语言模型和指令微调在生物医学关系抽取中的领域特定性有多重要？

Aviv Brokman, Ramakanth Kavuluru
University of Kentucky, Lexington, KY USA
avivbrokman@gmail.com, ramakanth.kavuluru@uky.edu

摘要

前沿的自然语言处理技术通常会被后续应用于高价值、数据丰富的生物医学领域。过去几年里，生成式语言模型（LM）、指令微调和少量样本学习成为了 NLP 研究的重点。因此，在生物医学语料库上预训练的生成式 LM 大量出现，并且已经尝试了生物医学领域的指令微调，所有这些都寄希望于特定领域的专业性能够提升下游任务的表现。鉴于训练此类模型所需的非同寻常的努力，我们调查了它们在关键的生物医学 NLP 任务——关系抽取中是否具有任何优势。具体来说，我们解答两个问题：（1）在生物医学语料库上训练的 LM 是否优于在通用领域语料库上训练的？（2）在生物医学数据集上进行指令微调的模型是否优于那些在各种数据集上进行微调或仅仅是预训练的模型？我们利用现有的 LM 来解答这些问题，在四个数据集上进行了测试。令人惊讶的是，通用领域的模型通常表现优于特定于生物医学领域的模型。然而，尽管用于生物医学领域指令微调的指导数量少得多，但其性能提升程度与通用领域指令微调相似。我们的研究结果表明，将研究精力集中在对通用 LM 进行更大规模的生物医学指令微调可能比构建特定领域的生物医学 LM 更有成效。

Keywords: 关系抽取，语言模型，生物医学信息提取，指令微调

1. 介绍

生物医学实体及其之间的关系是生物医学知识发现的前沿。生物医学实体之间的新关系信息通常通过科学文献或临床记录中的文本传达。例如，蛋白质-蛋白质相互作用（以理解疾病的病因和进展）、基因-疾病关联（以识别潜在的药物靶点）、药物-疾病治疗关系（以发现非适应症使用情况或评估再定位潜力）以及药物-基因相互作用（以设计针对性疗法）经常在文献中被讨论。提取的关系通常用于构建知识图谱，以促进知识发现和基于知识的搜索系统 (Bakal et al., 2018; Lever et al., 2018; Zhang et al., 2021)。不良事件关系（链接药物和副作用）常在临床文本中报告，并且其提取有助于更有效的上市后药物警戒 (Liu et al., 2019; Botsis et al., 2012)。因此，生物医学关系抽取（RE）是 BioNLP 社区经常追求的任务之一，在许多共享任务系列（例如，BioCreative、i2b2、n2c2）中都有体现。我们当前的贡献在于关于生物医学领域特定语言模型（LMs）和指令微调在生物医学 RE 中的价值的高层次问题。在本节其余部分，我们将概述最近与 RE 相关的方法格局，并在此背景下阐明我们的贡献。

近几年，基于变压器的生成模型（编码器-解码器和仅解码器）语言模型激增。部分原因是诸如总结、释义和代码生成等任务本质上是生成性的。但在某些以前使用后者解决的任务中，生成模型甚至比仅有编码器的语言模型更有优势。编码器模型主要是为分类问题设计的，在 [CLS] 标记的最后一层嵌入上训练分类器。它们还以各种其他方式被使用——特别是在命名实体识别（NER）和 RE 任务中，其中跨度的嵌入被传递给分类器 (Eberts and Ulges, 2020, 2021; Zhong and Chen, 2021)。这些方法从头开始微调权重，因此需要大量的数据。但少样本和零样本学习已成为 NLP 研究的重点领域；它们的发展有望提高 NLP 的可扩展性，因为手动创建大型数据集非常耗时。在编码器模型中实现少/零样本学习通常是通过将 NLP 任务的目标函数与预训练目标函数对齐来完成的。使用包含待预测 [MASK] 标记的模板进行提示是仅解码器模型 (Schick and Schütze, 2021; Wang et al., 2022) 少/零样本方法。采用这种策略很大程度上取决于所需的 NLP 任务是否为分类任务或可以被表述成这类任务。总之，仅有编码器的语言模型非常适合处理（与少样本设置相对）各种非生成性任务的全微调场景。相比之下，在 RE 中相对于编码器模型所需额外分层和定制可以在生成模型中省略掉，因为它们使用较为灵活的方法通过简单自然语言模板来表述 RE。这在少样本和零样本学习中特别有用。

在过去两年中，通用领域的指令微调（IFT）

已成为生成式语言模型 (Ouyang et al., 2022; Wei et al., 2022; Sanh et al., 2022; Ouyang et al., 2022; Wei et al., 2021; Touvron et al., 2023) 的自然语言处理研究的热点领域。在基本形式下，通过将各种带标签数据集中的示例转换成自然语言指令对（连同输入文本）和响应（包含正确的输出），并在此基础上训练一个生成式的语言模型来进行 IFT。这通常通过使模型朝向人类用户的目标进行对齐来提高少量样本的表现性能。然后，对于新任务，相应的指令一般会比没有经过 IFT 时更符合分布。

大多数自然语言处理的进展最初都是为通用领域开发的，后来才适应生物医学领域。最近，这一点在流行通用领域语言模型的生物医学版本的发展中得到了体现。直观上，这样做是有道理的，因为生物医学 NLP 任务中的文本分布与生物医学文本更相似，而且生物医学特定的分词器可以更简洁地表示生物医学实体。同样，在 IFT 取得显著成功的基础上，他们在一个自己整理的生物医学元数据集（称为 BoX）上对 Parmar et al. (2022) 指令微调了 BART-Base(Lewis et al., 2019) 模型以创建 In-BoXBART。

生物医学生成语言模型正在迅速训练，我们可以安全地假设更大的生物医学模型将在近期被训练。鉴于创建特定于生物医学的语言模型在劳动力、货币和努力成本上的高昂代价，询问它们是否值得是明智的。在某种程度上，这已经在 T5(Raffel et al., 2020; Phan et al., 2021) 和 BART(Lewis et al., 2019; Yuan et al., 2022) 的生物医学版本中通过全面微调进行了测试，结果分数提升并不显著。然而，所测试的任务大多需要极短且简单的生成。可以推测，在更复杂的任务中，特别是在少样本场景下，这些模型可能会与它们的一般领域对应物有所区别。我们调查了在高价值的 RE[1] 任务中是否确实如此。具体来说我们问：

- 预训练于基于 PubMed 的科学文本上的生成模型是否比那些在通用领域文本上训练的模型表现更好？
- 生物医学和通用领域指令微调模型的表现与它们的基础模型相比如何？
- 这些问题的答案在完全微调和少量样本设置中是否不同？

我们在四个生物医学关系抽取数据集上使用多种开放的生物医学和通用领域的语言模型测试了这些问题。（我们提供了代码供审核[2]，并计划在被接受后公开。）

2. 相关工作

2.1. 生成关系抽取

关系提取（RE）传统上通过两步流水线进行——首先是命名实体识别（NER），然后是在来自 NER 的实体对之间进行关系分类（RC）。然而，从 Miwa and Sasaki (2014) 开始，这种流水线方法被端到端的方法所补充，这些方法为 NER 和 RC 阶段采用了联合损失函数。在这些方法中，Zeng et al. (2018) 使用复制机制来处理无需中间 NER 步骤的生成式 RE；随后有几个扩展也采取了同样的做法 (Nayak and Ng, 2020; Hou et al., 2021; Nayak and Ng, 2020; Zeng et al., 2020; Giorgi et al., 2022)。他们的策略的一个优点是，生成可以自然地处理不连续的实体（例如，在短语*手和手臂疼痛*中的*手痛*）。不连续实体给命名实体识别带来了如此大的挑战，以至于有一整条研究线专门致力于解决这个问题 (Wang et al., 2021; Dirkson et al., 2021)。但这些基于复制的模型需要从头开始训练神经网络组件，因此不适合少量样本学习。

Huguet Cabot and Navigli (2021) 使用 BART 直接生成关系三元组，使用特殊标记来划分三元组中的角色；这仍然需要训练随机初始化的特殊标记。Luo et al. (2022) 在自然语言模板上微调 BioGPT，使当前 NLP 任务的输出形式与预训练策略最为一致。在测试时，从生成的文本中使用正则表达式提取关系。我们在此研究中采用 Luo et al. (2022) 这种一般策略。Wadhwa et al. (2023) 使用将关系转换为结构化数据的目标序列而非自然语言的模板来微调 Flan-T5(Wei et al., 2021)。我们注意到，在初步实验中，使用此类形式的模板性能较低。

2.2. 生物医学语言模型

早期基于变换器的生物医学语言模型采用仅编码器架构，通常基于 BERT(Devlin et al., 2019)，

[1] 这里我们明确处理端到端的关系抽取，其中实体必须由关系抽取系统提取，而不是提供先验的。

[2] 代码可在 https://anonymous.4open.science/r/domain_specificity-BCB8 获取

并且在 PubMed 摘要、PubMed Central 全文文章 (PMC) 和 MIMIC III(Johnson et al., 2016) 临床记录的一些组合上进行训练。BioBERT(Lee et al., 2019) 使用来自 BERT 的权重进行初始化，然后继续在 PubMed 和 PMC 上预训练。Alsentzer et al. (2019) 继续在 MIMIC-III 数据集上对 BioBERT 进行预训练，从而生成了 ClinicalBERT。同样，Gururangan et al. (2020) 持续预训练了 RoBERTa(Liu et al., 2020) 来自 S2ORC 的额外全文生物医学文章 (Lo et al., 2020)，获得了 BioMed-RoBERTa。从头开始使用生物医学文本训练的模型通常比其持续训练的同类表现更好。基于 BERT 的 BlueBERT(Peng et al., 2019) 和 PubMedBERT(Gu et al., 2020)，以及基于 RoBERTa 的 Bio-LM(Lewis et al., 2020) 是此类模型中的几个例子，其中 PubMedBERT 作为许多下游应用的流行基础模型。Shin et al. (2020) 训练版本的 BioMegatron，分别从零开始训练和通过继续预训练 Megatron (Shoeybi et al., 2020)。BioELECTRA(Kanakarajan et al., 2021)，一个基于高性能 ELECTRA(Clark et al., 2020) 的生物医学版本，在 PubMed 和 PMC 上从零开始训练。

随着生成模型（具有编码器-解码器和仅解码器架构）的日益普及，基于 PubMed 的生成模型也从头开始训练。T5(Raffel et al., 2020) 的生物医学版本—SciFive、(Phan et al., 2021) 和 GPT—BioGPT(Luo et al., 2022) 和 BioMedLM(Bolton et al., 2022) 均已在 PubMed/PMC 语料库上从头开始训练。BioBART(Yuan et al., 2022) 是一个使用 PubMed 文本持续预训练的 BART(Lewis et al., 2019) 模型。SciFive 和 BioBART 在大多数生物医学任务上显示出微乎其微的改进，在少数任务中显示出适度的改进，优于其通用领域对应模型。然而，我们认为 SciFive 和 BioBART 的生成能力在 Phan et al. (2021) 和 Yuan et al. (2022) 中并未得到充分探索，因为所解决的任务要么需要适度的生成（例如，多项选择题答案），或者通过 ROUGE 分数进行评估，这与人类评价相比仅能提供粗略的质量估计。Luo et al. (2022) 在三个端到端关系提取数据集上测试了 BioGPT，其表现显著优于同样规模的 GPT-2 Medium(Radford et al., 2019)。我们无法找到 BioGPT-Large 在端到端关系提取上的任何性能得分。BioMedLM(Bolton et al., 2022) 在三个问答任务中得分远高于同等规模的 GPT-Neo 2.7B 模型。

总结来说，有证据表明在非常短的生成任务中，在完全微调设置下，生物医学编码器-解码器语言模型优于通用领域模型。但是对于更复杂的生成任务中解码器-only 模型性能的有效性信息很少，对于编码器-解码器模型更是如此。少样本性能模式尤其未知。

2.3. 指令微调

Wei et al. (2022) 通过指令微调在 60 个任务上探索 IFT 对零样本性能的影响，每个任务构建 10 个模板，发现性能一致且显著的提升。Sanh et al. (2022) 在不同的数据集集合上进行类似的实验，生成了 T0 模型。Chung et al. (2022) 基于之前的尝试，将微调规模扩大到 1600 个任务，并随后发布了 Flan-T5 系列指令微调模型。InstructGPT(Ouyang et al., 2022) 使用带有人类反馈的强化学习（RLHF）(Christiano et al., 2017; Stiennon et al., 2020) 对指令进行训练以使 GPT-3(Brown et al., 2020) 与人类偏好对齐。Wang et al. (2023) 使用 GPT-3 创建用于 IFT 的合成指令数据，达到了与成本更高的 InstructGPT 相当的性能。In-BoXBART(Parmar et al., 2022)（据我们所知唯一的生物医学指令微调尝试）是基于 BART 基础模型在 32 个生物医学任务上训练得到的 IFT 版本，每个任务一个模板，因此其 IFT 规模远小于其他 IFT 模型。

3. 任务设置

3.1. 训练

RE 实例通常由源文本和一组在其中声明的半结构化关系组成。为了在一个关系抽取数据集上微调生成性语言模型，我们将其训练示例转换成自然语言序列，然后再以常规方式微调模型。本质上，这些自然语言序列包含源文本以及一条用于抽取关系的指令，随后是用自然语言形式表达的关系。Figure 1 说明了我们在下文形式化表述的从科学文献中抽取组合药物疗法（DCE）任务中的方法 (Tiktinsky et al., 2022)。

令 (x, y) 为一个训练样本，其中 x 是输入文本，y 表示其存在的关系。为了微调语言模型，我们为每个数据集构造一个单一的提示模板 $T = (T_s, T_t)$。一个模板由源转换函数 T_s 和目标转换函数 T_t 组成，它们返回包含源文本 x 以及指导生成的指令的源序列 x^T，以及传达关系 y 以自然语言形式的目

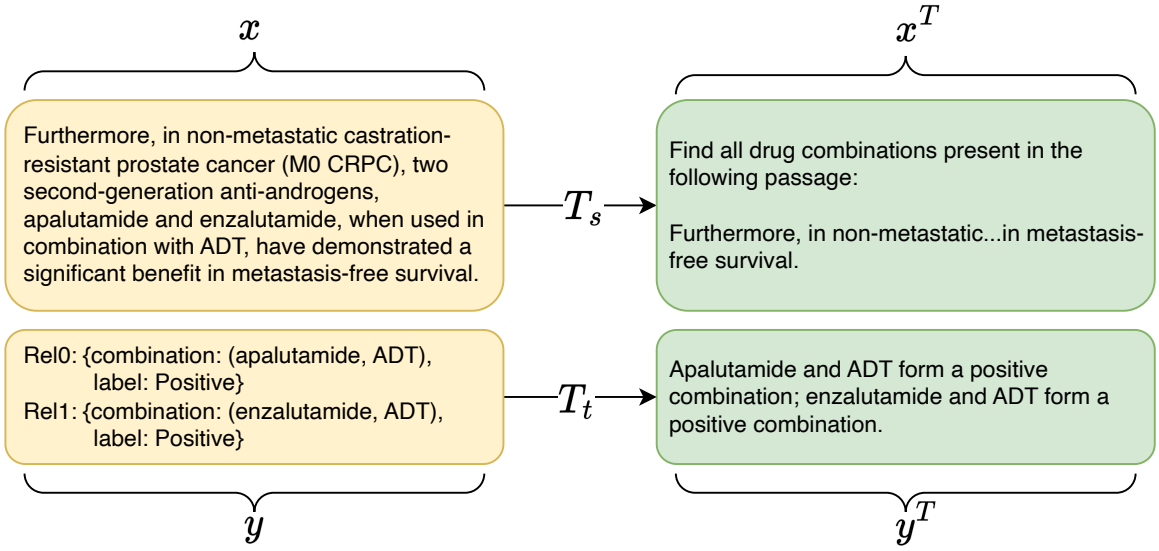

图 1: 数据集的训练示例转换为自然语言序列。

标序列 y^T:

$$T_s(x) = x^T$$
$$T_t(y) = y^T$$

对于编码器-解码器模型，x^T 和 y^T 分别提供给编码器和解码器模块。对于解码器模型，序列 x^T 和 y^T 通过函数 J 被合并为一个序列，然后作为单一输入供模型使用。我们取 J 为

$$J\left(x^T, y^T\right) = x^T \| \backslash n \backslash n \| y^T$$

其中 $\|$ 表示连接。我们的 J 形式用两个换行符分隔源序列和目标序列，尽管也可以使用其他形式。

3.2. 评估

对于验证和测试示例，我们只有源序列，我们的目标是生成目标序列。对于编码器-解码器模型，x^T 再次作为输入传递给编码器。对于解码器模型，输入序列为

$$J\left(x^T, \cdot\right) = x^T \| \backslash n \backslash n.$$

模型 $\hat{y}^T$ 生成的输出序列必须转换为结构化格式，以便可以将预测的关系与标注的关系进行比较以计算性能指标。我们使用正则表达式为每个模板设计一个信息提取器。

4. 实验设置

我们进行了两项实验以研究领域特定性的影响。首先，我们将生物医学领域的模型与通用领域的模型进行比较。其次，我们将经过生物医学指令微调的、通用领域指令微调的和基础模型性能进行比较。所有实验均在完全微调和少量样本设置下重复进行。我们将包括每个数据集的模板以及建模选择（学习率、批量大小、...）在内的附录在最终定稿版本中（如果被接受），鉴于会议提交指南限制了这些内容在初步审阅期间的包含。

4.1. 实验

4.1.1. 生物医学与通用领域预训练

在我们的第一个实验中，我们比较了生物医学语言模型及其最接近的通用领域语言等效模型在架构和规模方面的效果。我们评估了生物医学解码器语言模型 BioGPT（350M）(Luo et al., 2022)、BioGPT-Large（1.5B）(Luo et al., 2022) 和 BioMedLM（2.7B）(Bolton et al., 2022) 的表现，并将其与它们的近似通用领域等效模型进行了比较，分别是 GPT2-Medium（350M）、GPT2-XL（1.5B）(Radford et al., 2019) 和 GPT-Neo（2.7B）(Black et al., 2021)。我们还评估了生物医学编码器-解码器语言模型 SciFive-Base（220M）、SciFive-Large（770M）（均为 PubMed 版本）(Phan et al., 2021)、BioBART-Base（140M）和 BioBART-Large（400M）(Yuan et al., 2022)，并与通用领域的等效模型 T5-Base（220M）、T5-Large（770M）(Raffel et al., 2020)、BART-Base（140M）和 BART-Large（400M）(Lewis et al., 2019) 进行了比较。

4.1.2. 指令微调与基础模型

在我们的第二个实验中，我们比较了指令微调模型及其基础模型的有效性。在文献回顾中没有生物医学模型被进行指令微调。通用领域的模型使用通用领域的数据集进行了指令微调，并单独使用生物医学数据集进行了微调。对于通用领域指令，我们将 Flan-T5(Wei et al., 2021) 系列的模型与它们训练的基础 T5 模型 (Raffel et al., 2020) 进行了比较。对于生物医学领域的指令，我们比较了 In-BoXBART(Parmar et al., 2022) 和基础 BART(Lewis et al., 2019)，它是从中训练而来的。

4.1.3. 少量样本设置

我们在第 4.1.1 和 4.1.2 节重复实验，但在少量样本设置下进行。我们强调的是，我们微调了模型，而不是执行上下文学习。我们试验了 16-shot 和 64-shot 的学习。少量样本性能值是通过对五次运行的不同随机种子选择训练样例的平均值得到的。除了一个例外，完全微调的仅解码器模型性能远不如编码器-解码器模型（F1<30）；因此我们没有对仅解码器模型进行少量样本实验。

4.2. 数据集

我们使用四个公开的生物医学关系抽取数据集进行实验。对于没有预设训练/验证分割的数据集，我们选择训练数据集中的 20%作为验证集。

4.2.1. CDR

CDR(Li et al., 2016) 是一个 BioCreative V 数据集，包含 PubMed 摘要，并对所有化学物质和疾病的提及进行了标注，以及它们之间描述由化学物质引起疾病的所有关系。所有的关系都属于单一类别并且在实体级别上进行标注。1500 篇摘要被等分为训练集、验证集和测试集。

4.2.2. 化学保护蛋白

ChemProt(Krallinger et al., 2017) 是一个 BioCreative VI 数据集，包含 PubMed 摘要，在这些摘要中所有提到的化学化合物/药物和基因/蛋白质都被注释了，并且这两类实体之间的所有句内关系都根据一组感兴趣的关联类型进行了标注。关系属于*激动剂*、*拮抗剂*、*上调*、*下调*和*衬底*这几类，并且在提及层面进行了标注。由于关联是在提及层面上进行标注的，一个示例可能包含多个概念上相同但提及不同的已标注关联。数据集被划分为 1,020 个训练样本、612 个验证样本和 800 个测试样本。

4.2.3. 动态电路等效

药物组合提取（DCE）(Tiktinsky et al., 2022) 数据集由从 PubMed 摘要中选取的包含多种药物的句子组成，这些句子标注了给定的联合用药。一组药物被标注为具有正面效果的药物组合、文本中未明确说明效果的药物组合或不是药物组合。尽管一些研究将后两种关系类型合并，(Tiktinsky et al., 2022; Jiang and Kavuluru, 2023) 我们仍按原样建模它们。该数据集分为 1362 个训练实例和 272 个测试实例。

4.2.4. 药物相互作用

DDI 数据集 (Herrero-Zazo et al., 2013) 包含来自 DrugBank 和 Medline 的文本，并对药物/化学品及其共现药物之间的关系进行了标注。药物的提及被分类为通用药品名称、品牌药品名、药品组或非药品活性物质。对于药品对的关系被标注为*药代动力学机制*、*组合效应*、*关于交互的建议或缺乏支持信息时的交互*。关系在提及级别上进行标注。该数据集由 784 份 DrugBank 文档和 233 篇 Medline 摘要组成，分为 714 个训练样本和 303 个测试样本。

4.3. 评估

我们计算所有任务的精度、召回率和 F1 值。评估是在严格设置下进行的：只有当实体全部正确且关系类型（如果适用）也正确时，才会认为关系是正确的。一个正确的实体由精确字符串匹配组成。对于 CDR，预测的实体必须与黄金实体对应的 MeSH ID 中的某个提及相匹配。

5. 结果

所有结果均显示在表 1（第 8 页）中，主要按模型系列（如 T5）分组，并按模型大小进行内部排序。在进入主要结果之前，我们注意到数据集和模型系列中有两个高层次的趋势：(1) 总体而言，CDR 和 DCE 的表现优于 ChemProt 和 DDI，这可能是由于后者的关联类型数量较多所致。（此后，我们将 CDR 和 DCE 称为高表现，将 ChemProt 和 DDI

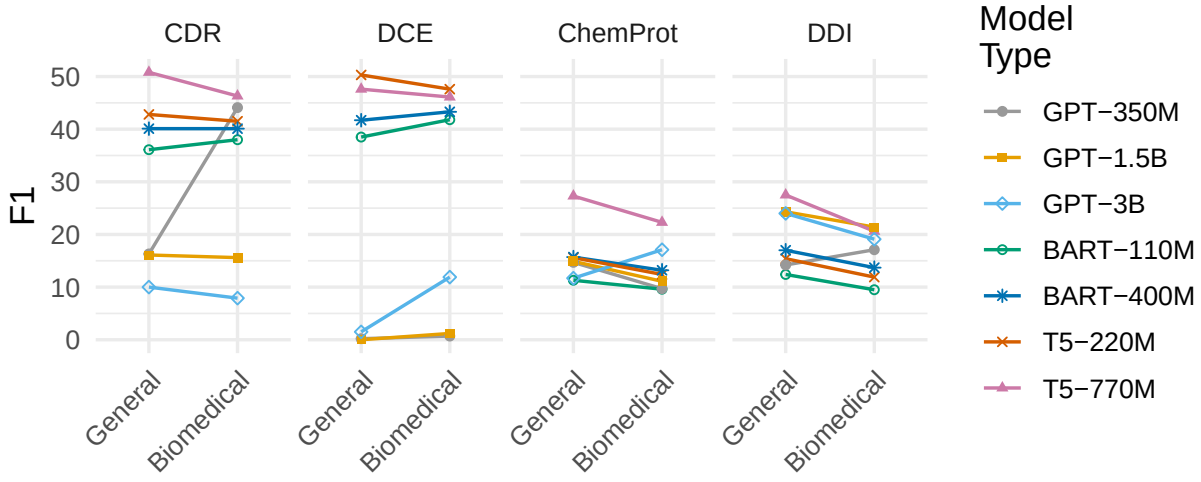

图 2: 通用领域和生物医学文本预训练的完全微调语言模型的比较。

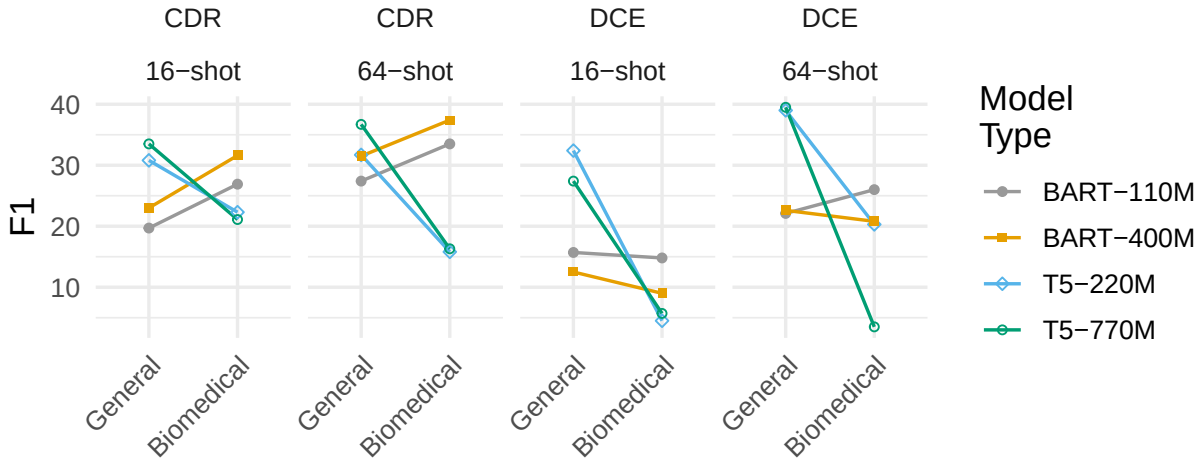

图 3: 在通用领域和生物医学文本上预训练的少量样本微调语言模型的比较。

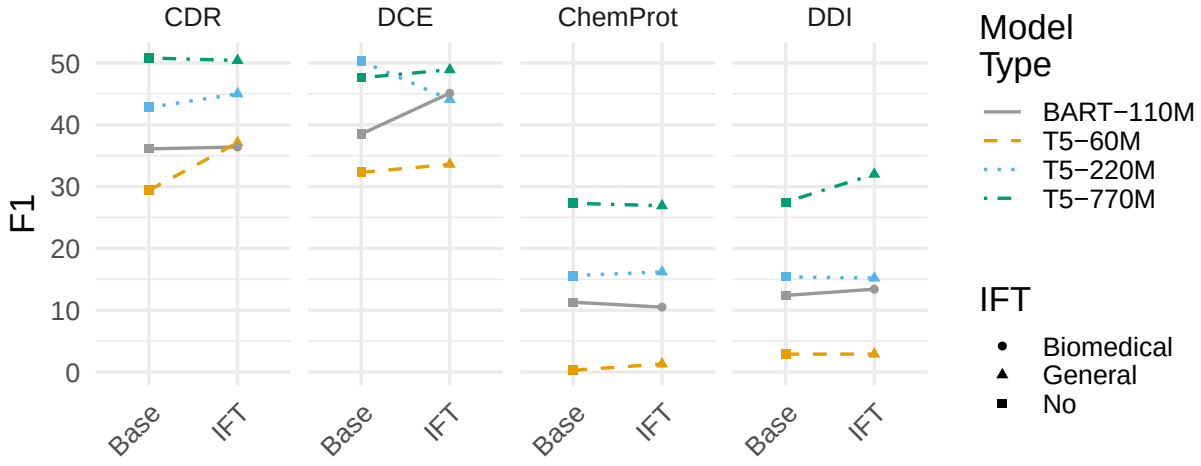

图 4: 不同数据集上完全微调基础模型和指令微调语言模型的比较。

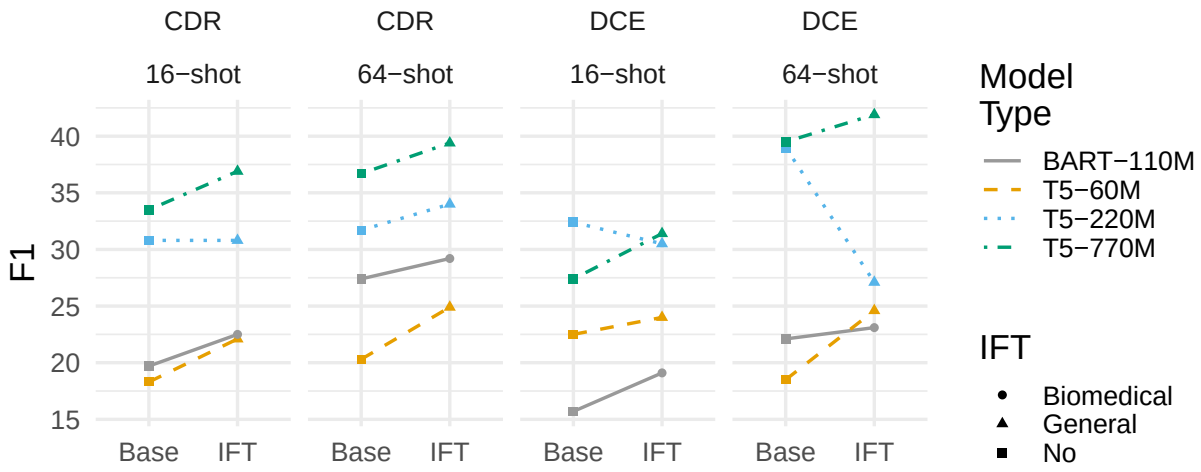

图 5: 少量样本对比基础模型和指令微调语言模型在高性能数据集上的表现。

称为低表现，以解释其他结果。）(2) 编码器-解码器模型的表现远好于仅有解码器的模型，BioGPT 在 CDR 上的例外除外。事实上，仅有解码器的模型表现如此之差，以至于我们在进一步讨论中主要限于编码器-解码器模型。这可能是因为解码器模型更适合生成性输出（如摘要），而不是关系抽取。

5.1. 生物医学与通用领域预训练

总体而言，在完整和少量样本微调实验中，通用领域预训练优于生物医学领域预训练（Figure 2 和 Figure 3）。在完全微调中，这种模式适用于所有数据集中的 T5 模型以及低性能数据集中的 BART 模型。值得注意的例外是高性能数据集中的 BART 模型，在这些模型中，生物医学模型的表现优于或与相应的通用领域模型相当，以及 BioGPT，其表现显著优于其 GPT2-Medium 对应版本。然而，BioGPT 在所有其他情况下表现不佳。在少量样本微调中，生物医学 T5 模型的表现远低于通用领域模型（Figure 3）。生物医学预训练对 BART 模型的影响不一致，提高了 CDR 的性能，但通常降低了 DCE 的性能。

5.2. 指令微调的影响

在全微调中，IFT 版本的模型通常表现相似或略优于其基础版本（Figure 4）。IFT 的小幅优势主要体现在高性能数据集上。在少样本微调中，通用领域指令和生物医学指令一致提升了性能并且提升程度相似（Figure 5），尽管生物医学指令数据集的数量少了几个数量级。

6. 讨论

我们惊讶地发现解码器-only 模型表现如此糟糕，远远不如编码器-解码器模型。Luo et al. (2022) 发现 BioGPT 在 CDR、DDI 和 KD-DTI 上达到了最先进的（SOTA）性能 (Hou et al., 2022)。虽然我们在数据集的某些预处理步骤中与 Luo et al. (2022) 存在差异（例如，我们没有移除包含无关系的例子，训练了多少个周期以及如何确定预测实体是否匹配黄金标准实体），但在 DDI 上的表现远低于他们的结果，并且在两个额外的数据集 DCE 和 ChemProt 上，我们的结果也远远落后于 SOTA。一个不寻常的发现是，只有在 CDR 中，BioGPT 的表现优于 BioGPT-Large 和 BioMedLM（27 亿参数模型）。初步使用 Llama2-7b(Touvron et al., 2023) 的实验同样没有成果，因此我们不会报告任何初始实验令人失望的结果得分。由于编码器-解码器模型表现明显更好，未来实用的生成式关系抽取工作可能在采用编码器-解码器架构时更具成效；如果需要对生物医学语料库（PubMed）进行预训练，则编码器-解码器模型可能是更保险的选择。

然而，在 PubMed 上训练额外的生成式 LM（甚至是编码器-解码器类型）通常会产生较差的模型，至少在 RE 方面是如此。这一结果在完全微调和少样本微调情境中都成立。我们发现这种模式令人难以理解，因为我们实验中的大多数数据集由来自 PubMed 的标注文本组成——人们可能会期望直接在 PubMed 上训练的模型特别适合这些任务，因

		CDR			动态电路等效			ChemProt	DDI
Model	# Par.	16	64	Full	16	64	Full	Full	Full
GPT-2-Medium	350M	-	-	16.3	-	-	0.2	14.7	14.2
BioGPT[b]	350M	-	-	44.1	-	-	0.7	9.7	17.1
GPT-2-XL	1.5B	-	-	16.1	-	-	0	14.9	24.3
BioGPT-Large[b]	1.5B	-	-	15.6	-	-	1.2	11.1	21.4
GPT-Neo-2.7B	2.7B	-	-	10.0	-	-	1.5	11.7	24.0
BioMedLM[b]	2.7B	-	-	7.9	-	-	11.9	17.1	19.1
BART-Base	110M	17.6	27.4	36.1	8.9	30.6	38.5	11.3	12.4
BioBART-Base[b]	110M	29.0	33.1	38.0	8.9	24.1	41.8	9.6	9.5
In-BoXBART†	110M	21.4	28.9	36.4	10.4	26.7	45.1	10.5	13.4
BART-Large	400M	26.4	32.5	40.1	9.6	30.4	41.7	15.7	17.0
BioBART-Large[b]	400M	30.7	33.0	40.1	8.7	29.2	43.3	13.2	13.7
T5-Small	60M	18.3	20.3	29.4	22.5	18.5	32.3	0.3	2.9
Flan-T5-Small*	60M	18.9	22.2	37.1	10.7	21.2	33.6	1.3	2.9
T5-Base	220M	30.3	29.8	42.8	30.0	39.3	50.3	15.6	15.4
SciFive-Base[b]	220M	28.8	10.1	41.5	0	21.6	47.6	12.4	11.9
Flan-T5-Base*	220M	30.5	33.1	45.0	30.0	33.8	44.1	16.2	15.2
T5-Large	770M	33.9	37.7	50.8	25.7	42.4	47.6	27.3	27.5
SciFive-Large[b]	770M	22.7	5.9	46.3	3.1	3.7	46.1	22.3	20.6
Flan-T5-Large*	770M	35.9	38.4	50.4	27.4	45.1	48.9	26.9	32.0

表 1: 模型 F1 分数比较 跨越四种生物医学关系抽取数据集的不同模型类型和规模。少量样本值是通过选择训练示例的五个不同随机种子计算得到的平均值。标记有 “b” 的模型是经过生物医学预训练的。标记有 * 的模型是使用通用领域指令进行指令微调的，而那些标记为 † 的则是使用生物医学指令进行指令微调的。

为它们完全位于预训练语料库分布之内。我们假设（1）用于一般领域文本训练的令牌数量之多超过了领域特定性的优势，和/或（2）生物医学领域的模型可能学习到较差的语言表示，因为生物医学模型是在英语语言分布的一个相对狭窄子集中进行训练的，而 RE 需要语言上的复杂性。因此，即使是对于编码器-解码器模型，从头开始训练生物医学 LM 也不一定是明智的选择。

关于 IFT 实验，我们的结果并不表明领域特定训练完全没有意义。在完全微调的情况下，IFT 模型的表现大致与它们的基础模型相同，尽管在训练过程中收敛所需的时代要少得多。在少量样本设置中，我们发现小型生物医学 IFT 产生的性能增益与大规模通用领域 IFT 相当。In-BoXBART 在一个任务一个模板的基础上对 32 个生物医学任务进行了微调。相比之下，Flan-T5 模型在一个任务多个模板的基础上对 1,836 个任务进行了微调，并且 Wei et al. (2021) 发现性能持续上升直到~300 个任务。这表明使用额外的生物医学数据集进行 IFT 可能会产生显著的好处。虽然标注的生物医学数据集远远少于通用领域数据集，但包含超过 100 个当前数据集的生物医学元数据集 BigBIO(Fries et al., 2022)

将是生物医学 IFT 的一个宝贵资源。或者，也可以生成合成的生物医学指令示例依照 Wang et al. (2023)。在解释 In-BoxBART 在 CDR 数据集上的少量样本性能时需要一些细微差别。In-BoxBART 被指示微调的任务之一是 CDR 的 NER 任务。有可能 In-BoxBART 在 CDR 上的 RE 能力有一部分来自这种泄漏。除了我们之前的进一步研究建议，我们的结论仅限于 RE，这是 LM 的一个特别有价值的用途。还应对其他生物医学任务进行评估。

总之，公共语言模型及其在许多领域的快速演变自然导致了领域特定模型和 IFT 的产生。在这篇论文中，我们迈出了评估这些模型在生物医学关系抽取任务中对领域特异性需求严格评估的第一步。看看我们的发现是否也适用于其他任务（例如，生物医学摘要）和其他领域（例如，法律）将会很有趣。

7. 参考文献

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Gokhan Bakal, Preetham Talari, Elijah V Kakani, and Ramakanth Kavuluru. 2018. Exploiting semantic patterns over biomedical knowledge graphs for predicting treatment and causative relations. *Journal of biomedical informatics*, 82:189–199.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large scale autoregressive language modeling with mesh-tensorflow.

Elliot Bolton, David Hall, Michihiro Yasunaga, Tony Lee, Chris Manning, and Percy Liang. 2022. BioMedLM. https://crfm.stanford.edu/2022/12/15/biomedlm.html. Accessed: 2023-08-18.

Taxiarchis Botsis, Thomas Buttolph, Michael D Nguyen, Scott Winiecki, Emily Jane Woo, and Robert Ball. 2012. Vaccine adverse event text mining system for extracting features from vaccine safety reports. *Journal of the American Medical Informatics Association*, 19(6):1011–1018.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human*

Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Anne Dirkson, Suzan Verberne, and Wessel Kraaij. 2021. FuzzyBIO: A proposal for fuzzy representation of discontinuous entities. In *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis*, pages 77–82, online. Association for Computational Linguistics.

Markus Eberts and Adrian Ulges. 2020. Span-based joint entity and relation extraction with transformer pre-training. *Proceedings of the 24th European Conference on Artificial Intelligence (ECAI 2020)*, page 2006 – 2013.

Markus Eberts and Adrian Ulges. 2021. An end-to-end model for entity-level relation extraction using multi-instance learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3650–3660, Online. Association for Computational Linguistics.

Jason Alan Fries, Leon Weber, Natasha Seelam, Gabriel Altay, Debajyoti Datta, Samuele Garda, Myungsun Kang, Ruisi Su, Wojciech Kusa, Samuel Cahyawijaya, et al. 2022. Bigbio: A framework for data-centric biomedical natural language processing. *arXiv preprint arXiv:2206.15076*.

John Giorgi, Gary Bader, and Bo Wang. 2022. A sequence-to-sequence approach for document-level relation extraction. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 10–25, Dublin, Ireland. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

María Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, and Thierry Declerck. 2013. The ddi corpus: An annotated corpus with pharmacological substances and drug–drug interactions. *Journal of biomedical informatics*, 46(5):914–920.

Yutai Hou, Yingce Xia, Lijun Wu, Shufang Xie, Yang Fan, Jinhua Zhu, Wanxiang Che, Tao Qin, and Tie-Yan Liu. 2021. Discovering drug-target interaction knowledge from biomedical literature.

Yutai Hou, Yingce Xia, Lijun Wu, Shufang Xie, Yang Fan, Jinhua Zhu, Tao Qin, and Tie-Yan Liu. 2022. Discovering drug – target interaction knowledge from biomedical literature. *Bioinformatics*, 38(22):5100–5107.

Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. REBEL: Relation extraction by end-to-end language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2370–2381, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yuhang Jiang and Ramakanth Kavuluru. 2023. End-to-end n-ary relation extraction for combination drug therapies.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Kamal raj Kanakarajan, Bhuvana Kundumani, and Malaikannan Sankarasubbu. 2021. BioELECTRA:pretrained biomedical text encoder using discriminators. In *Proceedings of the 20th*

Workshop on Biomedical Language Processing, pages 143–154, Online. Association for Computational Linguistics.

Martin Krallinger, Obdulia Rabal, Saber A Akhondi, Martın Pérez Pérez, Jesús Santamaría, Gael Pérez Rodríguez, Georgios Tsatsaronis, Ander Intxaurrondo, José Antonio López, Umesh Nandal, et al. 2017. Overview of the biocreative vi chemical-protein interaction track. In *Proceedings of the sixth BioCreative challenge evaluation workshop*, volume 1, pages 141–146.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Jake Lever, Sitanshu Gakkhar, Michael Gottlieb, Tahereh Rashnavadi, Santina Lin, Celia Siu, Maia Smith, Martin R Jones, Martin Krzywinski, and Steven JM Jones. 2018. A collaborative filtering-based approach to biomedical knowledge discovery. *Bioinformatics*, 34(4):652–659.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Patrick Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. 2020. Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 146–157, Online. Association for Computational Linguistics.

Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016. Biocreative V CDR task corpus: a resource for chemical disease relation extraction. *Database J. Biol. Databases Curation*, 2016.

Feifan Liu, Abhyuday Jagannatha, and Hong Yu. 2019. Towards drug safety surveillance and pharmacovigilance: current progress in detecting medication and adverse drug events from electronic health records. *Drug safety*, 42(1):95–97.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Ro{bert}a: A robustly optimized {bert} pretraining approach.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6). Bbac409.

Makoto Miwa and Yutaka Sasaki. 2014. Modeling joint entity and relation extraction with table representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1858–1869.

Tapas Nayak and Hwee Tou Ng. 2020. Effective modeling of encoder-decoder architecture for joint entity and relation extraction. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8528–8535. AAAI Press.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex

Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Mihir Parmar, Swaroop Mishra, Mirali Purohit, Man Luo, Murad Mohammad, and Chitta Baral. 2022. In-BoXBART: Get instructions into biomedical multi-task learning. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 112–128, Seattle, United States. Association for Computational Linguistics.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. In *Proceedings of the 2019 Workshop on Biomedical Natural Language Processing (BioNLP 2019)*, pages 58–65.

Long N. Phan, James T. Anibal, Hieu Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltekian, and Grégoire Altan-Bonnet. 2021. Scifive: a text-to-text transformer model for biomedical literature.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Tali Bers, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask prompted training enables zero-shot task generalization.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Hoo-Chang Shin, Yang Zhang, Evelina Bakhturina, Raul Puri, Mostofa Patwary, Mohammad Shoeybi, and Raghav Mani. 2020. BioMegatron: Larger biomedical domain language model. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4700–4706, Online. Association for Computational Linguistics.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2020. Megatron-lm: Training multi-billion parameter language models using model parallelism.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021. Curran Associates, Inc.

Aryeh Tiktinsky, Vijay Viswanathan, Danna Niezni, Dana Meron Azagury, Yosi Shamay, Hillel Taub-Tabib, Tom Hope, and Yoav Goldberg. 2022. A dataset for n-ary relation extraction of drug combinations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3190–3203, Seattle, United

States. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Somin Wadhwa, Silvio Amir, and Byron C. Wallace. 2023. Revisiting relation extraction in the era of large language models.

Han Wang, Canwen Xu, and Julian McAuley. 2022. Automatic multi-label prompting: Simple and interpretable few-shot classification. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5483–5492, Seattle, United States. Association for Computational Linguistics.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Yucheng Wang, Bowen Yu, Hongsong Zhu, Tingwen Liu, Nan Yu, and Limin Sun. 2021. Discontinuous named entity recognition as maximal clique discovery. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 764–774, Online. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. 2022. BioBART: Pretraining and evaluation of a biomedical generative language model. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 97–109, Dublin, Ireland. Association for Computational Linguistics.

Daojian Zeng, Ranran Haoran Zhang, and Qianying Liu. 2020. Copymtl: Copy mechanism for joint extraction of entities and relations with multi-task learning.

Xiangrong Zeng, Daojian Zeng, Shizhu He, Kang Liu, and Jun Zhao. 2018. Extracting relational facts by an end-to-end neural model with copy mechanism. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 506–514, Melbourne, Australia. Association for Computational Linguistics.

Rui Zhang, Dimitar Hristovski, Dalton Schutte, Andrej Kastrin, Marcelo Fiszman, and Halil Kilicoglu. 2021. Drug repurposing for covid-19 via knowledge graph completion. *Journal of biomedical informatics*, 115:103696.

Zexuan Zhong and Danqi Chen. 2021. A frustratingly easy approach for entity and relation extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 50–61, Online. Association for Computational Linguistics.