doi:10.3772/j.issn.2095-915x.2021.05.001

基于 SciBERT 模型的引文上下文识别系统优化

郭晨睿[1] 王佳敏[1] 崔浩冉[1] 武健[2]

1. 武汉大学信息管理学院 武汉 430072；
2. 欧道明大学计算机科学系 诺福克 23529

开放科学（资源服务）标识码（OSID）

摘要：[目的 / 意义] 本文介绍一个用于从英文学术文献中提取显性引文句和隐性引文上下文的系统。该系统基于 SmartCiteCon（SCC）系统对识别模型进行了优化，本文称该系统为 SmartCiteCon2.0（SCC2.0）。[方法 / 过程] 该系统的识别模型部分通过 Python 语言编写，基于经过微调的 SciBERT 模型构建而成。该模型基于国际计算语言学协会（ACL）会议论文集中人工标注的 27,832 个引用上下文句子对进行训练，并以 SCC 系统使用的 SVM 模型作为对照组以验证效果。[结果 / 结论] 实验结果表明，微调后的 SciBERT 模型的 F1 值为 90%。相比仅使用句子对文本特征的 SVM 模型 F1 值提升了 11%，相比于使用全部特征 SVM 模型的 F1 提升了 3%。在该模型基础上，本文构建的 SCC2.0 系统开箱即用，支持 PDF，以及符合特定标准的 XML 和 JSON 格式的学术文献。该 API 同时支持单文档处理和多文档并发处理，源代码发布于 https://gitee.com/Lan_Tianchen/smart-cite-con。

关键词：引文识别；隐性引文上下文；SciBERT；引文识别系统

中图分类号：TP311.52；G35

作者简介 郭晨睿（1995-），硕士研究生，研究方向为文本挖掘与信息检索；王佳敏（1992-），博士研究生，研究方向为文本挖掘与知识网络，E-mail: wangjm@whu.edu.cn；崔浩冉（1999-），硕士研究生，研究方向为文本挖掘与信息检索；武健（1981-），博士，助理教授，研究方向为文本挖掘与机器学习。

引用格式 郭晨睿 , 王佳敏 , 崔浩冉 , 等 . 基于 SciBERT 模型的引文上下文识别系统优化 [J]. 情报工程 , 2021, 7(5): 3-14.

doi:10.3772/j.issn.2095-915x.2021.05.001

Optimization of Citation Context Recognition System Based on SciBERT Model

GUO Chenrui[1] WANG Jiamin[1] CUI Haoran[1] Wu Jian[2]

1. School of Information Management Wuhan University, Wuhan 430072, China;
2. Department of Computer Science Old Dominion University, Norfolk 23529, USA

Abstract: [Objective/ Significance] This article introduces a system for extracting explicit citation sentences and implicit citation context from English academic literature. The system is based on the SmartCiteCon (SCC) system to optimize the recognition model. This article calls the improved system SmartCiteCon2.0 (SCC2.0). [Methods/Process] The recognition part of the system is written in Python and constructed based on the fine-tuned SciBERT model. The model is trained based on 27,832 pairs of citing context sentences manually annotated in ACL Anthology, and the SVM model is used as a control group to verify the effect. [Results /Conclusions] The experimental results show that the F1 value of the fine-tuned SciBERT model is 90%. Compared with the SVM model using only sentence-to-text features, the F1 value is increased by 11%, and compared with the F1 using all-feature SVM model, it is increased by 3%. Based on this model, the SCC2.0 system constructed in this paper can be used out of the box, supporting PDF, as well as academic documents in XML and JSON formats that meet specific standards. The API supports both single-document processing and multiple-document concurrent processing. The source code is published at https://gitee.com/Lan_Tianchen/smart-cite-con.

Keywords: Citation recognition; implicit citation context; SciBERT; citation recognition system

引言

引文上下文是指围绕在引证文献的引文标记周围用来描述被引文献的一个或多个句子[1]。这些句子表征了被引文献有代表性价值的观点和方法。对引文上下文进行分析对了解学术文献引证时的引文习惯、被引原因、分析论文信息流向特点等有着重要价值。在引文分析领域，引文上下文使得仅依赖计量学所产生的引文分析偏差得到修正，是引用动机识别[2]、被引片段识别[3, 4]、被引文献自动摘要生成[5]等研究的基础。随着机器学习技术的飞速发展，引文上下文也在学术文献检索优化[6]、主题识别[7]、增强和改进关键字短语提取[8]等研究中发挥重要价值。

引文上下文分为两种类型，其中显性引文上下文是包含引用标记的句子[9]，隐性引文上下文是与被引文章在语义上相关但不包含引文标记的句子[10]。在图 1 的示例中，包含引用标记 “(Vogel et al., 1996)” 的显性引文上下文以绿色突出显示，隐性引文上下文句子以黄色突出显示，非突出显示的句子不是给定引用的引用上下文。

ParsCit[11]、GROBID[12]、ConSyn①等工具可以帮助研究者识别显性引文上下文。但多数情

① http://consyn.elsevier.com

况下，作者都会使用多个句子来总结对被引文献的评述。同时，很多学者在研究中发现使用完整的引文上下文能够很大程度上提升下游任务的效果。例如，Ritchie 等 [13] 发现与仅使用显性引文上下文相比，使用完整的引文上下文在文献检索任务上能取得更好效果。

The more reasonable constraints are imposed on this process, the easier the task would become. For instance, the most relaxed IBM Model-1, which assumes that any source word can be generated by any target word equally regardless of distance, can be improved by demanding a Markov process of alignments as in HMM-based models (Vogel et al., 1996), or implementing a distribution of number of target words linked to a source word as in IBM fertility-based models (Brown et al., 1993). Following the path, we shall put more constraints on word alignment models and investigate ways of implementing them in a statistical framework.

图 1 显性和隐性引文上下文示例

造成引文上下文提取工具缺乏的原因重要有以下三点。首先，学术文献数据往往是多源异构的，不同类型和文件格式的文献处理方式和难度都不同；其次，引文分析等任务往往需要同时处理大批量的学术文献；最后，与带有引用标记的显性引文上下文句子不同，缺少引用标记使隐性引用上下文识别具有一定的挑战性，仅通过简单的正则表达式无法实现对隐性引文上下文的识别。受限于引文上下文识别过程中的诸多瓶颈，科研工作者花费了大量的时间却不一定能获得有一定质量的引文上下文数据 [14]。SCC 引文上下文识别系统 [15] 一定程度上解决了上述问题。

SCC 系统采用了人工构造特征基于 SVM 模型进行文本分类的传统方法。传统方法中人工进行特征工程成本较高，特征表达能力弱，所获得模型往往效果有限，且缺乏拓展性。不同于传统的基于特征工程的方法，预训练语言模型在文本表示能力上效果更优，可以自动获取低维度、高密度的文本特征，且经过大型语料库上的预训练后会包含一定先验知识，可在其他自然语言处理任务上进行迁移学习。

基于上述背景，本研究对 SCC 系统核心识别算法进行了优化，开发了 SmartCiteCon2.0（SCC2.0）系统。该系统基于预训练语言模型 SciBERT 设计，将显性和隐性引文上下文的识别相结合，可以对 PDF、XML 和 JSON 等多种格式的文献进行处理，支持批处理等多种处理模式。可以有效地为引文分析、自动摘要、文献检索等引文上下文的相关任务提供支持，以弥补传统工具在隐性引文上下文识别上的不足。

1 相关研究

1.1 引文上下文的识别方法

研究人员在引文上下文识别方法上已有很多探索。1999 年 Nanba 等 [16] 将引文上下文的范围定义为带有引文标记的句子（即引文句）前后的几个连续句子。在另一项工作中，研究者使用了马尔可夫模型来识别引文上下文 [17]。2010 年 Sugiyama 描述了一种支持向量机（SVM）和最大熵（ME）模型，通过使用一些浅层特征（例如专有名词等）来对引文句前后的句子进行分

doi:10.3772/j.issn.2095-915x.2021.05.001

类以识别引文上下文[18]，其训练样本中的正例是通过使用正则表达式匹配含有引文标记的句子，并不包括隐性引文上下文。2016 年雷声伟等[10]对引文上下文的特征进行了细致的总结和分析，使用 CRF 模型和 SVM 模型对引文上下文进行识别，发现基于文本分类思想的 SVM 模型效果要优于基于序列标注思想的 CRF 模型的实验效果。虽然基于特征工程的有监督学习算法，已经在引文上下文识别上取得一定效果。但是受限于自然语言的复杂性，该方法难以涵盖引文上下文的全部特征。

1.2 引文上下文识别工具

引文上下文识别工具为科研人员提供了工具支持。2008 年开发的 ParsCit 就是一款用于引文解析和引文上下文提取的开源软件[11]，其使用条件随机场（CRF）模型来解析引文字符串，通过提取引用标记的任一侧固定窗口长度的字符串（默认为 200 个字符）作为引文上下文。2009 年开发的 GROBID 是一个从学术文献中提取信息的工具[12]，该工具在显性引文句子解析的 F1 值约为 75%，它既可以正确识别引文标记，又可以将其与参考文献列表正确关联。Elsevier 自 2011 年以来开始提供 XML 格式的论文，并提供了 ConSyn 工具以识别和提取含有引文标记的引文句。2013 年 Angrosh 等使用词汇特征基于 CRF 技术开发了 CitContExt 工具[19]，该工具支持对隐性引文上下文的识别，但其使用的模型仍基于传统的人工特征构建，且仅支持纯文本类型文献，不支持批处理功能。综合来看，多数的引文上下文提取工具都侧重于提取显性引文上下文，对隐性引文上下文的关注不足，但是后者亦包含了与被引论文在语义上相关的信息。由于受到不同标准和文件格式的限制，多数工具仅支持符合特定标准的纯文本类型的论文数据。

1.3 预训练语言模型

预训练技术将编码好的数据输入到预先设计好的深度网络结构中进行训练，以提升模型的泛化能力。经过预训练的模型含有大量的先验知识，可以基于其在下游任务进行微调，无须从零开始训练。2000 年，Alex 等人尝试将神经网络引入到语言模型中，并创造性地提出了词向量的概念，实验表明他们设计的 NNLM 模型相较于 N-gram 模型有更好的性能[20]。Collobert 等[21]发现在未标记的数据中嵌入预训练的词可以明显提升许多 NLP 任务的效果。Word2Vec 模型正是基于词嵌入技术提出的，与 NNLM 模型相比该模型更多地利用了词的上下文信息，但无法解决一词多义的问题。ELMo 模型[22]采用双向 LSTM 进行预训练，可以结合上下文的语境对词进行建模，很好地解决了一词多义、句法结构理解等问题。BERT 模型[23]将无监督学习的预训练融合有监督学习的微调的模式推广到了更深层次的双向结构中，在多项 NLP 任务中取得较好表现。SciBERT[24]模型基于 BERT 模型的结构，使用大量的多领域学术文献作为无监督预训练语料，在基于学术文本的序列标记、句子分类等任务中取得了比 BERT 更好的效果。

综上所述，引文上下文是引文标记周围描述被引文献的句子集合，既应包含显性引文上下文也应包含隐性引文上下文。当前主流的识

别工具侧重于识别显性引文上下文，忽视了隐性引文上下文的识别，隐性引文上下文识别工具的缺失影响了下游任务发挥更好的效果。预训练语言模型近年来在文本表示和分类等任务中取得较好表现，但基于这类模型进行隐性引文上下文识别的研究还较少。因此，本文基于预训练语言模型 SciBERT 设计了一个能够识别显性和隐性的引文上下文，且支持多种标准和文件格式论文的引文上下文识别工具，为引文上下文及相关下游任务的研究者提供支持。

2 引文上下文识别实验

2.1 任务特点分析

引文上下文识别是句子级别的分类任务。句子级的分类任务有两种类型：（1）基于句子对的分类任务；（2）基于单个句子的分类任务。基于预训练语言模型的引文上下文识别任务应属于句子对分类任务，下面做进一步的说明。

传统的引文上下文识别方法需要进行大量的特征工程工作，这为总结引文上下文的特征打下了良好的基础。2016 年雷声伟 [10] 通过总结前人研究和进行实例分析较为完整地总结了五大类 19 种引文上下文的特征。在传统基于特征工程的识别方法中，引文上下文识别任务被定义为：构建适合的模型，使得在给定的目标引文标记 XREF，及给定候选上下文句集合 CANDIDATE_SET=$\{s_1, s_2, \cdots\cdots, s_n\}$ 情况下，可以给 CANDIDATE_SET 中每个句子一个标记 label，如果 s_i(∈CANDIDATE_SET) 属于目标引文的上下文，则 label 为 1，否则为 0。该定义并没有明确说明引文上下文识别任务是基于句子对的分类任务还是基于单个句子的分类任务，在基于预训练语言模型的方法中需要对该问题进行进一步的明确。因此，如表 1 笔者将上述 19 个特征从句子对分类的视角重新进行归纳总结。

引文上下的特征主要分为句子特征，句间关系特征，引用关系特征和篇章或结构特征四种类型的特征。句子特征是指描述句子本身所含有信息以及语法语义的特征。句间关系特征是指描述当前句子与另一句相关连的语法语义的特征。引用关系特征是指当前句与被引文献信息相关连的语法语义特征。篇章与结构特征是指当前句与章节相关的位置或语法信息。

传统特征工程所使用的 19 类特征，与上述四类特征的对应关系如表 1。可以看出，传统特征工程虽然将引文上下文分类看作基于单个句子的分类任务，实际上是将许多句子间关系特征抽象到了单个句子上。引文上下文更多地是一种引文句和隐性引文上下文的关系，因此引文上下文识别应当看作基于句子对的分类任务。

同时，上述四大类特征可以进一步被划分为句子级文本特征与篇章或结构特征。受限于文本特征的表达的能力，传统的特征工程方法需要引入非常多篇章结构特征来提升分类效果。辩证来看，相对于传统特征工程方法，预训练语言模型的劣势在于不能直接对篇章或结构特征进行建模，但优势在于其对文本特征的表达能力更强，除了表 1 中包含的句子级特征外，其还能表达更丰富和多层次的文本语义特征。

doi:10.3772/j.issn.2095-915x.2021.05.001

表 1 句子对分类视角下的引文上下文特征归类

类型	特征
句子语义特征	候选上下文句中是否只包含其他引文标记
	引文句中的引文标记个数
	句子对中包含了Work Nouns词组。
	句子对的内容相似度
句子对语义特征	候选上下文句是否起始于指定的连接副词
	候选上下文句是否包含第三人称代词
	候选上下文句是否包含引文句中目标引文标记前面相邻的名词短语
引用关系特征	候选上下文句中包含被引文献作者的名字
	候选上下文句包含Lexical hooks 词表中的词
	句子对在文章中的距离
	句子对是否在同一个段落内
	候选上下文句是否为所在章节的第一句
篇章或结构特征	候选上下文句是否为所在章节的最后一句
	候选上下文句的前一句是否为章节的第一句
	候选上下文句前面一句是否是非目标引文句
	候选上下文句后面一句是否是非目标引文句
	句子在文章中的区域

综合前述的分析，本文所定义的引文上下文识别任务为：构建适合的模型，使得在给定含有目标引文标记的引文句 *csent*，及给定候选上下文句集合 $\text{CANDIDATE}_{\text{SET}}=\{sent_1, sent_2, \cdots\cdots, sent_n\}$ 情况下，判定 CANDIDATE_SET 中任意句子 $sent_i$ 与引文句 *csent* 所构成的句子对 (csent, sent_i) 是否存在引文上下文关系 f，如果句子对 (csent, sent_i)，$s_i(\in\text{CANDIDATE}_{\text{SET}})$ 存在引文上下文关系，则 $f(csent, sent_i)=1$，否则 $f(csent, sent_i)=0$。

2.2 数据集

为了验证本研究所使用的预训练语言模型在引文上下文识别任务上的有效性，本研究 SCC 系统中使用的 SVM 模型 [10]（下文简称：Lei_SVM）作为基准研究，并对其进行了复现。同时，为了尽可能地保证实验有较强的对比价值，本研究所使用预训练语言模型也基于 Lei_SVM 模型所使用的数据集（下文简称：Lei_SVM 数据）进行训练。

该数据来自国际计算语言学协会（Association for Computational Linguistics，ACL）数据集，该协会是计算语言学（CL）和自然语言处理（NLP）领域最重要的国际学术组织，包含 34,000 篇 PDF 格式会议论文。使用 OCR 技术将原始 PDF 文件转换为包含完整参考文献、段落和章节信息的 XML 格式 [25]。Lei_SVM 模型随机选取了其中 130 篇文献，对 XML 数据进行了清洗、句子分段、引文标签识别与标注等工作。后由 13 名信息管理相关专业的研究生对

该数据集进行了引文上下文标注，标注结果已通过 Cohen 提出的 Kappa 系数（κ= 0.937）进行了测试。

2.2.1 数据清洗

由于 Lei_SVM 数据是由 PDF 文件通过 OCR 技术识别获得的，该数据存在部分数据缺失和异常问题。Lei_SVM 模型的特征是人工构造的，包含较多非语义层面的特征且相对稀疏，因此部分数据的缺失和异常并不会对实验效果产生很大的影响。但是，预训练语言模型对语义层面的特征进行建模，且特征相对稠密，数据集的质量将对实验效果产生较大的影响。因此，笔者对全量数据集进行了人工复查，并对有缺失问题的数据进行了补全，对异常问题的数据进行了修改或删除。在对全量数据的人工复查中，共发现 ref_num 属性缺失项 173 个，人工匹配到 45 个，删除非引文标记的 < ref/ > 标签 21 个。

2.2.2 预处理

为了更加完整地复现 Lei_SVM 模型，本研究参考其预处理流程对数据进行了预处理。Lei_SVM 模型的预处理流程依次为节点过滤、句子切分、引文标记处理、分词、非句法成分替换和词性标注。本研究在对 Lei_SVM 模型复现时对原预处理流程做了部分调整，这些调整包括：使用 Standford Core NLP[②]工具替代原有的正则表达式和 Stanford Parser 进行分词和词性标注，效果相对更优；用括号替换流程中识别转义字符。由于 Standford Core NLP 工具在分词后，会将左括号转义为“-LRB-”，右括号转义为“-RRB-”，这会导致原预处理流程中括号替换失效，因此需要用括号替换识别“-LRB-”和“-RRB-”两个转义字符。

由于预训练语言模型的输入为纯文本信息，因此将 Lei_SVM 数据应用于预训练语言模型的预处理流程相对简单，仅需要进行节点过滤、句子对匹配和非句法成分处理操作即可。

2.2.3 数据采样

相关理论表明，隐性引文上下文信息主要出现在以引文句为中心的前后四句的范围内[26]。Lei_SVM 模型和本研究在数据采样的过程中均沿用了上述方法。Lei_SVM 模型的过程包含了特征识别工作，采样为结果含有 19 个特征及其对应值的 libsvm 格式文件，而预训练语言模型的采样结果为包含句子对的 CSV 格式文件。为保证 Lei_SVM 模型和预训练语言模型所对应的原始语料的一致性，两者的数据采样是同步进行的。采样后的数据按 8:1:1 的比例划分为训练集、评估集和测试集，为保证三个集合中正例和负例数据比例的相对一致，在切分数据集前对数据进行了随机化，同时为保证不同的数据集对应的原始语料的一致性，随机化种子 SEED 固定为 100。为保证采样结果中正负样本的比例与实际样本分布一致，本研究未对采样后正负样本数量进行均一化，采样结果如表 2 所示。

表 2 数据采样结果

	正例	负例	总计
训练集	5789	16475	22264
评估集	719	2064	2783
测试集	741	2044	2785
总计	7249	20583	27832

② https://stanfordnlp.github.io/CoreNLP/

doi:10.3772/j.issn.2095-915x.2021.05.001

2.3 实验过程

2.3.1 实验环境

本研究进行引文上下文识别的实验环境如表 3 所示。

表 3 实验环境

配置项	配置参数
CPU	AMD Ryzen 9 5900X 12-Core Processor 24 线程
内存	32GB
硬盘	2TB
显卡	Nvidia GeForce RTX 3090 24G显存
操作系统	Ubuntu 20.04.2 LTS
WEKA	3.8.5
LibSVM	1.0.10
CUDA	CUDA toolkit v11.1.74
Python	3.9.1
Pytorch	1.7.1
Simple Trans-formers	0.60.9

2.3.2 实验结果分析

为了验证预训练语言模型在引文上下文识别任务上的有效性，本研究以 Lei_SVM 模型作为基准研究，在相同的数据集上进行同步采样。由于预训练语言并未对篇章或结构特征进行表达，因此本研究使用了 Lei_SVM_10 和 Lei_SVM_19 两组模型做对照组。Lei_SVM_10 模型训练时仅包含了表 1 中的句子级文本特征，Lei_SVM_19 训练时包含了全部的特征。

实验结果通过召回率（recall，R）、准确率（precision，P）和调和平均值（F1-score，F1 值），以及按照正负样本比例加权后的相应 Weight 值作为评价指标。实验结果如表 4 所示。

表 4 实验结果

model	Lei_Svm_10	Lei_Svm_19	Scibert
P	92%	88%	93%
R	75%	87%	87%
F1	79%	87%	90%
Weight-P	89%	90%	92%
Weight-R	87%	90%	92%
Weight-F1	85%	90%	92%

从表 4 可以看出，SciBERT 模型在 P、R、F1 三个指标上全面超越两个 SVM 模型。相比于 Lei_Svm_10 模型，F1 上提升了 11%，Weight-F1 值提升了 7%。这说明在仅使用句子对文本特征情况下，SciBERT 模型的分类效果大幅度高于 SVM 模型。表明 SciBERT 预训练模型有更好的文本特征表示能力，深度学习的神经网络捕获了更多隐含的尚未被人为总结的特征，并且这些特征有效地提升了对引文上下文关系的识别效果。同时，相比于 Lei_Svm_19 模型，F1 提升了 3%，Weight-F1 提升了 2%，表明使用句子对文本特征的 SciBERT 模型较使用全部特征的 SVM 模型也有部分提升。综合来看，基于 SciBERT 的引文上下文识别模型比传统 SVM 模型效果更好，验证了本文提出方法的有效性，为隐性引文上下文识别提供了新的有效模型和方法，也为类似的任务提供了一定的参考价值。

3 系统思路与构建

在前述 SciBERT 模型的基础上，本研究设计了 SCC2.0 系统，该系统在 SCC 系统 [15] 基础上进行加强，在识别效果和效率上相对于 SCC 系统都有较大的提升，且提供了简单易用的图

形化界面服务。

3.1 系统架构

如图 2 所示，SCC2.0 通过四个流程完成引文上下文的识别工作：（1）文件类型识别；（2）预处理；（3）特征提取；（4）句子分类。输出是一个包含显性引文上下文和隐性引文上下文以及其他与引用相关的信息的 JSON 文件，该 JSON 文件的结构可在源代码 Readme 文件中查看。该工具基于 Springboot 框架，采用 Java 语言编写而成，识别模型部分是基于 Simple Transformers[③]框架实现的。

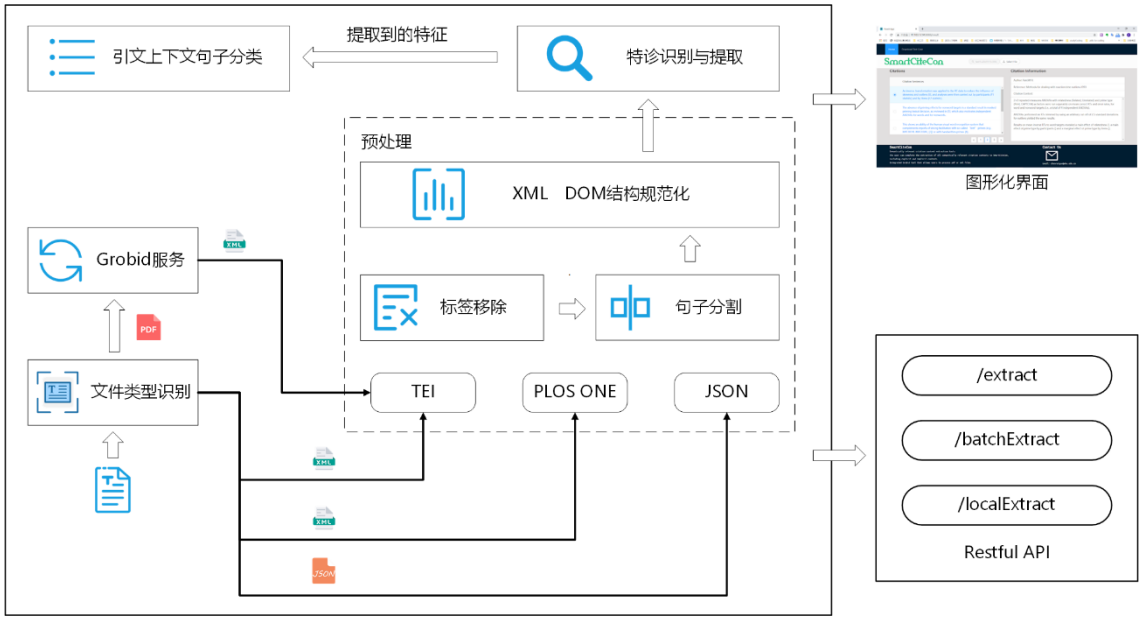

图 2 SCC2.0 工具系统框架图

文件类型识别和预处理模块基于 SCC 系统设计，支持 PDF、符合 TEI 或 PloS ONE 标准的 XML 和 Semantic Scholar 发布的 S2ORC 标准 [27] 的 JSON 格式的文献。对多源异构数据的处理，主要依赖于预处理流程。预处理流程将不同的文件类型规范为统一的 XML 格式用于后续的特征提取工作。SCC2.0 使用 Standford Core NLP 工具替代 SCC 的正则表达式和 Stanford Parser 进行分词和词性标注。各预处理器虽然针对的文件类型不同，但主要的处理步骤是类似的，主要包括标签移除、句子分割和 DOM 结构规范化。

SCC2.0 系统使用的特征提取算法是基于第 2 小节中获得的 SciBERT 模型，该模型将句子分类为隐性引文上下文和显性引文上下文。此模块从规范化的 XML 文件中提取出句子对，通过内置 Restful 接口将句子对信息传入 Simple Transformer 框架进行分类，并输出指示句子对是否为隐性引文上下文关系的结果。输出的 JSON 文件包含引用标记及其位置，引用语句和归类为隐性引文上下文的语句。

③ https://simpletransformers.ai/

doi:10.3772/j.issn.2095-915x.2021.05.001

3.2 系统性能

本研究复现了SCC测试系统的性能的实验。在表3的所示的设备环境下测试SCC和SCC2.0系统。分别用同样的PDF、XML、JSON格式的文献各10篇，在4GB堆内存线程数Np=8和Np=1的条件下运行系统并记录平均处理单篇文献的时间，结果如表5所示。可以看出，SCC2.0系统性能在处理各类文档的效率上相对于SCC提升约10倍。

表5 SCC和SCC2.0在Np=8和Np=1时平均处理单篇文献所需时间（秒）

	XML		PDF		JSON	
	MAX	MIN	MAX	MIN	MAX	MIN
SCC	164	43	177	45	39	12
SCC2.0	1.44	0.91	7.10	2.21	0.59	0.49

3.3 系统应用

为方便研究人员基于不同的需求使用本系统，本研究既开发了便于集成和部署的Restful API服务，同时也提供了图形化的使用界面。Restful API接口支持单文档、批处理、本地超大批量处理三种处理模式。在单文档和批处理提取模式下，API将返回含有引文上下文的JSON对象和执行状态。对于本地提取模式，API将返回执行状态，并且结果将保存在JSON文件中。图形化交互界面的实现采用的React前端框架实现，主界面包含导航栏、文件检索框、引文句列表栏、引文上下文列表栏。导航栏提供了便于用户测试的案例文献，用户可以点击“Download Test Case”链接下载案例文献。使用图形化交互界面时仅需点击检索框，上传需要处理的文献。上传后，前端会向后台请求/extract服务，后台会将识别引文上下文的结果返回前端。

用户可以非常便捷地在本地计算机上部署SCC2.0，部署方法可以参考源代码中Readme文件。部署后可通过接口模式使用本研究所提供的引文上下文识别服务。本研究的源代码发布于https://gitee.com/Lan_Tianchen/smart-cite-con，同时用户也可通过http://47.117.112.104:8090/使用在线服务。

如图3所示，SCC2.0对文献*Guiding Statistical Word Alignment Models With Prior Knowledge.XML*进行分析后，页面左侧的“Citation Sentences”列表中展示分析后获得的引文句列表。用户点击其中任意一句引文句后“Citation Information”列表中展示对应引文的作者、参考文献标题和引文上下文信息。图3示例中获得了两句引文上下文句子，其中一句为引文当前句，另一句为隐性引文上下文。可以看出，引文句陈述了“研究表明，人类通过标注的数据可以显著提升模型的性能”的研究事实，所识别出的隐性引文上下文包含了“大多数模型取决于训练数据的质量和数量”的观点，两个句子具有较高的相关性，共同构成了该参考文献的引文上下文。

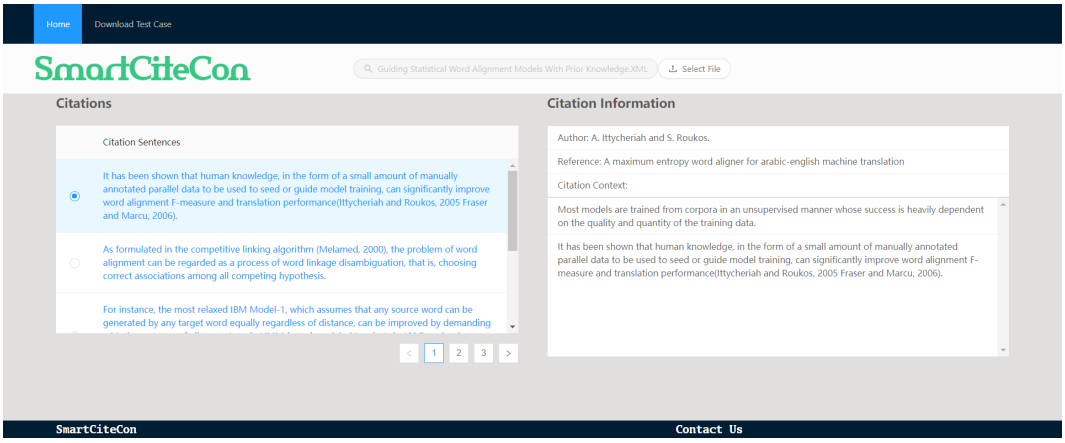

图 3 **SmartCiteCon** 前端可视化界面

4 总结与展望

本文在 SCC 系统基础上开发了引文上下文识别工具 SmartCiteCon2.0（SCC2.0），用于从学术文献中提取显性和隐性的引用上下文。该系统使用的微调后的 SciBERT 模型 F1 值达到 90 %。SCC2.0 接受 PDF、XML（符合 PLoS ONE 或 TEI 标准）和 JSON（符合 S2ORC 标准）格式的学术文献。SCC2.0 的输出是一个 JSON 文件，其中包含对应引文标记的引用上下文和论文的元数据。

SCC2.0 的局限性在于该模型是在计算机语言学相关的论文上进行训练的，因此将模型应用于其他领域时，应进行更仔细的评估和特征分布分析。后续将会在其他领域的数据上进行评估分析，使得该系统在更多学科领域有更为广泛的适用性。

参 考 文 献

[1] Hernández-Alvarez M, Gomez J M. Survey about citation context analysis:Tasks, techniques, and resources[J]. Natural Language Engineering, 2016, 22(3):327-349.

[2] Le M H, Ho T B, Nakamori Y. Detecting emerging trends from scientific corpora[J]. International Journal of Knowledge and Systems Sciences, 2005, 2(2):53-59.

[3] 章成志 , 徐津 , 马舒天 . 学术文本被引片段的自动识别研究 [J]. 情报理论与实践 , 2019, 42(9):139-145.

[4] 徐健 , 李纲 , 毛进 , 等 . 文献被引片段特征分析与识别研究 [J]. 数据分析与知识发现 , 2017, 1(11):37-45.

[5] Qazvinian V, Radev D R. Scientific paper summarization using citation summary networks[C]. Proceedings of the 22nd International Conference on Computational Linguistics. 2008:689-696.

[6] 牛海波 , 赵丹群 , 郭倩影 . 基于 BERT 和引文上下文的文献表征与检索方法研究 [J]. 情报理论与实践 , 2020, 43(9):125-131.

[7] 祝清松 , 冷伏海 . 基于引文内容分析的高被引论文主题识别研究 [J]. 中国图书馆学报 , 2014, 40(1):39-49.

[8] Caragea C, Bulgarov F, Godea A, et al. Citation-enhanced keyphrase extraction from research papers:A supervised approach[C]. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. 2014:1435-1446.

[9] Abu-Jbara A, Ezra J, Radev D. Purpose and polarity

doi:10.3772/j.issn.2095-915x.2021.05.001

of citation:Towards nlp-based bibliometrics[C]. Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics:Human language technologies. 2013:596-606.

[10] 雷声伟 , 陈海华 , 黄永 , 等 . 学术文献引文上下文自动识别研究 [J]. 图书情报工作 , 2016(17):78-87.

[11] Councill I G, Giles C L, Kan M Y. ParsCit:an Open-source CRF Reference String Parsing Package[C]. Proceedings of the language resources and evaluation conference. 2008:661-667.

[12] Lopez P. Grobid:Combining automatic bibliographic data recognition and term extraction for scholarship publications[C]. International conference on theory and practice of digital libraries. 2009:473-474.

[13] Ritchie A, Robertson S, TEUFEL S. Comparing citation contexts for information retrieval[C]. Proceedings of the 17th ACM conference on Information and knowledge management. 2008:213-222.

[14] Tahamtan I, Bornmann L. What Do Citation Counts Measure? An Updated Review of Studies on Citations in Scientific Documents Published between 2006 and 2018[J]. Scientometrics, 2019(5):1635-1684.

[15] Guo C, Cui H, Zhang L, et al. SmartCiteCon:Implicit Citation Context Extraction from Academic Literature Using Supervised Learning[C]. Proceedings of the 8th International Workshop on Mining Scientific Publications. 2020:21-26.

[16] Nanba H, Okumura M. Towards Multi-paper Summarization Using Reference Information[C]. Proceedings of 16th International Joint Conferences on Artificial Intelligence. 1999.

[17] Qazvinian V, Radev D. Identifying Non-Explicit Citing Sentences for Citation-Based Summarization[C]. Proceedings of the 48th annual meeting of the association for computational linguistics. 2010:555-564.

[18] Sugiyama K, Kumar T, Kan M Y, et al. Identifying citing sentences in research papers using supervised learning[C]. 2010 International Conference on Information Retrieval & Knowledge Management. 2010:67-72.

[19] Angrosh M A, Cranefield S, Stanger N. Conditional random field based sentence context identification:enhancing citation services for the research community[C]. Proceedings of the First Australasian Web Conference. 2013:59-68.

[20] Xu W, Rudnicky A. Can artificial neural networks learn language models?[C]. Sixth international conference on spoken language processing, 2000.

[21] Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch[J]. Journal of machine learning research, 2011(11):2493-2537.

[22] Peters M E, Neumann M, Iyyer M, et al. Deep contextualized word representations[C]. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2018:2227-2237.

[23] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. arXiv:1810.04805

[24] Beltagy I, Lo K, Cohan A. SciBERT:A Pretrained 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. 2019:3165-3120.

[25] Schäfer U, Weitz B. Combining OCR outputs for logical document structure markup. Technical background to the ACL 2012 Contributed Task[C]. Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries. 2012:104-109.

[26] Teufel S, Siddharthan A, Tidhar D. Automatic classification of citation function[C]. Proceedings of the 2006 conference on empirical methods in natural language processing. 2006:103-110.

[27] Lo K, Wang L L, Neumann M, et al. S2ORC:The Semantic Scholar Open Research Corpus[C]. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 2020.