

基于机器学习与深度学习的临床术语规范化分析

一、竞赛背景

近年来，临床术语（Clinical Terms）的规范化已成为门诊处方审核亟需解决的关键问题。统一规范的临床术语不仅是诊断的基础，也是疾病索引、分类统计以及门诊付费的关键。然而，临床术语的不规范化严重阻碍计算机的识别和分类效率，影响了医保审核和用药审核的准确性。传统上，医疗机构和卫生系统主要依赖人工对这些不规范的临床术语进行规范化整理，但随着医疗数据量的激增，人工处理已难以满足需求。因此，亟需借助先进的技术手段，特别是机器学习和深度学习方法，实现对临床术语的自动化规范化处理，以提高医疗服务的效率和质量。

二、竞赛数据

SNOMED CT（Systematized Nomenclature of Medicine - Clinical Terms）是全球最全面、多语言的临床医学术语系统，覆盖了临床发现、手术操作、身体部位、观察实体等在内的19个临床领域，超过35万个医学概念。凭借其强大的术语体系和广泛的应用价值，SNOMED CT已被超过80个国家的医疗机构和专业人士广泛采纳，成为全球医疗信息化领域的重要基石。

本次竞赛的数据集由神州医疗科技股份有限公司提供，内容丰富、结构严谨，涵盖临床医学多个关键领域，旨在为参赛者提供具有挑战性且具实用价值的数据资源，助力参赛者在竞赛中充分发挥专业能力，探索医学数据的深度应用与创新潜力。

完整竞赛数据集文件data_train.xlsx将在3月9日正式报名结束后通过电子邮件发给参赛队伍。表1所示为少量竞赛数据样本，相关说明如下：

表1 竞赛数据样本

ID	原始诊断名称	映射结果SCTID	映射结果中文同义词
1	1型糖尿病	46635009	1型糖尿病
2	1型糖尿病性不变的持续的蛋白尿性肾病	243421000119104	1型糖尿病导致的蛋白尿
3	1型糖尿病性低血糖	84371000119108	1型糖尿病致低血糖
4	1型糖尿病性前期肾病	90721000119101	1型糖尿病致慢性肾病1期
5	1型糖尿病性肾病	421893009	1型糖尿病致肾脏疾病

- **原始诊断名称：**以自然语言描述的疾病诊断名称，即医生或护士未规范表述的诊断，例如“糖尿病性不变的持续的蛋白尿性肾病”。

- **映射结果 SCTID：**每个诊断名称对应的系统化医学命名法—临床术语（SNOMED CT）唯一标识符，例如“243421000119104”。

- **映射结果中文同义词：**对应的规范化中文疾病名称，例如“糖尿病导致的蛋白尿”。

三、竞赛任务

1. 任务目标

本竞赛旨在开发基于机器学习和深度学习的模型，将原始诊断名称准确映射为规范化的临床术语，即映射结果中文同义词。通过数据预处理、模型设计与优化，确保模型在准确率、BLEU分数和ROUGE分数等指标上达到预期目标。此外，任务还将建立持续优化机制，以适应不断变化的医疗数据和需求，最终实现医疗数据的标准化处理，提升医疗信息化水平和临床决策支持能力。

2. 任务要求

根据原始诊断名称，构建机器学习和深度学习模型，能够准确输出规范化的映射结果中文名。

输入：词向量化后的原始诊断名称。

输出：规范后的映射结果中文同义词。

3. 模型评估与优化

- 使用BLEU分数、ROUGE分数和准确率等指标评估模型性能。
- 模型超参数调节的结果展示，表格或者图片（如条形图，折线图）。

四、作品提交要求

参赛选手最终至少需提交两份材料：一份**主题分析报告**和一份**源程序**。

- **主题分析报告：**详细描述数据处理、模型构建、评估结果和分析结论。以PDF形式提交。

- **源程序：**包含完整的源代码程序文件和readme.txt文件。源代码需配有必要的注释和说明，运行结果需和主题分析报告一致（包括并不限于图表、统计数字等），否则相应结果不计入作品评审中。

● **数据集：**数据集将由主办方提供，包含原始诊断名称和标准化映射结果。无需重复提交。

五、作品评审

作品评审将主要从以下四个方面进行综合评估：

1. 综合评估指标：如BLEU分数、ROUGE分数、准确率等。
2. 分析深度：分析报告的深度与广度，模型设计的合理性。
3. 技术实现：所选模型的实现效果，代码的清晰度与规范性。
4. 报告展示：报告的结构、清晰度、条理性及展示效果。

另外，作品必须使用给定的数据集，否则不予评审。