
CCF BDCI
CCF BIG DATA & COMPUTING INTELLIGENCE CONTEST
2020 8th
CCF 大数据与计算智能大赛
非结构化商业文本信息中隐私信息识别
队伍名称：犹豫就会败北

CCF BIG DATA & COMPUTING INTELLIGENCE CONTEST

目录

团队介绍

陈智垚（哈尔滨工业大学）

- 基于 Adversarial Attack 的问题等价性判别比赛（2019）第 13 名
- 中国大学生保险数字科技挑战赛华南赛区（2020）第 2 名

黎洋（中国平安数据挖掘算法工程师）

- 热衷于将机器学习算法应用于金融场景
- 曾在CCF、CCKS、ICDM等多个赛事取得TOP

傅勇昊（哈尔滨工业大学）

- CCKS事件抽取评测（2020）第 2 名

聂才（哈尔滨工业大学）

- BDCI 大数据时代的Serverless工作负载预测（2020）第 8 名

田嘉豪（哈尔滨工业大学）

- 基于 Adversarial Attack 的问题等价性判别比赛（2019）第 13 名

参赛历程

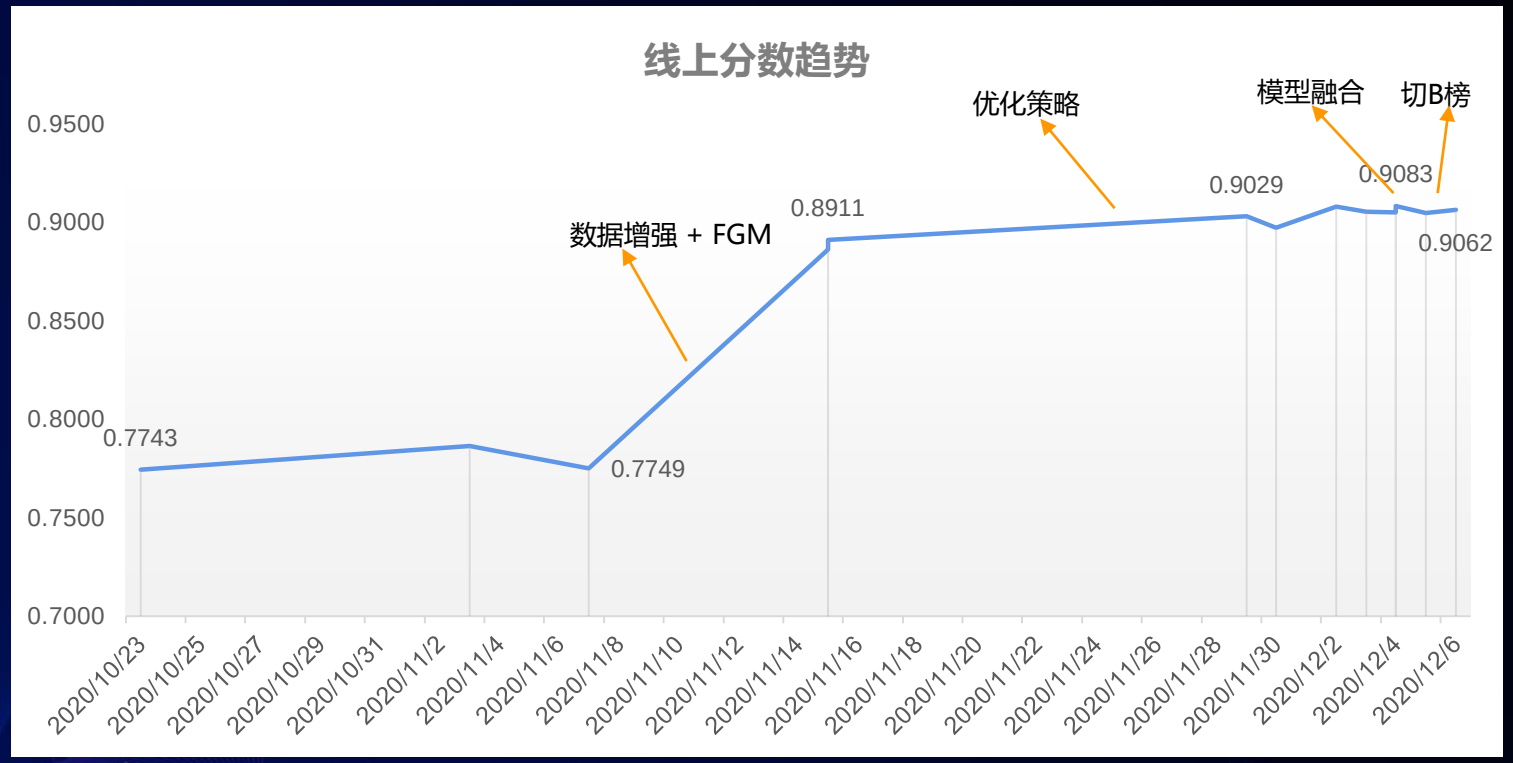

表1 线上排名结果

赛程	排名	成绩	备注
A榜最终评测	5	0.9083097	存在test数据未预测
B榜自动评测	5	0.904476	存在test数据未预测
B榜最终评测	2	0.9062074	最终得分

赛题描述与分析

CCF BIG DATA & COMPUTING INTELLIGENCE CONTEST

赛题背景：随着社交网络、移动通讯等技术的迅速发展，网络中存在大量包含隐私数据的文本信息，如何在非结构化的本文信息中精准识别隐私数据并对其进行保护已经成为隐私保护领域中亟需解决的问题。

赛题目的： 本赛题将关注点集中在隐私属性的识别问题中，针对非结构化的本文信息进行分析，对中所涉及到的隐私信息精准提取。

数据描述：我们将这个赛题建模一个序列标注任务：

评价指标：Micro F1

赛题描述与分析

数据分析

- 由图1可知，赛题实体类型分布不均，少样本实体类型在后续建模中可能受限于数据大小，影响模型性能。
- 图2反映出实体的长度分布基本集中在1-10的长度区间，10以上长度的分布呈一个长尾状态，根据我们团队对序列标注问题的经验，在后续过程中有必要针对实体长尾分布问题进行建模优化。
- 由图2可知，82%的文本长度在200以内；最大的文本长度超过1000；以512作为文本最长限制，可达到98.5%的覆盖率。

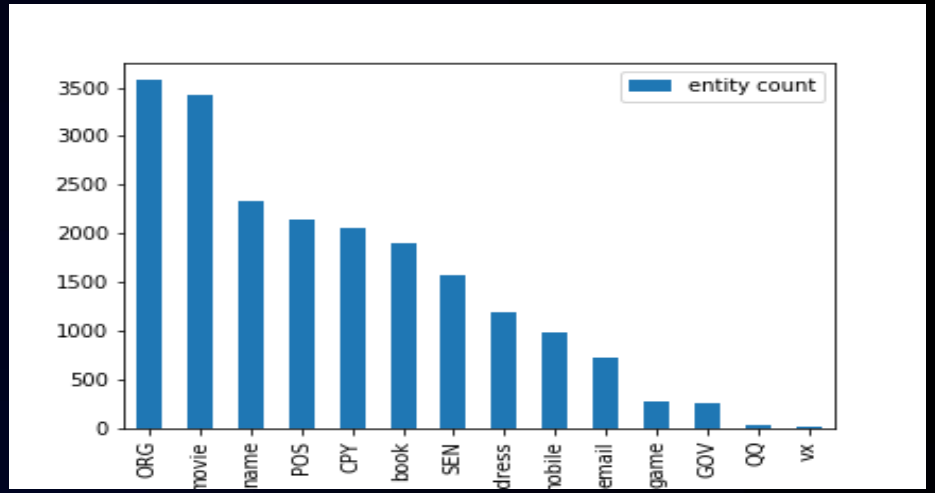

图1 实体类型分布

难点总结

- 数据存在实体类别不平衡情况；
- 实体长度呈现长尾分布
- 文本数目只有2515条，样本较少；
- 长文本中语义不连贯较为明显，上下文语境存在明显的断裂；
- 个别类别数据存在标注标准不统一及漏标的情况。

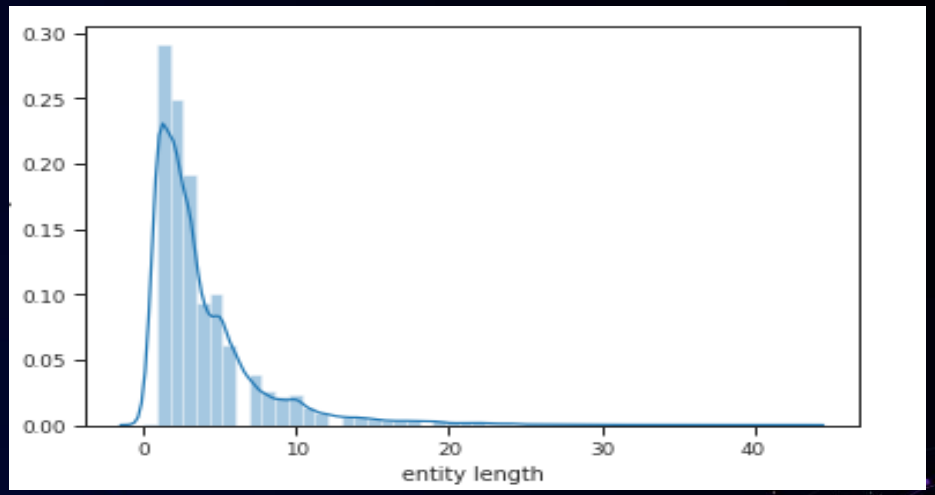

图2 实体长度分布

算法模型-数据处理

- 采用动态滑动窗口策略做数据增强，解决部分样本长度过长的问题

划分策略过程

1） 在滑动窗口过程中以标点符号优先级对句子进行切割和重组，尽可能保留上下文信息；

2）当重组句子长度超过阈值时，则生成一条新的子数据；

3） 阈值根据不同模型对于文本长度的敏感性有所区分；

4） 数据动态滑窗，解决BERT数据过长的问题和充分挖掘数据信息。

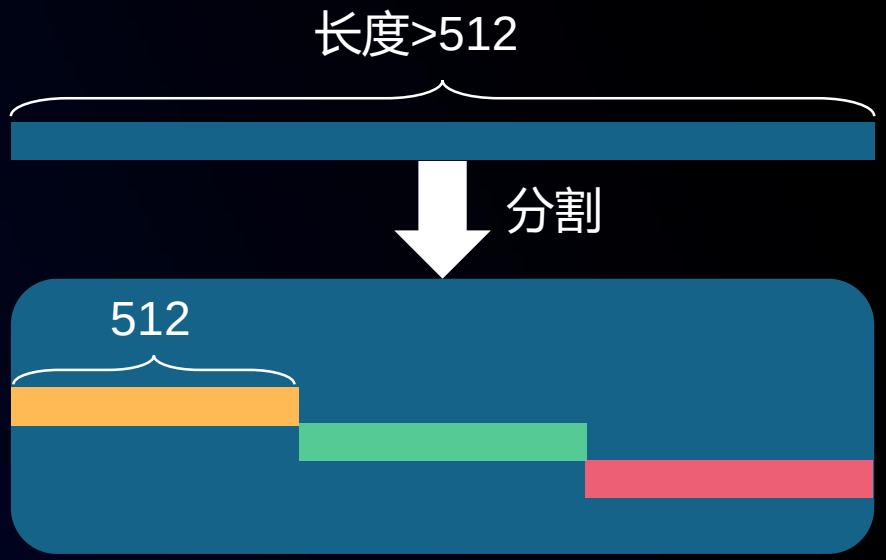

图3 数据增强

- 使用外部数据Cluener进行数据增强，提高模型的泛化性
- 使用BIOS标注法标注实体

标注过程：

1）每一类实体都对应各自独立的 B、I 起止标签；

2）对于长度为1的实体另设S标签；

3）非实体设置O标签。

在我们的后续实验中证明，BIOS标签比BIO标签有一定的提升

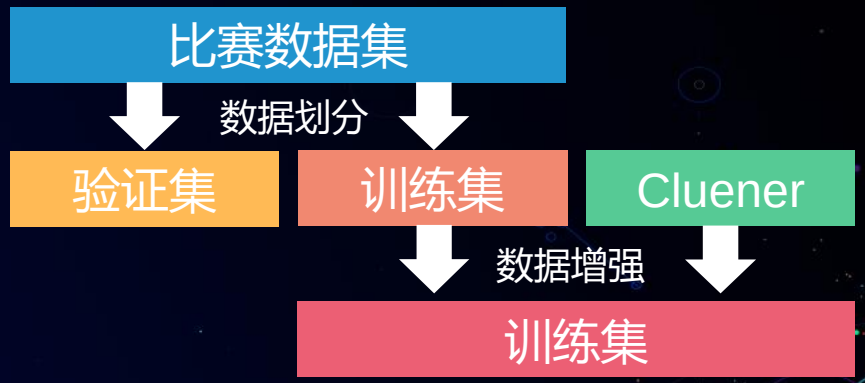

图4 数据划分

算法模型-模型构建

BERT-CRF模型

- 建模思路
  - 我们团队在Bert后接一层CRF，对实体标签预测加以标签转移约束，以解决下层模型仅能捕捉上下文特征的问题，确保最终的预测结果准确、有效。
- 模型优势
  - 训练速度快，对短实体预测效果好

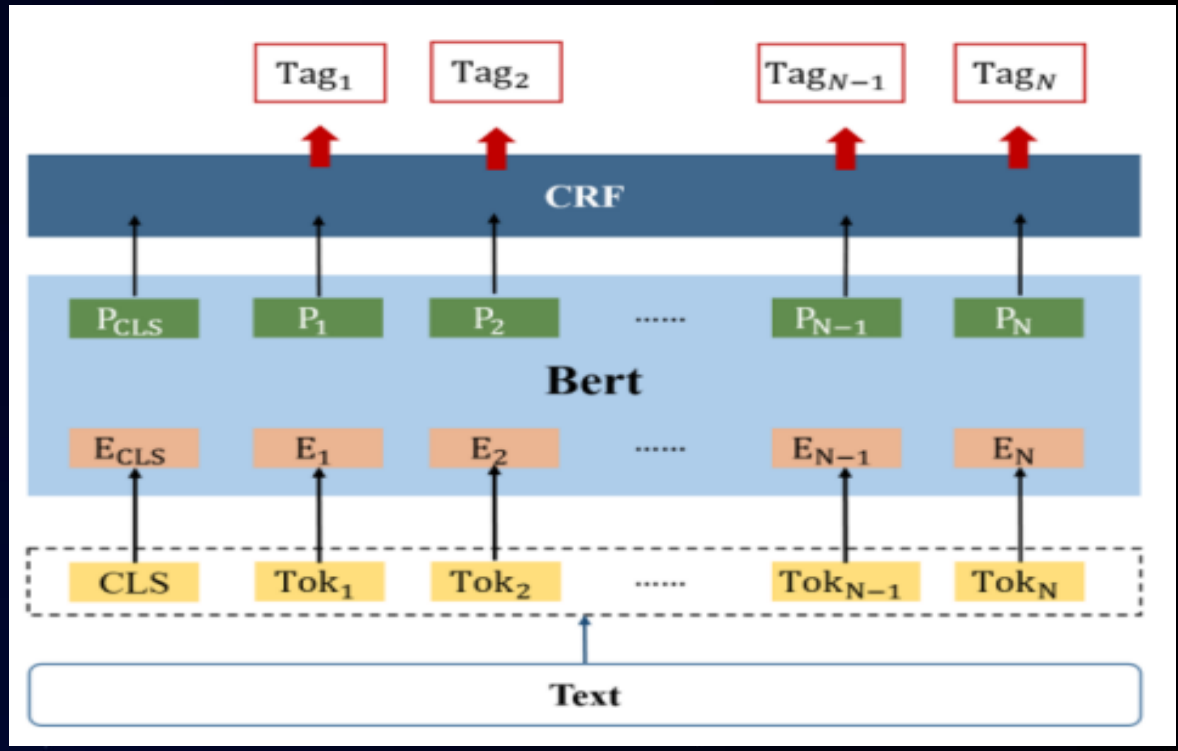

图5 BERT-CEF原理图

算法模型-模型构建

BERT-SPAN模型

- 建模思路
  - 借鉴于MRC式的指针网络模型和赛题对于多片段实体抽取的需求，我们构建了图6所示模型；
  - 该模型将多片段抽取问题转化为N个K分类问题（N为序列长度，K为实体类型数目）；
  - 通过后续实体边界判定模块，实现对实体的起始边界的判定，从而完成实体的抽取。
- 模型优势
  - BERT-SAPN模型可以解决上文提及的实体长尾分布问题。针对例如position这类长尾实体，抽取效果好。

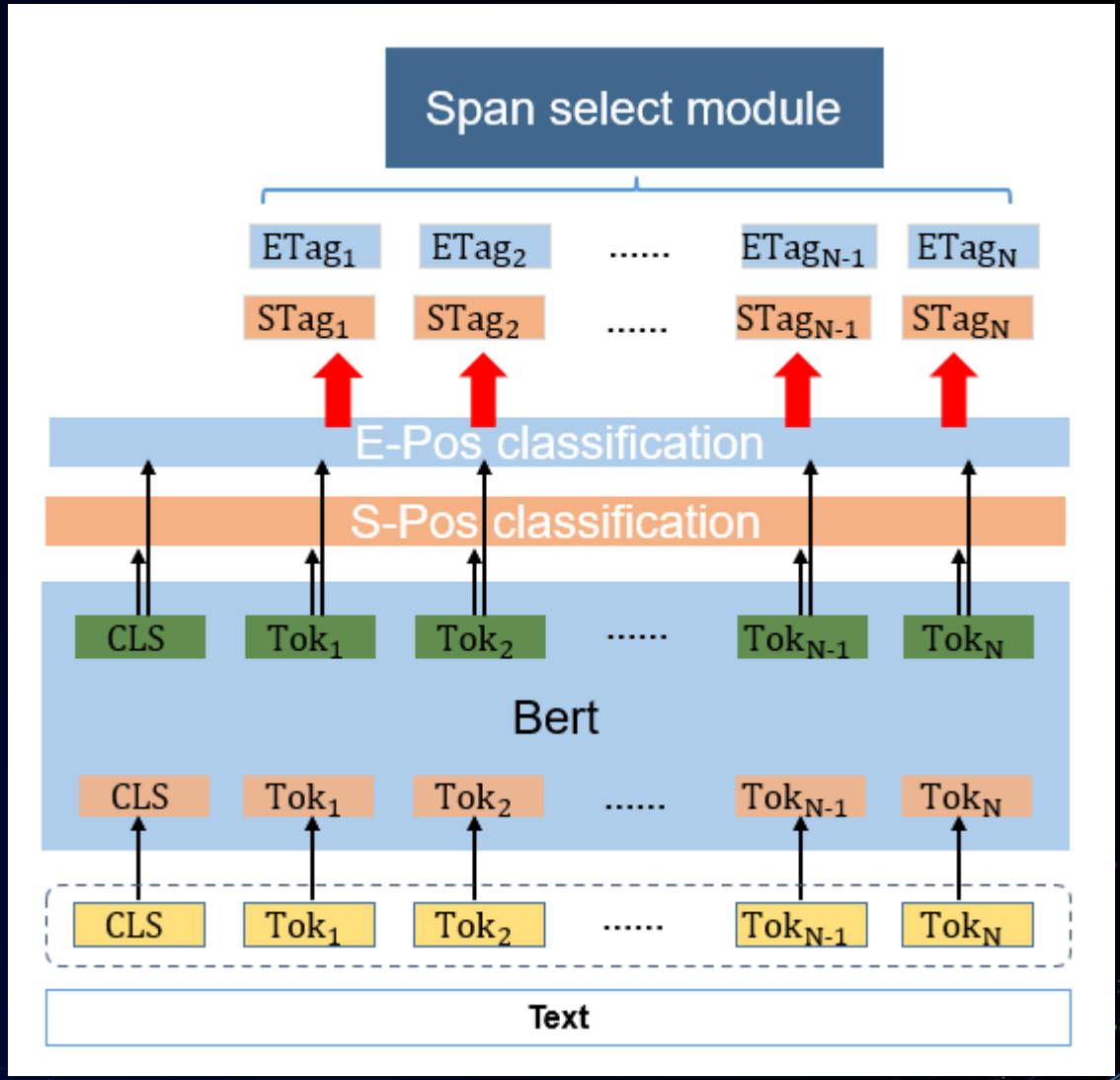

图6 BERT-SPAN原理图

算法模型-优化策略

K折交叉验证

- 多次划分数据集来构造服从同分布的训练集和验证集，生成的结果进行投票预测最为最终的结果；
- 减少模型预测的方差，减少线上和线下不同分布导致的训练偏差，从而提高模型的预测稳定性；

使用后效果：模型表现提升约0.1%-0.18%

学习率调整策略

- CRF学习率调节

设置其学习率为BERT的10-100倍，让CRF层学习率更大一些，使得CRF层能快速学习；

- 分层学习率
- Warmup + AdamW + LR线性调节

1）Warmup：减小模型训练初期由于权重随机初始化带来的不稳定（振荡），使得模型收敛速度变得更快；

2）LR线性调节：在训练后期调节学习率至一个较小的值，减少训练后期的振荡现象，从而找到相对更优的解。

对抗学习

- 使用FGM对抗训练方法，对下层的预训练模型的Embedding层进行对抗训练，提升模型泛化性。

使用后效果：模型表现提升约5个千分位点

算法模型-模型融合

投票策略

- 不同的解码方式
- 不同的预训练模型
- K折交叉验证
- 简单投票机制

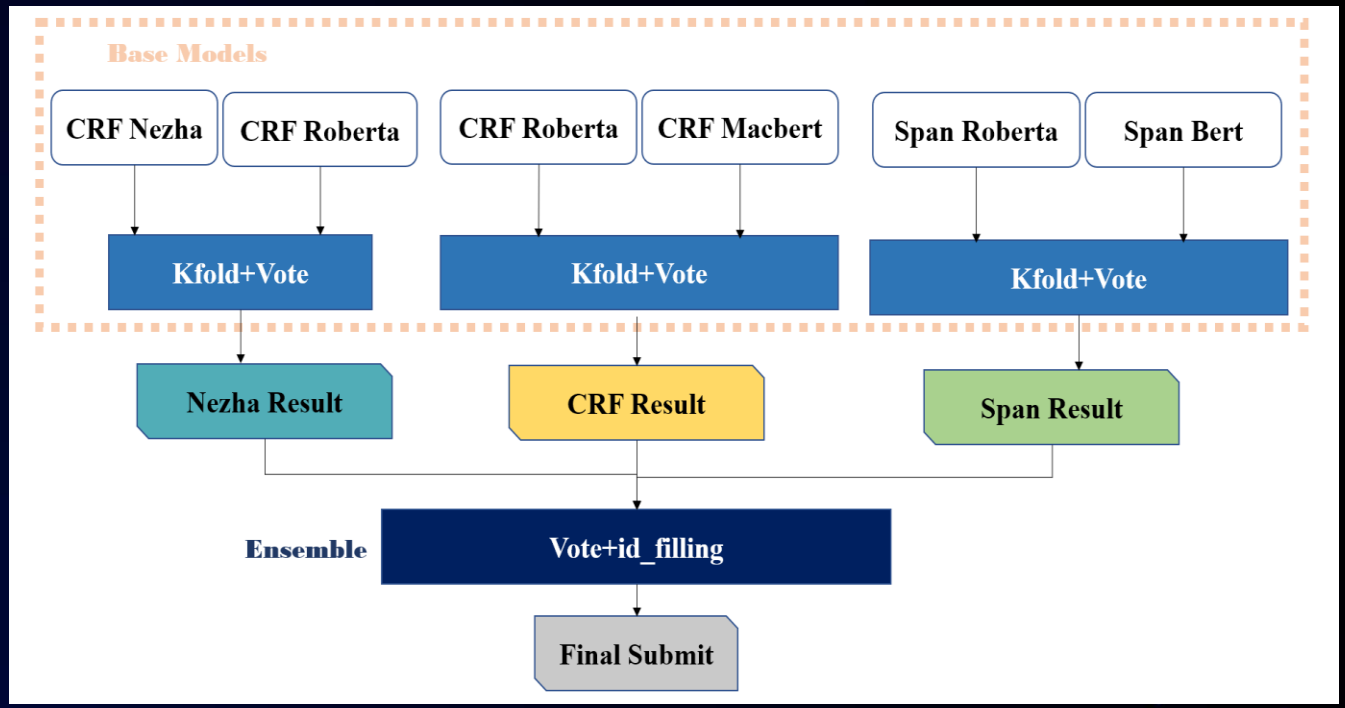

图7 模型融合框架图

算法模型-后处理策略

缺失实体填充策略

- 如果一个测试样例没有被模型识别到任何实体，就把BERT-CRF模型中的高置信度实体填充到最终结果里面；
- 可以有效缓解缺失实体的问题。

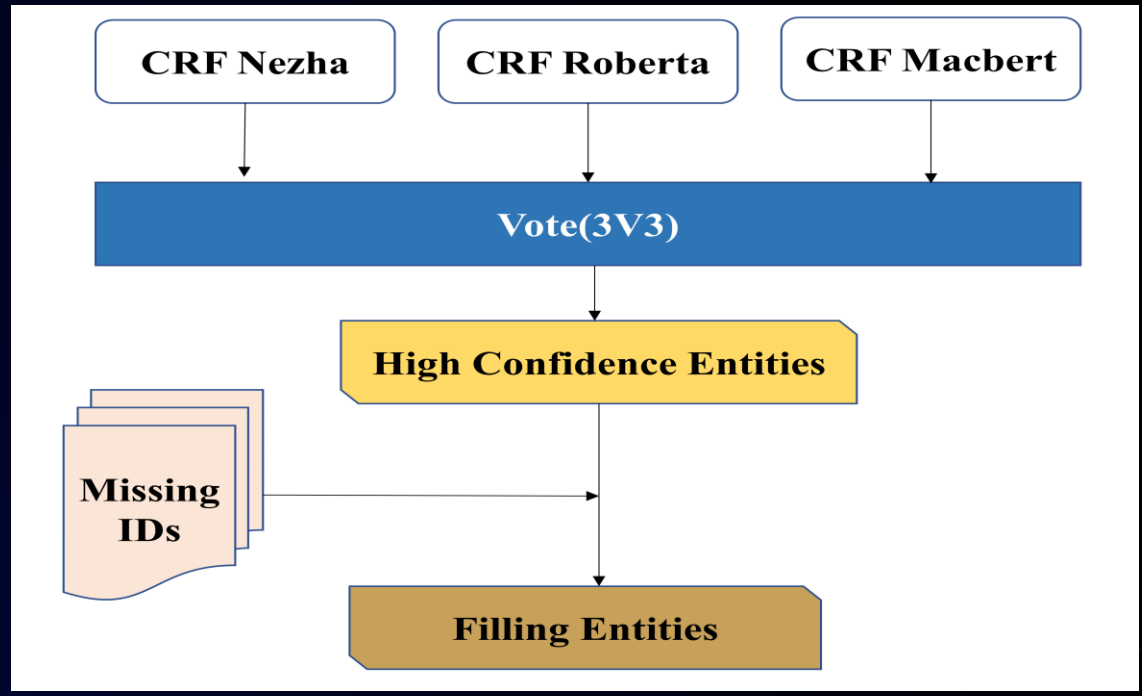

图8 缺失实体填充

算法模型-模型效果

表2 不同模型对比

模型	A榜成绩	B榜成绩
BERT-CRF	0.78587079	\
BERT-SPAN	0.77465582	\
BERT-BILSTM-CRF+数据增广	0.86322454	\
BERT-CNN-CRF+数据增广	0.86284245	\
BERT-CRF+数据增广	0.88767766	\
BERT-CRF+数据增广+ FGM	0.89173406	\
BERT-CRF+数据增广+ FGM +学习率策略+后处理	0.90160871	\
BERT-SPAN+数据增广	0.89114185	\
BERT-SPAN+数据增广+ FGM	0.89707029	\
BERT-SPAN+数据增广+ FGM +学习率策略+后处理	0.90292764	\
BERT-CRF融合	0.90384513	\
BERT-SPAN融合	0.90584513	\
异构模型融合(BERT-CRF+BERT-SPAN)	0.9083097	0.90620744

方案潜力与应用价值

方案潜力

- 综合数据处理、模型框架、后处理策略三者权衡，模型精度高、性能稳定和泛化性强；
- 采用BERT方案，单模就能到达分数0.902，完全满足工业界的精度要求；
- 采用工业界代码规范，可复用性高，易于训练方便部署，满足工业界落地要求；

应用价值

- 使用知识蒸馏等模型裁减技术，减少模型中的参数量和模型结构复杂度，可满足工业界高精度、高效率等要求；
- 满足通用的NER解决方案，技术可支持知识图谱、智能客服和知识库等场景建设。

总结

制胜要诀

- 数据：详细的分析和难点分析、数据增强
- 模型：序列标注模型框架、指针标注模型框架、优化策略（学习率调整策略、对抗学习、融合策略和后处理策略等）
- 性能：使用外部数据，单模训练时间大约3h，精度损失不到1%

其他尝试

- **BERT-LSTM-CRF、BERT-CNN-CRF**

相比于同类型的BERT-CRF，效果有所降低，同时模型训练时间增加了20%

- **Focal loss损失函数**
- 使用Focal Loss对少量失衡实体有提升，整体提升不明显

模型落地 = 数据 + 模型 + 性能

THANKS