arXiv:2304.11922v1 [cs.CL] 24 Apr 2023

利用科学出版物生成科学概念主题页面

Hosein Azarbonyad、Zubair Afzal 和 George Tsatsaronis

Elsevier，荷兰{h.azarbonyad, zubair.afzal, g.tsatsaronis}
@elsevier.com

摘要。 本文介绍 *Topic Pages*，这是一个从大量科学书籍和期刊中提取的科学概念及相关信息库。 *Topic Pages*的主要目标是向读者提供理解他们在阅读任何科学领域的学术内容时遇到的科学概念所需的全部信息。 *Topic Pages* 是一系列使用自然语言处理和机器学习自动生成的信息页面，每个页面对应一个科学概念。每个页面包含三类信息：定义、相关概念和最相关的片段，均提取自科学同行评审出版物。本文将详细讨论提取上述各项内容的不同组件。现已上线的页面集合包含超过 360，000 *Topic Pages* ，覆盖 20 个不同的科学领域，月均独立访问量达 2300 万次，使其成为广受欢迎的科学信息来源。

关键词： 科学文档处理 · 定义抽取 · 多文档摘要。

1 引言

技术术语是科学出版物中的重要组成部分 [6,7]。科学家和研究人员使用技术术语和概念来简洁地传达信息。因此，在任何科学领域中，科学概念的数量都十分庞大且不断增长，这也增加了科学家跟进不断扩充的技术概念清单和新内容的难度。诸如维基百科 这样的知识来源能够在很大程度上提供有关技术和科学概念的有用信息。然而，由于其“众包智慧”的创建方式，其中存在许多遗漏和错误，因此它们未必总是理解和查阅科学概念的可靠来源。我们的主题页面[1] 方案以“专家智慧”的方式构建知识来源，从该领域的经典科学书籍或该主题上具有高影响力的同行评审科学出版物中抽取科学概念信息。

每个 主题页面都围绕一个科学概念展开，并包含该概念的定义、一组相关概念，以及一组相关片段

[1] https://www.elsevier.com/solutions/sciencedirect/topics

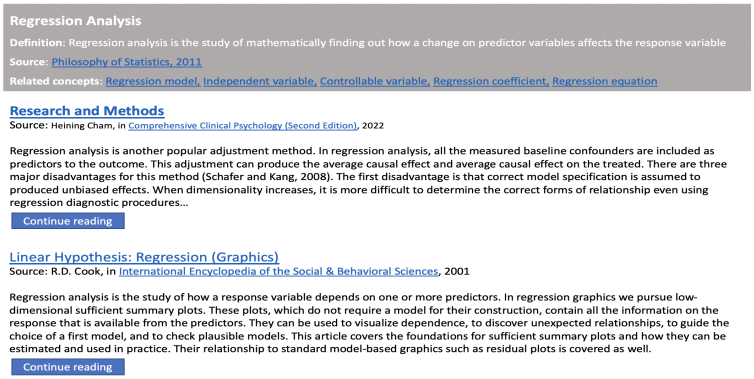


图 1. 一个示例 *Topic Page*，展示了“回归分析”这一概念，并提供了定义、相关概念，以及从文章和书籍中提取的一组相关片段。

这些内容均提取自科学同行评审文章和书籍。定义由从书籍和期刊中提取的一句话组成，用于简要而准确地描述该概念。片段是与该概念相关的书籍或期刊中的文本摘录，并提供有关该概念的上下文信息。相关概念是一组与给定概念最相关的概念，可帮助用户探索围绕其感兴趣概念的相关术语。

该合集包含 20 个不同科学领域的超过 360，000 个主题页面 。这些主题页面通过 ScienceDirect[2],上的出版物链接至原文。ScienceDirect 是规模最大的科学出版物搜索引擎和数据库之一，收录了超过 1,800 万篇全文文章，帮助用户只需单击一下，即可在文章中遇到不熟悉的科学概念时跳转至相应的主题页面 。我们已在科学文章与主题页面之间创建超链接，目前已有超过 580.万篇文章提供此类超链接。主题页面 每月吸引超过 2,300 万独立访客。

在本文余下部分中，我们将在第 2 节简要回顾相关工作，在第 3 节介绍生成主题页面 的技术流程，在第 4 节对处理流程中最具挑战性的模块——定义抽取——进行实证评估，并在第 5 节总结当前技术方案的一些局限，同时提出未来工作的方向。

2 相关工作

据我们所知，目前还没有与本文提出的方案类似、能够自动为科学概念生成主题页面的解决方案。大多数相关工作都属于定义抽取任务，这也是本节的重点。早期的定义抽取任务研究

[2] https://www.sciencedirect.com/

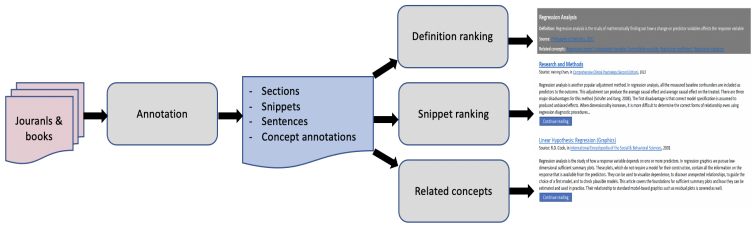


图 2. 主题页面生成处理流程概览，包含所有必要组件。

主要采用基于规则和模式匹配方法 [3,8,20], ，但由于覆盖范围有限，往往导致召回率较低。研究者也提出了监督模型，并证明在这项任务中，它们比基于规则的方法更有效 [6,14,15,16,9]。这些模型利用有关概念的统计信息，以及句子的结构信息（例如词性（词性）标签），来区分定义性句子和非定义性句子。近期的定义抽取研究则着重于使用神经网络模型完成这项任务[11,2,5,18,7,13,19]。值得注意的是，研究者曾使用$LSTM$ [11] 以及CNN 和$LSTM$的组合[2] 来学习定义性句子的结构。在我们的工作中，我们也提出并使用了结合$LSTM+CNN$ 的模型，但与[2],不同，我们同时捕捉句子的语义信息（从句子本身学习）和结构信息（从词性标签学习）。此前， [7], 曾使用一种对句子及其结构进行编码的联合模型；但与该研究所处理的任务不同，我们的定义抽取组件假定术语已知，并尝试判断给定的候选句子是否是该术语的良好定义。

3 主题页面处理流程

如图 2 所示，生成主题页面主要包含四个组件：文章注释模块、定义排序模块、摘要排序模块和相关概念提取模块。

3.1 文章注释

注释模块接收XML 格式的内容，在文章和书籍中查找概念的提及，然后将提及某个概念的句子和片段传递给后续组件。文章中的每个章节都被视为一个片段。完成句子切分后，我们通过对$Omniscience$ 分类体系 [12]进行简单的词典查找来注释句子中的概念，该分类体系是一个科学概念的分类体系。如果文本中提出了该概念的缩写，例如“机器学习（ML）”，那么该缩写（ML）也会作为该概念的别名添加，并在文章中进行查找。我们使用Schwartz和Hearst方法 [17]来检测此类缩写。如果多个概念部分共享某个跨度（注释范围），我们会使用最长的概念注释该跨度，并忽略较短的注释。

3.2 定义排序与提取

定义对概念进行了简洁的描述。对于每个概念和每个领域，我们会对所有注释了该概念的句子进行排名，并选择排名最高的句子作为该概念的定义。我们将机器学习任务简化为二分类：给定一个概念和一个候选句子，模型预测该句子是否是该概念的良好定义。对于目标概念，候选句子会根据分类器为其分配的分数进行排名，并将排名最高的句子用作定义。我们为定义分类任务使用两种不同的模型：$LSTM+CNN$ 和$SciBERT$ 模型。

LSTM+CNN 模型 先前的工作 [11,2] 使用 $LSTM$ [4] 和 CNN [10]模型对定义分类任务中的句子进行分类。我们采用一种结合方法，使用两个 ***LSTM*** 和两个 ***CNN***：一个 $LSTM$ 以实际句子作为输入，捕捉词项的序列模式；另一个 $LSTM$ 以句子中各词的 POS 标记作为输入，捕捉句子句法的序列模式。一个 CNN 以实际句子作为输入，捕捉词项的空间分布；另一个 CNN 以句子中各词的 POS 标记作为输入，捕捉语法元素的空间分布。我们将这些模型各自学得的表示拼接起来，并将其输入前馈机器学习感知机（***MLP***）层进行分类，训练时使用交叉熵损失。

SciBERT 模型 我们使用在科学文章上训练的 $SciBERT$ 模型 [1] 。作为输入，我们将概念与候选句子用特殊标记（[SEP]）分隔后输入 $SciBERT$ 模型，并获取 [CLS] 标记的表示。随后，将该表示输入一个简单的前馈层进行分类，并使用交叉熵损失进行优化。

3.3 摘要片段排序

对于给定概念，我们收集所有标注了该概念的片段，并通过摘要片段排序方法对其进行排序。排名前 10 的片段用于生成该概念的主题页面 。我们使用词汇匹配模型，利用简单的位置感知词频分数对片段进行评分，如下所示：

$$F(c,s) = \frac{tf(c,s)}{|s|} * (1 - \frac{l_1(c)}{|s|}) \tag{1}$$

其中， c和 s 分别表示一个概念和一个片段， $tf(c, s)$表示 c 在 s中的频率， $|s|$ 表示 s的长度，而 $l_1(c)$表示c 在 s中首次出现的位置。因此，概念在片段中出现得越早，该片段获得的分数就越高。

3.4 相关概念提取

为了找到与给定概念最相关的概念，我们检索片段中所有共现的概念。随后根据这些概念与目标概念的

模型	精确率	召回率	F1值
Jin等人 [6]	0.92	0.79	0.85
Li等人 [11]	0.90	0.92	0.91
Navigli和Velardi [14]	**0.99**	0.61	0.85
LSTM+CNN	0.94	**0.91**	0.92
SciBERT	0.94	**0.93**	**0.93**

表 1。不同定义分类模型在 WCL数据集上的宏平均精确率、召回率和 F1值表现。

共现次数进行排序，并选取排名前 5 的概念作为目标概念的相关概念。

4 结果

本节介绍作为*Omniscience*科学内容集合和所使用的分类体系，这是主题页面的基础。我们还将讨论不同定义排序模型在两个数据集上的结果，并提供生成的主题页面 的一些统计数据以及随时间变化的使用统计数据。我们将摘要片段排序和相关概念提取模块的大规模评估留作未来工作。

4.1 数据集和基线

我们使用一组发表于2，700多种期刊上的文章，以及43，000本书的内容来生成主题页面。该集合包含超过1,800万篇采用XML 格式的文章和书籍章节。所有期刊和书籍均属于不同的科学领域。我们使用$OmniScience$ 分类体系构建主题页面，其中涵盖20个领域的70多万个概念。

为评估定义排序模块的性能，我们使用WCL 数据集 [14] ，其中包含4,619个句子，这些句子围绕某个概念被标记为定义性（“good”）或非定义性（“bad”）句子。我们采用与 [11] 相同的设置，在该数据集上训练和评估模型。此外，我们还使用一个专有数据集，其中包含从分布于8个不同领域的文章和书籍中提取的43,368个句子。这些句子由学科专家针对定义评估任务进行标注，分别作为围绕某个概念的“good”或“bad”定义。我们在WCL 数据集上比较了多个模型的性能，包括$LSTM+CNN$、$SciBERT$、Navigli和Velardi [14],、Li等人 [11], 以及Jin等人 [6] 提出的模型。随后，我们在专有数据集上进一步评估表现最佳的模型。对于$LSTM+CNN$ 模型，批量大小设为32，$LSTM$ 模型的隐藏层数量设为128，并使用$GloVe$ 初始化词嵌入，在训练过程中进行微调。MLP 模块包含一个256维的隐藏层，训练10个周期。$SciBERT$ 模型训练8个周期，批量大小为16。我们进行10折交叉验证，并报告平均性能。

4.2 定义抽取模型的结果

表1展示了不同模型在 WCL 数据集上的性能。该数据集提取自维基百科，其中大多数基于维基百科的定义遵循相似的结构，因此易于分类。$SciBERT$ 模型在该数据集上取得了最佳的*F1*值。Navigli和Velardi [14] 的精确率高于所有模型，但与$SciBERT$相比，其召回率非常低。$SciBERT$ 模型相较于$LSTM+CNN$ 模型表现更佳，这表明$SciBERT$ 能够学习定义性句子中词语的序列分布和空间分布，以及此类句子中的结构信息。

我们进一步在专有数据集上评估了表现最佳的模型（$SciB\text{-}ERT$ 、$LSTM+CNN$）的性能。该数据集比 WCL 数据集大得多，结果如表 3 所示。该数据集包含来自各种来源的定义。与基于维基百科的定义不同，从不同书籍和期刊中抽取的定义没有遵循相似的结构，这使分类任务更加困难，因此这两个模型的表现低于在 WCL 数据集上的表现。在该数据集上，$SciBERT$ 模型同样在所有领域都优于$LSTM+CNN$ 模型。这再次证实了$SciBERT$ 模型在建模定义的语义和结构方面的能力。此外，除社会科学外，$SciBERT$ 在所有单个领域中的表现始终高于$LSTM+CNN$ 。由于$SciBERT$ 是在生物医学和计算机科学领域的出版物上进行预训练的，因此该模型在社会科学 等领域中的较低表现可能归因于这一事实。另一方面，结果显示，$SciBERT$ 在化学 和材料科学 等领域中的表现更好，因为这些领域与其训练领域更加接近。

概念	**定义**	**错误来源**
关联列表	关联列表只是一组名称-值对列表。	过于笼统
层次数据库	在层次数据库中，关系是由存储结构定义	过于笼统
资格认证	获得此前不具备的能力。	过于具体
样本空间	概率模型中所有可能结果的集合，部分合理	

表2。 基于 SciBERT 的模型所产生错误（假阳性）的示例。

除领域差异外，其他错误应归因于任务本身固有的难度。根据我们的分析，错误的最大来源是假阳性，这主要是由泛泛的定义、具体定义或部分良好定义造成的。表2展示了被$SciBERT$ 模型错误标注的定义示例，以及这些错误可能产生的原因。泛泛的定义本身是良好的定义，但涵盖了概念非常宽泛的方面。具体定义同样是良好的定义，但包含了不必要的额外信息。部分良好定义只涵盖了概念的一些关键方面。领域专家将上述情况全部标注为“糟糕的定义”，但

领域	SciBERT			LSTM+CNN		
	精确率	召回率	F1值	精确率	召回率	F1值
化学	0.78	0.80	0.79	0.69	0.68	0.68
地球科学	0.80	0.84	0.82	0.66	0.64	0.65
材料科学	0.80	0.88	0.83	0.50	0.49	0.49
计算机科学	0.56	0.60	0.58	0.43	0.48	0.45
社会科学	0.39	0.43	0.41	0.38	0.46	0.42
所有领域	**0.79**	**0.78**	**0.78**	0.70	0.69	0.69

表 3。 *LSTM+CNN* 和 *SciBERT* 模型在五个领域上的性能。

模型却将其识别为“良好的定义”。要处理这类情况，模型应当能够理解概念的普遍性或特异性，而这可能很难建模。

主题页面 产品包含 20 个不同科学领域中的超过 363，000 个主题页面。主题页面每月访问量超过 2,300 万次，使其成为研究人员和学生中最受欢迎的知识库之一。主题页面 上约有 63，000 个概念没有定义，这主要是由于当前生产模型（*LSTM+CNN*）在某些领域中的性能不佳。

5 结论与讨论

本文介绍了主题页面，这是一个公开可用的科学概念知识库，其中包含概念的定义、最相关的概念，以及提供更多背景信息的片段。我们介绍了构建这一资源所结合的所有主要组件。用于生成主题页面 的处理流程可应用于任何文档集合以及分类体系，从而在任何领域构建类似资源。目前，，20 个不同科学领域中已有超过 363,000 个主题页面，每月独立访客超过 2,300 万，主题页面 是研究人员和学生常用的知识库之一。我们介绍了处理流程中的所有主要组件，这些组件用于抽取生成页面所需的不同信息。本文主要聚焦于构建高性能的定义抽取模型。为此，我们使用了 *LSTM+CNN* 和 *SciBERT* 模型。实证评估表明，这两种模型在定义分类与抽取任务上的表现均优于现有模型。不过，*SciBERT* 模型在社会科学等领域仍有待改进。在这些领域使用*SciBERT* 的最大缺点是，该模型主要基于生物医学文章进行预训练，因此无法同样有效地对其他领域建模。作为未来工作，我们计划利用从维基百科中抽取的概念及其定义，并扩充数据集，以便针对这些领域进一步微调*SciBERT* 模型。另一项未来工作是利用我们收集的点击数据作为代理信号，为相关概念提取和摘要片段排序组件训练监督模型。

参考文献

1. Beltagy, I., Lo, K., Cohan, A.：SciBERT：一种用于科学文本的预训练语言模型。载于：EMNLP-IJCNLP，第 3615–3620 页（2019） 2. Espinosa-Anke, L., Schockaert, S.：用于定义抽取的句法感知神经网络架构。载于：NAACL，第 378–385 页（2018） 3. Hearst, M.A.：从大型文本语料库中自动获取下位词。载于：COLING，第 539–545 页（1992） 4. Hochreiter, S., Schmidhuber, J.：长短期记忆。Neural computation**9**(8), 1735–1780（1997） 5. Jain, A., Gupta, N., Mujumdar, S., Mehta, S., Madhok, R.：利用概念定义和应用丰富形式文本内容。载于：第 29 届超文本与社交媒体会议论文集，第 96–100 页（2018） 6. Jin, Y., Kan, M.Y., Ng, J.P., He, X.：科学术语及其定义的挖掘：ACL 选集研究。载于：EMNLP，第 780–790 页（2013） 7. Kang, D., Head, A., Sidhu, R., Lo, K., Weld, D.S., Hearst, M.A.：学术文档中的文档级定义检测：现有模型、错误分析与未来方向。载于：第一届学术文档处理研讨会论文集，第 196–206 页（2020） 8. Klavans, J.L., Muresan, S.：自动构建和评估词典资源的方法。载于：LREC，第 231–234 页（2002） 9. Kobyli´nski, ˛L., Przepi´orkowski, A.：基于平衡随机森林的定义抽取。载于：国际自然语言处理会议，第 237–247 页（2008） 10. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.：应用于文档识别的基于梯度的学习。IEEE 论文集**86**(11), 2278–2324（1998） 11. Li, S., Xu, B., Chung, T.L.：使用 LSTM 循环神经网络进行定义抽取。载于：基于自然标注大数据的中文计算语言学与自然语言处理，第 177–189 页（2016） 12. Malais´e, V., Otten, A., Coupet, P.：Omniscience 及其扩展——设计多领域、多用例知识表示系统的经验教训。载于：欧洲知识获取研讨会，第 228–242 页（2018） 13. Murthy, S.K., Lo, K., King, D., Bhagavatula, C., Kuehl, B., Johnson, S., Borchardt, J., Weld, D.S., Hope, T., Downey, D.：Accord：一种用于生成科学概念多样化描述的多文档方法。arXiv 预印本 arXiv:2205.06982 （ 2022 ） 14. Navigli, R., Velardi, P.：学习用于定义和上位词抽取的词类格。载于：ACL，第 1318–1327 页（2010） 15. Reiplinger, M., Sch¨afer, U., Wolska, M.：从学术文章中抽取术语表句子：模式自举与深度分析的比较评估。载于：ACL-2012 关于重温 50 年研究发现的特别研讨会，第 55–65 页（2012） 16. Roig Mirapeix, M., Espinosa Anke, L., Camacho-Collados, J.：定义抽取特征分析：从规范定义到自然出现的定义。载于：词汇认知方面研讨会论文集，第 81–91 页（2020） 17. Schwartz, A.S., Hearst, M.A.：一种用于识别生物医学文本中缩写定义的简单算法。载于：生物计算 2003，第 451–462 页（2002） 18. Veyseh, A., Dernoncourt, F., Dou, D., Nguyen, T.：一种结合句法连接与语义一致性的定义抽取联合模型。载于：AAAI，第 9098–9105 页（2020）

19. Veyseh, A.P.B.、Dernoncourt, F.、Tran, Q.H.、Nguyen, T.H.：这个缩略词是什么意思？介绍一个用于缩略词识别与消歧的新数据集。载于：COLING。第 3285–3301 页（2020） 20. Westerhout, E.：使用语言学和结构特征进行定义抽取。载于：《第 1 届定义抽取研讨会论文集》。第 61–67 页（2009）