SapBert：用于医学实体对齐的模型

📎 2021.naacl-main.334.pdfLiu, Fangyu, et al. "Self-alignment pretraining for biomedical entity representations." *arXiv preprint arXiv:2010.11784* (2020).

论文核心内容总结

本文提出了一种针对生物医学实体表示的自对齐预训练模型 **SAPBERT（Self-Alignment Pretrained BERT）**，旨在解决生物医学领域中实体异名（如同义词、缩写、品牌名）导致的语义表示难题。通过结合UMLS（包含400万+生物医学概念的本体库）和度量学习框架，SAPBERT将同一概念的不同名称在表示空间中聚类，从而提升实体链接（MEL）等任务的性能。实验表明，SAPBERT在6个MEL基准数据集上均取得了state-of-the-art（SOTA）结果，甚至在科学文本领域无需任务微调即可超越此前的监督学习方法。

可以学习的点

- 这个处理硬样本的方法，针对难的样本，同时使用 MS 损失函数，加强处理顽固样本。这里顽固样本是正样本 ==> 修改下损失函数，就可以用于处理顽固的负样本了

核心创新点

1. **自对齐预训练框架**
   设计了基于UMLS的度量学习方案，通过将同一概念的不同名称（如“Hydroxychloroquine”和“Plaquenil”）在表示空间中强制聚类，解决生物医学实体的异名问题。该框架利用在线硬样本挖掘（Online Hard Pairs Mining）动态选择难例训练样本，并采用Multi-Similarity（MS）损失函数优化表示空间，使同义词的向量距离更近，非同义词更远。
2. **端到端单模型解决方案**
   与传统复杂流水线方法（如结合tf-idf、字符串匹配的混合系统）不同，SAPBERT通过单一模型实现医学实体链接，测试时仅需最近邻搜索即可完成预测，大幅简化了流程并提升泛化能力。

3. **无需任务监督的高效迁移**
   在科学文本领域，SAPBERT无需针对具体任务微调，仅通过UMLS预训练即可超越此前依赖任务标签的监督学习SOTA模型（如BIOSYN），证明了预训练策略的通用性和鲁棒性。
4. **参数高效的适配器变体**
   引入ADAPTER模块，在保持性能的同时大幅减少训练参数（如1%参数版本仍可达到接近完整模型的效果），为资源受限场景提供了可行方案。

关键实验结果

- **性能提升**：在6个MEL数据集（如NCBI、BC5CDR、MedMentions等）上，SAPBERT相较于基线模型（如BIOBERT、PUBMEDBERT）提升显著，例如在社交媒体数据集COMETA上，Acc@1提升幅度可达31%。
- **可视化验证**：t-SNE结果显示，SAPBERT处理后同一概念的不同名称（如“Hydroxychloroquine”的别名）在表示空间中形成紧凑簇，而原始PUBMEDBERT的表示则较为分散。
- **消融实验**：关闭在线硬样本挖掘会导致性能显著下降（如COMETA数据集Acc@1从67.2%降至52.3%），证明该模块的必要性。

科学意义与应用价值

- **生物医学文本挖掘**：为实体链接、概念归一化等任务提供了更精准的表示，可应用于文献检索、临床决策支持和知识图谱构建。
- **跨领域迁移启示**：通过知识图谱（如UMLS）引导预训练的思路，为其他领域（如化学、基因科学）的表示学习提供了参考范式。

如需进一步探讨方法细节、实验对比或未来工作方向，可随时提出具体问题。

文章的一些技术细节:

1、训练阶段的数据配对(Formal Definition)

1.1 构建配对的元组

1. **元组（x, y）的含义**

- **x**：生物医学实体的“名字”，可以是正式名称、缩写、别名等。
  例如：“Remdesivir”（正式名）、“HCQ”（缩写）、“Plaquenil”（品牌名）。
- **y**：这个名字对应的“唯一身份标签”（在UMLS中称为CUI）。
  例如：所有指代“Remdesivir”的名字都对应同一个CUI（如C4726677），就像每个人有唯一的身份证号。

1.2 预训练阶段

2. **预训练（Self-Alignment Pretraining）阶段**
   - **目标**：让模型学会“不同名字如果对应同一个CUI，就应该在向量空间中靠得很近”。
   - **数据**：使用UMLS中的（名字，CUI）对，比如（“Remdesivir”, C4726677）、（“Plaquenil”, C4726677）。
   - **核心动作**：通过模型将名字转换成向量（函数f的作用），然后调整参数，让`同一CUI的 ID`向量尽可能相似（用余弦相似度衡量）。

1.3 微调阶段

3. **微调（Finetuning）阶段**
   - **目标**：将预训练的能力应用到具体任务，比如“把社交媒体中的‘嗓子痒’映射到医学概念”。
   - **数据**：使用（实体提及，本体映射）对，比如（“scratchy throat”, 102618009），其中102618009是“嗓子痒”在医学本体中的唯一标识。
   - **核心动作**：用具体任务的数据进一步调整模型，让模型学会将用户输入的“非标准名称”映射到正确的医学概念。

1.4 小总结

> 说人话

- **预训练阶段**：相当于教模型“背词典”——告诉它“HCQ”和“Hydroxychloroquine”都是同一种药，它们的“数字指纹”应该相似。
- **微调阶段**：相当于教模型“做应用题”——用具体场景中的例子（如“scratchy throat”对应某个医学概念），让它学会将日常用语映射到标准医学术语。

通过这种方式，模型就能解决生物医学中“同一概念多种叫法”的问题，比如在实体链接任务中，将不同的名称正确对应到同一个标准概念上。

2、 构建挖掘的硬样本，就是很难配对的样本

2.1、用「考试刷题」类比核心逻辑

- **普通训练**：像做100道题里80道简单题+20道难题，但模型总在练简单题，成绩难提高。
- **硬样本挖掘**：专门挑出20道难题（硬正例/负例）重点练，因为这些题最能暴露模型弱点。

2.2、三步拆解在线硬样本挖掘

1. 构造「三元组」：锚点+正例+负例

- **锚点（(x_a)）**：随便选一个名字，比如“Hydroxychloroquine”（羟氯喹）。
- **正例（(x_p)）**：和锚点同一概念的名字，比如“Plaquenil”（品牌名），它们的CUI相同（(y_a=y_p)）。
- **负例（(x_n)）**：不同概念的名字，比如“Vitamin C”（维生素C），CUI不同（(y_a≠y_n)）。

2. 用「距离条件」筛选难题

- **距离公式**：（ $|f(x_a) - f(x_p)|_2 < |f(x_a) - f(x_n)|_2 + \lambda$ ）
  - 白话翻译：**如果锚点到正例的距离 < 锚点到负例的距离 + 预设差距（λ），说明这是一道难题**。
  - 例子：正常情况下，“Hydroxychloroquine”到“Plaquenil”的距离应该小于到“Vitamin C”的距离。但如果模型学歪了，可能出现“Hydroxychloroquine”到“Vitamin C”的距离反而更近（违反条件），这就是需要纠正的难题。

3. 收集「硬样本对」

- **硬正例对**：锚点和正例本应相似，但模型没学好（距离太远）。
- **硬负例对**：锚点和负例本应不相似，但模型搞错了（距离太近）。
- **作用**：让模型重点学习这些“搞混”的例子，强制拉近正确对、推远错误对。

放一下文章里的源码

```python
from pytorch_metric_learning import miners, losses

self.miner = miners.TripletMarginMiner(margin=miner_margin, type_of_triplets=type_of_triplets)
self.loss = losses.MultiSimilarityLoss(alpha=1, beta=60, base=0.5) # 1,2,3; 40,50,60
# miner会返回一个hardpair,然后根据整个hard_pair汲取算这个embedding

query_embed = torch.cat([query_embed1, query_embed2], dim=0)

labels = torch.cat([labels, labels], dim=0)
if self.use_miner:
    hard_pairs = self.miner(query_embed, labels)
    return self.loss(query_embed, labels, hard_pairs)

```

2.3、为什么必须做硬样本挖掘?

- **数据特点**：生物医学中大部分名字容易区分（如“头痛”和“肺炎”），但少数难例（如“HCQ”和“Hydroxychloroquine”）才是性能瓶颈。
- **实验验证**：关闭该功能后，模型在COMETA数据集的准确率从67.2%暴跌到52.3%，说明不练难题就考不好。

如何处理硬样本的挖掘

1. 动态筛选机制

- 每批数据中自动生成所有可能的三元组，实时计算距离并筛选出违反条件的硬样本，避免人工标注。
- 优势：生物医学数据中硬样本分布不均，动态挖掘可自适应聚焦难点，比随机采样效率高 30%+。

2. 阈值 λ 的作用

- λ 为负数（文中 λ=−0.2），允许负例与正例的距离更接近，从而筛选出“更难”的样本对。
- 例如：若 λ=0，仅当正例距离 > 负例距离时才视为硬样本；λ=−0.2 时，正例距离比负例距离大 0.1 也会被视为硬样本（即更严格的筛选标准）。

3. 与 MS 损失的协同

- 硬样本挖掘为 MS 损失提供“高价值训练素材”，MS 损失则通过权重调整，让硬样本获得更多梯度更新信号（如 β=50 时，硬正例对的损失权重更高）因为硬正例子的 S_{ip}应该是很大的，但是实际值 太小了，-kesi 都有可能变复制了，负负得正，e 就变得很大了

2.4、一句话总结

从每批数据中找出模型最容易搞混的“同名异义”和“异名同义”例子，强迫模型纠正错误，就像用难题特训提高考试成绩。

3、构造一个损失

$$\mathcal{L} = \frac{1}{|\mathcal{X}_b|}\sum_{i=1}^{|\mathcal{X}_b|}\left(\frac{1}{\alpha}\log\left(1+\sum_{n\in\mathcal{N}_i}e^{\alpha(\mathbf{S}_{in}-\epsilon)}\right)\right. \\ \left. +\frac{1}{\beta}\log\left(1+\sum_{p\in\mathcal{P}_i}e^{-\beta(\mathbf{S}_{ip}-\epsilon)}\right)\right), \tag{2}$$

where α / β are temperature scaling ... an offset

3.1、用「班级座位表」类比相似度矩阵

- **输入**：一批名字（如“Hydroxychloroquine”“Plaquenil”“Vitamin C”），经BERT转化为向量。
- **相似度矩阵S**：像一张班级座位表，每个格子(S_{ij})记录第i个名字和第j个名字的“相似程度”（用余弦相似度计算）。
  - 例如：(S_{12})表示“Hydroxychloroquine”和“Plaquenil”的相似度（应接近1），(S_{13})表示它们和“Vitamin C”的相似度（应接近0）。

3.2、MS损失的核心：「朋友分组」与「陌生人隔离」

1. 处理负例（陌生人）：推开不相似的名字

- **公式第一部分**：$(\frac{1}{\alpha}\log\left(1+\sum_{n\in\mathcal{N}i}e^{\alpha(Sin-\epsilon)}\right))$
- **白话解释**：
  - 对于每个名字i（如“Hydroxychloroquine”），找到所有和它不同概念的名字n（负例，如“Vitamin C”）。
  - 如果i和n的相似度(S_{in})太高（说明模型误判它们相似），指数项会变大，导致损失增加，强迫模型降低它们的相似度。
  - **α的作用**：像“放大镜”，α越大，模型对“误判的陌生人”越敏感，惩罚越重。

> S_{in} 这值应该是要越小越好，这样 e 之后，就会小，所以 loss 也会小

2. 处理正例（朋友）：拉近相似的名字

- **公式第二部分**：$(\frac{1}{\beta}\log\left(1+\sum_{p\in\mathcal{P}i}e^{-\beta(Sip-\epsilon)}\right))$
- **白话解释**：
  - 对于每个名字i，找到和它同概念的名字p（正例，如“Plaquenil”）。
  - 如果i和p的相似度(S_{ip})太低（说明模型没认出它们是同一概念），指数项会变大，损失增加，强迫模型提高它们的相似度。
  - **β的作用**：类似α，但专门针对“没认亲的朋友”，β越大，模型越重视拉近正例。

> S_{ip} 这值应该是要越大越好，这样*-β 之后，指数部分是负的，就会 loss 变小

3.3、ϵ：给相似度“调基准线”

- **作用**：像给考试分数“划及格线”，ϵ决定了模型认为“多相似才算相似”。

- **例子**：若ϵ=0.5，当(S_{ip}=0.6)（略高于及格线），(S_{ip}−ϵ=0.1)，指数项(e^{−\beta*0.1})较小，损失也小；若(S_{ip}=0.4)（不及格），损失会显著增加，强迫模型调整。

3.4、MS损失的终极目标

- **动态平衡**：同时处理所有正负样本对，让模型学会：
  a. 把同一概念的名字（如“HCQ”和“Hydroxychloroquine”）的向量“粘在一起”；
  b. 把不同概念的名字（如“新冠病毒”和“流感病毒”）的向量“推开”。
- **为什么选MS损失**：在视觉识别中已证明能有效处理复杂相似关系，改编到生物医学领域后，比传统损失函数（如三元组损失）更能捕捉实体异名的细微差异。

3.5、一句话总结

MS损失就像一个“纪律委员”，一边惩罚“把陌生人当朋友”的错误（负例项），一边批评“把朋友当陌生人”的疏忽（正例项），通过调整惩罚力度（α、β）和标准（ϵ），让模型学会正确分组生物医学实体的不同名字。

模型训练阶段

1. SAPBERT 在不同的预训练的 Bert 模型上使用这里的 Self-aligning Pretrained 继续训练

1. 基于现有的 Bert 模型继续训练:
   - 选用现有 BERT 模型（如 PUBMEDBERT、BIOBERT 等）作为初始化参数，保留其已学习的语言表示能力。模型结构不变，
   - 仅通过预训练任务调整参数，使生物医学实体的表示空间自对齐（即同一概念的不同名称向量更接近）。
2. 基于 UMLS 的自对齐预训练数据输入：
   - 使用 UMLS 中的（实体名称，CUI）对，如（“Hydroxychloroquine”, C0019756）和（“Plaquenil”, C0019756），构建同义对样本。
   - 目标函数：通过度量学习强制同一 CUI 的名称向量相似，不同 CUI 的向量相异，具体通过以下模块实现：
   - 在线硬样本挖掘：从每批数据中筛选“难例”（如模型误判的非同义词对或未聚簇的同义词对），增强训练针对性。

- Multi-Similarity（MS）损失：同时优化正例（同义词）拉近和负例（非同义词）推开，动态调整表示空间。

3. 参数更新策略

- 完整模型微调：对基础模型的所有参数进行更新，适用于计算资源充足的场景。
- 适配器（ADAPTER）微调：在基础模型中插入轻量级 ADAPTER 模块，仅训练新模块参数（如 13% 或 1% 的参数），适用于参数高效场景。

2、SAPBERT训练及评估数据集表格

类别	数据集名称	数据来源/特点	关键统计信息
预训练数据集	UMLS 2020AA	全球最大生物医学本体库，包含150+受控词表（如MeSH、SNOMED CT、RxNorm等），覆盖400万+概念及1000万+同义词，支持多语言生物医学概念归一化。	处理后得到971万+（名字，CUI）对，进一步生成1179万+同义配对样本（每个概念最多保留50对同义词），用于自对齐预训练。
评估数据集	**科学文本领域**		
	NCBI disease	793篇PubMed摘要，6881个疾病提及，映射至MEDIC字典，用于疾病实体链接。	训练集5134例，验证集787例，测试集960例，涉及11915个概念、71923个表面形式。

	BC5CDR–d/c	1500篇PubMed文章，包含4409个化学实体、5818个疾病提及及3116个相互作用，疾病映射至MEDIC，化学实体映射至CTD字典。	疾病子任务（BC5CDR–d）：训练集4182例，测试集4424例；化学子任务（BC5CDR–c）：训练集5203例，测试集5385例，均涉及11915个概念。
	MedMentions	超4000篇摘要，35万+提及，链接至UMLS 2017AA，数据规模大（341万+概念、1481万+表面形式），传统方法难以处理。	训练集28.2万例，测试集7.04万例，概念覆盖范围极广，验证模型泛化能力。
评估数据集	**社交媒体领域**		
	AskAPatient	从askapatient.com收集的17324条药物不良反应注释，提及映射至SNOMED–CT和AMT，语言更口语化（如“headache”“nausea”）。	10折交叉验证，训练集约1.57万例，测试集约866例，涉及1036个概念。

	COMETA	从Reddit健康讨论中提取的2万+医学提及，映射至SNOMED-CT，包含“分层通用”（stratified general）和“零样本通用”（zeroshot general）两种拆分，后者更具挑战性（测试集概念未在训练集出现）。	分层通用拆分：训练集1.35万例，测试集4350例；零样本通用拆分：训练集1.41万例，测试集3995例，涉及35万+概念、91万+表面形式。

3、SAPBERT核心超参数表格

阶段	超参数名称	数值	作用及说明
预训练阶段	学习率（Learning Rate）	2e-5（0.00002）	控制参数更新步长，小学习率确保模型在大规模UMLS数据中稳定收敛，避免过拟合。
	批量大小（Batch Size）	512	每次迭代处理512个样本，平衡内存消耗与梯度稳定性，较大批量使训练更高效。
	训练轮次（Epochs）	1	完整训练1轮UMLS数据（约50k次迭代），因UMLS数据量极大，1轮已足够捕捉语义关系，多轮易过拟合。

	在线硬样本挖掘阈值（λ）	−0.2	筛选“难题”的标准：当锚点到正例的距离 < 锚点到负例的距离 + λ时，判定为硬样本。负数λ允许负例更接近正例，从而挖掘更具挑战性的样本对。
MS损失函数	温度系数α（Alpha）	2	调节负例惩罚力度：α越大，模型对“误判为相似的非同义词”惩罚越重，强迫不同概念的表示距离增大。
	温度系数β（Beta）	50	调节正例拉近力度：β越大，模型对“未聚在一起的同义词”惩罚越重，强迫同概念表示距离缩小。
	偏移量ϵ（Epsilon）	0.5	设定相似度基准线：相似度超过ϵ（0.5）才被视为“相似”，低于则触发损失优化。
微调阶段	训练轮次（Epochs）	科学文本：3 AskAPatient：15 COMETA：10	科学文本数据较规范，3轮即可；社交媒体数据更嘈杂（如口语化表达），需更多轮次（10–15轮）适应非标准术语。

	最大序列长度（Max Seq Length）	25	限制输入文本长度，截断超过25个token的句子，平衡计算效率与语义信息保留（生物医学实体多为短词或短语）。

结论

在这个基座Bert的基础上，继续使用这个配对后的UMLS数据进行训练，穿插了Adapter模块，只更新部分权重

model	scientific language								social media language			
	NCBI		BC5CDR-d		BC5CDR-c		MedMentions		AskAPatient		COMETA	
	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5	@1	@5
vanilla BERT (Devlin et al., 2019)	67.6	77.0	81.4	89.1	79.8	91.2	39.6	60.2	38.2	43.3	40.4	47.7
+ SAPBERT	91.6	95.2	92.7	95.4	96.1	98.0	52.5	72.6	68.4	87.6	59.5	76.8
BIOBERT (Lee et al., 2020)	71.3	84.1	79.8	92.3	74.0	90.0	24.2	38.5	41.4	51.5	35.9	46.1
+ SAPBERT	91.0	94.7	93.3	95.5	96.6	97.6	53.0	73.7	72.4	89.1	63.3	77.0
BLUEBERT (Peng et al., 2019)	75.7	87.2	83.2	91.0	87.7	94.1	41.6	61.9	41.5	48.5	42.9	52.9
+ SAPBERT	90.9	94.0	93.4	96.0	96.7	98.2	49.6	73.1	72.4	89.4	66.0	78.8
CLINICALBERT (Alsentzer et al., 2019)	72.1	84.5	82.7	91.6	75.9	88.5	43.9	54.3	43.1	51.8	40.6	61.8
+ SAPBERT	91.1	95.1	93.0	95.7	96.6	97.7	51.5	73.0	71.1	88.5	64.3	77.3
SCIBERT (Beltagy et al., 2019)	85.1	88.4	89.3	92.8	94.2	95.5	42.3	51.9	48.0	54.8	45.8	66.8
+ SAPBERT	91.7	95.2	93.3	95.7	96.6	98.0	50.1	73.9	72.1	88.7	64.5	77.5
UMLSBERT (Michalopoulos et al., 2020)	77.0	85.4	85.5	92.5	88.9	94.1	36.1	55.8	44.4	54.5	44.6	53.0
+ SAPBERT	91.2	95.2	92.8	95.5	96.6	97.7	52.1	73.2	72.6	89.3	63.4	76.9
PUBMEDBERT (Gu et al., 2020)	77.8	86.9	89.0	93.8	93.0	94.6	43.9	64.7	42.5	49.6	46.8	53.2
+ SAPBERT	92.0	95.6	93.5	96.0	96.5	98.2	50.8	74.4	70.5	88.9	65.9	77.9
supervised SOTA	91.1	93.9	93.2	96.0	96.6	97.2	OOM	OOM	87.5	-	79.0	-
PUBMEDBERT	77.8	86.9	89.0	93.8	93.0	94.6	43.9	64.7	42.5	49.6	46.8	53.2
+ SAPBERT	92.0	95.6	93.5	96.0	96.5	98.2	50.8	74.4	70.5	88.9	65.9	77.9
+ SAPBERT (ADAPTER$_{13\%}$)	91.5	95.8	93.6	96.3	96.5	98.0	50.7	**75.0**†	67.5	87.1	64.5	74.9
+ SAPBERT (ADAPTER$_{1\%}$)	90.9	95.4	**93.8**†	**96.5**†	96.5	97.9	**52.2**†	74.8	65.7	84.0	63.5	74.2
+ SAPBERT (FINE-TUNED)	92.3	95.5	93.2	95.4	96.5	97.9	50.4	73.9	**89.0**†	**96.2**†	75.1 (**81.1**†)	85.5 (**86.1**†)
BIOSYN	91.1	93.9	93.2	96.0	96.6	97.2	OOM	OOM	82.6	87.0	71.3	77.8
+ (init. w/) SAPBERT	**92.5**†	**96.2**†	93.6	96.2	**96.8**	**98.4**†	OOM	OOM	87.6	95.6	77.0	84.2