 FastText、融合词频-逆文本频率及隐含狄利克雷分布的短文本分类方法，在中文短文本分类上有较高精确率。

在以上针对短文本作者识别的研究中，文本特征提取起到至关重要的作用。在现有研究中，特征提取多以人工特征建模为主，需要复杂的设计处理。部分研究结合深度学习，使用词嵌入（word embedding）的方式取加权平均对短文本进行特征提取，但

该方法最大的弊端是无法处理多义词。结合近几年兴起的采用神经网络识别作者的思路，本文提出了采用 BERT 模型提取短文本特征生成词向量，利用带有注意力机制的双向 GRU 网络进行训练，最终通过 A-softmax 分类器进行分类的作者识别模型。

2 社交媒体作者身份识别模型

本文提出的身份识别模型结构分为 4 层：文本输入层、Bi-GRU 层、自注意力机制层以及 A-softmax 分类层。BERT-BiGRU-ATT 短文本作者识别模型结构如图 1 所示。

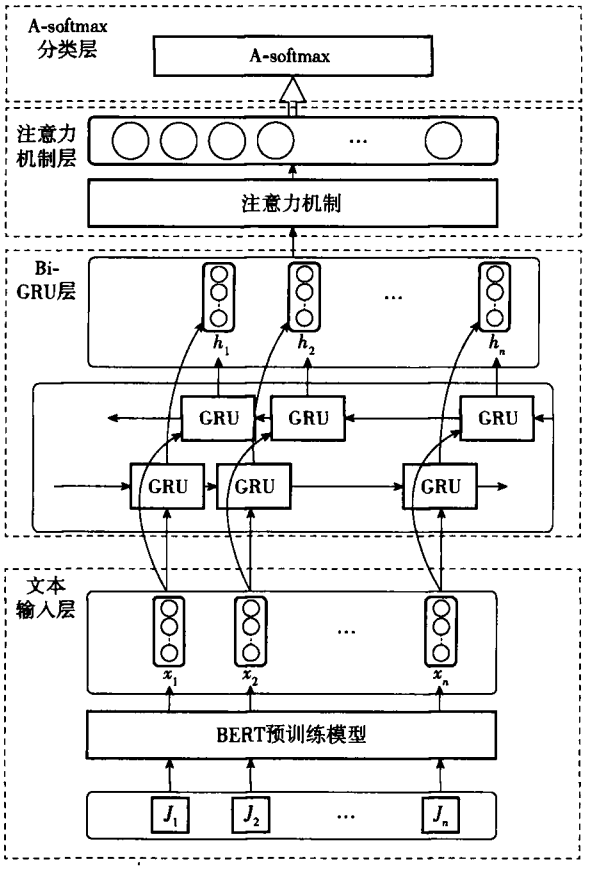

图 1 BERT-BiGRU-ATT 模型结构

文本输入层中针对短文本作者识别中文本特征提取难度较大的情况，为避免复杂的特征设计，针对中文文本利用预训练的 BERT 模型生成特征向量。将上述高质量的向量输入下游模型。

在文本深层次信息提取中采用了 Bi-GRU 神经网络作为下游模型，GRU 是目前较为流行的循环神经网络（RNN）的一种，在 LSTM 的基础上诞生，适用于学习长期依赖。它相较 LSTM 而言训练参数更少，训练更快且需要更少的数据来泛化，十分贴合短文本作者识别的特点，本文采用的 Bi-GRU 由正向 GRU 和逆向 GRU 组合而成，其优点在可很好地理解文本上下文信息，捕获文本语境特征。

注意力机制层对 Bi-GRU 提取到的特征向量加以优化，更好地为重要信息内容分配权重的同时获得文本的更深层特征。

最后将得到的特征向量输入 A-softmax 层进行分类，完成短文本作者的识别。

2.1 BERT 模型

BERT 模型是谷歌人工智能研究团队于 2018 年提出的里程碑式无监督预训练语言模型，其英文全称为 Bidirectional Encoder Representation from Transformers，即来自 Transformer 的双向编码器表示。Transformer 由编码器与解码器组成，是一种使用注意力机制搭建的序列到序列模型，能注意输入序列的不同位置以计算该序列表示能力。而 BERT 模型从名称上不难看出是一个用双向多层 Transformer 编码器作为特征提取器的预训练模型，其结构如图 2 所示。

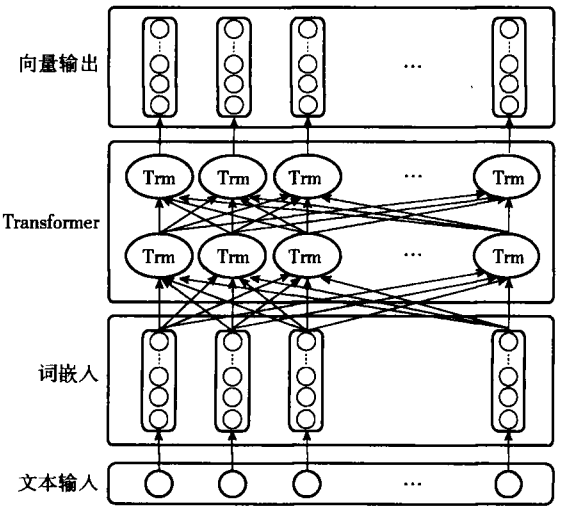

图 2 BERT 模型结构

此前的语言训练模型（例如 Word2Vec）都是单向的，只能从左至右或从右至左，无法得到整个文本的综合特征信息，这导致面对单词多义情况时容易出错，而 BERT 能很好解决这一问题。BERT 在预训练方法上采用两个非监督训练任务，分别为遮盖语言建模（Masked LM）与下一句预测（Next sentence prediction）。

2.1.1 遮盖语言建模

该任务可以简单概括为随机屏蔽部分输入的单词，然后根据未被屏蔽的内容对已屏蔽的单词实现

预测。在训练过程中随机屏蔽 15% 的单词，考虑到屏蔽标记对模型的影响，在这 15% 的单词中随机挑选十分之一替换成其他单词，五分之四被替换为“[MASK]”字符，剩下的维持原状。

2.1.2 下一句预测

该任务可以概括成判断连续的两句话中的第二句话是否紧随前句。其目的在于让模型更好理解两个句子之间的联系，提高上下文把控能力。

2.2 双向门控循环单元网络

GRU 是循环神经网络的一种[16]，是对 LSTM 的改进产物。GRU 对 LSTM 的结构进行精简，将 LSTM 中的输入门与遗忘门合并为更新门，并与重置门共同组成 GRU 单元。故相较 LSTM，GRU 在同等算力下训练时间大大减少，GRU 单元结构如图 3 所示。

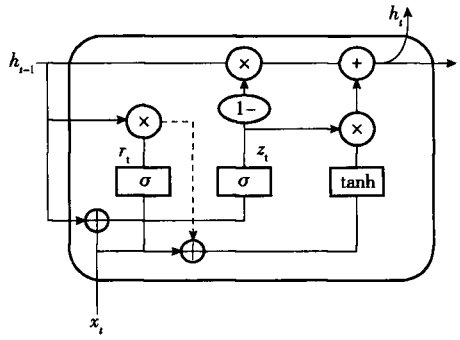

图 3 GRU 单元结构

图中的更新门表示为 z_t，r_t 代表重置门。更新门的作用在于控制上一时刻带至当前时刻的状态信息量，更新门的值大小与状态信息带入量成正相关。重置门的作用在于控制前一状态写入到当前的候选集 $\tilde{h}_t$ 量的多少，即 h_{t-1} 对 $\tilde{h}_t$ 的重要性，重置门大小与前一状态的信息写入量同样成正比。GRU 单元状态的计算公式如下：

$$z_t = \sigma(w_z[h_{t-1}, x_t] + b_z) \tag{1}$$

$$r_t = \sigma(w_r[h_{t-1}, x_t] + b_r) \tag{2}$$

$$\tilde{h}_t = \tanh(w_h[r_t h_{t-1}] + b_h) \tag{3}$$

$$h_t = (1 - z_t)h_{t-1} + z_t\tilde{h}_t \tag{4}$$

上述式中 σ 为 sigmod 函数，通过这个函数将数据转化为 0～1 区间的值以当门控信号。w_z、w_r、w_h 均为权重矩阵。$\tilde{h}_t$ 表示当前单元中需要更新的信息。t 时刻的输入向量为 x_t，h_t 为输出向量，包含了 t 时刻前的所有有效信息。

由于 GRU 网络中信息的传递是单向的，本文采用的 BiGRU 网络由一对方向相反的 GRU 单元组成，系双向传递的网络，弥补了普通 GRU 网络的单向传递缺陷，可以更充分捕获语句序列的文本特征。

式(5)、(6)分别表示 t 时刻前向、后向 GRU 单元隐含层输出，对输出拼接可得到 BiGRU 在该时刻的最终输出，如式(7)所示。

$$\overrightarrow{h_t} = GRU(x_t, \overrightarrow{h_{t-1}}) \tag{5}$$

$$\overleftarrow{h_t} = GRU(x_t, \overleftarrow{h_{t-1}}) \tag{6}$$

$$h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}] \tag{7}$$

2.3 注意力机制

注意力机制在 2014 年被 Mnih 等人[17]首次提出，是一种用来提升基于循环神经网络中 encoder + decoder 模型效果的机制，在自然语言处理等领域有着广泛应用。在自然语言处理中，注意力机制可以赋予句子中的每个词不同权重，能够更好地为重要信息分配权重，从而更加准确理解序列语义。首先利用激活函数形成对齐模型，随后获取注意力概率分布，最后将得到的权重矩阵与输入向量相乘得到最终输出结果。注意力机制公式如下：

$$m_t = \tanh(W_v h_t + b_v) \tag{8}$$

$$\alpha_t = \frac{\exp(m_t^{\mathrm{T}} k_v)}{\sum_t \exp(m_t^{\mathrm{T}} k_v)} \tag{9}$$

$$C = \sum_t \alpha_t^{\mathrm{T}} ht \tag{10}$$

上述式中，h_t 为 BiGRU 网络层的输出，W_v 是注意力模型可调节权重，b_v 为偏置项，式(9)计算结果 α_t 系注意力权重矩阵，其中权重值用 k_v 表示，C 为经过注意力模型计算后的特征向量。

2.4 A-softmax

A-softmax 可以看作 softmax 的增强版本，在较小的数据集合上有着良好的效果且具有不错的可解释性。与 softmax 相比，A-softmax 算法使得决策边界更加严格与分离，对更具区分性的特征学习有更大驱动力。关于 A-softmax 的损失函数定义如下：

$$L = \frac{1}{n}\sum_{n=1}^{N} -\log \frac{e^{\|x^n\|\phi(\theta_{y_n}^{(n)})}}{e^{\|x^n\|\phi(\theta_{y_n}^{(n)})} + \sum_{j \neq y_n} e^{\|x^{(n)}\|\cos\theta_j^{(n)}}} \tag{11}$$

其中 N 为训练样本的总数。$x^{(n)}$ 和 $y^{(n)}$ 分别表示第 n 个训练样本的特征向量和作者标签。$\theta_j^{(n)}$ 为 $x^{(n)}$ 与 wj 的夹角，$\theta_{y_n}^{(n)}$ 为 $x(n)$ 与权向量 W_{y_n} 之间的夹角。

3 实验与分析

3.1 实验环境及配置

为验证本文所提出模型的有效性，在如表 1 软硬件环境中进行实验。

表 1 实验环境及配置表

实验环境	详细信息
操作系统	Windows 10
处理器	Intel(R) Core(TM) i7-8750H
显卡	NVIDIA GeForce GTX 1060
内存大小	16G
硬盘	1TB
语言	Python
深度学习框架	Keras
框架后端	TensorFlow

3.2 数据集

本文数据集，分为微调 BERT 预训练模型所需的大量短文本博文与实验数据两部分，语料均来自微博。采用 python 的 scrapy 框架结合账号池与 IP 池基于 weibo.cn 站点进行微博信息爬取，共收集了 26.8 G 微博用户数据，经过清洗后的数据构成为用户名与该用户所有发文内容，从中挑选发文量超过 2 000 条的共 20 名用户制作测试集用作最后的模型准确率测试，共对应 51 249 条短文本，将上述 20 人的用户名作为该用户发文内容标签。剩余数据中的短文本内容用作训练语料，本文使用的是哈尔滨工业大学发布的基于全词遮罩(Whole Word Masking)技术的中文预训练模型 BERT-wwm，其语料为通用的中文维基，采用了哈尔滨工业大学 LTP 作为分词工具，对于微博这类灵活的短文本敏感度会稍差，故加以使用微博语料训练集进一步预训练。

3.3 评价指标

本文使用在作者识别中普遍使用的精确率(Precision)、召回率(Recall)以及调和平均数(F1 Score)3 项指标来测量各个模型的有效性。精确率表示所有预测正例样本的准确率，召回率用来度量有多少正例样本被分为正例，F1 则对精确率与召回率进行调和，得出整体评价。各指标定义公式如下：

$$\text{精确率} = \frac{TP}{TP + FP} \tag{12}$$

$$\text{召回率} = \frac{TP}{TP + FN} \tag{13}$$

$$\text{调和平均数} = \frac{2 \times \text{精确率} \times \text{召回率}}{\text{精确率} + \text{召回率}} \tag{14}$$

其中，TP 表示正确预测正例样本的数量，被误判为正例样本的负例样本的数量用 FP 表示，FN 代表被误判成负例样本的正例样本数。

3.4 实验结果及分析

接下来将进行两组实验分别对本文提出的 BERT-BiGRU-ATT 模型进行作者识别有效性验证。一个实验将比较 BERT 预训练模型与 Word2Vec、fastText 与 GPT 3 种不同词向量表示工具在对于短文本词向量提取方面效果的优劣，实验结果如表 2 所示；另一个实验将 BERT-BiGRU-ATT 模型作者识别效果结果同 SVM、TextCNN 与 BERT-BiGRU 3 种模型进行对比，实验结果如表 3 所示。

表 2 不同词向量提取效果对比实验结果

模型	精确率	召回率	F1 值
Word2Vec-BiGRU-ATT	0.846	0.839	0.842
fastText-BiGRU-ATT	0.851	0.849	0.850
GPT-BiGRU-ATT	0.919	0.907	0.913
BERT-BiGRU-ATT	0.947	0.926	0.936

表 3 不同模型效果对比实验结果

模型	精确率	召回率	F1 值
SVM	0.744	0.738	0.741
TextCNN	0.841	0.832	0.836
BERT-BiLSTM	0.918	0.922	0.920
BERT-BiGRU-ATT	0.947	0.926	0.936

最终结果表明，BERT 在词向量提取效果方面均优于其他 3 种方式。在模型效果上，相较其他 3 种模型，BERT-BiGRU-ATT 模型在精确率、召回率、F1 值上的表现均处于领先地位。

由第一个实验的结果发现，采用不同词向量提取方式对模型的效果存在不同程度的影响。fastText 方法优于 Word2Vec，是因为 fastText 在训练词向量时将 subword 纳入考虑范围，且引入了字符级 n-gram，使之更好地处理长词与低频词汇，在面对训练语料库以外的单词时也完成了词向量构建工作。GPT 是一个生成式预训练模型，其特征抽取器采用了多层 Transformer 解码器构成，与 fastText 方法相比 GPT 能够捕捉语义信息以及识别多义词，所以采用 GPT 作为词向量工具的模型各方面效果均优于采用 fastText 的模型，F1 值提升了 6.3%。虽然 BERT 与 GPT 均采用 transformer，但 BERT 使用的是双向编

码，相比单向捕获信息的 GPT 模型，BERT 能利用全部上下文信息，在词向量提取方面会更有优势。

从第 2 个实验结果来看，除 SVM 以外的 3 种模型在中文短文本作者识别任务中的精确率、召回率以及精确率与召回率的调和平均数数值上均超过 84%，这表明将深度学习的方法运用在作者识别领域是可行且有效好的效果。与 TextCNN 进行比较，BERT-BiLSTM 模型的 F1 值高出 8.4%，原因在于 TextCNN 采用窗口滑动的方式提取文本特征，是单向的，而 BiLSTM 可以联系上下文捕获文本特征，并且利用 BERT 预训练模型进行词嵌入可以提取更具表示能力的词向量，更好地满足下游任务。BERT-BiGRU-ATT 与 BERT-BiGRU 相比，F1 值提升了 1.6%，由此可见的是增加注意力机制可以赋予重要的信息更高权重从而进一步提取有效地提取文本特征。

4 结语

针对中文短文本作者的识别，本文运用 BERT 模型提取文本特征，采用混合了 BiGRU 网络与注意力的深度学习模型，对提取的文本特征进行深层次特征提取，最后利用 A-softmax 实现作者识别。与传统的作者识别模型对比发现，本文所提出的 BERT-BiGRU-ATT 模型在针对短文本作者的分类上效果更好，可以运用到实战中帮助互联网监管者打击网络犯罪。

本文的研究仍存在改进空间，其一是 BERT 预训练模型使用的短文本语料可以大量补充，但是需要足够的算力与资源支撑模型的训练，其二可以采取提高文本分类效果这一思路来提升作者识别准确度，如可以引入对抗学习网络对文本分类效果进行提升。

参 考 文 献

[1] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[J]. 【缺期刊名】, 2018, 【缺卷册页】.

[2] MENDENHALL T C. The characteristic curves of composition[J]. Science, 1887, 9(124): 237 - 249.

[3] MOHTASSEB H, AHMED A. Mining online diaries for blogger identification[J]. Lecture Notes in Engineering & Computer Science, 2009, 2176(1): 295 - 302.

[4] PILLAY S R, SOLORIO T. Authorship attribution of web forum posts[C]// Ecrime Researchers Summit, 2010: 1 - 7.

[5] CRISTANI M, ROFFO G, SEGALIN C, et al. Conversationally-inspired stylometric features for authorship attribution in instant messaging[C]// Proceedings of the 20th ACM international conference on Multimedia, 2012: 1121 - 1124.

[6] INCHES G, HARVEY M, CRESTANI F. Finding participants in a chat: Authorship attribution for conversational documents[C]// 2013 International Conference on Social-Computing, 2013: 272 - 279.

[7] HOLLINGSWORTH C. Using dependency-based annotations for authorship identification[C]// International Conference on Text, Speech and Dialogue. Berlin, Heidelberg: Springer, 2012: 314 - 319.

[8] 王少康，董科军，阎保平. 基于语句节奏特征的作者身份识别研究[J]. 计算机工程，2011，37(9)：4 - 5，8.

[9] 李晓军，刘怀亮，杜坤. 一种基于复杂网络模型的作者身份识别方法[J]. 图书情报工作，2015，59(18)：102 - 107.

[10] TANG X, LIANG S, LIU Z. Authorship attribution of the golden lotus based on text classification methods[C]// Proceedings of the 2019 3rd International Conference on Innovation in Artificial Intelligence, 2019: 69 - 72.

[11] 祁瑞华，杨德礼，郭旭，等. 基于多层面文体特征的博客作者身份识别研究[J]. 情报学报，2015，34(6)：628 - 634.

[12] YANG M, ZHU D, TANG Y, et al. Authorship attribution with topic drift model[C]// Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, 2017: 5015 - 5016.

[13] ZHANG R, HU Z, GUO H, et al. Syntax encoding with application in authorship attribution[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018: 2742 - 2753.

[14] 徐晓霖，蔡满春，芦天亮. 基于深度学习的中文微博作者身份识别研究[J]. 计算机应用研究，2020，37(1)：16 - 18，25.

[15] 冯勇，屈渤浩，徐红艳，等. 融合 TF-IDF 和 LDA 的中文 FastText 短文本分类方法[J]. 应用科学学报，2019，37(3)：378 - 388.

[16] CHO K, MERRIENBOER V B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation [J]. 【缺期刊名】, 2014, 【缺卷册页】.

[17] MNIH V, HEESS N, GRAVES A. Recurrent models of visual attention[C]// Advances in Neural Information Processing Systems, 2014: 2204 - 2212.

（责任编辑　于瑞华）