

# 基于章节段落的学术文献结构功能识别方法研究<sup>\*</sup>

刘忠宝<sup>1,2</sup> 王宇飞<sup>3</sup> 赵文娟<sup>1,2</sup>

- (1. 云计算与物联网技术福建省高等学校重点实验室(泉州信息工程学院), 泉州 362000; 2. 北京语言大学语言智能研究院, 北京 100083; 3. 中北大学软件学院, 太原 030051)

**摘要:**[目的/意义] 学术文献的结构功能识别问题是学术大数据知识挖掘与分析的研究热点, 研究如何从中挖掘有效的知识, 有助于从更深层次、更细粒度理解学术文献, 从而促进学术文献语义理解的发展。[方法/过程] 以学术文献的章节段落为研究对象, 对卷积神经网络(CNN)、长短期记忆网络(LSTM)、来自变换器的双向编码器表征量(BERT)等深度学习模型的学术文献章节段落的结构功能识别性能进行比较研究, 并与传统机器学习算法SVM进行了对比实验。[结果/结论] 中国知网语料集上的实验结果表明, 与SVM和LSTM、CNN模型相比, BERT模型具有更优的结构功能识别性能, 其在整体识别性能上F1值达到0.66, 在具体结构功能的识别性能上F1值最高达到0.79。此外, 通过引入混淆矩阵, 对功能结构误识情况进行分析。

**关键词:** 结构功能 学术文献 章节段落 深度学习 识别方法

**分类号:** G256

**DOI:** 10.31193/SSAP.J.ISSN.2096-6695.2020.03.04

## 0 引言

学术文献是研究人员进行科学研究和学术交流的重要载体。而学术文献的结构功能则是对学术论文篇章结构的概括与描述, 对学术文献进行功能划分有助于更细粒度地展示学术文献的逻辑结构, 便于研究人员进行更深层次的研究。因此在图书情报、信息科学等领域, 学术文献

<sup>\*</sup> 本文系福建省社会科学规划项目“大数据环境下面向图书馆资源的跨媒体知识服务研究”(项目编号: FJ2019B052)的研究成果之一。

[作者简介] 刘忠宝(ORCID: 0000-0002-0038-2462), 男, 教授, 博士, 研究方向为信息资源管理, Email: liuzb@nuc.edu.cn; 王宇飞(ORCID: 0000-0002-8504-9641), 男, 硕士研究生, 研究方向为知识组织与知识发现, Email: 1546563183@qq.com; 赵文娟(ORCID: 0000-0002-8389-768X, 通讯作者), 女, 讲师, 硕士, 研究方向为知识组织、信息检索, Email: 63896887@qq.com。

的结构功能识别成为学术大数据知识挖掘与分析的研究热点。特别是随着大数据时代的到来和人工智能技术的发展,如何将一些性能优良的智能化技术方法引入到学术文献结构功能识别,对于提高已有方法的识别效率具有现实意义。鉴于此,笔者在引入一系列深度学习模型的基础上,对学术文献的摘要、章节标题、章节内容、章节段落等结构功能识别方法展开研究。本文属于系列研究的一部分,着重探讨基于章节段落的学术文献结构功能识别方法。

## 1 研究进展

近年来,学术文献结构功能识别受到研究人员的关注,并取得了不少研究成果。Constantin 等<sup>[1]</sup>从学术文献的字体和布局出发,建立了基于规则的学术文献结构功能识别系统 PDFX,该系统具有对学术文献的标题、章节、表格、参考文献等结构功能的识别能力。真实语料集上的实验结果表明,该系统的 F1 值最高达到 77.45%。Tuarob 等<sup>[2]</sup>利用机器学习算法,对学术文献的章节边界进行自动划分,以实现学术文献的功能结构识别。黄永等从章节标题<sup>[3]</sup>、章节内容<sup>[4]</sup>、篇章段落<sup>[5]</sup>三方面对学术文献进行结构功能识别,并对三方面的识别效果进行了比较研究。Ren 等<sup>[6]</sup>结合学术文献结构,提出一种基于残差网络的特征提取器,用以识别学术文献的功能结构。王东波等<sup>[7]</sup>从学术文献的摘要出发,分别利用长短期记忆网络(Long Short-Term Memory, LSTM)、支持向量机(Support Vector Machine, SVM)、长短期记忆网络与条件随机场(Conditional Random Field, CRF)混合模型 LSTM-CRF、卷积神经网络(Convolutional Neural Networks, CNN)与条件随机场混合模型 CNN-CRF 等对 3672 篇情报领域的学术文献进行了摘要结构功能识别研究。此外,该作者<sup>[8]</sup>在真实语料集上,通过设计双向长短期记忆网络(Bidirectional LSTM, BiLSTM)、支持向量机 SVM 和条件随机场 CRF 等模型的学术文献结构功能识别实验,得知 CRF 具有最优的识别性能,其 F1 值能够达到 92.88%。沈思等<sup>[9]</sup>以学术文献摘要中的字为基本语义单位,构建了基于 LSTM-CRF 的摘要结构功能自动识别模型。与支持向量机 SVM、条件随机场 CRF、循环神经网络(Recurrent Neural Network, RNN)和长短期记忆网络 LSTM 相比,该模型的识别性能最优,其 F1 值最高可达到 85.47%。王佳敏等<sup>[10]</sup>提出多层次融合的学术文献结构功能识别模型,该模型首先利用深度学习模型对不同层次学术文献进行结构功能识别,然后利用投票法对识别结果进行融合。李楠等<sup>[11]</sup>着重探讨基于深度学习模型的结构功能识别方法对学科领域的适用性问题,6452 篇学术文献上的实验结果表明,学科差异性对深度学习模型识别性能具有显著影响。

由上述研究进展可以看出,机器学习算法逐渐被引入到学术文献结构功能识别研究,但这些方法的复杂度大多与文献规模呈指数级或对数级关系,随着学术文献规模的增大,这些方法的性能大幅下降,无法保证识别效果。有部分研究引入适用于大数据处理的深度学习模型,如 CNN、RNN、LSTM 等,但这些模型属于早期的深度学习模型,在大多数应用场景下,其学习能力明显低于近年来新出现的 BERT (Bidirectional Encoder Representation from Transformers) 模型,该模型在学术文献结构功能识别的适用性问题值得探讨。基于上述分析,本文以学术文献的章节段落为

研究对象, 在引入 BERT 深度学习模型的基础上, 对基于章节段落的学术文献结构功能识别方法进行研究。

## 2 模型引入

### 2.1 支持向量机

支持向量机 (SVM)<sup>[12]</sup> 是传统机器学习算法中常用的分类算法, 在文本分类中有着较好的分类效果。该模型通过构造一个超平面作为决策平面, 使两类样本之间的分类间隔最大, 其优化目标函数与约束条件如公式 (1) 所示。

$$\begin{aligned} \min & \left( \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i \right) \\ \text{s.t. } & y_i (w^T x_i + b) \geq 1 - \xi_i, i = 1, 2, \dots, n \end{aligned} \quad (1)$$

其中,  $x_i$  为样本点,  $w$  为超平面的法向量,  $b$  为偏置量,  $\xi_i$  为松弛变量, 表示对噪声的容忍度,  $C$  为惩罚系数。

SVM 一般应用于二分类任务, 而学术文献结构功能识别是一个多分类任务, 因此本文通过组合多个二分类支持向量机来解决学术文献的章节段落结构功能识别问题。

### 2.2 卷积神经网络

卷积神经网络 CNN<sup>[13]</sup> 是一种带有卷积结构的深度神经网络, 该网络的卷积结构能够大幅减少模型训练的参数数量, 避免“过拟合”现象的发生。如图 1 所示, CNN 由输入层、卷积层、池化层和输出层构成。

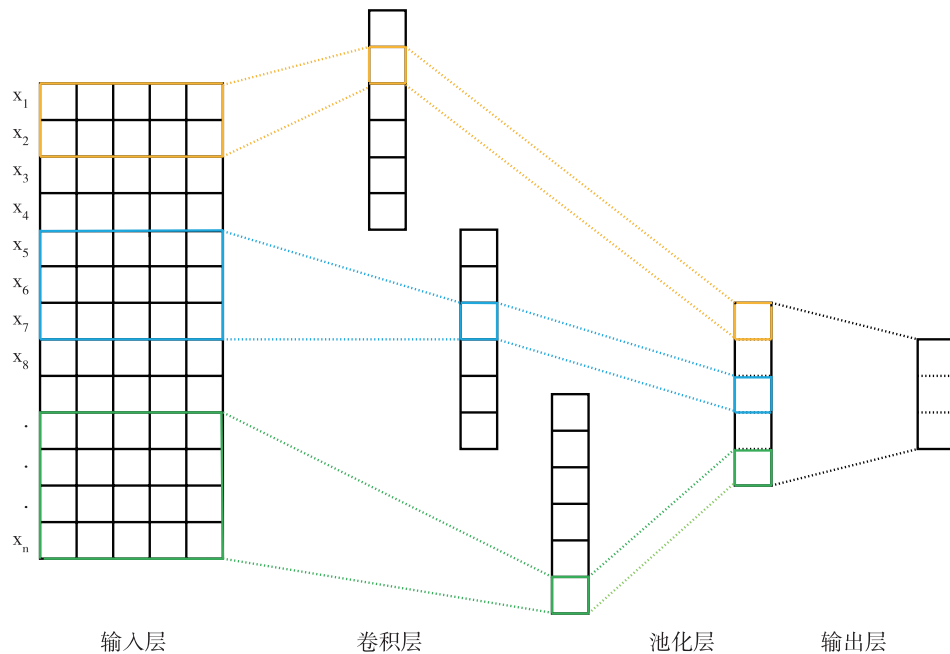


图 1 CNN 结构示意图

在图1中,输入层利用 Word2vec 模型对分词后的输入语料 ( $x_1, x_2, \dots, x_n$ ) 进行向量化表示,得到输入向量。卷积层利用一组不同大小的卷积核对输入向量进行卷积运算,以提取不同粒度的语义特征,本文采用大小分别为 2、3、4 的卷积核。池化层对提取的语义特征进行池化操作以降低该特征维度。输出层将池化后的语义特征通过全连接网络并利用 softmax 激活函数作为分类器得到识别结果。

### 2.3 长短时记忆神经网络

长短时记忆神经网络 (Long Short-Term Memory, LSTM) <sup>[14, 15]</sup> 是一种特殊的 RNN, 与传统 RNN 相比, 它设计了隐藏层单元结构。LSTM 利用两个门结构控制记忆单元  $c_t$  的信息, 其中遗忘门  $f_t$  决定上一时刻的记忆单元  $c_{t-1}$  有多少信息保留到当前时刻的  $c_t$ ; 输入门  $i_t$  决定当前时刻的输入  $x_t$  有多少信息保留到记忆单元  $c_t$ ; 输出门  $o_t$  用来控制记忆单元  $c_t$  有多少信息传递给当前输出值  $h_t$ 。LSTM 的结构如图 2 所示, 其中  $i_t$ 、 $f_t$ 、 $o_t$  分别表示输入门、遗忘门和输出门,  $\tilde{c}_t$  表示候选记忆单元,  $c_t$  表示记忆单元,  $\sigma$  和  $\tanh$  分别表示 sigmoid 激活函数和  $\tanh$  激活函数,  $x_t$  为输入向量,  $h_t$  为输出向量,  $W$  和  $b$  为对应的权重矩阵和偏置并可以作为参数在训练过程中学习得到。

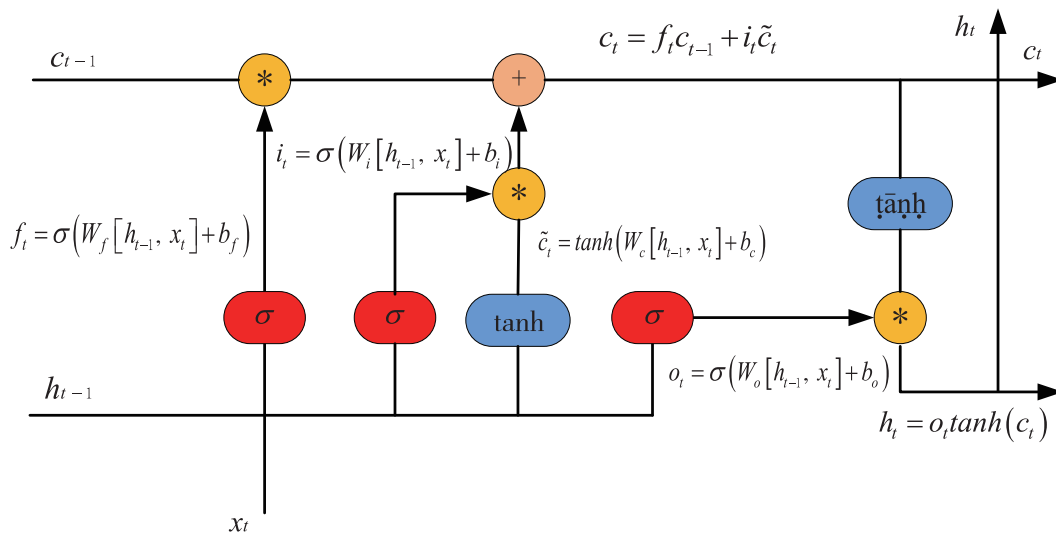


图2 LSTM 结构示意图

LSTM 模型的基本工作流程为: 首先, 输入层通过 Word2Vec 对输入语料 ( $x_1, x_2, \dots, x_n$ ) 进行向量化表示, 得到输入向量; 然后, 在隐藏层通过 LSTM 对输入向量进行特征提取, 得到特征向量; 最后, 输出层将特征向量通过全连接层并利用 softmax 函数得到最终的识别结果。

### 2.4 BERT 模型

BERT 模型 <sup>[16]</sup> 引入 Transformer 特征抽取器 <sup>[17]</sup> 进一步增强了特征提取的能力, 能够更充分地描述学术文献之间字符级、词级、句子级及句间关系特征。BERT 模型的结构如图 3 所示, 其

中 [CLS] 表示输入语料 ( $x_1, x_2, \dots, x_n$ ) 的起始标识,  $E_i (i=1, 2, \dots, n)$  表示模型的输入向量, 该向量由字向量、句向量和位置向量拼接而成。Trm 为 Transformer 编码器, C 和  $T_i (i=1, 2, \dots, n)$  表示模型输出的特征向量。

BERT 模型的基本工作流程为: 输入层对分词后的输入语料进行向量化表示 ( $x_1, x_2, \dots, x_n$ ), 并在起始位置标识 [CLS]; 嵌入层生成输入向量对应的词向量 ( $E_{[CLS]}, E_1, E_2, \dots, E_n$ ); 编码层利用多层双向 Transformer 编码器对输入向量进行特征提取, 生成特征向量 (C,  $T_1, T_2, \dots, T_n$ ); 输出层通过全连接网络并利用 softmax 分类函数得到识别结果。

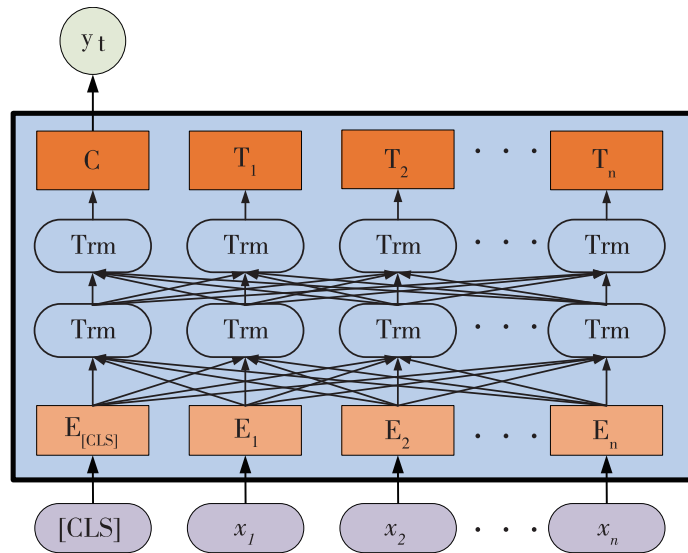


图3 BERT 模型结构示意图

### 3 实验分析

#### 3.1 实验环境

表 1 给出了实验环境与相应配置。

表 1 实验环境与配置

实验环境	配置
操作系统	Ubuntu16.04
内存	16G
编程语言	Python3.7
深度学习框架	TensorFlow1.14

续表

实验环境	配置
词向量训练工具	Word2vec
中文分词工具	jieba 分词
GPU	NVIDIA GeForce GTX 1060

### 3.2 实验语料集

对中国知网 (<https://www.cnki.net/>) 2016年至2019年之间在《情报学报》《情报理论与实践》《情报科学》《情报探索》《图书情报工作》《情报杂志》《图书情报知识》等期刊发表的学术文献进行全文抽取, 经过去重、删除异常值等人工处理后得到795篇学术文献, 作为本文的实验语料集。由于学术文献的结构功能存在不一致的情况, 因此, 对不同学术文献的结构功能标识进行归一化处理, 处理结果如表2所示。

表2 学术文献结构功能标识对照表

结构功能	对应的章节标题
摘要	摘要
引言	引言、前言、简介、概述
相关研究	相关研究、相关工作、相关研究综述、研究综述、国内外研究现状、文献回顾、文献概述、研究背景、相关研究与理论基础、研究现状、相关研究回顾、相关研究状况、文献综述、研究进展
方法	实验方法、数据来源与方法、模型介绍、模型引入、模型简介、模型构建、数据与方法、方法描述、研究方法综述、框架与方法、研究设计与研究方法、系统构建、数据采集与研究方法、研究方法、研究设计、研究方法设计、实验设计、数据获取与处理、研究框架、模型和方法、研究问题与方法、数据源和模型简介
实验	结果与分析、研究结果、实验、实验与结果分析、结果分析、数据分析、实验及结果、实验过程与分析、实验分析、实证研究、实验过程、实证分析、实验与效果评价
结论	结论、结束语、结语、总结、结论与讨论、结论与展望、总结与展望、结论与启示、总结与讨论、小结

如表2所示, 本文根据学术文献的章节标题关键词将不同学术文献的结构功能统一标识为{摘要, 引言, 相关研究, 方法, 实验, 结论}, 作为实验的分类标签。

### 3.3 参数设置

参数的选取对深度学习模型的性能至关重要。因此, 选取一小部分实验语料集进行预实验, 进而得到如表3所示的实验参数设置表, 其中batch\_size表示每一次训练所选取的语料集大小, epochs表示实验语料集循环使用的次数, dropout\_keep表示神经元的保留概率, 本文引入dropout技术以避免“过拟合”现象的发生, lr表示模型的学习率, hidden\_units表示隐藏层的神经元数。

表 3 实验参数设置表

实验参数	CNN	LSTM	BERT
batch_size	64	128	16
epochs	30	50	5
dropout_keep	0.5	0.8	0.9
lr	1e-3	1e-3	2e-5
hidden_units	128	128	768

### 3.4 实验性能评价

利用准确率 P (Precision)、召回率 R (Recall) 及调和平均值 F1 (F-measure) 等指标对 CNN、LSTM、BERT 等模型的结构功能识别性能进行评价。三项指标的计算公式如下:

$$P = \frac{TP}{TP + FP} \quad (2)$$

$$R = \frac{TP}{TP + FN} \quad (3)$$

$$F1 = \frac{2 \times P \times R}{P + R} \quad (4)$$

其中 TP、FP、FN 表示正确识别、错误识别、无法识别的结构功能数量。

### 3.5 实验结果

本文利用 5 折交叉验证法进行实验。CNN、LSTM、BERT 等模型的章节段落识别性能对比结果如表 4 所示。其中表 4 的“整体”为整体识别结果, 其值为各结构功能的平均值。

表 4 章节段落识别性能对比

模型	评价指标	摘要	引言	相关研究	方法	实验	结论	整体
LSTM	P	0.36	0.65	0.57	0.63	0.70	0.44	0.56
	R	0.31	0.56	0.59	0.70	0.70	0.37	0.54
	F1	0.33	0.60	0.58	0.66	0.70	0.40	0.55
CNN	P	0.64	0.60	0.59	0.61	0.74	0.48	0.61
	R	0.85	0.43	0.53	0.71	0.72	0.48	0.62
	F1	0.73	0.50	0.56	0.66	0.73	0.48	0.61
BERT	P	0.80	0.59	0.61	0.66	0.75	0.47	0.65
	R	0.77	0.50	0.48	0.70	0.77	0.75	0.66
	F1	0.79	0.54	0.58	0.68	0.76	0.58	0.66



由表4可以得出以下结论:

(1) 整体识别结果。从整体性能来看, BERT模型的P、R、F1值在各模型中均为最高, CNN模型次之, LSTM模型最低。其中BERT模型的F1值为0.66, 比CNN模型高0.05, 比LSTM模型高0.11。因此, 对学术文献的章节段落进行结构功能识别, BERT模型具有最优的整体识别性能。

(2) 各结构功能识别结果。从具体结构功能识别性能来看, 对“摘要”识别, BERT模型性能最优, LSTM模型性能最差, 两者F1值相差0.66; 对“引言”识别, LSTM模型性能最优, CNN模型性能最差, 两者F1值相差0.10, BERT模型的F1值为0.54, 与最优性能相差0.06; 对“相关研究”识别, BERT模型和LSTM模型均最优, CNN模型最差, 三者F1值相差0.02; 对“方法”识别, BERT模型性能最优, 其余两类模型性能相当, 三者F1值相差0.02; 对“实验”识别, BERT模型性能最优, LSTM模型性能最差, 两者F1值相差0.06; 对“结论”识别, BERT模型性能最优, LSTM模型性能最差, 两者F1值相差0.18。由此可见, 除“引言”外, BERT模型均具有最优的结构功能识别性能。此外, CNN模型对“实验”识别性能最优, 其次是“摘要”“方法”等; LSTM模型对“实验”识别性能最优, 其次是“方法”“引言”等; BERT模型对“摘要”识别性能最优, 其次是“实验”“方法”等。不难看出, 三类模型均对“方法”和“实验”的识别性能较好, 其主要原因是这两部分在学术文献中所占的篇幅较多, 有利于深度学习模型学习训练, 以提取更深层次的语义特征, 从而获得更优的功能结构识别结果。

### 3.6 对比实验

本实验采用Python机器学习工具包sklearn提供的SVM算法进行分类实验, 在实验中将惩罚系数设置为0.5, 并采用线性函数作为核函数。通过该实验对比传统机器学习模型与BERT模型在学术文献章节段落上的结构功能识别差异。表5为SVM的实验结果。

表5 基于SVM的章节段落识别结果

结构功能	P	R	F1
摘要	0.51	0.36	0.42
引言	0.56	0.18	0.28
相关研究	0.64	0.24	0.34
方法	0.74	0.51	0.60
实验	0.62	0.75	0.68
结论	0.56	0.44	0.49
整体	0.60	0.41	0.47

由表5可以看出, 相较于传统机器学习算法SVM, 本文采用的BERT模型在整体识别结果(P、R、F1)上分别提高0.05、0.25、0.19, 该结果说明BERT模型在学术文献章节段落结构功能



识别中的性能要优于传统的机器学习算法。以 SVM 为代表的机器学习算法依赖于分类规则与构建特征的质量, 在学术文献这种语料关联度较高的文本中, 其类别浅层特征不明显, 会严重影响算法的性能。而 BERT 模型能够有效学习词语、句子等关系特征, 在学术文本多分类任务中优势更大。

### 3.7 误识分析

本文选取综合识别性能最优的 BERT 模型为研究对象, 通过构建如表 6 所示的混淆矩阵, 来分析该模型的误识情况。表 6 中的行表示相应的结构功能被误识为其他类别的数量, 列表示其他结构功能被误识为该类别的数量。

表 6 混淆矩阵

结构功能	摘要	引言	相关研究	方法	实验	结论
摘要	123	12	6	3	3	12
引言	0	390	129	129	93	39
相关研究	9	153	474	144	156	57
方法	0	57	114	1473	441	9
实验	6	42	48	474	2073	36
结论	15	9	6	3	12	138

由表 6 的行可以看出, “摘要” 主要被误识为 “引言” 和 “结论”; “引言” 主要被误识为 “相关研究”; “相关研究” 主要被误识为 “引言” 和 “实验”; “方法” 主要被误识为 “实验”; “实验” 主要被误识为 “方法”; “结论” 主要被误识为 “摘要”。究其原因, 相关结构功能具有较强的关联度, 很多描述内容具有较大的相似性, 这增大了深度学习模型的识别难度, 进而出现某些结构功能误识的情况。

## 4 结 语

本文在学术大数据的背景下, 对基于章节段落的学术文献结构功能识别方法进行研究。通过设计实验分别对比分析了 BERT 模型与传统深度学习模型 CNN、LSTM 和传统机器学习算法 SVM 的识别性能表现, 实验结果表明, BERT 模型具有最优的识别性能。最后, 通过引入混淆矩阵, 给出了 BERT 模型的功能结构误识结果, 并对该结果进行了分析。

### 【参考文献】

[ 1 ] CONSTANTIN A, PETTIFER S, VORONKOV A. PDFX: fullyautomated PDF-to-XML conversion of scientific literature [ C ]. Proceedings of the ACM symposium on document engineering, Florence, Italy, 2013: 177-180.

- [ 2 ] TUAROB S, MITRA P, GILES C L. A hybrid approach to discover semantic hierarchical sections in scholarly documents [ C ]. Proceedings of the 13th International Conference on Document Analysis and Recognition, Nancy, France, 2015: 1081–1085.
- [ 3 ] 陆伟, 黄永, 程齐凯. 学术文本的结构功能识别——功能框架及基于章节标题的识别 [ J ]. 情报学报, 2014,33(9):979–985.
- [ 4 ] 黄永, 陆伟, 程齐凯. 学术文本的结构功能识别——基于章节内容的识别 [ J ]. 情报科学, 2016,35(3): 293–300.
- [ 5 ] 黄永, 陆伟, 程齐凯, 等. 学术文本的结构功能识别——基于段落的识别 [ J ]. 情报学报, 2016,35(5): 530–538.
- [ 6 ] REN X H, ZHOU Y, HUANG Z, et al. A novel text structure feature extractor for Chinese scene text detection and recognition [ J ]. IEEE Access, 2017, 5: 3193–3204.
- [ 7 ] 王东波, 陆昊翔, 周鑫, 等. 面向摘要结构功能划分的模型性能比较研究 [ J ]. 图书情报工作, 2018, 62(12):84–90.
- [ 8 ] 王东波, 高瑞卿, 叶文豪, 等. 不同特征下的学术文本结构功能自动识别研究 [ J ]. 情报学报, 2018, 37(10):997–1008.
- [ 9 ] 沈思, 胡昊天, 叶文豪, 等. 基于全字语义的摘要结构功能自动识别研究 [ J ]. 情报学报, 2019, 38(1):79–88.
- [ 10 ] 王佳敏, 陆伟, 刘家伟, 等. 多层次融合的学术文本结构功能识别研究 [ J ]. 图书情报工作, 2019, 63(13):95–104.
- [ 11 ] 李楠, 方丽, 张逸飞. 学术文本结构功能深度学习识别方法的多学科对比分析 [ J ]. 现代情报, 2019, 39(12):55–63+87.
- [ 12 ] HSU C W, LIN C J. A comparison of methods for multiclass support vector machines [ J ]. IEEE Transactions on Neural Networks, 2002,13(2):415–425.
- [ 13 ] HUGHES M, LI I, KOTOULAS S, et al. Medical text classification using convolutional neural networks [ J ]. Studies in Health Technology and Informatics, 2017, 235: 246–50.
- [ 14 ] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [ J ]. Neural Computation, 1997, 9(8): 1735–1780.
- [ 15 ] LIU P, QIU X, HUANG X. Recurrent neural network for text classification with multi-task learning [ C ]. Proceedings of the 25th International Joint Conference on Artificial Intelligence, New York, USA, 2016: 2873–2879.
- [ 16 ] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [ C ]. Proceedings of the 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, USA, 2019: 4171–4186.
- [ 17 ] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [ C ]. Proceedings of the Advances in Neural Information Processing Systems, Long Beach, USA, 2017: 5998–6008.

# Research on the Recognition Method of Academic Literature Structure Function Based on Paragraph

LIU Zhongbao<sup>1,2</sup> WANG Yufei<sup>3</sup> ZHAO Wenjuan<sup>1,2</sup>

- (1. Key Laboratory of Cloud Computing and Internet-of-Things Technology (Quanzhou University of Information Engineering), Fujian Province University, Quanzhou 362000, China;
2. Institute of Language Intelligence, Beijing Language and Culture University, Beijing 100083, China;
3. School of Software, North University of China, Taiyuan 030051, China)

---

**Abstract:** [ **Purpose/significance** ] The recognition of the academic literature structure function is an important research hotspot in the knowledge mining and analysis of academic big data. It is helpful to understand the academic literature from a deeper and more fine-grained level through the effective knowledge mining and to promote the development of semantic understanding of academic literature. [ **Method/process** ] This paper investigates the recognition method of academic literature structure function based on paragraphs by comparing the recognition performance of Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), Bidirectional Encoder Representations from Transformers (BERT), and makes a comparative experiment with the traditional machine learning algorithm SVM. [ **Result/conclusion** ] The experimental results on CNKI corpus show that, compared with SVM, LSTM and CNN models, BERT model has better structure function recognition performance, and its F1-value reaches 0.66 in the overall recognition performance and 0.79 in the specific structure function recognition performance. In addition, confusion matrix is introduced to analysis the misrecognition.

**Keywords:** Structure function; Academic literature; Paragraph; Deep learning; Recognition method

---

( 本文责编: 周 霞 )