基于BERT蕴含分数排序的术语标准化系统

认知医疗组
AI Labs
云知声

任务描述

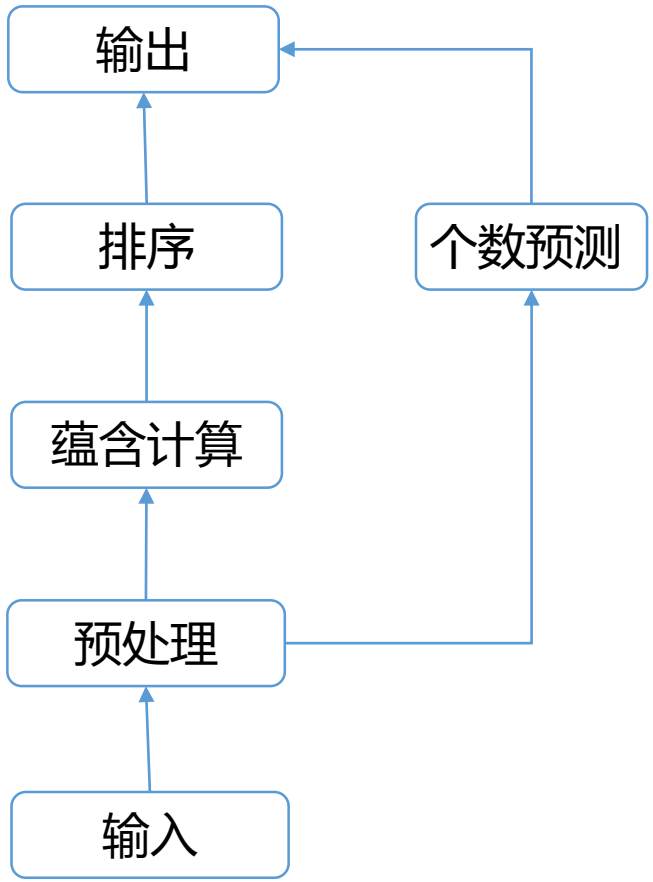
- 给定一手术原词，要求给出其对应的手术标准词（ICD9-2017协和临床版）
  - 比如
    - VVI型永久心脏起搏器植入术　　单腔永久起搏器置入术
    - 埋藏式单腔心脏起搏器安置术　　单腔永久起搏器置入术
- 问题抽象
  - 打分排序问题

问题抽象与难点分析

- 难点1：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
  - 硬脊膜外病损切除术 硬脊膜下病损切除术
- 难点2：标准名个数不确定
- 难点3：zero-shot以及few-shot
  - 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

整体方案

- 模块1：数据预处理
- **模块2：基于BERT蕴含推理**
- 模块3：回归排序
- **模块4：数量预测**



数据预处理

- 规范化
  - 全角转半角
  - 数字标准化
  - 去掉输入中的编码
- 根据训练数据中出现的英文数据，修改了标准名

输入	原始标准名	修改后标准名
PICC	经外周静脉穿刺中心静脉置管术	经外周静脉穿刺中心静脉置管术（PICC）

- 对包含不的标准名重新进行了修改

原始标准名	修改后标准名
手指骨折开放性复位术*不伴内固定*	手指骨折开放性复位术

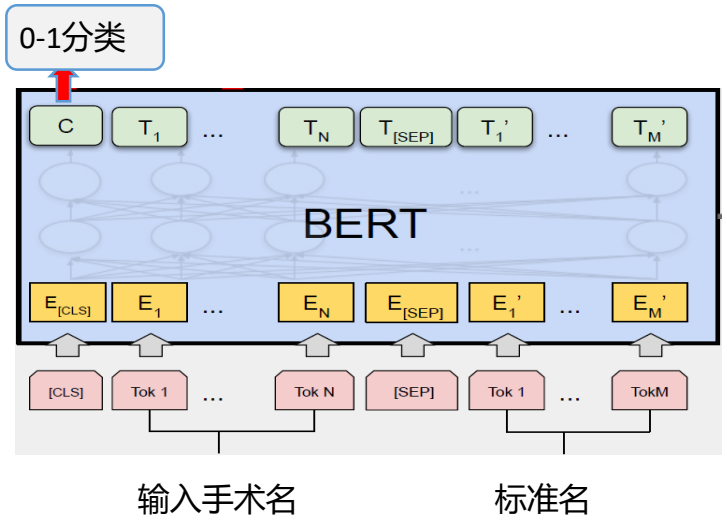
蕴含推理

- 任务
  - BERT fine-tuning 0-1分类任务
  - query为partA，标准名称为partB
- 训练集构造
  - 正例： <原始词$_i$,标准词$_i$,1>
  - 负例：标准词集合R， 对∀标准词$_j$∈ R ， 以及相似度阈值thres

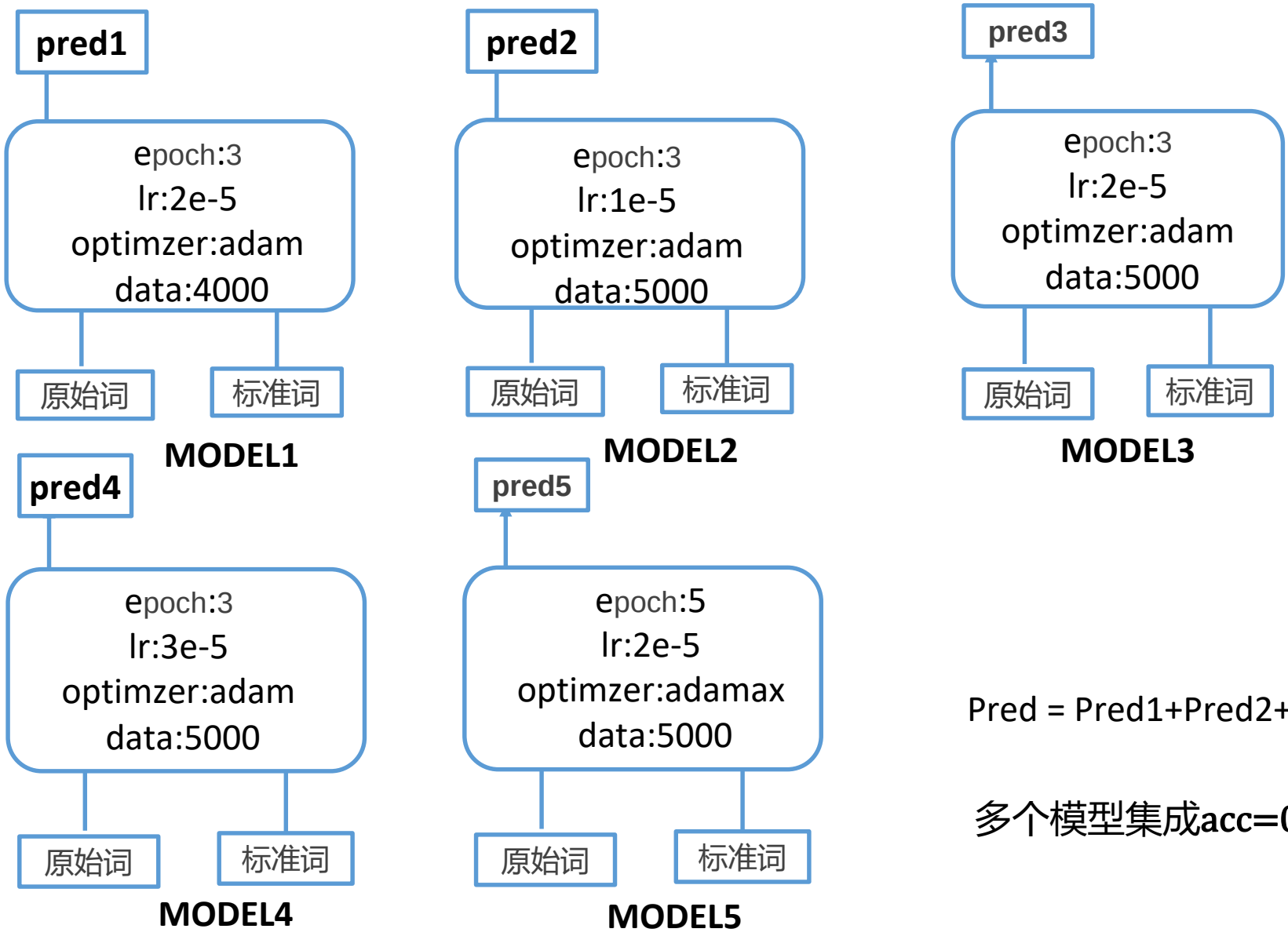
$$\textbf{if} \quad \frac{LCS(\text{标准词}_j,\text{原始词}_i)}{max(len(\text{原始词}_i),len(\text{标准词}_j))} \geq thres \;||\; \frac{LCS(\text{标准词}_j,\text{标准词}_i)}{max(len(\text{标准词}_j),len(\text{标准词}_i))} \geq thres$$

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

- 训练
  - **vocab中增加了一些生僻字，如髌、跗等**
  - 8gpu，batch-size=16，max-seq-length=128
- 结果
  - **acc=0.92**



模型集成及结果



Pred = Pred1+Pred2+Pred3+Pred4+Pred5

多个模型集成acc=0.932

排序

- Logistics回归
- 特征：蕴含分数、部位相似度、术式相似度
  - 对输入query和标准名称抽取部位和术式
  - 利用BERT的向量进行相似度计算
- 开发集 acc=0.941

基于BERT的数量预测

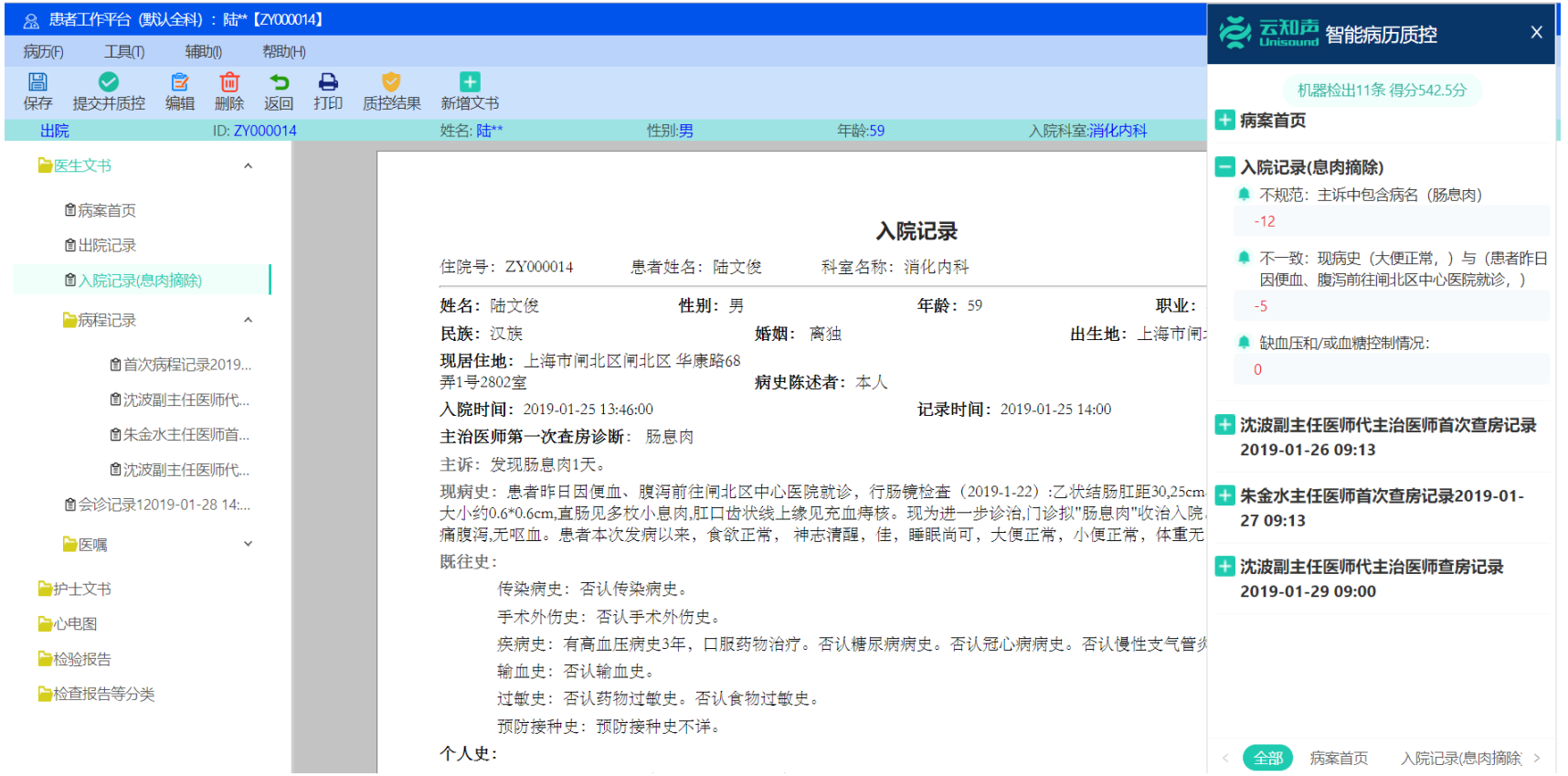
- 任务
  - BERT fine tuning 多分类任务
  - 标签为1、2、3、4
  - 对于数量超过4的按照分隔符个数确定数量
- 数据集构建
  - Step1：原始训练数据中case
  - Step2：加入icd中的带有“伴”的标准名称，数量为1
  - Step3：加入从官方训练集中拆分得到的1对2或者2对1的query
  - Step4：利用前三步得到的数据拼接数据，使得训练集标签为1、2、3、4的case数量相当
- 训练
  - 训练数据：2.8w
  - gpu=1，epoch=3，learning rate=0.00005，batch size=32，max seq length=128
- 结果
  - 开发集上数量预测**acc=0.988**；直接用简单规则acc=0.945；全部预测为1，acc=0.95

最终结果

- 盲测集acc=0.94825，排名第一

实际应用

- 工程化
  - 运行速度提升（响应时间从10分钟降低到400ms，acc=0.939）
- 产品化
  - 病案首页质控
    - 编码错误检查
  - 病历质控
    - 手术名称错误



结论

- 基于BERT蕴含分数和数量预测能够较好解决术语标准化问题
- 超多分类问题可以用BERT转换成二分类问题解决
  - 训练集尤其是负例的构造方法很重要
- BERT能够较为准确的预测标准名称的个数
- 部位和术式对手术术语标准化有所帮助

知音知医　智享医疗

招聘

病历质控