Privacy information recognition

宋和田 202028015029017 songhetian20@mails.ucas.ac.cn
王泽明 202028015059027 wangzeming20@mails.ucas.ac.cn
许婷婷 202028015059005 xutingting20@mails.ucas.ac.cn
夏文浩 202028015059029 xiawenhao20@mails.ucas.ac.cn
杜大钊 202028015029009 dudazhao20@mails.ucas.ac.cn

1 问题描述

本赛题取自 2020 CCF 大数据与计算智能大赛，题目是《非结构化商业文本信息中隐私信息识别》[1]。赛题数据基于真实商业交互数据，大赛对其中的隐私数据已进行脱敏处理并进行相应的标注，可用于隐私数据提取场景。本赛题要求参赛者从提供的非结构化商业文本信息中识别出文本中所涉及到的隐私数据。

该赛题训练数据包含（1）包含文本信息的.txt 文件（2）包含隐私标注信息的.csv 文件，其中 txt 文件和.csv 文件是一一对应的，且.csv 文件中的信息有：原文 ID、隐私信息文本、隐私信息的位置、隐私类别。

2 调研工作

该赛题是一道典型的自然语言处理序列标注中的命名实体识别任务 [4]，需要标注出某些字段是否为隐私数据，如果是隐私数据还需要标注出隐私的类别。

由于自然语言的特殊性，从零开始训练出完整的模型十分困难，需要大量数据集和高昂的计算成本。目前在 NLP 的相关应用中，大多采用预训练好的 BERT 模型或其变种作为任务的 Word2vec 部分并用于下游任务。BERT[3] 诞生于 2018 年，该模型一经提出就引起了 NLP 界的关注，可以将其作为下游任务的固定特征提取器，也可以直接对其进行微调。

3 备选方案

目前我们选用的建模方案是给 Bert 连接两个结构以适应命名实体识别任务，即 BERT+BiLSTM+CRF[5]，并准备试着将中间层 BiLSTM 替换为 LSTM 或者 GRU，对比查看效果。原始文本经过相应的清洗和预处理后输入到 Bert 模型，之后将 BERT 生成的词向量作为 BiLSTM-CRF 的输入，输出的是每个单元的标签，即完成了文本的序列标注。

其中 BERT 我们采用哈工大讯飞联合实验室的中文 RoBERTa 模型 [2]。该模型在 Google 官方的 Bert-Base 基础上将全词 Mask 的方法应用在了中文中，使用中文维基百科进行预训练。BiLSTM 是双向长短期记忆网络用于预测每个标签，CRF 是条件随机场用于为标签增加约束来保证预测的标签是合法。

在模型训练结束后，测试阶段，将测试集的文本逐条输入，并将输出转换成赛题要求的提交格式。

REFERENCES

[1] 2020. *CCF BIG DATA*. Retrieved Oct 23, 2020 from https://www.datafountain.cn/competitions/472

[2] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting Pre-Trained Models for Chinese Natural Language Processing. In *Findings of EMNLP*. Association for Computational Linguistics.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805 (2018). arXiv:1810.04805 http://arxiv.org/abs/1810.04805

[4] Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2018. A Survey on Deep Learning for Named Entity Recognition. *CoRR* abs/1812.09449 (2018). arXiv:1812.09449 http://arxiv.org/abs/1812.09449

[5] Ling Luo, Zhihao Yang, Pei Yang, Yin Zhang, Lei Wang, Hongfei Lin, and Jian Wang. 2018. An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition. *Bioinform.* 34, 8 (2018), 1381–1388. https://doi.org/10.1093/bioinformatics/btx761