非结构化商业文本信息中隐私信息识别

犹豫就会败北

陈智垚
计算机研二
哈尔滨工业大学(深圳)
中国-深圳
zhiyao_chen@163.com

黎洋
算法工程师
中国平安
中国-深圳
yangli_i1@163.com

聂才
计算机研二
哈尔滨工业大学(深圳)
中国-深圳
nie-cai@qq.com

田嘉豪
计算机研二
哈尔滨工业大学(深圳)
中国-深圳
jiahao_tian@outlook.com

傅勇昊
计算机研二
哈尔滨工业大学(深圳)
中国-深圳
2871375215@qq.com

团队简介

犹豫就会败北团队成员分别为来自中国平安的黎洋，以及来自哈尔滨工业大学的陈智垚，傅勇昊，田嘉豪，聂才。

团队成员傅勇昊曾在 2020CCKS 事件抽取评测取得第 2 名；黎洋现担任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；陈智垚和田嘉豪曾在 2019 基于 Adversarial Attack 的问题等价性判别比赛取得第 13 名，此外陈智垚还在 2020 中国大学生保险数字科技挑战赛华南赛区取得第 2 名；聂才在 2020BDCI 大数据时代的 Serverless 工作负载预测中取得第 8 名。

摘要

针对本次比赛提供的真实商业交互数据，结合赛题目标定义为一个多任务抽取任务，通过数据分析发现存在样本的长度较大、实体类别多、实体存在过长等问题，根据赛题数据特点，我们以动态数据划窗及数据增强两种方式重新构建训练数据，以多种 BERT 预训练模型作为文本表示，分别采用序列标注模型框架和指针标注模型框架两类实体识别模型框架作为基准模型。 我们团队对上述模型辅以模型微调，数据增强，对抗学习等手段，极大提升了单模型的性能。 同时我们团队根据数据分析和结果错例分析制定了一套充分挖掘两类模型(BERT-SPAN 和 BERT-CRF)互补性的模型融合及处理策略。本次比赛我们主要贡献包括：1、提供了长文本序列标注任务的有效文本处理方法； 2 针对模型泛化问题，设计多种的性能优化方案； 3、针对模型的差异性，设计了一套异构模型互补的模型融合方案。最终团队以 0.9062 分数取得 B 榜排名第二。

关键词

预训练模型，数据增强，对抗学习，序列标注，指针标注

1 数据分析

针对赛题数据，我们团队进行了详细的统计和分析，赛题任务需要识别 14 中不同类型的隐私数据。如图 1 所示，文本长度方差较大，数据中存在不少超长文本的现象。 其中文本长度大于 100 的数据占比 50%，文本长度超过 270 占比 10%，同时还有 1.5%的数据文本长度超过 512，最大的文本长度高达 1000 个字符。如图 2 的实体类型数量统计可以发现，实体数量较多的三类实体类型分别为：organization, movie 和 name。这三类实体在数据集中的分布远远高于其

他类型的实体，还有 government，game，QQ，VX 四类实体在数据集中分布较少，数量远远低于其他实体类型。

经过我们对数据分析，比赛数据共存在如下几点问题：

1. 文本中语义不连贯，语境断裂现象明显；
2. 文本长度普遍过长；
3. 数据存在实体类别不平衡情况；
4. 文本数目相对较少；
5. 个别类别数据存在标注标准不统一及漏标的情况。

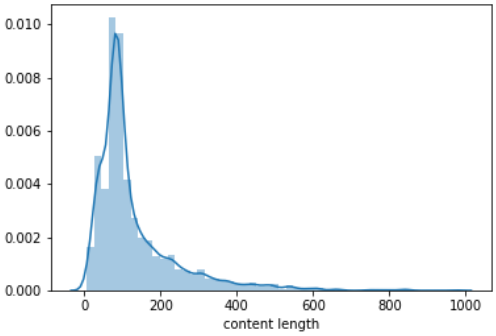

图 1：文本长度分布

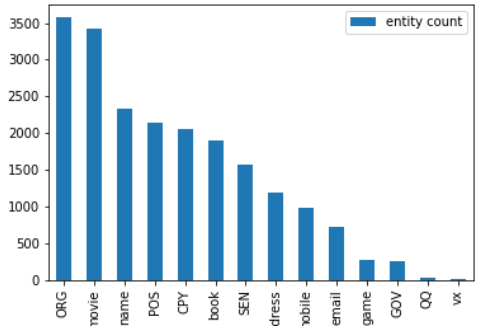

图 2：实体类型数量统计图

2 数据处理

2.1 数据预处理

在文本处理阶段，我们考虑了文本长度的影响，采用了数据划窗策略来优化模型性能。为了尽可能保留上下文信息，本团队在划窗过程中以标点符号优先级对句子进行切割并按原顺序重组，当重组句子长度超过阈值时，则生成一条新的子数据。 此处阈值根据两类模型对于文本长度的敏感性有所区分。 我们的后处理策略缓解了预训练模型无法建模长文本的问题，且较完整的利用了数据信息。

2.2 数据增强

由于训练数据较少，仅 2515 条，我们使用了标签相对一致的外部数据 cluener[2]作为数据增强的手段，通过实验发现外部数据可以显著提高模型准确性和泛化性。

3 模型框架

针对该任务本团队使用了指针标注框（BERT-SPAN）与序列标注框架（BERT-CRF）两套差异度较大的 NER 框架构建基准模型，同时采用了常见的开源预训练模型 Robert，NeZha，MacBert 提升下层模型的语言表征能力。

3.1 BERT SPAN（指针标注框架）

指针标注框架（BERT-SPAN）思想是采用对每个实体的开头和结尾的字符进行标记的方式完成实体的抽取。其原理如图 3 所示，对于一段文本序列，指针标注框架有等长的 start 序列（end 序列）来标记输入文本序列里的每一个字符是否为某一类实体的开始（结束）字符。该模型将多片段抽取问题转化为 N 个 K 分类（N 为序列长度，K 为实体类型数目），span 模型可以轻松预测出文本长度较长的实体，例如 position 类的实体。

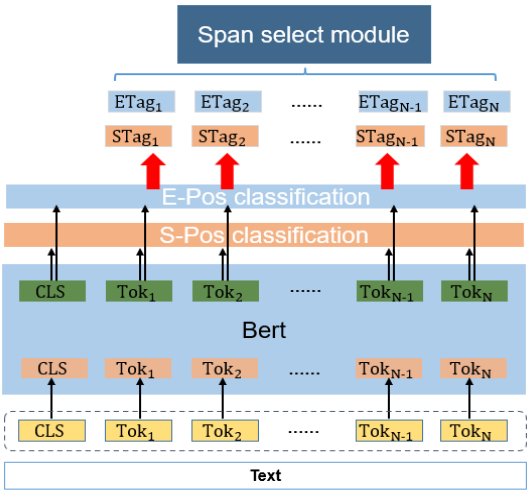

图 3：指针标注框架图

3.2 BERT CRF（序列标注框架）

结合 BERT 和全连接层对输入的每个 token 进行分类的方式可以解决序列标注问题， 但是 softmax 层的输出是相互独立的，即虽然 BERT 学习到了上下文的信息，但是输出相互之间并没有影响，它只是在每一步挑选一个最大概率值的 label 输出，会导致如 B-name 后再接一个 B-name 的问题。CRF 是一种经典的概率图模型，可以加入一些约束来保证最终的预测结果是有效的，这些约束可以在训练数据时被 CRF 层自动学习得到。 因此，本次比赛采用 BERT-CRF 框架作为第二个基准模型，其原理如图 4 所示。

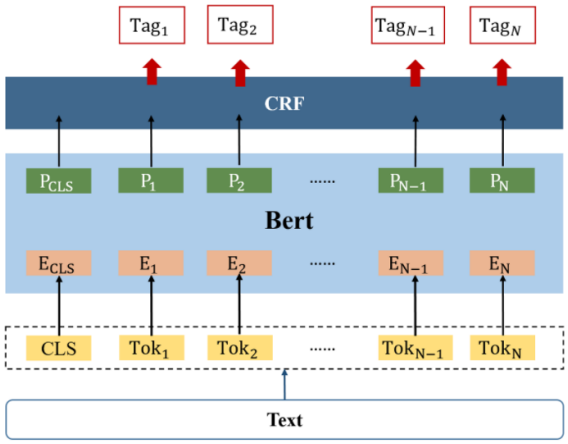

图 4：序列标注框架图

3.3 模型调优

- 分层学习率[3]：对于 bert 后接 crf 层，本团队设置 CRF 学习率为 bert 的 10-100 倍，可以有效提高 BERT-CRF 的训练效率。
- 学习率优化： 我们使用 adamw 结合动态学习率调整策略，并通过 warmup 增加训练过程的平稳性。
- 对抗学习： 对于上述的两个框架我们通过 fgm[1]提升对标注噪声的抵抗，提升其泛化性。我们在模型优化过程中给样本的词嵌入加入一个梯度上升方向上的微小扰动，将加入扰动后的样本看作是对抗样本，加入训练。具体如下面公式所示：

$$r_{adv} = \epsilon \cdot \frac{g}{\|g\|_2} \quad (1)$$

其中$g = \nabla_x L(\theta, x, y)$。

- Focal loss[4]： 我们在赛题初期使用 focal loss 配合 span 模型解决类别失衡问题，但是数据量少时 BERT-SAPN 模型本身效果弱于 BERT-CRF 模型；后期经过数据增强，使用 focal-loss 对少量失衡实体有提升，但对整体提分不明显故而未使用该策略。

4 模型集成及后处理

4.1 投票融合策略

为了让模型能够获得不同的归纳偏置，取得更好的融合效果，我们采用不同的预训练模型对语句进行编码，同时分别采用序列标注模型（CRF）和指针模型进行解码（Span）。经过组合后，我们一共有五个基础模型，分别是 BERT-CRF+Nezha，BERT-CRF+Roberta，BERT-CRF+Macbert，BERT-SPAN+Roberta，BERT-SPAN+Bertbase，如图 5 所示，我们将五个模型分为三组，每组分别进行 K（K=5）折投票融合得到三个不同的结果：Nezha Result，CRF Result 和 Span Result。最后将 3 份结果进行投票，留下票数不低于两票的结果。

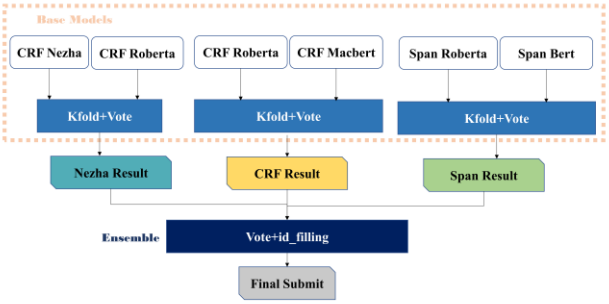

图 5：模型融合策略示意图

4.2 后处理策略

我们对指针标注框架（BERT-SPAN）和序列标注框架（BERT-CRF）得到的结果进行对比分析时发现，相对于 BERT-CRF，BERT-SPAN 得到的实体数量要少，甚至出现有些样例没有抽取到任何实体的现象。我们发现，在本赛题中 BERT-CRF 模型是一种高召回的模型，BERT-SPAN 模型是一

种高精确率的模型。当我们使用 BERT-SPAN 模型和 BERT-CRF 模型结果进行投票融合的过程中，会导致召回率越来越低，精确率越来越高。虽然投票融合能够使得模型的结果获得更高的精确率，但是低召回导致的实体缺失问题也在投票过程中凸显出来，因此我们采用的缺失实体填充策略。具体来说，如果一个测试样例没有被模型识别到任何实体，我们就把 BERT-CRF 模型中的高置信度（在所有 BERT-CRF 模型中都出现）实体填充到最终结果里面，可以有效缓解缺失实体的问题，其原理如图 6 所示。

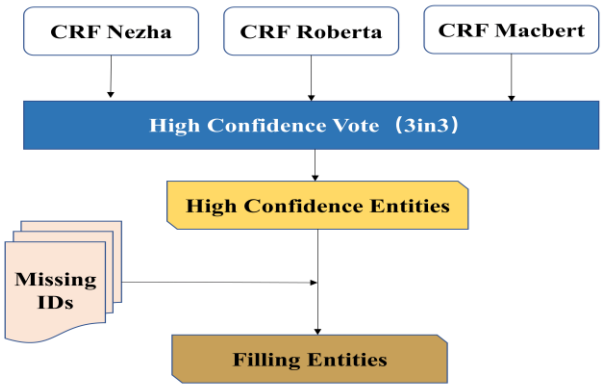

图 6：实体填充策略示意图

4.3 模型效果对比

表 1 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

通过不同模型对比，我们主要采用 BERT-CRF 和 BERT-SPAN 算法框架来构建模型，两者在基准模型上效果相近。我们还尝试加上双向 LSTM、CNN 等，但是对于本次 NER 任务的效果并不理想，同时会导致训练时间增加 20%。在基准模型基础上，增加数据增广（外部数据+动态数据滑窗）可以有效提高 0.08~0.12 的分数。除此之外，我们还对模型做各种优化策略，例如学习率调整、对抗学习和后处理策略等，单模型可以有 0.045~0.065 的提升。由于我们采用两套不同的模型框架（BERT-SPAN&BERT-CRF），这两套 NER 思想不同模型，经实验模型融合后有 0.005~0.006 分数的提升，得到比较大的融合收益。

总结展望

本次我们团队参赛尝试使用大量 NER 抽取技术，主要采用异构模型融合方式思路对不同隐私类别和类别信息进行抽取。结合赛题目标和数据的细致分析，我们制定一套完整的模型构建方案(序列标注模型框架和指针标注模型框架),通过数据增强、调参策略、对抗学习和模型融合与后处理等方法，可以有效的提高模型的效果。为了减少模型中的参数量和模型结构复杂度，可以使用蒸馏学习来缩减模型的规模达到应用效果。在实际应用场景中，可支持知识图谱、智能客服和垂直领域知识库建设等，对多个应用场景有着重要的技术价值和应用。

致谢

在本次比赛中，我们收获良多。首先我们想感谢主办方、赞助方、赛事组委会的精心筹备和辛苦付出，为我们提供了优质的赛题和便利的比赛平台；感谢一路走来为我们提供帮助的每一个人，感谢团队里的每一个人的付出和坚持，我们会永远铭记凌晨修改代码的艰辛，铭记 B 榜放榜前的忐忑，铭记得知进入决赛时的喜悦。

参考文献

[1]Miyato, T., Dai, A. M., & Goodfellow, I. (2016). Adversarial training methods for semi-supervised text classification. arXiv preprint arXiv:1605.07725.
[2]Xu, L., Dong, Q., Yu, C., Tian, Y., Liu, W., Li, L., & Zhang, X. (2020). CLUENER2020: Fine-grained Name Entity Recognition for Chinese. arXiv preprint arXiv:2001.04351.
[3] https://spaces.ac.cn/archives/7196
[4] Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision (pp. 2980-2988).