维基百科文章的链接预测是自然的语言推理任务

Chau-Thang Phan1,2,*, §、 Quoc-Nam Nguyen1,2, †, §和 Kiet Van Nguyen1,2, , ¶
[1]信息科学与工程学院,信息技术大学,胡志明市,越南
2越南国立大学,胡志明市,越南
{ *20520955, †20520644}@gm.uit.edu.vn, kietnv@uit.edu.vn

摘要:链接预测任务对于自动理解大型知识库的结构至关重要。在本文中,我们在 2023 年数据科学和高级分析竞赛“高效且有效的链接预测”(DSAA-2023 竞赛) [1]中展示了解决此任务的系统,该系统包含 948,233 个训练数据和 238,265 个用于公共测试的语料库。本文介绍了一种在维基百科文章中进行链接预测的方法,将其表述为自然语言推理 (NLI)任务。受到自然语言处理和理解领域最新进展的启发,我们将链接预测作为一项 NLI 任务,其中两篇文章之间存在链接被视为前提,任务是根据信息确定该前提是否成立文章中介绍。我们基于维基百科文章任务的链接预测的句子对分类来实现我们的系统。

我们的系统在公共和私人测试集上分别获得了 0.99996 Macro F1 分数和 1.00000 Macro F1 分数。我们团队UIT-NLP在私人测试集上的表现排名第三,与第一名和第二名的成绩持平。我们的代码1是公开用于研究目的的。

索引词 链接预测、自然语言推理、 DSAA2023竞赛

一、简介

维基百科2是世界上最大的协作百科全书,已成为获取广泛主题知识的宝贵资源。维基百科拥有数百万篇涵盖不同主题的文章,提供了一个不断扩展的庞大信息库。然而,尽管维基百科的规模令人印象深刻,但维基百科内的文章之间的相互链接并不总是全面的,导致信息连通性方面的差距。

链接预测是网络分析中的一项基本任务,旨在根据现有连接预测给定网络中缺失的链接。在维基百科的背景下,链接预测变得尤为重要,因为它可以帮助增强百科全书的导航性、提高信息可访问性并促进对相关主题的更深入理解。 DSAA 2023 挑战重点关注应用于维基百科文章的链接预测任务。此外, DSAA 2023 挑战赛[1]的重点是提出类网络数据结构中的链路预测方法,例如网络重建和网络开发,使用维基百科文章作为主要数据源。

自然语言推理 (NLI) 的任务是在给定“前提”的情况下确定“假设”是真(蕴含)、假(矛盾)还是不确定(中性) [2]。

维基百科文章任务的链接预测被定义为给出维基百科网络的稀疏子图并预测两个维基百科页面之间是否存在链接。

在本文中,我们利用自然语言推理任务和维基百科文章的链接预测之间固有的相似性。利用这种联系,我们将NLI 中常用的句子对分类 (SPC) 技术应用于链接预测任务的特定上下文。

本文通过结合基于 NLI 的 SPC 和预处理技术来解决维基百科文章的链接预测挑战。维基百科中链接预测的传统方法通常依赖于基于图的算法或文本相似性度量[3],这可能会忽略文章文本中嵌入的微妙关系。通过集成句子对分类和预处理技术,我们的目标是更有效地捕获维基百科文章中的语义和上下文信息,改进链接预测准确性。

我们的贡献总结如下:

- 首先,我们采用高效的数据预处理技术来清洗从维基百科获得的评论。这些技术的利用有助于提高数据的整体质量,并在为链接预测任务训练模型之前显着改进相关信息的提取。
- 其次,我们通过将广泛使用的SPC 技术从NLI 应用于链接预测,利用NLI 和维基百科文章链接预测之间的相似性。
- 最后,我们在这项任务上取得了最好的成绩,在公开测试中得分为 0.99996,在私人测试中得分为 1.00000 ,分别排名第八和第三。

二.相关作品

A. 链接预测

链接预测是网络结构数据中的一个关键问题[3],最初是作为监督学习任务引入的

[§]平等贡献 ¶
通讯作者
[1]https://github.com/phanchauthang/dsaa-2023-kaggle/
[2]https://www.wikipedia.org/

阿尔·哈桑等人。 [4]。他们还确定了监督学习框架内的关键性能特征。基于以图表形式分析用户-项目交互的概念， Huang 等人。 [5]利用最近网络建模文献中的链接预测技术来增强协同过滤推荐。相比之下,Liu和Lu- [6]提出了一种基于局部随机游走的方法,与其他基于随机游走的方法相比,该方法可以实现有竞争力甚至更好的预测,同时保持较低的计算复杂度。 Trouillon 等人提出的另一种方法。 [7]涉及通过潜在因式分解解决链接预测任务,从而能够扩展至具有线性空间和时间复杂度的大型数据集。此外,Zhang 和 Chen [8]探索了链接预测的启发式学习范式。此外,Negi 和 Chaudhury [9]将异构网络中的链路预测任务定义为多任务度量学习任务。

B. 自然语言推理

在人工智能领域,推理一直是一个突出的话题。虽然形式演绎的自动方法取得了重大进展,但自然语言推理（NLI)任务的进展相对较慢。 NLI 是指确定从自然语言前提(p)推断自然语言假设(h)的有效性的任务[10]。鲍曼等人。 [11]介绍了斯坦福自然语言推理语料库,该语料库由从事基于图像字幕的新颖任务的人类创建的标记句子对组成。该语料库可免费访问,是宝贵的资源。 Wang 和 Jiang [12]提出了一种基于长短期记忆 (LSTM) 网络的 NLI 专用架构。他们的方法利用匹配 LSTM在假设和前提之间进行逐字比较,从而增强了 NLI 的性能。陈等人。 [13]证明,使用链式 LSTM 仔细设计顺序推理模型可以在性能方面超越以前的模型。

三.算法技术

本节简要介绍了维基百科文章的链接预测任务定义、数据集、我们的预处理技术以及我们针对 DSAA 2023 挑战[1]提出的方法。

A. 任务定义

该任务涉及预测一对节点 (u, v) 之间是否存在边。在本次比赛中，我们的重点是维基百科文章的链接预测。更准确地说,给定维基百科网络的稀疏子图,目标是确定维基百科上下文中两个页面 u 和 v 之间是否存在链接。

B. DSAA-2023 数据集

比赛中使用的数据集是从维基百科中提取的,其中图形节点用文本注释。 DSAA-2023竞赛的数据包括以下内容：

- train.csv 文件:它包含节点对以及它们之间是否存在边的指示。该文件有四列： id、 id1、 id2和label。 id列表示配对标识符， id1和id2是节点 ID,label 列声明是否存在边（0 或 1）。
- nodes.zip 文件:此文件是nodes.tsv 的压缩版本,包含有关每个节点的信息。它由两列组成:id（节点标识符)和text （节点的文本描述）。
- sample_submission.csv 文件:此演示提交文件有两列:对 ID 和相应的标签（0 或 1）。

将train.csv和nodes.tsv文件组合在一起后训练集标签的统计数据如表1所示。

	不相关 (0) 相关 (1)	
频率 512,389 (%)	435,843 百分比 54.03	(%) 45.97

表 I:训练集标签的统计数据。

C. 预处理技术

该数据集由维基百科对象的主要内容和维基百科内的 CSS 代码组成。然而,这些 CSS 代码被认为是噪声,不需要进行训练。因此,去除它们至关重要。为了实现这一目标,我们采用两种算法,即平衡花括号算法和删除双花括号算法,分别为算法 1 和算法 2,以有效消除这些不必要的代码。标点符号（例如句号、逗号、问号、感叹号等)和冗余空格可能会在训练过程中引入噪音并扰乱文本的自然流动。因此,我们应用正则表达式库3来删除它们。我们针对 DSAA-2023 竞赛任务的预处理技术总结如下:平衡大括号、删除双大括号、删除所有标点符号以及删除多余空格。

D. 我们提出的方法

我们的方法从最近自然语言理解任务的成功中汲取了灵感,特别是在预训练的语言模型中。我们假设两篇维基百科文章之间的链接可以被视为前提,预测链接的存在类似于确定前提和假设之间的逻辑关系。我们将两篇文章的内容分别编码为前提和假设,并利用此 NLI 设置进行链接预测。

在句子对分类中,模型以两个句子作为输入,旨在确定它们的关系。它学习根据所需的任务将配对分为不同的类别。维基百科文章的链接预测任务是预测两个给定节点之间是否存在边或链接。

在本文中,我们实现了基于XLM-Roberta（一种著名的架构)的句子对分类模型

[3]https://docs.python.org/3/library/re.html

算法 1平衡大括号

```
程序BALANCECURLYBRACES（文本）
    opening_count ← text.count( { )
    结束计数 ← text.count( } )
    如果opening_count > close_count则
        当opening_count > opening_count时执行
            索引 ← text.find( { )
            如果索引 ≠ −1则
                文本 = 文本[:索引] + 文本[索引 + 1:]
                开盘计数 ← 开盘计数 − 1
    否则如果close_count > opening_count then
        当close_count > opening_count时执行
            索引 ← text.rfind( } )
            如果索引 ≠ −1则
                文本 = 文本[:索引] + 文本[索引 + 1:]
                关闭计数 ← 关闭计数 − 1
    返回文本.strip()
```

算法 2删除双花括号

```
程序REMOVEDOUBLECURLYBRACES（文本）
    堆栈 ← []
    clean_text ← for
    char in text do
        如果char = { 那么
            堆栈.push(字符)
        否则如果char = } 那么
            如果不是 isEmpty(stack)并且
                顶部（堆栈)= { 然后
                堆栈.弹出（堆栈）
        别的
            如果为空（堆栈)则
                clean_text ← clean_text + 字符
    返回干净的文本
```

Conneau 等人的工作中介绍了这一点。 （2020） [14],其中广泛用于自然语言推理（NLI）。我们的实施是针对通用和特定维基百科量身定制的文章链接预测任务。此外,我们还整合了一个XLM-Roberta 架构之上的线性预测层来计算最终的输出。我们建议的细节图 1 详细介绍了该方法的架构。

四.实验结果

A. 实验配置

我们遵循相当标准的微调实践,大多数其中[15]中有描述。我们使用的批量大小为 128，最大令牌长度为 128,学习率为 2e-5,并且AdamW 优化器,epsilon 为 1e-8。

我们凭经验为我们的 SentencePair 分类模型系统使用simpletransformers[4] 。我们训练我们的模型1×RTX4090 GPU 在 Vast.ai[5]平台上运行四个小时。十预测该数据集的测试集需要几秒钟的时间。

B. 实验结果

在本节中,我们将描述我们的实验和结果链接预测任务。仅宏 F1-score 用于评估。实验结果如表二所示

通过实施稳健且高效的数据预处理专门为清理所获得的评论而定制的技术来自维基百科,我们的目标是提高整体数据质量并大幅改进之前提取相关信息的能力训练链接预测任务的模型。结合这些句子对分类的预处理技术与未应用此类技术的场景相比,该模型带来了显着的性能改进。

我们提出的方法取得了显着的成果，公开测试的宏F1得分为0.99996,完美私人考试成绩为 1.00000。这些优异的成绩准确地表明我们方法的有效性预测给定上下文中的链接。从排名来看，我们在公开测试中获得了第八名,并且取得了令人印象深刻的成绩私人测试第三名。

	公开测试	私人测试
具有预处理技术		
我们的方法	0.99996	1.00000
无预处理技术		
我们的方法	0.97680	0.97663

表 II:以宏 F1 分数表示的实验结果。

五.结论

链接预测任务的意义在于它的能力理解广泛的知识库的结构自动地。在本文中,我们提供了详细的说明我们提出的解决这项任务的方法2023 年数据科学和高级分析的背景竞赛题为“高效且有效的链接预测”。为了解决维基百科文章的链接预测任务,我们实现了基于自然的句子对分类语言推理和高效的预处理技术管道。值得注意的是,我们的系统取得了卓越的性能公共测试集的宏观 F1 分数为 0.99996，私人测试集满分 1.00000。因此，我们的系统在私人测试集上排名第三,等于第一名和第二名的成绩。

致谢

这项研究得到了 VNUHCM 大学的支持信息技术科学研究支持基金。

参考

[1] AN Papadopoulos，“Dsaa 2023 竞赛”,2023 年。[在线的]。可用:https: //kaggle.com/competitions/DSAA-2023-竞赛

[2] S. Storks、Q. Gau 和 JY Chai，“最近自然语言推理的进展:一项调查基准、资源和方法”， arXiv

[4] https://simpletransformers.ai/（版本 0.63.11）

[5] https://vast.ai

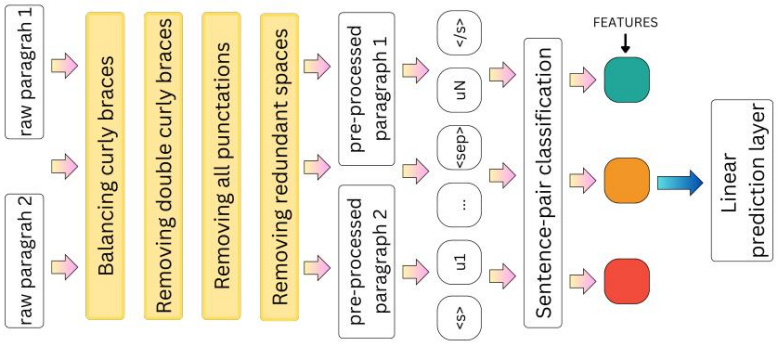

图 1:我们用于维基百科文章链接预测的方法架构。

预印本 arXiv:1904.01172,2019。 [在线]。可用： https: //arxiv.org/abs/1904.01172

[3] A. Kumar、SS Singh、K. Singh 和 B. Biswas，“链接预测技术、应用和性能:一项调查”， Physica A:统计力学及其应用,卷。 553,p。 124289, 2020。[在线]。

可用:https: //www.sciencedirect.com/science/article/pii/S0378437120300856 [4]

M. Al Hasan、V . Chaoji、S. Salem 和 M. Zaki，“使用监督学习进行链接预测”,载于SDM06 :链接分析、反恐与安全研讨会,卷。 30,2006 年,第 798–805 页。

[5] Z. Huang、X. Li 和 H. Chen，“协同过滤的链接预测方法”,第五届 ACM/IEEE-CS 数字图书馆联合会议论文集， 2005 年,第 141-142 页。 [在线的]。可用： https://iopscience.iop。 org/article/10.1209/0295-5075/89/58007/meta [6] W. Liu 和 L. Lu，“基于局部随机游走的链路预测”,欧洲物理学快报,卷。 89,没有。 5,p。 58007,2010。[在线]。可用： https： //doi.org/10.48550/arXiv.1001.2467 [7] T. Trouillon、J. Welbl、S. Riedel、E. Gaussier 和 G. Bouchard，“简单链接预测的复杂嵌入”,第33 届国际会议论文集机器学习会议,系列。机器学习研究论文集,MF Balcan 和 KQ

温伯格编辑,卷。 48. 美国纽约州纽约市:PMLR,2016 年 6 月 20-22 日,第 2071-2080 页。 [在线的]。
可用:https: //proceedings.mlr.press/v48/trouillon16。 html

[8] M. 张和 Y. Chen，“基于图神经网络的链接预测”,《神经信息处理系统进展》， S. Bengio、H. Wallach、H. Larochelle、K. Grauman、N. Cesa Bianchi,和 R. Garnett,编辑,卷。 31.Curran Associates，2018。

公司， [在线的]。 可用：
https: //proceedings.neurips.cc/paper_files/paper/2018/file/53f0d7c537d99b3824f0f99d62ea2428-Paper.pdf [9]

S. Negi 和 S. Chaudhury，“异构社交网络中的链接预测”系列。 CIKM 16。
美国纽约州纽约市:计算机协会,2016 年,第 14 页。 609–617。 [在线的]。可用的：
https://doi.org/10.1145/2983323.2983722

[10] B. MacCartney,自然语言推理。
斯坦福大学,2009。 [在线的]。
可用： https://www.proquest.com/openview/1f496dd128e01b6c0b2a030d2a2447f8/1 ?pq-origsite=gscholar&cbl=18750

[11] SR Bowman、G. Angeli、C. Potts 和 CD Manning，“用于学习自然语言推理的大型注释语料库”,《 2015 年自然语言处理经验方法会议论文集》。葡萄牙里斯本:计算语言学协会,2015 年 9 月,第 632–642 页。

[在线的]。可用:https: //aclanthology.org/D15-1075 [12] S. Wang 和 J. Jiang，“用 LSTM 学习自然语言推理”,计算语言学协会北美分会 2016 年会议记录:人类语言技术。加利福尼亚州圣地亚哥:计算语言学协会,2016 年 6 月,第 1442–1451 页。 [在线的]。

可用:https: //aclanthology.org/N16-1170 [13] Q. Chen、X. Zhu、Z. Ling、S. Wei、H. Jiang 和 D. Inkpen， “用于自然语言推理的增强型 lstm”,计算语言学协会第 55 届年会记录（ACL 2017）。温哥华： ACL，2017 年 7 月。

[14] A. Conneau、K. Khandelwal、N. Goyal、V. Chaudhary、 G. Wenzek、F. Guzmán、E. Grave、M. Ott、L. Zettlemoyer和 V. Stoyanov，“无监督跨语言表示”大规模学习” ,载于计算语言学协会第 58 届年会论文集。在线:计算语言学协会,2020 年 7 月,第 8440–8451 页。 [在线的]。

可用： https ://aclanthology.org/2020.acl-main.747 [15] J. Devlin, M.-W. Chang、K. Lee 和 K. Toutanova，“Bert:用于语言理解的深度双向转换器的预训练”,载于计算语言学协会北美分会 2019 年会议记录:人类语言技术,第 1 卷（长论文和短论文）， 2019 年,第 4171–4186 页。 [在线的]。可用：
https: //aclanthology.org/N19-1423/