基于多阶段分类的科研项目申请书结构功能识别*

林鑫　杜莹　罗宇
（华中师范大学信息管理学院，武汉 430079）

摘要：科研项目申请书蕴含丰富的科学知识，被广泛用作科技情报分析的基础数据，其中重复检测、分析挖掘等智能处理工作需要在明晰申请书结构功能的前提下展开。因此，构建一种基于多阶段分类的科研项目申请书结构功能识别模型。首先，对申请书进行预处理，识别申请书的正文内容及其包含的多模态要素，并将文本段落规范化；之后，基于BiLSTM-Attention模型，依次区分申请书中的章节标题与正文文本，基于标题识别正文文本的一级功能，进而识别申请书的细粒度结构功能。实验结果显示，所提方法的准确率与召回率分别达到93.7%和93.1%，该方法能较好支撑科研项目申请书的结构化解析，也能为其他类型学术文本的结构功能识别提供参考。

关键词：科研项目申请书；结构功能识别；多阶段分类；BiLSTM-Attention
中图分类号：G254　DOI：10.3772/j.issn.1673-2286.2024.03.003
引文格式：林鑫，杜莹，罗宇. 基于多阶段分类的科研项目申请书结构功能识别[J]. 数字图书馆论坛，2024，20（3）：25-33.

课题制管理模式下，科研项目申请书具有非常重要的价值，不仅是科研项目立项评审的核心依据，还体现了申请人对相关领域的认知，蕴含着丰富的科学知识，可以作为科技情报分析的重要基础数据。已有学者以科研项目申请书为基础数据进行了领域热点挖掘[1-2]、前沿主题探测[3-4]等研究。尽管有填写说明作指导，但实际提交的科研项目申请书的结构化程度仍然有限。一方面，项目申请书的正文主体部分大多以非结构化形态存在，篇幅较长、内容丰富，但只能线性阅读；另一方面，尽管申请书模板提出了填写建议，但实际撰写过程仍缺乏统一规范，学者会按照个人习惯自行拟定章节或小节标题，无论是标题用词还是内容设置都存在一定的个性化特点。这既不便于项目评审专家的高效阅读，也给重复检测、分析挖掘等智能处理带来一定的阻碍。为此，本文拟借鉴图情学科学术论文结构化的研究成果，构建科研项目申请书结构功能体系，并提出自动化结构功能识别模型，为科研项目申请书的数据化提供支持。

1　相关研究

学术文本结构功能识别是指识别各个句子、段落或章节在学术文本内容层次的功能作用[5]，该主题是近年来图情学科关注的重点领域，国内外学者从多个方面进行了研究探索。从对象角度来说，全文与摘要都是结构功能识别研究的重要对象，如赵旸等[6]围绕基金项目摘要的结构功能识别进行了研究，秦成磊等[7]、Ma等[8]、Asadi等[9]则关注学术论文全文的结构功能识别。从粒度角度看，段落与句子是两个重点关注的

收稿日期：2023-12-21
*本研究得到国家社会科学基金项目“面向多模态发布的学术论文语义标注与对象链接研究”（编号：23BTQ083）资助。

粒度。前者的典型研究是陆伟教授团队开展的系列研究，其将学术论文的各个段落分为引言、相关研究、方法、实验、结论5类[5,10-12]；后者的研究既包括各摘要句的结构功能识别[13]，也包括学术论文全文的句子粒度结构功能识别[14]。从结构功能识别目标看，除了全类型识别研究外，还包括针对具体功能单元的识别研究，包括创新句[15]、研究方法句[16]、问题句[17]、贡献句[18]、评价句[19]。从技术方法应用方面看，机器学习是最为流行的技术思路，早期以传统机器学习方法为主，包括条件随机场[20]、支持向量机[21]等，近年来以深度学习方法为主，包括CNN（Convolutional Neural Networks）、DNN（Deep Neural Networks）、LSTM（Long Short Term Memory）、BiLSTM、Attention等[22-26]。还有研究尝试将自然语言处理领域的预训练机制引入模型，如欧石燕等[27]提出了将深度学习预训练模型BERT与传统机器学习分类算法深度森林相结合的句子粒度结构功能识别混合模型。从技术模型构建目标看，除了普遍关注结构功能识别的准确率、召回率等效果指标提升外，也有研究开始关注不充分资源条件下的识别模型构建问题，以降低模型训练对人工标注语料的依赖程度，如陈果等[28]构建了基于主动学习的论文摘要结构功能识别模型，刘江峰等[29]则尝试利用数据增强技术进行模型构建。

总体来说，国内外围绕学术文本的结构功能识别，提出了多种技术模型，而且部分模型的效果良好。但是，相关研究主要以学术论文为对象，以科研项目申请书为对象的仅有1篇文献，且只考虑了申请书摘要的结构功能识别，但无论是学术论文还是申请书的摘要，其结构功能体系均与科研项目申请书正文差异较大。此外，既有研究主要针对简单结构功能体系，类目数量多为个位数，而科研项目申请书正文结构功能体系类目较多且结构较为复杂，直接照搬既有模型容易造成识别效果大幅下降，因此开展面向科研项目申请书的结构功能识别研究十分必要。本文拟选择多种类型的科研项目的申请书进行调研，构建申请书正文结构功能体系，进而结合申请书及结构功能体系的特点进行自动识别模型构建，以取得良好的识别效果。

2　科研项目申请书结构功能体系构建

鉴于各类科研项目申请书正文填写的要求、科研人员写作习惯不尽相同，为构建通用性强、全面系统的结构功能框架，采用调研归纳的方法进行申请书正文结构功能体系框架的构建。针对立项规模较大的国家自然科学基金面上项目与青年项目、国家社会科学基金一般项目与青年项目、教育部人文社会科学研究一般项目、中国博士后科学基金等四大基金项目，采用结合项目申请书的填写指南和科研人员的实际撰写情况的结构功能单元归纳方法，以确保所构建的结构功能框架能够适用于各主要项目类型的申请书。为确保调研的充分性，采用饱和抽样原则，先对5篇申请书进行分析，以获得初步的结构功能框架，之后以5篇申请书为单元，当连续5篇申请书中均未出现新增的结构功能类型时，停止对此类申请书的调研。实践发现，尽管科研人员制定的章节标题名称及写作顺序有所差异，甚至个别申请书未全部涵盖填写指南中的建议内容，但并未出现新增结构功能类型的情况，因此对每类申请书都只选用10篇作为调研对象。鉴于国家社会科学基金重大项目、各省级社会科学基金与自然科学基金项目、国家重点研发计划等类型项目申请书的获取难度较大，并参照前述四大国家级基金项目申请书的调研情况，以申请书模板为对象可以较全面地识别其结构功能体系，因此调研这些类型项目的申请书模板。根据网络搜寻结果，找到25类项目的申请书模板，涵盖了国家社会科学基金、国家自然科学基金除青年项目、一般（面上）项目之外的其他项目类型，以及国家重点研发计划、多个省级自然科学和社会科学基金项目，调研对象具备了较强的代表性。

在完成申请书样本及模板分析基础上，将内涵相近的结构功能类型进行合并，如将国家自然科学基金项目申请书中的拟解决的关键问题与国家社会科学基金项目申请书中的重点难点合并。部分申请书将研究内容作为与研究对象、研究目标等并列的二级结构功能，此处将其视为一级结构功能，以研究框架为二级结构功能的名称，最终形成了包括10个一级结构功能、18个二级结构功能的科研项目申请书结构功能体系框架（见表1）。

3　基于多阶段分类的科研项目申请书结构功能识别模型

为支撑结构功能识别模型设计，以40篇科研项目申请书为样本进行了内容特点分析，结果显示：①无论是一级还是二级结构功能，属于同一结构功能的内容均集中分布在一起；②科研项目申请书中同一功能的内容由章节标题与正文文本、图像、表格、公式5类要素构成，从样本数据看，所有结构功能单元都包含章节标题

表1　科研项目申请书结构功能体系框架

一级结构功能	二级结构功能
立项依据	研究背景、研究意义、研究现状、立项依据-其他
研究内容	研究对象、研究目标、研究框架、重点难点、研究内容-其他
研究方案	研究思路、研究方法、技术路线、可行性分析、研究方案-其他
研究基础	项目成员简历、项目成员相关观点、项目成员相关科研成果、研究基础-其他
创新之处	无
经费预算	无
研究计划	无
预期成果	无
工作条件	无
参考文献	无

与正文文本，但只有部分包含图像、表格、公式；③样本数据中，每一处的图像、表格、公式均有对应的正文文本内容，而且每一个正文文本段落均归属于唯一的结构功能单元；④与正文文本段落、图像、表格、公式相比，不同功能单元的章节标题差异更为显著。根据上述特点，构建基于多阶段分类的科研项目申请书结构功能识别模型（见图1），以章节标题为科研项目申请书结构功能识别的主要基础数据。在章节标题欠缺时以正文文本段落为补充基础数据，并将其转换为分类问题进行解决；同时，为避免类目过多影响识别效果，拟采用多阶段法进行结构功能识别，先识别一级结构功能（包含10个结构功能），之后对二级结构功能（共18个二级结构功能，涉及4个一级结构功能）进行细粒度识别，从而将单次分类的类目数量控制在较小范围内。

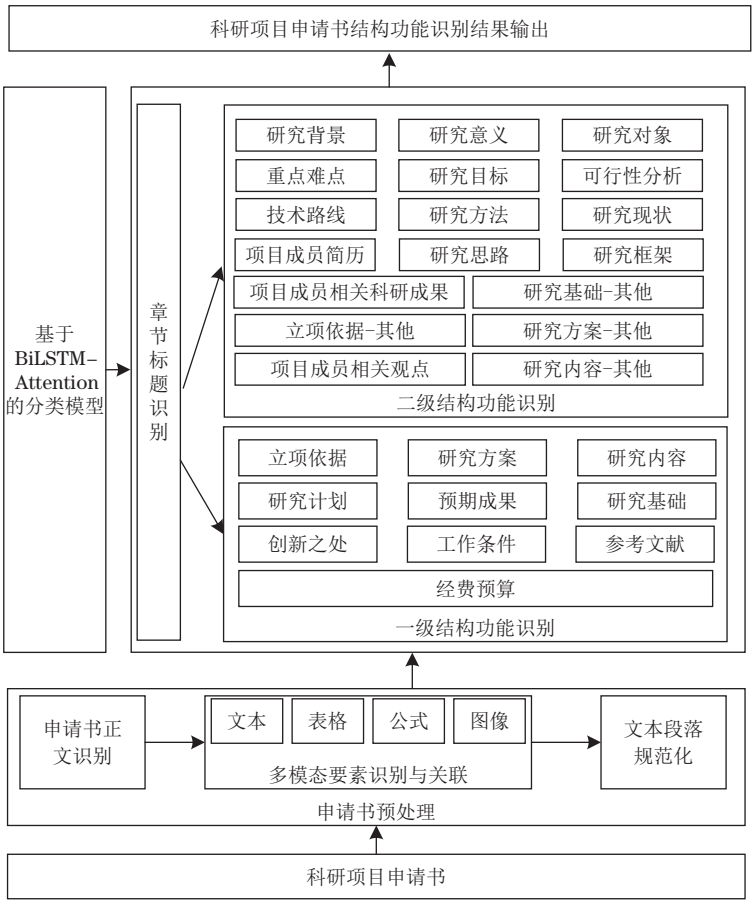

图1　基于多阶段分类的科研项目申请书结构功能识别模型

3.1 数据预处理

由于申请书结构功能识别只关注课题论证相关的正文部分，首先需要将此部分内容识别出来。结构功能识别中，图像、表格、公式等模态对象直接关联其对应文本所属的结构功能即可，因此需要对正文进行多模态要素识别与关联。此外，申请书中的章节标题常常由序号、标题内容和说明性文字（用以指导申请书内容的撰写）组成，而序号和说明性文字在结构功能识别中可能造成干扰，为此需要对申请书中的文本内容进行规范化处理。上述步骤也就构成了申请书预处理的主要内容。

（1）申请书正文识别。尽管各类科研项目或同一科研项目不同年度的模板有所区别，但同年度、同类型科研项目的申请书具有相同的模板，因此能够较容易实现正文与填写说明、基本信息表、签字盖章页等内容的区分。在实现过程中，可以对模板进行逐类分析，形成与其相适应的识别规则，将申请书正文部分精准识别出来。

（2）多模态要素识别与关联。多模态要素识别中，需要结合申请书的文档类型进行识别方法选择。如果是doc、docx等格式文档，可以根据文档中的语义标签信息进行多模态要素识别；如果是纸质文档扫描件等，为取得理想的效果，则需要采用机器视觉技术进行文档版式分割及模态类型识别。鉴于文本段落中常常有标识所关联的图像、公式或表格的特征词，可以采用基于规则的方法发现与图像、公式、表格相关联的文本段落。

（3）文本段落规范化。总体来说，申请书正文中各级章节标题的序号、说明性文字一般具有较为明显的特征，如序号一般为阿拉伯数字或表征数字的汉字，可采用规则法进行识别与过滤。为提升规则学习的效率，可以按如下方法半自动化地构建规则集合：①根据经验制定少量种子规则；②随机抽取规模较大的样本集合，按照已制定规则进行处理，剔除处理后仅包含汉字和英文字母的文本段落；③抽取少量包含阿拉伯数字或特殊字符、空格的文本段落，补充完善规则集合；④循环进行样本数据的处理和规则集合完善，直至样本数据集为空。

3.2 面向结构功能分析的章节标题识别

对于申请书中的文本内容，只有实现了章节标题与正文文本段落的区分，才能为章节标题优先的结构功能识别奠定数据基础。需要说明的是，此处的章节标题识别并非单纯将申请书中的全部章节标题识别出来，而是将表征申请书结构功能起始位置的章节标题识别出来，如某申请书的两个章节标题分别为“2.2 总体框架”“（1）社会网络用户认知与知识标注的关联关系”，后者尽管也属于章节标题，但可能对结构功能识别造成干扰，因此不应将其识别出来。鉴于章节标题的长度较短，科研项目申请书正文中的章节标题识别可以分两步进行：首先根据文本段落的长度，筛选出疑似章节标题的文本段落；其次，将章节标题识别视为文本段落分类任务，将剩余文本段落区分为章节标题与非标题。实施过程中，前一环节应避免将章节标题误判为非标题，因此应适当设置阈值；后一环节拟融合自然语言处理中的词向量技术与深度学习技术进行分类模型的设计。

3.3 科研项目申请书一级与二级结构功能识别

鉴于同一段落或章节的内容都属于同一个结构功能，可以将结构功能识别问题视为以章节标题或文本段落为输入的分类问题。分类器设计上，可以采用与章节标题识别相同的技术模型，下面先对一级与二级结构功能识别的思路进行阐述。

（1）一级结构功能识别。科研项目申请书中鲜有缺少一级结构功能单元标题的情况，因此可以仅采用章节标题作为结构功能识别的基础数据。同时，鉴于章节标题识别中可能存在误识别的情形，在类目体系上，除了立项依据、研究内容、研究方案、研究基础、创新之处、经费预算、研究计划、预期成果、工作条件、参考文献外，还应增加“其他”类目，用于容纳误识别的章节标题。处理流程上，在利用分类器完成各章节标题所属结构功能识别后，还可以根据同一结构功能单元内容集中分布的特征对识别结果进行校正：若两个非连续标题属于同一个结构功能单元，则两者间的其他章节标题也都属于该结构功能；若某一章节标题分类为“其他”，且前后两个章节标题属于不同的结构功能，则赋予其前一个章节的结构功能类型。

（2）二级结构功能识别。在完成一级结构功能识别基础上，需要对立项依据、研究内容、研究方案、研究基础4个一级结构功能的内容进行细粒度结构功能识别。受个人习惯的影响，部分申请书没有为二级结构功能专门设置章节标题，因此需要采用章节标题与正文段落相结合的方

式进行识别。首先以章节标题为基础数据进行识别，如果识别出的二级结构功能类型不少于2个，则不再以正文段落为基础数据进行识别；否则，以段落为单位，对每一个正文文本段落进行二级结构功能识别。需要说明的是，以章节标题为输入进行处理时，应采用与一级结构功能识别相同的处理方式，以取得更好的识别效果。

3.4　基于BiLSTM-Attention的分类器设计

基于深度学习的分类方法以其强大的数据拟合能力，能够在训练数据较为充足的前提下取得更好的分类效果，因此，越来越多的研究与实践采用基于深度学习的模型解决文本分类问题。BiLSTM作为一种深度学习方法，可以学习文本语义序列的双向特征，较好地捕捉文本序列信息，并预测序列的下一个状态，因此被广泛应用于文本分类任务。然而，尽管其可以较好地捕获文本的全局结构信息，但对关键模式信息不敏感，进而影响分类效果。而Attention机制能够捕获文本中的重要词汇，并赋予较高的权重，帮助机器学习算法抓住文本中的重点，但这一技术本身忽略了词汇的序列信息，导致文本的全文结构信息无法得到较充分的利用。因此，融合BiLSTM与Attention机制两种技术方法，构建BiLSTM-Attention网络模型（见图2），分为输入层、BiLSTM层、Attention层、输出层，以改进自动分类的效果。

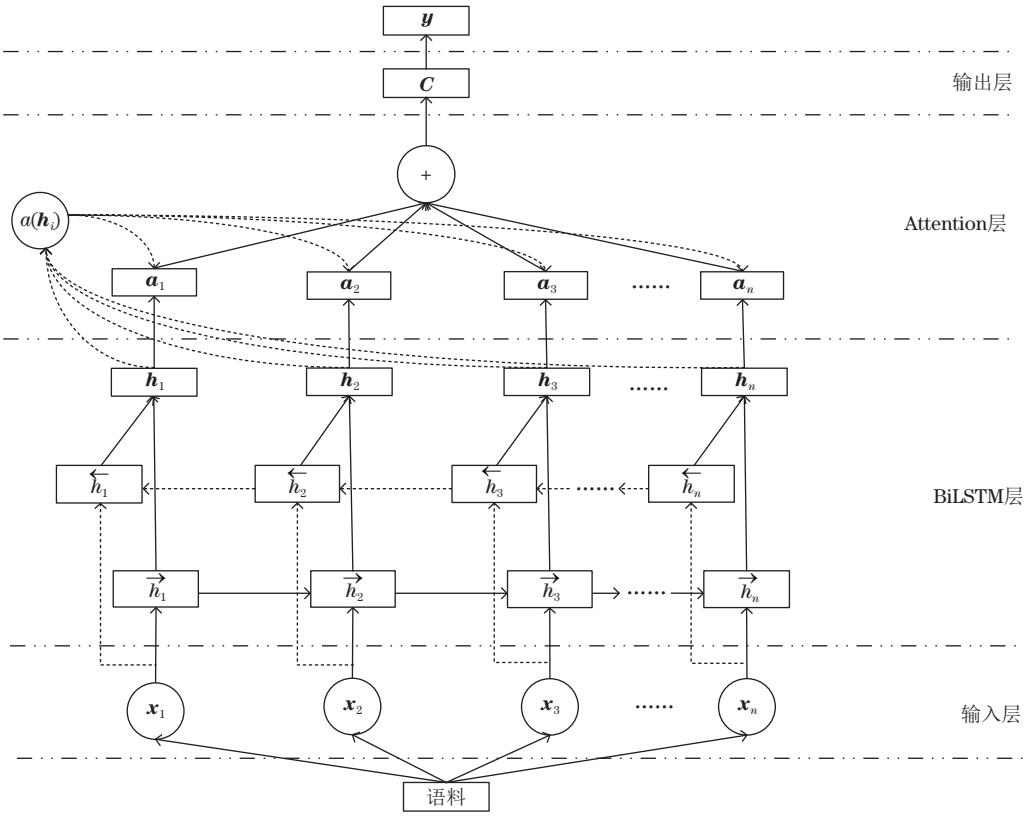

图2　BiLSTM-Attention网络模型结构

（1）输入层。由于申请书的文本序列无法直接被算法识别并进行运算，需要将其转化为向量的形式参与模型运算。借鉴国内外研究与实践，采用基于词向量的分布式表示方法，将每个词映射到一个较短的向量空间中，以实现文本的语义表示，并缓解传统词向量维度灾难和数据稀疏的问题[30]。Word2Vec是Google公司于2013年推出的包含词向量训练和计算等功能的高效工具，可以支持在百万数量级的词典和上亿的数据集上的高效训练，自开源以来，得到了工业界和学术界的广泛应用。因此选择Word2Vec作为工具训练词向量，对输入层传入的文本序列进行词向量表示，并将文本的词向量序列作为此层的输出，输入BiLSTM层。

（2）BiLSTM层。BiLSTM模型是在LSTM模型的基础上进行改良形成的，其输出由前向LSTM和后向LSTM共同决定。对于申请书正文的特征提取而言，除

了正向序列会影响文本的语义分析进而影响分类结果外，文本的后向序列也会对文本分类产生影响。因此采用BiLSTM模型同时捕获上下文的语义特征，以此为依据进行申请书正文结构功能分类。

（3）Attention层。申请书章节标题或正文中不同词汇在分类中的重要性存在明显不同，需要进行差异化赋权，为此引入词级的Attention机制，以提高分类结果的准确性。具体过程为：保留BiLSTM层对输入文本序列处理后的中间结果输出，训练Attention层以对来自BiLSTM层的输出结果进行选择性学习，并在Attention层输出时，将输出序列与BiLSTM层的中间输出结果进行关联，以突出具有重要作用的信息。计算过程如式（1）~（3）所示。

$$\boldsymbol{e}_t = a\left(\boldsymbol{h}_t\right) \tag{1}$$

$$\boldsymbol{a}_t = \frac{\exp(\boldsymbol{e}_t)}{\sum_{i=1}^{T}\exp(\boldsymbol{e}_i)} \tag{2}$$

$$\boldsymbol{C} = \sum_{t=1}^{T}\boldsymbol{a}_t\boldsymbol{h}_t \tag{3}$$

式中：t和i为某一时刻，T为最终时刻；$\boldsymbol{h}_t$为t时刻BiLSTM层的输出；a为计算$\boldsymbol{h}_t$梯度重要性的函数；$\boldsymbol{a}_t$为$\boldsymbol{h}_t$的权重分布；$\boldsymbol{C}$为经过Attention层处理后的文本序列向量。

（4）输出层。其任务是将经过Attention层处理后的文本序列向量映射到样本标注空间。由于章节标题识别属于二分类任务，采用sigmoid函数进行分类预测。一级及二级结构功能识别属于多分类任务，因此采用softmax函数对文本序列在各个分类上的概率进行预测，同时使用交叉熵作为损失函数，反向传播机制对模型中的参数进行更新，最后输出预测分类结果。

4 实验

为验证科研项目申请书结构功能识别模型的效果，自建申请书数据集进行实验验证，实验设置及效果说明如下。

4.1 数据集构建

借助百度搜索、小木虫社区、经管之家社区等渠道获取了300篇科研项目申请书的全文，涵盖了国家自然科学基金、国家社会科学基金、教育部人文社会科学研究等3类基金项目，包括45 002个正文文本段落，其中有7 745个结构功能标题类段落、37 257个非结构功能标题类段落。按前文所述的结构功能体系，人工对数据进行标注，其中所有的一级结构功能单元和二级结构功能单元均有标题。按照9∶1的比例，将样本申请书区分为训练集和测试集。

4.2 对照实验设置

为便于通过比较衡量模型的效果，研究中设置了4个对照实验，下面分别说明各模型与前文所提模型的区别：①基于LSTM的多阶段结构功能识别模型，区别之处在于采用了LSTM模型作为分类器，用以检验所选取分类器模型是否具有优越性；②基于TextCNN的多阶段结构功能识别模型，设置该对照实验的目的同前，区别之处在于采用了TextCNN模型作为分类器；③基于单阶段的结构功能识别模型，区别之处在于不采用多阶段的分类方法，直接进行二级结构功能识别，用以检验多阶段分类思路是否具有优越性；④基于文本段落的结构功能识别模型，区别之处在于以章节标题与正文文本段落为输入数据，用以检验章节标题优先的策略是否具有优越性。

实验过程中，首先以训练数据为基础建立了正文识别规则，以及序号、说明性文字过滤规则，包括去除文本段落中的阿拉伯数字、去除文本段落首个特殊标点符号（例如“、”“.”）及其之前的文本内容、去除文本段落中小括号及小括号中的文本内容，以及去除文本段落中括号之后的内容。在此基础上，以30个字符为阈值，将文本段落分为疑似章节标题与非标题两类，以满足基于BiLSTM-Attention的多阶段结构功能识别模型、基于LSTM的多阶段识别模型，以及基于TextCNN的多阶段识别模型等3个模型的输入要求。

4.3 实验结果及分析

实验效果评价方面，参照自动分类的常用评价指标，采用准确率、召回率以及F1值进行评价。5个实验最终的结构功能识别效果如表2所示。总体来说，以章节标题为基础数据，基于BiLSTM-Attention的多阶段识别模型的准确率、召回率、F1值最高，效果最为理想；在分类器相同的情况下，基于全部文本段落的结构功能识别效果较差，说明正文文本段落中的噪声会对结构功能

判断产生影响；基于章节标题的结构功能识别中，相较于单阶段识别策略，多阶段策略对最终效果的提升作用较为明显，说明减少单次分类时的类目数量有助于改善结构功能识别效果；在多阶段识别策略下，相比采用LSTM、TextCNN的模型，采用BiLSTM-Attention的模型的效果明显提升，说明分类器的选择具有合理性。

以章节标题为基础数据，基于BiLSTM-Attention的多阶段结构功能识别模型对申请书不同结构功能的识别效果如表3所示。从表3可以看出识别效果差异较为显著：功能单元“研究对象”和“经费预算”的F1值最高，达到了100.0%；而功能单元“研究方案-其他”的F1值最低，仅为77.9%。

表2　不同模型申请书结构功能识别效果

单位：%

类别	模型	准确率	召回率	F1值
基于章节标题的结构功能识别	基于BiLSTM-Attention的多阶段模型	93.7	93.1	93.4
	基于LSTM的多阶段模型	68.0	59.0	63.2
	基于TextCNN的多阶段模型	79.8	75.4	77.5
	基于BiLSTM-Attention的单阶段模型	87.1	83.4	85.2
基于文本段落的结构功能识别	基于BiLSTM-Attention的多阶段模型	84.1	90.5	87.2

表3　申请书不同结构功能识别效果

单位：%

序号	结构功能	准确率	召回率	F1值	序号	结构功能	准确率	召回率	F1值
1	立项依据-其他	96.9	93.9	95.4	13	参考文献	99.6	100.0	99.8
2	研究背景	89.0	95.5	92.1	14	工作条件	94.1	67.0	78.2
3	研究现状	100.0	75.0	85.7	15	研究计划	99.2	96.6	97.9
4	研究意义	99.2	96.2	97.7	16	研究方案-其他	94.9	66.1	77.9
5	研究内容-其他	98.5	97.7	98.1	17	研究方法	79.0	96.7	87.0
6	研究对象	100.0	100.0	100.0	18	研究思路	92.2	90.8	91.5
7	研究框架	99.5	85.2	91.8	19	技术路线	94.6	97.9	96.2
8	研究目标	73.5	100.0	84.7	20	可行性分析	98.8	89.7	94.0
9	重点难点	70.8	100.0	82.9	21	研究基础-其他	99.1	100.0	99.6
10	创新之处	100.0	96.3	98.1	22	项目成员简历	97.9	96.1	97.0
11	预期成果	88.3	100.0	93.8	23	项目成员相关观点	95.8	100.0	97.9
12	经费预算	100.0	100.0	100.0	24	项目成员相关科研成果	86.7	97.2	91.7

进一步地，为厘清基于多阶段分类的结构功能识别模型出现偏差的具体原因，对章节标题的识别效果进行统计，其召回率为99.6%、准确率为93.1%、F1值为96.2%。由此可见，未召回和误召回的比例分别为0.4%和6.9%。经过分析可知，未召回的有“与本项目相关的工作积累和已取得的工作成绩”“其他附件清单”等，这是由于其在样本数据中出现的频次不高，模型训练不充分导致其未被识别出来；误召回的有“时延测试包括以下主要内容”“基本解决了关键方法与技术”等，这是由于其包含了标题类文本中的高频词汇。

除了章节标题识别效果无法达到理想状态外，一级结构功能识别中的偏差也对二级结构功能识别效果有一定影响。包含二级结构功能的一级结构功能识别效果如表4所示。经过分析可知，未召回的有“本课题研究的主要内容、研究方法和创新之处”“立项依据与研究内容”等，根据其标题内容可能被判断为包含两个或两个以上的一级结构功能单元，训练时会对模型产生一定的干扰，从而导致识别不准确；误召回的如“研究方法的创新”“技术路径的创新”等，根据其标题内容应被识别为“创新之处”，但其标题内容中包含其他结构功能章节标题的高频词，如“研究方法”“技术路径”，导致识别错误。

表4　包含二级结构功能的一级结构功能识别效果

单位：%

一级结构功能	准确率	召回率	F1值
立项依据	97.7	82.8	89.6
研究内容	87.9	95.3	91.4
研究方案	91.5	85.1	88.2
研究基础	90.6	97.4	93.9

5 结语

识别正文部分的结构功能是科研项目申请书数据化的重要工作，进而能够为科研项目申请书的专家评审、重复检测、分析挖掘提供结构化的数据支持。为实现申请书正文结构功能自动识别，在构建两级申请书结构功能体系基础上，结合申请书的行文特点设计了优先选择章节标题作为基础数据的多阶段实施的结构功能识别模型。实验结果显示，所提出的方法在准确率、召回率与F1值上均具有更佳的表现。就局限性而言，BiLSTM-Attention分类器对易混章节标题的识别效果不够理想，存在一定的误识别情况，需要在以后的研究中加以突破。一方面拟尝试基于大语言模型更好地捕获文本语义，进而提升识别效果；另一方面拟探索融合章节标题与正文内容的识别策略，通过补充更多问题实现效果提升。

参考文献

[1] 时弘易，李华一，胡骏，等. 国家自然科学基金信息科学领域学科热点变化趋势展望[J]. 中国科学基金，2022，36（3）：497-505.

[2] 马赫，关心惠，沈思. 图书情报学项目研究现状与热点：基于“十三五”时期国家社科基金年度与青年项目的分析[J]. 情报科学，2022，40（4）：186-192.

[3] 刘自强，岳丽欣，方曙. 基于主题扩散演化滞后的研究前沿趋势预测方法研究[J]. 情报理论与实践，2023，46（6）：145-154.

[4] 曾文，王海燕，陈峰，等. 科技前沿探测中的科技信息融合方法研究[J]. 情报学报，2023，42（8）：906-914.

[5] 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（9）：979-985.

[6] 赵旸，张智雄，刘欢，等. 基金项目摘要的语步识别系统设计与实现[J]. 情报理论与实践，2022，45（8）：162-168.

[7] 秦成磊，章成志. 基于层次注意力网络模型的学术文本结构功能识别[J]. 数据分析与知识发现，2020，4（11）：26-42.

[8] MA B W，ZHANG C Z，WANG Y Z，et al. Enhancing identification of structure function of academic articles using contextual information[J]. Scientometrics，2022，127（2）：885-925.

[9] ASADI N，BADIE K，MAHMOUDI M T. Automatic zone identification in scientific papers via fusion techniques[J]. Scientometrics，2019，119（2）：845-862.

[10] 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.

[11] 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.

[12] 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.

[13] YU G H，ZHANG Z X，LIU H，et al. Masked sentence model based on BERT for move recognition in medical scientific abstracts[J]. Journal of Data and Information Science，2019，4（4）：42-55.

[14] 黄文彬，王越千，步一，等. 学术论文子句语义类型自动标注技术研究[J]. 情报学报，2021，40（6）：621-629.

[15] 曹树金，闫颂. 基于语义角色信息的科技论文创新段落定位及功能句识别方法研究：以中文情报学领域论文为例[J]. 情报理论与实践，2022，45（11）：1-9，20.

[16] 张颖怡，章成志. 基于学术论文全文的研究方法句自动抽取研究[J]. 情报学报，2020，39（6）：640-650.

[17] 李雪思，张智雄，刘熠，等. 科技文献研究问题句识别方法研究[J]. 图书情报工作，2023，67（9）：132-140.

[18] 罗卓然，蔡乐，钱佳佳，等. 学术论文创新贡献句识别研究[J]. 图书情报工作，2021，65（12）：93-100.

[19] 章成志，李铮. 基于学术论文全文的创新研究评价句抽取研究[J]. 数据分析与知识发现，2019，3（10）：12-19.

[20] LIAKATA M，SAHA S，DOBNIK S，et al. Automatic recognition of conceptualization zones in scientific articles and two life science applications[J]. Bioinformatics，2012，28（7）：991-1000.

[21] COTOS E，PENDAR N. Discourse classification into rhetorical functions for AWE feedback[J]. CALICO Journal，2016，33（1）：92-116.

[22] 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.

[23] 马晓慧，赵文娟，刘忠宝. 基于深度学习的多学科多层次学术论文结构功能识别方法比较研究[J]. 情报科学，2021，39（8）：94-102.

[24] BADIE K，ASADI N，MAHMOUDI M T. Zone identification based on features with high semantic richness and combining results of separate classifiers[J]. Journal of Information and Telecommunication，2018，2（4）：411-427.

[25] 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.

[26] 王末，崔运鹏，陈丽，等. 基于深度学习的学术论文语步结构分

类方法研究[J]. 数据分析与知识发现，2020，4（6）：60-68.

[27] 欧石燕，陈嘉文. 科学论文全文语步自动识别研究[J]. 现代情报，2021，41（11）：3-11.

[28] 陈果，许天祥. 基于主动学习的科技论文句子功能识别研究[J]. 数据分析与知识发现，2019，3（8）：53-61.

[29] 刘江峰，冯钰童，刘浏，等. 领域双语数据增强的学术文本摘要结构识别研究[J]. 数据分析与知识发现，2023，7（8）：105-118.

[30] CHEN Y Q，PEROZZI B，AL-RFOU R，et al. The expressive power of word embeddings[C]//Proceedings of the 30th International Conference on Machine Learning，2013：1-8.

作者简介

林鑫，男，博士，副教授，研究方向：数字信息资源管理与服务研究，E-mail：xinlin@ccnu.edu.cn。
杜莹，女，硕士研究生，研究方向：信息组织与检索研究。
罗宇，女，硕士研究生，研究方向：信息组织与检索研究。

Structure Function Recognition of Scientific Research Project Application Based on Multi-Stage Classification

LIN Xin　DU Ying　LUO Yu
(School of Information Management, Central China Normal University, Wuhan 430079, P. R. China)

Abstract: The research project applications contain rich scientific knowledge and are widely used as the basic data for scientific and technological information analyses. Some information analyses such as duplicate detection and analysis mining need to be carried out on the premise of clarifying the structure function of the applications. Therefore, this paper proposes a research project application structure function recognition model based on multi-stage classification. Firstly, the research project applications should be preprocessed, including identifying the main content and multimodal elements of the applications, and standardizing the text paragraphs. Afterwards, based on the BiLSTM-Attention model, the chapter titles and their text are distinguished, and the primary structure function is recognized based on the titles. Furtherly, the fine-grained structure function of the application is identified. The experiment shows that the precision and recall rate of the model reach 93.7% and 93.1%. The model can support the structured analysis of scientific research project applications and provide references for the structure function recognition of other types of academic texts.

Keywords: Scientific Research Project Application; Structure Function Recognition; Multi-Stage Classification; BiLSTM-Attention

（责任编辑：王玮）