基于多任务联合训练的事件主体抽取模型

任君翔，钟曦
万达信息股份有限公司，上海，201112
renjunxiang1215@sina.com

摘要：本文的目标，是解决金融领域的事件主体抽取问题。本文基于RoBERTa 作为预训练模型，采用文本分类、实体识别和阅读理解相结合的方案，通过共享语言模型、联合训练子任务的方式，兼顾运算效率和模型精度，从文本中准确的抽取事件类型和事件元素，达到良好的事件主体抽取效果。

关键词：文本分类，实体识别，阅读理解，联合训练

1 引言

“事件抽取”是舆情监控领域和金融领域的重要任务之一，“事件”在金融领域是投资分析，资产管理的重要决策参考；事件也是知识图谱的重要组成部分，事件抽取是进行图谱推理、事件分析的必要过程。本次比赛的任务是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。

基于机器阅读理解（MRC）的事件抽取方法和命名实体识别方法近年来随着深度学习的发展引起了广泛关注[1]。基于知识的机器阅读理解（KBMRC）[2]，多文章 MRC 任务[3]以及对话问答（CQA）[4]正逐渐成为研究热点。[3]提出了端到端的神经网络框架，该模型基于预测边界、答案建模及多文章间答案验证三个因素预测答案，使得不同文章得到的候选答案可以基于它们代表的内容互相验证，解决了多文章 MRC 从不同文章得到多个混淆候选答案的问题。Thenmalar[5]等提出了一种基于半监督 Bootstrapping 算法的 NER 方法，该方法使用识别的命名实体、单词和上下文特征来定义模式，每个命名实体类别的此模式用作种子模式，以标识测试集中的命名实体，分别对英语和泰米尔语进行 NER，两种语言的平均 F1 值达 75%。[6]使用基于机器阅读理解(MRC) 的框架代替序列标注模型，统一处理嵌套与非嵌套命名实体识别问题，解决了序列标注模型无法处理嵌套命名实体识别的缺陷，在 8 个中英数据集上取得良好的效果。Jiang [7]等用一种由跨度和跨度之间的关系组成的框架来表示大规模自然语言分析任务。该框架首先提取跨度并预测其标签，然后预测跨度之间的关系。通过使用此 SpanRel 模型尝试了 10 个任务证明该通用的与任务无关的模型可以为每个任务量身定制，并能取得 SOTA 结果。[8]提出了 LSTM-CRF 和 Stack LSTM 两种神经网络模型，模型输入利用有限的监督数据，使用词嵌入结合预训练的特征和基于字符的特征，通过简单的 CRF 架构或基于转移的算法来构建和标记输入块，实现了 SOTA 效果。[9]提出了一种 read-then-verify 的模型，该模型不仅能够利用神经网络从候选答案中进行抽取，并且可以产生无答案概率，同时利用答案验证器来决定预测的答案是否来源于输入的片段。此外，通过引入两个新的 loss 函数辅助 reader 模型更好地解决答案抽取过程中没有答案的情况。该模型在 SQuAD2.0 数据集上取得了优异的结果。[10]利用强化助记忆符阅读器增强先前的注意力阅读器。在多轮对齐架构中，提出重关注机制，即通过直接访问历史注意力来提炼当前注意力的计算，从而避免注意力冗余和注意力缺乏的问题。该阅读器不仅能够在 SQuAD 数据

集上达到 SOTA 效果，在其对抗样本上仍具有良好的性能。

本次比赛，我们在 BERT 预训练模型下游任务[111] 的基础上，将事件主体抽取拆解为文本分类（Classify，以下简称 CLF）、抽取式阅读理解（Machine Reading Comprehension，以下简称 MRC）、实体识别（Named Entity Recognition，以下简称 NER）和观点型阅读理解（Opinion Questions Machine Reading Comprehension，以下简称 OQMRC），通过共享预训练模型、联合训练子任务的方式实现事件主体抽取。

2 数据

本次比赛主要围绕篇章级新闻文本和公司公告进行事件主体抽取，我们对句子的长度、事件的类型等进行数据分析，以便于后续的模型设计和调优。

通过数据分析可以发现，训练集和验证集的文本长度 99 分位数约为 300（表 1），大多为 100 以内的短文本（图 1）。因此在训练过程中可以将文本长度限制在 300 以内，从而扩大训练的 batchsize。

表 1 文本长度分布

分位数	训练集	验证集
50	67	77
75	91	109
90	132	148
95	164	187
99	291	297.03
99.5	346.11	319.03
99.9	506.822	336.707
max	340198	343

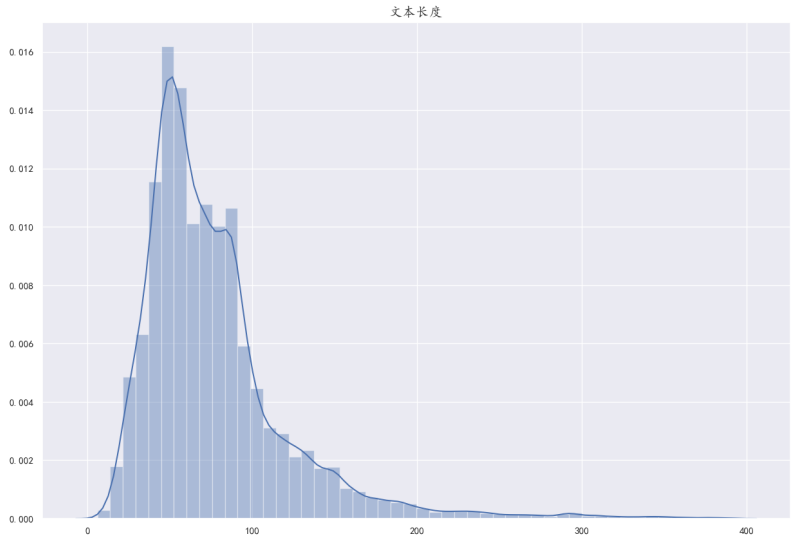

图 1 文本长度分布

通过对文本事件的频数统计和分布来看（表 2 和图 2），数据中事件类型分布不均匀，“业务资产重组”接近五分之一，前五个事件类型总和过半。因此，在后续训练是需要考虑事件类别的不平衡性，对损失进行调整。

表 2 事件类型分布

事件类型	频数	频率
业务资产重组	8000	19.48%
涉嫌非法集资	4210	10.25%
股票转让-股权受让	3414	8.31%
债务违约	3405	8.29%
涉嫌传销	3273	7.97%
实控人股东变更	2172	5.29%
交易违规	2062	5.02%
不能履职	1573	3.83%
涉嫌欺诈	1516	3.69%
涉嫌违法	1421	3.46%
实际控制人变更	1404	3.42%
重组失败	1283	3.12%
业绩下滑	834	2.03%
财务信息造假	723	1.76%
提现困难	710	1.73%
财务造假	705	1.72%
资金紧张	591	1.44%
商业信息泄露	554	1.35%
实际控制人涉诉仲裁	519	1.26%
歇业停业	458	1.12%
失联跑路	386	0.94%
高管负面	380	0.93%
资产负面	375	0.91%
资金账户风险	372	0.91%
债务重组	342	0.83%
投诉维权	297	0.72%
履行连带担保责任	88	0.21%

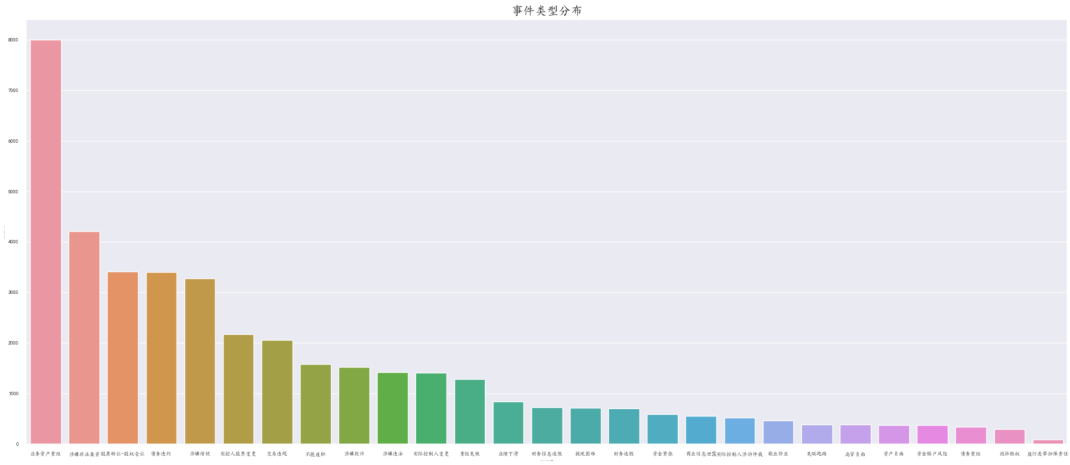

图 2 事件类型分布

通过对文本的主体数量、事件数量和相同实体不同类型数量进行分析可以发现（表 3），训练集中 10%以上的文本存在多个主体，5%以上存在多个事件，5%以上存在同一实体涉及多个事件类型。因此，在模型设计中，需要考虑文本多实体、实体多事件的情况。

表 3 标签分布

分位数	标签	事件	实体多类型
50	1	1	0
75	1	1	0
90	2	1	0
95	2	2	1
99	4	3	2
99.5	6	3	2
99.9	10	4	5
max	163	8	20

3 模型

3.1 事件类型-事件主体

通过文本数据的初步分析以及事件要素抽取的文献综述研究，我们设计了方案一，事件类型-事件主体（图 3）。

1. BERT 作为编码器，content 经 BERT 编码后输出为向量；
2. 选择 content 的[CLS]位置，经过全连接层，得到事件向量；
3. 通过 sigmoid 转为 0-1 的概率，选择概率 p>0.5 的位置索引，从而识别 type；
4. BERT 作为编码器，将 type 作为 query，拼接 content，通过 MRC 的方式，type+content 经 BERT 编码后输出为向量；
5. 抽取序列中 content 的位置向量，经过两个全连接层，输出首尾指针向量；
6. 利用 sigmoid 将 content 序列向量转为 0-1 的概率表示，计算文本序列首尾指针概率，当首尾指针概率 p 均大于 0.5，锁定 type 对应的 entity；

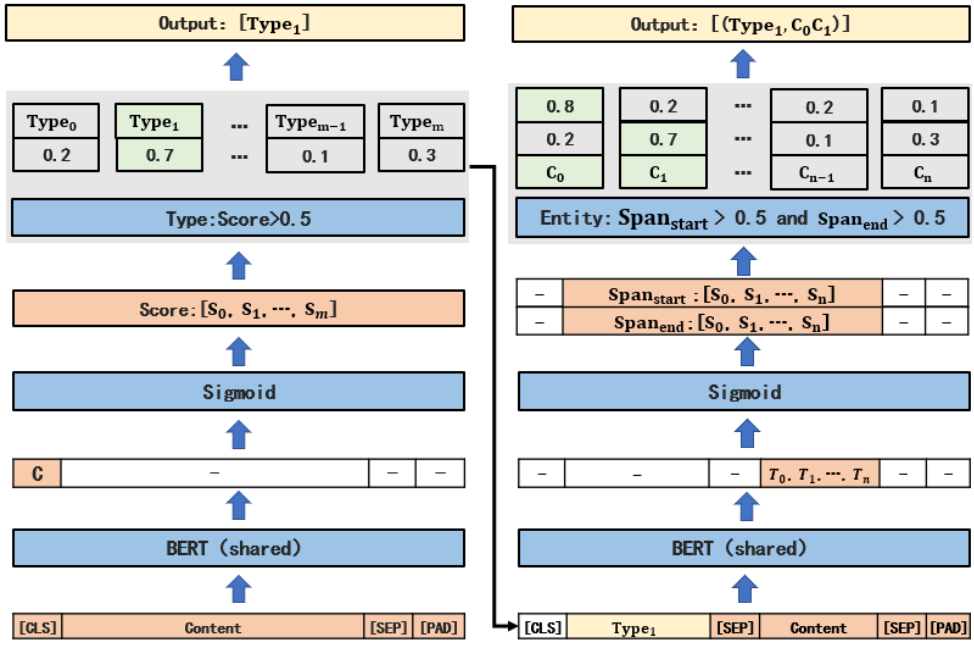

图 3 事件类型-事件主体的网络结构

两个任务共享预训练模型，激活函数均为 sigmoid，损失处于统一数量级。该方案的优点在于通过事件分类，可以在计算量较小的情况下过滤掉无关文本，尤其是复赛测试集庞大的情况下，具有较好的性能。

3.2 事件主体-事件类型

通过文本数据的初步分析以及事件要素抽取的文献综述研究，我们设计了方案二，事件主体-事件类型（图 4）：

1. BERT 作为编码器，content 经 BERT 编码后输出为向量；
2. 向量经过两个全连接层，输出首尾指针向量；
3. 利用 sigmoid 将 content 序列向量转为 0-1 的概率表示，计算文本序列首尾指针概率，当首尾指针概率 p 均大于 0.5，锁定 entity；
4. BERT 作为编码器，将 entity 作为 query，拼接 content，通过 OQMRC 的方式，type+content 经 BERT 编码后输出为向量；
5. 选择向量的[CLS]位置，经过全连接层，得到事件向量；
6. 利用 sigmoid 转为 0-1 的概率，选择概率 p>0.5 的位置索引，从而识别 entity 对应的 type；

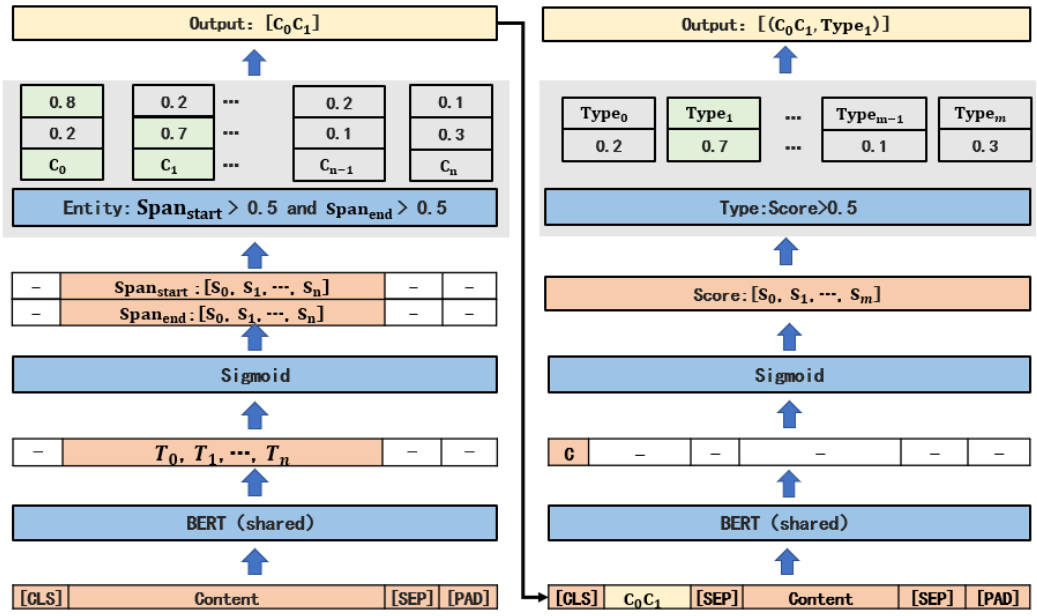

图 4 事件主体-事件类型的网络结构

两个任务共享预训练模型，激活函数均为 sigmoid，损失处于统一数量级。识别实体的任务相对于文本分类任务要更复杂，方案一的 MRC 对 query 比较敏感，有可能会出现事件相似的情况下，实体抽取混淆的情况。该方案的优点在于先识别 entity 的方式，避免了人为构造 query，也更贴近人类阅读新闻时从词到句进行理解分析的方式。

3.3 模型训练及融合

最终我们采用将事件抽取拆解为文本分类、抽取式阅读理解、实体识别和观点型阅读理解几个子任务，通过共享预训练模型、联合训练子任务的方式实现事件主题抽取，使得模型

更加合理和高效。模型输出以“事件”+“实体”为单位，形式为（type，entity，span）的三元组，方便后续模型融合时进行加权投票以及后处理优化，见图 5。

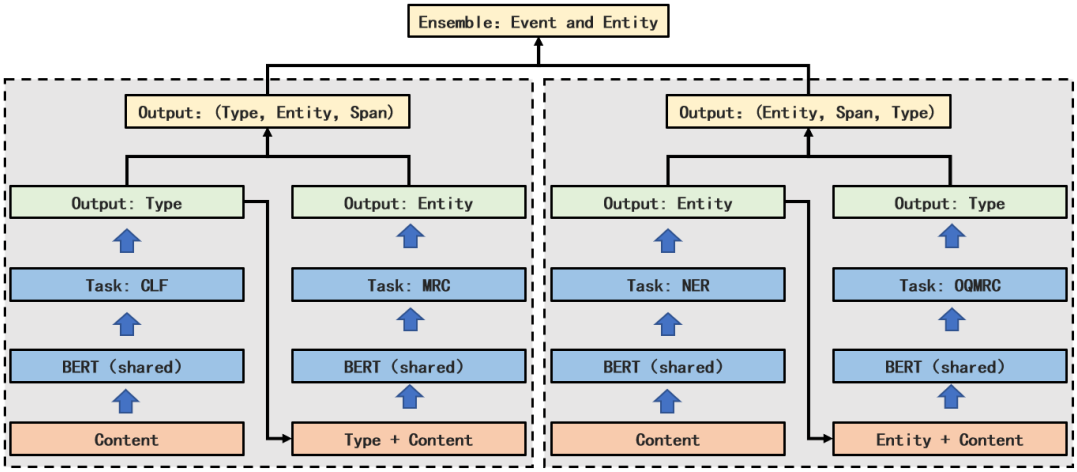

图 5 模型整体结构

训练过程，我们最终采用 RoBERTa-wwm-ext[12]作为预训练模型。为了让三个任务的损失在同一个量级，且考虑到事件类型的不平衡性，我们采用 Focal loss 作为损失函数。Focal loss 参数为默认值 gamma=2、alpha=1，其他超参分别为 optimizer=Adam、lr=1e-5、batchsize=2。考虑到文本长度在 300 的情况下，batchsize 较小模型不易收敛至最优，我们采用梯度累积的方法来扩大 batchsize 至 8。

模型输出以 entity 为基本要素，在融合的时候通过模型的 F1 进行加权投票，超过阈值的 entity 被保留。如果一条文本的在各个模型的输出完全一致，我们将该条文本和预测结果保存为伪标签，作为半监督学习的伪标签训练集再次训练。

4 实验结果

为了测试本文所提出的模型在事件抽取方面的效果差异，在使用相同测试样本数据集的基础上，分别采用上述方案以及融合之后的模型进行测试，模型效率（表 4）及得分（表 5）如下。可以发现，预训练采用 large 的效果要优于 base，任务层采用 CLS+MRC 和 NER+OQMRC 得分接近，模型融合较单模得分可以提升约 4 个百分点。

其他说明：由于复赛榜单和初赛相差过大，我们提供初赛模型评估结果。初赛参与时间较短，每类模型仅提交 1-2 次结果测试。

表 4 模型效率（2080ti）

模型	训练（约 2 万条）	推断（约 35 万条）
BERT-base	5 分钟	23 分钟
BERT-large	20 分钟	90 分钟

表 5 模型得分

预训练	模型方案	线上得分
RoBERTa-wwm-ext	CLS+MRC	0.739
RoBERTa-wwm-ext	NER+OQMRC	0.741
RoBERTa-wwm-ext (ensemble)	CLS+MRC	0.773
RoBERTa-wwm-ext (ensemble)	NER+OQMR	0.769

RoBERTa-wwm-ext (ensemble)	ALL	0.775
RoBERTa-wwm-ext-large	CLS+MRC	0.745
RoBERTa-wwm-ext-large	NER+OQMRC	0.744
RoBERTa-wwm-ext-large (ensemble)	CLS+MRC	0.775
RoBERTa-wwm-ext-large (ensemble)	NER+OQMR	0.776
RoBERTa-wwm-ext-large (ensemble)	**ALL**	**0.778**

5 结论

针对比赛任务，本文提出了基于“文本分类（CLF）+抽取式阅读理解（MRC）”、“实体识别（NER）+观点型阅读理解（OQMRC）”的联合训练方案，实现了从金融领域文本中抽取事件主体的需求。该方案的四个算法可以独立拆解使用，也可以根据业务场景自由组合，具有较高的灵活性和模型性能，能够胜任多种事件抽取场景的集成和部署。

金融领域的篇章级事件主体抽取，是 NLP 领域一个非常重要的课题，结合多种任务实现联合训练，兼顾效率和精度是业界的发展方向。我们将尝试更多的方案，不断提高模型的整体性能。

参考文献

1. Kaixuan Li, Xiujuan Xian, Jiafu Wang, et al. Neural Machine Reading Comprehension: Methods and Trends. Computation and Language. (2019)
2. Yibo Sun, Daya Guo, Duyu Tang et al. Knowledge Based Machine Reading Comprehension. Computation and Language. (2018)
3. Yizhong Wang, Kai Liu, Jing Liu, et al. Multi-Passage Machine Reading Comprehension with Cross-Passage Answer Verification. ACL 2018. (2018)
4. Rajarshi Das, Shehzaad Dhuliawala, Manzil Zaheer, el al. Multi-step retriever-reader interaction for scalable open-domain question answering. arXiv preprint arXiv:1905.05733. (2019)
5. S. Thenmalar, J. Balaji and T.V. Geetha. Semi-supervised Bootstrapping approach for Named Entity Recognition. International Journal on Natural Language Computing (IJNLC). (2015)
6. Xiaoya Li, Jingrong Feng, Yuxian Meng, et al. A Unified MRC Framework for Named Entity Recognition. ACL 2020. (2020)
7. Zhengbao Jiang, Wei Xu, Jun Araki and Graham Neubig. Generalizing Natural Language Analysis through Span-relation Representations. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. (2020)
8. Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, et al. Neural Architectures for Named Entity Recognition. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. (2016)
9. Minghao Hu, Furu Wei, Yuxing Peng, et al. Read + Verify: Machine Reading Comprehension with Unanswerable Questions. AAAI-19. (2018)

10. Minghao Hu, Yuxing Peng, Zhen Huang, et al. Reinforced Mnemonic Reader for Machine Reading Comprehension. 27th International Joint Conference on Artificial Intelligence (IJCAI). (2018)
11. Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL .(2019)
12. Cui, Yiming and Che, Wanxiang and Liu, Ting and Qin, Bing and Wang, Shijin and Hu, Guoping: Revisiting Pre-Trained Models for Chinese Natural Language Processing. Findings of EMNLP .(2020)