

专题三：实体识别挑战赛 进阶串讲

导师：Gauss

目录

1/ 数据增强

2/ 标签不平衡问题

3/ 线上线下不一致问题

4/ 模型融合

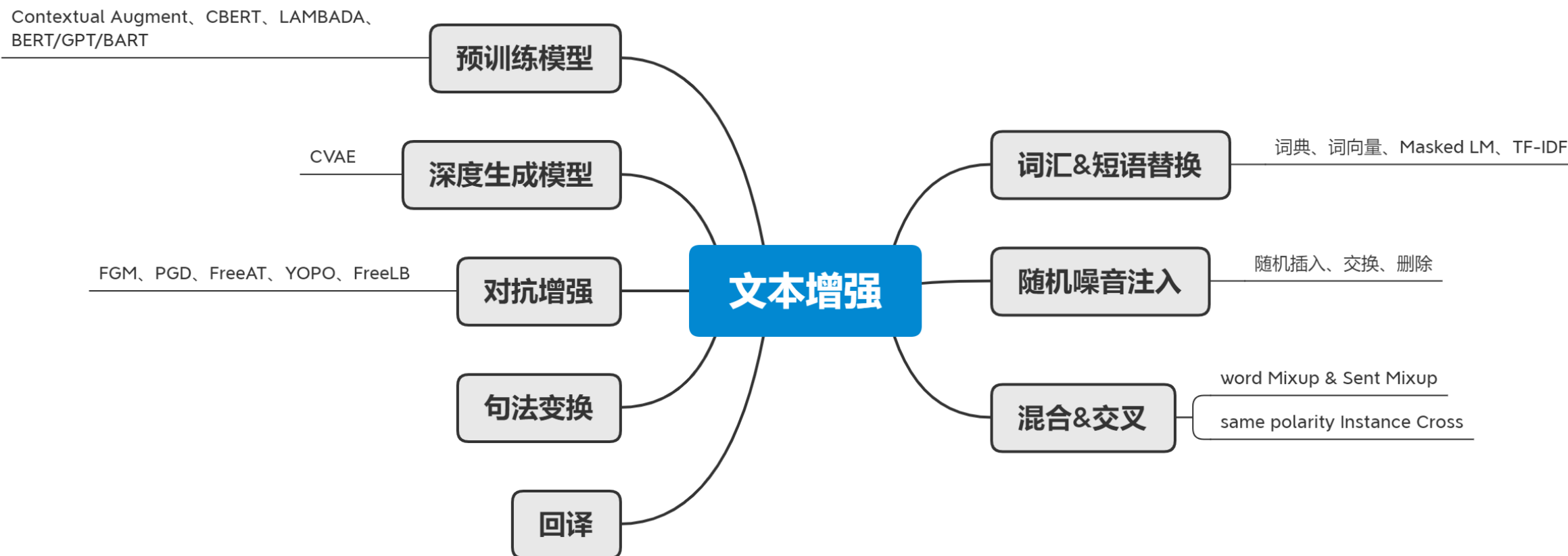
1、数据增强

文本增强

什么才是一种好的解决少样本困境的方案？

- 文本增强
- 弱监督学习（半监督学习）

文本增强



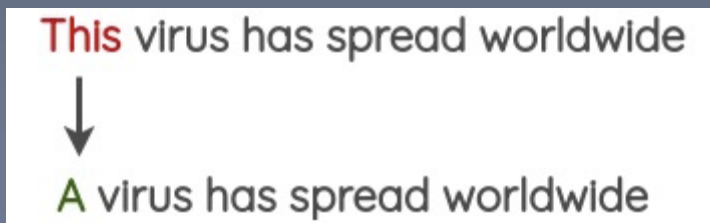
文本增强

词汇&短语替换

- 基于词典：主要从文本中选择词汇或短语进行同义词替换，词典可以采取WordNet或哈工大词林等。著名的EDA(Easy Data Augmentation)就采用了这种方法。
- 基于词向量：在嵌入空间中找寻相邻词汇进行替换，我们所熟知的TinyBERT (TinyBERT: Distilling BERT for Natural Language Understanding)就利用这种技术进行了数据增强。
- Masked LM：借鉴预训练语言模型（如BERT）中的自编码语言模型，可以启发式地Mask词汇并进行预测替换。

文本增强

- TF-IDF: 实质上是一种非核心词替换, 对那些low TF-IDF scores进行替换, 这一方法最早由Google的UDA提出:



This virus has spread worldwide
↓
A virus has spread worldwide

- 随机噪音注入
 - 随机插入: 随机插入一个词汇、相应的拼写错误、占位符等, UDA[4]则根据Uni-gram词频分布进行了采样。
 - 随机交换: 随机交换词汇或交换shuffle句子。
 - 随机删除: 随机删除(drop)词汇或句子。

文本增强

- 回译：基于机器翻译技术，例如从中文-英文-日文-中文；我们熟知的机器阅读理解模型QANet和UDA都采用了回译技术进行数据增强。
- 句法交换：通过句法树对文本句子进行解析，并利用相关规则进行转换，例如将主动式变成被动式句子。
- 对抗增强：不同于CV领域利用GAN生成对抗进行数据增强，NLP中通常在词向量上添加扰动并进行对抗训练，文献[10]NLP中的对抗训练方法FGM, PGD, FreeAT, YOPO, FreeLB等进行了总结。

文本增强

深度生成模型

预训练语言模型 (CBERT、LAMBADA)



半监督学习

监督学习往往需要大量的标注数据，而标注数据的成本比较高，因此如何利用大量的无标注数据来提高监督学习的效果，具有十分重要的意义。这种利用少量标注数据和大量无标注数据进行学习的方式称为半监督学习(Semi-Supervised Learning, SSL)。

SSL如何在少量标注样本下达到或超越大量标注样本下监督学习的效果，SSL如何在大量标注样本下也不会陷入到“过拟合陷阱”，是SSL研究者面临的一个挑战。

半监督学习

参考资料：

向右老师的讲解

<https://zhuanlan.zhihu.com/p/146777068>

NER数据增强

句子组合

'补气养血、调经止带，用于月经不调、经期腹痛非处方药物（甲类），国家基本药物目录（2012）如果服用任何其他药品请告知医师或药师，包括任何从药房、超市或保健品商店购买的非处方药品。本药内所含人参、白芍，反藜芦，忌与含藜芦的药物同用。本药内所含甘草，反甘遂、大戟、海藻、芫花，忌与含甘遂、大戟、海藻、芫花的药物同用。服药期间避免与生冷、辛辣、荤腥油腻、不易消化食品同用，戒烟酒，以防助湿化热，加重病情。服药期间不宜喝茶和吃萝卜，不宜同时服用五灵脂、皂荚或其制剂。'

'补气养血、调经止带，用于月经不调、经期腹痛非处方药物（甲类），国家基本药物目录（2012）如果服用任何其他药品请告知医师或药师，包括任何从药房、超市或保健品商店购买的非处方药品。本药内所含人参、白芍，反藜芦，忌与含藜芦的药物同用。5,高血压,糖尿病患者或正在接受其他药物治疗的患者应在医师指导下服用。服药期间避免与生冷、辛辣、荤腥油腻、不易消化食品同用，戒烟酒，以防助湿化热，加重病情。服药期间不宜喝茶和吃萝卜，不宜同时服用五灵脂、皂荚或其制剂。'



NER数据增强

实体词替换：

'补气养血、调经止带，用于月经不调、经期腹痛非处方药物（甲类），国家基本药物目录（2012）如果服用任何其他药品请告知医师或药师，包括任何从药房、超市或保健品商店购买的非处方药品。本药内所含人参、白芍，反藜芦，忌与含藜芦的药物同用。本药内所含甘草，反甘遂、大戟、海藻、芫花，忌与含甘遂、大戟、海藻、芫花的药物同用。服药期间避免与生冷、辛辣、荤腥油腻、不易消化食品同用，戒烟酒，以防助湿化热，加重病情。服药期间不宜喝茶和吃萝卜，不宜同时服用五灵脂、皂荚或其制剂。'

'理气、调经止带，用于大便溏薄形寒肢冷、乳腺胀痛非处方药物（甲类），国家基本药物目录（2012）如果服用任何其他药品请告知医师或药师，包括任何从药房、超市或保健品商店购买的非处方药品。本药内所含知母地黄、五灵脂，山茱萸，忌与黄柏的药物同用。本药内所含当归，甘草、大戟、芫花、苦参，忌与含大戟、枯矾、五灵脂、含藜芦的药物同用。服药期间避免与油腻、油腻、油腻难消化油腻、辛辣酸食品同用，戒烟酒，以防助湿化热，加重病情。服药期间不宜喝茶和吃萝卜，不宜同时服用五灵脂、皂荚或其制剂'

2、标签不平衡问题

标签不平衡

通过对于标签类型数据分析，类别严重不平衡，长尾分布很明显。

如何解决标签不平衡问题？

SYMPTOM	6090
DRUG_EFFICACY	3257
PERSON_GROUP	1718
SYNDROME	1206
DRUG_TASTE	1133
DISEASE	1104
DRUG_DOSAGE	1016
DRUG_INGREDIENT	728
FOOD_GROUP	641
DISEASE_GROUP	623
DRUG	156
FOOD	71
DRUG_GROUP	14

标签不平衡

学术角度解决问题：

NER in Long-tailed Datasets

- 优化损失函数
- 辅助任务联合训练

标签不平衡

比赛中如何快速解决问题？

数据增强：

常规的欠采样和过采样方法在序列标注任务上，稍微思考一下就是不靠谱的。在这个任务上如果直接在样本上操作，不能欠采样，导致样本多的类别也会表现不好。

少实体类别替换，丰富少数实体类别样本。具体讲，将少数实体类别的实体随机替换为同类别的其他实体。

标签不平衡

比赛中如何快速解决问题？

损失函数：

Focal loss, 由何恺明在论文《Focal Loss for Dense Object Detection》提出, 最初用于解决目标检测下的样本不均衡问题。

Dice loss, 由VNET提出, 香侬科技的2020ACL《Dice Loss for Data-imbalanced NLP Tasks》对原Dice Loss进行了改动, 进而解决NLP中样本不均衡问题。

3、线上线下不一致问题

Bad case分析

这应该是一个比较普适的处理问题的方法，确认目标，剖析问题，处理问题，检验结果。

首先，要分析这bad case，**知道bad case产生的原因**，只有知道病因才能对症下药。

确定这个case是否需要解决，即评估这个case的影响面，有多少相似的case存在，解决后收益有多大，毕竟我们需要把资源花在最有收益的地方。

提出解决方案并进行试验。

校验case是否解决，解决程度如何（其实有时候能解决一个问题的80%已经很不错了，不见得要完全搞定）

校验，这个解决方案有没有引入新的问题（一般要做回归测试）。

Bad case分析

代码中演示

线上线下一致

验证集指标和线上指标不一致问题：

- 验证集标签缺失严重
- 验证集和线上识别难度不一致

线下过高，线上过低：

- 过拟合
- 训练集和验证集部分重合

4、模型融合

模型融合

主要分为两大块：

- 概率融合
- 实体投票

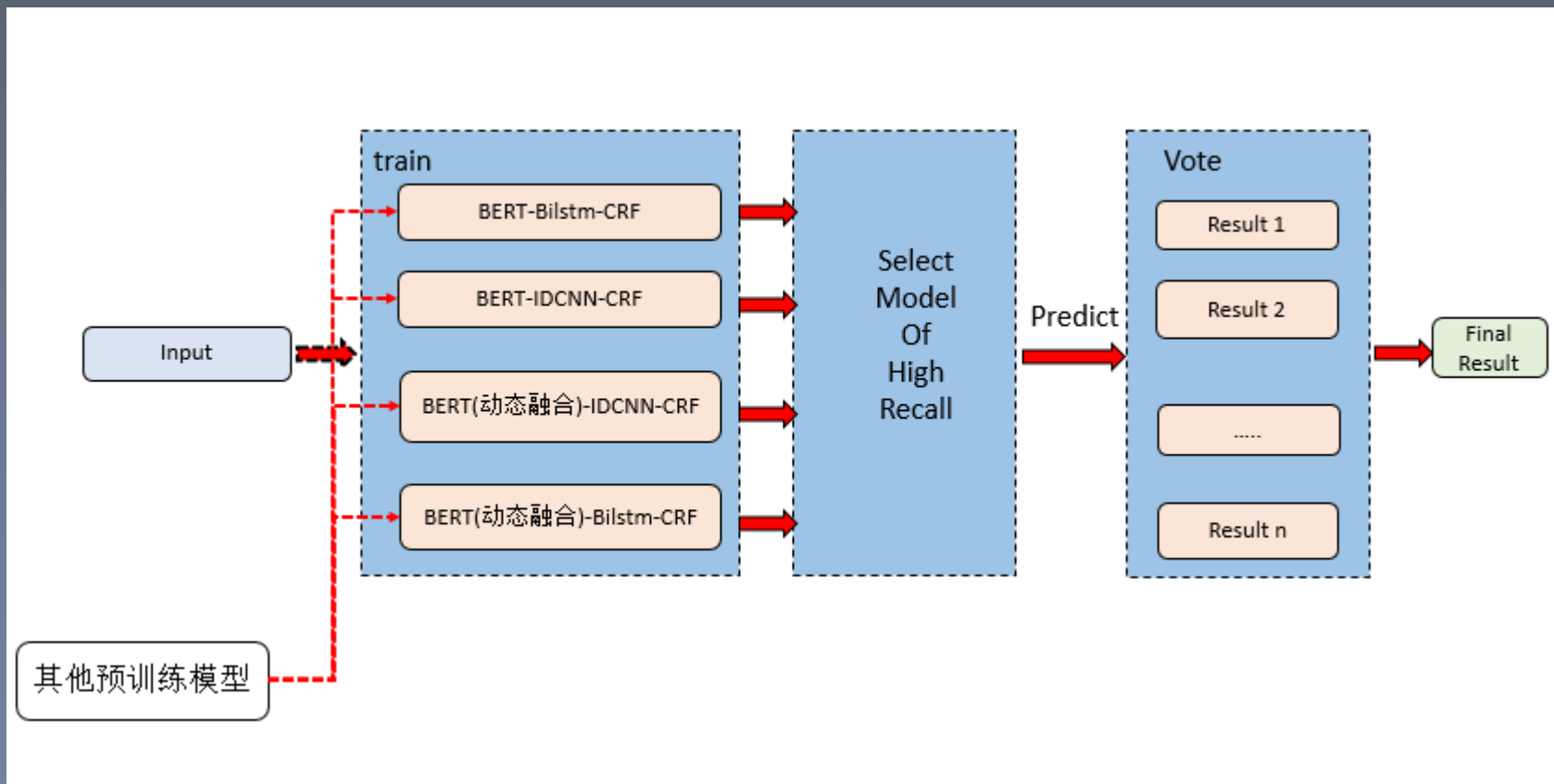
模型融合

概率融合



模型融合

结果投票：比如可以通过多个模型的结果进行投票。



——结 语——

感谢大家参加本次的比赛直播

课下请一定要

亲自动手操作一次！





deepshare.net

深度之眼

联系我们：

电话：18001992849

邮箱：service@deepshare.net

Q Q：2677693114



公众号



客服微信

