



# 基于BERT蕴含分数排序的术语标准化系统

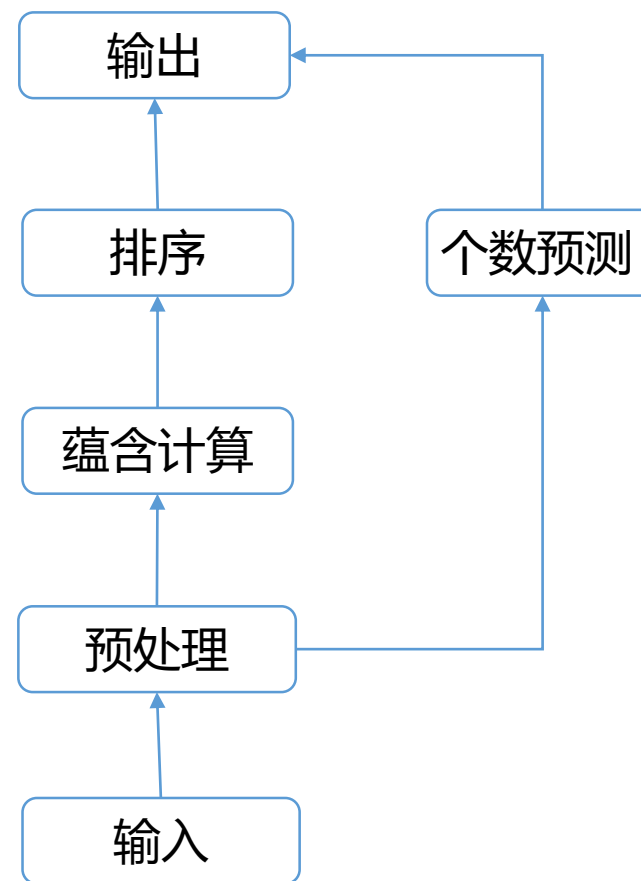
---

认知医疗组  
AI Labs  
云知声

- 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

- 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

- 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**



- 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 <del>不伴内固定</del>	手指骨折开放性复位术

## • 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:  $\langle \text{原始词}_i, \text{标准词}_i, 1 \rangle$
- 负例: 标准词集合 $R$ , 对 $\forall \text{标准词}_j \in R$ , 以及相似度阈值 $\text{thres}$

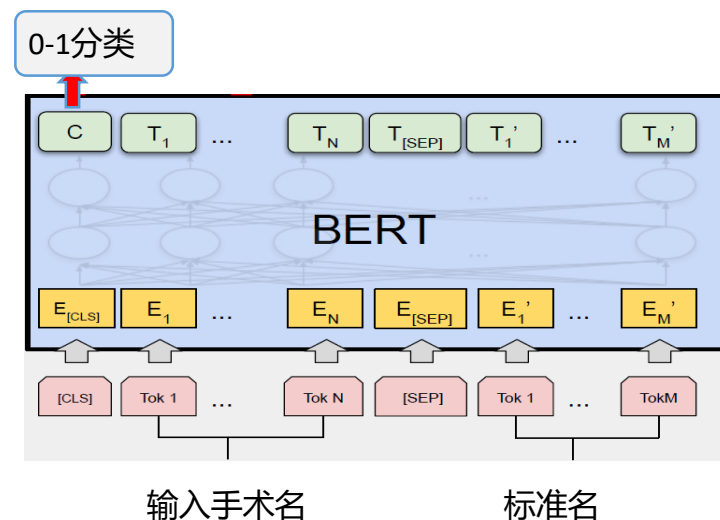
**if**  $\frac{LCS(\text{标准词}_j, \text{原始词}_i)}{\max(\text{len}(\text{原始词}_i), \text{len}(\text{标准词}_j))} \geq \text{thres} \parallel \frac{LCS(\text{标准词}_j, \text{标准词}_i)}{\max(\text{len}(\text{标准词}_j), \text{len}(\text{标准词}_i))} \geq \text{thres}$   
**then** add  $\langle \text{原始词}_i, \text{标准词}_j, 0 \rangle$

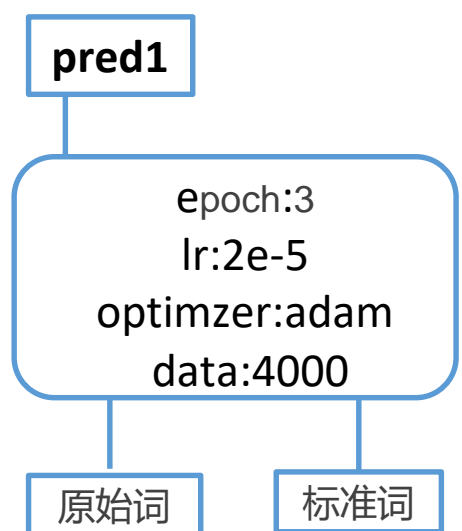
## • 训练

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

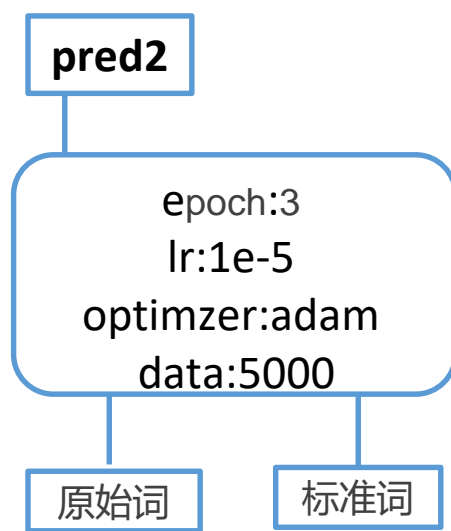
## • 结果

- acc=0.92

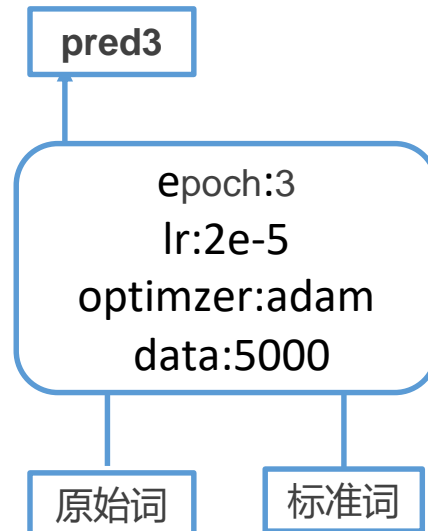




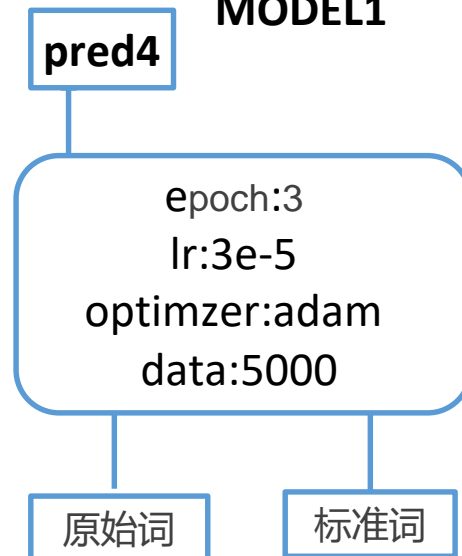
MODEL1



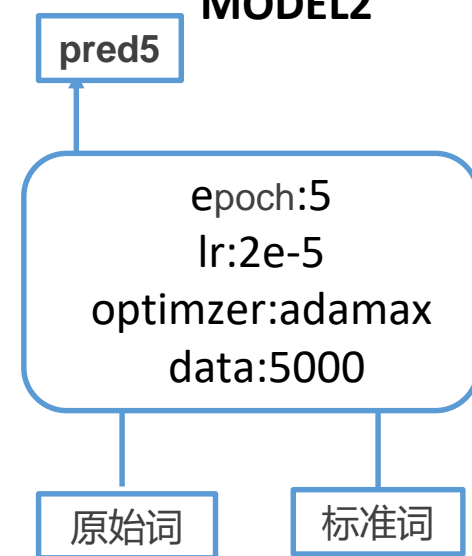
MODEL2



MODEL3



MODEL4



MODEL5

$$\text{Pred} = \text{Pred1} + \text{Pred2} + \text{Pred3} + \text{Pred4} + \text{Pred5}$$

多个模型集成acc=0.932

- 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  $\text{acc}=0.941$



- 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

- 盲测集acc=0.94825, 排名第一

- 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

患者工作平台 (默认全科) : 陆\*\* [ZY000014]

病历(F)

工具(T)

辅助(O)

帮助(H)

保存

提交并质控

编辑

删除

返回

打印

质控结果

新增文书

出院

ID: ZY000014

姓名: 陆\*\*

性别: 男

年龄: 59

入院科室: 消化内科

医生文书

病案首页

出院记录

入院记录(息肉摘除)

病程记录

首次病程记录2019...

沈波副主任医师代...

朱金水主任医师首...

沈波副主任医师代...

会诊记录2019-01-28 14:...

医嘱

护士文书

心电图

检验报告

检查报告等分类

入院记录

住院号: ZY000014

患者姓名: 陆文俊

科室名称: 消化内科

姓名: 陆文俊

性别: 男

年龄: 59

职业:

民族: 汉族

婚姻: 离独

出生地: 上海市闸...

现居住地: 上海市闸北区闸北区 华康路68弄1号2802室

病史陈述者: 本人

入院时间: 2019-01-25 13:46:00

记录时间: 2019-01-25 14:00

主治医师第一次查房诊断: 肠息肉

主诉: 发现肠息肉1天。

现病史: 患者昨日因便血、腹泻前往闸北区中心医院就诊,行肠镜检查(2019-1-22):乙状结肠距肛30,25cm大小约0.6\*0.6cm,直肠见多枚小息肉,肛口齿状线上缘见充血痔核。现为进一步诊治,门诊拟“肠息肉”收治入院。腹痛腹泻,无呕血。患者本次发病以来,食欲正常,神志清醒,佳,睡眠尚可,大便正常,小便正常,体重无

既往史:

传染病史: 否认传染病史。

手术外伤史: 否认手术外伤史。

疾病史: 有高血压病史3年,口服药物治疗。否认糖尿病病史。否认冠心病病史。否认慢性支气管炎

输血史: 否认输血史。

过敏史: 否认药物过敏史。否认食物过敏史。

预防接种史: 预防接种史不详。

个人史:

云知声 Unisound 智能病历质控

机器检出11条 得分542.5分

病案首页

入院记录(息肉摘除)

不规范: 主诉中包含病名 (肠息肉)  
-12

不一致: 现病史 (大便正常,) 与 (患者昨日因便血、腹泻前往闸北区中心医院就诊, )  
-5

缺血压和/或血糖控制情况:  
0

沈波副主任医师代主治医师首次查房记录  
2019-01-26 09:13

朱金水主任医师首次查房记录2019-01-27 09:13

沈波副主任医师代主治医师查房记录  
2019-01-29 09:00

全部 病案首页 入院记录(息肉摘除)

- 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

