# 临床术语标准化——基于LLM实现

作者：WMJ, CRJ

## 背景与方法调研

1. 项目目标

本次大作业的主题是实现一个面向语义相似度匹配的临床术语标准化方法。具体目标包括：

1. 针对中文电子病历中的诊断实体，通过给定的诊断原词，预测对应的标准词；
2. 通过对不同算法的比较和分析，选择最适合本项目的算法；
3. 评估预测结果的准确性，主要通过**Micro-F1**得分来衡量。
4. 任务背景

临床术语标准化方法是医学统计中不可或缺的一项方法，具有至关重要的作用。由于临床术语的多样性，同一种诊断、手术、药品、检查、化验、症状等可能有多种不同的写法，这给后续的统计分析带来了困难。通过标准化，可以确保各种术语的一致性，从而提高统计分析的准确性和效率。

临床术语标准化方法的发展经历了多个阶段。最初，人们主要依赖于人工操作和规则制定来进行标准化。然而，随着自然语言处理技术的发展，机器学习方法开始被广泛应用于临床术语标准化。近年来，深度学习模型如循环神经网络（RNN）、Transformer等也在该领域取得了显著的成果。本质上，临床术语标准化方法也是语义相似度匹配任务的一种。但是由于原词表述方式过于多样，单一的匹配模型很难获得很好的效果。

1. 可行方法

以下是五种可能适用于临床术语标准化的方法：

1. 基于**规则**的方法：根据已有的医学知识和术语规则，人工制定一系列的映射规则，将原词映射到标准词。这种方法需要大量的医学知识和人力投入；
2. 基于**统计**的方法：利用词频统计和机器学习算法（如朴素贝叶斯、支持向量机等）进行分类或回归。这种方法需要对大量的语料库进行预处理和特征工程；
3. 基于**深度学习**的方法：利用深度神经网络（如RNN、Transformer等）进行端到端的训练和预测。这种方法需要大量的计算资源和数据；
4. **微调预训练语言模型**：利用大规模的预训练语言模型（如BERT、GPT等）进行微调，使其适应临床术语标准化的任务。这种方法需要大量的计算资源和数据；
5. **集成**方法：将不同的算法进行集成，利用各自的优点，提高预测精度。这种方法需要更多的设计和实现工作。

## 方法间比较与分析

1. 方法优缺点与实现难度分析

对上一节五种方法的优缺点和实现难度分析：

1. 基于**规则**的方法：优点是规则一旦制定就可以直接使用，且结果直观易懂；缺点是需要投入大量的人力物力去制定规则，且对于新的术语或复杂的情况可能无法很好地处理。实现难度一般，主要取决于规则的数量和复杂度；
2. 基于**统计**的方法：优点是对大规模语料库中的术语可以进行较好的分类或回归；缺点是需要大量的预处理和特征工程工作，且对于新的术语或复杂的情况可能无法很好地处理。实现难度一般，主要取决于语料库的大小和特征工程的复杂性；
3. 基于**深度学习**的方法：优点是可以端到端地训练模型，避免复杂的特征工程；缺点是需要大量的计算资源和数据，且对于新的术语或复杂的情况可能无法很好地处理。实现难度较高，主要取决于模型的复杂性和训练数据的大小；
4. **预训练语言模型**：优点是可以利用大规模的预训练语言模型进行微调，提高模型的泛化能力；缺点是需要大量的计算资源和数据，且对于特定的领域可能需要额外的调整和优化。实现难度较高，主要取决于预训练模型的大小和微调的复杂性；
5. **集成**方法：优点是可以综合利用不同算法的优点，提高预测精度；缺点是需要更多的设计和实现工作，且可能会增加模型的复杂性。实现难度较高，主要取决于集成方案的设计和实施。
6. 方法可行性分析

根据本项目的要求和条件，我们需要针对中文电子病历中的诊断实体进行标准化，并评估预测结果的准确性。考虑到临床术语的多样性和语义相似度匹配任务的特点，单一的匹配模型很难获得很好的效果。因此，我们可以选择基于**ChatGPT Training Pipeline**[1]对预训练好的**大型语言模型(Large Language Model, LLM)**进行增量预训练和有监督微调的方法，使其适应临床术语标准化的任务，同时减少从头开始训练的工作量。

## 使用方法

本次大作业是基于ChatGPT Training Pipeline训练LLM实现的，即第一阶段进行**PT(Continue PreTraining)增量预训练**，第二阶段进行**SFT(Supervised Fine-tuning)有监督微调**，后续强化学习阶段不太适合于我们的任务和数据集，因此只要进行PT和SFT就好了。我们选用的模型是HuggingFace上的**shibing624/ziya-llama-13b-medical-merged**[2]，它在240万条中英文医疗数据集**shibing624/medical**[3]数据集上SFT微调了一版**Ziya-LLaMA-13B**[4]模型，医疗问答效果有提升，发布微调后的完整模型权重(单轮对话)，更适合于临床术语标准化任务。

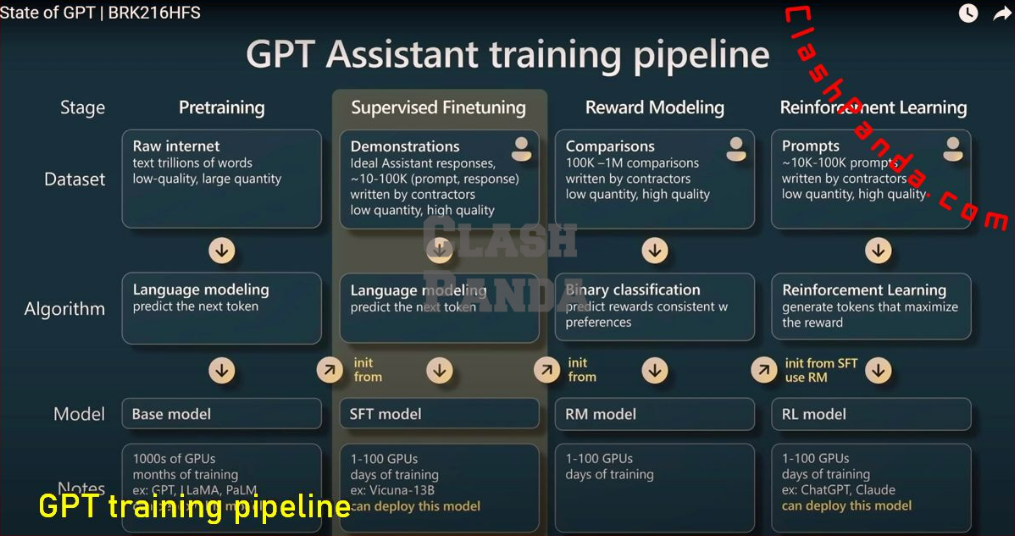


图 1 GPT Training Pipeline[1]

该项目的算法实现难度较高。为了节省时间，便基于Github开源项目代码**MedicalGPT**[2]对模型进行PT和SFT。需要针对临床术语标准化任务的特点进行定制化的数据处理和预处理，以供LLM进行训练，其中1万个标准词表用于PT，归一化标准化训练集用于SFT。测试的时候用类似ChatGPT问答的方式完成临床术语标准化任务，最后使用Micro-F1分数进行模型预测效果评估。

也可以不直接在问答就标准化，询问LLM，让其逐个进行匹配，但这种方法效率过低，，对1万个标准词逐个匹配的时间成本过高，因此没有采用这种方式。

## 主要步骤介绍

1. 项目完成主要步骤

本次大作业项目的完成主要有以下7个步骤：

1. **选用服务器**：LLM的训练与测试必须使用高性能AI服务器，在个人电脑上训练与测试是很困难的，除非愿意接受频繁的显存爆炸、龟速的训练与差劲的效果（至少1小时）；
2. **配置环境**：下载部署LLM，并根据要求配置环境，确保所有依赖项都已正确安装，以便顺利进行后续的实验步骤（至少6小时）；
3. **数据采集与预处理**：下载临床术语标准化数据集，划分训练集与验证集，并预处理成大模型能使用的格式（至少6小时）；
4. **增量预训练**：使用标准词表《国际疾病分类ICD-10北京临床版v601.xlsx》二次预训练LLM，以注入临床术语标准词知识（至少12小时）；
5. **有监督微调**：使用临床术语标准化样例数据集构造指令微调数据集，在预训练模型基础上做指令精调，以对齐指令意图，让大模型学习临床术语标准化工作（至少24小时）；
6. **模型测试**：用验证集对模型进行测试，对比真实数据与预测数据，计算Micro-F1，验证模型效果（至少6小时）；
7. **撰写文档与演讲**：撰写“README.md”、“复现教程.ipynb”、”大作业报告.docx”等文档，制作PPT进行演讲，向他人介绍工作（至少24小时）。

更详细的步骤说明请见”...代码.zip”压缩包根目录的“复现教程.ipynb”。

1. 难点及解决办法

本次大作业项目的完成主要有以下3个难点，并分别给出了解决或缓解问题的方法：

1. 算力资源困难：网上租用AutoDL高性能AI服务器训练；
2. 资金困难：通过学生认证、费用均摊、平时节省等方法缓解，并适时升降配置，在非必要时不启用GPU环境，在测试时不启用多GPU训练，仅在正式训练时租用高配置服务器；
3. 数据集过少：调整训练集、验证集比例，保证足够的验证集数据，尽量争取更多的训练集数据；
4. 任务过于复杂：使用网上开源代码，借助AI工具，参考一切能够利用的现成资料。

3. 项目分工

WMJ：完成了代码压缩包里包含的全部工作（模型选用、部署、训练、预测，相关处理代码编写，“复现教程.ipynb”撰写等），演讲，PPT制作，撰写大作业报告；

CRJ：思路讨论，演讲，PPT制作。

## 代码实现与解释性标注

见”...代码.zip”压缩包根目录的“复现教程.ipynb”和“README.md”，若需要更详细的信息可访问[5]。

## 项目总结评估及未来发展展望

1. 项目总结评估

在PT阶段，eval loss值随训练epoch变化情况进行可视化如图 1 所示，其中y轴是对数刻度。

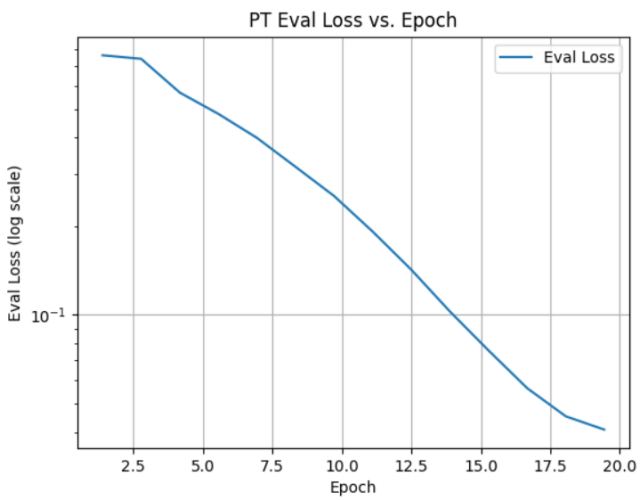


图 2 PT阶段loss-epoch变化情况

在SFT阶段，eval loss值随训练epoch变化情况进行可视化如图 2 所示，其中y轴是对数刻度。

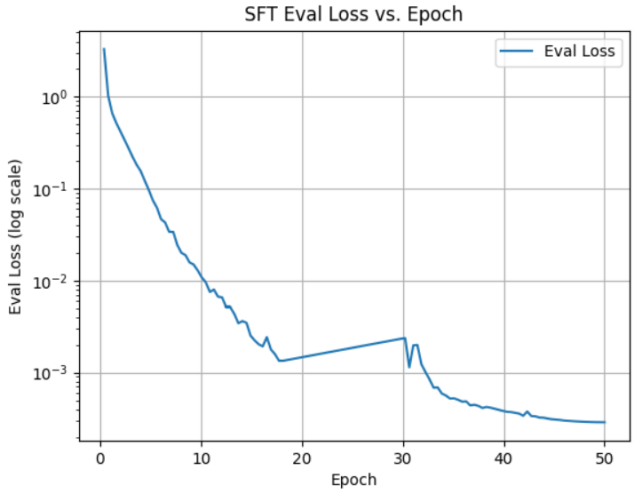


图 3 SFT阶段loss-epoch变化情况

由loss变化情况，基本可以看出模型的训练已经比较充分，SFT的loss经过50轮的训练，已经比一开始时降低了至少上千倍。

选取了验证集中7条数据组合的真实数据与预测数据对比作为模型预测效果的示例，如表 1 所示。

|  |  |  |
| --- | --- | --- |
| **诊断原词** | **归一化后的标准词（真实）** | **归一化后的标准词（预测）** |
| 盲肠腺瘤 | 盲肠良性肿瘤##腺瘤 | 盲肠良性肿瘤##腺瘤 |
| 冠心病-不稳定心绞痛 | 冠状动脉粥样硬化性心脏病##不稳定性心绞痛 | 冠状动脉粥样硬化性心脏病##不稳定性心绞痛 |
| 窦房结功能不全 | 窦房结功能低下 | 窦房结功能紊乱 |
| 右附睾头囊肿 | 附睾囊肿 | 附睾囊肿 |
| 双跟骨骨折 | 跟骨骨折 | 跟骨骨折 |
| G周宫内妊娠引产后 | 引产 | 妊娠状态 |
| 前列腺动态未定肿瘤 | 前列腺动态未定或动态未知的肿瘤 | 前列腺恶性肿瘤 |

表 1 预测效果示例

Micro-F1分数能够有效地衡量临床术语标准化任务的模型预测效果。因为Micro-F1考虑了各种类别的数量，所以更适用于数据分布不平衡的情况。在这种情况下，数量较多的类别对F1的影响会较大[6]。

本次实验在同一份训练集和验证集下调整不同参数训练了两次，不同的训练参数的Micro-F1分数如表 2 所示。

|  |  |  |  |
| --- | --- | --- | --- |
| **训练Epochs轮数** | **精确率** | **召回率** | **Micro-F1** |
| PT: 1, SFT: 10 | 0.684 | 0.705 | 0.694 |
| PT: 20, SFT: 50 | 0.681 | 0.684 | 0.682 |

表 2 不同训练参数的模型效果评价

第一次测试的时候PT训练了1个epochs，SFT阶段训练了10个epochs，训练集和验证集比例一样。它得到的Micro-F1分数约比PT训练20个epochs、SFT训练50个epochs的分数还高大约0.01，说明在该数据集该基座模型下，训练出来的模型的预测能力已经饱和，加多了59个epochs模型预测能力不仅没有增加，还衰减了一点点。分析此结果，可猜测不管再怎么加多epochs，只能达到大约0.7的极限Micro-F1，很难再提升。如果想要突破这个值，最好的也是最根本的方法是**扩大数据集**，其次是更换更好的模型。

1. 未来发展展望

本次实验选用了LLM来完成，一般的个人电脑的配置根本无法撑起这个训练。租用了AutoDL上的服务器，中途曾多次更换不同配置的服务器，最终使用了有4个RTX 4090(24GB)的服务器进行训练，每小时10元，最终约花费了400元。

经过本次实验，深有体会NLP工程对算力资源要求之高。没有足够的资金，没有足够的算力，根本撑不起先进的NLP实验。随着大模型的应用范围越来越广，NLP科研可能会越来越集中于对大模型的研究，包括多模态跨领域的大模型，它们一个模型就可以优秀地完成原本需要多种小模型完成各种特定小任务。只要资金与算力充足，用于完成特定任务的小模型存在的必要性会越来越少，对NLP研究的重要突破会越来越集中在资金与算力充足的企业手中，高校的NLP科研主要限于提供部分理论支持。

美国禁止RTX 4090等高端显卡在中国的销售是一件很严重的事情，中国的AI研究很难不会因此受到冲击。由此看出，中国自主研发高性能GPU是很有必要的，应该加快研发，为AI科研提供支持，中国科技才能赶上世界前列。

同样从算力来看，量子计算也是个很有前景的方向，一旦这个领域研究完善，AI算力将极大提高，通用人工智能（AGI）很有可能将成为现实。2020年3月9日，Google宣布推出TensorFlow Quantum量子机器学习框架[7]；同年5月20日，百度研究院量子计算研究所也宣布推出国内第一个也是全球第二个量子机器学习开源框架“量浆”（Paddle Quantum）[8]。可以看出业界前沿早已预见到量子计算在AI领域的极大潜能，不过量子计算的计算方式与电子计算有极大区别，因此量子机器学习的方式与传统机器学习也可能有极大差别，未来可能会换一种更适合于量子计算的AI理论，用量子计算来实现NLP任务，也许是未来有前景的研究发展方向之一。



图 4 2020年国产量子机器学习框架“量浆”发布会[8]

## 参考资料

1. Andrew Nguyen,GPT Training Pipeline: Unleashing the Power of Language Models,clashpanda,2023.<https://clashpanda.com/gpt-training-pipeline-unleashing-the-power-of-language-models/>.
2. <https://huggingface.co/shibing624/ziya-llama-13b-medical-merged>.
3. <https://huggingface.co/datasets/shibing624/medical>.
4. <https://huggingface.co/IDEA-CCNL/Ziya-LLaMA-13B-v1>.
5. Ming Xu,MedicalGPT: Training Medical GPT Model,github,2023.<https://github.com/shibing624/MedicalGPT/tree/main>.
6. 眼睛里进砖头了,模型评价指标：串起来看TP、FP、FN、TN，micro-f1、macro-f1，ROC、AUC,知乎,2022.<https://zhuanlan.zhihu.com/p/514732098?utm_id=0>.
7. 扬奇智能社区,论文|TensorFlow Quantum：用于量子机器学习的软件框架,知乎,2020,<https://zhuanlan.zhihu.com/p/115143609>.
8. 雷峰网leiphone,量桨，国内首个量子机器学习开源框架,百度百家号,2020.<https://baijiahao.baidu.com/s?id=1667189344613644452&wfr=spider&for=pc>.