基于神经网络中文短文本作者识别研究

李孟林[1], 罗文华[1], 李绍鸣[2]
(1. 中国刑事警察学院网络犯罪侦查系, 辽宁沈阳 110854;
2. 沈阳航空航天大学人机智能研究中心, 辽宁沈阳 110136)

摘 要 随着互联网应用的日益普及,短文本作为电子数据证据在法庭科学中日益重要,法院亟需对大量网络聊天内容作者归属进行同一认定。传统机器学习方法对特征选取非常敏感,因为在实践中较难提取到准确的作者写作习惯特征,所以影响了传统机器学习方法的实践效果。针对文本短、特征少、特征提取困难的缺点,提出了融合多属性的神经网络中文短文本作者识别方法。首先将文本的结构特征、语义特征、发送时间、发送位置、发送频率等属性融合进文本序列,对文本序列进行词向量化表示,采用卷积层和 Bi-LSTM 层自动提取局部特征和上下文关系特征,通过注意力机制动态调整特征权重,使用 Softmax 分类器得到文本作者。以最大熵模型做对比实验,实验结果表明卷积层和 Bi-LSTM 层能"学习"到短文本上下文特征,注意力机制能更多"学习"到文本序列不同位置的关键特征,融合多属性的神经网络方法的作者识别精度比传统模型大约提高了 5%。

关键词 短文本; 多属性; Bi-LSTM; 最大熵; 作者识别
中图分类号 TP393.08

Research on Author Recognition of Chinese Short Text Based on Neural Networks

LI Menglin[1], LUO Wenhua[1], LI Shaoming[2]
(1. Department of Cyber Crime Investigation, Criminal Investigation Police University of China, Shenyang 110854, China;
2. Shenyang Aerospace University, Human-computer Intelligence Research Center, Shenyang 110136, China)

Abstract: With the increasing popularity of Internet applications, short text as electronic data evidence is increasingly important in forensic science. The court urgently needs to identify the author of a large number of online chat content. Traditional machine learning methods are very sensitive to feature selection, because it is difficult to extract accurate author style recognition features in practice, so it affects the practical effect of traditional machine learning methods. In view of the shortcomings of short text, including few features and difficult feature extraction, a Chinese short text author recognition method based on a neural network with multi-attribute fusion was proposed. Firstly, the text structure features, semantic features, sending time, sending location, sending frequency and other attributes are integrated into the text sequence, and the text sequence is represented by word vectorization. Local features and context features are extracted automatically by convolutional layer and Bi-LSTM layer, and the feature weight is adjusted dynamically through the attention mechanism, and the text author is obtained by Softmax classifier. Using the maximum entropy model as a comparative experiment, the results show that the convolution lay-

收稿日期 2019-12-20
作者简介 李孟林(1994—),男,湖北宜昌人,在读硕士研究生。研究方向为网络安全执法技术、自然语言处理。
通讯作者 罗文华(1977—),男,教授。E-mail:luowenhua770404@126.com

er and the Bi-LSTM layer can "learn" the short text context features, and the attention mechanism can "learn" the key features of different positions of the text sequence. The author's recognition accuracy of the neural network method with multi-attribute fusion is improved by about 5% compared with the traditional model.

Key words: short text; Multi-Attribute; Bi-LSTM; maximum entropy; authorship attribution

0 引言

文本作者的身份识别一直以来都是法庭科学的重点。通常情况下，文本作者的身份识别多数以笔迹鉴定的方式在法庭呈现。但是随着信息技术在日常生活中的普及，犯罪分子为了便利和逃避侦查以电子书写方式代替手写方式来隐藏身份，如勒索信、暴恐信、诈骗信、举报信等，在没有笔迹的情况下，如何判断文本作者显得越发迫切和重要。

随着互联网的发展，短文本大量涌现。短文本通常是指长度较短，一般在 160 个字符以内的电子文本，包括了微博、电子邮件、手机短信(SMS)、即时聊天记录(微信/QQ/MSN/Skype)等。由于短文本在日常生活中的普遍使用使得短文本作为电子数据证据的案例越来越多，法庭也亟需对大量的短文本进行作者识别。因此，基于短文本的犯罪嫌疑人写作习惯乃至身份特征的分析成为法庭科学日益关注的热点与难点。

文本作者的身份识别来源于作品作者识别，国际上针对此类问题的研究已经比较丰富，并积累了一定的成功经验。Shunichi Ishihara[1]借助语言模型工具，针对英文短信，利用似然比对作者进行判断；Sarah R. Boutwell[2]则针对 Twitter 文本语料库，为每名作者构建统计模型，实现对文本作者的识别；Monika Nawrot[3]提出了一种混合算法，通过函数为英文电子邮件的不同特征赋以不同的权重，进而识别出作者。

国内虽然对此起步较晚，但在文本作者识别方面已经进行了大量探索。武晓春等[4]依据文体学理论，充分利用功能词以外的其他词汇，提出一种新的基于词汇语义分析的相似度评估方法。年洪东等[5]使用以词汇为基础的多种统计量作为识别特征对现代文学作品进行了作者身份识别研究。祁瑞华等[6]探索性的建立了由词汇特征、浅层句法特征、深层句法特征和结构特征组成的多层面文体风格特征模型，为网络文本作者身份的自动识别提供了新的技术思路。廖志芳等[7]以 HowNet 为语料库，以 Standford 为语法解析工具，结合中文语句语义相似性以及语法相似性，提出一种基于语法语义的短文本相似度算法。卢玲等[8]基于 Word Embedding 文本语义扩展方法，通过构造卷积神经网络(CNN)来提取扩展文本的特征，提高了中文新闻标题分类准确性。范亚超等[9]采用降噪自编码器深度模型提取文本结构特征，通过支持向量机分类器完成作者识别，准确率最高达到了 78.2%。米硕等[10]提出了一种新的基于循环神经网络(RNN)和卷积神经网络(CNN)的网络架构，对电子邮件的作者识别取得了不错的效果。

现有模型研究多是针对长文本，无法直接应用于短文本中。而短文本的模型方法均是针对特定语料库(新闻标题、微博、电子邮件)，识别结果依赖于特征的选择。此外，中文与英文的巨大差异，西方国家主要以英文为应用场景的研究成果在中文应用场景下并不能很好地适用，因此研究适合于中文应用场景的网络短文本作者识别模型非常有现实意义。

1 融合多属性的作者识别系统

为了克服短文本噪声大、特征稀疏、特征提取困难等缺陷，提出融合多属性的作者识别系统。该系统思想是通过提取主谓宾结构特征、语气词特征、附属信息特征，为文本引入更多的外部特征，将短文本做一个特征延伸。利用卷积神经网络(CNN)特征提取能力强的特点，提取文本序列特征，进一步得到内部特征和外部特征相融合的文本特征表示，并将其输入双向长短时记忆网络(Bi-LSTM)[12]，发挥 Bi-LSTM 对序列数据进行建模的优势，得到上下文关系特征的文本表示。通过注意力机制(Attention)对文本不同位置特征信息赋以不同权重，从而对短文本作者进行有效识别。融合多属性的作者识别系统架构如图 1 所示。

1.1 预处理与特征提取

通过对短文本内容进行分析，发现短文本中大

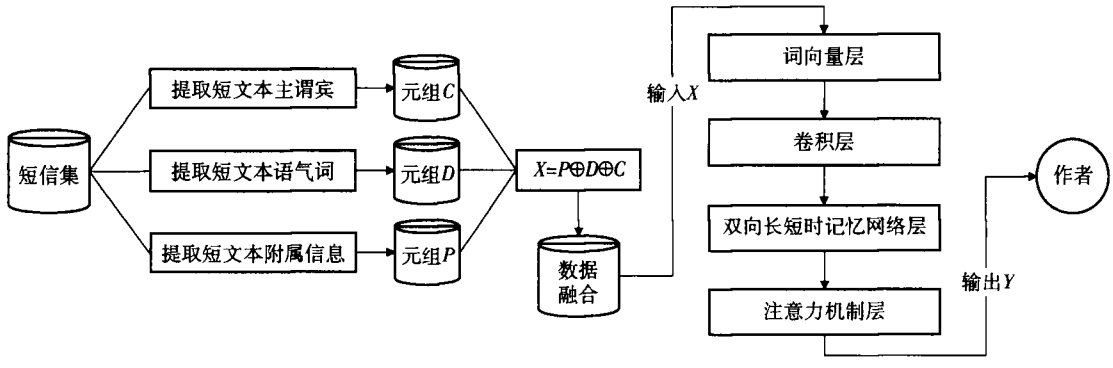

图1 融合多属性的作者识别系统架构

量出现语气词和省略指代的情况，说明短文本虽然长度有限，但语言表达习惯却因人而异。首先，有些人在表达时习惯带上语气词，如“走，吃饭啦”，而有些人在表达时就不习惯使用语气词，如“走，吃饭”，同样的表达，即使是都使用语气词，也可能有所不同，比如“走，吃饭呀”。其次，受地域影响，有些人不按常规的主谓宾方式表达，习惯省略某一结构，甚至出现倒装，这在短文本中大量出现，如“晓不得”“晓得不”和“不晓得”的表达。最后，除文本自身外，文本的附属信息在一定程度上体现了作者的身份特征。因此，提取短文本语气词特征、短文本主谓宾特征、文本附属信息特征进行作者识别。为了提取这些特征，首先就要对短文本进行预处理，主要包括中文分词和词性标注。

中文分词就是对短文本按词切分的过程，目的是为了词性标注。例如，“他去北京了呀”经过中文分词后变成“他/去/北京/了/呀”。

词性标注就是对分词后的结果按照其上下文意思标记词性。词性包括名词(n)、代词(r)、动词(v)、形容词(a)、连词(c)、助词(u)等词性。例句词性标注后的结果如图2所示，其中“r”表示代词，“v”表示动词，“ns”表示名词中的地名，“u”表示助词。

提取短文本语气词。根据自己制定的语气词表提取出文本中语气词，语气词表如表1所示，从表1可以发现，“了”不在语气词表里，而“呀”在语气词表里，因此提取“呀”作为语气词特征：

表1 语气词表

分类	语气词
元音语气词	阿、啊、呃、哇、呀、也、耶、哟、噢
辅音单音节	吧、罢、嘛、呗、啦、嘞、哩、咧、咯、喽、吗、嘛、么、哪、呢、呐、呵、哈、噢、兮、哉
辅音双音节	罢了、不成、得了、而已、的话、来着、了得、也罢、已而、着呢、着哩、着呐、也好、是的、一般、再说、不过
辅音多音节	就是了

提取文本主谓宾特征。利用文献[13]中使用的哈尔滨工业大学 pyltp 依存句法分析器提取文本的主语、谓语、宾语等文本主体结构。依存句法分析结果如图3所示，其中“HED”表示核心关系，指整个句子的核心(Root)；“SBV”表示主谓关系，指“他”和“去”是主谓关系；“VOB”表示动宾关系，指“去”和“北京”是动宾关系；“RAD”表示右附加关系，指“了”、“呀”和“去”是右附加关系。根据该依存句法关系，可以提取出主语“他”，谓语“去”，宾语“北京”。

他 去 北京 了 呀

r v ns u u

图2 分词后词性标注的结果

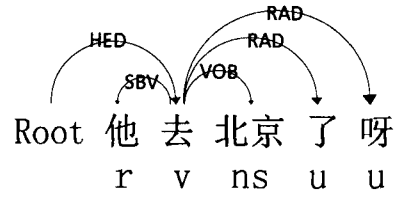

图3 依存句法分析结果

最终，例句“他去北京了呀”，经过语气词提取得到语气词“呀”，经过依存分析提取找到主语“他”，谓语“去”，宾语“北京”，最终将其转化为 D = {他，去，北京，呀}。

提取文本附属信息特征。短文本除了文本自身以外，通常还包含有一定附属信息，文本的附属信息也更能反映出作者的书写习惯和生活习惯，例如：文

本发送频率、发送地点、发送时间、性别、年龄、输入法、系统版本号等，这里提取文本发送频率、发送地点、发送时间作为文本附属信息特征。

1.2 多属性融合

由于双向循环神经网络层是对序列数据进行建模，很难从短文本中学习到文本的有效特征，即使卷积神经网络自动提取特征能力很强，面对着长度较短的短文本也是力不从心，为此将文本的语气词特征、主谓宾结构特征、附属信息特征融合进原始文本序列，在一定程度上延长了神经网络捕获短文本特征的时间序列长度，从而能够让双向循环神经网络更充分地捕获文本特征。

首先，将文本分词并提取语气词后的短文本语义特征放入元组 P 中，然后将依存分析后提取的主谓宾结构特征放入元组 D 中，最后将短文本附属信息特征放入元组 C 中，C 可以简单的表示为集合 $C=\{$发送频率，发送地点，发送时间$\}$。作为对比实验，一方面将该 3 组特征作为最大熵模型对文本进行作者识别的特征直接输入。另一方面将提取的短文本语义特征 P、主谓宾结构特征 D 和短文本附属信息特征 C 做一个拼接，如公式(1)：

$$x=P\oplus D\oplus C \tag{1}$$

其中⊕代表相邻两个元素的连接符，x 作为神经网络的输入文本序列。通过引入文本的外部属性特征，增加了短文本的文本结构长度，从而能够让神经网络学习到更多的文本特征。

1.3 作者识别模型

在将数据特征融合的基础上，借鉴文献[14]提出的问句分类方法架构图，设置了词向量层、卷积层、双向长短时记忆网络层、注意力机制层。如图 4 所示，首先，将融合多属性后的短文本序列以词向量的形式来表示并输入神经网络；接下来将进入卷积层，充分发挥卷积层特征提取能力强的优势，更好的提取句子的特征，将提取的特征和分词后的文本放入循环神经网络层，循环神经网络能够很好捕获数据变化规律；接着利用注意力机制来识别文本主要特征；最后经过分类器得出作者识别结果。

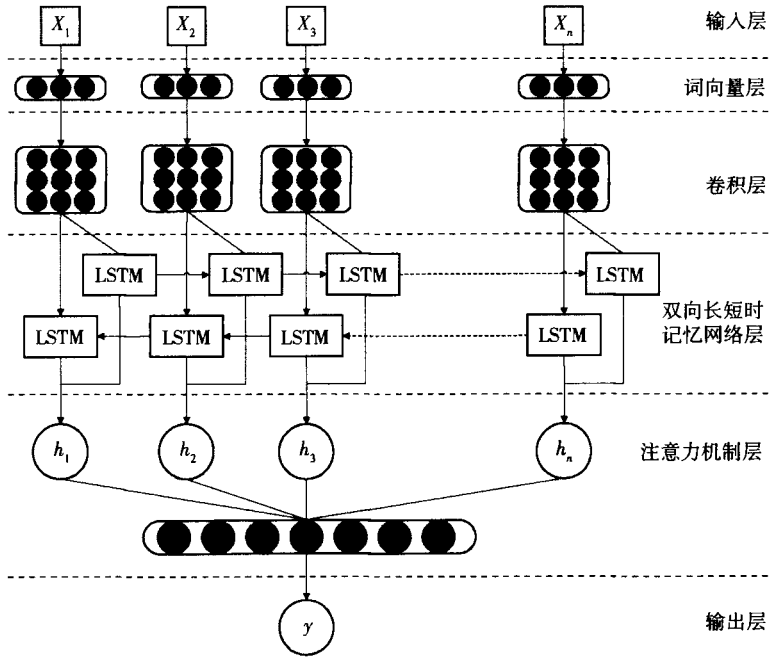

图 4 基于 Attention 的 CNN + Bi-LSTM 模型图

1.3.1 词向量层

首先，对输入层的中文短文本进行分词，并通过 Word2Vec[15] 将文本中的词转化为词向量形式，这些词向量蕴含了文本的信息，将融合后的属性信息同样进行向量化表示。接下来，在词向量层加入了文本更多的特征信息，假设文本 Q 包含 n 个单词，$Q=\{x_1,x_2,\cdots,x_n\}$，x_i 代表文本中的第 i 个词，在文本信息后边加入该文本对应的发送时间、发送频率、发送地点等特征信息。公式(2)所示，首先根据文本建立一个词典 Dic，初始化一个词向量矩阵 $\boldsymbol{E}_w$ 来获得词向量，根据单词在词典中的位置 v_i，可以将词转变为词向量 $\boldsymbol{e}_i$：

$$e_i = E_w v_i \tag{2}$$

其中，v_i 是采用独立热编码的形式，在模型训练过程中不断更新。经过这个步骤，文本将以 $embeddings_Q = \{e_1, e_2, \cdots, e_n\}$ 的形式进入下一层网络。

1.3.2 卷积层

在经过词向量层后，每个文本 t 可以表示成如下形式，其中 T 为句子长度：

$$t = [e_1, e_2, \cdots, e_n]^{\mathrm{T}} \tag{3}$$

卷积过程中每次选取不同维度的卷积核提取文本中的特征，每次特征提取可以由卷积核在文本上进行一次卷积操作，每次选取窗口大小为 m 的核对文本 t 进行如下操作：

$$c_i = f(wh_{i:i+m-1} + b) \tag{4}$$

其中 w 是过滤器，$h_{i:i+m-1}$ 是词向量，b 是一个偏置项，f 是一个非线性函数，文本最后被表示为：

$$c^* = [c_0, c_1, \cdots, c_{n-m}] \tag{5}$$

1.3.3 双向长短时记忆网络层

长短时记忆网络主要由 3 部分构成：(1)输入门；(2)输出门；(3)遗忘门。长短时记忆网络通过“门”的结构让信息有选择性地影响循环神经网络中每个时刻的状态，使用 sigmoid 函数(σ)作为激活函数的全连接神经网络层会输出一个 0 到 1 之间的数值，描述当前有多少信息量可以通过这个结构。输入门决定哪些信息加入到当前状态来生成新的状态信息，遗忘门的作用是让网络“忘记”之前没有用的信息，神经网络在得到新状态后产生新的输出是通过输出门完成的。不妨设输入门(i_t)的权重矩阵为 W_{x_i}、W_{h_i}、W_{c_i}、b_i；遗忘门(f_t)的权重矩阵为 W_{x_f}、W_{h_f}、W_{c_f}、b_f；输出门(o_t)的权重矩阵为 W_{x_o}、W_{h_o}、W_{c_o}、b_o；候选信息(g_t)的权重矩阵为 W_{x_c}、W_{h_f}、W_{c_c}、b_c。在 t 时刻，当前时刻网络的输入值为 x_t，上一时刻 LSTM 的输出值为 h_{t-1}，以及上一时刻的单元状态为 c_{t-1}，而当前时刻 LSTM 输出值是 h_t，b_i、b_f、b_o、g_t 分是输入门、遗忘门、输出门以及候选信息的偏置项，具体每个“门”的公式定义如下：

$$i_t = \sigma(W_{x_i}x_t + W_{h_i}h_{t-1} + W_{c_i}c_{t-1} + b_i) \tag{6}$$

$$f_t = \sigma(W_{x_f}x_t + W_{h_f}h_{t-1} + W_{c_f}c_{t-1} + b_f) \tag{7}$$

$$g_t = \tanh(W_{x_c}x_t + W_{h_f}h_{t-1} + W_{c_c}c_{t-1} + b_c) \tag{8}$$

$$c_t = i_t g_t + f_t c_{t-1} \tag{9}$$

$$o_t = \sigma(W_{x_o}x_t + W_{h_o}h_{t-1} + W_{c_o}c_t + b_o) \tag{10}$$

$$h_t = o_t \tanh(c_t) \tag{11}$$

因此，当前结构单元状态是由之前单元状态的权重和当前单元所生成的当前信息决定。在经典的循环神经网络中，状态的传输是从前往后单向传递，只考虑到上文中的信息，而忽略了下文中的信息。Bi-LSTM 由两个单向的循环神经网络结合，每一时刻的输入会提供两个相反的循环神经网络，这样每一时刻的输出，都考虑到上下文信息。

1.3.4 注意力机制层

为了更好的捕捉文本中的有效信息，抓住文本重点信息，本文在作者识别模型中加入了注意力机制，该注意力机制的权重矩阵通过如下公式得到：

$$M = \tanh(\boldsymbol{H}) \tag{12}$$

$$\boldsymbol{\alpha} = softmax(\boldsymbol{w}^{\mathrm{T}}M) \tag{13}$$

$$r = \boldsymbol{H}\boldsymbol{\alpha}^{\mathrm{T}} \tag{14}$$

其中，$\boldsymbol{H}$ 表示由上层 Bi-LSTM 网络输出向量所组成的矩阵，$\boldsymbol{w}^{\mathrm{T}}$ 是一个参数向量。向量矩阵 $\boldsymbol{H}$ 通过 $tanh$ 函数得到隐层表示 M，M 和 $\boldsymbol{w}^{\mathrm{T}}$ 通过 softmax 函数得到权重矩阵 $\boldsymbol{\alpha}$。在向量矩阵 $\boldsymbol{H}$ 的基础上乘以该权重矩阵，就得到了句子的文本的表示 r。最后用于识别文本作者的向量 $\boldsymbol{c}^*$ 表示如下：

$$\boldsymbol{c}^* = \tanh(r) \tag{15}$$

1.3.5 分类器

这一层网络结构，使用 softmax 分类器，在 y 中预测 x 所属的作者，$\boldsymbol{w}$ 是参数向量，b 是偏置项，分类器利用隐藏状态 c^* 作为输入：

$$p(y|x) = softmax(Wc^* + b) \tag{16}$$

$$y = \arg\max p(y|x) \tag{17}$$

损失函数如下：

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} t_i \log(y_i) + \lambda \| \theta \|_F^2 \tag{18}$$

其中，t 是 one-hot 表示，m 是作者的数量，y 代表估计每个作者的概率，θ 表示正则化参数。

2 实验研究

2.1 实验数据

研究采用新加坡国立大学收集的短信库(NUS SMS Corpus)，使用 2015.03.09 版本进行，该版本包含有 31 465 条中文短信，分别归属于 594 位作者。每条短信都伴随有相应的附属信息，例如：date(发送日期)、time(时间)、text(内容)、UserID(发送人唯一识别号)、manufactuer(手机厂商)、age(年龄)、sex(性别)、city(发送短信时所在的城市)、experience(手机使用时间)、frequency(每天发送短信的数量)、inputMethod(输入法)等。

2.2 实验设置

为使实验结果更具有普遍性，从实验数据中随机抽取80%作为训练集，其余20%作为测试集，采用机器学习方法中最大熵模型作为对比模型。实验设置2组对照实验。第1组分别采用最大熵模型和神经网络模型进行对比，第2组对神经网络模型和融合后的神经网络模型进行对比。第1组设置4个模型，分别是最大熵模型、长短期记忆神经网络(LSTM)、卷积神经网络+长短期记忆神经网络(CNN+LSTM)、卷积神经网络+长短期记忆神经网络+注意力机制(CNN+LSTM+Attention)。第2组设置3个模型，这3个模型均是在融合多属性情况下进行的，分别是LSTM、CNN+LSTM、CNN+LSTM+Attention。

本次实验最大熵模型选取7个特征，分别是短文本的主语、谓语、宾语、语气词、发送时间、发送频率、发送地点，最大熵模型中的参数估计使用GIS算法，迭代100次后结束。由于文本长度较短，本次实验在卷积层设置窗口大小为3，在训练时使用随机梯度下降算法，batch_size大小设置为50，droupout rate设置为0.5，epoch大小设置为1 000，使用的词向量是谷歌通过Word2Vec预先训练好的包含1 000亿词汇量的谷歌新闻语料。

2.3 评价指标

采用3个评价指标对本次实验结果进行评价，即准确率(P)、召回率(R)、F_1值(F_1)，计算公式如下：

$$P=\frac{\text{分类正确的正例数目}}{\text{分类器判为正例的数目}} \tag{19}$$

$$R=\frac{\text{识别正确的数目}}{\text{所有正例的数目}} \tag{20}$$

$$F_1=\frac{2\times P\times R}{P+R} \tag{21}$$

2.4 结果分析

相同数据集在融合多属性前后对比实验的准确率、召回率、F_1值如表2、表3所示。

表2 数据集在融合多属性前各个模型实验的准确率、召回率、F_1值

Model	P	R	F_1
最大熵	0.870 0	0.864 2	0.867 1
Bi-LSTM	0.905 0	0.896 3	0.900 6
CNN+Bi-LSTM	0.914 8	0.903 7	0.909 2
CNN+Bi-LSTM+Attention	0.935 4	0.921 7	0.928 5

表3 数据集在融合多属性后各个模型实验的准确率、召回率、F_1值

Model	P	R	F_1
Bi-LSTM	0.912 8	0.905 6	0.909 2
CNN+Bi-LSTM	0.923 4	0.895 6	0.909 3
CNN+Bi-LSTM+Attention	0.947 8	0.937 9	0.942 8

(1)通过对比最大熵和Bi-LSTM模型的实验结果可知，结合上下文信息的Bi-LSTM模型比传统的最大熵模型更优，说明深度神经网络捕获了更深层次文本特征，F_1值提高了3.35%。

(2)通过对比Bi-LSTM和CNN+Bi-LSTM模型的实验结果可知，CNN层很大程度上获取了短文本的内部语义特征，F_1值提高了30.86%。

(3)通过对比CNN+Bi-LSTM和CNN+Bi-LSTM+Attention模型实验结果可知，引入注意力机制，很大程度上获取了句子不同位置的特征信息。F_1值提高了1.93%。

(4)通过对比Bi-LSTM和融合多属性的Bi-LSTM模型实验结果可知，融合多属性的Bi-LSTM捕获到了引入的文本外部特征，F_1值提高了0.86%。

(5)通过对比CNN+Bi-LSTM模型实验结果可知，融合多属性的CNN+Bi-LSTM，融合多属性的CNN+Bi-LSTM模型聚焦于文本序列深层次的语义特征。尽管F_1值仅提高了0.01%，但实验的准确率提高了0.86%。

(6)通过对比CNN+Bi-LSTM+Attention和融合多属性的CNN+Bi-LSTM+Attention模型的实验结果可知，通过引入外部属性特征，文本序列融入了更多的特征信息，Attention机制的加入，让模型更多聚焦于文本不同位置特征信息。模型的准确率、召回率、F_1值分别提高了1.24%、1.62%、1.43%。

3 结语

本文提出了融合多属性的神经网络中文短文本作者识别方法，通过对短文本语气词特征、主谓宾结构特征的提取，结合文本发送时间、发送位置、年龄、发送频率等附属信息特征，使用最大熵模型与传统的神经网络模型进行作者识别的对比实验，在此基础上采用了融合多属性的神经网络模型进一步提高了实验的准确率，在实验数据集上验证了融合多属性的神经网络方法的有效性。

参 考 文 献

[1] ISHIHARA S. A forensic authorship classification in SMS messages: A likelihood ratio based approach using N-gram[C] // The Australian National University School. Proc of Australasian Language Technology Association Workshop 2011, Canberra: The Australian National University, c2011:47 - 56.

[2] BOUTWELL S R. Authorship attribution of short messages using multimodal features [D]. California: Naval Postgraduate School, 2011.

[3] NAWROT M. Automatic author attribution for short text documents [C] // Human Language Technology LTC 2009. Proceedings of the 4th Language and Technology Conference, Berlin: Springer-Verlag, c2011:468 - 477.

[4] 武晓春，黄萱菁，吴立德. 基于语义分析的作者身份识别方法研究[J]. 中文信息学报，2006，20(6)：63 - 70.

[5] 年洪东，陈小荷，王东波. 现当代文学作品的作者身份识别研究[J]. 计算机工程与应用，2010(4)：230 - 233.

[6] 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(6)：628 - 634.

[7] 廖志芳，周国恩，李俊锋，等. 中文短文本语法语义相似度算法[J]. 湖南大学学报(自然科学版)，2016，43(2)：135 - 140.

[8] 卢玲，杨武，杨有俊，等. 结合语义扩展和卷积神经网络的中文短文本分类方法[J]. 计算机应用，2017(12)：160 - 165.

[9] 范亚超，罗天健，周昌乐. 基于降噪自编码器特征学习的作者识别及其在《西游记》诗词上的应用[J]. 厦门大学学报(自然科学版)，2018，57(6)：150 - 155.

[10] 米硕，孙瑞彬，李欣，等. 基于循环神经网络(RNN)和卷积神经网络(CNN)对电子邮件的作者识别[J]. 科技创新与应用，2018(1)：24 - 25.

[11] 李荣陆，王建会，陈晓云，等. 使用最大熵模型进行中文文本分类[J]. 计算机研究与发展，2005，42(18)：94 - 101.

[12] HOCHREITER S, SCHMINDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735 - 1780.

[13] 周娜，何润奇. 基于文本情感分析的文化综艺节目综合评价——以央视文化类综艺节目《国家宝藏》为例[J]. 中南民族大学学报(人文社会科学版)，2019，39(5)：175 - 180.

[14] 史梦飞，杨燕，贺樑，等. 基于 Bi-LSTM 和 CNN 并包含注意力机制的社区问答问句分类方法[J]. 计算机系统应用，2018，27(9)：157 - 162.

[15] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. Computer Science, 2013: 1 - 12.

（责任编辑　于瑞华）