● 蔡 乐[1,2]，罗卓然[1,2]，陆 伟[1,2]
（1. 武汉大学信息管理学院，湖北 武汉 430072；2. 武汉大学信息检索与知识挖掘研究所，湖北 武汉 430072）

学术论文科研贡献类型自动识别研究*

摘 要：［目的/意义］学术论文中的科研贡献是论文中最有价值的信息类型之一。［方法/过程］文章将学术论文贡献内容从理论层面分为三个主要维度，即贡献功能、贡献重要性、问题—方法贡献。以此为指导，设计了一套包含五大贡献类别的标注框架，其中包括揭示贡献类型抽象性质的贡献分类标注体系及描述贡献内容的多层术语词汇功能标注体系。在此基础上，以SCI-BERT为基础模型，引入了学术论文的章节功能和结构化的术语信息，提出了语义角色标注增强下的科研贡献识别模型CNSC，并将其与过往的文本分类方法进行对比。［结果/结论］实验结果表明，文章提出的CNSC模型充分利用了论文的术语结构和贡献句的章节信息，对贡献类型的识别要优于其他模型。

关键词：贡献内容；学术论文；文本分类；预训练模型
DOI：10.16353/j.cnki.1000-7490.2023.06.023
引用格式：蔡乐，罗卓然，陆伟．学术论文科研贡献类型自动识别研究［J］．情报理论与实践，2023，46（6）：168-175.

Research on Automatic Recognition of Scientific Research Contribution Types of Academic Papers

Abstract：［Purpose/significance］Scientific research contributions in academic papers show the most valuable types of information. ［Method/process］To further explore the rich connotation of scientific research contribution, we divide the contribution content of academic papers into three main dimensions from the theoretical level, namely, contribution function, contribution importance, and problem method contribution. Under this guidance, we design a annotation framework including five contribution categories, including a contribution classification annotation module that reveals the abstract nature of contribution types and a multi-level glossary function annotation module that describes the content of contributions. Based on this, we introduce the chapter function and structured terminology information and propose the CNSC, a research contribution recognition model under semantic role annotation enhancement. ［Result/conclusion］The experimental results show that the CNSC model proposed in this paper makes full use of the term structure of the paper and the chapter information of the contributed sentences, and the recognition results of the contributed sentences are better than those of other models.

Keywords：contributed content; academic papers; text classification; pre-trained model

0 引言

学术论文作为科研人员从事科研探索中研究成果的一种重要载体，以标准化的结构提供了跟踪科学领域进展的线索。德国著名统计学家、Derek de Solla Price纪念奖章获得者沃尔夫冈·格兰泽尔在其“Bibliometrics as a Research Field a Course on Theory and Application of Bibliometric Indicators”中谈到原创贡献构成了科学的核心价值[1]。原创贡献在科学的认可体系中受到高度重视，并与资金分配、雇佣、任期评估和科学奖评选等关键科学决策高度相关。但科学发现的重要性、独创性和贡献价值往往很难衡量[2-4]，该观点先后得到了国内外许多有影响力学者的支持。通常来说，原创贡献通常通过同行评议的方式进行评估，这仅在小范围内可行，而在大范围内评估贡献则仍是一个巨大的挑战。同行评议系统可能会引发与研究活动一致的从众行为，这可能会缩小知识多样性的范围[5]。在信息计量学的研究中，学者们从不同角度开展了学术论文内容测度[6-7]，但由于科研贡献本身很难被识别与测度，为此如何准确地识别学术论文中的科研贡献依然是一个有待解决的问题。近年来，对不同层面的学术文本信息开展内容挖掘越来越受到研究人员的关注。然而，现有研究大多集中于关键短语提取、引文内容分析、修辞结构分析和相关句子抽取等，学术论文中的科研贡献类型识别仍处在初

* 本文为国家重点研发计划课题“服务内容资源知识表示、分类与编码和自动编目技术研究”的成果，项目编号：2019YFB1404702。

期探索阶段。

科研贡献类型表明研究论文与之前的主题研究相比如何提供新知识或新理解，能够帮助研究人员把握论文的主要贡献价值。一般来说，学术论文的主要组成部分包括摘要、介绍、相关工作、方法、实验和结果以及结论，与之对应的，科研贡献通常涉及研究问题、研究方法或研究成果的一种。例如，“本文提出基于学术全文本的创新贡献抽取方法，成功抽取出的‘创新点’和‘贡献点’对于科学评价中的学者评价、趋势预测等领域将起到基础资源的作用”[8] 和“我们的工作同以往的 Baseline 相比取得了巨大的进步”是两种不同的科研贡献，前者是研究方法贡献，后者是实验结果贡献。学术论文种的贡献句子通常包括一些典型的语言特征，如果将这类贡献描述句进行分类标注并实现自动分类，将有助于知识推荐、结构化生成以及科学进化分析等。因此，定义科研贡献类别、科研贡献的标注方案和构建类型标注数据库是科研贡献分类主要任务。

为进一步探索面向文本内容层面的科研贡献识别方法，本文对现有的学术文献标注方案进行调研和梳理，总结了目前科研贡献标注体系的优势和局限，发现已有分类体系缺乏对贡献术语内容及类型区分度的重视。为此，本文以计算协会年会会议记录（ACL）和《信息处理与管理》（IP&M）的开源数据集为切入点[9]，提出一套科研贡献标注框架，一方面结合科研贡献本身的重要性，对类似的功能类目重新组织；另一方面，在框架中加入语义信息标注，利用本文提出的框架进一步标注实验，并提出语义角色标注增强下的科研贡献识别模型 CNSC（A Cascade of Neural Models for SRL and Classification），通过模型对比实验，验证了新框架下的贡献类型区分更加明显，同时本文提出的 CNSC 表现效果更好，表明语义角色标注可以更好地帮助对科研贡献进行分类。

1 相关研究

科研贡献识别任务属于文本内容识别的任务，与此相关的文本内容层面的句子功能识别在科技文献研究中较为普遍，国内外相关学者从不同的视角对学术论文中的句子进行标注。Hao Wenke 等[10] 从未来研究入手，提出了一个针对未来工作的句子级标注方案，其中包括 6 个主要类别和 17 个子类别。J. D’Souza 等[11] 从知识组织视角，总结了学术论文贡献句的 10 个核心信息单元，分别是：研究问题、方法、目标、实验设置、结果、任务、实验、消融分析、基线和代码。从引文功能的视角出发，S. Teufel 等[12] 提出了一个引用功能标注方案，其中包含 4 个类别和 12 个细粒度类别，并对 320 篇会议论文进行了标注实验，同时使用 kappa 的一致性来检测标注结果的可信度。B. A. Lipetz[13] 定义了 4 组（施引文献的原创贡献、非原创贡献、一致性关系、施引文献对被引文献的情感）共 29 项特征的贡献指标用于提高学术引文索引中。陆伟等[14] 结合引文重要性，设定 5 个重要性等级，15 个功能类目，设计了一个较小粒度的功能体系，展现了一个全面的引文情景。综上，相关研究从未来句、贡献信息、引文功能等角度开展了学术论文句子级标注研究，为本文如何标注贡献类型及如何评估标注体系可靠性提供了良好的借鉴。

由于科研贡献内容的形态和特征各异，因此在实施贡献内容分析之前需有针对性地制订一个学术论文科研贡献标注方案，并按照贡献的对应特征对科研贡献进行归类和分析。S. Auer 等[15] 构建了开放研究知识图谱 ORKG，总结了每篇论文的基本贡献特性和价值，用以比较不同论文之间的科研贡献。L. Vogt 等[16] 使用知识图谱单元表示了学术知识图谱中的研究分布，与 ORKG 相比，他们提出的科研贡献模型 RCM 可以生成知识单元，其内容更易于维护和理解。J. D’Souza 等[11] 制定了一个科研贡献语义角色标注方案〈主语、谓语、宾语〉，以识别研究自然语言处理文献的贡献。Le Xiaoqiu 等[17] 对施引句进行分析，识别问题、方法、结论，同时通过情感分析和主题聚类反映被引文献的贡献程度。S. Minaee 等[18] 认为开发建立合适的科研贡献分类模型是一个反复实验的过程，但通常可以分为以下 5 个步骤：①预训练模型的选取；②目标域适应[19-21]；③适用特定任务的模型设计；④特定任务的微调；⑤模型压缩。此外，一些文本发现研究没有把目标聚焦于科研贡献识别上，而是瞄准了类似的任务——亮点句规律分析[4]、研究亮点提取[22]、亮点生成[23] 等研究中，科研贡献识别结果也被用于创新评价[24]、贡献图谱分析[25]、研究影响评估[26] 等任务中。

综上，不少学者通过文本内容对句子功能及其类型进行了拓展研究，已有相关研究为本文科研贡献识别研究提供了启发，被广泛认可的 S. Minaee 等提出的文本分类建模流程为本文提供了重要思路指导。同时本文也发现，耗时的手工标注、格式化数据获取困难等问题阻碍其进一步发展和应用。如今，随着自然语言处理和人工智能技术的发展，基于深度学习的文本分类模型从自动化角度为贡献类型的识别奠定了基础[8,27-28]。

2 贡献类型标注框架设计

从章节功能及术语信息理论视角分析贡献特征、类型以及两者之间的关系，即分析学术论文中科研贡献的“形态”与“效果”。一方面，科研贡献类型取决于文献中包

含的结构化术语知识；另一方面，贡献类型取决于贡献句在文献中的章节位置与体现的效果。因此，本文通过分析贡献句在文献中的位置与结构化术语，探寻作者贡献行为、章节位置在文献中的实际功能以及影响贡献类型的相关因素，研究思路框架如图1所示。

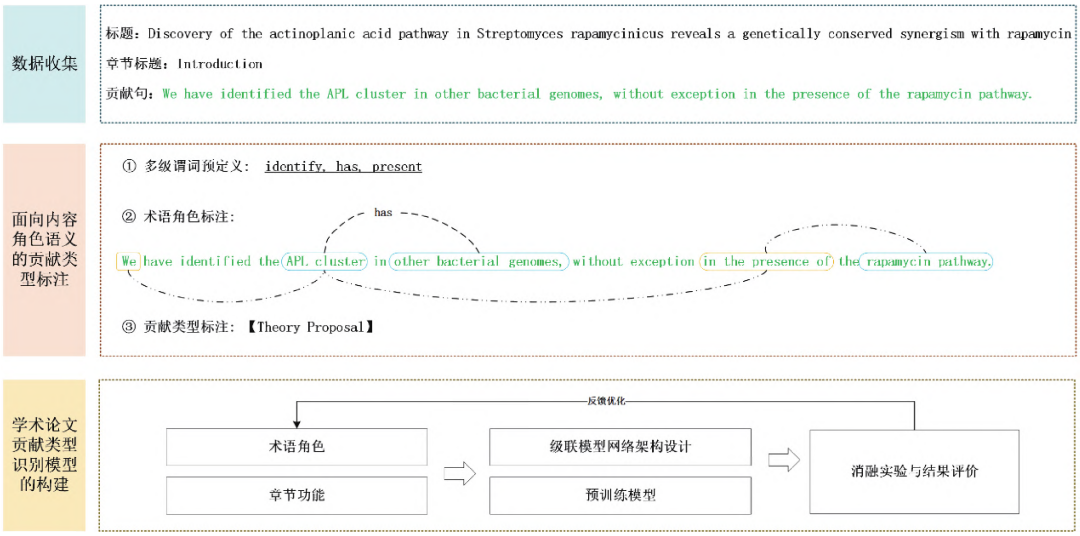

图1 研究思路框架

Fig. 1 Research idea framework

2.1 贡献类型标注体系

J. D'Souza等[11]在研究中指出，实施贡献内容分析首先对科研文献中包含贡献内容的句子进行识别和提取，其次需要制订一个为支持不同粒度进一步分析的标注体系。然而，制订一个全面、清晰、可执行的标注体系并非易事。一个组织分类合理且能够指导贡献类型自动分类的标注框架尤为重要，本文对目前影响力较大的贡献类型体系研究成果进行整理，并按照其分类所依据的标准维度，将已有体系归为3类。

1) 贡献类型。体现科研工作者在领域内哪些维度做出了科研贡献，是最主要的分类维度。Chen Haihua等[29]提出数据集创建、理论提出、模型构建、模型优化等11类贡献类型，罗卓然等[28]提出问题、理论、方法、成果4类贡献类型。

2) 贡献重要性。体现科研工作对研究领域推动的重要性。周海晨等[8]将科研贡献分为3个类别：解决、探明、建立的某一具体问题、规律、方法等为创新贡献，对领域有积极作用或有助于解决潜在问题为一般贡献，以及其他贡献。曹树金等[30]通过预训练模型先识别情报学期刊论文全文中的创新句，并通过语义角色信息匹配创新段落，将创新段落中的功能句分为“创新过程”与“创新贡献”。

3) 问题—方法贡献。体现科研工作提出了哪些新问题或者新方法，B. Qasemizadeh等[31]、张颖怡等[32]、丁睿祎等[33]、程齐凯等[34]、J. D'Souza等[11]围绕科研文献内的科研贡献，通过信息抽取的方式围绕研究过程中的问题、方法、模型等展开探究。

上述3种体系中，贡献的重要性倾向于从评价者的主观视角进行研究，与本文关注的维度并不一致。其他两类体系从抽象层面分析贡献内容与所属维度间的联系，均解释了贡献类型的典型特点，是本文设计的主要方向。本文很大程度上受到J. D'Souza等[11]成果的启发，认为细粒度的术语抽取对于科研贡献类型的研究具有重大价值，然而目前的成果很少有针对贡献类型自动化划分的研究。基于上述考虑，本文提出一个全面支持贡献类型分析的标注框架，主要包括：①一个揭示贡献类型抽象性质的分类体系；②一个描述贡献具体内容的多层词汇功能标注体系。本文参考先前工作中的类目设计并在自动化实验中进行调整，最终确定了5个贡献类目，如表1所示。

2.2 标注实验设计

为了评估上述贡献标注方案的可靠性，并创建一个可靠的用于自动科研贡献分类的数据集，本文在Chen Haihua等[9]公开的数据集上邀请三位拥有自然语言处理和机器学习的背景注释者参与进一步的标注实验，尽可能保证术语的专业性和完整性，标注过程具体可以分为3个步骤。

1) 标注贡献句的类型。在第一步骤中，来自ACL(2022)和IP & M(2022)公开数据集上的5024条句子被视为候选贡献句，作为实验的标注样本，将从5种不同的类型中选择一种进行标注，其中从ACL收集到3374个贡献句，从IP & M收集到1650个贡献句。如果句子中同时包含多种类型，则将句子进一步拆分，确保对每一条候选贡献句，有且只有一种创新类型。IP & M在投稿的过程中会要求总结归纳重点内容，因此哪些句子表明了科研贡献是明确且易于提取的。同时数据ACL是计算语言学顶

表 1　结合词汇功能的科研贡献分类体系
Tab. 1　Classification system of scientific research contribution combined with vocabulary functions

类型	描述	示例	术语
数据集/实验样本	创建或扩展新的数据集/样本	to apply the discovered patterns in the creation of a larger annotated dataset for training machine readers of research contributions	谓词: create 一级对象: a larger annotated dataset 二级对象: dataset
理论提出	提出一种新的理论来解决/改善现有的问题	We suggest viewing learning event embedding as a multi-relational problem	谓词: create 一级对象: learning event embedding 二级对象: multi-relational problem
模型构建或优化	构建模型或提出优化现有模型的策略	We used the BERTBASE model pre-trained on English Wikipedia and BooksCorpus	谓词: use 一级对象: BERTBASE model 二级对象: Wikipedia
效果评估	评估对比 baseline/现有效果	For NER, S-LSTM givesan *F*1-score of 91.57% on the CoNLL testset	谓词: for 一级对象: NER 二级对象: *F*1-score
成果运用	将提出的模型或理论在其他方面运用	to integrate the machine readers into the ORKG to assist users in the manual curation of their respective article contributions	谓词: integrate 一级对象: machine readers 二级对象: ORKG

级的会议，IP & M 是计算机和信息科学领域的顶级期刊，来源涉及信息检索、数据挖掘、自然语言处理等多个主题的理论与应用研究，保证了实验样本的高质量、普适性与广泛性。

2) 主体、谓词、对象实体的短语分块。然后对所选句子进行科学知识实体标注。实体由人工和标注器[35]共同进行标注，标注器对它们是否在每个三元组上下文中扮演主体、谓词和对象角色有一个隐式关系的识别。需要注意的是，根据本文的方案，谓词不一定是动词，也可以是名词。另外，在理想情况下，短语分块中的主体、谓词、对象实体应该在相应的句子中找到。但是，对于谓词来说可能并不总是在文本中找到如表 1 效果评估中的示例所示，因此有时会根据注释者的判断对其进行额外的注释。然而，即使是这种开放式的选择仍然从一组预定义选项中选择。它包括{ “has”, “on”, “by”, “for”, “has value”, “has description”, “based on”, “as”}。

3) 建立多层贡献序列。在步骤二标注的三元组中标注到的关联主题或对象本身，可能涉及关联关系，如图 2 所示，如果在细粒度的详细级别上执行注释，一个三元组中的对象可以是另一个三元组中的主体。为了灵活地表达这种三元组的嵌套关系，选用 Neo4j 图数据库储存数据。

3　科研贡献自动分类

SCI-BERT[20]是一种基于 Semantic Scholar 的科学论文训练的预训练语言模型。如前所述，预训练的文本嵌入在某些领域特定的数据集中效果较差，考虑到标注数据主要来自计算机领域，为避免原始预训练过程中带来的噪音，本文选择 SCI-BERT 作为预训练模型。如本文 2.1 节所述，科技文献的贡献内容可以划分为谓词及其对象形成的结构

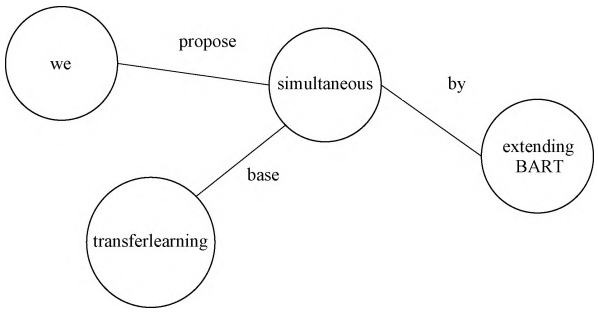

图 2　多层谓词嵌套示例
Fig. 2　Example of multilayer predicate nesting

化术语，但是现有的语言模型 ELMo，GPT 等只利用了上下文的语义信息来获得句子的嵌入表示，缺少结构化信息的考虑。因此，为了弥补现有语言模型的不足，获取更加丰富的结构化术语信息，提高学术论文贡献识别的效果，从预先标注的多层谓词序列中引入显式的术语信息，本文基于 SCI-BERT 提出了一种语义角色标注增强下的科研贡献识别模型 CNSC (A Cascade of Neural Models for SRL and Classification)，架构如图 3 所示。

CNSC 能够处理多个序列输入，在 CNSC 中，将输入序列中的单词及其对应的术语标签传递到语义角色编码器中，来获取多层谓词派生下的显式术语语义，并在一个线性层后聚合形成相应的术语语义嵌入表示。同时，利用 SCI-BERT 实现对输入序列中句子整体的编码，获得文本的上下文表示。此外，陆伟等[36]、黄永等[37]认为科技文献中每句话是否描述了该文献的一个贡献不仅仅基于句子的语义，还与它在文件中的位置有关，科技文献的章节标题和章节功能也为句子的贡献类型提供重要线索。如标题“相关工作”表明这部分的句子很可能讨论之前的科研贡献。因此，在 CNSC 中同时利用 SCI-BERT 对章节标题进

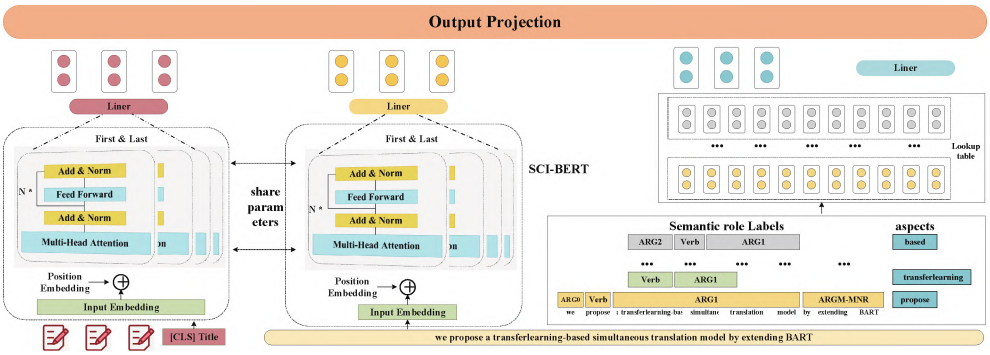

图3 语义角色标注增强下的科研贡献识别模型架构图

Fig. 3 Framework of scientific research contribution recognition model enhanced by semantic role tagging

BERT中，进行编码获得嵌入向量。输入句子 $X=\{x_1,x_2,\cdots,x_n\}$ 是长度为 n 的单词序列，它首先被标记为单词片段。随后，编码器通过transformer的encoder产生一系列上下文嵌入。章节标题 $F=\{f_1,f_2,\cdots,f_n\}$ 是长度为 m 的单词序列，以同样的方式行编码。最后，将术语表示、语义嵌入、章节功能连接起来，形成下游任务的联合表示，并使用最终的分类层进行分类。

3.1 语义角色编码器

为了在文本表示的基础上，进一步获取多层谓词派生下的术语语义，利用贡献句标注过程中的显式结构化术语信息，提高类型识别效果，在这一模块中，本文参照Zhang Zhuosheng等的研究[38]，对于与每个谓词 m 相关的术语序列，$T=\{t_1,t_2,\cdots,t_m\}$，其中每个 t_i 是一个长度为 n 的标签序列 $\{\text{label}_1,\text{label}_2,\cdots,\text{label}_n\}$，值得注意的是，在标注的过程中部分谓词没有在句中出现。针对这种情况，将额外标记的谓词拼接到原始序列中，并用特殊符号［SEP］进行分隔，使标签序列与单词序列得到对齐，标签序列与原始序列长度均为 n。随后，使用lookup table将术语标签映射成为向量 $\{v_1^i,v_2^i,\cdots,v_n^i\}$，并输入一个BiLSTM层中来获得术语标签序列的嵌入化表示 $e_{t_i}=\text{Bilstm}\{v_1^i,v_2^i,\cdots,v_n^i\}$。将 m 个谓词对应的标签序列拼接起来，用 L_i 来表示句子的术语编码得到 $e(L_i)=\{e_{t_1},e_{t_2},\cdots,e_{t_m}\}$。最后通过一层全连接网络来获得 m 个序列的术语表示 e^t，语义角色编码流程如图4所示。

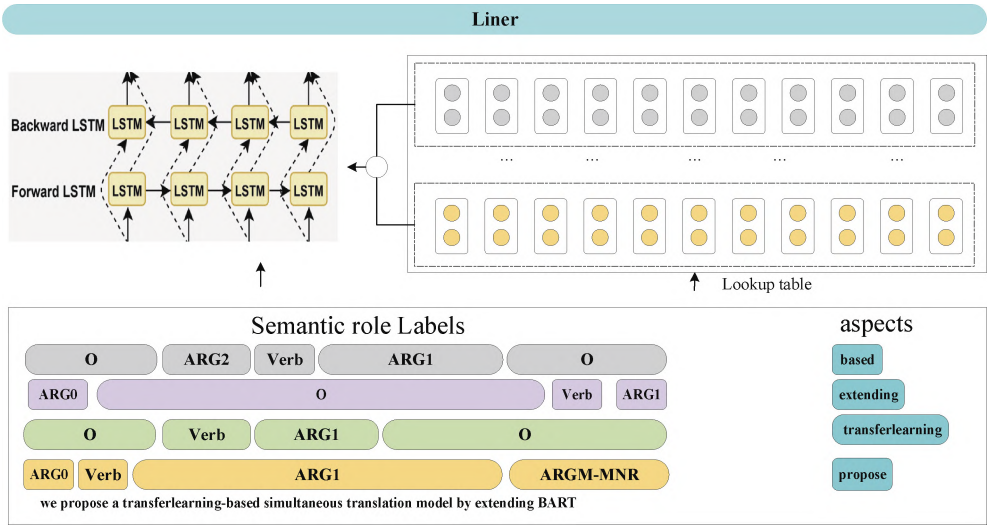

图4 科研贡献句术语语义角色编码器

Fig. 4 Semantic role encoder of scientific research contribution terms

3.2 章节功能及文本序列编码器

为了将章节功能信息引入模型中，本文将原始文本序列及该句所对应的章节标题，分别输入预训练模型SCI-BERT共享参数，随后将章节标题表示与文本序列表示拼接起来，获得句子的语义嵌入 e^s。章节功能及文本序列编码器流程如图5所示。

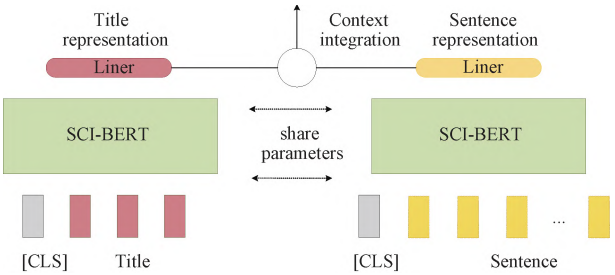

图5 章节功能及文本序列编码器

Fig. 5 Chapter function and text sequence encoder

3.3 实验结果分析

图6统计了数据集中不同类别贡献句的数量，从贡献句子类型分布上看，“应用”类与其他类相比数据是严重

不平衡的。严重的不均衡数据会导致模型倾向预测数量较多的类别，降低模型的效果。因此，本文使用对比学习的方式一定程度上缓解了样本数量过少，同时对数量较多的类别采用随机下采样的方式，来缓解样本不均衡问题。值得注意的是，在模型目标域适应的过程中，采用全量的文本数据来训练模型，以获得更好的文本表示。

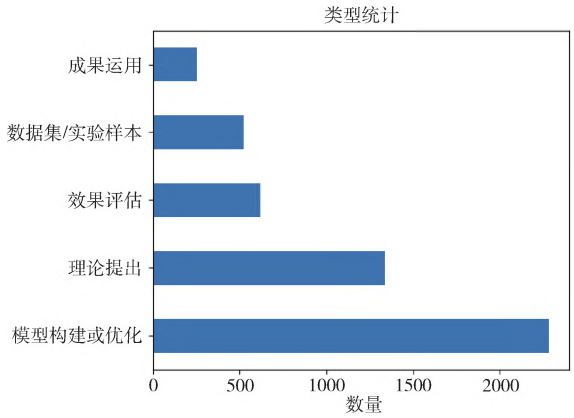

图 6　不同贡献类型分布

Fig. 6　Distribution of different contribution types

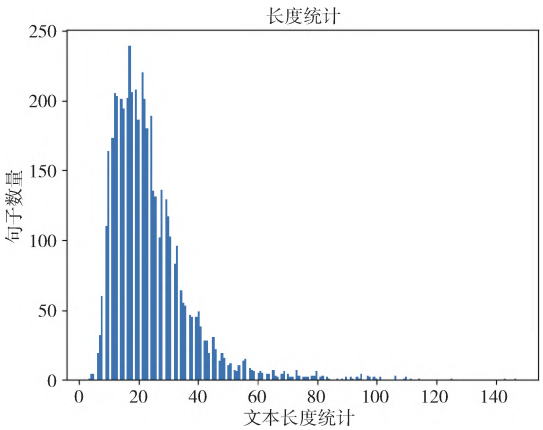

图 7　贡献句句长分布情况

Fig. 7　Distribution of contribution sentence length

本文基于 PyTorch 版本的 SCI-BERT 实现模型的构造。由于额外的术语表示嵌入维度远小于语义及章节功能表示的维度，因此可以直接使用 SCI-BERT 预训练的权重，并遵循与 SCI-BERT 相同的目标域适应及微调过程，不做任何修改，并在适度增加模型尺寸的情况下调整所有层。将初始学习率设为 {5e－6,1e－5,2e－5,3e－5}，warm-uprate 设置为 0.1，L2 warm-up 设置为 0.01。如图 7 所示，对贡献句的规模进行统计，贡献句多以短句为主，95% 以上的贡献句句长小于 128，因此实验过程中，max-length 设置为 128，batch size 大小在 {32,64,128} 中选择，术语嵌入维数设置为 64，默认的多层谓词的最大数量 m 被设置为 3。

使用文本分类问题上的常用指标召回率、精确度和 $F1$ 分数来评估每个类别的性能，总体结果如表 2 所示。

表 2　贡献句类型分类效果

Tab. 2　Classification effect of contribution sentence type

Model	Accuracy	Precision	Recall	$F1$ score
Manual features + LR	0.49	0.49	0.49	0.49
Manual features + RF	0.52	0.53	0.51	0.52
Word2Vec + LR	0.51	0.49	0.52	0.50
Word2Vec + SVM	0.54	0.52	0.52	0.52
BERT	0.54	0.55	0.57	0.56
SCI-BERT	0.59	0.56	0.56	0.55
SemBERT	0.58	0.53	0.64	0.58
CNSC (ours)	0.67	0.61	0.62	0.61

表 2 显示了在上文提到的贡献句类型识别数据集上的结果，结果表明，CNSC 比 SCI-BERT 有明显提高，并且优于 Chen Haihua 等[9] 的研究中传统手工特征提取下的机器学习模型。由于 CNSC 仍以 SCI-BERT 为骨干，计算过程相同，因此增益完全归功于新引入的章节功能及显式术语语义。虽然目前的主流模型在多任务处理、知识蒸馏、转移学习或集成等方面都取得了进步，但我们的单一模型具有轻量级和竞争力，设计简单甚至能得到更好的结果。

3.4　性能评估

本文通过消融实验来评估模型的各个组成部分在贡献句分类中的作用以及可能的改进方案。表 3 显示了在使用所有组成部分一起使用，仅使用章节功能及句子，仅使用语义角色编码器及句子以及仅使用句子时，模型在验证集上的性能。

表 3　贡献句类型分类消融实验效果

Tab. 3　Experimental results of ablation of contribution sentence type classification

Settings	P	R	$F1$
Sentense + Title + SRL	61.46	61.53	61.17
Sentense + SRL	57.28	62.38	57.89
Sentense + Title	60.87	56.24	60.03
Sentence only	56.39	56.01	54.83

从表 3 中可以看出，章节功能信息显著提高了性能，术语语义角色也起到了一定作用但作用较小，这可能是在分词的过程中因词汇本身被切分为更细粒度的 token 进而使语义角色与词汇本身未严格对齐导致，可以在未来的研究过程中进一步提高。综上所述，结合术语语义和章节功能信息对贡献句进行分类效果最好。为了进一步分析模型效果中存在的问题，本文对分类性能进行了评估。相关的混淆矩阵如图 8 所示。从图 8 中可以看出，CNSC 模型在数据量较大的类别中效果较好，但是在“成果应用”类别的效果较差，它的数据占数据集的 5.07%，这不足以

让模型充分学习类型特征，尽管采用了对比学习、数据增强、下采样等方式，但仍未得到充分解决。严重的混淆主要发生在模型与效果评估、成果应用中。例如，有些论文数据旨在讨论一个抽象的想法对模型的结构进行改动，有些专注于指标上的提升，但大多数论文处于两者之间的中间地带，在贡献句中近似同时包含两种类型。为了解决这一问题，本文也尝试一些手工特征的引入，但效果并不理想。此外，本文发现在实验中进行类别合并可以起到一定改善效果，这表明虽然本文的CNSC分类模型具有良好的准确性，但相似单元区分方面仍有一些改进空间，特别是在模型和效果评估之间。

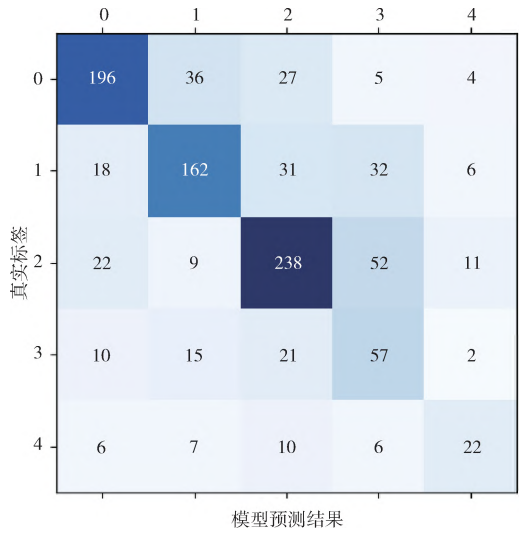

图8 CNSC模型结果混淆矩阵

Fig. 8 CNSC model result confusion matrix

4 结束语

本文提出了一个包含5类科研贡献及术语词汇功能的细粒度注释方案，基于上述方案，对开源数据集重新标注，用于科研贡献类型识别。此外，本文针对贡献类型识别任务，基于SCI-BERT提出了文本分类和序列标记模型的级联模型CNSC。实验证明了CNSC模型的优越性能，同时在消融实验中可以观察到章节标题信息显著提高了模型的效果，进一步验证了学术文本结构功能在科技文献细粒度解析中的重要作用。本文暂未对贡献类型间的细微差异做进一步细化与讨论，对贡献类型划分提出更细粒度的分类标准是下一步研究的方向。由于数据集的规模、来源有限，本文将在未来的工作中进一步扩大数据来源，增大数据集的全面性，并为科研贡献类型表示及类型划分挖掘更多更有效的特征。本文提出的细粒度注释方案可用于对学术文献中的科研贡献进行大规模计量分析，构建的科研贡献识别模型为自动化科研贡献及知识片段提取、推荐提供了基础。□

参考文献

[1] GLANZEL W. Bibliometric as a research field: a course on theory and application of bibliometric indicators [EB/OL]. [2023-01-20]. Course Handouts. http://nsdl.niscair.res.in/jspui/bitstream/123456789/968/1/Bib_Module_KUL.pdf.

[2] GASTON J. Originality and competition in science: a study of the British high energy physics community [M]. Chicago: University of Chicago Press, 1973: 5-6.

[3] Nigel Gilbert G. Referencing as persuasion [J]. Social Studies of Science, 1977, 7 (1): 113-122.

[4] 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.

[5] PARTHA D, DAVID P A. Toward a new economics of science [J]. Research Policy, 1994, 23 (5): 487-521.

[6] UZZI B, MUKHERJEE S, STRINGER M, et al. Atypical combinations and scientific impact [J]. Science, 2013, 342 (6157): 468-472.

[7] LEE You-Na, WALSH J P, WANG Jian. Creativity in scientific teams: unpacking novelty and impact [J]. Research Policy, 2015, 44 (3): 684-697.

[8] 周海晨，郑德俊，郦天宇．学术全文本的学术创新贡献识别探索 [J]．情报学报，2020，39 (8)：845-851.

[9] CHEN Haihua, NGUYEN H, ALGHAMDI A. Constructing a high-quality dataset for automated creation of summaries of fundamental contributions of research articles [J]. Scientometrics, 2022: 1-15.

[10] HAO Wenke, LI Zhicheng, QIAN Yuchen, WANG Yuzhuo, ZHANG Chengzhi. The acl fws-rc: a dataset for recognition and classification of sentence about future works [C] //Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020. Association for Computing Machinery, Online, 2020: 261-269.

[11] D'SOUZA J, AUER S. NLP contributions: an annotation scheme for machine reading of scholarly contributions in natural language processing literature [C] //EEKE 2020-Workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents, 1. August 2020, Virtual Event. Aachen: RWTH. Association for Computing Machinery, Online, 2020: 16-27.

[12] TEUFEL S, SIDDHARTHAN A, TIDHAR D. An annotation scheme for citation function [C] //Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue. Sydney, Australia: Association for Computational Linguistics, 2006: 80-87.

[13] LIPETZ B A. Improvement of the selectivity of citation indexes to science literature through inclusion of citation relationship indicators [J]. American Documentation, 1965, 16 (2): 81-90.

[14] 陆伟，孟睿，刘兴帮．面向引用关系的引文内容标注框架研究 [J]．中国图书馆学报，2014，40 (6)：93-104.

[15] AUER S, KOVTUN V, PRINZ M, et al. Towards a knowledge graph for science [C] //Proceedings of the 8th International

Conference on Web Intelligence, Mining and Semantics. New York, NY, USA: Association for Computing Machinery, 2018: 1-6.

[16] VOGT L, D'SOUZA J, STOCKER M, et al. Toward representing research contributions in scholarly knowledge graphs using knowledge graph cells [C] //Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020. New York, NY, USA: Association for Computing Machinery, 2020: 107-116.

[17] LE Xiaoqiu, CHU Jingdan, DENG Siyi, et al. Citeopinion: evidence-based evaluation tool for academic contributions of research papers based on citing sentences [J]. Journal of Data and Information Science, 2019, 4 (4): 26-41.

[18] MINAEE S, KALCHBRENNER N, CAMBRIA E, et al. Deep learning-based text classification: a comprehensive review [J]. ACM Computing Surveys (CSUR), 2021, 54 (3): 1-40.

[19] GU Yu, ROBERT T, CHENG Hao, et al. Domain-specific language model pretraining for biomedical natural language processing [J]. ACM Transactions on Computing for Healthcare (HEALTH), 2021, 3 (1): 1-23.

[20] BELTAGY I, LO K, COHAN A. SciBERT: apretrained language model for scientific text [C] //Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, 2019: 3615-3620.

[21] HOO-CHANG Shin, ZHANG Yang, BAKHTURINA E, et al. BioMegatron: larger biomedical domain language model [C] //Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, Online, 2020: 4700-4706.

[22] WANG Waiming, SEE-TO E W K, LIN Hongtao, et al. Comparison of automatic extraction of research highlights and abstracts of journal articles [C] //Proceedings of the 2nd International Conference on Computer Science and Application Engineering. New York, NY, USA: Association for Computing Machinery, 2018: 1-5.

[23] REHMAN T, SANYAL D K, CHATTOPADHYAY S, et al. Automatic generation of research highlights from scientific [C] //2nd Workshop on Extraction and Evaluation of Knowledge Entities From Scientific Documents (EEKE'21), Collocated with JCDL. Association for Computing Machinery, Online, 2021: 21.

[24] 李贺, 杜杏叶. 基于知识元的学术论文内容创新性智能化评价研究 [J]. 图书情报工作, 2020, 64 (1): 93-104.

[25] KOK M O, SCHUIT A J. Contribution mapping: a method for mapping the contribution of research to enhance its impact [J]. Health Research Policy and Systems, 2012, 10 (1): 1-16.

[26] MORTON S. Progressing research impact assessment: a "contributions" approach [J]. Research Evaluation, 2015, 24 (4): 405-419.

[27] 田亮, 李博闻, 章成志. 基于学术论文全文的跨语言研究方法自动分类研究 [J]. 图书馆建设, 2022 (1): 75-86.

[28] 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.

[29] CHEN Haihua, KANUBODDU B N. A fine-grained annotation scheme for research contribution in academic literature//Proceedings of the 18th International Conference on Scientometrics and Informetrics. Scientometrics, Online, 2021: 241-248.

[30] 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.

[31] QASEMIZADEH B, HANDSCHUH S. The ACL RD-TEC: a dataset for benchmarking terminology extraction and classification in computational linguistics [C] //Proceedings of the 4th International Workshop on Computational Terminology (Computerm). Dublin, Ireland: Association for Computational Linguistics and Dublin City University, 2014: 52-63.

[32] 张颖怡, 章成志, HE Daqing. 学术论文中问题与方法识别及其关系抽取研究综述 [J]. 图书情报工作, 2022, 66 (12): 125-138.

[33] 丁睿祎, 王玉琢, 章成志. 基于学术论文全文内容的特定领域算法实体抽取研究 [J]. 数字图书馆论坛, 2022 (3): 2-14.

[34] 程齐凯, 李鹏程, 张国标, 等. 学术文本词汇功能识别——基于标题生成策略和注意力机制的问题方法抽取 [J]. 情报学报, 2021, 40 (1): 43-52.

[35] QI Peng, ZHANG Yuhao, ZHANG Yuhui, et al. Stanza: a Python natural language processing toolkit for many human languages [C] //Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. AAAI Press Association for Computational Linguistics, Online, 2020: 101-108.

[36] 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.

[37] 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.

[38] ZHANG Zhuosheng, WU Yuwei, ZHAO Hai, et al. Semantics-aware BERT for language understanding [C] //Proceedings of the AAAI Conference on Artificial Intelligence. 2020. Palo Alto, California, USA: AAAI Press, 2020, 34 (5): 9628-9635.

作者简介: **蔡乐**, 男, 1998 年生, 硕士生。研究方向: 数据挖掘, 深度学习。**罗卓然** (通信作者, Email: zoraluo@whu.edu.cn), 女, 1993 年生, 博士生。研究方向: 创新评价, 数据挖掘。**陆伟**, 男, 1974 年生, 教授, 博士生导师。研究方向: 信息检索与可视化, 数据智能与创新评价, AI 人机协同等。

作者贡献声明: **蔡乐**, 数据收集与分析, 代码与论文撰写。**罗卓然**, 提出研究思路, 论文修改。**陆伟**, 提出研究思路, 论文修改。

录用日期: 2022 - 12 - 29