【手术】医疗术语标准化

1.问题定义

临床术语标准化任务是医学统计中不可或缺的一项任务。标准化（归一）要解决的问题就是为临床上各种不同说法找到对应的标准说法。数据集样本举例：

手术原词	归一化标准词	ICD编码
右肾上腺巨大肿瘤切除术	肾上腺病损切除术	07.2100
右眼硅油取出联合人工晶体II期植入术	玻璃体硅油取出术 + 眼内人工晶状体二期置入	14.6x02/13.7200
右叶甲状腺切除 + 左叶甲状腺部分切除术	单侧甲状腺切除伴他叶部分切除术	06.2x03
右眼白内障超声乳化抽吸术 + 人工晶体置入术	置入人工晶状体 + 白内障晶状体乳化和抽吸	13.7000/13.4100

本质上，临床术语标准化任务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。

2.数据分析

CHIP2019提供的标准化任务是「手术名称」，采用的是协和2017版ICD-9-CM-3编码。

2.1 数据统计

数据集相关指标统计如下：

	训练集	验证集	测试集
数据条数	4,000	1,000	2,000
最大长度	144	120	93
平均长度	12.33	12.31	12.35
标准词平均个数	1.07	1.06	1.06
标准词最大个数	7	4	5
“一一对应”比例	0.918	0.913	0.927

2.2 任务难点

通过对数据分析，发现此任务有以下几个难点

- **标准名数量大，字面相似度高**：协和2017版ICD-9-CM-3共有9467个标准名，训练集和开发集有1067个标准名，例如「硬脊膜外病损切除术」和「硬脊膜下病损切除术」。
- **匹配个数不确定，难度陡增**：手术原词"和"归一化标准词"之间存在4种匹配对应关系（参照上述「数据集样本举例」）：
  - "一对一"关系：一个手术原词对应一个归一化标准词
  - "一对多"关系：一个手术原词对应一个多个归一化标准词
    - 右眼硅油取出联合人工晶体II期植入术 ==> 玻璃体硅油取出术 + 眼内人工晶状体二期置入
  - "多对一"关系：多个手术原词对应一个归一化标准词
    - 右叶甲状腺切除 + 左叶甲状腺部分切除术 ==> 单侧甲状腺切除伴他叶部分切除术
  - "多对多"关系：多个手术原词对应多个归一化标准词
    - 右眼白内障超声乳化抽吸术 + 人工晶体置入术 ==> 置入人工晶状体 + 白内障晶状体乳化和抽吸
- **存在zero-shot/few-shot问题**：
  - 开发集480个标准名中有111个标准名没有出现在训练集中，66个出现1次。

3.总体策略

通过多次策略迭代后，我构建了「全量召回+深度匹配+个数预测」的总体策略如下，共分为3个阶段

- 召回阶段：召回通常应采用轻量级算法，加快计算性能，同时保证topK的准确率（⚠️假设标准编码共有N个，这里K<<N）。如若topK的准确率都不到90%， 那即使排序阶段top1的准确率为100%，整体指标都不会达到90%，更何况排序阶段准确率也较难实现高指标。对于本次任务，召回阶段的做法至关重要。经过多次尝试，我最终选择了**「全量召回」**的策略，**即同时查询历史编码和标准编码**，指标大幅提升。此外，语料清洗十分重要，例如，手术原词中会包含很多编码数字，应该做清洗处理。
  - 两种编码方式：
    - 标准编码：查找与待归一化的"手术原词"最相近的"标准词"。
    - 历史编码：在标注数据上查找与待归一化的"手术原词"最相近的数据，并取该条数据的归一化"标准词"作为候选答案 。
- 排序阶段：排序的本质是是精准排序，为实现top1的高指标通常会采用**「深度匹配模型」**（⚠️详见附录）。在排序阶段最重要的一点就是如何为**「深度匹配模型」**构建**负样本**，负样本的选择至关重要。
- 预测阶段：根据上述数据发现，手术原词"和"归一化标准词"之间存在4种匹配对应关系，如果采用规则方法明显过于武断，因此，参照竞赛第1名的解决方法，我将个数预测模型化，指标提升明显。

4.实验结果与分析

4.1 召回实验

在召回阶段，先后采取Tf-Idf/Word2Vector/最长公共子序列/编辑距离等进行相似度计算，主要实验指标如下：

(零氪数据：冷启动，共1101条数据)

TF-IDF/Top	1	2	5	10	20	50
历史编码+标准编码	23.83	30.87	40.33	49.1	54.46	58.77

(调研数据：冷启动，共8000条数据)

TF-IDF/Top	1	2	5	10	20	50
标准编码	47.88	56.72	68.96	78.08	85.2	92.0
历史检索	69.33	79.3	85.23	87.21	87.6	88.2
历史编码+标准编码	76.29	86.26	93.98	96.14	97.27	98.21
历史编码+标准编码（清洗）	77.33	87.49	94.73	96.90	98.21	98.97

Word2Vec/Top	1	2	5	10	20	50
标准编码	36.78	45.91	55.22	63.97	76.76	85.42
历史检索	63.40	73.94	81.84	84.57	86.07	87.67
历史编码+标准编码	70.37	80.99	89.75	92.29	94.54	96.33

由实验结果可以看出，采用**「全量召回」**的**历史编码+标准编码**策略可以将top20的准确率提升近10%。

4.2 排序+个数预测实验

直接采取BERT模型做深度文本匹配和个数预测，对于构造负样本的方法主要为：

- 第一种方法：取Top20候选+10倍正例
- 第二种方法：通过最长公共子序列进行相似度计算，通过相似度阈值进行负样本构建（参考自第1名解决方案）：
  - 标准词集合R， 对∀标准词$_j \in R$ ， 以及相似度阈值

    if $\frac{LCS(\text{标准词}_j,\text{原始词}_i)}{max(len(\text{原始词}_i),len(\text{标准词}_j))} \geq thres \;||\; \frac{LCS(\text{标准词}_j,\text{标准词}_i)}{max(len(\text{标准词}_j),len(\text{标准词}_i))} \geq thres$

    then add $<\text{原始词}_i, \text{标准词}_j, 0>$

最终的实验指标为：

任务	Dev	Test
个数预测：BERT+数据增强	97.8	98.15
排序：取Top20候选+10倍正例	92.92	91.76
最终指标	91.8	91.2

在「个数预测任务」中，如直接采用规则，准确率dev为94.4%，而模型化后达到97.8%，提升3.4%。

上述指标与第1名的指标92%相比较为接近（92%为Dev集合指标，未做模型融合），与最初的baseline（准确率为50%+）相比有了明显提升。在最终20支提交结果的队伍当中，平均指标在70.7%，而前三名在92%+，整体差距较大，可见本次任务的挑战性。

心得感悟

- 本次「医疗术语标准化」任务探索的主要指标提升点在于：
    1. 采用全量召回，召回阶段top20的准确提升10%，这是指标提升的关键
    2. 采用个数预测模型化，指标提升3.4%。
    3. 构造负样本，这主要还是得益于全量召回指标的大幅提升。
- 本次「医疗术语标准化」任务只是简单的策略探索，在实际工程应用中应考虑：
    1. 考虑具体的应用场景和业务，具备明确的问题定义。
    2. 兼顾计算性能，本次只是直接应用了BERT，同时BERT词表中缺少医疗重要的关键字，如**髌**、**跗**等。
    3. 补充上下文信息：回溯手术原词所在的长文本，提取更充足的上下文信息；
    4. 拓展相关特征：如对手术原词输入抽取细粒度的**部位**和**术式**等实体。