工业知识图谱关系抽取-高端装备制造知识图谱自动化构建

反卷局来围观

团队简介

本团队是一只由热爱算法、热爱竞赛的小伙伴们组成的NLP 团队，分别由来自重庆邮电大学的三位研二同学、一名大三同学以及来自哈尔滨工业大学的一名研一同学组成，团队曾获多项专业竞赛 top 奖项。

摘要

命名实体识别和关系抽取是知识图谱构建中两项重要的基础任务，自动构建高端装备制造业故障知识图谱对于实现高端装备制造的智能化检修和诊断具有重大意义。各种高端装备领域的故障案例文本是由业务专家或者专业维修人员撰写的描述相关设备异常、以及故障排查步骤的记录，该记录包括故障现象、故障原因、解决方法以及排故过程等，这些故障案例知识的利用受到数据结构化程度的影响，因而识别数据中的部件单元、性能表征、故障状态、故障检测工具等核心实体及其之间的组成关系至关重要。

本团队选用GP[1]和GRTE[2]两种差异性较大的模型对本赛题数据进行实体和关系联合抽取，主要选择了 RoBERTA[3]、NeZha[4]等最新的预训练模型。针对该赛题中文本长度不均匀、长文本较多这一点进行分析和处理。对外部 CCL 数据集进行合并处理转化为本赛题格式，并且加入 A 榜伪标进行训练。此外，本团队还使用了分层学习率，FGM[5]对抗训练、SWA[6]等技术增强模型的鲁棒性。采用差异化多级模型融合，分别对 GP 和 GRTE 选取不同的预训练模型，第一级进行概率融合，第二级进行投票。最后，将结果进行后处理过滤掉不可信三元组。我们的结果在工业知识图谱关系抽取任务中复赛成绩 0.6671，排名第二。

关键词

关系抽取，预训练模型，数据处理，差异化多级模型融合

1 赛题分析

1.1 任务解读

本次评测任务的数据采用人工标注和专家复核的方式，确保语料标注样本质量。本任务提供的训练数据集和评测数据集均为文本文件格式。

本任务需要从故障案例文本自动抽取 4 种类型的关系和 4 种类型的实体。关系类型为：部件单元的故障状态、性能表征的故障状态、部件单元和性能表征的检测工具、部件单元之间的组成关系。

1.2 测评标准

本次评测任务采用微 F1 值（micro-F1）来评估关系抽取效果。对于每一种关系，相关的定义如下：

识别关系的精确率 = 识别关系与标注相同的数量 / 识别关系总数量

识别关系的召回率 = 识别关系与标注相同的数量 / 标注关系总数量

关系抽取的 F1 = 2 * (识别关系的精确率* 识别关系的召回率)/(识别关系的精确率+ 识别关系的召回率)

识别关系与标注相同指两个三元组的 h.name、t.name、h.pos、t.pos 和 relation 都相同，即主体、客体、关系类型都需要识别正确。

最终结果 F1 定义为各个实体的 F1 的微平均

1.3 数据分析

以下是本团队的数据分析，主要从文本长度分布、数据标签分布方面进行分析。

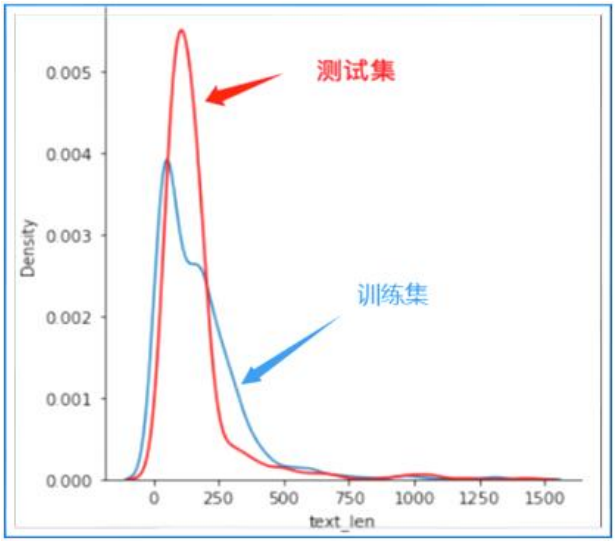

图 1：训练集和测试集长度分布

如图一所示，本赛题文本长度集中在 200 左右，不过有很多长度超过 BERT 等模型可接受的最大长度。

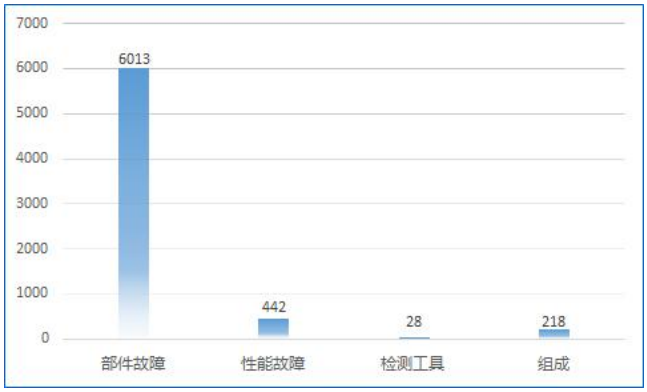

图 2：数据集标签分布

如图二所示，数据标签分布极不均衡，“部件故障”标签占比接近 90%，而“检测工具”只有 28 个。

在赛题进行过程中多了一个外部 CCL 数据集，该数据集跟此次赛题数据有一定的重合，但只有“部件故障”及“性能故障”两类标签，且文本长度短很多，标注质量欠佳，如何利用好该数据也十分重要，是一把双刃剑。

经过以上分析，本团队认为该赛题的难点有以下几点：

一是文本长度不均匀，长文本较多，怎样把长文本处理为短文本，同时不丢失过多的上下文信息十分重要。

二是该题数据量较少，对超参数很敏感，容易造成模型过拟合。

三是比赛过程中多了一个 CCL 外部数据，该数据质量较差且对 A 榜数据存在一定泄露，怎样将这份数据利用起来十分关键。

2 方案介绍

2.1 整体方案

本团队的最终解决方案，是由 5 个多折融合后的 GRTE 模型预测出的结果和 6 个多折融合后的 GP 模型预测出的结果投票而得，再加上针对本次赛题数据特点的后处理，整体流程如图所示：

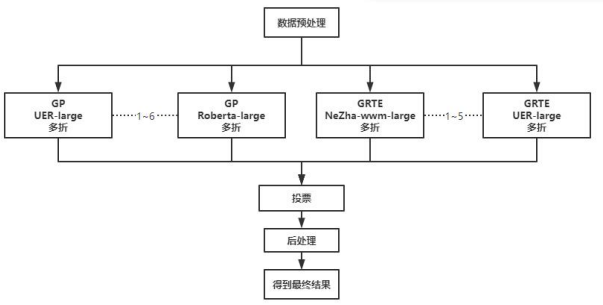

图 3：解决方案流程图

2.2 数据预处理

本赛题数据中文本长度不均匀，长文本较多，有的文本长度达到了 1000+，我们根据一些对三元组跨符号信息影响较小的标点符号对文本进行了拆分，然后将这些短文本按长度限制进行拼接，去除掉一些不合理和重复的文本，并限制原始数据切分后的若干数据同时出现在训练集或验证集，从而将长文本上下文信息的丢失降到最小。

2.3 GRTE 模型方案

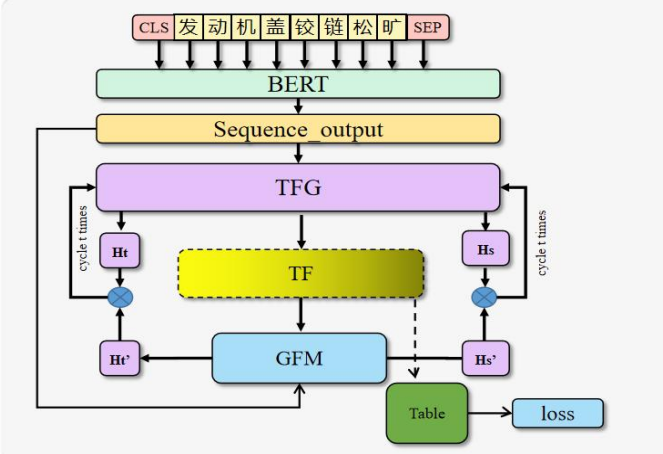

图 4：GRTE 方案

选取 Roberta-wwm-large、UER-large 和 Nehza-large 等预训练模型，共产出 5 个结果，其中 3 个加入 CCL 数据和 A 伪标数据，2 个只使用原始训练数据。采用多折交叉验证，每折结果进行 logits 融合。训练过程中使用了混合精度训练、分层学习率、梯度裁剪、EMA、SWA 等常见 trick，生成的 5 个结果用于后面和 GP 联合投票。

2.4 GP 模型方案

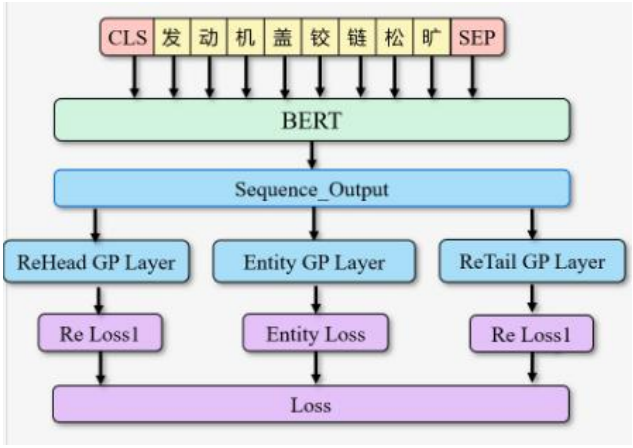

图 5：GP 方案

选取 Roberta-wwm-large、UER-large 和 Nehza-large 等预训练模型，共产出 6 个结果，其中 4 个加入 CCL 数据和 A 伪标数据，2 个只使用原始训练数据。采用多折交叉验证，每折结果进行 logits 融合。训练过程中使用了混合精度训练、fgm 对抗训练、EMA、SWA 等常见 trick，生成的 6 个结果用于后面和 GRTE 联合投票。

2.5 投票

对 GRTE 的 5 个结果和 GP 的 6 个结果进行投票，根据线上分数，阈值设为 5，使得三元组的召回高些，整体表现较好。

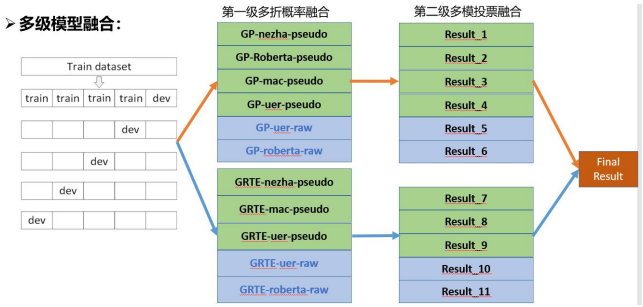

图 6：融合方案

2.6 后处理

此赛题数据特点明显，通过观察一些 Badcase 发现可以用规则过滤掉一些三元组，线上分数提升明显。具体方案如下：一是去除掉头实体和尾实体具有重叠文本的三元组。二是对于一条文本中识别出来的所有三元组，若存在两个三元组其中一边界相同而头实体重叠或尾实体重叠的情况，则过滤掉实体长度短的三元组。三是过滤掉实体存在空格的情况。其他的方案在投票前有一定效果，投票后基本规避掉了这些情况，这里不一一列举。

➢ **case 1：**

{'h': {'name': '变压器绝缘', 'pos': [119, 124]}, 't': {'name': '绝缘降低', 'pos': [122, 126]}, 'relation': '部件故障'}（去除此三元组）

➢ **case 2：**

{'h': {'name': '柱塞式喷油泵的柱塞弹簧', 'pos': [29, 40]}, 't': {'name': '断裂', 'pos': [43, 45]}, 'relation': '部件故障'}（去除此三元组）

{'h': {'name': '柱塞弹簧', 'pos': [36, 40]}, 't': {'name': '断裂', 'pos': [43, 45]}, 'relation': '部件故障'}（保留此三元组）

➢ **case 3：**

{'h': {'name': '电能表 RS-485 ', 'pos': [18, 30]}, 't': {'name': '无电压', 'pos': [34, 37]}, 'relation': '部件故障'}（去除此三元组）

图 7：结果后处理

3 其他尝试方案

（1）添加新词：该数据集中有部分 token 不在预训练模型的词表中，我们尝试加入新词，线上并无提升。

（2）Pipline 方案：我们尝试多一种异构方案加入后期融合，GRTE 和 GP 均使用联合抽取方式，我们先提取实体后提取关系，表现并不理想且加入融合没有什么提升。

（3）加权 loss：本赛题数据标签分布极不均匀，因此我们尝试对低频标签损失赋予更高的权重，防止模型过度倾向于学习高频标签样本，线上未提升。

（4）持续预训练：我们尝试了最基本的 MLM 任务对数据进行预训练，数据太少造成了过拟合，线上未提升。

（5）Prompt：尝试最新 NLP 范式 Prompt，使预训练和下游任务统一，该方案在线下有一定提升，线上无提升。

4 总结

此次比赛多了 CCL 外部数据，对 A 榜数据存在一定的信息泄露，因此 A 榜最终结果对 B 榜成绩参考意义不大，部分实验结果记录如表 1 所示：

表 1：部分实验结果记录表

操作	F1
改进 Baseline（数据处理）	0.660
FGM 对抗训练	0.665
EMA	0.672
增加训练轮数	0.675
更换 large 预训练模型	0.685
五折融合	0.708
增加 CCL 数据	0.746
后处理	0.749
与 GRTE 多模投票	0.7585
加入 A 榜伪标	0.7593

从表一可以看出，该赛题数据处理尤为重要，在增加 CCL 数据后，A 榜分数提升明显（存在一定泄露），我们采用两种不同模型来增加融合结构的差异性，增加 A 榜伪标质量，并且训练一些不加伪标的模型来防止过拟合，最终在 B 榜线上测评成绩 0.6671，居第二。我们的模型方案较大程度的缓解了模型容易过拟合的问题，效果好，泛化能力强。

致谢

感谢每个队友的辛苦付出！感谢老师的悉心指导！感谢本次大赛的所有工作人员！

参考

[1] 苏剑林. (Jan. 30, 2022). 《GPLinker：基于 GlobalPointer 的实体关系联合抽取》[Blog post]. Retrieved from https://spaces.ac.cn/archives/8888

[2] Feiliang Ren, Longhui Zhang, Shujuan Yin, Xiaofeng Zhao, Shilei Liu, Bochao Li, and Yaduo Liu. 2021. A Novel Global Feature-Oriented Relational Triple Extraction Model based on Table Filling. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 2646–2656, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

[3] Liu v , ott M , Goyal N , et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach[J]. 2019.

[4] Wei J, Ren x , Li x , et al. NEZHA: Neural Contextualized Representation for Chinese Language Understanding]. 2019.

[5] Miyato T, Dai A M， Goodfellow l . Adversarial Training Methods for Semi-Supervised Text Classification[C]// International Conference on Learning Representations. 2016.

[6] lzmailov P, D Podoprikhin,Garipov T , et al.Averaging Weights Leads to Wider Optima and Better Generalization[J]. 2018.