重新训练用于迁移学习的 BERT 模型需求工程：初步研究

Muideen Ajagbe 和赵丽萍
曼彻斯特大学计算机科学系
牛津路，曼彻斯特，M13 9PL，英国

2022年IEEE第30届国际需求工程会议（RE）| 978-1-6654-7000-1/22/$31.00 ©2022 IEEE | 978-1-6654-7000-1/22/$31.00 DOI：10.1109/RE54965.2022.00046

抽象的——近年来，BERT、ELMO、ULMFiT、GPT等先进的深度学习语言模型在许多通用自然语言处理（NLP）任务上表现出了强大的性能。特别是 BERT 在一些特定领域的任务上也取得了可喜的结果，包括需求分类任务。然而，尽管 BERT 潜力巨大，但它在特定领域任务上的表现却不佳。在本文中，我们提出了 BERT4RE，一种基于 BERT 的模型，在需求文本上进行了再训练，旨在支持广泛的需求工程 (RE) 任务，包括需求分类、检测语言问题、识别关键领域概念以及建立需求可追溯性链接。我们通过对其识别关键领域概念的任务进行微调，展示了 BERT4RE 的可转移性。我们的初步研究表明 BERT4RE 取得了比 BERT 更好的结果根据演示的 RE 任务的模型。

索引术语—需求工程、需求 分类、语言模型、BERT、特定领域语言模型、迁移学习、深度学习、机器学习、自然语言处理。

1. 我简介

基于 Transformer 的语言模型 (LM) [1] 的最新发展，如 BERT [2]、ELMO [3]、GPT [4]、ULM-FiT [5] 代表了自然语言处理 (NLP) 的重大突破。这些模型用上下文化的单词表示或上下文嵌入代替单词的静态向量表示（例如，word2vec [6]），几乎在每个 NLP 任务上都取得了显着的改进 [7]，在许多领域取得了最先进的结果。最常见的 NLP 基准任务 [2]、[8]。特别是 BERT，它的性能优于其他 LM [7]。

实际上，这些先进的 LM 已经实现了迁移学习 [9]、[10] 的长期梦想，因为它们可以使用*无监督学习*大量的*未标记的，通用的*文本语料库，*然后*对*菲微调*执行下游任务，通过*监督学习*在*带标签的*、*特定于任务的文本*。语言模型的迁移学习能力对于寻求使用这些模型来处理文本密集型任务（例如需求工程（RE）中的任务）的研究人员来说非常有吸引力[11]，因为这些模型可以（重新）用于除它们本身的任务之外的任务。进行了训练，几乎没有进行微调[12]。

然而，尽管这些 LM 取得了强劲的性能，但研究人员发现通用 LM 在特定领域任务上表现不佳，因为它们无法识别高度特定领域的词汇 [13]-[16]。为了解决这一不足，人们进行了多次尝试*再训练*针对特定领域文本的通用 BERT 模型，这些尝试提高了特定领域任务的性能 [13]-[16]。

在本文中，我们提出BERT4RE，A再训练的特定领域的语言模型[1]旨在支持广泛的需求工程（RE）任务，例如需求分类、语言问题检测、领域概念识别以及需求追溯链接的建立。BERT4RE 在通用 BERT 上重新训练根据模型 [2] 使用公开的 RE 相关文本。为了证明 BERT4RE 的可转移性，在这项初步研究中，我们针对特定 RE 任务对 BERT4RE 进行了微调，即从需求文本中识别关键领域概念（因此称为*概念提取*）。然后我们比较 BERT4RE 和 BERT 的性能根据通过在同一标记数据集上微调两个模型。

在本文中，我们通过以下贡献为构建 RE 特定 LM 奠定了基础：

- 我们提出了 BERT4RE，一种用于 RE 的特定领域的 LM，并且我们描述了如何从 BERT 重新训练该模型根据使用 RE 相关数据的模型。
- 我们对 BERT4RE 和 BERT 进行了微调根据使用带标签的数据集，然后将它们应用于识别九个需求概念的多类分类任务。
- 我们证明 BERT4RE 优于 BERT根据并讨论每个模型获得的结果。我们在 Zenodo [18] 上提供 BERT4RE，供 RE 研究人员和从业者进行实验。

本文的结构如下：第 2 部分讨论针对特定领域任务的再训练和微调 BERT 模型的相关工作。第 3 节描述了我们如何使用 RE 相关数据集重新训练 BERT4RE。第 4 节介绍了一项实验研究，其中我们对 BERT4RE 和 BERT 进行了微调根据在通用数据集上比较它们的性能。第 5 节报告了我们的研究结果，第 6 节讨论了我们研究的潜在有效性威胁。最后，第七节总结了本文并概述了我们计划如何推进这项研究。

[1]我们区分预训练和再训练，前者是指在通用数据上预训练 LM，后者是指在特定领域数据上重新训练预训练模型。模型再训练的想法类似于领域适应[17]。

978-1-6654-7000-1/22/$31.00 ©2022 IEEE
DOI 10.1109/RE54965.2022.00046

2.R兴高采烈瓦奥克

在本节中，我们将简要回顾一些密切相关的工作，重点关注针对特定领域任务的 BERT 模型的再训练和微调工作。

A. 再训练 BERT 模型

如前所述，BERT 模型已针对多个领域进行了重新训练，以提高其在特定领域任务上的性能。在本节中，我们简要总结了重新训练特定领域 BERT 模型的一些努力。

据我们所知，Sainani 等人。[13] 报道了 RE 中重新训练 BERT 的唯一努力。作者重新训练了 BERT根据模型将合同义务分为不同的需求类型。他们使用相同的软件合同数据集来重新训练和微调 BERT 模型。他们发现保留的 BERT 略微优于其他学习技术（SVM、随机森林、Naive Baye 和 BiLSTM）。然而，由于作者使用相同的数据集进行模型重新训练和微调，因此重新训练的模型是特定于任务的，而不是特定于领域的，因为该模型仅在特定于任务的数据集上进行了重新训练。

在 RE 之外，还有更多的努力来重新训练 BERT 模型。其中包括 BioBERT [14]，其中 BERT根据模型在 pubMEd 的生物医学文章语料库上进行了 470,000 次迭代的重新训练。事实证明，与通用 BERT 模型相比，BioBERT 可以提高生物医学文本分类的性能。

CLINICALBERT [19] 和 CLINICALBioBERT [20] 根据 MIMIC-III 数据集 [21] 的临床文本进行重新训练，并分别在 BioBERT 上微调 15 万步。这些模型的性能优于 BERT根据模型。

此外，SCIBERT [15] 是另一个 BERT 模型变体，在语义学者语料库中的科学文本上进行了重新训练。具体来说，该模型在 114 万篇生物医学和计算机科学文本上进行了重新训练。SCIBERT是通过再训练BERT而开发的根据SCIVOCAB 模型使用其原始词汇和另一个变体，是通过从头开始重新训练 BERT 使用领域内词汇开发的。

BERTweet [22] 是 BERT 的一个变体根据模型在英语推文上进行了重新训练。BERTweet 被重新训练了 950,000 次迭代。与 RoBERTa 和 XLM-R 等其他 LM 相比，该模型在 POS 和 NER 任务上的性能得到了提高。

另一种 BERT 变体模型是 LEGAL-BERT [16]，它是通过在 BERT 上运行额外的再训练步骤而开发的根据带有英文法律文本。LEGAL-BERT 在带注释的法律数据集上进行了微调，用于文本分类和序列标记任务，取得了令人印象深刻的结果。

B. 微调 BERT 模型

大多数 RE 领域已发表的作品都专注于使用微调的 BERT 模型来执行 RE 任务。其中一件作品是 Hey 等人的作品。[12]，他提出了 NoRBERT，一种用于需求分类的迁移学习方法。NoRBERT 应用微调的 BERT 模型（BERT根据和伯特大的）适用于二元分类和多类分类任务。实验表明 NoRBERT 在最常见的非功能性需求 (NFR) 类别（例如安全性和性能）上可以达到 90% 以上的 F1 分数；NoRBERT 通常优于传统的非迁移学习方法。

李等人。[23]提出了DEMAR，一种用于需求分析的深度多任务学习方法。DEMAR 对 BERT 进行了微调根据模型来识别和注释来自开放论坛的需求。实验表明，在 8 个开源项目中，DEMAR 在需求发现任务上取得了 91% 的精确度和 83% 的召回率，优于传统的机器学习方法。

此外，王等人。[24]提出了 DEEPCOREF，一种用于需求中共指检测的迁移学习方法。DEEPCOREF 微调 BERT根据模型来检测和解决需求中的实体共指。对需求文档的实验表明，DEEPCOREF在行业合作伙伴数据上可以达到96%以上的精确率和召回率。DEEPCOREF 优于传统和非迁移学习方法。

Araujo 和 Marcacini 的 RE-BERT [25] 对 BERT 进行了微调根据从应用程序评论中提取需求语义表示的模型。对八个不同应用程序的手动标记数据进行的实验表明，RE-BERT 的 F 分数为 62%，优于现有的传统和非迁移学习 SOTA 结果。

同样，Pota 等人。[26]提出BERT的微调根据支持意大利推文情绪分析的模型。实验表明，所提出的方法在带注释的意大利情绪数据上优于其他迁移学习和非迁移学习方法，F 分数为 75%。

3.BERT4RE：R电子培训伯特为了关于

A. 再训练方法

BERT LM 的通用版本在维基百科（25 亿字）和 BooksCorpus（8 亿字）上进行了预训练 [2]。预训练涉及让 BERT 模型学习两个无监督任务，称为“掩蔽 LM (MLM)”和“下一句话预测 (NSP)”。BERT 以两种基本模型大小进行预训练：BERT根据（L=12，H=768，A=12，总参数=110M）和BERT大的（L=24，H=1024，A=16，总参数=340M），其中“L”是Transformer块或层的数量，“H”是隐藏大小，“A”是自学习的数量注意头。

在我们目前的初步研究中，我们遵循了之前涉及再训练 BERT 的工作（参见第 2.1 节）并选择了 BERT根据作为我们的模型尺寸。我们通过无监督学习在 RE 相关数据集上重新训练该模型，以便模型学习这些数据集给出的特定领域词汇。这个再训练过程，如图1所示，与训练原始BERT的过程基本相同根据模型。我们一层一层地重新训练了所有 12 个注意力层。再训练还需要模型学习特定领域词汇上的 MLM 和 NSP 任务。再训练过程的结果是 RE 特定的 BERT根据模型称为 BERT4RE。

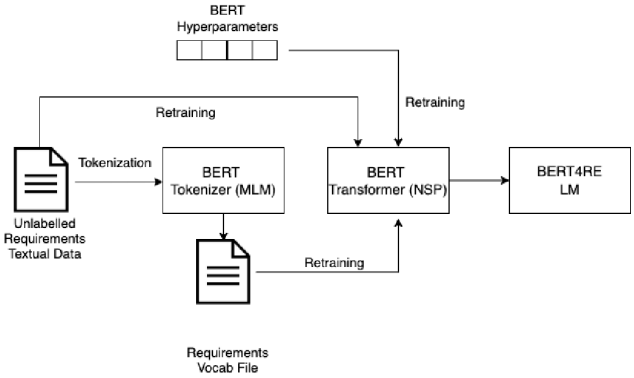

图 1. RE 重新训练 BERT 的无监督学习过程。再训练的结果是一个名为 BERT4RE 的 RE 特定模型。

以下小节描述了我们的再培训中使用了哪些数据以及我们如何准备它们。

B. 再训练数据

*1）数据收集：*为了重新训练 BERT 模型，我们从不同来源收集了以下四个与需求相关的数据集：

- PROMISE NFR 数据集 [27]：这是一个流行的数据集，被 RE 研究人员广泛用于训练和测试他们的机器学习分类器 [12]、[28]–[30]。该数据集包含来自 15 个软件开发项目的 625 个需求（255 个 FR 和 370 个 NFR）。

- PURE 数据集 [31]：这是最大的 RE 数据集之一。该数据集包含 522,444 个词汇和 865,551 个标记。我们只收集了与 FR 相关的部分数据集，因为这些需求陈述结构良好且易于识别。提取的需求包含 29,000 个唯一单词。

- 应用程序评论数据集：我们从 Appfollow 网站 (appfollow.io) 下载了 400 万个应用程序评论。这些评论的提交时间为 2010 年 1 月至 2021 年 4 月。我们从这些评论中提取了超过 300 万个独特单词。

- Google Play 商店应用程序评论[2]：我们从该应用程序商店收集了大约 60 万条应用程序评论，并从这些评论中提取了超过 200 万个独特单词。

这些数据集总结在表 I 中，总共包含 7,758,173（超过700万）字。我们将所有这些数据集组合成一个数据集（语料库）以进行再训练
伯特根据。

*2）数据预处理：*再训练数据集需要是在用于重新训练 BERT 模型之前，将其清理并处理为单词标记和句子。为此，我们将以下 NLP 管道应用于数据集中的文本：

- 标记化：此步骤将文本转换为标记。我们使用了斯坦福核心 NLP 分词器[3]打破我们的输入

2https://www.kaggle.com/gauthamp10/google-playstore-apps
3https://stanfordnlp.github.io/CoreNLP/tokenize.html

表一
右要求D文件为右电子培训BERT4RE。

数据集	# 字	描述
承诺 NFR数据集	10,629	652 标记功能性和非功能性需求
纯数据集	38,420	健康、铁路等多个应用领域的 79 份需求文件。
应用程序评论	5,101,685	社交媒体应用程序（例如 twitter、whatsapp 和 snapchat）上大量缓存的应用程序评论集合。
应用商店 谷歌应用套件 用户评论	2,607,439	Google Play 商店应用程序上的应用程序的用户评论数据集。

将需求文本转换为单词标记。我们承认应用程序评论在措辞上更接近推文，因此我们使用 NLTK TweetTokenizer 工具包对与应用程序需求相关的应用程序评论文档进行标记化，因为缺乏专门为它们量身定制的标记器。

- 表情符号/表情符号转换：应用程序评论中使用表情符号和表情符号来传达和代表评论者的情绪。这意味着表情符号是评论的强烈表达，需要翻译成文本。为了处理应用程序评论中的表情符号，我们使用 python emoji 包[4]将表情符号翻译成文本字符串。在这里，每个表情符号图标都被翻译成一个单词标记。

- 规范化：此过程涉及将文本或图标转换为自然语言形式，从而减少其随机性。为了实现这一点，我们使用 langdetect python 库过滤应用程序评论文档[5]来自 Google 的语言检测库，仅检测英语评论。规范化的第二步包括将文本转换为小写，删除数字和特殊字符、标点符号和空格。少于 2 个单词的句子被删除，URL 链接被转换为特殊标记。

- 句子拆分：此步骤使用斯坦福自然语言处理句子拆分器根据句点分隔符将文本拆分为句子[6]。

该管道的输出由一组预处理的需求和需求相关文本组成。这些文本数据现在可用于重新训练 BERT 模型。

C. 再培训程序

为了重新训练 BERT4RE 模型，我们执行以下步骤：

首先，为了生成每个训练句子，我们自动扫描数据集以确定句子的最大长度（即最大单词数）。我们

4https://pypi.org/project/emoji/ 5https://pypi.org/project/langdetect/ 6https://stanfordnlp.github.io/CoreNLP/ssplit.html

发现最长的句子包含 90 个单词。继 Nguyen 等人之后。[22]，我们使用 FastBPE [32] 使用 32,000 个子词生成的词汇来分割数据集中的每个句子。每个子词都是单词的一部分，传达某种含义。例如，单词“subword”有两个子词：“sub”和“word”。我们的数据集中每个句子的平均子词标记数量为 25。将单词拆分为子词的原因是为了使单词更细粒度，以便更容易标记化，并为 LM 提供更多未知或罕见的单词。

接下来，我们按照 RoBERTa [33] 重新训练 BERT根据并定义用于重新训练的超参数。我们将批量大小设置为 128 个序列和 3200 个标记/批次（128 * 25 子字标记）。对于迭代步骤，我们将数据集中的单词数量（700 万）乘以子单词标记的数量（25），然后除以批量大小（128）以生成序列块（7M * 25/ 128 = 1.3） M 序列块）。我们重新训练 BERT根据40 个 epoch 相当于 1.3M * 40/ 3200 = 17, 000 个步骤。

最后，我们使用 Adam [34] 进行优化，学习率为 1e-4，权重衰减为 0.01，dropout 率为 0.1。重新训练是使用带有单个 8 核处理器的 Google Cloud TPU v3-8 执行的。

4.F内科-T尤宁BERT4RE和伯特根据

一、*研究方法*

本节介绍了一项实验研究，其中我们对 BERT4RE 和 BERT进行了微调根据执行从需求文本中识别关键领域概念的多类分类任务。本研究的目的是证明 BERT4RE 的可迁移性，并将再训练模型与预训练模型的性能进行比较。下面，我们描述用于微调这两个模型的数据集、微调过程和评估方法。

B. 微调数据集

用于微调的数据集基于 PURE 数据集 [31]。我们一共收集了1,055 个 FR 句子从这个数据集中。提取的 FR 包含22,916 字，句子的最大长度为 90 个单词。提取的 FR 会进行词形还原、词性标记和解析，以便轻松标记。然后，我们根据从 Fillmore 案例框架 [35] 和 SOM 模式 [36] 中提取的九个语义概念（或案例）手动标记每个需求。下面我们简单介绍一下这九个概念。

- 代理（AGT）：这是导致对象发生变化的无生命的发起者或动作的主体。例如，在“系统应将用户信息保存到数据库”中，“系统”一词就是一个代理。
- 对象（OBJ）：这是执行操作的关键对象。例如，“系统应将用户的详细联系方式保存到具有指定角色的数据库中”中，术语“用户详细联系方式”就是关键对象。
- 仪器（INT）：这是代理用来执行操作的工具。例如，在“视力受损的用户应该能够阅读文本”

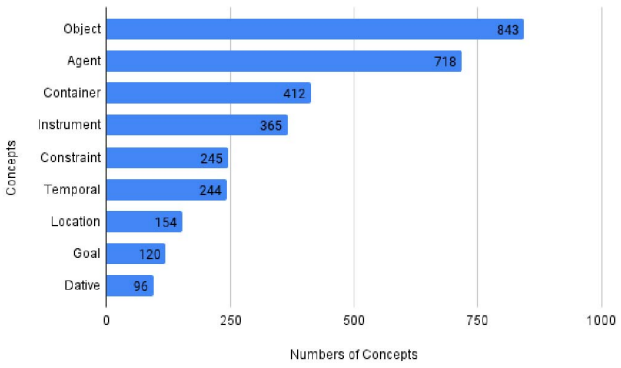

图 2. 标签中使用的各个关键领域概念的实例
我们的微调数据集。

特殊放大镜功能”，术语“特殊放大镜功能”就是仪器。
- 容器（续）：这是存储代理使用的信息的结构对象。例如，在“系统应确保用户输入自动存储在文件中”中，术语“文件”就是容器。
- 与格（DAT）：与格是表示动作的间接宾语的情况。例如，在“系统收到用户付款后应向用户发送电子邮件确认”中，术语“电子邮件确认”就是与格。
- 地点（LOC）：这是由动作标识的位置。例如，在“项目团队每周二上午 9 点至 10 点在员工室开会”中，术语“员工室”就是地点。
- 时间（TEMP）：这是通过动作表达的时间、日期或频率的发生。例如，在“项目团队应每周二上午 9 点至 10 点在员工室开会”中，术语“每周二上午 9 点至 10 点”是时间性的。
- 目标（目标）：这是通过行动要达到的预期效果。例如，在“系统将为每个注册用户分配一个唯一的标签”中，术语“唯一标签”就是目标。
- 约束（CONST）：这是动作的条件。例如，在“用户的个人信息发生变更时，系统应当通知用户”中，“个人信息发生变更时”就是约束条件。

图 2 显示了这些概念在提取的 FRS 及其实例中的分布。

C. 微调程序和评估方法

微调 BERT4RE 和 BERT根据模型来执行具有 9 个类别的多类别分类任务 (即概念），我们采取以下步骤：

首先，我们使用Python的模块Pytorch Dataset Class准备标记数据集，将数据集中的句子表示为1055 ×90大批 （第1055章 要求句和一个 每句话最多 90 个单词）。

接下来，我们在原始BERT的基础上定义各个模型的超参数进行微调根据[2]。具体来说，我们将数据集的批量大小设置为16，句子的最大长度为 90，标签数量为 9，学习率为 1e-4，历元为 20，权重衰减为 0.01，dropout 率为 0.1。我们还将每个模型的分词器定义为 BertTokenizerFast.from pretrained('bert-base-uncased')。

然后，我们将数组随机分成 70-30 份作为训练集和测试集。最后，我们使用训练集微调每个模型，然后使用测试集测试每个微调模型。微调和测试都涉及让这些模型根据九个标签将每个输入句子分类为一个或多个类别。每个模型的微调都是使用定义的超参数完成的。

对两个模型进行微调后，我们通过用于确定模型如何适合微调数据集的指标来评估其训练损失。训练损失越小意味着模型拟合越好。我们发现BERT4RE的平均训练损失为0.0007，明显低于BERT的训练损失根据，平均值为 0.005。

衡量 BERT4RE 和 BERT 的性能根据
对于单个类别（即概念），我们使用未加权精度（P）、召回率（R）和F1分数指标来衡量这些模型的测试结果。更具体地说，为了精度，我们测量正确分类的概念出现次数相对于已识别概念总数的百分比。为了回忆，我们测量正确分类的概念出现的百分比。对于 F1-Score，我们通过调和平均值聚合精度和召回率来衡量微调模型的性能 [37]。

为了衡量每个模型在所有类别上的整体性能，我们使用*宏观平均*和 *微观平均*[37]。具体来说，宏观平均独立计算每个类别的精度、召回率和 F1，然后取每个指标的平均值，从而平等地对待所有类别，无论其大小如何。因此，宏观平均值减少了不平衡类对性能结果的影响。另一方面，微平均聚合所有类别的贡献来计算平均精度、召回率和 F1。因此，微平均精度、召回率和 F1 都等于*准确性*。

表 II 展示了经过微调的 BERT4RE 和 BERT 取得的性能结果根据各个类别的模型及其宏观和微观平均值。该表按 BERT4RE 的 F1 分数排序。在下一节中，我们将讨论这些结果。

5.R结果

我们可以从表二中得出以下观察结果：

- BERT4RE的宏观平均P、R、F1比BERT高7%根据，表明 BERT4RE 比 BERT 更好地处理微调数据集中的不平衡类别根据。

表二
磷Fine 取得的绩效结果-调整两者伯特根据和BERT4RE关于识别9要求概念。

概念	伯特根据 磷（%）	电阻率（%）	F1 (%)	BERT4RE 磷（%）	电阻率（%）	F1 (%)
AGT	0.94	0.88	0.91	0.96	0.94	0.95
继续	0.88	0.94	0.91	0.92	0.94	0.93
INT	0.86	0.89	0.88	0.92	0.89	0.91
温度	0.74	0.95	0.83	0.87	0.95	0.91
OBJ	0.85	0.91	0.88	0.90	0.92	0.91
LOC	0.80	0.92	0.85	0.86	0.91	0.88
目标	0.75	0.73	0.74	0.93	0.79	0.86
常量	0.89	0.89	0.89	0.90	0.82	0.86
数据传输技术	0.88	0.45	0.59	0.89	0.81	0.85
准确性			0.88			0.95
宏观平均	0.84	0.86	0.85	0.92	0.91	0.92
微平均	0.88	0.88	0.88	0.95	0.95	0.95

- BERT4RE的微平均P、R、F1比BERT高7%根据，表明 BERT4RE 优于 BERT根据并且更加准确。

当查看这两个 mod8ls 在各个类别上的性能结果时，表 II 还揭示了以下发现：

- BERT4RE 优于 BERT根据在 9 个类别（主体、容器、工具、与格、位置、对象、目标和时间）中的 8 个上取得了超过 90% 的 F1 分数，并在五个类别（主体、容器、工具、对象和时间）上获得了超过 90% 的 F1 分数。由于 Location、Goal 和 Dative 是较小的类别，BERT4RE 的性能优于 BERT根据在这些类别上的结果表明，BERT4RE 不仅优于最先进的模型，而且还提高了不平衡类别的性能。特别是，BERT4RE 优于 BERT根据在最小类与格上，F1 分数为 85% – 比 BERT 高 26%根据。

- BERT4RE 的 F1 分数比 BERT 低 3%根据在约束类上。这可能是因为 Constraint 类通常具有“If ..., then ...”或“While ...”等句子结构，并且这些结构已经由原始的预训练 BERT 模型学习。相比之下，我们的再训练数据集中的大多数句子都是简单的句子，没有“If”、“then”和“While”等词，因此再训练的模型比原始的预训练 BERT 模型的词汇量有限。

上述结果清楚地证明了重新训练 BERT 的好处和潜力根据模型，即使数据量很小。

6.T威胁到V活力

本节讨论我们工作的潜在有效性威胁。这些威胁涉及我们的再训练数据集和微调数据集、模型再训练和微调的实验设计、模型验证的评估方法、我们的研究结果以及结果的普遍性。

*再训练数据集：*我们的再训练数据集包含 700 万个单词，与用于的再训练数据集相当

SCIBERT（114 万字）[15]。然而，其他特定领域的 LM 使用比我们大得多的数据集进行重新训练，其中 BiOBERT [14] 使用 135 亿个单词进行训练，BERTweet [22] 使用 8.5 亿个单词进行训练，LEGAL-BERT [16] 使用 60 亿个单词进行训练。尽管如此，我们相信我们的再训练数据集不会对我们的研究结果构成严重威胁。

*微调数据集：*由于没有适合我们概念提取任务的标记数据集，我们自己开发了一个微调数据集。为了最大限度地减少偏差，我们使用来自不同领域的需求文档组成了数据集（第 3.21 节）。为了确保概念清晰可辨，我们仅选择标签的功能要求。为了保证标注的可靠性，我们聘请了两名具有丰富领域知识的需求工程师来独立对数据集进行标注，并通过讨论解决对标注结果的分歧。为了确保注释者对要标记的概念有共同的理解，我们通过定义用于标记的不同类别的语义角色和关系标签来提供清晰的注释方案。尽管我们付出了努力，标记数据集的大小仍然受到限制，因为它仅包含 1055 个标记句子，最大句子长度为 90。为 RE 任务生成更多标记数据集仍然是在 RE 中使用监督学习方法的公开挑战[31]。

*实验设计：*一个潜在的内部有效性威胁涉及我们的实验设计，为了减轻这种威胁，我们遵循 Kitchenham 等人的实验设计指南。[38]。对于再训练和微调实验，我们确保我们的实验程序得到清晰解释并且设计选择合理。因此，我们相信我们的实验设计不会受到严重威胁。

*评价方法：*由于我们的数据集很小，我们没有验证 BERT4RE 和 BERT根据使用标准的型号 k 倍交叉验证[39]；相反，我们采用了常见的保留方法，将数据集分成 70-30 个，使用 70% 的数据进行训练，剩余 30% 的数据进行测试。我们使用标准精度、召回率和 F1 分数指标来衡量模型的性能。我们注意到其他RE研究人员也应用了项目级特定的交叉验证方法，例如p 折叠 [30] 和留一项目 [12]，到他们的机器学习模型。这些项目特定的验证方法可以通过将模型暴露于不同组合的数据的不同方面来确保模型得到彻底验证。然而，我们发现这些方法只有在数据集很大的情况下才有用。对于像我们这样的小数据集，这些方法（包括 k-fold）只会使数据集过于碎片化并降低模型的性能。

*结果的局限性和普遍性：*我们的研究获得的结果基于一个数据集，其中包含有限数量的标记需求文本。此外，数据集是从在线爬取的公开需求文档中收集的[31]。因此，我们无法确定这些结果是否可以在不同的数据集上重复，特别是来自行业标准项目的数据集。为了应对这一威胁，我们正在提供 BERT4RE 模型 [18] 供可再生能源研究人员和从业者尝试，我们欢迎您的反馈。

7.C结论和F优图磷局域网

在本文中，我们提出了 BERT4RE，一种针对 RE 的重新训练的 BERT 模型。BERT4RE 在 RE 相关数据集上进行重新训练，旨在通过模型微调转移到执行不同的 RE 特定任务。作为第一步，本文演示了 BERT4RE 在识别 9 个需求概念的任务上的可迁移性。为了评估其有效性，我们将其性能与 BERT 进行比较根据通过使用具有九个类别的标记数据集对这两个模型进行微调。我们的研究表明 BERT4RE 优于 BERT根据识别 9 个概念中的 7 个，从而证明了重新培训 LM 的好处。

为了推进这项研究，我们打算开展以下四个项目：

第一个项目涉及展示在各种 RE 任务中使用具有迁移学习能力的特定领域 LM 的有效性。我们打算针对其他 RE 任务对 BERT4RE 进行微调，例如使用 PROMISE 数据集进行分类、语言问题的可追溯性和检测等。

第二个项目是探索我们提出的技术对于开发支持识别需求文本中关键概念的自动工具的有效性。通过定义建模框架，BERT4RE 模型将得到扩展，以识别关键概念并创建将需求文本表示为一流工件的模型。该框架将被概念化为一种语义方法，用于识别*语义角色*句法特征并将这些角色（例如代理、动作、目标等）及其关系转化为概念模型。

第三个项目将建立在开发 BERT4RE 模型所使用的流程和方法的基础上。我们希望通过升级其架构来训练其他高级 LM 模型变体（例如 ALBERT [40]），从而使我们的模型更加稳健，其中使用两参数缩减方法来降低内存消耗并提高 BERT 的训练速度。我们相信这种方法将能够训练和扩展特定领域模型，例如我们的 BERT4RE 模型。此外，我们采用其架构的另一个 LM 是 ELECTRA LM [41]。该架构使我们能够训练一个判别模型，与仅适用于输入标记屏蔽的 BERT 模型 [2] 不同。我们希望使用一种样本高效的方法来取代标记检测，而不是屏蔽输入标记。我们设想的这种方法可以在输入标记上定义预训练任务，而不是屏蔽小输入子集。

本研究的第四个项目将寻求升级和验证我们的 BERT4RE 模型的性能。这将通过对特定领域的数据以及具体的法律和社交媒体相关数据作为案例研究进行微调来进行。微调的目的是通过几个下游 NLP 任务提取语义信息。

相信这将有助于解决 RE 等特定领域领域缺乏训练数据的问题。

右参考文献

[1] A. Vaswani、N. Shazeer、N. Parmar、J. Uszkoreit、L. Jones、AN Gomez、L. Kaiser 和 I. Polosukhin，“注意力就是你所需要的”*ArXiv*，卷。绝对/1706.03762，2017。

[2] J.德夫林，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”*arXiv 预印本 arXiv:1810.04805*，2018。

[3] ME Peters、M. Neumann、M. Iyyer、M. Gardner、C. Clark、K. Lee 和 L. Zettlemoyer，“深度语境化单词表示”，*arXiv 预印本 arXiv:1802.05365*，2018。

[4] A. Radford、K. Narasimhan、T. Salimans 和 I. Sutskever，“通过生成预训练提高语言理解”
https://www。CS。不列颠哥伦比亚省。加州/~amuham01/LING530/论文/雷德福2018改进。pdf，2018。

[5] J. Howard 和 S. Ruder，“文本分类的通用语言模型微调”，*arXiv 预印本 arXiv:1801.06146*，2018。

[6] T. Mikolov、I. Sutskever、K. Chen、GS Corrado 和 J. Dean，“单词和短语的分布式表示及其组合性”，*神经信息处理系统的进展*，卷。2013 年 26 日。

[7] K. Ethayarajh，“语境化的单词表征的语境如何？比较 bert、elmo 和 gpt-2 嵌入的几何形状，” *arXiv 预印本 arXiv:1909.00512*，2019。

[8] C. Raffel、N. Shazeer、A. Roberts、K. Lee、S. Narang、M. Matena、Y. Zhou、W. Li 和 PJ Liu，“利用统一的文本到文本转换器探索迁移学习的局限性”*arXiv 预印本 arXiv:1910.10683*，2019。

[9] SJ Pan 和 Q. Yang，“迁移学习调查”，*IEEE 知识与数据工程汇刊*，卷。22、没有。10，第 1345–1359 页，2009 年。

[10] F. Zhuang，Z. Qi，K. Duan，D. Xi，Y. Zhu，H. Zhu，H. Xiong，和问：He，“迁移学习的综合调查”*IEEE 会刊*，卷。109，没有。1，第 43-76 页，2020 年。

[11] L.Zhao、W.Alhoshan、A.Ferrari、KJ Letsholo、MA Ajagbe、E.-V。Chioasca 和 RT Batista-Navarro，“需求工程的自然语言处理：系统映射研究”，*ACM 计算调查 (CSUR)*，卷。54，没有。3，第 1-41 页，2021 年。

[12] T. Hey、J. Keim、A. Koziolek 和 W. Tichy，“Norbert：需求分类的迁移学习”，*2020年IEEE第28届国际需求工程会议（RE）*，第 169–179 页，2020 年。

[13] A. Sainani、PR Anish、VN Joshi 和 S. Ghaisas，“从软件工程合同中提取和分类需求”，*2020年IEEE第28届国际需求工程会议（RE）*，第 147–157 页，2020 年。

[14] J. Lee、W. Yoon、S. Kim、D. Kim、S. Kim、CH So 和 J. Kang，“Biobert：用于生物医学文本挖掘的预训练生物医学语言表示模型”，*生物信息学*，卷。36、没有。4，第 1234–1240 页，2020 年。

[15] I. Beltagy、K. Lo 和 A. Cohan，“Scibert：科学文本的预训练语言模型”，*arXiv 预印本 arXiv:1903.10676*，2019。

[16] I. Chalkidis、M. Fergadiotis、P. Malakasiotis、N. Aletras 和 I. Androutsopoulos，“Legal-bert：法学院毕业的布偶”，*arXiv 预印本 arXiv:2010.02559*，2020。

[17] M. Wang 和 W. Deng，“深度视觉域适应：一项调查”，*神经计算*，卷。312，第 135-153 页，2018 年。

[18] M. Ajagbe 和 L. Zhu，《需求工程中迁移学习再训练 bert 模型的补充材料》：一项初步研究”，2022 年 3 月。[在线]。可用：https://zenodo。组织/记录/6354280

[19] E. Alsentzer、J. Murphy、W. Boag、W. Weng、D. Jin、T. Naumann 和 MBA McDermott，“公开可用的临床 bert 嵌入”，*ArXiv*，卷。绝对/1904.03323，2019。

[20] K. Huang、J. Altosaar 和 R. Ranganath，“Clinicalbert：临床记录建模和预测再入院”，*ArXiv*，卷。绝对/1904.05342，2019。

[21] AEW Johnson，T. Pollard，L. Shen，L. wei H. Lehman，M. Feng，M. Ghassemi、B. Moody、P. Szolovits、L. Celi 和 R. Mark，“Mimic-iii，一个可免费访问的重症监护数据库”*科学数据*，卷。2016 年 3 日。

[22] DQ Nguyen、T. Vu 和 AT Nguyen，“Bertweet：英语推文的预训练语言模型”*arXiv 预印本 arXiv:2005.10200*，2020。

[23] M. Li、L. Shi、Y. Yang 和 Q. Wang，“一种从开放论坛进行需求发现和注释的深度多任务学习方法”，*2020年第35届IEEE/ACM自动化软件工程国际会议（ASE）*，第 336–348 页，2020 年。

[24] Y. Wang、L. Shi、M. Li、Q. Wang 和 Y. Yang，“自然语言要求中的共指检测的深度上下文方法”，*2020年IEEE第28届国际需求工程会议（RE）*，第 180–191 页，2020 年。

[25] AF Araujo 和 RM Marcacini，“Re-bert：使用 bert 语言模型从应用程序评论中自动提取软件需求”，*第 36 届 ACM 应用计算年度研讨会论文集*，2021 年。

[26] M. Pota、M. Ventura、R. Catelli 和 M. Esposito，“基于 bert 的有效 Twitter 情绪分析管道：意大利语案例研究”，*传感器（瑞士巴塞尔）*，卷。2021 年 21 日。

[27] J. Cleland-Huang、S. Mazrouee、H. Liguo 和 D. Port，“NFR”，2007 年 3 月。[在线]。可用：https://doi。组织/10。5281/泽诺多。268542

[28] J. Cleland-Huang、R. Settimi、X. Zou 和 P. Solc，“非功能性需求的自动分类”，*需求工程*，卷。12、没有。2，第 103-120 页，2007 年。

[29] Z. Kurtanović 和 W. Maalej，“使用监督机器学习自动分类功能和非功能需求”，*在2017年IEEE第25届国际需求工程会议（RE）*。IEEE，2017 年，第 490–495 页。

[30] F. Dalpiaz、D. Dell'Anna、FB Aydemir 和 S. Çevikol，“利用可解释的机器学习和依存分析进行需求分类”，载于*2019 IEEE第27届国际需求工程会议（RE）*。IEEE，2019 年，第 142–152 页。

[31] A. Ferrari、GO Spagnolo 和 S. Gnesi，“Pure：公共需求文档数据集”，载于*2017年IEEE第25届国际需求工程会议（RE）*。IEEE，2017 年，第 502–505 页。

[32] R. Sennrich、B. Haddow 和 A. Birch，“带有子词单元的稀有词的神经机器翻译”，*arXiv 预印本 arXiv:1508.07909*，2015。

[33] Y. Liu、M. Ott、N. Goyal、J. Du、M. Joshi、D. Chen、O. Levy、M. Lewis、L. Zettlemoyer 和 V. Stoyanov，“Roberta：一种稳健优化的 bert 预训练方法”*arXiv 预印本 arXiv:1907.11692*，2019。

[34] DP Kingma 和 J. Ba，“Adam：一种随机优化方法”，*钻RR*，卷。绝对/1412.6980，2015。

[35] CJ菲尔莫尔*等人*。，“框架语义和语言的本质”，*纽约科学院年鉴：关于语言和言语的起源和发展的会议*，卷。280，没有。1. 纽约，1976 年，第 20-32 页。

[36] K. Letsholo，L. Zhu，和 E.-V。Chioasca，“Tram：将文本需求转换为分析模型的工具”*2013年第28届IEEE/ACM自动化软件工程国际会议（ASE）*，第 738–741 页，2013 年。

[37] M. Grandini、E. Bagli 和 G. Visani，“多类分类指标：概述”*arXiv 预印本 arXiv:2008.05756*，2020。

[38] BA Kitchenham、SL Pfleeger、LM Pickard、PW Jones、DC Hoaglin、K. El Emam 和 J. Rosenberg，“软件工程实证研究的初步指南”*IEEE 软件工程汇刊*，卷。28、没有。8，第 721–734 页，2002 年。

[39] P. Refaeilzadeh、L. Tang 和 H. Liu，“交叉验证”。*数据库系统百科全书*，卷。5，第 532–538 页，2009 年。

[40] Z. Lan、M. Chen、S. Goodman、K. Gimpel、P. Sharma 和 R. Soricut，“Albert：语言表征自我监督学习的 lite bert”，*arXiv 预印本 arXiv:1909.11942*，2019。

[41] K.克拉克，M.-T。Luong、QV Le 和 CD Manning，“Electra：将文本编码器预训练为判别器而不是生成器”*arXiv 预印本 arXiv:2003.10555*，2020。