一、专业实践训练整体情况

<table>
<tr><td>实践单位
名称</td><td colspan="3">阿里健康科技(杭州)有限公司</td></tr>
<tr><td>实践单位
地点</td><td colspan="3">浙江省杭州市余杭区五常街道文一西路 969 号 6 幢 2 层 213 室</td></tr>
<tr><td>实践岗位
名称</td><td colspan="3">算法专家</td></tr>
<tr><td rowspan="2">专业实践
训练时间</td><td rowspan="2">集中进行</td><td>2020 年 10 月 01 日开始　至　2021 年 12 月 31 日结束</td></tr>
<tr><td>专业实践训练累计　456　天（单位考核前），其中项目研究天数　456　天（单位考核前）</td></tr>
<tr><td colspan="4">（1）基本概况（含实践单位简介、实习实践内容等）

阿里健康是阿里集团在大健康领域的科技公司，主要业务集中于医药电商、互联网医疗、消费医疗、智慧医疗等领域。我在阿里健康算法部门参与医学知识图谱的构建和应用，主要负责医学实体识别的应用和研究。</td></tr>
<tr><td colspan="4">（2）项目研究概述（含项目名称、项目来源、项目经费、主要研究目标和技术难点等）

本次实践的项目是“知识图谱构建”，项目来源阿里健康公司。主要目标是通过构建医学知识图谱并将其应用在互联网问诊中，辅助用户全面描述病情，提升在线问诊质量。其中，主要的技术难点是医学知识图谱的构建。</td></tr>
</table>

（3）项目开展情况（含项目研究内容、研究方案及技术路线，研究团队分工、本人承担任务及完成情况，存在问题与改进建议等，不少于 500 字。）

本研究的主要内容是构建中文医学知识图谱，整体方案采用人机结合的有效方式，即通过算法自动构建为主、人工质检为辅的人机结合方式，高质量、高效率地构建知识图谱。 整个技术方案分为知识处理、知识融合、质量检测三大模块。其中，知识处理又细分为实体识别、关系抽取、实体对齐等技术点，我主要负责命名实体识别技术的研究与应用。命名实体识别（Named Entity Recognition，NER）是自然语言处理领域重点研究方向之一，也是构建知识图谱必不可少的组成部分。支撑着下游关系抽取、实体链接等任务，处于构建中文医疗知识图谱的基础地位。命名实体识别任务一直存在几个难解决的问题：实体难定义、实体有歧义、边界难确定、数据难标注，在医疗领域下这几个问题尤为突出。 在医疗领域中，由于医疗术语词语义丰富、表述结构复杂，实体词的边界有时候会存在交错的情况，根据边界的交错情况，我们将这种情况称之为复杂粒度的实体识别，如：“咳黄痰”要识别为“咳痰”和“黄”，“饮食睡眠尚可”要识别为“饮食尚可”和“睡眠尚可”。常规解决细粒度实体识别的方法是将其转化为序列标注的问题，但是这种方法对于复杂粒度实体的识别的效果有限。我们借助 MRC 的思想，通过构建实体起始、实体终止索引矩阵、实体类别片段矩阵这三个任务，构建多任务学习的方法，同时在文本字面信息上引入词法信息，成功的解决了复杂粒度实体识别问题。其中这种字词混合训练学习的方法，经过验证效果均有提升。

二、专业实践训练收获

（一）围绕考核评价指标体系，举例说明以下收获（不少于 800 字）

通过参与知识图谱构建项目，我对于命名实体识别技术有了系统而全面的了解。命名实体识别是自然语言处理领域重点研究方向之一，也是构建知识图谱必不可少的组成部分。在任务启动的初期，我对于医疗实体识别认识不够深刻。通过大量阅读文献资料，同时对医疗数据进行统计分析，逐渐了解了命名实体的发展和医疗命名实体识别的不同之处。早期命名实体识别基于词典和规则，但是这种方式不具备可迁移的特性；之后，基于传统机器学习的方法，NER 被当做序列标注任务去解决。采用的方法主有:隐马尔可夫模型(HMM)、最大熵马尔可夫模型(MEMM)、条件随机场 (CRF)等；随着深度神经网络的快速发展，NER 的研究重点放在深度神经网络上，比较经典的方法有 LSTM-CRF、B1LSTM-CRF、BERT、BERT-CRF、BERT-MRC 等。通常我们所接触到的场景以细粒度实体识别为主，但是由于医疗领域属于专业、表述结构复杂，实体词的边界有时候会存在交错的情况，实体难定义、实体有歧义、边界难确定这几个问题相比细粒度实体识别而言更为突出，因此我的主要挑战目标是解决复杂粒度的实体识别。实体识别的主要解决方法有序列标注、片段排列、多头标注和指针网络。通过构建索引学习和类别学习的多任务学习方法，将片段排列和指针网络相结合，同时融入词法特征语义向量，最终解决了这个问题，并且验证有效。

通过参与本次项目的研究，对于 NLP 有了更深的了解，算法设计、开发水平有了一定的进步，对于快速掌握一个新领域的能力也有了一定提升的。在开展一个我们不熟悉的技术项目时，可以在期刊、顶会中找寻近几年相关的论文进行阅读，借助参综述性的文献以及引用次数多的文献可以快速了解技术项目的大致轮廓。接下来对学习到的知识进行应用，设计初步的技术方案，把我们要实现的项目拆解成多个技术点，对每个技术点的背景和学术界的效果要有了解，搞清楚每个技术点的难点在哪里，攻克的成本有多大。接下来调整技术方案，拆解成更鲁棒性更强的技术方案，同时思考我们的方案和现有研究方案不同之处在哪里，我们改进了什么。如此，可以更好的借助学者们宝贵的实践经验，更快、更对的达到我们的目标。

通过参与本次项目研究，自身素质也有了很大提升。首先对于复杂技术项目的把控性有了认识，能够在今后工作中更加沉稳一些；工作中经常对技术沟通交流，锻炼了自己的逻辑能力和表达能力；同时，团队伴们对负责的事情一丝不苟的态度也深深的感染着我，让我能够持之以恒坚持做对的事情，并且保持对学术研究严谨、实事求是的态度，在学习和生活中真正的做到了：踏踏实实做人，认认真真做事。

（二）取得成效

随着人工智能与互联网技术的迅猛发展，在线问诊这种新的互联网就医形式逐渐被人们认可。目前各互联网公司在线问诊功能均会要求患者在问诊前对当下的病情进行简要的描述，但是由于医学极具专业性，患者很难清晰、准确、全面的描述自身病情，因此，医生仍需再次询问患者相关情况。医生再次询问的过程中，有时会忽略患者已经描述过的部分信息，这样会造成患者已经描述过的信息还需要再次被询问，用户体验不够好；医生耗费较大精力阅读患者不全面、不准确的病情描述却没有获得有效输入。因此，我们希望辅助患者全面、准确地书写自身症状情况，同时去除病例描述中无效信息，把关键的医疗信息从非结构化的文本中提炼出来,清楚直观地展示给医生。因此，我们需要构建知识图谱，同时构建医疗命名实体识别的能力。

构建医疗知识图谱确实不容易，从海量的文献、病历、百科数据中挖掘医疗实体、实体关系是一个庞大的工程，需要大量的医学专家耗费极大的精力来完成。我们将整体技术方案划分为知识处理、知识融合和质量检测三个部分。知识处理是构建知识图谱的基础，而命名实体识别是高效率知识处理必不可少的组成部分，也支撑着下游关系抽取、实体链接等任务，同时还需要为线上的应用提供服务。借助 NLP 领域的命名实体识别技术来对海量非结构化数据进行自动挖掘，极大地提高了生产效率，减轻医学专家的工作量。我们从医疗场景出发，分析医疗术语的特殊性，结合现有的实体识别技术的同时还引入了中文独特的词法特征，更出色地解决了医疗复杂粒度命名实体识别问题。同时，我们把从非结构化文本中识别医疗实体词的能力进行复用，既支持数据处理场景从海量数据中自动挖掘医疗实体词，又支持在线服务实时把线上用户描述的关键医疗信息进行结构化，清晰地呈现给医生。最终，我们构建的图谱可以有效的辅助患者全面、准确的描述自身病情；还可以自动理解用户已经描述过的信息，减少重复提问，提升用户体验；同时还把准确、直观的医疗信息提供给医生，提高医生的工作效率。

本次实践项目研究方向与学位论文完全相关，学位论文方向为“一种高效的中文医学知识图谱的构建方法”，本次实践受益匪浅。

3.在校期间主要研究成果【含产品与样机、专利（含申请）、著作、软件著作权、论文、标准、获奖、成果转化等】

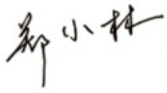
成果名称	类别[含产品与样机、专利（含申请）、著作、软件著作权、论文、标准、获奖、成果转化等]	发表时间/授权或申请时间等	刊物名称/专利授权或申请号等	本人排名/总人数	学校排名/总参与单位数

本人承诺

在专业实践训练及考核报告撰写过程中，如实提供材料，严守学术道德、遵循学术规范。

签字：  2022年 5 月 26日

三、考核评价

校外合作导师(或现场导师) 评价	**重点对研究生项目研究开展情况、职业素养、行业知识掌握、环境和岗位适应能力、工程实践能力、团队协作能力，以及通过技术应用创新、成果转化、解决工程实际问题等取得的经济和社会效益等方面的评价：** 该同学有积极的心态，能够乐观的应对变化，快速适应环境；很强的主动学习意愿，在医疗实体识别任务上取得了一定的技术积累；扎实的技术基础能力，能够快速解决工程问题。 校外合作导师（或现场导师）签字：  2022年 5 月 27日
校内导师 评价	**重点对研究生科学素质、基础及专业知识掌握、技术应用创新能力、取得的研究成果、项目研究与学位论文撰写的相关程度等方面的评价：** 马瑞祥同学在知识图谱领域进行了深入的基础理论学习，并在阿里健康公司结合互联网问诊场景开展实践命名实体识别技术，将理论与实践紧密结合。该工作将对后续要开展的毕业设计奠定扎实的理论与实践基础。 校内导师签字：  2022年 5 月 27日

实践单位过程考核意见	**实际实践开始时间：**2020年 10月 1日 **实际实践结束时间：**2021年 12月 31日 **专业实践训练累计天数：** 456天 **其中项目研究天数：** 456天 **实践单位过程考核结果：**☑优秀 ☐良好 ☐合格 ☐不合格 审核签字并盖公章： 2022年 5月 30日
最终考核结果审核备案	**考核总成绩（含现场答辩考核成绩 90%+单位过程考核成绩 10%组成）：** **是否重修：☐是 ☐否** 教学管理部（或相关分院）审核签字（公章）： 年 月 日

四、相关支撑材料

在校期间主要研究成果【含产品与样机、专利（含申请）、著作、软件著作权、论文、标准、获奖、成果转化等】证明材料原件扫描件，具体提交要求如下：

1.产品与样机扫描件包含企业证明材料（含产品与样机功能及创新性介绍、社会经济效益、个人贡献说明及相关照片等）。

2.授权专利扫描件包含专利证书授权页；未授权专利扫描件包含专利受理书扫描件和专利请求书扫描件。

3.著作扫描件包含封面、封底和版权页。

4.软件著作权扫描件包含著作权证书和登记申请表。

5.论文扫描件包含封面、封底、目录和论文全文（含收录证明）。

6.标准扫描件包含封面、版权页、发布公告、前言和目次。

7.获奖扫描件包含显示单位和个人排名的获奖证书。

8.成果转化扫描件包含企业证明材料（含成果技术说明、社会经济效益、个人贡献说明及相关照片等）。