平安医疗科技
PING AN HEALTHCARE TECHNOLOGY

CHIP

基于BERT与提升树模型的语义匹配方法

团　队：wzm

汇报人：吴梓明

团队简介

吴梓明　华南理工大学　研究生三年级

- 贵在联通——“联创黔线”杯大数据应用创新大赛 1st
- 蚂蚁金服风险大脑：支付风险识别(内部赛) 1st
- DigSci 科学数据挖掘大赛 2019 3rd
- 同盾科技声纹识别建模大赛 4th
- 第三届融360天机智能金融算法挑战赛 5th
- 智源&计算所-互联网虚假新闻检测挑战赛 7th
- 美年健康AI大赛——双高疾病风险预测 9th

任务说明

任务描述

本次评测任务的主要目标是针对中文的疾病问答数据，进行病种间的迁移学习。具体而言，给定来自5个不同病种的问句对，要求判定两个句子语义是否相同或者相近。所有语料来自互联网上患者真实的问题，并经过了筛选和人工的意图匹配标注。

评价指标

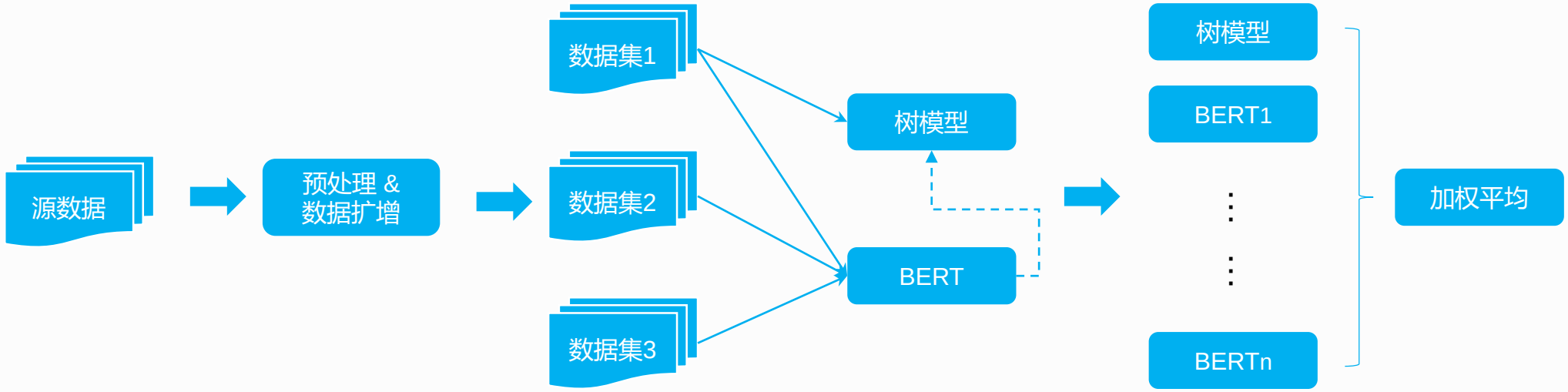
本任务的评价指标包括准确率(Precision)，召回率(Recall)和F1值。最终排名以F1值为基准。

$$precision = \frac{TP}{TP+FP}$$

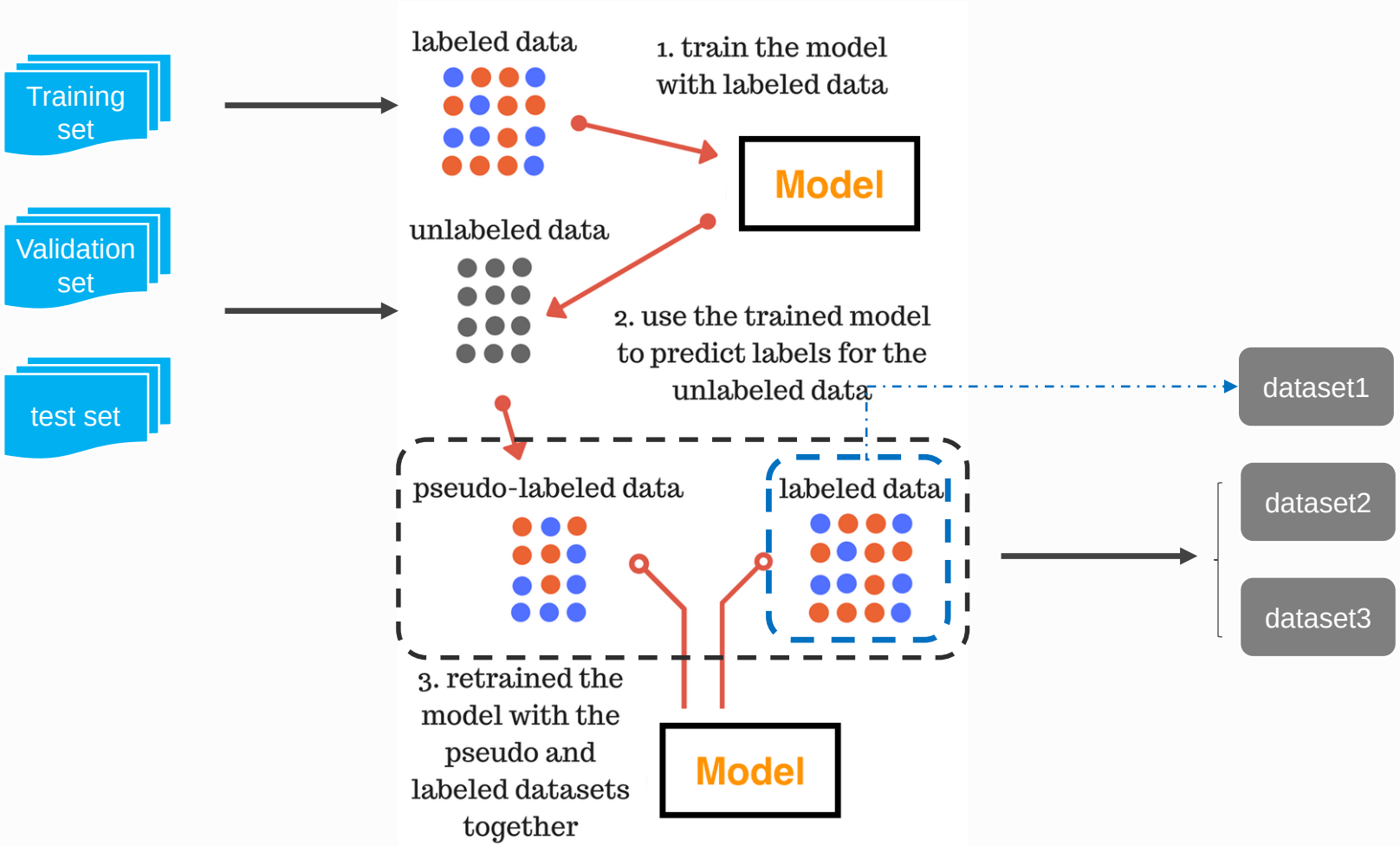
$$recall = \frac{TP}{TP+FN}$$

$$F1 = \frac{2*precision*recall}{precision+recall}$$

整体方案设计



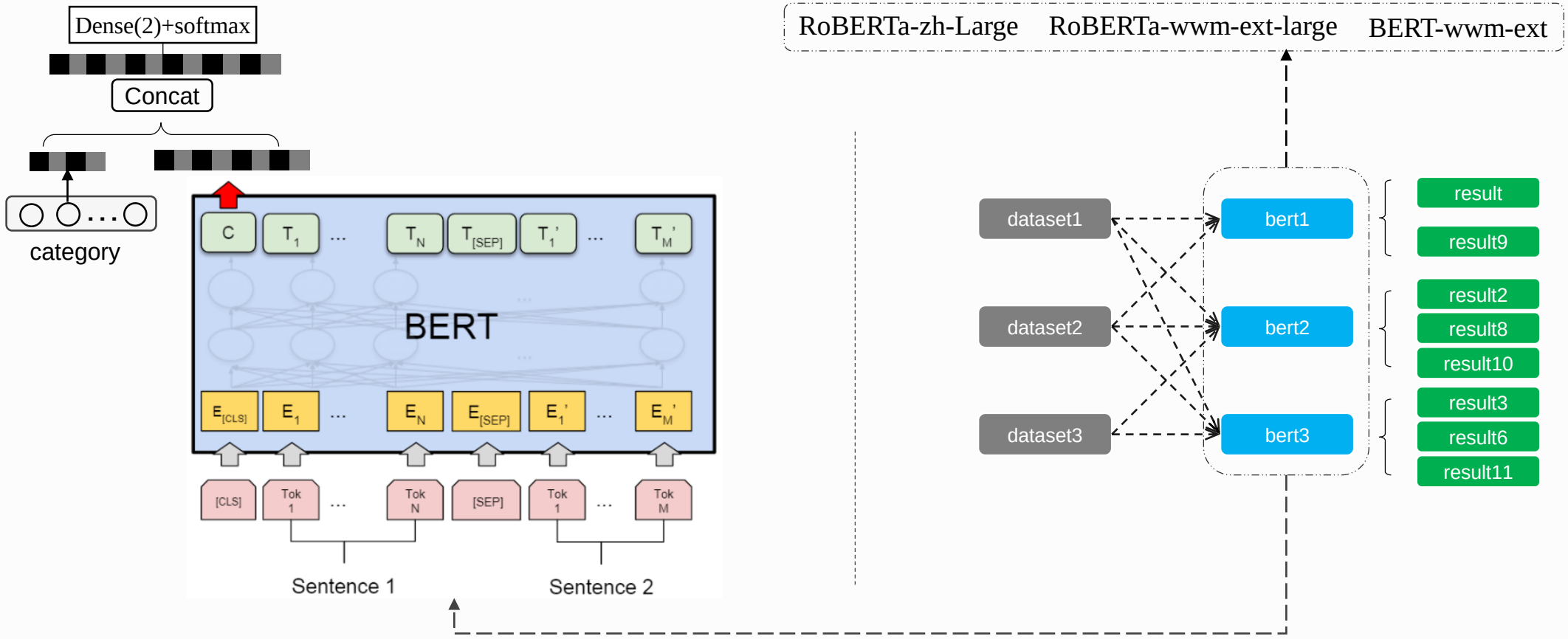
数据扩增



来源：

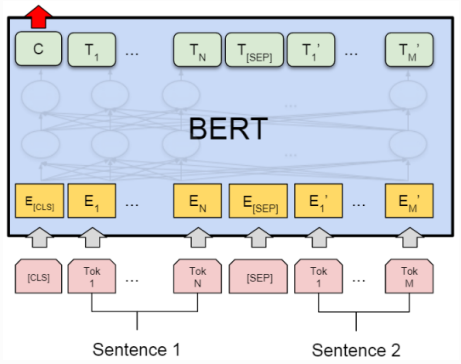
[1] SHUBHAM J. Introduction to Pseudo-Labelling: A Semi-Supervised learning technique. Analytics Vidhya[J]. 2017.

BERT模型

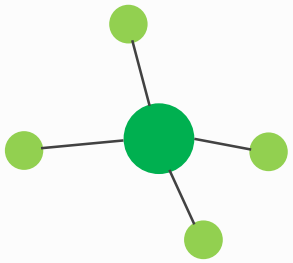


[1] Liu Y, Ott M, Goyal N, et al. Roberta: A robustly optimized bert pretraining approach[J]. arXiv preprint arXiv:1907.11692, 2019.
[2] Cui Y, Che W, Liu T, et al. Pre-Training with Whole Word Masking for Chinese BERT[J]. arXiv preprint arXiv:1906.08101, 2019.

树模型



sentence vector



graph features

sentecne1 vs sentecne2

interaction features

XGBoost/ LightGBM

主要特征

类别	特征	说明
隐式表征	pool_nsp_x	经过bert之后的隐式表征
图特征	out_degree	出度
	in_degree	入度
	degree	出度+入度
	pr	pagerank值
交互特征	distance	编辑距离
	ratio	莱文斯坦比
	jaro	jaro距离
	jaro_winkler	Jaro–Winkler距离

为什么有效

示例

	句子1	句子2
1	高血压症的患者又时而出现低血压是什么原因	以前是高血压现在又是低血压，是什么原因
2	请问我血糖查出来是15.4，我是不是有糖尿病。？	我的尿糖是2个加号，是不是糖尿病

- BERT模型学习出来的概率在0.45-0.55之间，不确定性非常大，
  但是其实直观上的观察就能发现，编辑距离较小，两者极有可能相关。

- 同时，传统模型存在缺点，非常容易过拟合，对数据集依赖大，举例：假如训练集中大量label的句子对，都出现了“血糖”这个词语，那么很有可能学习出来的模型就会认为只要句子中含有“血糖”，那么标签就极有可能为1，而偏离了我们需要考虑的语义特性，因此也不能仅使用树模型，同时在本方案中避免了使用类似tfidf等词频特征，就是为了避免模型出现上述的过拟合现象。

还有什么能做?

1. 数据扩增：传递性，对称性…
2. TTA（test time augmentation）
3. Focal Loss
4. 更细致的特征工程

平安医疗科技
PING AN HEALTHCARE TECHNOLOGY

THANKS

团　队：wzm

汇报人：吴梓明