新冠知识图谱构建与问答评测子任务一：新冠百科知识图谱类型推断 评测报告

作者：王欢，李雨茗，夏茂晋，余强，王屹东

北京清博大数据科技有限公司
Email: {wanghuan; liyuming; xiamaojin; yuqiang; wangyidong}@gsdata.cn

Abstract. 本届CCKS新冠知识图谱构建与问答任务中的子任务一主要为新冠百科知识图谱构建中的实体类型推断问题提供解决方案。其具体目标为从给定的数据中推断相关实体的类型，即输入一系列实体名称，在指定类别范围中返回每个实体所对应的类别。我们使用基于roberta-large预训练模型的方法，并结合爬虫和相应人工技术对训练集进行扩充及修正，最终模型达到了0.997的F1-score，在测评中取得了第二名的成绩。

Keywords: 实体类推断, RoBERTa, 知识图谱.

1 引言

1.1 背景

随着互联网软硬件技术的飞速发展，人们逐渐从信息时代进入智能时代。知识图谱作为底层承载海量知识、支持上层智能应用的重要载体，在智能时代发挥着极其重要的作用。然而，由于非结构化文本与结构化知识的巨大差异，在自动构建知识图谱和利用知识图谱支持上层应用方面还存在许多挑战。因此，有必要对知识图谱的构建及其核心应用进行研究。时值2020年新型冠状病毒疫情爆发时期，愈来愈多的相关企业单位及院校使用自动化的技术，以新型冠状病毒为核心构建了包括新冠百科、健康、防控等多个高质量的知识图谱。通常构建知识图谱需要在命名实体识别任务之后，为每一个实体分配预定义的类型，这个过程被称作“实体类推断”。实体类推断广义上指识别出给定实体所属类别的一项自然语言处理任务，是信息抽取与知识图谱构建的重要环节之一。

1.2 任务

类型信息在知识库中具有非常高的价值，实体类型推断的研究一直是领域的热点。然而，大量类型信息以非结构化文本形式呈现于网络页面中，文本处理难度大，抽取结果同时保证高准确度和覆盖率仍然是个极大的挑战。针对实体的通用类型推断，近年来已有若干解决方案，如使用统计机器学习方法及利用外部知识（通向其他数据源的链接或文本信息）等。

本评测任务围绕新冠百科知识图谱构建中的实体类型推断（Entity Type Inference）展开。评测从实体百科（包括百度百科、互动百科、维基百科、医学百科）页面出发，从给定的数据中推断相关实体的类型。即输入为需要推断类型的相关实体，其中包括噪音（即不属于任何限定类型的实体），最终输出为实体所属的类别。

1.3 数据描述

本任务总共包含17.5W左右的实体百科页面。其中训练集中实体分为七类，分别是病毒、细菌、疾病、药物、医学专科、检查科目、症状，以及未定义类NoneType。
其部分数据样例如表一所示：

目标类型	样例1	样例2	样例3
病毒	EB病毒	流感病毒	意大利蝗痘病毒
细菌	红球菌属	门多萨假单胞菌	芽孢杆菌
疾病	史密斯骨折	急性肺水肿	多形性腺瘤
药物	消旋卡多曲	肝康颗粒	噻乙啶
医学专科	精神心理科	口腔科	消化内科
检查科目	棘形红细胞	53蛋白	T8淋巴细胞亚群
症状	高热寒战	肝内管梗阻	凹陷瘢痕
NoneType	上海精神医学	N95型口罩	公共卫生事件

表一：原始数据集目标类型及样例展示

在此基础上，我们通过在OpenKG，百度百科，互动百科，医学百科，39健康网和有问必答网站的爬虫对原始实体的描述进行扩充，其中扩充数据案例如表二所示。最后将两部分数据整合后作为整体训练资料。

目标类型	数据样例
病毒	EB病毒（Epstein-Barrvirus，EBV）是疱疹病毒科嗜淋巴细胞病毒属的成员，基因组为DNA。EB病毒具有在体内外专一性地感染人类及某些灵长类B细胞的生物学特性。人是EB病毒感染的宿主，主要通过唾液传播。无症状感染多发生在幼儿，3～5岁幼儿90%以上曾感染EB病毒，90%以上的成人都有病毒抗体。EB病毒是传染性单核细胞增多症的病原体，此外EB病毒与鼻咽癌、儿童淋巴瘤的发生有密切相关性，被列为可能致癌的人类肿瘤病毒之一。目前所测EB病毒抗体，主要有针对病毒的衣壳抗原（CA）、早期抗原（EA）和核抗原（EBNA）。
细菌	鼠伤寒沙门菌（S.typhimurium）属多价O抗血清的B群，也是一种临床较常见的沙门菌[1] 。鼠伤寒沙门菌是一种重要的人畜共患病原菌，其感染发病率居沙门菌感染的首位，约占人源沙门菌感染的40%～80%。多见于婴幼儿，可导致医院感染和暴发性食物中毒，病死率较高。

疾病	骨旁骨肉瘤起源于骨周围的骨膜，可向骨外生长，包绕骨干。本病较为罕见，可发生在任何年龄，但30岁以上患者多见，男女发病率相似。临床表现为无痛性肿块生长缓慢，可反复发作，晚期有转移。治疗以手术切除加化疗为主。
药物	消旋卡多曲是一个脑啡肽酶抑制剂，为白色或类白色结晶性粉末，可选择性、可逆性的抑制脑啡肽酶，从而保护内源性脑啡肽免受降解，延长消化道内源性脑啡肽的生理活性。
医学专科	脑外科一般指的是神经外科，是利用神经外科学，并以检查为主，手术为径，综合治疗，全面评估的学科。医学中最年轻、最复杂而又发展最快的一门学科。以前由于技术的限制，人的脑部手术可以说是一个禁区，然而随着科技的发展，已经可以借助先进的显微外科设备开展各种显微神经外科手术。
检查科目	尿沉渣管型检查，尿沉渣管型检查是尿沉渣检查的内容之一。管型是蛋白质在肾小管内凝聚而成的，尿出现管型一般是肾实质病变的证据，在其形成的过程中，若含有细胞，则为细胞管型;如含退行性细胞碎屑，即为颗粒管型;若含脂肪滴，则为脂肪管型。
症状	心肌灰白而松弛是心肌损害的一种，可能与病毒感染后发生的免疫性心肌损害有关，一般见于扩张型心肌病的超声检查。可作为与其他心肌病的鉴别诊断。
NoneType	《上海精神医学》杂志创刊于1959年，1989年正式公开发行，是国内第一本精神科专业的学术期刊。

表二：补充数据集及样例展示

2 模型及方法介绍

2.1 数据补充

由于数据本身只包含实体信息，因此我们分别通过对原始实体使用爬虫进行描述性补充以及查询OpenKG网站开源数据新冠病毒百科类数据进行补充，最终两种方式数据混合后，进而使用roberta_large预训练模型。

2.2 预训练模型

因为本次任务是实体类型推断，我们采用roberta_large[1]序列分类的下游任务模型，构建实体的序列，在处理为张量后，进行10折交叉验证，形成多个模型，以多个模型预测的结果进行加权处理，形成综合的result结果。

RoBERTa (Robustly Optimized BERT Pretraining Approach) 模型是BERT[2]的改进版，其建立在BERT的语言掩蔽策略的基础上，修改BERT中的关键超参数，包括删除BERT的下一个句子训练前目标，以及使用更大的bacth size和学习率进

行训练。RoBERTa也接受了比BERT多一个数量级的训练，时间更长。这使得RoBERTa表示能够比BERT更好地推广到下游任务。

为了进一步提升中文自然语言处理任务效果，我们在此使用哈工大讯飞联合实验室、认知智能国家重点实验室发布RoBERTa-large的中文预训练模型。

2.3 数据增强

数据增强的方法有很多，传统的方法有翻译回译方法、同义词替换方法等，本组采用的数据增强的方式类似同义词替换的思路，只是获取同义词的方式不同于以往的词典或者Word2vec模型的方式，而是采用BERT模型天生的MaskedLM能力，对文本中的Token随机进行遮挡预测，并选取可能性最大的两个预测结果替换原文中的Token，最终从一个文本中获得多个生成文本，并控制总体的数据比例，减少数据不均衡带来的影响。但从实验结果来看，这样的技巧没有为我们带来提升，分析增强的数据可以看到，BERT预测出来的字不是很符号语言规律，融合在原文本中使原来的句子变得晦涩难懂。

2.4 loss权重

训练集数据分布不均衡，针对样本不均衡，训练时针对不同的标签loss设置权重，不断迭代调整权重值得到最优组合，class_weight：[1,1,1,2,1.5,1.5]。

2.5 其他技巧

除此之外，在用不同预训练模型训练的时候，对原样输入做了不尽相同的预处理，例如微博昵称的去除和文本内网址的去除。具体包括哪些处理手段可参考训练代码。

为了完整的利用所有训练数据，我们采用了10折交叉验证的训练方式，同时也减少了预测结果的波动。

在检查数据的时候，发现训练集和测试集之间有数据泄露，对于泄露的数据，我们选取了训练集的标注结果对预测结果进行修正，对最终的测试集的分数也带来了一定提升。

对于模型融合，本组认为相同的预训连模型的概率输出具有一致性，应该采用概率加权的方式融合模型，而对于不同预训练模型之间，相同的损失函数带来了预测结果的一致性，应该采用投票的方式进行融合，且考虑每个模型在验证集上的得分当做投票权重。最终融合结果在验证集上取得了远超单模型的得分，具体提升见下表。

3 实验结果及分析

我们将实验分为5个阶段逐步优化，具体优化过程体现在训练集数量扩充和label修正上，并结合人工检查使之更为准确。其分为以下5个大阶段：

- 基础base阶段：无添加信息的train与test
- 加载描述信息阶段：训练集以实体文本文件开始，从官方给的百科信息中添加描述信息，再加上 xml提取+人工补充的5000条训练集，以及在百度百科与互动百科添加描述信息，构建第一版数据。
- 加载OpenKG新冠百科阶段：通过引入外部数据OpenKG上新冠百科类数据，对训练集进行补充，加载到训练集。
- Loss权重阶段：训练集数据分布不均衡，针对样本不均衡，训练时针对不同的标签loss设置权重，不断迭代调整权重值得到最优组合。
- 模型融合阶段：将模型进行微调和修正，以得到最终模型。

其中每个具体阶段的F1如表3所示：

实验阶段	F1-Score
基础base	0.946
加载描述信息	0.949
OpenKG新冠百科	0.995
Loss权重	0.996
模型融合	0.997

表三：实验阶段及每阶段结果

由表三可知，随着训练数据数量的扩充和描述信息的丰富，结果呈现递增趋势，加上人工修正和模型微调，最终达到0.997的F1-score。

4 总结

本次测评的任务在全球疫情爆发的大环境之下具有很高的实际价值，将人工智能相关技术落实到实际任务之中，既能以一种高效且智能的方法对疫情相关问题提出合理且可行的解决方案，又是机器学习相关算法模型在实际应用层面上的证实。我们针对新冠百科知识图谱类型推断任务提出了一种基于roberta-large预训练模型的方法，并结合爬虫和相应人工技术对训练集进行扩充及修正，最终模型达到了0.997的F1-score，在测评中取得了第二名的成绩。

Reference

[1] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

[2] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.