上市公司公告事件要素抽取研究

纪梦兰[1]，李婷[1]，彭本[1]，张士松[1]

[1]广州华资软件技术有限公司
963570357@qq.com

摘要：本文是对本团队在 CCSK 2020 中面向金融领域的篇章事件要素抽取任务的提交报告。本文设计了针对篇章要素抽取的管道（pipeline）模式的方案，先后通过事件分类、句子要素分类、句子要素提取和规则处理来获得最终的要素信息。最终，方案在 A 榜中成绩为 0.784，排名第四，在 B 榜中成绩为 0.65，排名第二。

关键词：BERT,信息抽取,阅读理解,文本分类

1 引言

事件抽取通常以一张丰富的事件图谱来呈现其价值，它能支撑起日常的信息检索、兴趣推荐、智能问答以及其他行业的特殊应用。在金融领域里，事件抽取能辅助投资者的决策分析，帮助监管部门保持警惕，因此，事件抽取必然将成为一个重要的研究课题。

事件抽取要求我们用人工或者自动的方法，从半结构化或者非结构化数据中，识别一个与我们的目标相关的事件的重要元素识别出来。按模式分，事件抽取方法分为 pipeline 和 joint 两种模式。按步骤分，事件抽取一般包含触发词抽取（Trigger）、事件分类和论元（Argument）抽取。其中触发词识别是指根据上下文识别出触发词；事件分类是指根据触发词及上下文判断事件类型；论元抽取则是根据事件类型，抽取出参与事件的论元（即事件的参与者，或称元素、要素），并识别出对应的论元角色（Argument role，即事件论元在事件中的角色）。

随着深度学习的进一步发展，传统的事件抽取从依赖手工特征 Li and Zhou(2012)[1]过渡到神经网络领域。在神经网络的不断发展中，各种神经网络被用来用低维向量自动表示文本语义，并进一步基于这些语义向量提取事件参数，包括卷积神经网络 Chen et al.(2015)[2]和递归神经网络 Nguyen et al.(2016)[3]。在 BERT[7]的横空出世后，预训练语言模型得到了极大的发展，深度学习提高了许多自然语言处理的性能。很多自然语言理解的任务得以转换为机器阅读理解任务 Mccann et al.(2018)[4]。对于篇章级别的要素抽取任务来说，句子当中只有少数几个关键句能够反映该事件的中心（中心句），Hang Yang et al.(2019)[5]以文档中某句为事件中心句进行论元补充，为了解决篇章包含

的多个事件，Shun Zheng et al.(2019)[6]将事件抽取转化为端到端构建基于实体的有向无环图。

受到刚刚结束不久的LIC2020事件抽取任务的启发，本文最终采用了预训练模型BERT作为特征提取，融合了自然语言处理中的分类模型和阅读理解模型，通过pipeline模式完成事件要素的抽取。

2 数据

2.1 数据来源

本文的课题来源于CCKS2020面向金融领域的篇章级事件主体与要素抽取，其中任务二为篇章事件要素抽取。该任务旨在从文本中抽取事件类型和对应的事件要素。目标为给定文本 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 的事件要素。

2.2 类型说明

训练数据共有9个篇章类型，每个篇章类型有各自包含的要素角色，分别为：

事件类型	要素角色
破产清算	公司名称、公司行业、受理法院、裁定时间
股东减持	减持开始日期、减持的股东、减持金额
股东增持	增持开始日期、增持的股东、增持金额
股权冻结	冻结开始日期、冻结结束日期、冻结金额、被冻结股东
股权质押	接收方、质押开始日期、质押方、质押结束日期、质押金额
重大安全事故	伤亡人数、公司名称、其他影响、损失金额
重大对外赔付	公司名称、赔付对象、赔付金额
重大资产损失	公司名称、其他损失、损失金额
高层死亡	公司名称、死亡/失联时间、死亡年龄、高层人员、高层职务

表 1 篇章类型及其要素角色

2.3 格式说明

数据格式包括{doc_id,events,content}，具体训练样本如表2：

content	股票简称：S*ST 长岭 股票代码：000561 公告编号：2009—084 长岭（集团）股份有限公司破产程序进展公告本公司董事会全体成员保证公告内容的真实、准确和完整，没有虚假记载、误导性陈述或者重大遗漏。根据宝鸡市中级人民法院《民事裁定书》（[2007]宝市中法破字第14-14号）裁定的《重整计划》，公司债权总额进行了调整，其中职工债权、税款债权按100%清偿，普通债权按18%清偿，调整后公司应清偿债权总额为2.5亿元。目前公司已进入重整计划

	执行期，偿债资金已到位，现已偿还债务19821万元，尚有5179万元债务未清偿。公司将根据重整计划执行情况及时履行信息披露义务。特此公告长岭（集团）股份有限公司董 事 会二〇〇九年六月二十九日
doc_id	2159480
events	[{"event_id": "4219323", "受理法院": "宝鸡市中级人民法院", "公司名称": "长岭（集团）股份有限公司", "公告时间": "二〇〇九年六月二十九日", "event_type": "破产清算"}]

表 2 一条训练样本

content 可根据文本结构大致分为两类：一类为内地发布的公告信息，一类则为香港交易所发布的公告信息。通常需要表格说明的的事件类型有：股东增持、股东减持、股权冻结、股权质押；未包含表格说明的事件类型有：破产清算、重大对外赔付、重大安全事故、高层死亡、重大资产损失。

通过分析训练数据发现，每个公告 content 几乎都存在固定格式（开头标题+正文+公告公司名称+公告时间），并且不同类型的事件要素提取存在各自的规则，因此，在后处理的时候需要针对不同事件类型做对应的处理。

各个事件类型数量分布如图 1 所示：

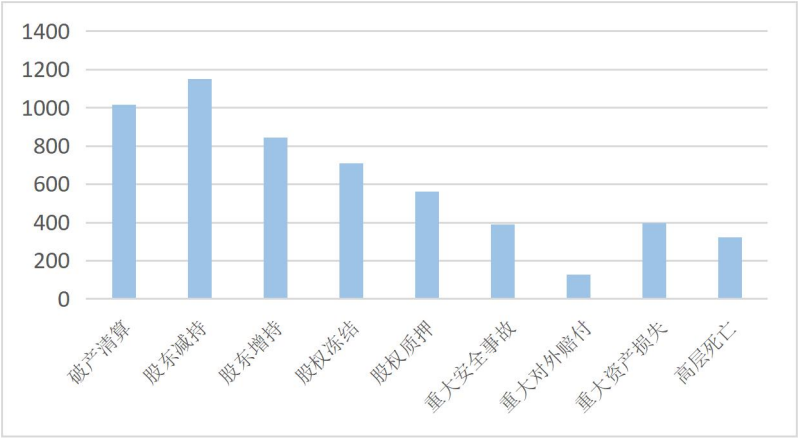

图 1 事件类型分布

3 方法描述

本文的事件要素抽取流程框架具体如下：

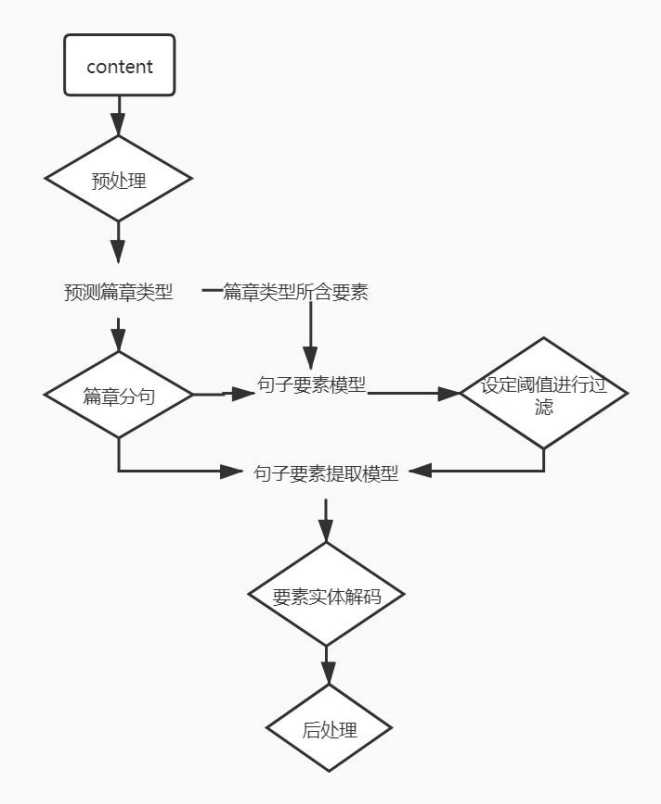

图 2 事件要素抽取流程框架

3.1 篇章预处理

由于导入的篇章文本结构比较混乱，因此需要对文本进行预处理：

(1) 去掉多余的解析符号
(2) 繁转简
(3) 去掉多余的公告信息
(4) 优化文本内容

3.2 预测篇章类型

从分析数据上看，篇章的标题一般位于开头位置，而标题往往可以预判篇章类型，因此，通过预处理后，提取开头的前三个句子，输入模型中进行训练和预测。具体结构如图 3:

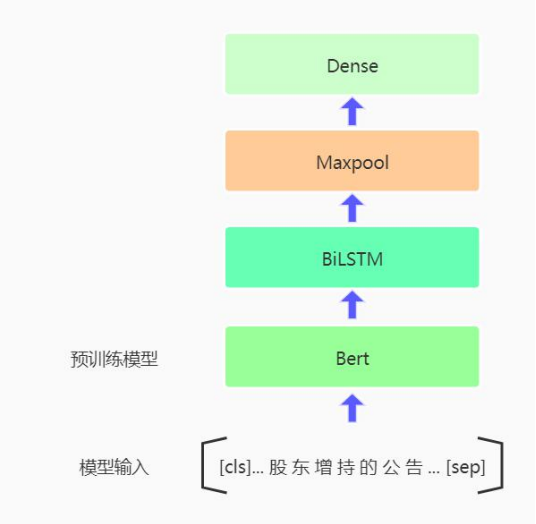

图 3 篇章类型模型

将文本输入 bert 模型当中，将得到的模型特征输入到双向 LSTM 层学习新的特征，接着输入最大池化层，最后传送至输出层进行分类，得到每个类别的概率。

3.3 预测句子所含要素

- 对 content 中的简称替换成全称，如将“重庆秦安机电股份有限公司（下称“秦安股份”）”中的"泰安股份"全部用"重庆秦安机电股份有限公司"进行替换，此处采用人民日报的机构数据以及训练集中出现的特殊机构名称做训练，采用的是 bert-crf[8]模型
- 优化文本内容，如去掉一些造成干扰的括号内信息等等
- 根据公告的说明结构对篇章进行分句
- 筛选出可能包含要素的句子
- 对长句子进行切分
- 对部分要素角色内容修改，比如：“股权质押的接收方”改为“股权被质押给谁”，“重大安全事故的其他影响”改为“公司行政处罚罚款缴纳金额”等等
- 句子拼接篇章类型所包含的所有要素角色，具体结构如图 4

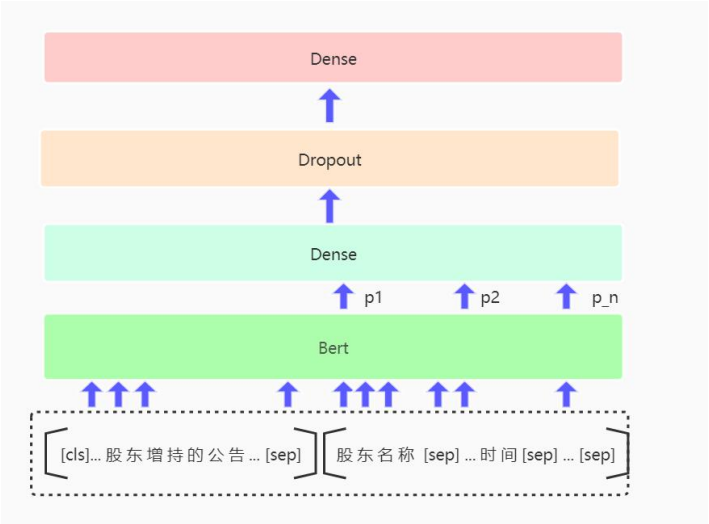

图 4 句子要素模型

在句子要素模型中，将分句后的句子以及 n 个要素角色（$ele_1, ele_2, ..., ele_n$）拼接在一起，其中 ele_1 如图为股东名称，输入 bert 模型后提取最后一层的输出作为特征，提取该特征层中的 n 个要素角色首位的 bert 特征（$p_1, p_2, ... p_n$），再输入全连接层和 dropout 层，最后得到输出层的输出结果，即句子包含各个要素的概率值。

- 根据阈值进行句子过滤，得到可能包含要素的句子

3.4 句子要素提取

3.4.1 开始和结束的位置预测

将需要提取的要素当做问题 Q，将句子当做材料 C，通过阅读理解模型获得事件句子要素实体。模型结构如图 5 所示。

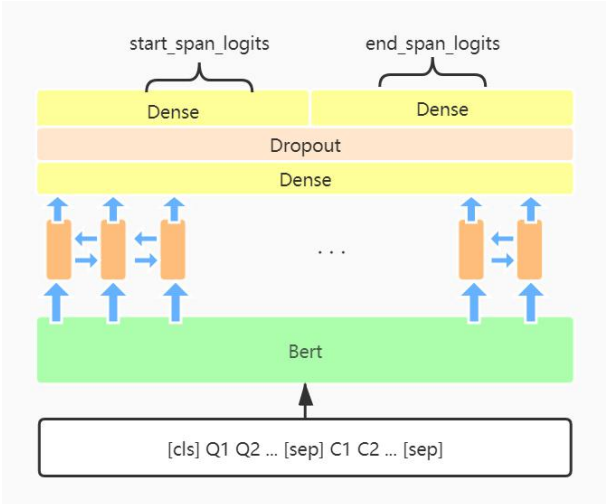

图 5 要素提取模型

构建好输入序列后，输入到 BERT 的预训练网络中，经 BERT 编码，序列中的每个字被编码成了 768 维的向量输出。将该向量经过 2 层双向 GRU 层的编码获得新的特征，将其依次输入 256 层的连接层和 dropout 层，然后将输出分成两个向量，分别计算开始位置和结束位置的字符得分。得分越高，表明越有可能是要提取的事件主体在事件描述中的开始位置和结束位置。

3.4.2 解码过程

将预测的结果进行解码，提取开始位置向量大于阈值 t_1的，对每个符合阈值的开始位置 a，判断该位置后的结束位置向量大于阈值 t_2的，提取符合阈值 t_2的首个位置 b，通过文本截取 ab 位置，即得到要素实体。对于部分要素实体明显长度过长或者不全的情况，则需要通过一些规则重新确定位置。

3.5 后处理

通过以上步骤，我们得到：句子与其提取出来的要素实体，接下来针对有表格的四种类型（股东增持、股东减持、股权冻结、股权质押），需要进行比较复杂的后处理：

① 单句子的信息处理
② 句子合并的信息处理

对没有表格的其他五种类型（破产清算、重大对外赔付、重大安全事故、高层死亡、重大资产损失），后处理则比较简单：对所有句子提取出来的要素实体进行投票，当票数一样时则比较对应要素实体的总得分，得分分数为要素提取模型的开始概率加上结束概率，票数和得分高的即为提取结果。最后对缺乏公司名称的提取结尾备注的公司名称，缺乏公告日期要素的提取结尾公告日期，方法是截取 content 尾部片段，输入 bert-crf 预测机构以及正则匹配结果相结合，补充进该句子要素。

3.5.1 单句子的信息处理

- 对提取出错的句子进行过滤
- 设置提前停止提取的触发词
- 针对一个句子多种要素的情况，根据所在前后位置进行合并成该句子要素
- 对部分表格中，人物信息位于合并单元格，解析过后只在开头出现一个人物信息的情况，则进行人物补齐
- 如果要素出现互相包含的情况，或者去掉相同元素之后没有共同元素，对要素进行融合合并，如“ [{'股权质押的股权被质押给谁': '中国银行股份有限公司无锡锡山支行', '股权质押的质押金额': '2000000', '股权质押的质押登记日质押开始日': '2018 年 2 月 23 日', '股权质押的质押实际股东控制人': '红豆集团有限公司'}, {'股权质押的质押解除日质押到期日': '2020 年 12 月 12 日', '股权质押的质押登记日质押开始日': '2018 年 2 月 23 日', '股权质押的质押实际股东控制人': '红豆集团有限公司'}]”，合

并的结果为：[{'股权质押的股权被质押给谁': '中国银行股份有限公司无锡锡山支行', '股权质押的质押实际股东控制人': '红豆集团有限公司', '股权质押的质押登记日质押开始日': '2018 年 2 月 23 日', '股权质押的质押解除日质押到期日': '2020 年 12 月 12 日', '股权质押的质押金额': '2000000'}]

3.5.2 句子合并的信息处理

- 判断句子是否是表格信息
- 判断增持和减持事件里，含有表格的句子是否缺乏股东人物信息，如果缺乏则补齐出现次数最多的一个；冻结事件同理
- 如果整个篇章中有包含表格信息的话，那么只需要提取表格句子所提取的句子要素，剩下句子不需要提取；如果篇章中没有包含表格信息，那么根据句子先后顺序提取出句子要素信息
- 同单句子的信息处理一样，为了避免句子要素信息之间出现互相包含，对其进行融合合并处理
- 如果要素之间只有金额不一致，并且金额是同个概念的不同表示，如“1000 万”和"10000000"，则只提取一个，日期同理
- 过滤掉要素信息不齐全的句子要素，比如该句子只提取了“股东名称”
- 对于特殊结构，如“预披露公告”特殊处理
- 对冻结和质押事件缺乏的日期要素实体设置为空
- 对所有要素恢复原文对应的简繁体

4 实验过程

BERT 模型的最长序列长度为 512，优化算法为 Adam 算法，学习率初始设 1e-5，每隔 2-3 个 epoch 衰减 1e-1，batch_size 为 16。

篇章类型模型

文本经过简单的预处理后输入模型训练，双向 LSTM 层的隐层节点数为 128 层。由于训练数据中出现一个篇章多种类型的数据并不多，因此此处只考虑一个篇章只有一种类型。

句子要素模型

通过句子分句之后，判断句子所含要素角色，该模型由两个模型组成，其一是利用较为粗糙规则过滤的数据训练而得，其二是利用精细规则过滤的数据训练而得，通过两个模型融合预测，能够避免部分要素角色漏选。

要素提取模型

一篇文章不同句子包含的要素信息有多有少，对要素信息含量进行排序，取该篇章中要素信息含量前四的句子作为正样本，同时随机采样没有包含任何要素角色的句子作为负样本，负样本数为八个，对个别少样本的要素做数据增强（比如英文格式的公司名称，人工随机生成英文替换增强），修改为模型的输入格式进行训练。模型双向 GRU 层的隐层节点数为 200 层。采用 focal_loss 作为损失函数。由于在 B 榜数据中出现很多日期间杂空格的现象，此处模型训练时并未对日期出现空格的情况做特殊训练，因此提取日期的结果并不太好。

5 结论

本文方法仍然存在不足之处，由于训练数据的多样性不够丰富，因此此前如篇章多标签等部分问题并未得到解决；预处理和后处理采用的规则较多，在处理上容易出现差错；由于提取规则不够规范，所以无法判断提取是否准确，如原文出现的公司名称中间出现空格，提取时是否去掉；部分实体的边界仍然无法准确识别；没有尝试 joint 方法去实现。

本文提供了一种 pipeline 式的篇章要素抽取方法，在最终测试集的验证下证明了其有效性。在未来的工作中，我们还将尝试更多的模型和思路。

References

1. Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event Peifeng Li and Guodong Zhou. 2012. Employing morphological structures and sememes for Chinese event extraction. In Proceedings of the 24th International Conference on Computational Linguistics, pages 1619–1634.
2. extraction via dynamic multi-pooling convolutional neural networks. In Proceedings of ACL-IJCNLP, pages 167–176.
3. Thien Huu Nguyen, Kyunghyun Cho, and Ralph Grishman. 2016. Joint event extraction via recurrent neural networks. In Proceedings of NAACL-HLT, pages 300–309.
4. Bryan Mccann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering. arXiv: Computation and Language.
5. Hang Yang, Yubo Chen , Kang Liu , et al.: DCFEE: A Document-level Chinese Financial Event Extraction System based on Automatically Labeled Training Data. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics-System Demonstrations. 2018: 50-55.
6. Shun Zheng, Wei Cao, Wei Xu, et al.: Doc2EDAG: An end-to-end document-level framework for chinese financial event extraction. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. 2019: 337-346.
7. J. Devlin, M.-W. Chang, K. Lee, and K. J. a. p. a. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2018.
8. N. Pang, L. Qian, W. Lyu, and J.-D. J. a. p. a. Yang, "Transfer Learning for Scientific Data Chain Extraction in Small Chemical Corpus with BERT-CRF Model," 2019.
9. https://github.com/qiufengyuyi/event_extraction
10. https://github.com/bojone/kg-2019