基于对比学习的通信领域事件共指消解模型

陆轩韬[1]，王欢[2]，夏茂晋[2]，余强[2]，黄金凤[2]，刘井平[1]，肖仰华[1]

[1] 复旦大学
[2] 北京清博智能科技有限公司
{xtlu20,jpliu17,shawyh}@fudan.edu.cn
{wanghuan,xiamaojin,yuqiang,huangjinfeng}@gsdata.cn

Abstract. 本届CCKS面向通信领域的过程类知识抽取的子任务之一是为通信领域中的事件共指消解问题提供解决方案。其具体目标为给定两个自由文本及其中的事件判定两个事件是否为同一个事件，即给定文本T1和T2以及T1中的事件e1和T2中的事件e2，判定e1和e2是否相等。我们分别尝试了Nezha[1]、Roformer[2]等预训练语言模型。此外，我们首次尝试使用当下最新的对比学习方法作为使模型适应领域数据的手段，最终得到的模型在测试集上达到了0.9102的F1-score，在评测中取得了A榜第一，B榜第二的成绩。

Keywords: 共指消解，预训练语言模型，对比学习

1 引言

1.1 任务背景

通信领域存在多种的过程类知识，如硬件安装（基站主设备安装操作步骤）、参数配置（配置网元开通与对接相关的参数）、集成调测（网元开通调试和功能验证）、故障处理（修复网元开通或正常运行中出现的故障）等，其中故障处理过程类知识尤为重要。通信运维过程中，通过“事件”及“事件关系”对故障过程知识进行梳理，给用户呈现故障发生的逻辑，提供故障排查和故障恢复方案，指导一线处理现网故障。在故障知识整理过程中，“事件共指消解”是实现故障脉络、排查步骤和恢复步骤梳理的重要手段之一。其难点在于事件元素表述多样化和事件元素缺损（漏抽取、文本描述缺损）。本次评测任务的语料来源主要是华为公司的公开故障处理案例。事件类型包括：指标恶化

类、软硬件异常、采集数据、核查类、配置类故障、外部事件、调整机器、操作机器等。

1.2 数据描述

本任务总共含有15000条训练数据，2000条A榜验证数据，29000条B榜测试数据（其中只有2000条参与结果评测，其余为干扰数据）。其数据样例格式如下：


```
{
    "pair_id": "1",
    "eventA": {
        "text": "PUCCH扩张失败导致RRC建立失败",
        "trigger": ["SoftHardwareFault", 14, "建立"],
        "argument": [
            ["Subject", 11, "RRC"],
            ["State", 16, "失败"]
        ]
    },
    "eventB": {
        "text": "图 1RRC建立失败（原因值为无资源导致）和小区最大用户数",
        "trigger": ["SoftHardwareFault", 6, "建立"],
        "argument": [
            ["Subject", 2, "1RRC"],
            ["State", 8, "失败"]
        ]
    },
    "label": true
}
```


图1 数据样例格式

其中，“pair_id”为事件对ID，“eventA”和“eventB”为事件的文本及其元素(包含text，trigger，argument三个字段)，label为标签（true表示共指，false表示非共指，验证集和测试集不包含此字段）。

2 模型及方法介绍

2.1 数据增强

数据增强的方法有很多，比较通用的方法包括回译、同义词替换等。但考虑到本次任务自身的特点，我们采用了如下两种数据增强方法：首先是对偶数据增强，由于eventA和eventB两个事件是可以交换顺序的，所以A-B pair可以变成B-A pair。其次是负样本采样，我们从两个不同的pair中各自抽取一个事件，组成一个新的pair，并认为该pair的label为false（即非共指），以此增加训练

数据中的负样本个数。需要注意的是数据增强的过程中要保证正负样本的比例不能过度失衡，否则可能会导致数据增强后的效果变差。

2.2 模型架构

我们使用当下最为流行的预训练语言模型作为主要架构。在具体使用时，有两种主流方法。一种可以称作cross-encoder类：即对一组句对中的两句话进行拼接后再编码，编码过程中可以进行句内及句间的信息交互。另一种可以称作bi-encoder类：即对一组句对中的两句话分别进行编码，再通过网络结构进行表示间的交互和计算，典型的模型有Siamese network[3]、Sentence-Bert[4]等。两种方法的图例如下：

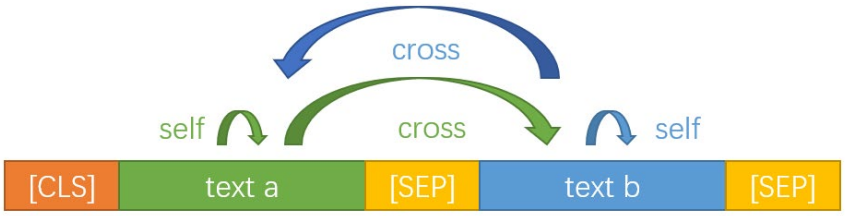

图2 cross-encoder类模型输入格式

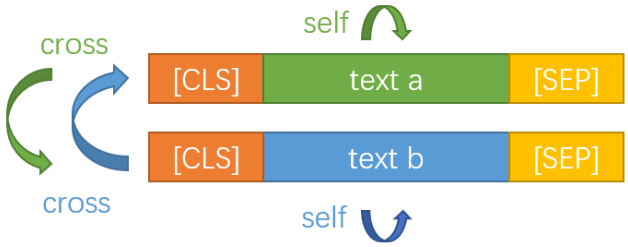

图3 bi-encoder类模型输入格式

通常来说，cross-encoder类模型具有更加充分的信息交互，所以会得到更好的效果，但是加剧了对文本长度的限制。而bi-encoder类模型具有更大的文本长度上限，而且可以通过离线储存文本表示，在线计算相似度的方式提升效率。考虑本次任务中的数据大多为短文本，故而我们采用cross-encoder类模型的方案来取得更好的效果，不同的数据字段使用[SEP]符号进行拼接后送入模

型。

在预训练语言模型的选择上，我们尝试了多种流行的预训练语言模型，并从中选出两个作为重点尝试。其一是Nezha，其二是Rotary Transformer（RoFormer）。Nezha相较于BERT的改进主要包括：相对位置编码、全词掩盖、混合精度训练和使用LAMB Optimizer进行参数优化。而RoFormer的主要改动则是其独特的旋转式位置编码方式。

在预训练语言模型后接结构的选择上，考虑到预训练语言模型本身很强的特征提取能力，因此后接结构应设计的尽量简单。我们尝试了以下五种不同的后接结构，如下所示：

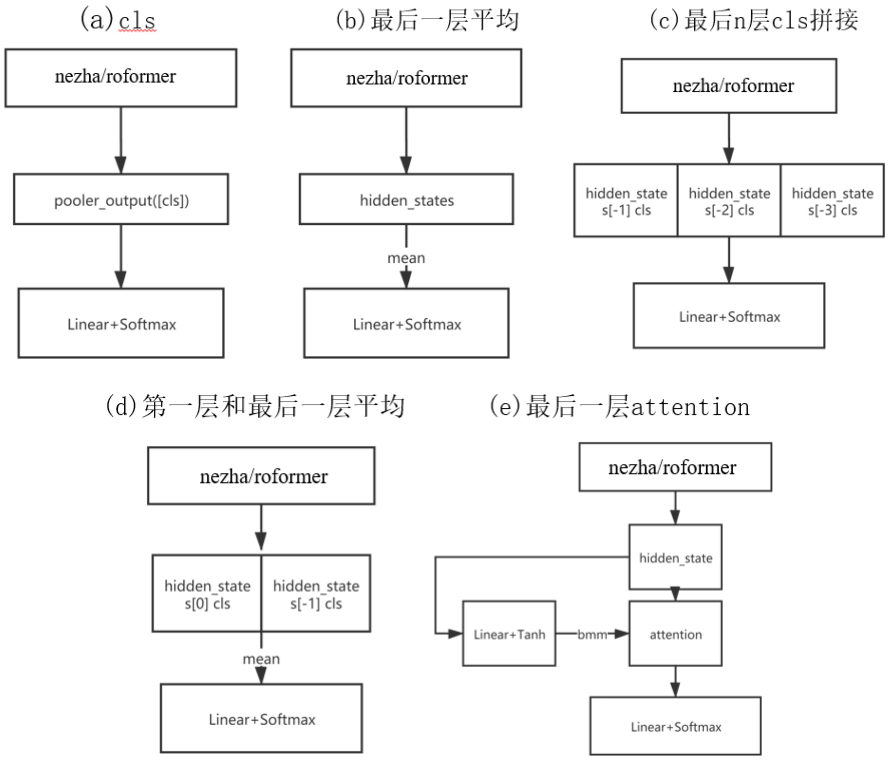

图4 预训练语言模型的不同后接结构

经过实验发现，五种不同后接结构的效果差距很小，最终实际采用的是表现最好的cls后接结构。

2.3 对比学习

对比学习在最近一年的迅速流行启发了我们尝试使用对比学习来进一步提升模型的效果。何恺明团队提出的MoCo[8]通过对比学习的方法，使得无监督学习在ImageNet上的分类效果超过了有监督学习。Ting Chen等人提出的SimCLR[9]与MoCo相比，关注的重点是正负样例的构建方式，同时还探究了非线性层在对比学习中的作用，并分析了batch size大小、训练轮数等超参数对对比学习的影响。

在众多的对比学习方法中，SimCSE[5]发现利用预训练语言模型中自带的Dropout作为增强手段得到的句子表示，其质量远好于传统方法，并在无监督语义任务上达到SOTA。具体来说，对同一个句子进行两次前向传播，由于模型中Dropout的存在，可以得到两个不同的embeddings向量，将同一个句子得到的向量对作为正样本对，对于每一个向量，选取其他句子产生的embeddings向量作为负样本，以此来训练模型，其训练损失函数如下：

$$l_i = -\log \frac{e^{\frac{\mathrm{sim}(h_i, h_i^+)}{r}}}{\sum_{j=1}^{N} e^{\frac{\mathrm{sim}(h_i, h_j^+)}{r}}} \tag{1}$$

其中，$\mathrm{sim}(h_1, h_2) = \frac{h_1^T h2}{\|h_1\| \cdot \|h_2\|}$用来度量两个embedding表示之间的相似度。

在以往的研究和比赛中，为了使预训练语言模型适应领域数据，通常采用的方式是在领域数据上继续做pre-training[6]，即mask一部分输入再预测出mask部分是什么，但是这种方式需要大量无标注的领域数据，在领域数据不充足的情况下往往会存在限制。我们想到，可以使用SimCSE替代传统的预训练方式，而且该方法是完全无监督的，不需要任何标注信息，故而可以同时利用训练集，验证集和测试集中的数据。

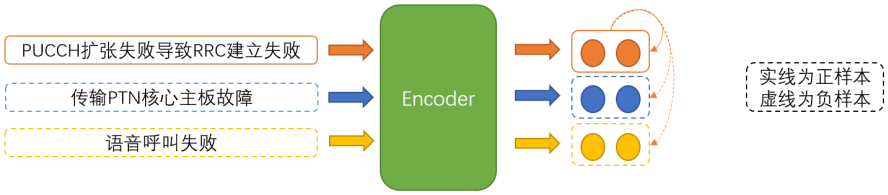

图5 使用SimCSE进行对比学习

2.4 训练技巧

为了完整的利用所有训练数据，我们采用了五折交叉验证的训练方式。此外，在训练过程中使用Fast Gradient Method (FGM)[7]进行对抗训练，即对embedding层在梯度方向添加扰动，通过这种方式对参数进行正则化，缓解模型鲁棒性差的问题，进一步提升模型的泛化能力。

3 实验结果及分析

我们的实验在Nezha-large和Roformer-base两个模型上分别展开，并逐步优化，具体优化方法在前文中已全部提及。主要分为以下几个阶段：

- 加入对偶数据
- 加入FGM对抗训练
- 五折交叉验证
- 加入负采样数据
- 使用SimCSE进行对比学习
- 模型融合

表1. 验证集(A榜)上不同模型和方法的实验结果（F1 Score）

模型	方法	F1 Score
Roformer-base	+对偶数据	0.7790
Roformer-base	+对偶数据、FGM	0.7846
Roformer-base	+对偶数据、FGM、五折	0.7871
Roformer-base	+对偶数据、FGM、五折、负采样	0.7914
Roformer-base	+对偶数据、FGM、五折、负采样、SimCSE	0.7945
Nezha-large	+对偶数据、FGM	0.7877
Nezha-large	+对偶数据、FGM、五折	0.7979
Roformer-base + Nezha-large	最佳模型融合	**0.8020**

表2. 测试集(B榜)上不同模型和方法的实验结果（F1 Score）

模型	方法	F1 Score
Roformer-base	+对偶数据、FGM、五折、负采样、SimCSE	0.9059
Nezha-large	+对偶数据、FGM、五折	**0.9102**

4 总结

本文提出的模型在本届CCKS面向通信领域的过程类知识抽取的子任务之一：通信领域中的事件共指消解任务上达到了0.9102的F1-score，在评测中取得了A榜第一，B榜第二的成绩。此外，我们首次提出可以通过当下流行的对比学习方法替代以往的In-domain pre-training（领域内预训练），以此缓解领域数据不足的问题，使得模型能够适应领域数据，并进一步提升模型在领域数据上的效果。

参考文献

1. Wei J, Ren X, Li X, et al. Nezha: Neural contextualized representation for chinese language understanding[J]. arXiv preprint arXiv:1909.00204, 2019.
2. Su J, Lu Y, Pan S, et al. Roformer: Enhanced transformer with rotary position embedding[J]. arXiv preprint arXiv:2104.09864, 2021.
3. Chicco D. Siamese neural networks: An overview[J]. Artificial Neural Networks, 2021: 73-94.
4. Reimers N, Gurevych I. Sentence-bert: Sentence embeddings using siamese bert-networks[J]. arXiv preprint arXiv:1908.10084, 2019.
5. Gao T, Yao X, Chen D. SimCSE: Simple Contrastive Learning of Sentence Embeddings[J]. arXiv preprint arXiv:2104.08821, 2021.
6. Gururangan S, Marasovic A, Swayamdipta S, et al. Don’t Stop Pretraining: Adapt Language Models to Domains and Tasks[J].
7. Miyato T, Dai A M, Goodfellow I. Adversarial training methods for semi-supervised text classification[J]. arXiv preprint arXiv:1605.07725, 2016.
8. He, Kaiming, et al. "Momentum contrast for unsupervised visual representation learning." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.
9. Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." International conference on machine learning. PMLR, 2020.