无名稿件编号。（将由编辑插入）

评估用于数字图书馆馆藏学术知识图谱构建的基于 BERT 的科学关系分类器

Ming Jiang · Jennifer D'Souza · S¨oren Auer · J. Stephen Downie

arXiv:2305.02291v1 [cs.DL] 3 May 2023

收稿日期：日期 / 接受日期：日期

摘要 研究出版物的快速增长，对数字图书馆（DL）的先进信息管理技术提出了巨大需求。为满足这些需求，业界正倡导采用依赖知识图谱结构的技术。在此类基于图谱的流程中，推断相关科学概念之间的语义关系是关键步骤。近年来，基于 BERT 的预训练模型被广泛用于自动关系分类。尽管取得了显著进展，但其中大多数模型是在不同场景下进行评估的，这限制了它们之间的可比性。此外，现有方法主要在干净文本上进行评估，忽略了早期学术出版物在机器扫描和光学字符识别（OCR）方面的数字化背景。在此类情况下，文本可能包含光学字符识别噪声，进而导致人们无法确定现有分类器的性能。为解决这些局限，我们首先基于三个干净语料库创建了含有光学字符识别噪声的文本。利用这些平行语料库，我们围绕三个因素，对八种基于 BERT 的分类模型进行了全面的实证评估：（1）BERT 变体；（2）分类策略；（3）光学字符识别噪声的影响。对干净数据的实验表明，领域专用的预训练 BERT 是识别科学关系的最佳变体。通常，每次预测单一关系的策略优于同时识别多个关系的策略。在噪声语料库上，最优分类器的 F 分数可能下降约 10% 至 20%。本研究探讨的见解有助于数字图书馆相关方选择适当技术，构建基于知识图谱的最优系统。

关键词 数字图书馆 · 信息抽取 ·学术文本挖掘 · 语义关系分类 ·知识图谱 ·神经机器学习。

1 引言

数字图书馆（DL）在促进学术知识传播与探索方面发挥着重要作用，而这正是学术交流的关键部分。虽然数字图书馆中海量学术记录的可访问性为众多研究社群的学术发展带来了实际益处，但这一现象也给数字图书馆员和数据策展人管理来自不同来源的学术信息带来了新的挑战。尤其是，现有的基于文献的学术交流模式给研究人员带来了难题：他们难以针对特定研究主题获取全面、细粒度且具有上下文敏感性的学术知识，跨学科研究尤其如此 [23]。根据 [4,5], 当前基于关键词的方法，

资助：本研究得到美国国家科学基金会第 OAC 1939929 号资助，以及欧洲研究委员会 ScienceGRAPH 项目（资助协议编号：819536）的支持。**利益冲突**：不适用**数据/代码可用性**：干净语料库和所选分类模型的源代码均已公开。相应的噪声语料库及相关代码将在论文发表后发布。**伦理审批**：不适用**参与同意**：不适用**发表同意**：不适用

M. Jiang 美国伊利诺伊大学厄巴纳-香槟分校
电子邮箱：mjiang17@illinois.edu

J. D'Souza 德国汉诺威 TIB 莱布尼茨科学与技术信息中心

S. Auer 德国汉诺威 TIB 莱布尼茨科学与技术信息中心
汉诺威莱布尼茨大学 L3S 研究中心

J.S. Downie 美国伊利诺伊大学厄巴纳-香槟分校

用于索引学术文章的方法会使相关文献彼此分散，并且强调集中关注选定的关键词，而不是文章中的全部知识。鉴于这些局限，为了优化数字图书馆中的学术知识组织与表示，一些倡议 [23,52] 主张结合人工策展与机器学习技术，构建相互关联且语义丰富的知识图谱结构。

在基于学术出版物构建知识图谱的过程中，一个关键步骤是建立已识别科学术语之间的语义关系。鉴于释义是自然语言中的常见现象，许多已识别的语义关系实例实质上是用不同表达方式指示同一种关系，这可能导致大型语料库所构建的图谱中出现信息冗余。为避免这一问题，科学关系分类任务（即从一组预定义关系中为每对相关概念确定恰当的关系类型）不可或缺。近期，自然语言处理（NLP）领域的一些研究人员基于对学术文章的实证标注，定义了七种常见的科学关系类型 [6,16],，包括 HYPONYM-OF、PART-OF、USAGE、COMPARE、CONJUNCTION、FEATURE-OF 和 RESULT。标注采用广义关系三元组的形式：⟨实验⟩ 比较 ⟨另一项实验⟩；⟨方法⟩ 用途 ⟨数据⟩；⟨方法⟩ 用途 ⟨研究任务⟩。

在这个“深度学习浪潮”时代，先前的研究已开始利用强大的神经网络技术构建科学关系分类器，以提升性能 [32]。随着 BERT（即双向 Transformer 编码器表示） [13]等 Transformer 模型的近期问世，构建性能更强的机器学习系统的可能性进一步凸显。尽管先前的研究 [8,50] 已取得较高的分类性能，但这些研究报告的结果不太具有可比性，因为所提出的方法通常是在不同的评估语料库上进行评估的。这一问题使得难以获得可比较的结果，也难以对所开发的分类器在现实应用中的有效性得出确切结论。尤其是在学术数字图书馆的情境下，数字图书馆员可能难以根据以往评估的发现，选择满足学术知识组织需求的最佳工具包（例如，底层数据在内容、规模和多样性方面可能各不相同）。此外，现有研究主要聚焦于未扫描的干净语料库 [8,30,16], 中的科学关系分类技术开发，这表明研究人员往往着重于机器自动识别以数字格式产生的最新学术出版物中科学概念之间关系的能力。然而，在更广泛的实践中，大量数字图书馆馆藏最初以印刷形式出版，之后通过机器扫描和光学字符识别进行数字化。这些文本不可避免地会包含独特的数字化噪声，例如 OCR 错误，这可能会对先进的、基于学习的关系分类技术构成挑战。这一差距进一步引发了人们对最先进分类模型能否在进行预测时稳健处理数字化所造成文本噪声的疑虑。

为解决上述局限，我们着重在全面且可比较的环境下，对最先进的基于 BERT 的分类模型进行全面考察，同时涵盖干净数据和光学字符识别噪声数据场景。为帮助数字图书馆领域的利益相关者选择构建学术知识图谱的最佳工具，我们提出了一个两阶段考察流程：(1) 确定适用于干净文本的最佳模型设置；(2) 进一步考察所确定的最佳模型设置在含有光学字符识别错误的噪声文本上的稳健性，而这正是数字图书馆馆藏中最普遍的局限之一。关于模型设置，我们特别关注两个关键因素的影响：(1) 分类策略（即一次预测单个关系还是多个关系）；(2) BERT 模型变体，具体涉及其预训练语料库的领域（即通用文本或科学文本）和词汇大小写（即区分大小写或不区分大小写）。为衡量最佳模型抵御光学字符识别噪声的稳健性，我们首先依据 HathiTrust研究中心提供的现有词级光学字符识别错误词典，随机向每个干净实例中添加一定比例的光学字符识别错误，从而生成对齐的噪声—干净实例对 [9]。获得平行数据后，我们比较了最佳模型在噪声文本与干净文本上的性能差异。

考虑到现实世界中的数字图书馆可能采用各种数据设置，且这种多样性很可能会影响模型的表现，我们在三个语料库上评估了每个模型的性能，包括：(1) 一个单领域语料库，其中包含自然语言处理领域学术出版物摘要上稀疏的关系标注 [16]；(2) 一个多领域语料库，涵盖从各类人工智能（AI）会议论文集中收集的出版物摘要上更丰富的关系标注 [30]；以及 (3) 前两个语料库的组合，其中数据领域的分布不均衡，且标注由两个不同的标注人员群体提供。构建这一语料库的目的是模拟数字图书馆中的真实数据设置。出于类似考虑，我们为每个干净语料库准备了两个含噪版本，重点考察 OCR 错误的数量，分别为低噪声版本（18%）和高噪声版本（49%）。

在三个不同的干净评估语料库上进行的实验表明，预训练于学术领域的 uncased BERT 模型。

针对特定学术领域文本的变体最适合用于分类科学关系的类型。关于分类策略，总体而言，每次预测单一关系的策略比同时识别多种关系类型的策略具有更高的分类准确率。进一步考察最优分类器在三个对应噪声语料库上的表现后，我们发现，分类器的可预测性明显下降了 10% 至 20% 的 F 值，这表明，随着语料库中 OCR 错误数量的增加，下降幅度也会进一步增大。

总之，按照我们的考察流程，本研究回答了以下研究问题：

RQ1：在干净数据上，八种基于 BERT 的分类器用于科学关系分类时表现如何？**研究问题2**：在所研究的七种关系类型中，哪些容易被选定的最优分类器识别？**RQ3**：经常出现哪些类型的预测错误？**RQ4**：光学字符识别错误如何影响最优分类器的整体表现？

RQ5：所研究的七种关系类型中，哪些对光学字符识别错误更具鲁棒性？

2 相关工作

2.1 从科学出版物中挖掘的关系

总体而言，数字图书馆中的知识主要依据数字馆藏的两个方面进行组织：(1) 元数据；以及 (2) 内容 [44,21]。后者还可进一步分为自由形式内容和本体化内容（即通过公理定义的形式化内容）。在此背景下，学术出版物中探讨的关系主要分为两类。

其中一类是元数据关系，例如作者身份、共同作者关系和引用 [49,43]。这类研究主要关注学术交流的社会维度，例如研究影响力评估 [36,40,18,10,12,34,46,56],、共同作者预测 [43,47,20,51,37]以及学术社群分析 [49] 这一类研究探讨的热门数据资源包括 Microsoft Academic Graph [42]和 Pubmed [1]。

第二组包括基于内容的语义关系，既可以是根据给定的自由形式内容 [28,26] 经验性定义的语义类别，也可以是通过对某一学科领域内概念属性进行系统的概念分析而正式定义的语义类别 [41,38]。在自动化系统的框架下，自由形式的基于内容的关系已从以下方面进行了研究：（1）关系识别（即识别相关的科学术语对） [16,26]；以及（2）关系分类（即，确定每对术语的关系类型，其中关系类型通常是预先定义的) [50,8,30]。关于本体定义的属性，先前的研究主要考虑基于形式概念分析的概念层次结构 [41,38]。

我们尝试对基于自由形式内容的语义关系进行分类。鉴于数字图书馆有意创建关联数据 [19],，我们尝试完成的任务可直接促进学术知识图谱的构建，并提供结构化数据，以支持图书馆员生成关联数据。

2.2 语义关系分类技术的发展

关系分类既有基于规则的方法 [2]，也有基于学习的方法 [11,54]。传统上，基于学习的方法通常依赖人工设计的语义和／或句法特征 [2,11]。在各种方法中，基于知识数据库 [33,39,22] 应用远程监督的策略已得到广泛采用，并在此基础上进一步改进。这种方法的主要优势在于，能够解决模型训练缺乏人工标注真实标注这一难题。

近年来，深度学习技术受到广泛研究，因为它们能够更有效地学习潜在特征表示，以区分不同关系。基于注意力机制的双向长短期记忆网络（BiLSTM） [54] 是最早表现出色的系统之一，它利用神经注意力机制捕捉每个句子中的重要信息，以进行关系分类。另一种先进系统 [31] 采用基于 BiLSTM 的动态跨度图框架，同时提取术语并推断它们之间的成对关系。除了这些考虑词序的神经方法之外，基于 Transformer 的模型（如 BERT） [13]利用自注意力机制量化每个词与其上下文之间的语义关联，现已成为关系分类领域的当前最先进方法。除了使用书籍和维基百科进行训练的通用 BERT 模型外，Beltagy 等人 [8]最近还开发了 SciBERT，这是一种使用学术出版物进行训练的 BERT 模型。

在分类策略方面，以往研究通常采用单关系逐次分类（SRC），每次识别实体对的关系类型 [54,31,8]。为提高分类效率，Wang 等人 [50] 设计了一种基于 BERT 的分类器，可一次识别多个实体对之间的关系，这可以视为多关系同时分类（MRC）。不同于以往着重于改进分类的研究，我们专注于对现有资源进行细粒度分析，以选择合适的工具

来提取和组织数字图书馆中的科学信息。

2.3 噪声数据上的关系分类

一般而言，先前针对噪声数据中实体关系分类的研究通常关注噪声标注，包括实体层面的噪声（如未纠正的实体边界） [16,17]，或关系层面的噪声（如错误的关系标签） [25,15,53]。这类噪声通常来自两个方面：（1）由于个人对细粒度文本语义的理解存在差异，人工标注产生偏差；（2）基于远程监督、依赖通用知识库的机器标注出现错误，因为其中缺少目标内容特定的上下文信息。

为解决标注中的这类噪声，一些方法着重开发基于远程监督学习机制的基于注意力机制的神经网络模型，使模型能够学习如何将来自现有大型知识库的通用结构信息与语料库
特定的语义信息，从而提高关系分类器的鲁棒性 [17,25]。此外，一些研究 [15,53] 提出采用强化学习策略：先筛选高质量的标注数据，再将筛选出的实例输入关系分类器进行训练。

尽管先前的工作在提升学习模型应对噪声标注的鲁棒性方面做出了显著贡献，但这些研究所依赖的文本数据仍然是干净的。因此，我们对噪声数据的考量与先前工作不同：我们关注文本中的基于内容的噪声，尤其聚焦于数字化图书馆馆藏中的 OCR 内容。

2.4 OCR 错误对下游自然语言处理的影响 sks

随着将自然语言处理技术应用于数字图书馆文本资源以开展宏观层面的计算研究日益普及[24,4]，人们近来愈发关注自然语言处理技术在处理数字化图书馆馆藏时的可靠性 [27,45,48]。根据我们的文献综述，影响自然语言处理技术处理 OCR 文本时可靠性的主要问题之一，是文本可能包含 OCR 过程产生的错误 [27,45,48]。

常见的 OCR 错误包括字符替换、词语分离、词语连写和错误符号。鉴于 OCR 错误普遍存在、分布不均且性质各异，即使采用最先进的 OCR 校正技术，也难以彻底清除这类文本噪声。因此，标准自然语言处理技术应用于含有 OCR 错误的文本时，其性能存在不确定性。

总体而言，现有关于 OCR 错误对自然语言处理任务影响的研究可分为

Id	关系类型	SemEval18		SciERC		组合	
		总计	%	总计	%	总计	%
1	**USAGE**：一种用于其他科学实体或被其他科学实体使用的科学实体-tific 实体。例如，*MT* 系统应用于日语	658	42.13%	2,437	52.43%	3,095	49.84%
2	**特征特征-抽特**：实体是特征或抽象模型 另一个实体的。例如 合一的 合一	392	25.10%	264	5.68%	656	10.56%
3	**合取**：通过词汇合取关系（ 即使用“和”或“或”连接）关联的实体。例如，来自*GoogleVideo* 和 *NatGeo* 纪录片	-	-	582	12.52%	582	9.37%
4	**部分-整体**：处于部分-整体关系中的科学实体—— 关系。例如，描述话语篇中话语的处理过程	304	19.46%	269	5.79%	573	9.23%
5	**结果**：一个实体会影响或产生一个结果。例如，仅使用 12 训练说话人 进行说话人无关识别，取得了 7.5% 的词 错误率	92	5.89%	454	9.77%	546	8.79%
6	**下位词**：一个实体，其语义范畴包含于另一个实体的语义范畴之中 例如，图像匹配 是一个问题 ，属于计算机视觉	-	-	409	8.80%	409	6.59%
7	**比较**：一个实体与另一个实体进行比较。例如，*con-* 对话记录具有显著不同的特征，与其相比。 整洁的文本	116	7.43%	233	5.01%	349	5.62%
	总体	1,562	100%	4,648	100%	6,210	100%

表 1：语料库统计概览（也可通过 https://www.orkg.org/orkg/comparison/R38012 访问）。“总计”和“%”列分别显示所有摘要中标注了相应关系的实例数量和百分比。

两类。第一类基于定量分析 [27,45]。在这类研究中，研究人员通常会衡量并比较同一种自然语言处理工具应用于干净文本和 OCR 识别文本时的性能差异。第二类研究基于定性分析。例如， [48] 对研究人员进行了一系列访谈，收集他们使用自然语言处理技术分析数字档案的反馈。基于学术用户的反馈，该研究分析并总结了 OCR 错误对自然语言处理技术的影响。这两类研究都考察了多种自然语言处理任务，包括分词、句子切分、词性标注、命名实体识别、主题建模、文档级信息检索、文本分类、词语搭配和作者归属判定。根据以往研究的发现，未经纠正的 OCR 错误会持续对自然语言处理任务产生负面影响，其中一些任务甚至会“受到 OCR 错误的不可挽回的损害” [45]。

在本研究中，我们通过探索 OCR 错误对学术出版物关系分类的影响，拓展了先前的研究。具体而言，我们旨在系统地考察基于 BERT 的关系分类器在干净文本和含 OCR 噪声文本上的表现。

3 语料库准备

本研究的源数据包括两个公开可用的数据集 [16,30],，每个数据集均包含 500 篇原生数字化学术摘要，并对以下内容进行了人工标注：(1) 科学术语；(2) 每对相关术语的语义关系类型。为全面考察基于 BERT 的关系分类器从学术出版物中识别科学关系的能力，并着重关注真实世界的数字图书馆场景，我们准备了干净和含噪语料库。下面介绍各版本语料库的详细信息。

3.1 干净语料库

关于干净语料库，除了直接将选定的两个原始数据集用作两个实验语料库外，我们还将这两个语料库组合成第三个新语料库。该语料库提供了更贴近现实的评估设置，因为它能更全面、多样地呈现任务。表 1 展示了实验语料库的统计信息，以下各小节将分别详细介绍每个语料库。

3.1.1 C1：*SemEval18* 语料库。

该语料库是为在 SemEval-2018 [16] 举办的第七届共享任务创建的。数据收集自 ACL Anthology [1] 中的学术出版物：共计 500 篇摘要，其中 350 篇被划分为训练数据，其余 150 篇为测试数据。最初，该语料库中的标注包含六种离散的语义关系，用于捕捉主要的信息内容。由于关系“主题”的标注数量远少于其他关系类型，因此在评估中我们省略了这一关系类型，并考虑以下五种关系类型：用途、结果、模型、部分-整体和比较。

3.1.2 C2：*SciERC* 语料库

尽管第二个评估语料库 SciERC [30]所标注的摘要数量相同，但与 SemEval18 语料库不同，该语料库涵盖了多种底层数据领域，其摘要取自五个研究领域的 12 个人工智能（AI）会议/研讨会论文集：人工智能、自然语言处理、语音、机器学习和计算机视觉。这些摘要标注了以下七种关系：比较、部分关系、合取、评估、特征关系、使用关系和下位词关系。与 *C1* 类似，该语料库由创建者预先划分。他们采用了 350/50/100 训练集/开发集/测试集划分。将*C2* 与*C1*进行比较后，我们发现，*C2* 中的五种关系（不包括合取和下位词关系）与*C1*中标注的关系在语义上相同。

3.1.3 C3：组合语料库

最后，将*C1*和*C2*合并，创建了这个评估语料库。在合并过程中，我们对一些语义相同但标签不同的关系重新命名。首先，将*C2* 中的USED-FOR与*C1* 中的USAGE统一为USAGE。此外，通过观察*C1*和*C2*中的关系标注，我们发现*C1* 中的RESULT与*C2* 中的EVALUATE-FOR表达的含义基本相似，但这两种关系的论元顺序相反。例如，在*C2*中，“[accuracy] for [semantic classification]”被标注为“accuracy” → EVALUATE-FOR → “semantic classifica- tion”，可视为“semantic classifica- tion” → RESULT → “accuracy”。因此，我们将语料库*C2* 中所有标注为EVALUATE-FOR的实例都改为RESULT，并调换其论元顺序。通过整合来自两种资源的共1000篇带有人工标注的摘要，我们构建的第三个评估语料库呈现了一个规模更大、异构性更强且相对更贴近现实的评估场景。

[1] https://aclanthology.org/

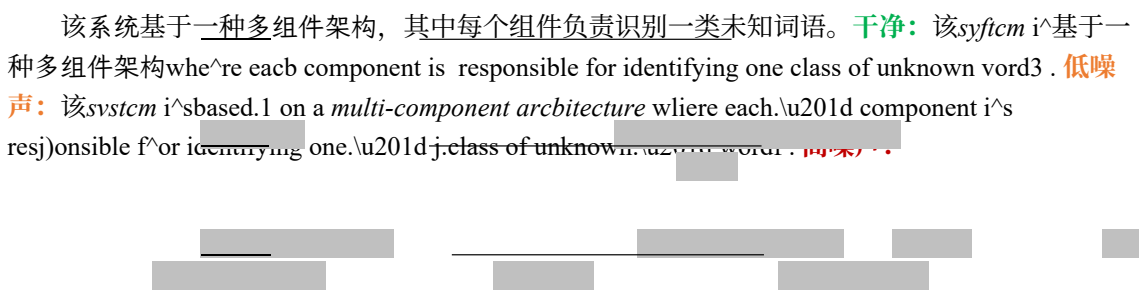

图 1：一个句子三个版本的示例：(1) 干净；(2) 低噪声；以及 (3) 高噪声。带下划线的短语是原始无噪声数据集中经人工标注的科学术语；灰色高亮短语是 OCR 错误引入的文本。

3.2 含 OCR 错误的噪声语料库

鉴于现有关于科学关系分类的研究主要聚焦于利用人工清理的纯文本开发技术，据我们所知，迄今似乎尚无研究利用自然语言处理技术探讨 OCR 错误对科学关系分类的影响。因此，对我们开展这项研究而言，一个主要挑战是缺乏可用的数据资源。理想情况下，此类数据集应满足两个要求：(1) 包含内容上相互平行的文本的干净版本和含噪版本；以及 (2) 在干净数据上具有经人工标注的关系类型作为真实标注。

考虑到直接准备此类语料库的成本很高，不仅耗时，而且需要大量人工劳动来执行手动检查文本对齐和标注关系类型等任务，我们为本研究提出了一种替代策略：利用上述现成的、带有人工标注的干净语料库，并将其中部分内容替换为未经校正的 OCR 文本。

为采用这一策略，我们使用了一个现有的词元级 OCR 错误词典 [9]，其中列出了大量经常未被纠正的 OCR 识别词元及其对应的纠正结果。据 [9]，，这些词元对收集自 1700 至 1922 年间出版的大规模数字化英语文献（178,381 卷）。该词典的总词汇量为 43,955。为细致分析 OCR 错误对基于 BERT 的分类器性能的影响与文本噪声数量之间的关系，我们没有替换每篇学者摘要与词典之间所有重合的词语，而是设置了一个参数来控制替换比例。此外，

	低噪声	高噪声
SemEval18	18.09% ± 0.073	49.23% ± 0.103
SciERC	17.75% ± 0.070	48.88% ± 0.099
组合	17.96% ± 0.071	48.97% ± 0.100

表 2：每个噪声语料库中的 OCR 错误分布

考虑到扫描文本中未经校正的真实世界光学字符识别（OCR）错误通常分布不规则，为了模拟这种模式，我们为每篇学术摘要从预设值组中随机选取一个比例值，并基于词典与摘要内容交集中的词元候选项进行随机抽样替换。在我们的实证实现中，我们准备了两个级别的噪声语料库：（1）低噪声语料库，其中文本的替换比例值从集合 {0.4、0.5、0.7} 中随机选取；（2）高噪声语料库，其中文本的比例值为 {0.9 或 1.0}。表 2 展示了各语料库中 OCR 错误的分布。平均而言，低文本噪声语料库中的每篇文本约有 18% 的 OCR 错误，而高噪声语料库中的每篇摘要约有 49% 的错误。

为了更好地了解我们构建的噪声语料库，我们在图 1 中给出了一个示例。不出所料，文本噪声较高的句子比噪声较低的句子包含更多 OCR 错误。我们可以看到，即使是相同的词语/短语，文本噪声也可能有所不同。例如，在低噪声句子中，词语“architecture”被替换为“**d’architettura**”；而在高噪声句子中，该词语则变成了“**arcbitecture**”。鉴于词典中的许多干净词语都有多个 OCR 错误版本，我们为每个词语词元的替换随机选取一个版本，这有助于使 OCR 错误的异质性与真实场景相近。

4 基于 BERT 的科学关系分类器

BERT [13] 是一系列基于尖端神经网络技术构建的预训练语言表示；它为自然语言处理从业者提供高质量、开箱即用的语言特征，从而提升多种自然语言处理任务的性能。这些模型会生成上下文化的 词嵌入，可直接用作下游任务的特征。此外，只需在核心 BERT架构上进行少量特定任务的扩展，就能以相对低廉的成本针对当前任务微调这些嵌入，进而进一步大幅提升任务性能。

在本研究中，我们采用 BERT 嵌入，并使用两种分类策略对其进行微调：（1）逐一关系分类（SRC）；（2）多关系同时分类（MRC）。本节余下部分将首先介绍我们采用的 BERT 模型，然后分别介绍经过微调的 SRC 和 MRC 分类器。

4.1 预训练 BERT 变体

作为预训练语言表示，BERT 模型有多种变体，具体取决于：(1) 模型配置参数，例如模型大小和预训练任务；以及 (2) 预训练数据设置，例如语言、词汇大小写和文本领域。本研究根据预训练语料库的两个关键因素组合，选取了以下四种核心变体，每个因素均包含两种类别：(1) 文本领域（即通用或科学）；以及 (2) 词汇大小写（即区分大小写或不区分大小写）。

BERT基础 [2] 我们使用的前两个模型属于预训练 BERT基础模型。这些模型使用 BooksCorpus（8 亿个词语）[55] 和英文维基百科（25 亿个词语）文本数据中的数十亿个词语进行预训练。我们选取的两个模型分别是：(1) 区分大小写的模型（训练 BERT基础时保留了底层词语的大小写）；以及 (2) 不区分大小写的模型（训练 BERT基础时，所有底层词语均转换为小写）。

SciBERT [3] 本研究采用的接下来两种模型属于预训练科学 BERT（SCIBERT）类别。它们是基于 BERT 的语言模型，在大型科学文本语料库上进行训练。具体而言，预训练语料库是从 Semantic Scholar [3] 随机抽取的 114 万篇论文，包含已发表论文的全文，其中 18% 来自计算机科学领域，82% 来自广义生物医学领域。与 BERTBASE一样，对于 SCIB
ERT，我们同时使用其**区分大小写**和**不区分大小写**的变体。

4.2 微调后的基于 BERT 的分类器

我们在两个神经系统扩展中实现了上述 BERT 模型，分别采用不同的分类策略。

单关系逐次分类（SRC）为 SRC 构建的分类模型通常会在核心 BERT架构的基础上扩展一个额外的线性分类层，其维度为 $K \times H$，其中 K 是标签数量（即关系类型），而 H 表示.....的维度。

[2] https://github.com/google-research/bert
[3] https://github.com/allenai/scibert

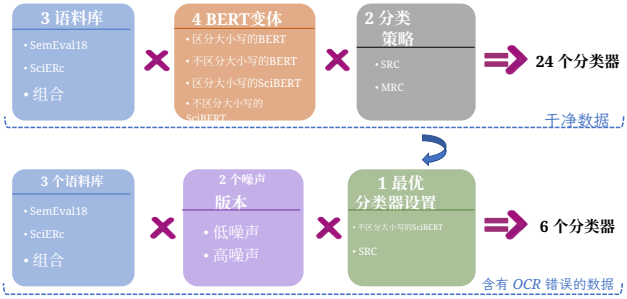

图 2：分类器统计数据。

词嵌入空间。标签概率进一步通过 softmax 函数进行归一化，分类器将概率最高的标签分配给每个相关概念对。

多关系同时分类（MRC）该策略是分类问题中较新的一项创新，在这种方法中，分类器可以一次使用句子中的所有关系实例进行训练，或一次性预测所有实例，而不是分别处理每个实例。不过，在这种情况下，核心 BERT架构的自注意力机制经过修改，可以高效地考虑科学术语 [50], 相对位置的表示，从而使新型多关系同时分类问题的编码成本变得可控。为获得分类概率，与 SRC 策略类似，MRC 也扩展了一个线性分类层。不过，此时该层侧重于同时计算句子中每个相关术语对所分配标签的概率。最后，每对术语的标签分配与 SRC 中相同，基于 Softmax 函数进行。

5 实验

5.1 实验设置

图 2 简要总结了本研究中考察的分类器。我们总共在干净语料库上构建了 *24* 个分类器，并在存在光学字符识别错误的语料库上构建了 *6* 个分类器。每个语料库均由原始数据集创建者划分为训练集／开发集／测试集。为获得每个语料库上的最优分类器，我们针对 η 将学习率参数调至 {2e-5、3e-5、5e-5}。对于训练轮次等其他参数，我们采用 SciBERT 和 BERT 模型中的默认值。

在评估分类器的性能时，我们采用了标准的分类评估指标，包括：精确率（P）、召回率（R）、F1 分数（$F1$）和准确率（Acc）。

	SRC						MRC						平均值±标准差	
	SemEval18		SciERC		组合		SemEval18		SciERC		组合			
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
BERT基础版（无大小写）	76.42	71.74	84.6	77.25	81.75	77.38	80.4	79.98	83.42	74.84	80.84	76.29	81.24±2.84	76.25±2.78
BERT 基础版（区分大小写）	73.58	71.14	85.32	77.92	78.73	74.38	79.55	78.44	83.72	75.07	79.42	74.8	80.05±4.14	75.29±2.65
区分大小写的SciBERT	73.58	69.72	**86.86**	**79.65**	**84.46**	**81.60**	80.11	78.32	83.42	74.35	**81.80**	**77.68**	81.71±4.60	76.89±4.25
不区分大小写的SciBERT	**80.97**	**79.42**	86.14	79.49	83.11	80.27	**81.82**	**80.54**	**84.33**	**77.44**	81.06	76.76	**82.91±2.04**	**78.99±1.54**
平均分数	准确率 84.10 F1 80.22						准确率 82.35 F1 77.52							

表 3：在三个数据集（SemEval18、SciERC 和组合数据集）、四种 BERT 模型变体（区分大小写与不区分大小写的 BERT；区分大小写与不区分大小写的 SciBERT）以及两种分类策略（SRC 和 MRC）上的科学关系分类结果。*Acc.* 表示准确率，$F1$ 表示宏平均 F1 分数；最高分以粗体显示。

5.2 分类结果与分析

5.2.1 RQ1：八种基于*BERT*的分类器在干净数据上的科学关系分类性能如何？

表3概述了基于 SRC 和 MRC 分类策略，在干净语料库上训练和测试的八种分类器的分类结果。我们的比较主要侧重于以下三个关键方面的分类器。

分类策略，即 ***SRC*** 与 ***MRC*** 的比较。根据表 3 中所示的 Acc 和 $F1$，我们观察到，除 SemEval18 外，SRC 在两个数据集上的表现均优于 MRC。SemEval18 数据集的一个特征是，其标注数量明显少于另外两个数据集。鉴于此，我们推断，新颖的 MRC 策略比 SRC 更具鲁棒性，因为即使标注数量减少，其性能水平也未受到影响。

词嵌入特征，即 **BERT** 与 **SCIBERT**。关于不同基于 BERT 的模型编码的词嵌入特征，SciBERT 在三个语料库上的准确率和 F1 分数均高于 BERT。由于我们的实验语料库均为学术数据，正如预期，领域专用 BERT 模型编码的词嵌入能够比通用 BERT 模型编码的嵌入特征更好地捕捉词元级语义关联，从而支持域内语料库中的关系分类。

BERT模型的词汇大小写，即区分大小写与不区分大小写。我们观察到，不区分大小写的BERT 模型（SCIBERT：82.91 BERT：81.24）平均而言，比其区分大小写模型对应版本（SciBERT：81.71，BERT：80.05）表现出更高的分类准确率。此外，不区分大小写模型的准确率总体标准差（SciBERT：2.04，BERT：2.84）低于区分大小写模型（SciBERT：4.60，BERT：4.14）；对 $F1$ 的比较也大致类似。因此，我们的结果表明不区分大小写的BERT模型能够获得比区分大小写的变体更稳定的性能。

总之，就分类策略而言，我们观察到 SRC 的表现优于 MRC （见表 3 最后一行的平均分数）。尽管如此，先进的 MRC 策略表现始终稳健，与 SRC 相比，其性能受较小数据集规模的影响相对较小（例如，SemEval18 语料库上的 SRC 与 MRC 结果）。另一方面，就 BERT 词嵌入变体而言，根据表 3 最后一列的平均分数，不区分大小写的 SciBERT模型在学术文章文本特征编码方面表现最佳。为进一步验证我们的发现，我们对分类结果进行了一系列统计检验。考虑到本研究使用了多个数据集，且任意两组样本均值之差的分布可能并非正态分布，我们决定使用 Wilcoxon 符号秩检验，探究任意两种模型设置变体之间预测差异的显著性。我们逐行比较了三种数据集在两种分类策略下任意两种 BERT 词嵌入变体的表现（分别依据表 3 中的 F1 分数和准确率），观察到基于不区分大小写的 SciBERT的分类器显著优于基于区分大小写和不区分大小写BERT的分类器（$p<0.05$）。为探究 SRC 和 MRC 之间的性能差异，我们在三个数据集上进行了逐列比较。同样，该比较分别依据 F1 分数和准确率。检验结果显示，在来自 SciERC 和组合语料库的测试数据上，SRC 和 MRC 之间存在统计学显著差异（$p<0.05$）。尽管分类结果显示，在 SemEval18 中，MRC 比 SRC 更有利于基于 BERT 的分类器，但这种性能差异在统计学上并不显著。

5.2.2 研究问题*2*：所选最优分类器容易或难以识别所研究的七种科学关系类型中的哪些？

针对每个语料库的每种分类策略，我们选取了最优分类器（即基于SCIBERT的模型，可以是区分大小写的或

关系类型	SRC			MRC		
SemEval18	P	R	F1	P	R	F1
用途	**87.22**	89.71	**88.45**	90.53	**87.43**	**88.95**
结果	78.26	**90.00**	83.72	**100.00**	75.00	85.71
比较	85.71	85.71	85.71	75.00	85.71	80.00
模型-特征	66.67	75.76	70.92	70.83	77.27	73.91
部分-整体	79.25	60.00	68.29	70.83	72.86	71.83

表4：在 SemEval18 上采用 underSRC 和 MRC 策略的最佳 BERT 变体的按关系分类结果。

关系类型	SRC			MRC		
SciERC	P	R	F1	P	R	F1
用于	**93.30**	91.37	**92.32**	**88.75**	90.24	**89.49**
合取	87.97	**95.12**	91.41	80.69	**95.12**	87.31
下位词关系	92.31	89.55	90.91	80.00	82.93	81.44
评估	82.29	86.81	84.49	84.44	83.52	83.98
比较	72.73	84.21	78.05	83.87	68.42	75.36
部分关系	66.04	55.56	60.34	65.52	60.32	62.81
特征关系	59.02	61.02	60.00	73.68	47.46	57.73

表5：在SciERC上采用SRC和MRC策略的最佳BERT变体的按关系分类结果。

关系类型	SRC			MRC		
组合	P	R	F1	P	R	F1
合取	**92.56**	**91.06**	**91.80**	85.07	**92.68**	**88.72**
用途	91.30	88.98	90.13	**87.96**	87.71	87.84
下位词关系	89.39	88.06	88.72	83.12	78.05	80.50
比较	86.89	89.83	88.33	73.85	81.36	77.41
结果	76.36	75.68	76.02	84.69	74.77	79.43
部分关系	75.86	66.17	70.68	68.33	61.65	64.82
特征关系	58.02	75.20	65.51	60.28	68.00	63.91

表6：Combined语料库上，最佳 BERT 变体在 SRC 和 MRC 策略下的按关系分类结果。

不区分大小写的），随后进一步考察这些分类器识别真实标注数据中标注的各类关系的能力。表4、表5和表6分别展示了SemEval18、SciERC和组合语料库上的结果。

关系类型敏感性概述。总体而言，三个表格中的结果表明，在这两种分类策略下，与其他关系类型相比，USAGE（USED-FOR）关系更容易被识别。对此现象的一种可能解释是，USAGE 是所有语料库中的主要类型，因此，与其他类型相比，分类模型能够更容易地学习这种关系类型的潜在语言模式。

对于较具挑战性的关系，我们发现，FEATURE-OF（MODEL-FEATURE）和PART-WHOLE（PART-OF）更难识别，在全部三个表中，其F1 分数均低于其他关系类型。我们的观察可能由两个方面解释。首先，这两类关系的语言表达多样性较高，使基于 BERT 的分类模型难以捕捉这两类关系的语言模式。例如，通过查看标注为 PART-WHOLE（PART-OF）的实例，我们发现，这种关系类型的关键信号包括多种形式，例如 ⟨A⟩“由” ⟨B⟩ “组成”、 ⟨A⟩ “.....位于” ⟨B⟩，以及 ⟨A⟩ “B ⟨的一个核心实例”⟩。此外，这两类关系在语料库中的标注数量相对较少，尤其是在SCIERC语料库中，这降低了模型预测这两类关系时的泛化能力，从而导致模型在未见过的测试样例上的F1 分数较低。

分类策略对各关系类型分类的影响。通过考察 SemEval18 语料库的分类策略（见表4），我们发现，对于用途、模型-特征和部分-整体，SRC 分类器和 MRC 分类器的分类排名顺序相同；但对于结果和比较，排名顺序则相反。具体而言，SRC 分类器在识别比较方面表现更好，而 MRC 分类器则更擅长识别结果。至于 SciERC（见表5），值得注意的是，从 SRC 策略改为 MRC 策略后，识别下位词关系的能力显著下降，这表明当下位词关系类型与其他类型混合时，其语言模式难以被捕捉。

异质人工标注对逐关系类型分类的影响。 将 SemEval18 和 SCI- ERC 组合成 Combined语料库，使该语料库更加异质且贴近实际。鉴于此，我们进一步探究了来自不同语料库来源的两组人工标注的混合，是否会对分类器识别各关系类型的能力产生影响。

有趣的是，与表4（SemEval18）和表5（SCIERC）所示结果不同，在这两张表中，USAGE（用于）的预测 F1 分数在 SRC 和 MRC 策略中均排名第一；我们观察到，在 Combined语料库中，对于 SRC 和 MRC 分类器而言，合取是最容易分类的关系类型（见表6）。此外，表5中合取的 F1 分数高于表6中的分数。由于 SemEval18 语料库中不存在合取关系，我们可以推断，这类关系的测试样例应与 SCIERC语料库中的相同。鉴于此，我们的结果表明，两组人工标注的关系 USAGE 可能存在不一致之处，这影响了模型对这种关系类型的可学习性。另一方面，SciERC语料库中对合取的一致标注，可能有助于模型捕捉 Combined语料库中此类关系一致的潜在语言模式。

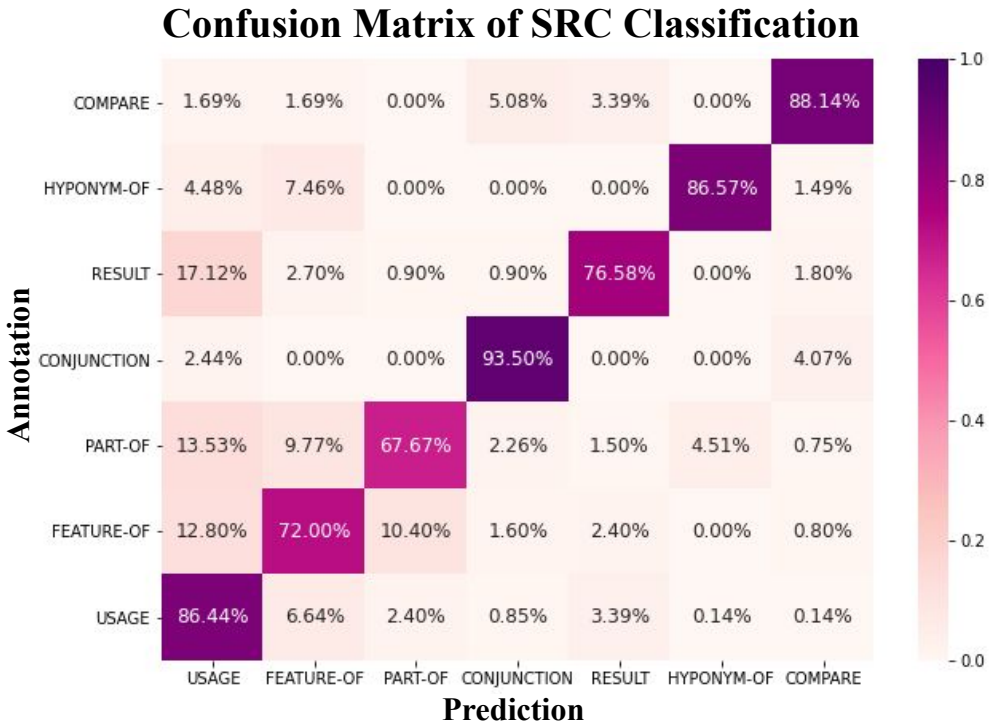

图3：Combined语料库上的SRC分类混淆矩阵。

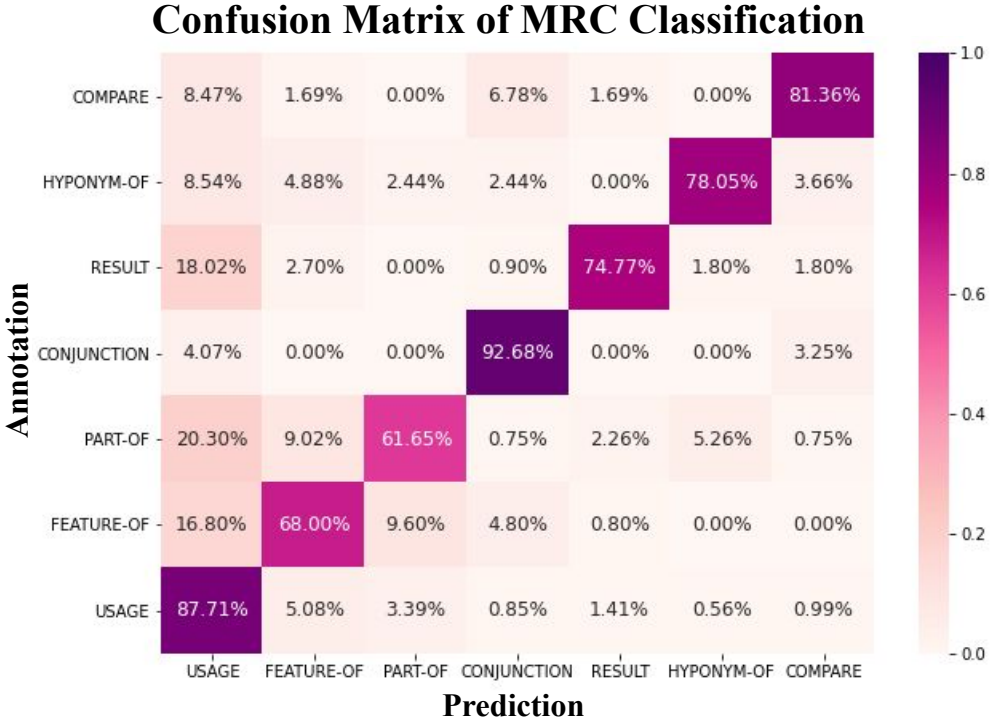

图4：Combined数据集上的MRC分类混淆矩阵。

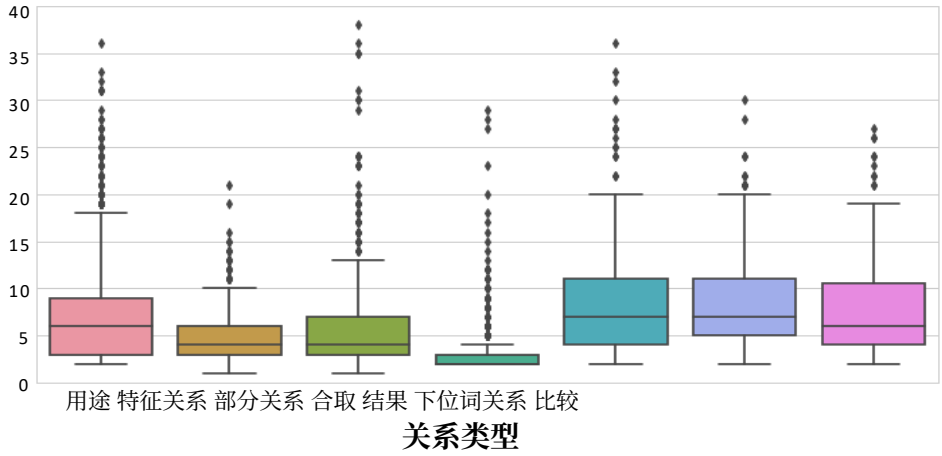

图 5：Combined语料库摘要中科学术语对之间词距的分布。

语料库，因此使其能够更有效地分类。不过，我们可以看到，USAGE 的分类 F1 分数几乎等于各个语料库得分的平均值。这为预测来自多个来源的异质语料库上的分类行为提供了一个潜在指标。

5.2.3 RQ3：常见的预测错误有哪些？

SRC 和 ***MRC*** 分类的混淆矩阵。图 3 和图 4 分别展示了 Combined语料库中 SRC 分类器和 MRC 分类器的混淆矩阵，便于更仔细地考察误分类情况。两图的结果均表明，分类器往往会将其他关系类型（如 RESULT、PART-OF 和 FEATURE-OF）主要误分类为 USAGE，MRC分类中尤为明显。一个可能的原因是，USAGE 关系的标注实例数量最多，导致分类器倾向于这一关系类型。总体而言，训练样本分布不均衡（详见第 3 节）是机器学习系统产生混淆的主要原因之一。同时，对多种关系类型同时进行预测的策略，可能会加剧这种偏差，使模型在学习时受到影响。

具体而言，在两种分类策略中，特征关系、部分关系和用途都极易彼此混淆，这表明这三种关系之间的语义关联比其他关系类型更紧密，因此机器难以区分这三种关系类型。我们进一步参照 SemEval18 和 SciERC 的标注指南，验证了这一发现。例如，在 SciERC 的指南中，实例 ⟨B “属于” A⟩ 在语义上与实例 ⟨B“是 A 的一部分”⟩相似。然而，前一种关系被定义为特征关系，而后一种关系则属于部分关系。我们还发现，下位词关系（约 6%）更容易被误分类为特征关系（以及用途），但反向情况并未出现。这种单向关系在一定程度上表明，机器可能学到了这两种关系类型之间的层级关系，即下位词关系涵盖特征关系，但反之则不成立。

总之，我们的研究结果表明，最优的基于 BERT 的分类器在科学关系分类中出现错误，可能由三个因素造成：（1）数据分布不均衡；（2）预定义关系类型存在语义歧义；（3）预定义关系类型具有层级语义。

词距分布。 为了从另一个相关角度分析分类器的错误，我们计算了 Combined 语料库中每种关系类型下相关科学术语对之间的词距分布。结果如图 5 所示。总体而言，图 5 中的大多数箱线图都呈偏态，上须较长而下须较短。这种模式表明，文本中成对科学术语之间的距离通常较近。与其他关系相比，合取的词距要短得多，这很合理，因为具有这种关系的术语对通常由单个连接词（如“和”和“或”）连接。这种一致的模式可能也是合取关系相较于其他关系更容易分类的另一个原因。此外，特征关系、部分关系、下位词关系和比较关系的平均词距比其他关系更接近下四分位数。这种分布差异可能给分类器识别这些关系带来挑战。值得注意的是，特征关系和部分关系的中位数及分布范围相近，这或许可以解释为什么分类模型难以识别这两种关系。

5.2.4 研究问题 *4*：*OCR* 错误如何影响整体最优分类器的性能？

在前文对干净语料库上的分类结果进行分析后，我们发现，在总体上，基于 underSRC 构建的未区分大小写的 SciBERT 最适合识别科学关系。为了使不同语料库的结果具有可比性，本节采用这一最优模型，并进一步分析其在所有评估语料库的噪声版本上的分类性能。我们的目标是探究 OCR 噪声对预测的影响是否存在一致的模式。

表 7、8 和 9 分别展示了在 Se-mEval18、SciERC 和 Combined语料库上的分类结果。我们的分析主要聚焦于两个方面：（1）分类中各种带噪训练集和测试集划分组合所产生的 OCR 错误的影响；（2）与 OCR 错误影响相关的其他潜在数据因素，例如语料库规模。下面将详细说明各个方面。

测试数据中 ***OCR*** 噪声的影响。三张表中的结果表明，测试数据中 OCR 错误的增加会导致分类器的预测性能迅速下降，而无论训练数据中的文本噪声有多少。例如，在各语料库的干净文本上训练的分类器（见表 7、8、9），在 OCR 错误较少的测试数据上的 F1 分数损失约为 5%，而在文本噪声较多的测试数据上的损失约为 20%。在训练数据相同的情况下，我们观察到，随着文本中 OCR 错误的增加，预测损失也随之增加。

训练数据中 ***OCR*** 噪声的影响。 关于训练数据中的 OCR 错误数量，我们观察到，在每个语料库中，增加训练数据中的文本噪声有助于分类器提高其鲁棒性，使其在测试样例上进行预测时的 F1 分数损失更低，尤其是对于文本噪声量较高的训练集。如三张表所示，使用干净数据训练的分类器性能损失约为 20% 的 F1 分数，而基于高噪声训练数据构建的模型，平均 F1 分数损失降至 6%。我们的观察表明，文本中的 OCR 错误会明显影响基于 BERT 的分类器识别句子中科学关系的性能。有趣的是，训练数据中的文本噪声具有一种正则化效应，作用于 the

基于 Transformer 的神经网络模型在训练过程中，这本质上有助于模型提升其处理含噪未见数据的泛化能力。

不同语料库中 ***OCR*** 噪声规模的影响。进一步比较三个语料库的分类结果后，我们发现，分类器在以下语料库上的可预测性

训练 / 测试	干净		低噪声		高噪声	
	F1	Loss	F1	Loss	F1	Loss
干净	**79.42**	-	**75.11**	-	**64.82**	-
低噪声	73.16	5.26↓	72.21	2.90↓	62.72	2.10↓
高噪声	56.35	23.07↓	58.35	16.76↓	58.26	6.56↓

表7：SRC 中未区分大小写的 SciBERT 在含噪 SemEval18 语料库上的分类结果。最高 F1 分数以粗体显示。

训练 / 测试	干净		低噪声		高噪声	
	F1	Loss	F1	Loss	F1	Loss
干净	**79.49**	-	77.51	-	**78.05**	-
低噪声	75.16	4.33↓	**77.92**	-	76.35	1.70↓
高噪声	60.11	19.38↓	69.94	7.57↓	72.11	5.94↓

表 8：SRC 中无大小写区分的 SCIBERT 在含噪 SciERC 语料库上的分类结果。最高 F1 分数以粗体显示。

训练 / 训练	干净		低噪声		高噪声	
	F1	Loss	F1	Loss	F1	Loss
干净	**80.27**	-	**78.64**	-	**77.10**	-
低噪声	72.84	7.43↓	77.41	1.23↓	74.94	2.16↓
高噪声	57.43	22.84↓	69.29	9.35↓	70.93	6.17↓

表9：SRC中噪声环境下Combined语料库上未区分大小写的SciBERT分类结果。最高F1 分数以粗体显示。

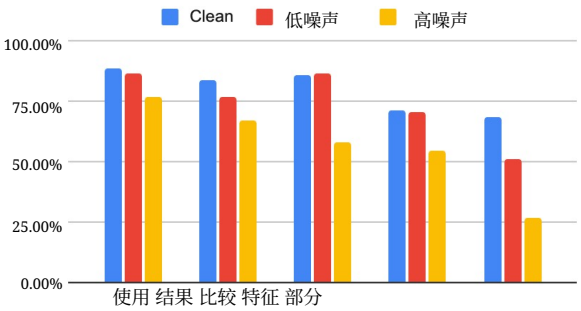

图6：干净、低噪声和高噪声SemEval18 语料库中各关系类型的分类结果。

SemEval18 语料库的表现要差得多：当训练数据存在噪声时，其 F1 分数低于分类模型在另外两个语料库上的表现。例如，当训练数据处于高噪声状态时，对于每种测试集，预测差异可能达到约 10%。鉴于 SemEval18 语料库的规模小于另外两个语料库，我们的观察表明，较大的语料库规模可能会减轻文本噪声对用于识别科学关系的基于 BERT 的分类模型的影响。

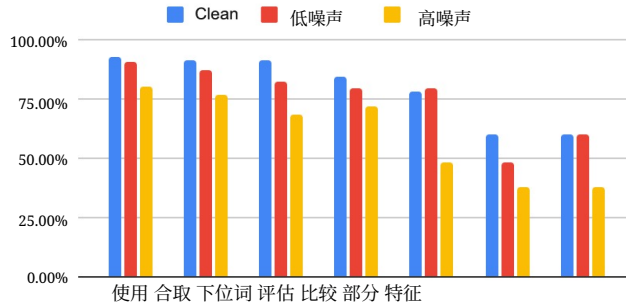

图 7：在干净、低噪声和高噪声 SCIERC语料库上的逐关系类型分类结果。

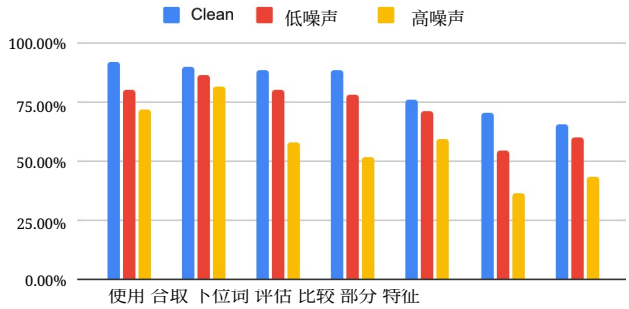

图 8：在组合语料库的干净、低噪声和高噪声数据上，各关系类型的分类结果。

5.2.5 RQ5：所研究的七种关系类型中，哪些对 *OCR* 错误具有更强的鲁棒性？

鉴于当前最先进的科学分类技术大多基于干净语料库开发，为了研究此类技术针对每种关系类型对文本噪声的脆弱性，我们考察了在干净语料库上训练的最优分类器从存在或不存在 OCR 错误的文本中识别各关系类型的能力。图 6、图 7 和图 8 分别展示了在三个语料库上的结果。

总的来说，对三个语料库的结果表明，无论测试文本中的噪声量多少，分类器识别各关系类型的难度往往是一致的。相比之下，比较、部分关系（部分-整体）和下位词关系是三种表现对大量光学字符识别错误更为敏感的关系类型：与其他关系类型相比，这些类型的 F1 分数损失更高。这表明，科学概念之间的这三种关系容易受到文本噪声的破坏。相反，各分类器对用途和结果（评估）的预测能力，比对其他关系类型更能抵御文本噪声。因此，我们推断，用途关系与结果关系之间可能存在很强的语义关联，能够在一定程度上克服光学字符识别错误的干扰。

进一步考察光学字符识别错误数量所导致的文本噪声影响后，我们发现，对于大多数关系类型，分类器在干净文本与低噪声文本之间的识别能力差异，远小于低噪声文本与高噪声文本之间的差异。唯一的例外是部分关系（部分-整体）。这一结果表明，在对大多数预定义关系进行预测时，我们的最优分类器能够适应少量文本噪声（约为 10%），但部分关系（部分-整体）除外。

5.3 科学关系分类在学术知识图谱构建中的应用案例

最后，作为本文研究的关系三元组的实际示例，我们根据 Combined语料库中的人工标注构建了一个知识图谱。图 9 展示了生成的知识图谱的可视化示例，其中包括：(1) 语料库层面的图（见右上方）；以及 (2) 两个概念层面的图，例如概念节点“machine translation”的自我网络（见左上方）和“词语”（见底部）。

观察语料库级图谱后，我们发现“method”“approach”和“system”等通用科学术语是连接最密集的节点，这符合预期，因为这些通用术语通常见于各个研究领域。在“machine translation”的放大自我网络中，我们可以看到，下位词关系因其连接“machine translation”及其同级节点（即其他研究任务，包括“speech recognition”和“natural language generation”）与上位节点“自然语言处理问题”而得到突出体现。“lexicon”这一概念通常用于（即，用途）关于“machine translation”和“operational foreign language”的研究。合取关系连接了“machine translation”和“speech recognition”这两个术语，它们都旨在将信息从一个来源翻译到另一个来源。关于“词语”的自我网络，我们发现这一术语主要涉及两类语义关系：（1）作为其他自然语言组成部分的一部分，例如“语料库”“song lyrics”和“句子”；或（2）在自然语言处理计算中作为特征关系中的特征，例如获取“part of speech information”和捕捉词语级的“co-occurrence patterns”。

除了单独考察节点之间的语义关联以及自我网络中的链接表示，并进一步考虑网络结构外，我们还可以推断某些相关科学概念的语义层级。例如，“machine translation”与“ranking tasks”均通过下位词关系与“NLP problems”相连，而“ranking tasks”用于解决（即，用途）“NLP problems”。这一三角结构表明，“ranking tasks”是比“machine translation”更细

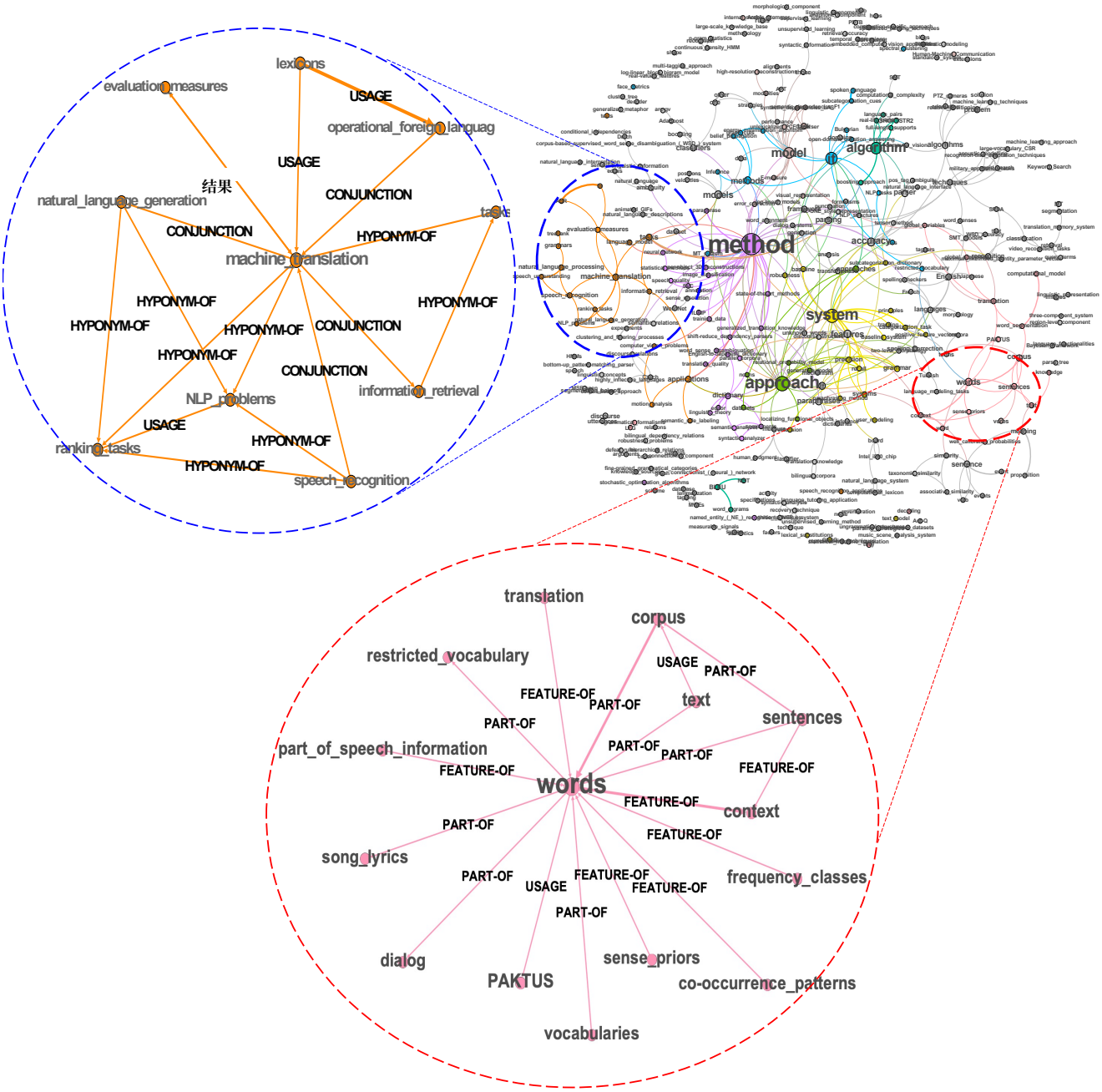

图 9：根据 Combined语料库中的关系三元组构建的知识图谱。以“机器翻译”和“词语”的自我网络作为两个示例。节点大小由节点的加权度决定。颜色表示根据图结构划分的模块度类别。该图使用 Gephi（https://gephi.org/）生成。

粒度的上位词，而不是“NLP problems”。此外，通过比较“machine-translation”与“words”的自我网络结构，我们观察到以“machine-translation”为中心的图比以“words”为中心的图更稠密。从本质上说，这一观察表明，“words”是比“machine translation”更通用的概念，因为它往往与各种彼此独立的概念相关联。

总之，给定一对具有关联的科学概念及所识别出的关系类型，我们可以构建知识图谱，以表示科学概念之间的语义关联：在宏观层面上涵盖整个语料库，或在微观层面上关注与特定概念相关的细粒度连接。

6 讨论

6.1 科学关系分类的基于 BERT 的模型建议

基于本研究的发现，我们向数字图书馆的利益相关者提出以下建议，以应用最优技术，自动从学术文章中分类科学关系：

– 与通用领域相比，使用专门由学术出版物构成的语料库预训练 BERT 模型，有助于构建能够识别科学概念之间关系的分类器。– 就分类策略而言，SRC 总体上优于 MRC。– 总体而言，不区分大小写的 BERT 模型在准确率更高、预测表现更稳定方面优于区分大小写的模型。– 对于包含 OCR 错误的噪声文本，如果文本噪声量较少（例如约为 10%-20%），那么在干净语料库上训练的最优分类模型（即使用不区分大小写的 SciBERT 进行 SRC 分类构建的模型）在进行预测时能够抵御此类文本噪声。– 对于 OCR 错误数量不确定的噪声文本，使用噪声数据训练分类器，有助于提升模型处理不同 OCR 错误数量文本的泛化能力。– 在基于 SciBERT 的 SRC 分类中，规模较大的语料库和丰富的关系标注有助于：(1) 提高预测准确率；(2) 减轻训练数据中的文本噪声对分类器构建的影响。– 对于关系类型标注，训练集中某种关系类型的大量标注有助于分类模型提高对该关系类型的可学习性。– 对于每一种预定义关系，表达中的固定句法结构有助于分类器区分不同的关系类型。– 与其他关系相比，预测 USAG 和 RESULT (EVALUATE-FOR) 等具体科学关系时，文本噪声对分类器的影响较小；而对于 COMPARE 和 HYPONYM-OF 等表示概念层级的某些关系，情况则相反。

6.2 研究局限与挑战

通过对评估流程进行系统性审查，我们发现了若干研究挑战及由此产生的研究局限，可概括为三个方面：(1) 人工标注收集；(2) 噪声语料库的准备；以及 (3) 评估方法。各方面介绍如下。

6.2.1 人工标注的收集

标注的成本与资源 正如先前研究 [16,30],所提到的，学术信息抽取研究（包括从学术记录中抽取实体和关系）面临的一项常见挑战，是收集人工标注作为基于学习的方法训练和/或测试所需真实标注的成本高昂且资源有限。这主要是因为，对科学信息进行标注不仅需要耗费时间和人力，与通用领域（如新闻文本）信息抽取任务类似，还要求标注人员具备目标学术文章所属特定科学领域的更高专业水平，因此较难找到合适的标注人员。此外，这类标注的费用也应当很高。由于这些挑战，科学信息的人工标注数量通常有限，而现有可获取的人工标注所涉及的研究领域数量也勉强够用。

受上述问题的影响，我们的评估主要集中于涵盖人工智能相关研究领域的语料库，这可能会导致我们对基于BERT的分类器在生物学、物理学等其他研究领域识别科学关系时的性能存在一定的不确定性。

作为真实标注的标注质量 根据我们的研究发现，关系类型定义的模糊性是导致分类器误分类的关键因素之一。这种模糊性主要由两个方面造成。首先，某些关系类型定义的含义使人们在标注数据时难以清楚理解这些关系类型。例如，在 SCIERC 数据集的标注指南中 [30], ⟨B⟩ “属于” ⟨A⟩ 被定义为 FEATURE-OF 的一个实例，而 ⟨B⟩ “是.....的一部分” ⟨A⟩ 则被定义为 PART-OF 的一个实例。从本质上说，这两个实例在语义上十分相似。其次，不同的标注系统可能会导致人工标注不一致。例如，对于具有层级语义关联的两种关系，即 HYPONYM-OF 和 PART-OF，SemEval18 数据集的指南只考虑 PART-OF，而 SCIERC 的指南则要求标注人员标注这两种关系。由于指南存在差异，标注人员很可能会在 SemEval18 数据集中用 PART-OF 标签标注真实的 HYPONYM-OF 关系。

除了标注歧义之外，语料库中还存在标注偏差，例如关系标签分布不平衡，这可能导致分类器偏向于识别某些代表性较强的关系（例如，

使用）。由于这些挑战导致标注质量存在不确定性，我们的评估可能会低估基于 BERT 的分类器识别科学关系的能力。

6.2.2 基于数字化的噪声语料库准备 *on*

为了研究数字化造成的文本噪声对关系分类技术的影响，构建理想的噪声语料库通常需要满足两个条件：（1）语料库应同时包含内容平行的干净文本和真实世界的OCR文本；（2）干净语料库应具有较高的质量。然而，这两项要求并非总能满足。一个主要原因是，数字化学术记录的原始版本通常以印刷页面的形式出版，这意味着这类资源没有现成的电子版本。因此，很难获取这些学术资源的完全干净文本。此外，人工标注面临的上述挑战也存在于这一数据准备过程中。

尽管本文提出的噪声语料库构建策略可以作为应对上述挑战的另一种方法，但基于词典在干净文本中添加常见的词级OCR噪声，可能无法充分反映真实数字化图书馆馆藏的数据模式。这一局限可能涉及我们噪声语料库中的工程伪影，有待进一步研究。

6.2.3 评估方法

遵循以往研究的标准做法 [50,8],，我们将现有的热门预训练BERT模型用作现成工具，并使用我们的每个语料库进一步微调这些模型，以构建用于科学关系分类的分类器。我们的研究主要关注微调后的基于BERT的模型为分类器带来的益处。关于 BERT 或 SciBERT 预训练设置的特性及其对下游应用的影响，仍存在一些尚待解答的问题。例如：

（1）用于预训练 BERT 的语料库质量如何；（2）这一因素对用于识别科学关系的基于BERT的分类器有何影响？这些问题值得在未来进一步探讨。此外，本研究主要关注涵盖人工智能相关研究领域的语料库。对于其他领域中用于识别科学关系的分类器，还需要进一步研究。

6.3 潜在的未来工作

为了进一步帮助希望在学术 коллекции上构建结构化语义表征的数字图书馆设计者和图书馆员—科学关系分类器在学术论文上构建结构化语义表示的数字图书馆设计者和图书馆员，还有三个值得未来探索的主要方向。

6.3.1 重新思考人工标注指南

确保人工标注质量所面临的挑战表明，科学概念及其关系的定义过程仍是一个开放问题，专家可以对此进行进一步探索。对此领域感兴趣的研究人员可以从两个方向开展后续研究。

一方面，正式理解特定研究领域中术语之间的语义关系，对于提高该领域学术信息抽取标注指南的清晰度至关重要。特别是，我们建议在考虑语义关系类型时，不仅要涵盖学术术语之间的语义关联，还应要求关系类型具有层次结构。

另一方面，如果设计标注指南的策略能够遵循最终应用目标，将会很有帮助。在实践中，构建学术知识图谱可用于组织单个或多个研究领域中的一般学术知识，也可用于管理具有特定侧重点的知识，例如科学著作贡献的演变。学术知识图谱的应用各不相同，相应的科学术语标注规则和预定义关系类型也可能不同。例如，图 9 所示的一般科学概念（如“方法”“途径”“系统”和“算法”）的标注，仅限于提供信息，以表明每项研究工作的具体贡献。

6.3.2 多领域基准语料库

尽管数字图书馆中的学术记录通常涵盖各种领域，但用于开发学术信息抽取技术的公开语料库主要集中在生物学 [29,35]和计算机科学 [31,16]等少数特定领域。鉴于不同领域的学术出版物可能涉及各种关系类型，而且词语选择、写作风格等语言表达方式也日益有所不同，因此亟需构建一个涵盖广泛领域学术记录的基准语料库，以便全面评估为抽取学术语义信息而开发的最先进关系分类技术。

6.3.3 对开放式信息抽取技术的评估

除了构建用于识别预定义关系的分类器之外，在开放信息抽取范式下开发的技术也能识别更多样的关系三元组，“无需任何针对特定关系的人工输入” [7] ，这可以成为抽取学术信息的一种替代但颇具前景的策略，尤其适用于从涵盖多个领域、且事先存在大量不确定语义关系的学术记录中识别科学关系元组。进一步考察这一领域用于科学关系识别的最先进技术，以及这些技术与基于 BERT 的预定义关系分类方法之间的权衡，或许是一个值得探索的方向。

7 结论

我们研究了科学关系分类任务，以支持基于数字图书馆馆藏构建学术知识图谱。我们全面评估了八种基于 BERT 的分类模型，涉及三个干净语料库；这些语料库在规模和标注指南方面各有特点。此外，考虑到现实世界中的数字图书馆有许多学术记录经过数字化处理时存在 OCR 错误，我们进一步准备了三个分别对应上述干净语料库的噪声语料库，并研究了 OCR 错误对基于干净数据确定的最优基于 BERT 的分类器的影响。本文所呈现的实证研究有助于数字图书馆相关方了解用于从学术出版物中识别语义关系的先进自然语言处理技术，从而为选择最优自然语言处理工具、构建数字图书馆馆藏的学术知识图谱带来实际益处。

我们的观察表明，分类器在干净文本上的表现主要与两个方面有关。首先，从训练算法的角度来看，包括分类策略、预训练语料库领域和词汇大小写在内的三个主要因素，决定了实践中应采用的最佳模型。其次，对于用于训练的科学关系标注，有两个关键因素会影响基于 BERT 的分类模型识别各类关系的能力：（1）每种关系类型的标注数量；（2）各类关系句法语境的规律性。进一步探究 OCR 噪声的影响后，我们发现，数字化造成的文本噪声会明显降低基于 BERT 的分类器识别科学关系的表现，尤其是在噪声比例较高时（例如 49%）。相比之下，USAGE 和 RESULT 等语义较为具体的关系，比其他关系更有利于提升分类器对 OCR 噪声的鲁棒性；而 COMP 和 HYPONYM-OF 等体现概念层级的关系，则更容易使分类器暴露出对噪声的脆弱性。

这些分类器对 OCR 噪声的鲁棒性；而 COMP 和 HYPONYM-OF 等体现概念层级的关系，则更容易使分类器暴露出对噪声的脆弱性。

本研究的总体发现表明，在 SRC 策略下构建的不区分大小写的、基于 SciBERT 的分类模型，通常是科学关系分类的最佳选择。对于含有 OCR 噪声的语料库，我们建议数字图书馆相关方使用噪声数据构建分类器，因为训练数据中 OCR 噪声的异质性有助于提升分类模型处理未见数据时的泛化能力。

致谢 作者谨向匿名审稿人致谢，感谢他们对本文提出的建设性意见。我们也感谢伊利诺伊大学厄巴纳-香槟分校的 Alaine Martaus、Jacob Jett 和 Yuerong Hu 帮助校对论文。

参考文献

1. Pubmed. https://pubmed.ncbi.nlm.nih.gov/ 2. Agichtein, E., Gravano, L.：Snowball：从大型纯文本语料库中抽取关系。载于：第5届ACM数字图书馆会议论文集。第85–94页（2000） 3. Ammar, W., Groeneveld, D., Bhagavatula, C., Beltagy, I., Crawford, M., Downey, D., Dunkelberger, J., Elgohary, A., Feldman, S., Ha, V.等：Semantic Scholar中的文献图谱构建。载于：2018年计算语言学协会北美分会会议：人类语言技术论文集，第3卷（行业论文）。第84–91页（2018） 4. Auer, S., Kovtun, V., Prinz, M., Kasprzik, A., Stocker, M., Vidal, M.E.：构建科学知识图谱。载于：第8届网络智能、挖掘与语义国际会议论文集。第1–6页（2018） 5. Auer, S., Mann, S.：迈向开放的知识研究图谱。《连续出版物图书馆员》**76**，1–7（2019） 6. Augenstein, I., Das, M., Riedel, S., Vikraman, L., McCallum, A.：SemEval 2017任务10：ScienceIE——从科学出版物中抽取关键短语和关系。载于：第11届国际语义评测研讨会（SemEval-2017）论文集。第546–555页（2017） 7. Banko, M., Etzioni, O.：开放式关系抽取与传统关系抽取之间的权衡。载于：计算语言学协会第46届年会：人类语言技术论文集。第28–36页（2008） 8. Beltagy, I., Lo, K., Cohan, A.：Sci BERT：面向科学文本的预训练语言模型。载于：2019年自然语言处理实证方法会议暨第9届自然语言处理国际联合会议论文集（EMNLP-IJCNLP）。第3615–3620页。ACL，香港，中国（2019年11月） 9. Center, H.R.：HathiTrust数字图书馆178,381册藏书的体裁专属词数统计 [v.0.1] (2015), https://wiki.htrc.illinois.edu/display/ COM/Word+Frequencies+in+English-Language+Literature%2C+1700-1922 10. Chakraborty, T., Kumar, S., Goyal, P., Ganguly, N., Mukherjee, A.：探索分层学习方法以预测未来引用次数。载于：IEEE/ACM数字图书馆联合会议。第351–360页。IEEE（2014）

11. Culotta, A., Sorensen, J.：用于关系抽取的依存树核。载于：计算语言学协会第42届年会论文集（ACL-04）。第423–429页。ACL（2004） 12. Davletov, F., Aydin, A.S., Cakmak, A.：利用时间特征和拓扑特征预测高影响力学术论文。载于：第23届ACM信息与知识管理国际会议论文集。第491–498页（2014） 13. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.：BERT：用于语言理解的深度双向变换器预训练。载于：2019年计算语言学协会北美分会会议：人类语言技术论文集，第1卷（长短论文）。第4171–4186页。ACL，明尼阿波利斯，明尼苏达州（2019年6月） 14. Esakov, J., Lopresti, D.P., Sandberg, J.S.：光学字符识别错误的分类与分布。载于：《文档识别》。第2181卷，第204–216页。国际光学与光子学学会（1994） 15. Feng, J., Huang, M., Zhao, L., Yang, Y., Zhu, X.：利用强化学习从噪声数据中进行关系分类。载于：第三十二届AAAI人工智能大会论文集。第32卷，第5779–5786页（2018） 16. G´abor, K., Buscaldi, D., Schumann, A.K., QasemiZadeh, B., Zargayouna, H., Charnois, T.：SemEval-2018任务7：科学论文中的语义关系抽取与分类。载于：第12届国际语义评测研讨会论文集。第679–688页（2018） 17. Gao, T., Han, X., Liu, Z., Sun, M.：用于噪声小样本关系分类的混合注意力原型网络。载于：第三十三届AAAI人工智能大会论文集。第33卷，第6407–6414页（2019） 18. Gonc.alves, G.D., Figueiredo, F., Almeida, J.M., Gonc.alves, M.A.：学者受欢迎程度特征分析：计算机科学研究社区的一项案例研究。载于：IEEE/ACM数字图书馆联合会议。第57–66页。IEEE（2014） 19. Hallo, M., Luj´an-Mora, S., Mat´e, A., Trujillo, J.：数字图书馆中关联数据的现状。《信息科学杂志》**42**(2)，117–127（2016） 20. Hashemi, S.H., Neshati, M., Beigy, H.：书目网络中的专长检索：一种主题主导性学习方法。载于：第22届ACM信息与知识管理国际会议论文集。第1117–1126页（2013） 21. Haslhofer, B., Isaac, A., Simon, R.：图书馆与数字人文领域的知识图谱。Sakr S., Zomaya A.（编）《大数据技术百科全书》。第1–8页（2018） 22. Hoffmann, R., Zhang, C., Ling, X., Zettlemoyer, L., Weld, D.S.：用于重叠关系信息抽取的基于知识的弱监督。载于：计算语言学协会第49届年会：人类语言技术论文集。第541–550页（2011） 23. Jaradeh, M.Y., Oelen, A., Farfar, K.E., Prinz, M., D'Souza, J., Kismih´ok, G., Stocker, M., Auer, S.：开放研究知识图谱：新一代语义学术知识基础设施。载于：第10届知识捕获国际会议论文集。第243–246页。ACM，纽约，美国（2019） 24. Jett, J., Cole, T.W., Han, M.J.K., Szylowicz, C.：用于图书馆特藏的关联开放数据（LOD）。载于：2017年ACM/IEEE数字图书馆联合会议（JCDL）。第1–2页。IEEE（2017） 25. Jia, W., Dai, D., Xiao, X., Wu, H.：ArnOR：基于注意力正则化的远程监督关系分类噪声消减方法。载于：计算语言学协会第57届年会论文集。第1399–1408页（2019） 26. Jiang, M., Diesner, J.：一种基于成分句法分析树的学术出版物摘要关系抽取方法。载于：《第十三届基于图的方法自然语言处理研讨会论文集》（TextGraphs-13）。第186–191页（2019年） 27. Jiang, M., Hu, Y., Worthey, G., Dubnicek, R.C., Capitanu, B., Kudeki, D., Downie, J.S., 等：Gutenberg-HathiTrust平行语料库：用于研究未经校正的光学字符识别文本中噪声的真实世界数据集。在iConference 2021上的海报（2021年），http://hdl.handle.net/2142/109695 28. Klampfl, S., Kern, R.：从数字科学文章中提取正文和目录的无监督机器学习方法。在：数字图书馆理论与实践国际会议。第144–155页。Springer（2013年） 29. Kruiper, R., Vincent, J.F., Chen-Burger, J., Desmulliez, M.P., Konstas, I.：面向仿生工程的科学信息抽取数据集。arXiv预印本 arXiv:2005.07753（2020年） 30. Luan, Y., He, L., Ostendorf, M., Hajishirzi, H.：面向科学知识图谱构建的实体、关系和指代消解多任务识别。在：2018年自然语言处理实证方法会议论文集。第3219–3232页（2018年） 31. Luan, Y., Wadden, D., He, L., Shah, A., Ostendorf, M., Hajishirzi, H.：使用动态跨度图进行信息抽取的通用框架。在：2019年自然语言处理协会北美分会会议论文集：人类语言技术，第1卷（长篇和短篇论文）。第3036–3046页（2019年6月） 32. Manning, C.D.：《计算语言学》与深度学习。《计算语言学》**41**(4)，701–707（2015年） 33. Mintz, M., Bills, S., Snow, R., Jurafsky, D.：无需标注数据的关系抽取远程监督。在：ACL第47届年会与AFNLP第4届自然语言处理国际联合会议联合会议论文集。第1003–1011页（2009年） 34. Mohapatra, D., Maiti, A., Bhatia, S., Chakraborty, T.：拓宽视野，深入分析：通过影响力扩散树量化科学论文的影响。在：2019年ACM/IEEE数字图书馆联合会议（JCDL）。第305–314页。IEEE（2019年） 35. Nasar, Z., Jaffry, S.W., Malik, M.K.：科学文章中的信息抽取：综述。《科学计量学》**117**(3)，1931–1990（2018年） 36. Penfield, T., Baker, M.J., Scoble, R., Wykes, M.C.：研究影响力的评估、评价与定义：综述。《研究评估》**23**(1)，21–32（2014年） 37. Pradhan, T., Pal, S.：基于多层融合的学术合作者推荐决策支持系统。《知识库系统》**197**，1–23（2020年） 38. Quan, T.T., Hui, S.C., Fong, A.C.M., Cao, T.H.：学术语义网本体的自动生成。在：语义网国际会议。第726–740页。Springer（2004年） 39. Riedel, S., Yao, L., McCallum, A.：无需标注文本对关系及其提及进行建模。在：欧洲机器学习与数据库知识发现联合会议。第148–163页。Springer（2010年） 40. Saggion, H., Ronzano, F.：学术数据挖掘：理解科学文献。在：2017年ACM/IEEE数字图书馆联合会议（JCDL）。第1–2页。IEEE（2017年） 41. Silvescu, A., Reinoso-Castillo, J., Honavar, V.：从异构、分布式、自主的生物数据源中进行本体驱动的信息抽取与知识获取。在：IJCAI-2001关于从异构、分布式、自主、动态数据与知识源中进行知识发现的研讨会论文集。第1–11页（2001年） 42. Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.J., Wang, K.：微软学术服务（MAS）及其应用概述。在：第24届万维网国际会议论文集。第243–246页（2015年）

43. Sivasubramaniam, A., Debnath, S., Li, H., Lee, W.C., Bolelli, L., Giles, C.L., Zhuang, Z., Councill, I.G.: 在在线引文匹配方案中从证据中学习元数据。载于：第六届 ACM/IEEE-CS 数字图书馆联合会议论文集。第276–285页。IEEE（2006） 44. Soergel, D.: 数字图书馆与知识组织。载于：语义数字图书馆，第9–39页。Springer（2009） 45. van Strien, D., Beelen, K., Ardanuy, M.C., Hosseini, K., McGillivray, B., Colavizza, G.: 评估 OCR 质量对下游自然语言处理任务的影响。载于：第12届智能体与人工智能国际会议论文集——第1卷：ARTIDIGH。第484–496页（2020） 46. Tahamtan, I., Afshar, A.S., Ahamdzadeh, K.: 影响引用次数的因素：文献综合评述。《科学计量学》**107**(3)，1195–1225（2016） 47. Tang, J., Wu, S., Sun, J., Su, H.: 跨领域合作推荐。载于：第18届 ACM SIGKDD 知识发现与数据挖掘国际会议论文集。第1285–1293页（2012） 48. Traub, M.C., Van Ossenbruggen, J., Hardman, L.: OCR 质量对数字档案研究任务的影响分析。载于：数字图书馆理论与实践国际会议。第252–263页。Springer（2015） 49. Vahdati, S., Palma, G., Nath, R.J., Lange, C., Auer, S., Vidal, M.E.: 揭示知识图谱中的学术社群。载于：数字图书馆理论与实践国际会议。第103–115页。Springer（2018） 50. Wang, H., Tan, M., Yu, M., Chang, S., Wang, D., Xu, K., Guo, X., Potdar, S.: 使用预训练 Transformer 单次提取多种关系。载于：第57届计算语言学协会年会论文集。第1371–1377页。ACL，意大利佛罗伦萨（2019年7月） 51. Wang, W., Xia, F., Wu, J., Gong, Z., Tong, H., Davison, B.D.: Scholar2vec：用于预测终身合作者的学者向量表示。《ACM知识发现数据汇刊（TKDD）》**15**(3)，1–19（2021） 52. Weigl, D.M., Kudeki, D.E., Cole, T.W., Downie, J.S., Jett, J., Page, K.R.: 合并还是连接：查询图书馆关联数据的实践经验。载于：信息科学与技术协会论文集。第56卷，第296–305页。Wiley 在线图书馆（2019） 53. Yang, K., He, L., Dai, X., Huang, S., Chen, J.: 利用噪声数据进行远程监督关系分类。载于：2019年计算语言学协会北美分会会议：人类语言技术论文集，第1卷（长文与短文）。第3216–3225页（2019） 54. Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., Xu, B.: 用于关系分类的注意力双向长短期记忆网络。载于：第54届 ACL 年会论文集（第2卷：短文）。第207–212页（2016年） 55. Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., Fidler, S.: 对齐书籍与电影：通过观看电影和阅读书籍实现类故事的视觉解释。载于：IEEE 国际计算机视觉会议论文集。第19–27页（2015） 56. Zuo, Z., Zhao, K.: 从社会网络视角理解和预测不同职业阶段的未来研究影响力。《信息科学与技术协会期刊》**72**(4)，454–472（2021）