基于 NEZHA 与 MRC 的命名实体识别方法

陈东来, 何锦源, and 王震

联易融数字科技集团有限公司，广东深圳
{chendonglai,hejinyuan,wangzhen}@linklogis.com

摘要 事件抽取是舆情监控的重要任务之一。舆情监控整合互联网上的公开信息进行爬虫抓取，进行智能分类。舆情监控可以跟踪特定的企业，行业，产品的新闻热点，形成图表，简报或报告，为舆情分析决策提供依据。本次竞赛任务为篇章级事件元素抽取，要求从给定的长文本中抽取金融诈骗事件的 13 个要素。我们用 NEZHA 模型和阅读理解模型得出预测值，对多个模型的结果进行投票，取得 B 榜 0.7726 的结果。

关键词 命名实体识别 · NEZHA · MRC · BERT.

1 任务定义和数据集

事件抽取是舆情监控的重要任务之一。舆情监控整合互联网上的公开信息进行爬虫抓取，进行智能分类。舆情监控可以跟踪特定的企业，行业，产品的新闻热点，形成图表，简报或报告，为舆情分析决策提供依据。

本次竞赛任务为篇章级事件元素抽取，数据来自互联网上公开的新闻，公众号，论坛以及法院判决书。本次竞赛要求从给定的长文本中抽取金融诈骗事件的 13 个要素，主要包括嫌疑人，受害人，案发城市，案发时间，资损金额。每条数据还提供三级事件类型供参赛者参考。

本次任务可以当作命名实体识别（Named Entity Recognition，简称 NER）。即在文本中标注出感兴趣的实体，比如人物，公司，地名，机构名等。但与一般命名实体识别不同的是，模型需要理解上下文的关系，比如文本中有多个人物，时间，地点，需要判断哪个是嫌疑人，哪个是受害人，哪个是案发时间，哪个是案发地点，哪些是无关的实体。

2 评价指标

本次竞赛采用 F1 值作为评价指标，即精确率 (Precision, P) 与召回率 (Recall, R) 的调和平均。其中精确率为识别正确的个数与总共识别数之比。召回率为识别正确的个数与标准答案数之比。只有当识别的实体的名称，种类与标准答案相同，而且识别的位置在文本中确实对应该实体，则记为识别正确。写成数学公式，即为：

$$F1 = \frac{2PR}{P+R} \tag{1}$$

3 训练数据集统计

竞赛数据的特点是文本长度比较长。带标签的文本或称为正文本，有 3869 条，长度的中位数为 451，有极少数长度可达 3876。不带标签的文本，或称为负文本，有 1131 条，长度的中位数为 63，有极少数长度可达 2886。具体统计数据见表 1。可以看到正文本的长度总体上比负文本的长。

训练集中各种实体出现的次数为：资损金额 6726 次，支付渠道 3915 次，受害人 2686 次，嫌疑人 2237 次，案发时间 2099 次，案发城市 1867，涉案平台 1000 次。剩下出现次数比较少的实体为：受害人身份 297 次，银行卡号 20 次，手机号 13 次，交易号 3 次，订单号 2 次，身份证号 2 次。假设训练集与验证集的分布相似，那么后面 5 个实体在验证集中出现的次数极少。这五个类别没有足够的数据去训练，预测的准确度不高。本次竞赛的训练集主要分为两类事件，欺诈风险和盗用风险。其中盗用风险占 2747 条，欺诈风险占 2209 条。比较明显的特点是，如果一个事件是亲属盗用了受害人的账号并进行消费，那么嫌疑人可能不出现在文中。

概率值	所有文本长度	正文本长度	负文本长度
0.01	10	22	5
0.05	23	38	11
0.1	34	65	16
0.2	68.8	119	24
0.3	118	186	33
0.4	182	286	44
0.5	297	451	63
0.6	496.4	634.8	96
0.7	714.3	826.6	136
0.8	986	1113	269
0.9	1564.2	1708.4	803
0.95	2165.3	2302.6	1418
0.99	3732.15	3876.36	2886.8
总数	5000	3869	1131

Table 1. 文本长度统计

4 方法概述

我们的方法如图 1所示。我们首先对文本更行预处理成能被 NEZHA 模型读入的格式。然后我们利用 NEZHA 模型和阅读理解模型得出预测值。最后我们将预测结果进行投票并利用规则剔除一些预测得到最终的预测结果。我们根据预测的提交结果对模型进行调整。经过不断的提交结果和模型评估，我们选出了最合适的模型与模型融合的方式。

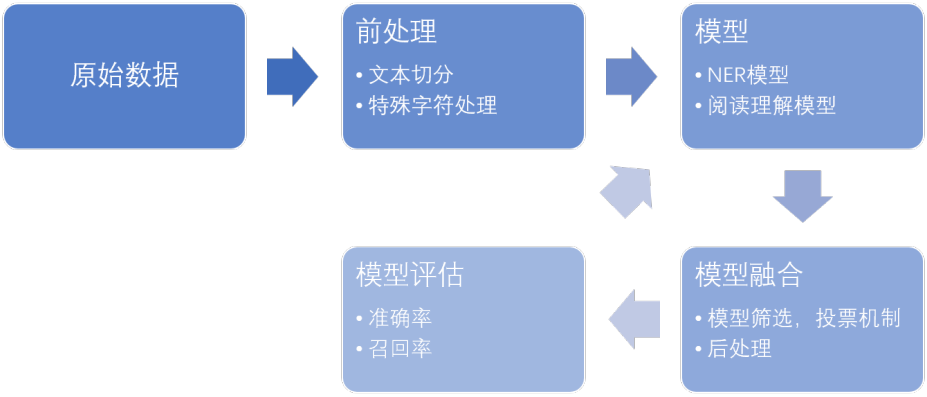

Fig. 1. 数据建模流程

5 文本预处理与后处理

我们采用滑动窗口来处理长文本。因为 NEZHA 等预训练模型最大能处理文本长度为 512，我们将训练集划分为若干个长度为 512 的片段，片段之间间隔 300 个字符。我们将数据转换成 BIO 格式，即要素的起始位置标 B，要素的其他位置标 I，要素之外的字符标 O。如果一个要素被滑动窗口截断，则这个要素标 O。我们也使用过不同的间隔长度，以产生多样性的预测结果。我们对验证集也做同样的处理。我们对切分的验证集作出预测之后再将预测结果拼接在一起，作为单个模型提交的结果。在拼接的过程中，我们对嵌套的实体进行处理。即对两个位置重叠的实体，一般取最长的实体。我们也对训练集中实体名称与位置对不上的进行了手工修改。我们将比赛的数据 json 格式转换成 NEZHA 读入的格式并删除了特殊字符。

6 序列标注模型

6.1 条件随机场

条件随机场 [1]（Conditional Random Fields，CRF）是一种无向概率图模型，属于判别式模型，广泛用于分词，词性标注和命名实体识别等任务中。在本任务具体指的是线性链条件随机场。模型结构如图 2所示。其中 $x = \{x_1, x_2, \ldots, x_n\}$ 是长度为 n 的字符序列，x_i 是第 i 个字符的特征所组成的向量。其中 $y = \{y_1, y_2, \ldots, y_n\}$ 是标签序列，在本任务中是实体的类别标签。条件随机场的联合概率可以写成

势函数乘积的形式。此模型在 NER 任务中用于寻找概率值最高的标签序列，用于学习标签之间的序列相关性。可以直接用 Bert 或 Nezha 等模型输出 NER 结果，但是常常有一些不符合 BIO 格式的结果出现。CRF 用于去除那些不符合 BIO 格式规范的输出结果，提高准确率。

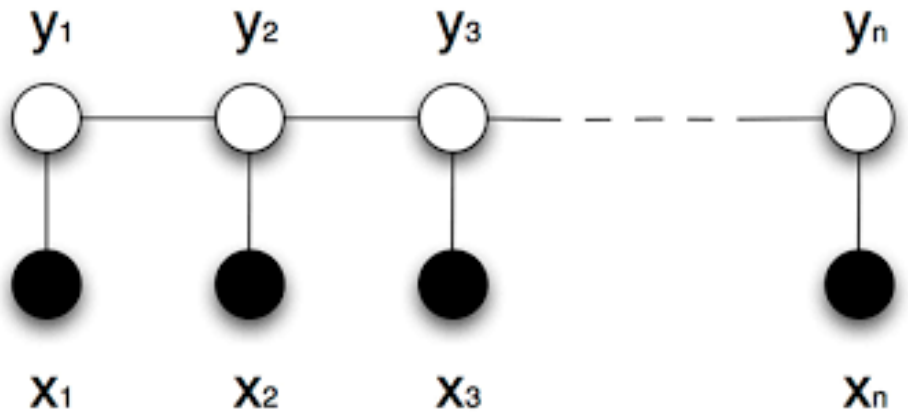

Fig. 2. 条件随机场

6.2 BERT

BERT（Bidirectional Encoder Representation from Transformers）[2] 模型由 Google AI 团队提出，是一个利用海量语料进行预训练得到的模型，用于各种下游任务，包括 NER，文本分类，阅读理解，文本相似度。此模型在各种自然语言处理任务上取得了优秀的成绩，被誉为划时代的工作。BERT 主要的原理是训练模型来随机遮盖词从而学习词之前的语义关系，以及句子对任务来学习上下句的关系。在此基础上，研究者们提出了 BERT 的各种改进版本，包括 NEZHA 和 Roberta。

6.3 NEZHA

NEZHA 全称为 NEural contextualiZed representation for CHinese lAnguage understanding[3]，是一种基于 Transformer 的预训练模型。NEZHA 大体结构上与 BERT 相似，也是一种预训练语言特征表示模型，可以利用这个模型完成各种下游任务。BERT, NEZHA 等预训练模型使用无监督学习，训练任务为预测被遮盖的词的信息。这一类的预训练模型可以利用海量的无标注互联网语料，只需要少量有标注的语料进行微调即可完成模型训练。它对 BERT 模型进行如下改进：

- 增加相对位置编码函数
- 全词掩码

- 混合精度训练
- 利用 LAMB 优化器进行预训练

其中增加相对位置编码函数是 NEZHA 的一大亮点。BERT 模型有绝对位置编码，但是很多时候数据的长度太不到模型的最大长度，因此靠后位置的位置向量得不到充分的训练。NEZHA 模型考虑各 token 之间的相对位置关系，可以更好地学习字符之间的相互关系。我们选择了 NEZHA large wwm 是使用了 whole word masking 的预训练版本，即预训练时对全词进行 mask。

6.4 RoBERTa

RoBERTa[4] 是一个被调整到最优的 BERT，主要的改进如下：

- 修改 Adam 的超参数
- 加入混合精度
- 加大 Batch size
- 在更长的序列上训练，移除 NSP，预测下一个句子的任务
- 把 BERT 静态遮掩改为动态遮掩
- 增大语料库

RoBERTa 在各项自然语言处理任务中取得了优秀的成绩，也证明了基于 BERT 的语言模型的有效性。

6.5 双向 LSTM

LSTM 的全称是 Long Short-Term Memory[1]，是一种序列模型。它的特点是对每一个结点计算有用的信息，同时丢弃无用的信息。这使它可以更好地捕捉到较长距离的依赖关系。但是单个 LSTM 只能捕捉单向的依赖关系。BiLSTM 由前向 LSTM 和后向 LSTM 组成，在自然语言处理任务中，广泛用于学习序列的依赖关系，尤其是命名实体识别任务。

6.6 R-Transformer

R-Transformer[5] 在 transformer 的基础上加了 RNN 来处理位置的信息，它把每三个词接入一个 RNN 获得隐藏状态，然后输出到多头注意力层，这与 Transformer 中的处理一致，最后经过全连接层。

6.7 模型

我们利用以上模型作为组件，建立以下模型：

- NEZHA large wwm
- NEZHA large wwm + CRF
- NEZHA large wwm + BiLSTM + CRF
- RoBERTa large wwm + R-transformers + CRF
- RoBERTa large wwm + BiLSTM + CRF

其中一部分 NEZHA 模型在竞赛训练集和测试做了预训练，为了更好地学习语料的语义信息。

7 阅读理解模型

我们也可以把命名实体识别当成阅读理解模型 [6]。由于竞赛数据提供了每条文本的三级分类信息，我们可以利用这些信息当作阅读理解模型的输入，来取得答案的起始位置和终止位置。我们每次向模型提问以找出某种类型的实体。我们可以设定不同的阈值让模型舍弃一些置信度不高的答案。

8 调整损失函数

8.1 Focal loss

Focal loss[7] 用于解决分类问题中类别不平衡、分类难度差异大的问题。主要的思想是关注那些分得不准的样本而不必过于训练那些分类准确度高的样本。一般在二分类问题是的交叉熵损失函数是:

$$L_{ce}(\hat{y}) = -y \log \hat{y} - (1-y) \log(1-\hat{y}) \tag{2}$$

其中 y 是真实标签，$\hat{y}$ 是预测概率。

在 Focal loss 中，损失函数为:

$$L_{fl}(\hat{y}) = -y\alpha(1-\hat{y})^{\gamma} \log \hat{y} - (1-\alpha)(1-y)\hat{y}^{\gamma} \log(1-\hat{y}) \tag{3}$$

在很多任务中，负样本的个数比正样本的个数大得多，这种损失函数可以降低负样本权重, 让模型更关注正样本。按照原论文的实验, 一般取 $\alpha = 0.25$,$\gamma = 2$ 的时候效果最好。

8.2 对抗训练

对抗训练（adversarial training）[8] 是增强神经网络鲁棒性的重要方式。在训练中，对损失函数进行一些微小的扰动，让神经网络适应这种改变，从而增加模型的鲁棒性和泛化性。

9 多折训练

我们将 5000 条数据随机抽样成 5 个 1000 条数据，取 4 个 1000 条用于训练 NER 模型。剩下 1000 条用于评价模型的精确度。我们用这 5 个模型有验证集中得出预测结果。这样我们得到 5 个相似但略有差异的模型预测结果。我们做了两次抽样，最终每一类基础模型可以得到 10 个模型。

10 模型融合

经过不断的调试，我们选择了以下几个模型用于投票，包括 NEZHA large, Roberta 加对抗训练，NEZHA large 加 CRF 以及 MRC。我们采取投票的方法来融合不同的模型预测结果。我们记实体名与类别相同的为一票。我们调整不同的投票阈值来取得最优的结果。具体的步骤为：我们根据一个投票结果的准

确率和召回进行判断，调整投票的阈值，尽量使准确率和召回尽量相近，从而得到最优的 F1 值。我们通过验证集的指标得到每个模型对于某个特定类别的准确率。对于某个模型在某个类别准确率高的我们加大这个模型在这个类别的权重，反之我们减少权重或者剔除这个模型。对于位置重叠的实体，我们取较长的实体作为最终结果。投票的结果比单个模型在精确度与召回率方面有较大提升。

11 单模型和模型融合实验结果

11.1 A 榜实验结果

以下是我们各种单模型和模型融合的 A 榜实验结果。

模型	精确度 P	召回率 R	F1 值
NEZHA Base	0.6306	0.7180	0.6715
Roberta 加对抗训练	0.70431	0.71784	0.71101
NEZHA large	0.68879	0.73374	0.71056
MRC	0.71091	0.80589	0.75542
NEZHA large 加 CRF	0.72109	0.76198	0.74097
24 个模型投票	0.75089	0.79829	0.77387
27 个模型投票	0.76951	0.81206	0.79021

11.2 B 榜实验结果

我们模型融合的 B 榜实验结果如下表

模型	精确度 P	召回率 R	F1 值
30 个模型投票	0.76402	0.78142	0.77262

12 结论

我们采用了多种基于 NEZHA 的模型用来完成金融事件的要素抽取，取得 B 榜 0.77262 这个比较满意的结果。最终结果显示各个模型在不同类型的实体方面有所侧重。最终用于提交的结果中，多个模型融合的效果比较好。利用模型融合方法来吸收不同模型的优点并尽量减少错误的预测。本次结果说明，对抗训练与 MRC 的效果比较好。目前处理长文本的模型并非完美，未来的工作将针对如何更好的理解长文本及提取要素。

References

1. Zhiheng Huang, W. Xu, and Kailiang Yu. Bidirectional lstm-crf models for sequence tagging. *ArXiv*, abs/1508.01991, 2015.
2. J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.
3. Junqiu Wei, Xiaozhe Ren, Xiaoguang Li, Wenyong Huang, Yi Liao, Yasheng Wang, J. Lin, Xin Jiang, Xiao Chen, and Qun Liu. Nezha: Neural contextualized representation for chinese language understanding. *ArXiv*, abs/1909.00204, 2019.
4. Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692, 2019.
5. Z. Wang, Y. Ma, Zitao Liu, and Jiliang Tang. R-transformer: Recurrent neural network enhanced transformer. *ArXiv*, abs/1907.05572, 2019.
6. Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. A unified mrc framework for named entity recognition. In *ACL*, 2020.
7. Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42:318–327, 2020.
8. Takeru Miyato, Andrew M. Dai, and I. Goodfellow. Adversarial training methods for semi-supervised text classification. *arXiv: Machine Learning*, 2017.