《CCKS 2020：面向金融领域的篇章级事件主体与要素抽取（一）事件主体抽取》测评报告

队伍名称：BH
成员（昵称）：***（brobear）
***（hexiaoqing）
单位：华东师范大学

目录

一、 任务描述

事件主体抽取任务旨在从文本中抽取事件类型和对应的事件主体。即给定文本 T，抽取 T 中所有的事件类型集合 S，对于 S 中的每个事件类型 s，从文本 T 中抽取 s 的事件主体。其中各事件类型的主体实体类型为公司名称或人名或机构名称。

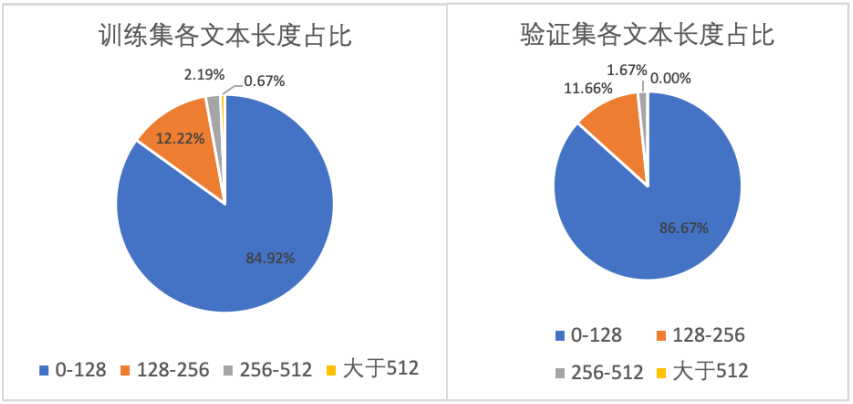
二、 主要思路

事件主体抽取任务分为两步进行，先识别出新闻文本中可能的事件主体，然后根据文本内容和事件主体通过分类获得事件类型。对于事件主体的识别，考虑到新闻中可能存在多个实体，故将其作为一个序列标注任务，即简化版的命名实体识别任务。第二步事件类型分类则是关于文本的多标签分类任务。

三、 事件主体的识别---命名实体识别

1. 数据处理

原始训练数据的格式为“文本 id\t 文本内容\t 事件类型\t 事件主体”，在此任务中近用到文本内容和事件主体两列，首先重复数据和事件主体为 NaN 的训练数据以及删除图片引用、网址、网页标签、连续的问号等无意义字符串。最后将文本内容相同事件主体不同的数据合并，并使用序列标注模式 BIOES 获得标签，用于训练。最终获得 33270 条训练数据。



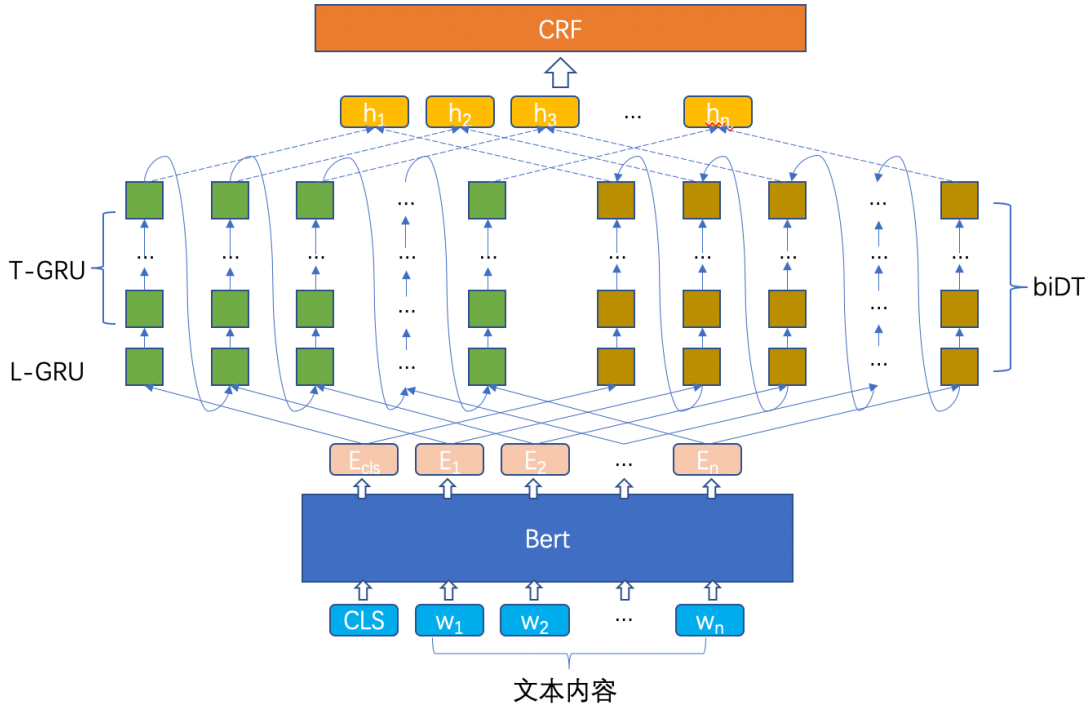
文本内容长度状况

考虑到文本内容长度大于 256 个字符的数据在训练集中仅占 2.86%，在验证集中仅占 1.67%，所以设置 Bert 输入的文本最大长度为 256，batch size 设为 16。对于超过最大长度的句子，通过截取实体前后的子字符串进行拼接得到符合长度要求的文本输入。

2. NER 模型搭建

尝试多种模型后，最终选择了 Bert＋biDT＋CRF 模型。biDT 是 biGRU 的变种，其中 DT 是指 Deep Transition RNN，旨在通过加深 GRU 状态转移网络的深度来提高模型性能。DT 分层模型中的深层过渡块由两个关键组件组成，即线性变换增强 GRU（L-GRU）和 Transition GRU（T-GRU）。 在每个时间步长，L-GRU 首先使用输入嵌入的附加线性变换对每个令牌进行编码，然后将 L-GRU 的隐藏状态传递到仅通过隐藏状态连接的 T-GRU 链中。 然后，将当前时间步长的最后一个 T-GRU 的输出“状态”作为下一个时间步长的第一个 L-GRU 的“状态”

输入。Bert 的预训练模型最终选择了哈工大讯飞联合实验室发布的 Bert-wwm-ext 和 Roberta-wwm-ext。



NER 模型结构图

3. 后处理

通过观察模型预测的结果，发现识别出的部分事件主体存在缺失“*ST”，“ST”以及少数实体存在“xxx 公司公司”、 “xxx 股份公司股份公司”等问题，通过代码进行处理。

四、 事件分类----文本的多标签分类

1. 数据处理

在删除图片引用、网址、网页标签、连续的问号等无意义字符串后，将文本内容、事件主体相同，事件类型不同的数据合并，构建用于训练多标签分类模型的标签，保留事件主体为 NaN 的数据，最终获得 67002 条训练数据。为减少训练集/验证集划分带来的影响，根据样本的事件标签对训练数据进行分成抽样，获得 5 份不相交数据，每次选取其中 1 份作为验证集，4 份作为训练集。

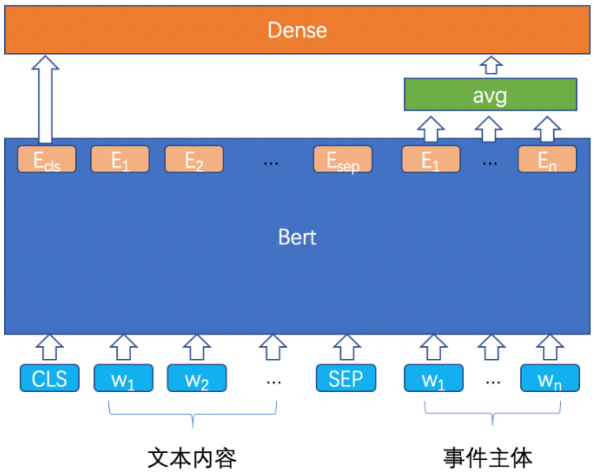
2. 添加白噪声

构造形如“文本内容\tNaN\tA”的数据，其中 A 为 text 的文本内容的子字符串。其来源有三个：使用 NER 模型识别的结果、文本内容的随机子字符串和通过企业实体字典匹配的实体。实现方式：读入一个训练样本，按概率随机构建负样本。首先文本内容不变，与训练样本相同。其次，事件主体标签按概率从三个来源随机获取一个。最后，事件类型标签的构造规则如下：如果新的事件主体和原有标签是嵌套关系，则分配原有的事件类型标签，否则分配 NaN。

3. 分类模型搭建

事件分类使用了 Bert+全连接层的模型，为了突出事件主体，在 Bert 和全连接层之间加入了平均与拼接操作。模型结构如下图所示。在模型融合方面，选

用投票机制，仅保留票数过半的结果。在 Bert 的预训练模型最终选择了 google 发布的 BERT-base。

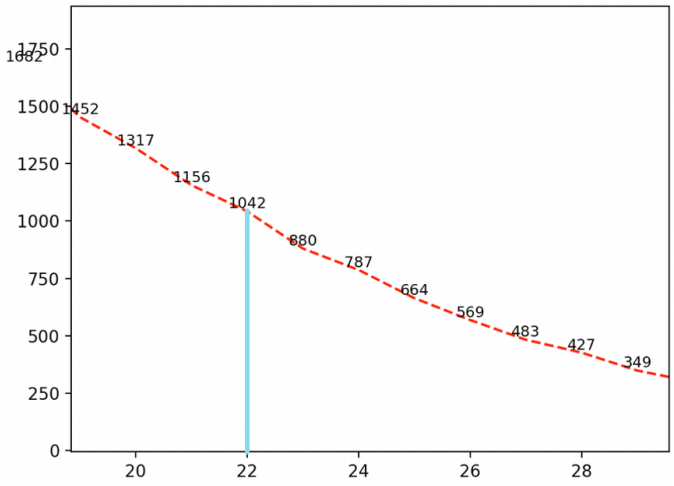


分类模型结构图

五、 针对 B 榜数据的处理

1. 主要流程

B 榜给了 29 万条左右的数据但仅有 1 千条有效数据，故需要筛选出最可能的答案。筛选规则如下，NER 模型选择获得票数最多的前 1 万个事件实体，分类模型对这 1 万个事件主体进行分类获得置信度最高的前 1 千、2 千、3 千个分类结果。最后对获得的结果进行投票，根据票数情况选择阈值为 22，从而获得结果。



票数阈值选择

2. 规则补充

发现结果中缺失“重组失败”、“ 履行连带担保责任”等标签，所以通过关键词匹配补充结果。例如按关键词“连带”、“责任”筛选出 171 条与“承担连带责任”相关的数据，通过 NER 模型提取的事件主体与关键词的相对位置及顿号等标点符号，筛选出结果，进行补充。

3. 后处理-规则过滤

因为主体识别的结果中存在一些地区名称、政府机构、球队，人名等非企业实体。所以根据非企业实体词典过滤掉“特朗普”、“C 轮”等不合理的结果。然后再根据从训练集中抽取的匹配规则，检查事件类型与所对应的文本是否存在相应的匹配规则，并将不匹配的结果去除。

六、 实验结果

1. 在 A 榜上的实验结果

在本次比赛中尝试了多种模型并进行了比较，同时也尝试了几种中文预训练模型。关键实验的结果如下表所示。从实验 1、2、3、5、6 可以看出事件主体识别模型 Bert+biDT+CRF 明显优于另外 3 个模型，且 bert-wwm-ext 优于 google 的 bert-base。其中 crossweight 框架是一种通过调整样本权重来减少错误样本对模型影响的框架，从实验 3、4、7、8 可以发现 crossweight 框架仅在前期提高了分数。表中“ner 标签补全”是指通过事件主体词典补全训练数据的 ner 标签。在事件分类模型中 Bert+entity+Dense + 白噪声的组合虽然得分不是最高的，但从实验 11 和 12 可以看出，其在事件主体较多的情况下性能相对更好。最终选择了 7、9、10、12 四个模型的结果并通过投票机制获得了融合后的结果，取得了 0.7959 的 F1 值。

序号	事件主体识别模型	事件分类模型	A 榜 F1 值
1	Bert+Dense	Bert+Dense	0.7597
2	Bert+CRF	Bert+Dense	0.7633
3	Bert+biLstm+CRF	Bert+Dense	0.7716
4	Bert+biLstm+CRF + crossweight 框架	Bert+Dense	0.7744
5	Bert-ext+biLstm+CRF	Bert+Dense	0.7766
6	Bert-ext+biDT+CRF	Bert+Dense	0.7808
7*	Bert-ext+biDT+CRF	Bert+Dense + 后处理	0.7894
8	Bert-ext+biDT+CRF + crossweight 框架	Bert+Dense	0.7758
9*	Bert-ext+biDT+CRF	Bert+entity+Dense + 白噪声	0.7710
10*	Roberta+biDT+CRF	Bert+Dense	0.7773
11	Bert-ext+biDT+CRF + ner 标签补全	Bert+Dense	0.6588
12*	Bert-ext+biDT+CRF + ner 标签补全	Bert+entity+Dense + 白噪声	0.7170
13	对带*的 4 个模型进行融合		0.7959

2. 最终结果

测试集	F1 值	排名
A 榜	0.79597	27
B 榜	0.21224	7

七、 其他尝试

1、将事件主体抽取任务分解为文本分类任务和机器阅读理解任务，先获取新闻文本的事件类型，然后通过回答“xx 事件的主体是谁？”获得事件主体。初步尝试发现效果不太好。

2、在训练分类模型时添加标签权重，在 A 榜上效果不好。

八、 实验环境

Google Colab

GPU 型号：P100

GPU 显存：16G

框架：keras+tensorflow1.14

九、 参数设置

主体识别模型参数	取值	分类模型参数	取值
batch size	16	batch size	16
max length	256	max length	256
learning_rate	5e-5	learning_rate	3e-5
max epoch	15	max epoch	15
earlystop patience	3	earlystop patience	3
earlystop monitor	val_acc	earlystop monitor	val_f1_metric

参考文献

1. Meng, Fandong, and Jinchao Zhang. "DTMT: A novel deep transition architecture for neural machine translation." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 2019.
2. Liu, Yijin, et al. "GCDT: A Global Context Enhanced Deep Transition Architecture for Sequence Labeling." Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019.
3. Wang, Zihan, et al. "Crossweigh: Training named entity tagger from imperfect annotations." arXiv preprint arXiv:1909.01441 (2019).
4. Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2014. How to construct deep recurrent neural networks. In ICLR.