用于预测缺失书目链接的文本采样策略 s

F. V. 克拉斯诺夫[a], I. S. 斯马兹涅维奇[a]*, E. N. 巴斯卡科娃[a]

[a] *NAUMEN R&D*，俄罗斯联邦叶卡捷琳堡市塔季谢娃街*49A*号，邮编*620028**

ismaznevich@naumen.ru

克拉斯诺夫·费奥多尔·弗拉基米罗维奇——工学博士。NAUMEN R&D 语义系统部专家。研究兴趣：智能文本分析。已发表70余篇科学论文。ORCID：0000-0002-9881-7371，电子邮箱：fkrasnov@naumen.ru

斯马兹涅维奇·伊рина·谢尔盖耶夫娜——NAUMEN R&D 语义系统部业务分析师。毕业于莫斯科国立罗蒙诺索夫大学计算数学与控制论系。研究兴趣：智能算法在应用信息系统中的应用。科学论文数量：8篇。ORCID：0000-0002-5996-4635，电子邮箱：ismaznevich@naumen.ru

巴斯卡科娃·叶лена·尼古拉耶夫娜——NAUMEN R&D 语义系统部首席系统分析师。毕业于鲍曼莫斯科国立技术大学信息学与控制系统系。研究兴趣：智能算法在应用信息系统中的应用。科学出版物数量：5篇。

ORCID：0000-0002-7071-8961 电子
邮箱：enbaskakova@naumen.ru

字数 – 3199

用于预测缺失书目链接的文本采样策略 s

本文提出了在自动进行句子分类以检测缺失书目链接时，对文本数据进行采样的多种策略。我们以句子作为文本的语义单元来构建样本，并加入由若干相邻句子组成的即时上下文。我们考察了多种采样策略，这些策略在上下文的大小和位置上各不相同。

该实验以 STEM 科学论文集为对象开展。将句子的上下文纳入样本，可提高其分类效果。通过对以不同方式采样的同一数据进行分类并实施集成投票，我们可以自动确定给定文本集的最佳采样策略。采用硬投票程序、考虑句子上下文的采样策略，可达到 98% 的分类准确率（F1 分数）。这种检测缺失书目链接的方法可用于应用型智能信息系统中的推荐引擎。

关键词：文本采样、采样策略、引文分析、书目链接预测、句子分类。

1. 引言

科学研究离不开将研究成果与其他科学家的工作进行关联。文章应通过插入书目链接来提及其他研究成果。科学计量学专家阐释了在研究之间建立此类链接的必要性，并提出了各种引文理论。

引文规范理论以 Merton（1973年）提出的科学伦理原则为基础，认为科学论文中的参考文献旨在指出作为研究基础或与研究主题相关的著作，描述所采用的研究方法，并为讨论研究结果所必需。根据反思理论，科学著作之间的链接体现了科学现状，并有助于构建其形式化表述，例如科学知识图谱（Akoev 等人，2014年）。

因此，科学引文正确性的受益者是整个科学界，包括撰写文章报告研究成果的研究人员，以及监督各科学领域成果的管理人员。在科学文本中提及其他科学家的相关重要成果，是基本要求之一，尤其从科学期刊编辑的角度来看更是如此。这些要求见于学术写作指南（Emerson等，2005年；Gray等，2008年；Pears和Shields，2019年），并在实践中得到印证，例如顶级国际期刊出版活动研究的结果（Arsyad等，2020年）。

科学论文的作者自行选择引文来源和链接的位置，目前这一过程尚未实现自动化。在本研究中，我们探讨了开发一种推荐算法的可行性，该算法能够在科学文章中找到缺失的参考文献，也就是识别出需要提及另一项研究成果的文本片段。为此，我们采用半监督机器学习方法，估算文本片段中存在链接的概率。所研究问题的正式表述如下：使用一组带有链接和不含链接的带标签片段作为训练数据，自动找出科学文章文本中缺少链接但有必要添加链接的片段（句子）。

根据文本片段中是否包含链接对其进行分类，这项任务在方法上类似于情感分析：情感分析会根据文本的情感特征，主要将文本自动分类为正类和负类。除了将片段划分为正类和负类，情感分析方法还用于区分其他类别，包括引文重要性检测（Aljuaid等，2021年；Prester等，2021年；Varanasi等，2021年；

Färber和Ashwath，2019年）。识别文本中缺失或不必要的链接这一问题与情感分析相似；这里要识别的情感是作者确认已表述观点的需求。

另一项密切相关的研究方向是利用分类器进行预测的命名实体识别（NER）。Fu 等人（2021）研究了一个类似问题，并采用跨度预测方法解决命名实体识别问题。命名实体识别可分两个阶段进行：找出很可能包含实体的片段，并确定这些实体的确切位置（Ziyadi 等人 2020；Li 2021）。一些命名实体识别方法还会考虑实体的局部、全局或外部上下文（Wang 等人 2021）。

许多研究都探讨了结合最近上下文进行句子分类的任务。Fiok 等人（2020）使用语言模型生成的上下文化嵌入，其高质量是以速度为代价的。Glazkova（2020）研究了主题分类，并指出将上下文作为输入的模型优于不考虑上下文的模型。在这些研究中，上下文大小是根据某种偏差预先[illegible]特定文本语料库而言，未必是最优的。

本文提出的方法也可以被视为一种重采样技术。迄今为止，重采样主要用于平衡训练数据集中的类别分布，以提高类别预测的准确率；不平衡数据会对准确率产生负面影响。重采样方法分为三类，即欠采样、过采样和混合技术。欠采样会删除多数类中的部分数据（例如，Maya 和 Jayasudha，2017；Akkasi 等人，2017），而过采样则会复制少数类中的现有样本，或创建新的样本（Luo 等人，2019；Li 等人，2018；

Chawla 等，2002），混合重采样技术旨在结合两者的优点（Taha 等，2021）。

现代文本分析神经网络架构会同时考虑局部上下文和全局上下文。由于文本是由词项构成的单向序列，上下文通常被理解为所考察词项前后相邻的一些词语（Gallant，1991；Huang，2012）。在卷积神经网络中，增大上下文范围会显著增加张量的维度，进而增加模型的参数数量，因此需要扩大语料库的规模。在称为 Transformer 的深度学习模型中，通过注意力机制来探索上下文，并将局部上下文与更广泛的上下文结合起来（BERT [Devlin 等人 2018], GPT-3 [Brown 等人 2020]）。

重要的是，上述所有算法都没有考虑文本的自然结构单元（即句子和段落），因为这些算法针对特定大小的上下文进行调整，而该大小由固定数量的词语构成；句子和段落的长度却各不相同。

2. 方法

确定缺失链接的任务被形式化为：找出链接缺失但必不可少，或反过来链接存在但并非必需的文本片段。

我们通过自动二元分类来解决这一问题，分类分为正类和负类两种。对于科学文章中的每个片段，我们的算法都会确定其中包含书目链接的概率。给定一组文本文件，其中每个文档由若干片段组成。片段是由不同数量的词（术语）构成的序列。片段之间可以重叠，大小也各不相同。每个片段都是一个样本，并被标记为两种可能类别之一，即正类或负类。类别标签表示给定的

片段是否包含书目链接。本研究的任务是寻找一种片段样本的构建策略，使特定分类器判定类别标签时达到最高准确率。

我们研究的假设如下：考虑上下文的文本采样策略能够提高句子分类的准确率，从而用于预测科学文章中缺失的书目链接。

我们认为，正样本由一个书目链接及其在原文中周围的上下文构成，而负样本则是不含书目链接的片段。为避免样本重复，对于包含两个或更多链接的句子，我们只将其计为一次。链接上下文可以仅限于包含该链接的句子，也可以扩展至包含相邻句子。

最理想的情况是，链接上下文的边界与该链接所属的完整作者陈述的边界一致。在这种情况下，文本的语义单元可能由一个或多个句子构成，因此难以确定上下文的大小。不过，为了在所提算法中实现既定目标，我们将上下文视为一个片段，其大小由句子数量而非词语数量决定，这一点与神经网络算法不同。因此，在我们的算法中，上下文是根据文本的自然结构单元构建的。

特征空间基于词袋模型（BoW）中的词汇统计自动构建。该模型的词汇包括单词以及所有原始标点符号和排版符号。我们还将命名实体作为附加特征。

算法

该算法由以下几个阶段组成。

1. 文本预处理

- 文本清洗：移除控制字符（制表符、换行符等）、词语（期刊标题、国际标准书号等）和章节（基金资助、参考文献）；• 分词：o 将文本拆分为句子；o 术语规范化。

2. 数据标注

- 为每篇文档（期刊论文）添加起始和结束标记；• 对于每个句子：o 如果句子中包含书目链接（引文标记），则将其标记为“含链接”类别；o 如果句子中没有引文标记，则将其标记为“不含链接”类别。o 标注句子后，移除引文标记。

3. 命名实体处理

- 检测文本中的命名实体；• 用特殊标记替换
命名实体。

4. 样本构建

Sa样本会根据类别（正类或负类）以不同方式构建 :

- 构造正样本时，我们取一个“含链接”的句子，并添加前面的n个句子和后面的m个句子，所有句子均按原文顺序排列。

- 负样本由原文中连续的k个“不含链接”的句子构成（相邻句子），其中$k = n+1+m$。

样本内容的可视化如图1所示。

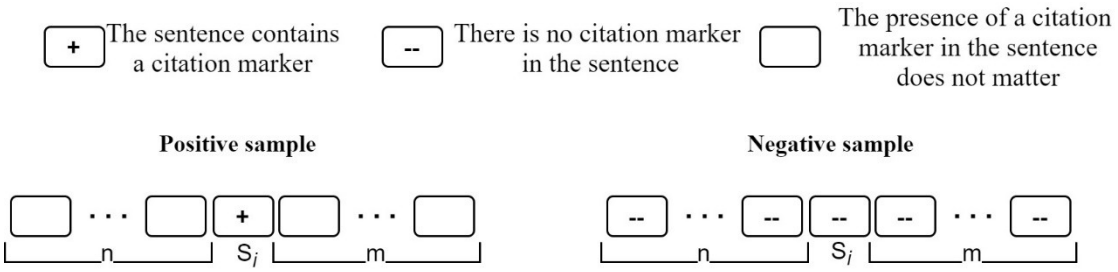


图1. 样本构建。

5. 样本分类

- 通过随机欠采样平衡训练集中的类别分布。- 对于每个样本，使用计数向量化器构建向量模型，这是速度最快且计算效率最高的文本表示方法。- 使用分类器处理向量化后的样本集。

6. 最优采样策略确定

此外，使用集成方法自动确定最优采样策略。我们将以不同方式采样的相同数据输入同类型的估计器，并实施投票程序。

整个算法的流程图如图2所示。每个词袋模型j对应一种采样策略，并且对于每种采样策略，我们都运行各自的估计器。所有估计器都实现相同的分类方法，但将不同类型的样本作为输入数据。

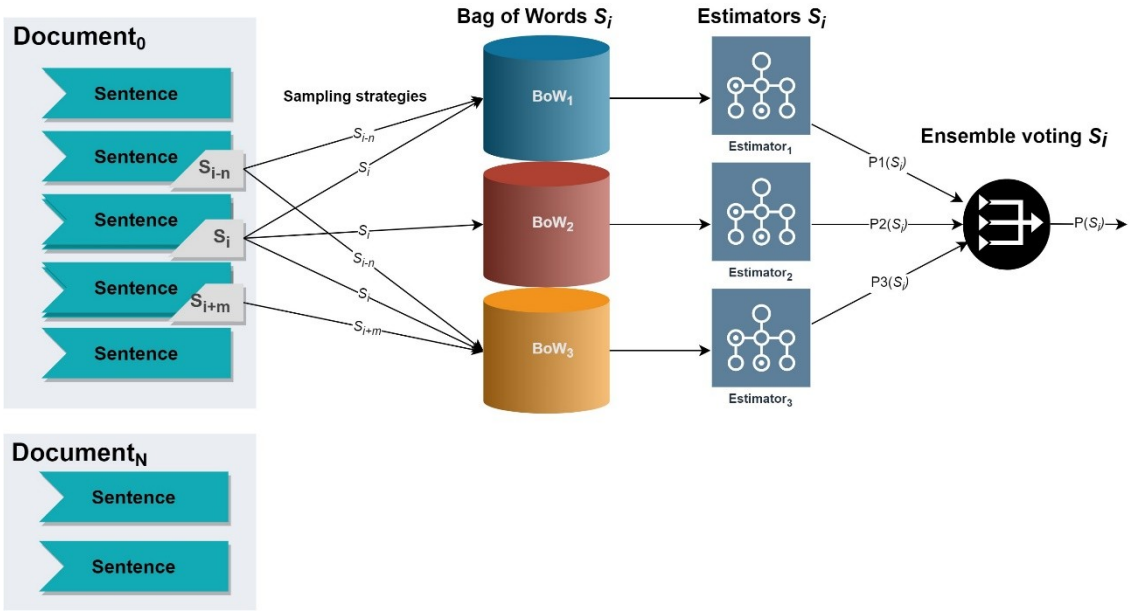


图2. 使用不同采样策略将句子分类为“含链接”和“不含链接”的算法流程图（针对 n=m=1）。

3. 实验

为了通过实验检验这一假设，我们采用了 Cohan等人（2018年）从科学存储库 arXiv.org 收集的 STEM期刊文章数据集。该数据集中的文档仅包含文本，图和表格均已移除。数学公式和引文标记被替换为特殊标记 @xmath<number> 和 @xcite。文档仅包含结论部分之前的章节，结论部分之后的所有章节均已移除。

数据集规模如下：文档数量为 215K，文档平均长度为 4938 个词，摘要平均长度为 220 个词。

这些文件采用 jsonlines格式，其中每一行都是一个对应于一篇科学文章的 JSON对象。每行包含文章的摘要、各章节和正文，并且这些文本均已进行句子分词。

在我们的实验中，我们考虑长度超过 30 个单词的句子。在这一限制下，我们总共得到 458774 个句子。

包含引文标记 @xcite 的句子被分配到正类（“含链接”），随后移除引文标记。不含引文标记的句子被标记为负类（“不含链接”）。类别比例为：24% 的正类句子，76% 的负类句子。这被视为采样策略 #0，而按此方式采样的数据上的分类结果被视为基线：通过 F1 分数衡量的采样策略 #0 的分类准确率为 0.7866。

建立基线后，我们测试了各种数据采样策略，以提高分类准确率。采样的主要思路是考虑包含链接的句子的上下文。不同的采样策略会采用不同的上下文方向、位置和范围，这些由周围句子的数量决定。在不同的采样策略中，每个句子 [*i*]都会被纳入不同类型（变体）的样本中。

在实验中，我们使用以下参数测试了 10 种策略：*n*：[0, 1, 23 4 5], *m*：[0, 1, 2, 3, 4], *k*：[1, 3]。所选采样策略对应的所有样本类型均列于表 1 中。

表 1. 实验中测试的采样策略。

#	**采样策略（样本构建算法）**	**每个样本的句子数**	
		正类	**负类**
0	句子 [*i*]	2640	12352
1	句子 [*i*： *i* +2]	2640	10639
2	句子 [*i* -1： *i*+1]	2640	10639
3	句子 [*i* -1: *i*+2]	2640	9376
4	句子 [*i* -2: *i*+2]	2640	8384
5	句子 [*i* -3: *i*+2]	2640	7574

6	句子 [*i* -3: *i*+3]	2640	6880
7	句子 [*i* -4: *i*+3]	2640	6287
8	句子 [*i* -4: *i*+4]	2640	5776
9	句子 [*i* -5: *i*+4]	2640	5326

图3展示了不同类型正类和负类样本的长度（单词数）分布。

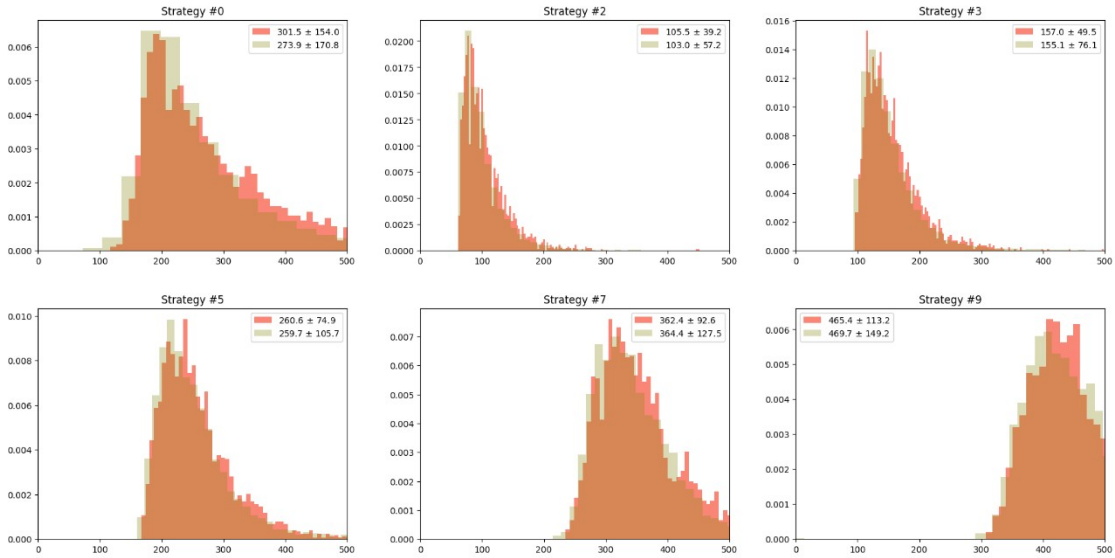


图3. 不同类型正样本和负样本中词数长度的分布（“鲑鱼色”表示正类，“橄榄色”表示负类）。

通过随机欠采样使类别均衡后，数据按照 test_size=0.33 的比例参数划分为训练集和测试集。

向量表示使用 Scikit-learn 库中的 CountVectorizer 方法构建。词汇表包含一元词组和二元词组，并通过参数 min_df=3、max_df=0.7 根据词频进行精简。

我们使用神经网络多层感知机进行分类（即 Scikit-learn 库中的 MLPClassifier 方法）。不同采样策略下的分类性能见表2。

表2. MLPClassifier 根据采样策略进行句子分类的结果（加权平均）。

y	采样策略	F1	精确率	召回率
	0	0.7866	0.7866	0.7866
	1	0.8882	0.8881	0.8881
	2	0.8884	0.8881	0.8881
	3	0.9214	0.9214	0.9214
	4	0.9444	0.9443	0.9443
	5	0.9410	0.9409	0.9409
	6	0.9601	0.9598	0.9598
	7	**0.9640**	**0.9639**	**0.9639**
	8	0.9593	0.9593	0.9593
	9	0.9581	0.9581	0.9581

从表中可以看出，采用采样策略 #7 时结果最佳。进一步增加每个样本的句子数并不能显著提高准确率，因为准确率趋于渐近线。

对于每个句子，我们比较采用所有采样策略得到的分类结果，并通过投票程序进一步改进结果，包括软投票和硬投票。对于软投票，我们将平均预测概率的阈值设为 0.5。对于硬投票，我们汇总所有预测概率，并将其总和与

阈值 3 进行比较。我们测试了不同数量的估计器。为构成采样策略组合，我们从估计器 #7、#8 和 #9 这一组开始，然后按逆序逐个添加估计器。不同估计器数量下的投票结果如图4所示。

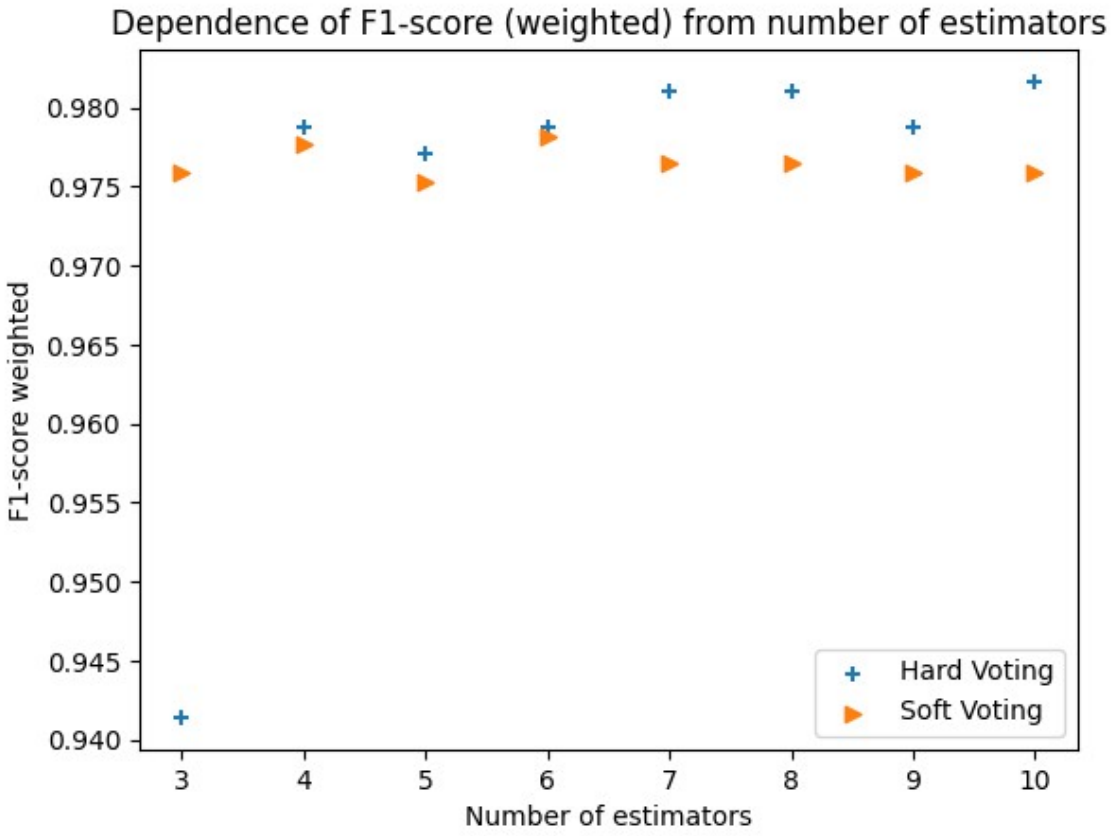


图4. 根据纳入的估计器数量，采用硬投票和软投票进行分类的结果。

图表表明，投票程序进一步提升了分类性能，使 F1 分数提高了 1,5%。在测试过的所有估计器组合中，结果始终很高，但对长样本进行分类时，使用 7、8 或 10 个估计器的硬投票取得了最佳结果（>98%）。

4. 结果

The 提出的研究假设已通过实验得到证实。我们已 e

s证明，采样策略的选择会影响文本分类的结果 .

以采样策略 #0 建立基线。在这种情况下，使用 F1 分数 衡量的分类性能仅为 79%，不足以满足工业信息系统实际应用的需求。

这一改进得益于数据采样策略，该策略能够自动确定最优样本类型。这通过对不同估计器所作出的决策应用投票程序来实现。所提算法的准确率（F1 分数）达到 98%，与采用自动分类的命名实体识别以及其他文本分类任务的最先进成果相当。重要的是，所提算法不仅准确率高，而且实现时不需要大量计算资源。

5. 结论

本文提出了一种新方法，用于确定科学文章片段中存在书目链接的概率。该方法采用集成投票进行句子分类，其中不同的数据采样策略对应于实现相同分类方法的估计器。作者提出的问题与命名实体识别和情感分析等研究充分的领域相近，但从实际应用角度来看仍属新颖。

所提方法的主要创新在于，找出对检测句子中缺失书目链接的概率影响最大的链接上下文。在所提算法中，上下文的最佳大小和位置由算法自动确定。其大小依据文本语义单元的边界，并以句子数而非词数来衡量，因此我们利用了这样一个事实：与词语相比，句子是语义容量更大（意义更丰富）的单位。大多数现有的文本分类方法并未将片段上下文视为具有重要意义的因素，但

本研究表明，将其纳入考虑至关重要。上下文对分类性能的显著影响表明，与书目链接相关的语义可以分布在长度各异的片段中。

所提算法的准确率达到 98%（F1 分数）。值得注意的是，与卷积人工神经网络相比，所述方法具有较高的计算效率。这一优势得益于样本规模较大。所研究的文本分析方法拓展了注意力机制的原理，旨在训练语言模型理解全局和局部上下文的影响。自动确定上下文边界的做法，与人工神经网络中自动选择显著特征的思路相呼应。

所提方法可用于应用型智能信息系统的推荐引擎，包括协助构建文档、撰写可能包含指向其他文档的链接的文本，或帮助检查文档的正确性。这些功能在许多领域都很有用，例如科学、法律或新闻业；在这些领域，文档中的陈述需要通过引用法律文件或其他来源加以证实。

根据公司的政策，我们不公开源代码。

参考文献

1. Akkasi, A., E. Varoğlu 和 N. Dimililer。2018年。《平衡欠采样：一种用于提高化学和生物医学文本中命名实体识别效果的新型基于句子的欠采样方法》。*Applied Intelligence* 48(8)：1965–1978。doi：https://doi.org/10.1007/s10489-017-0920-5

2. Akoev, M., V. Markusova, O. Moskaleva, and V. Pislyakov. 2014年. *Rukovodstvo po Naukometrii:Indikatori Razvitiia Naukii Tehnologii* [《科学计量学手册：科学技术发展指标》

（俄语）。叶卡捷琳堡：乌拉尔大学出版社。doi: https://doi.org/10.19181/smtp.2019.1.1.5 3. Aljuaid, H., R. Iftikhar, S. Ahmad, M. Asif, and M. T. Afzal. 2021年“利用文中引文的情感分析识别重要引文。”《电信与信息学》 56: 101492. doi: https://doi.org/10.1016/j.tele.2020.101492 4. Arsyad, S., S. Ramadhan, and I. Maisarah, 2020.“印度尼西亚讲师在社会科学与人文学科领域撰写国际期刊研究论文时遇到的修辞问题。”《亚洲应用语言学杂志》 7(1): 116–129. 5. Brown, T. B., B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, et al. 2020.“语言模型是少样本学习者”。*arXiv*预印本 arXiv:2005.14165. 6. Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002.“SMOTE：合成少数类过采样技术。”《人工智能研究杂志》 16: 321–357. doi: https://doi.org/10.1613/jair.953 7. Cohan, A., F. Dernoncourt, D. S. Kim, T. Bui, S. Kim, W. Chang, and N. Goharian. 2018.“用于长文档抽象式摘要的话语感知注意力模型。”*arXiv*预印本 arXiv:1804.05685. J. Devlin, M. W. Chang, K. Lee, and K. Toutanova. 2018 “Bert：用于语言理解的深度双向变换器预训练。”*arXiv*预印本 arXiv:1810.04805. 9. Emerson, L., M. T. Rees, and B. MacKay. 2005.“培养学术诚信：为教授参考文献引用技能创建学习环境。”《大学教学与学习实践杂志》 2 (3): 3. doi: https://doi.org/10.14453/jutlp.v2i3.3.10. ExplainaBoard—— 命名实体识别。http://explainaboard.nlpedia.ai/leaderboard/task-ner/，最后访问日期：2021/12/13。11. Farber, M., and S. Ashwath. 2019. 确定引文在引文语境中的用法。“数字图书馆理论与实践国际会议。施普林格，Cham。doi: https://doi.org/10.1007/978-3-030-30760-8_3812. Fiok, K., et al. 2020.“比较现代语言模型进行句子分类的质量与速度。”《应用科学》 10.10: 3386. doi: https://doi.org/10.3390/app10103386 13. Fu, J., X. Huang, and P. Liu. 2021.“SpanNer：将命名实体识别作为跨度预测。”*arXiv*预印本 arXiv:2106.00641.

14. Gallant, S. I. 1991. “一种利用神经网络表示上下文并执行词义消歧的实用方法。” 神经计算 3(3): 293–309. doi: https://doi.org/10.1162/neco.1991.3.3.293 15. Glazkova, A. V. 2020. “考虑最近上下文的文本片段主题分类。” 自动化与远程控制 81: 2262–2276. doi: https://doi.org/10.1134/S0005117920120097 16. Gray, K., C. Thompson, R. Clerehan, J. Sheard, and M. Hamilton. 2008. “Web 2.0 作者身份：学术诚信方面的参考文献标注与引文问题。” 《互联网与高等教育》 11(2): 112–118. doi: https://[illegible] E. H., R. Socher, C. D. Manning, and A. Y. Ng. 2012. “通过全局上下文和多个词原型改进词表示。” 计算语言学协会第*50*届年会论文集 1: 长篇论文：873–882. 18. John, M., and J. S. Jayasudha. 2017. “增强基于深度学习的文本摘要器的性能”。国际应用工程研究期刊 12 (24): 15986–15993. 19. Li, B. 2021 年 . “ 以目标检测的方式进行命名实体识别”。 *arxiv*预印本 arxiv:2101.11122. 20. Li, Y., H. Guo, Q. Zhang, M. Gu, and J. Yang. 2018年. “利用通用知识和领域特定知识进行不平衡文本情感分类。” 知识库系统 160: 1–15. doi: https://doi.org/10.1016/j.knosys.2018.06.019 21. Luo, Y., H. Feng, X. Weng, K. Huang, and H. Zheng. 2019. “一种新颖的过采样

g

用于不平衡文本分类的 SeqGAN 方法。” *2019*年 *IEEE*国际大数据会议（*Big Data*）：2891–2894。IEEE。doi: 10.1109/BigData47090.2019.9006128. 22. Merton, R. K. 1973年。科学社会学：理论与实证研究。芝加哥大学出版社。23. Pears, R., 和 G. J. Shields. 2019年。正确引用文献：基本参考文献指南。麦克米伦国际高等教育。24. Prester, J., G. Wagner, G. Schryen, 和 N. R. Hassan. 2021年。“信息系统综述文章观念影响的分类：一种内容增强深度学习方法。” 决策支持系统140：113432。doi：https://doi.org/10.1016/j.dss.2020.113432

25. Taha, A. Y. h, A. 2021年。“通过类别对齐进行多标签过采样与欠采样，以实现不平衡多标签文本分类。”《信息与通信技术杂志》20（3）：423–456。doi：https://doi.org/10.32890/jict2021.20.3.6

26. Varanasi, K. K., T. Ghosal, P. Tiwary, and M. Singh. 2021年。“Iitp-cu 的监督方法。”第二届学术文献处理研讨会论文集：140-145。27. Wang, Y., Y. Jiang, N. Bach, T. Wang, Z. Huang, F. Huang, and K. Tu. 2021年。“通过外部上下文检索与协同学习提升命名实体识别。” *arXiv* 预印本 arXiv:2105.03654。28. Ziyadi, M., Y. Sun, A. Goswami, J. Huang, and W. Chen. 2020。“基于示例的命名实体识别。” *arXiv* 预印本 arXiv:2008.10570。