doi:10.3772/j.issn.2095-915x.2020.03.005

基于层次注意力网络的论证区间识别研究

王鑫[1,2] 程齐凯[1,2] 马永强[1,2] 罗卓然[1,2]

1. 武汉大学信息管理学院 武汉 430072；
2. 武汉大学信息检索与知识挖掘研究所 武汉 430072

开放科学（资源服务）标识码（OSID）

摘要：学术文本论证区间识别是一项论证学术文献内容和分析修辞结构的研究，针对当前研究过多依赖人工经验来构建规则和特征的现状，以及论证区间识别研究存在时效性差、泛化能力弱等问题，本文采用基于层次注意力机制的 HAN 深度学习模型对学术文本论证区间识别进行了研究。本文首先构建了一个基于层次注意力机制的论证区间识别模型，阐述了该模型的整体架构和作用机制。其次，针对生物医学领域提出了一个论证区间 9 分类体系，在 PubMed 生物医学数据集上，用 LSTM 和 SVM 两种文本分类算法与 HAN 模型进行对比实验。研究结果表明，本文所采用的 HAN 模型在各个类别的论证区间识别上效果均为最优，F1 值达到了 0.90，能够较好的完成论证区间识别研究。最后，对实验结果进行错误总结和分析，并指出了下一步的研究方向。

关键词：层次注意力网络；论证区间；深度学习；文本分类

中图分类号：G35

Research on Argumentative Zoning Recognition Based on Hierarchical Attention Network

WANG Xin[1,2] CHENG Qikai[1,2] MA Yongqiang[1,2] LUO Zhuoran[1,2]

1. School of Information Management, Wuhan University, Wuhan 430072, China;
2. Information Retrieval and Knowledge Mining Laboratory, Wuhan University, Wuhan 430072, China

基金项目：国家自然科学基金面上项目“基于多语义信息融合的学术文献引文推荐研究”（7167030644）和国家自然科学基金青年项目“基于深度语义挖掘的引文推荐多样化研究”（71704137）。

作者简介：王鑫（1996-），硕士研究生，研究方向：文本挖掘、信息检索，E-mail: x.wang@whu.edu.cn；程齐凯（1989-），博士，讲师，研究方向：文本挖掘、信息检索、科技情报分析；马永强（1997-），硕士研究生，研究方向：机器学习，信息检索；罗卓然（1993-），博士研究生，研究方向：文本挖掘、网络表示学习。

Abstract: Argumentative zoning recognition of academic texts is an analysis of the argumentation and rhetorical structure of academic literature content. Aiming at the problems that current research mostly relies on traditional artificial experience to build rules and features with poor effect and weak generalization ability in argumentative zoning identification, this paper uses the HAN deep learning model based on the hierarchical attention mechanism to identify argumentative zoning of academic text. The article first introduces the HAN model, and describes the overall structure and mechanism of the model. Secondly, for the biomedicine field, this paper proposes a 9-classification system and compares the text classification algorithms LSTM and SVM with the HAN model on the PubMed biomedical dataset. The results show that the HAN model used in this paper has the best effect, and the F1 value reaches 0.90, which can well complete the research on the recognition of argumentative zoning. Finally, we analyze the misclassification of the experimental results and point out the future research direction.

Keywords: Hierarchical attention network; argumentative zoning; deep learning, text classification

引言

随着大数据时代的到来和互联网的普及，学术文献发表的速度越来越快，使得学术资源总量呈现出一种井喷式增长的趋势。以微软学术为例，截止 2020 年 1 月，其收录数据量已达到 2.3 亿，而且自 2013 年以来其每年新增收录的论文数都突破 1000 万。面对浩如烟海的学术论文，科研人员从学术文本中获取学术信息时，其往往只对学术文本中特定的章节部分感兴趣 [1]，而且有研究表明不同的章节内容对不同科研人员的吸引力和重要性也是不同的 [2]。

学术文本的篇章结构通常都包含丰富的信息，例如文章的层次结构通常也代表着其要表达的逻辑结构 [1]。不同的学术文本可能使用同样的篇章结构来陈述以往的研究、阐述自身研究的问题以及研究方法等内容，因此科研人员可以通过合理有效安排篇章结构帮助其传达核心信息。实验物理学、生物学和心理学等领域的学术文章的显著特征之一是文本拥有类似的分区结构，比如生物学中典型的 IMRAD 的论文结构 [3]，以及计算机等其他学科中在 IMRAD 结构基础上新增结论或者相关研究等分区结构以适应本学科领域的特点。国内外不同学科领域的学者在研究学术文本篇章结构和语义内容的关系时，使用了例如论证区间（argumentative zoning）[4]、论证结构（argumentative structure）[5]、结构功能（structure function）[6] 等不同的表达形式对此类研究进行相关探索。

本文为进一步挖掘学术论文结构与内容之间的语义关系，使用论证区间来对学术文本的篇章结构进行研究。论证区间选取“论证”的角度去探究学术文本的结构性质和内容信息的关系。“论证”本意旨在提高或降低某个有争议的观点的可接受性 [7]，而论证区间旨在分析作者在文本中组织语句陈述观点的方式，其最初从修辞学 (rhetoric) 中衍生而来，现在主要是指对学术文本内容的论证和修辞结构的分析，已在计算语言学、文本挖掘等多个领域中被使用。论证区间已多次被证明对于解决有关学术文本的各类信息访问获取任务 [8,9]、信息检索任务 [10,11] 以及自动文摘任务 [12] 都是有效的，现已成为学术文本特征分析研究中的一个重要分支。因此，对学术文本进行论证区间自动识别，具

doi:10.3772/j.issn.2095-915x.2020.03.005

有重要的研究意义和实用价值。

目前，学术文本论证区间识别主要采用基于规则或传统机器学习的方法，该类方法一方面需构建大量的规则以提取文本特征，严重依赖人工经验和相关领域知识。另一方面，随着训练数据量的增加，该方法本身呈现出的过拟合、泛化能力差等局限性更加突出。随着人工智能技术在自然语言处理方面引用的深入，深度学习技术也愈来愈多地被应用于学术文本挖掘的研究中。深度学习利用不同类型的深层次神经网络，可对大规模的数据进行自动的特征提取和表示学习，非常善于表达大规模复杂数据的内在结构[13]。针对当前研究现状，本研究以篇章结构较为典型的生物医学领域的数据集为研究对象，并使用不同的机器学习模型去探索论证区间识别任务，最后对这些模型在论证区间识别上取得的实验结果的进行分析和评价。

1 相关研究

本研究主要是从论证区间文本内容的角度，利用深度学习方法对其进行自动分类识别。因此本章节将从论证区间分类研究、论证区间自动识别和深度学习文本分类三个维度来介绍本文的相关研究。

1.1 论证区间分类研究

1999 年 Teufel 提出论证区间概念时，其依据学术文本中句子的修辞地位将论证区间定为 7 类：研究目标、结构介绍、研究内容、研究背景、研究对比、基础研究和其他[4]。2009 年 Teufel 针对化学和计算机语言学科特点对论证区间的类型进行细化和扩展，重新将其划分为 15 类，命名为 AZ-II 分类体系[14]。Liakata 等人[15]针对生物医学学科的学术文本特点提出 CoreSCs 的体系，将内容划分为假设、动机、目标、对象、背景、方法、实验、模型、观察、结果和结论共 11 种类别。Antonio 等人[5]根据 MEDLINE/PubMed 数据集的特点将论证区间划分为背景、结论、方法、目标以及结果 5 种。陆伟等人[6]针对计算机领域文献结构的特点将学术文本论证区间分为引言、相关研究、方法、实验和结论 5 种类型。

1.2 论证区间自动识别

关于论证区间的研究，很重要的一环就是确定文献中某段区间文本内容所属的论证类别，即论证区间的自动识别。Liakata 等人[15]基于生物医学领域的文献数据集和 CoreSCs 体系，使用 SVM 算法对论证区间进行分类，结果表明效果最好的特征分别是 n-gram 词组、句法分析特征以及章节标题。Liu[16]借用 Word2Vec 的思想在 AZ-II 分类体系下分别使用句子平均向量化、段落向量以及特殊词向量三种方法对论证区间进行识别。黄永、陆伟等人分别基于段落内容[17]和章节内容[18]，使用 CRF 和 SVM 算法实现了学术文本各种区间功能的自动识别，同样取得了不错的效果。王东波等人[19]利用 JASIST 上发表的 1579 篇论文作为数据集，使用 CRF、双向长短时记忆网络模型和 SVM 三种模型对论证区间识别结果进行对比，最终选择 CRF 模型并人工选取特征进行干预，F1 值达到了 92.88%。王佳敏等人[20]以 ScienceDirect 数据集为例，使用投票法对多个模型的论证区间识别结果进行

多层次的融合，最终整体准确率、召回率和 F1 值分别达到 86%、84% 和 84%。李楠等人 [21] 对不同的学科领域的论证区间识别结果进行对比分析，发现学科的差异性对实验结果有显著的影响，医学领域的识别效果相对较好。

1.3 深度学习文本分类

近些年来，深度学习在文本分类领域取得了很大进展。基于深度学习的文本分类模型使用分布式词向量表示文本语义特征，不依赖人工经验而使用隐藏层自动组合这些特征，最终利用卷积神经网络、循环神经网络等深度学习模型以及结合注意力机制对文本进行分类。

Liu 等人 [22] 利用循环神经网络结合多任务学习的框架进行分类，在多个数据集上进行测试取得了不错的效果。Zhou 等人 [23] 利用长短时记忆网络模型（long short term memory，简称 LSTM），结合注意力机制对跨语言的情感分类任务进行了探索。Yang 等人 [24] 提出了层次注意力模型（hierarchical attention networks，简称 HAN），利用双向的循环神经网络，并分别在单词层面和句子层面添加注意力机制，提高了文本分类的准确性和模型的可解释性。

2 论证区间识别模型构建

论证的抽象结构应是较为宏观的，多个句子连贯合并在一起才能进行完整的论证 [25]。因此学术文本的论证区间本身即具有层次结构，即多个单词形成完整的句子，而多个句子形成基本的论证区间。基于此，本文构建了一个基于层次注意力机制的论证区间识别模型，用来进行学术文本的论证区间识别，该模型的整体结构如图 1 所示。

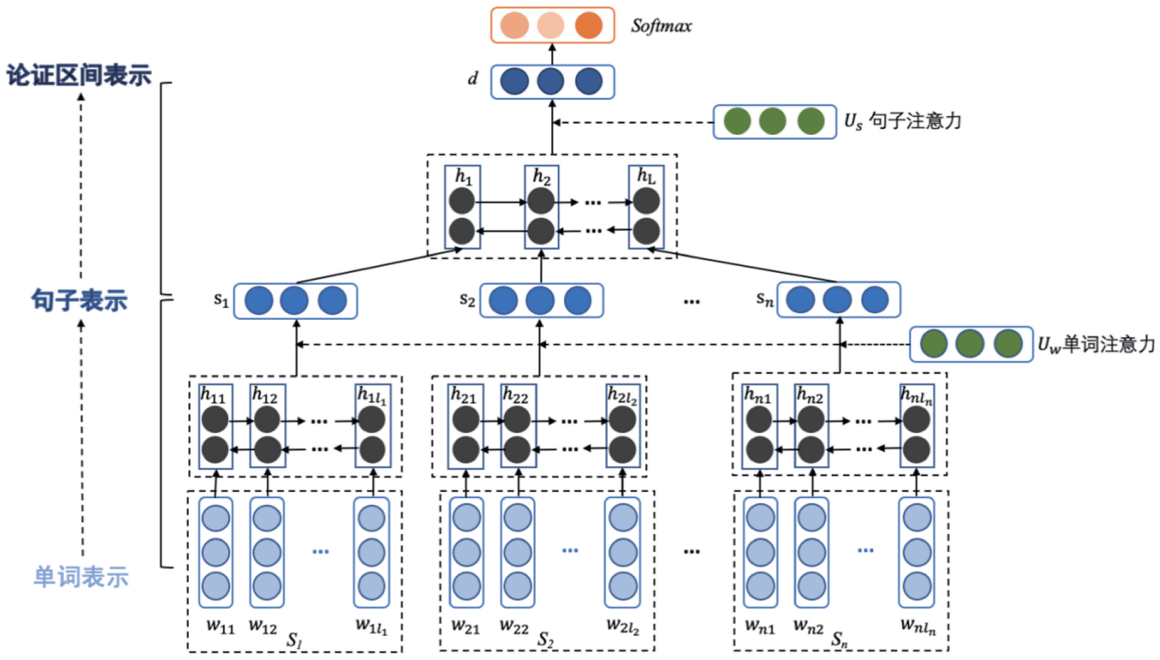

图 1 基于层次注意力机制的论证区间识别模型整体结构

doi:10.3772/j.issn.2095-915x.2020.03.005

基于层次注意力机制的论证区间识别模型由两个主要部分组成，即从单词到句子的表示和从句子到论证区间的表示，该模型生成的论证区间表示将作为该区间文本的最终特征进行分类。具体从结构上分析时，该模型可以分为输入层、词嵌入层、词编码层、词注意力层、句子编码层、句子注意力层和输出层 7 层。

（1）输入层对原始的论证区间文本进行分句和分词处理。假设某个论证区间文本被分为 L 个句子，每个句子由 T 个单词组成，则可使用 w_{it} 来表示整个论证区间的所有单词，其中 $i \in [1,L], t \in [1,T]$。

（2）词嵌入层将输入层的每一个单词 w_{it} 转换为词向量表示 x_{it}，本研究使用 TensorFlow 框架，模型运行前首先会初始化具体的词向量参数矩阵 W_e，该矩阵参数会随着网络的更新而更新，本层的转换过程如公式 1 所示：

$$x_{it}=W_e w_{it}, i \in [1,L], t \in [1,T] \quad (1)$$

（3）词编码层主要使用循环神经网络中的双向门控循环单元（Gated Recurrent Unit，简称 GRU）[26] 进行编码。GRU 和 LSTM 均是常用的循环神经网络门控算法。对比 LSTM 的结构，GRU 使用隐藏状态和候选隐藏状态的线形关系代替了 LSTM 中单元状态和隐藏状态的复杂联系，结构简单更易于计算训练。GRU 控制信息传输的结构也较为简单，主要由更新门（Update Gate）和重置门（Reset Gate）组成。更新门决定着上一时刻的隐藏状态信息保留到当前时刻隐藏状态的比例和当前候选隐藏状态信息保留到当前隐藏状态的比例，而重置门则决定着上一时刻的隐藏状态信息保留到当前时刻候选隐藏状态的比例。某时刻 t，GRU 的隐藏状态、更新门、候补隐藏状态和重置门的更新公式依次如下：

$$h_t = (1-z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \quad (2)$$

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \quad (3)$$

$$\tilde{h}_t = \tan h(W_h x_t + r_t \odot (U_h h_{t-1}) + b_h) \quad (4)$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \quad (5)$$

其中 h_t、$\tilde{h}_t$、h_{t-1} 代表 t 时刻的隐藏状态和候补隐藏状态以及 t–1 时刻的隐藏状态，x_t 代表输入文本的词向量表示，z_t 和 r_t 代表更新门和重置门，tan$h$ 和 σ 分别代表 tan$h$ 和 sigmoid 两种非线性激活函数，W_z、W_h、W_r、U_z、U_h、U_r 代表 GRU 神经元对应的权重参数，b_z、b_h、b_r 代表了 GRU 神经元对应的偏差参数，⊙代表了元素乘法（Hadamard product）。

如果使用 $\overrightarrow{GRU}$ 代表对句子 s_i 正向进行 GRU 神经元编码，$\overleftarrow{GRU}$ 表示反向的 GRU 神经元编码，词编码层的结构可用下面的公式表示：

$$\vec{h}_{it} = \overrightarrow{GRU}(x_{it}), t \in [1,T] \quad (6)$$

$$\overleftarrow{h}_{it} = \overleftarrow{GRU}(x_{it}), t \in [T,1] \quad (7)$$

其中，$\vec{h}_{it}$、$\overleftarrow{h}_{it}$ 分别代表单词 w_{it} 的正向隐藏状态和反向隐藏状态，两者应进行合并，使用 $h_{it} = [\vec{h}_{it}, \overleftarrow{h}_{it}]$，代表单词 w_{it} 的双向隐藏状态。

（4）词注意力层首先对词编码层得到的 w_{it} 的双向隐藏状态 h_{it} 进行基本多层感知器网络操作得到隐层向量 u_{it}，然后引入 softmax 操作得到权重 α_{it}，经过求和操作后并最终可得到句子特征表示 s_i，其公式具体可表示如下：

$$u_{it}=\tan h(W_w h_{it}+b_w) \quad (8)$$

$$\alpha_{it}=\text{softmax}(u_w,u_{it}) \quad (9)$$

$$s_i=\Sigma_t\alpha_{it}h_{it} \quad (10)$$

其中，W_w、u_w、b_w 分别为词注意力层对应的权重参数和偏差参数。

（5）句子编码层使用词编码层类似的方式对句子特征表示 s_i 进行编码，其结构可用下面公式表示：

$$\vec{h}_i=\overrightarrow{GRU}(s_i),i\in[1,L] \quad (11)$$

$$\overleftarrow{h}_i=\overleftarrow{GRU}(s_i),i\in[L,1] \quad (12)$$

同样可使用 $h_i=[\vec{h}_i,\overleftarrow{h}_i]$ 代表句子 s_i 的双向隐藏状态。

（6）句子注意力层与词注意力层结构类似，生成的最终论证区间特征可通过下面公出得出：

$$u_i=\tanh(W_sh_i+b_s) \quad (13)$$

$$\alpha_i=\text{softmax}(u_s,u_i) \quad (14)$$

$$d=\Sigma_i\alpha_ih_i \quad (15)$$

其中，W_s、u_s、b_s 分别为句子注意力层对应的权重参数和偏差参数。

（7）输出层使用论证区间特征 d 进行最后的分类，主要是采用 softmax 函数进行各个类别的概率预测，然后使用交叉熵函数作为损失函数进行训练。对于某个类别 o，其概率 p_o 可表示为：

$$p_o=\text{softmax}(W_od+b_o) \quad (16)$$

对于某个论证区间样本 a，其损失函数 L_a 可以表达为：

$$L_a=-\Sigma_jy_{aj}\log(p_{aj}) \quad (17)$$

其中，W_o 和 b_o 分别为输出层的权重参数和偏差参数，j 代表某个具体的分类，p_{aj} 代表模型预测的样本 a 在类别 j 的概率，而 y_{aj} 代表的样本 a 在类别 j 的真实标签，值为 0 或 1。

3 实验和结果分析

3.1 实验环境

本文中所有的实验均在如表 1 所示的实验环境中完成。

表 1 实验环境及配置

实验环境	环境配置
操作系统	Ubuntu16.04
GPU	NVIDIA GeForce GTX 2080 Ti
内存	32G
编程语言	Python3.6
深度学习框架	TensorFlow1.14

3.2 数据集

本文实验数据主要来源于 PubMed 生物医学数据库的期刊论文。笔者调研了多种论证区间分类体系，结合生物医学的文献论证特点并综合其他学者的研究结论，最终提出并使用了引言、研究背景、研究方法、实验结果、实验设计、病例介绍、结论、相关研究和研究不足的 9 分类体系。根据上述确定的分类体系，笔者通过关键词自动识别和人工标注校准相结合的方式进行数据类别的确定。由于样本数据分布不均匀，笔者使用了欠采样（Under-Sampling）的方法以保证数据均衡，即减少数量较多类别的样本数进行抽样，最终确定了 305635 条数据，每条数据均由一段论证区间文本以及对应的类别标签组成，其中测试集由每个类别随机抽取约 10%~15% 的数据产生，剩余的数据则作为训练集，最终训练集样本数量为 261635，测试集样本数量为 44000。

doi:10.3772/j.issn.2095-915x.2020.03.005

3.3 评价指标

本文采用广泛用于信息检索和统计学领域的分类评价体系——准确率（Precision，简称 P）、召回率（Recall，简称 R）和调和平均值（F-score，简称 F 值），各指标的计算公式如下：

准确率 P = 正确识别的论证区间数 / 识别出的论证区间数 （18）

召回率 R = 正确识别的论证区间数 / 实际论证区间数 （19）

调和平均值 F 可表达为：

$$F = \frac{(\alpha^2 + 1)P * R}{\alpha^2(P + R)} \quad （20）$$

本次实验的 α 参数取值为 1，即调和平均值 F 使用常用的 F1 值。因此，本次实验采用整体准确率、召回率和 F1 值，作为衡量各个模型的评价指标。

3.4 实验结果及分析

本文分别采用两种不同的深度学习模型 HAN、LSTM，在 Google 开源的 TensorFlow 框架进行论证区间识别的实验，并选取在文本分类任务中多次取得优异表现的 SVM 算法作为基准模型。

笔者首先对数据集进行基本统计，得出论证区间文本的平均长度为 488.88，区间包含平均单句数量为 18.61，而单句的平均长度为 26.28。基于此调整模型参数设置，HAN 模型在输入层将文本序列统一处理成 30*30 的矩阵，而 LSTM 模型则将文本序列最大长度设为 500 进行训练，两种模型的学习率均为 0.001，使用 Adam 算法更新模型参数。对于 SVM 模型，本文利用的 Python 常见的机器学习库 Scikit-Learn 进行训练。三种模型的具体实验参数设置如表 2 所示：

表 2 实验参数设置

HAN	LSTM	SVM
学习率0.001 词嵌入维度200 批尺寸64 隐层GRU节点数50 输入文本30*30 epoch次数25	学习率0.001 词嵌入维度200 批尺寸128 隐层LSTM节点数128 输入文本 500 丢弃率0.2 epoch次数50	输入特征 TF-IDF 惩罚系数 1.0 核函数 RBF

上述三种模型在训练集上进行训练后，分别在测试集上进行论证区间的分类测试实验。笔者统计了三种模型在测试集上不同类别上的准确率、召回率和 F1 值，测试实验结果如表 3 所示。

从表 3 可以看出，HAN 识别效果最好，整体的准确率、召回率和 F1 值均为最高，达到了 0.90。其次是 LSTM 和 SVM，两种深度学习模型均比传统的 SVM 模型表现要好。对于论证区间类别的识别，表现最好的类别为“病例介绍”和“研究方法”，其 F1 值均超过了 0.95，“实验设计”和“参考文献”两种类别的 F1 值也均超过了 0.9，F1 值最低的两个类别为“引言”和“研究背景”，分别为 0.74

和0.80。具体来看，“引言”的召回率仅为0.65为所有类别最低，“研究背景”的准确率为0.74为所有类别最低，导致两者的识别效果均不太理想。

表 **3** 论证区间测试实验结果

	HAN			LSTM			SVM		
	P	R	F1	P	R	F1	P	R	F1
引言	0.85	0.65	0.74	0.80	0.63	0.70	0.69	0.59	0.64
研究背景	0.74	0.88	0.80	0.70	0.88	0.78	0.67	0.88	0.76
研究方法	0.93	0.98	0.96	0.90	0.98	0.94	0.74	0.98	0.84
实验设计	0.99	0.90	0.94	0.97	0.84	0.90	0.99	0.52	0.68
实验结果	0.86	0.93	0.89	0.85	0.93	0.89	0.82	0.89	0.85
病例介绍	0.98	0.98	0.98	0.99	0.98	0.98	0.97	0.94	0.95
结论	0.87	0.89	0.88	0.87	0.88	0.87	0.85	0.84	0.84
相关研究	0.95	0.89	0.92	0.96	0.79	0.87	0.99	0.60	0.75
研究不足	0.89	0.84	0.87	0.89	0.79	0.84	0.92	0.64	0.76
整体	**0.90**	**0.90**	**0.90**	0.88	0.86	0.87	0.83	0.80	0.81

本文采用的基于层次注意力机制的HAN模型，对比基准实验SVM算法，整体F1值提高了9个百分点，而对比深度学习文本分类的另一代表算法LSTM，整体F1值也提高了3个百分点。关于所采用模型在该数据集上取得效果全优的原因，笔者分析总结为以下几点：

（1）数据集噪声较小。本次实验所用数据为生物医学领域的期刊论文，该学科的不同类型论证区间的区分度明显，人工标注数据集一致性较好。另一方面，所得的期刊全文数据为XML格式，通过预处理后可获得数据质量较高的数据集，这为本文实验结果的可靠性奠定了较好的数据基础。

（2）特征提取策略符合论证区间文本的结构。论证区间文本本身具有一定的层次结构，因此从句子和单词两种维度对文本提取特征进行层次建模的效果优于纯单词特征的单一维度的模型。

（3）HAN模型具有更好的特征表示能力。SVM算法采用各种统计信息作为文本表示，其本身忽略了文本内部丰富的语义信息，面对学术文本这种相似度较高的语料，难以提取到有效的支持向量。而HAN模型使用多种隐藏层自动提取并组合内在语义特征，所得到的这种更加优化的特征表示也保证了HAN模型的效果下限。

（4）HAN模型解决了长距离依赖问题。改良的LSTM模型通过遗忘门、输入门和输出门来控制序列信息传递过程中的保留与更新，尽量避免长距离依赖问题所带来的影响，但并不能根本解决。而HAN模型所基于的自注意力机制并不考虑序列的位置信息，直接对文本序列中任意两个单词之间的隐藏关系进行计算，等价于其将任意两个单词之间的距离看作为1，

doi:10.3772/j.issn.2095-915x.2020.03.005

因此不存在循环神经网络中的长距离依赖问题。

3.5 错误分析

为了进一步探究本文所提出方法的实验结果，笔者针对其实验结果进行了错分统计，结果如表 4 所示。其中行标签代表预测的论证区间类别名称，列标签代表真实的论证区间类别名称，具体的每一行代表每种论证区间的数据被划分各个类别的比例，每一列代表各种类别的论证区间数据被划分为该类别的比例。

表 4 HAN 模型实验结果错分比例表

	引言	研究背景	研究方法	实验设计	实验结果	病例介绍	结论	相关研究	研究不足	合计
引言	65.03%	25.80%	2.25%	0.18%	1.80%	0.43%	0.15%	2.05%	2.33%	100.00%
研究背景	7.63%	87.63%	1.28%	0.28%	1.30%	0.05%	0.03%	0.88%	0.95%	100.00%
研究方法	0.13%	0.16%	98.23%	1.13%	0.17%	0.04%	0.08%	0.03%	0.03%	100.00%
实验设计	0.10%	0.28%	2.93%	92.98%	0.88%	0.38%	0.13%	0.63%	1.73%	100.00%
实验结果	1.18%	1.55%	2.58%	1.58%	89.38%	0.13%	0.00%	0.45%	3.18%	100.00%
病例介绍	0.13%	0.15%	0.35%	0.55%	0.08%	98.43%	0.18%	0.05%	0.10%	100.00%
结论	0.20%	0.10%	7.43%	1.00%	0.08%	0.30%	90.05%	0.28%	0.58%	100.00%
相关研究	1.38%	1.10%	3.88%	2.53%	0.60%	0.15%	0.13%	89.20%	1.05%	100.00%
研究不足	0.40%	0.70%	1.35%	5.65%	7.62%	0.05%	0.03%	0.33%	83.89%	100.00%
合计	76.15%	117.46%	120.26%	105.86%	101.89%	99.94%	90.76%	93.88%	93.80%	900.00%

从表 4 可以看出，“引言”被错分为“研究背景”的比例最高为 25.80%。其次是“研究背景”错分为“引言”的占 7.63%，究其原因，生物医学的相关文献中，引言和研究背景并不总是分开的，引言部分经常会被融合到研究背景当中，确实两者也均是用来介绍当前论文的基本信息，相似度本身也较高。“结论”类的数据有 7.43% 的比例被错分为“研究方法”类，本文认为产生这种错分情况的产生是由于研究人员在进行研究结论撰写时，经常会同时介绍对应的研究方法所导致的。“研究不足”类的识别准确率为 83.89%，其中 7.62% 的数据被预测为“实验结果”类，5.65% 的数据被预测为“实验设计”类。本文认为该结果归结于两个原因：一方面，很多研究人员会在撰写实验结果时进行对比分析，得到现有结果的不足和改进的地方，与研究不足类别数据相似度较高；另一方面，并不是所有的学术文献都会专门去写本研究的不足，导致数据集中所含数量较少，数据分布的不均衡也是其准确率较低的原因之一。

根据上述分析结果，本文认为关于论证区间分类识别的研究可从以下两个方面可以进行改进：第一，增加数据集的数量并平衡不同论证区间类别文本的数量，减少数据量小以及数据倾斜所带来的影响。第二，对于论证区间类别体系进行调整和细化，如将“引言”和“研究背景”合并成一种类别去进行识别。

4 结语

本文将深度学习理论和技术引入到学术文本的论证区间识别研究中，针对生物医学领域

提出了论证区间 9 分类体系，并使用基于层次注意力机制的论证区间识别模型在 PubMed 生物医学数据库数据集上进行了实验探索。实验证明，与常见的 LSTM 和 SVM 两种文本分类算法进行比较，本文所采用的 HAN 模型在各个类别的论证区间识别效果上均为最优，F1 值达到了 0.90。整体来看，本文提出的基于层次注意力机制的论证区间识别模型准确率较高，在论证区间识别问题上取得了很好的效果。

本研究在取得了有益突破的同时也存在较大的可提升空间，未来可从以下两个方面进行更深层次的探索。一是增加语料规模，平衡数据分布，以增强所提出模型的识别效果和泛化能力；二是对于论证区间分类体系的研究，既针对生物医学领域的论证区间体系进行细化和改进，也包括对不同学科领域的论证区间分类的研究。

参考文献

[1] Rahman M M, Finin T. Deep Understanding of a Document’s Structure[C]. Proceedings of the Fourth IEEE/ACM International Conference on Big Data Computing, Applications and Technologies. ACM, 2017: 63-73.

[2] Alzahrani S, Palade V, Salim N, et al. Using structural information and citation evidence to detect significant plagiarism cases in scientific publications[J]. Journal of the American Society for Information Science and Technology, 2012, 63(2): 286-312.

[3] Agarwal S, Yu H. Automatically classifying sentences in full-text biomedical articles into Introduction, Methods, Results and Discussion[J]. Bioinformatics, 2009, 25(23): 3174-3180.

[4] Teufel S. Argumentative zoning: Information extraction from scientific text[D]. University of Edinburgh, 1999.

[5] Yepes A J, Mork J, Aronson A. Using the argumentative structure of scientific literature to improve information access[C]. Proceedings of the 2013 Workshop on Biomedical Natural Language Processing. 2013: 102-110.

[6] 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):985.

[7] Van Eemeren F H, Grootendorst R, Johnson R H, et al. Fundamentals of argumentation theory: A handbook of historical backgrounds and contemporary developments[M]. Routledge, 2013.

[8] Mizuta Y, Korhonen A, Mullen T, et al. Zone analysis in biology articles as a basis for information extraction[J]. International Journal of Medical Informatics, 2006, 75(6):468-487.

[9] 方龙 , 李信 , 黄永 , 陆伟 . 学术文本的结构功能识别——在关键词自动抽取中的应用 [J]. 情报学报 , 2017, 36(6):599-605.

[10] Tbahriti I, Chichester C, Lisacek F, et al. Using argumentation to retrieve articles with similar citations: an inquiry into improving related articles search in the MEDLINE digital library[J]. International Journal of Medical Informatics, 2006, 75(6):488-495.

[11] 黄永 , 陆伟 , 程齐凯 , 桂思思 . 学术文本的结构功能识别——在学术搜索中的应用 [J]. 情报学报 , 2016, 35(4):425-431.

[12] Teufel S, Moens M . Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status[J]. Computational Linguistics, 2002, 28(4):409-445.

[13] LeCun Y, Bengio Y, Hinton G. Deep learning[J]. Nature, 2015, 521(7553): 436.

[14] Teufel S, Siddharthan A, Batchelor C. Towards discipline-independent argumentative zoning: evidence from chemistry and computational linguistics[C]. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3. Association for Computational Linguistics, 2009: 1493-1502.

[15] Liakata M, Saha S, Dobnik S, et al. Automatic

doi:10.3772/j.issn.2095-915x.2020.03.005

recognition of conceptualization zones in scientific articles and two life science applications[J]. Bioinformatics, 2012, 28(7): 991-1000.

[16] Liu H. Automatic Argumentative-Zoning Using Word2vec[J]. arXiv preprint arXiv:1703.10152, 2017.

[17] 黄永 , 陆伟 , 程齐凯 , 桂思思 . 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.

[18] 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]. 情报学报 , 2016, 35(3): 293-300.

[19] 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.

[20] 王佳敏 , 陆伟 , 刘家伟 , 程齐凯 . 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.

[21] 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.

[22] Liu P, Qiu X, Huang X. Recurrent neural network for text classification with multi-task learning[J]. arXiv preprint arXiv:1605.05101, 2016.

[23] Zhou X, Wan X, Xiao J. Attention-based LSTM network for cross-lingual sentiment classification[C]. Proceedings of the 2016 conference on empirical methods in natural language processing. 2016: 247-256.

[24] Yang Z, Yang D, Dyer C, et al. Hierarchical attention networks for document classification[C]. Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies. 2016: 1480-1489.

[25] Palau R M, Moens M F. Argumentation mining: the detection, classification and structure of arguments in text[C]. Proceedings of the 12th international conference on artificial intelligence and law. ACM, 2009: 98-107.

[26] Cho K, Van Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. arXiv preprint arXiv:1406.1078, 2014.