大连理工大学

专业学位硕士研究生学位论文开题报告

论文题目：基于图卷积网络的文档级别关系抽取研究

姓　　名：徐奕斌

学　　号：31909185

专业/领域：计算机技术

培养类型：☑全日制 ☐ 在职学习

指导教师：姚卫红

实践导师：周惠巍

入学日期：2019.9

报告日期：2020.9.30

报告地点：创新园大厦 A802

研究生院制表

说　明

学位论文开题考核是硕士研究生课程学习结束后开展学位论文工作的基本要求，是保证学位论文质量、工作进度和研究生培养质量的首要环节。专业学位硕士研究生学位论文可以是产品研发、工程设计、专题应用研究、工程/项目管理、调研报告等形式，具体要求应参照全国相关专业/领域专业学位教育指导委员会制定的专业学位标准。

一、考核内容：首先，考查硕士生对本专业/领域课程学习、校内实践（实验）情况，对专业知识、技能、标准规范等掌握程度；其次，考查学位论文工作准备情况，包括论文选题是否适合专业学位研究生培养、论文的目的意义及国内外发展状况、论文内容设置的合理性、方法的科学性、工作量与工作难度、预期成果的实用性和新颖性、文献阅读等；第三，还要考查学生校外企业实践情况及时间安排、研究生的学习和工作态度等；此外，还要考查该生实践环节的实践条件是否有保障、校内外实践安排是否合理等。

二、考核时间：硕士生的开题报告应在第 2 学期末或第 3 学期初进行。

三、报告撰写：开题报告正文字数不少于 5000 字；参考文献数量不少于 20 篇（其中外文文献不少于 40%）；正文及参考文献等撰写要求参见《大连理工大学硕士学位论文格式规范》。

四、考核办法：开题考核由学部（学院）集中组织 3 名以上本学科领域专家（至少一名专家来自企业，导师和企业导师除外）以答辩的方式进行。学生进行口头陈述时间不得少于 10 分钟。专家组给出考核成绩和是否通过的意见。

五、报告保存：开题报告一式两份，签字后分别由学部（学院）和学生保存。

六、信息登录：研究生开题后登录研究生信息管理系统上传开题报告（PDF 文档）及考核结果。

开题报告正文

1. 课程学习情况（附成绩单）、参加科研和学术活动等情况

1.1 课程学习情况

应修 32 学分，目前已修 25 学分，其中必修课程已修 18 学分，选修课程已修 7 学分，有一门学科未录入成绩。大多数课程取得了不错的成绩。通过研一一年的学习，我学到了很多专业方面的知识，并且进行了校内实验，我的成绩单见表 1.1。

表 1.1 成绩单

必修课程	课程学分	选修学期	成绩
分布式数据库	2	1	81
计算机技术前沿	2	2	86
高级操作系统	2	1	83
工程伦理	1	1	80
知识产权	1	1	96
信息检索	1	1	P
中国特色社会主义理论与实践研究	2	1	78
阅读与写作 I（基础读写技能）	2	1	88
矩阵与数值分析	3	1	72
数理统计	2	1	80
选修课程	**课程学分**	**选修学期**	**成绩**
数据仓库技术	2	2	P
现代网络管理	2	2	P
中国古代文学专题	1	1	P
机器翻译基础	2	2	85

1.2 参加科研

①每周参加小组例会交流，汇报科研进展。

②参与 BioNLP2019 共享任务：生物医学问答的研究，实现基于 MT-DNN 预训练模型和 Transformer 模型的自动问答系统。

③参与构建文档级别关系抽取系统。

1.3 学术活动情况

① 2019 年 8 月 25 日，参加 Mini-Workshop on Computational Science(MWCS 2019)，听取相关研究报告。

② 2020 年 5 月 23 日，参加 ACL-IJCAI-SIGIR 顶级会议论文报告会，听取了相关的研究报告。

2. 学位论文研究背景、目的和意义

近年来，随着移动通信技术的发展，人类社会进入大数据时代。信息抽取（Information Extraction, IE）作为自然语言处理（Nature Language Processing, NLP）的一项基础任务越来越受到关注，它旨在从海量的非结构化文本中抽取出有价值的信息并且结构化成下游自然语言处理任务可用的格式。信息抽取的两个基本子任务是命名实体识别（Named Entity Recognition, NER）和关系抽取（Relation Extraction, RE），命名实体对应真实社会中存在的人名、地名、组织名和时间等具体对象，关系即为这些对象中存在的联系，比如“武汉”和“湖北省”都是命名实体，他们之间的关系为“省会”。关系抽取对于知识发现[1]、自动问答[2][3]、医药信息学[4]和本体关系学习[5]等任务都有着重要的意义。关系抽取分两步，一步是判断实体对之间是否存在关系，而另一步则是判断有关系的实体对之间的关系属于哪种。这两步也可以归并为一步，即把无关系当作一种特殊的关系，来进行多类别分类。

以下是两个实例：

实例 1：Seizures were induced by pilocarpine injections in trained and non-trained control groups.

在实例 1 中，Seizures（癫痫）是标注出的疾病实体，pilocarpine（匹罗卡品）是标注出的药物实体。Pilocarpine 和 Seizures 之间存在着药物诱导疾病的关系（chemical-induced disease, CID relations），关系抽取任务就是根据药物-疾病实体对以及它们所在的上下文来确定它们之间是否存在这样的关系。由于是以句子作为输入样例，则称该任务为句子级别关系抽取或提及关系抽取，句子级别关系抽取往往需要手动抽取、过滤样例，这会导致部分样例丢失，对实验结果产生影响。

实例 2：[1]“Ikwilalles met je delen” (“I want to share everything with you”) was the Dutch entry in the Eurovision Song Contest 1990, performed in Dutch by Maywood. [2] The English language version was entitled “No more winds to guide me”. [3] The song is a ballad, with the singer telling her lover that she wants to share everything with him-including the hard times in life. [4] She sings that... [5] The song was performed fifth on the night, following Turkey’s Kayahan with Gözlerinin

Hapsindeyim and preceding Luxembourg's Céline Carzo with “Quand je terêve”. [6] At the close of voting, it had received 25 points, placing 15th in a field of 22...

实例 2 是一段由六个句子组成的文本，实体对可能出现在一个句子中也可能出现在不同的句子中，例如< Céline Carzo , Eurovision Song Contest 1990 >实体对中，头实体出现在第五个句子中，而尾实体出现在第一个句子中，这样以文章作为输入，对文章中所有出现的实体对（包含跨句的）进行关系抽取的任务被称为文档级关系抽取或全局关系抽取。

文档级别关系抽取相比于句子级别关系抽取任务难度更大，因为更多的实体数量和更长的上下文序列使得对实体之间的关系建模变得更加困难。在上面的实例中，为了判断< Céline Carzo , participant of , Eurovision Song Contest 1990>这样一个关系事实，首先要从第一个句子得到“Ikwilalles met je delen”是 Eurovision Song Contest 1990 的一首参赛歌曲，然后从第五个句子得知这首歌是在 Céline Carzo 演唱的歌曲之后演唱的，最终可以推导出 Céline Carzo 是 Eurovision Song Contest 1990 参与者的事实。

越来越多的组织和个人致力于使用机器学习和深度学习的方法使机器能够自动的学习和抽取实体之间的关系。与此同时，很多关系抽取数据集发布如 SemEval-2013 Task 9: Extraction of Drug-Drug Interactions from BioMedical Texts[6]、BioCreative IV发布的蛋白质-蛋白质相互作用（PPI）关系抽取任务[7]和 Bio creative V 发布的药物诱导疾病（CDR）关系抽取任务[8]。这些评测任务对实体关系抽取问题的研究起到了极大的推动作用，许多来自国内外的队伍都参加了这些评测，并进行了深入的研究。

虽然评测任务的不断开展使得实体关系抽取技术取得了进步，但是目前的技术也体现出了一些局限性，例如现有的神经网络模型对于较长序列的建模存在困难，因而大多数研究者首先对给出的文章进行样例筛选，对每个不同的实体对分别构建属于它们自己的上下文序列。抽取和筛选样例不仅会造成相隔较远实体对的丢失还会损失大量的上下文信息，这会导致模型召回率的降低对实验结果产生很大的影响。这样的操作费时费力，如何以整篇文章作为输入，并高效地编码实体、对实体间复杂的交互进行建模，即实现文档级别的关系抽取是本文的研究重点。

除了不断提出的数据集和评测任务，许多不同种类的知识库也被构建，如 Wikipedia 包含了大量通用领域的结构化信息，Uniprot、BioGrid、IntAct 和 CTD[9] 等生物医学领域的知识库构建了大量的蛋白质、基因以及药物之间的关系。这些知识库包含的结构化的有效信息对关系抽取任务有着极大的帮助。Bio creative V CDR 评测任务中，最好的参赛结果 F 值仅为 57.03%，然而在引入知识信息后 F 值达到了 70.20%，可见融合外部知识信息对于关系抽取任务有着非常大的帮助。如何有效地将知识库中的结构化信息与文本信息相融合来进行文档级别的关系

也是本文的研究内容。

3. 国内外研究现状及发展动态分析

在近几年的研究中，实体关系抽取常用的方法大致可分为两种：基于规则的方法和基于机器学习的方法。基于规则的方法是利用句法结构的方法进行关系抽取。Lowe 等人[10]使用了一些能够表示 CID 关系的关键字以及其他句法结构信息识别疾病和药物之间的关系，在 BioCreative V CDR 任务中，基于标准的实体集取得了 60.75%的 F 值，基于自己识别出的实体集取得了 52.2%的 F 值。基于规则的方法具有很强的解释性，且简单有效。但构建规则需要投入大量的人力，且涉及到生物医学等专业性强的领域，需要相关领域专家的专业知识指导，不仅成本高，且这些人工制定的规则也很难应用到新的领域或数据集中。

基于机器学习的方法可以有效地缓解以上的问题。机器学习方法分为传统机器学习方法和最近取得更好结果的深度学习方法。传统的机器学习方法又包括基于特征的方法和基于核的方法。Gu 等人[11]使用了大量的语言学特征，基于最大熵分类器抽取疾病药物关系。分别在句内和跨句两个级别上做了实验，在 BioCreative V 任务中基于给定标准实体的基础上达到 58.3%的 F 值。Xu 等人[12]使用了许多药物知识库的资源生成基于知识库的特征，并通过支持向量机模型[13]（Support Vector Machines, SVM）进行分类。最终在 BioCreative V 任务中在标准的实体标注下，达到了 67.16%的 F 值。

基于核的方法先计算句法结构树的相似度然后进行分类，Zhou[14]等人探索使用多种树核函数来捕捉实体之间的句法信息，显示出树核函数对于关系抽取任务的有效性，Panyam[15]等人在 Zhou 的基础上又进一步引入了图核函数，获取了实体间的深层句法关系，在 BioCreative V 任务上 F 值达到了 60.3%。

目前在各类自然语言处理任务中，深度学习的方法被广泛应用，并且展示出一些相对于传统机器学习方法的优越性，越来越多的研究者探索利用深度学习模型来进行实体关系抽取。较为常用的两个深度学习模型是卷积神经网络（Convolutional neural network，CNN）和循环神经网络（Recurrent neural network，RNN）。Zeng 等人[16]使用 CNN 来学习句子级别的特征，并融合词级别特征来进行关系分类，他们的模型在 SemEval-2010 Task 8 中取得了 82.7%的 F 值。Gu 等人[17]除了用 CNN 编码了上下文特征，还利用另一个独立的 CNN 抽取了依存特征，通过融合上下文特征和依存特征进行关系抽取，在 BioCreative V 关系抽取任务中取得了 61.3%的 F 值。RNN 在序列学习上展现了自身的强大优势，但是输入序列过长往往会导致梯度弥散和梯度爆炸的问题[18]，长短时记忆网络（Long short term memory, LSTM）通过在 RNN 中加入门控单元解决了这一问题。

Sunil 和 Ashish[19]将句子中词的嵌入拼接上词的位置信息，然后输入到双向的长短时记忆网络，以此建模实体间长距离的关系信息。Yi 等人[20]使用了改进的 LSTM 模型——门控循环网络（Gated recurrent unit, GRU），将拼接了位置信息的句子作为输入，利用双向 GRU 模型进行训练，并且还在单词和句子两个不同级别使用注意力机制，这使得模型能够获得不同层级的有用特征。

各领域知识库的构建与完善为实体关系抽取提供了更多的可用资源，表示学习提供了如何表示知识库中结构化知识的方法。知识表示学习通过将知识库中三元组信息映射到低维向量空间来揭示知识库中结构化信息之间的相互关联。Bordes 等人[21]提出的结构表示模型（Structured Embedding，SE）是最早的知识表示模型，但该表示方法的头尾实体间的协同性较差，不能准确表示两个实体之间的语义关系。Bordes 等人[22]受词向量平移不变的特性的启发，又提出了 TransE 模型，将知识库中的关系看成是实体向量在空间中的某种平移，这也可以理解为关系是实体到实体间的翻译，因此这种表示学习方法又被称为翻译模型。为了解决 TransE 模型一对多、多对一、多对多复杂关系的局限性，Wang 等人[23]提出了 TransH 模型，对于每一种关系，该模型使用一个关系向量和一个关系平面法向量来表示，但是还是基于实体和关系在同一个向量空间的假设，这在某种程度上限制了它的表示能力。为了解决这个问题，Lin 等人[24]提出了 TransR 模型，做出了实体和关系应当属于不同语义空间、一个实体拥有不同的属性、不同关系关注实体不同属性的假设。基于翻译模型的知识表示方法将知识库的实体和关系表示成向量，可以和现有的深度学习技术紧密结合，更好地利用知识库中的结构化信息，提升实体关系抽取任务。

近年来，图卷积网络[25]（Graph Convolutional Networks，GCN）在很多自然语言处理任务上获得了广泛的关注，并且在很多句子级别和文档级别的关系抽取任务上被证明是有效的。当前应用于句子级别关系抽取任务的 GCN 大多是构建在输入句子的依存结构上的。Zhang 等人[26]使用 GCN 去建模从输入句子提取出的经过剪枝的依赖树，然后提取以实体为中心的表示来进行关系预测。Zhu 等人[27]提出在非结构的文本输入上生成图神经网络的参数，而不是对依赖关系进行建模，他们构造了以实体为节点的全连通图，利用生成的转移矩阵将节点的隐藏状态传播到相邻节点，用于后续的多跳推理。图卷积网络可以通过传播和聚合两个步骤将邻接节点的信息更新到目标节点中，这十分有利于对文档中实体间的复杂关系进行建模，因而利用图卷积网络完成文档级别关系抽取任务是一个可行的研究方向。

4. 主要研究内容、研究目标、拟解决的关键问题

4.1 研究内容

实体关系抽取指的是在给定的文本中，对于存在的实体对判断它们之间是否存在关系，存在怎样的关系。本文主要研究文档级别关系抽取任务，因此选用人工标注好实体的文档级别关系抽取语料来进行研究，本文的研究内容主要分为以下两个方面：

（1） 基于图卷积网络的文档级别关系抽取模型

拟利用上下文编码器对输入文档进行语义信息编码，通过平均操作获得各个实体的语义表示。同时以实体为节点建初始图，并利用图卷积操作更新节点的表示。在大规模文档级别关系抽取数据集上进行训练，最终构建基于图神经网络的文档级别关系抽取模型。

（2） 融合了知识表示的用于文档级别关系抽取任务的图卷积网络模型

通过抽取知识库中的知识信息，如知识三元组信息，通过知识表示学习方法将知识库中结构化信息转变为低维的稠密向量表示，即知识表示。将知识表示与文本信息表示融合，构建融合了知识表示的用于文档级别关系抽取任务的图卷积网络模型。

4.2 研究目标

文档级别关系抽取任务需要在文档中根据特定的实体来抽取它们之间的关系，这个关系通常是实体间上下文要表达的具体含义。因此文档级别关系抽取要充分利用文本语义信息对实体间的复杂交互进行建模。同时，知识库中包含的大量结构化信息蕴含着人类累积的经验知识，这对实体关系抽取有着很大的帮助作用。鉴于此本文的研究目标可以分为以下两点：

（1）利用文档自身包含的语义信息，例如词、词性、上下文信息等，通过建立高效可行的图卷积神经网络来构建适用于文档级别关系抽取任务的模型。

（2）抽取知识库中的结构化信息，例如（实体 1，关系，实体 2）的三元组信息。采用 TransE 等知识表示学习方法获得其相应的知识表示。融合文本信息和知识表示信息，通过图卷积网络构建高性能的文档级别关系抽取模型。

4.3 拟解决的关键问题

本文主要研究基于图卷积网络的文档级别关系抽取，同时在图卷积网络中进一步引入领域知识提高模型的性能。有以下几个问题需要解决①如何构建初始图，也就是获得怎样的邻接矩阵。②图卷积网络存在过渡平滑的缺点，如何在图卷积操作时关注到全局信息却不引入过多的噪声是需要考虑的。③由于知识表示与词向量表示不在同一语义空间中，选择合适的方法将知识表示和文本信息相融合也

是需要解决的问题。

5. 学位论文的研究方法、技术路线、试验手段、关键技术等论述

5.1 学位论文的研究方法

基于图卷积网络的文档级别关系抽取研究，主要通过文本编码器编码文本获得实体的向量表示，并通过图神经网络对实体间的复杂交互进行建模，获得最终的实体关系抽取模型。

（1）基于图卷积网络的文档级别关系抽取模型

首先通过 Glove 模型编码文本，并通过 LSTM 对上下文信息进行交互，或直接使用 Bert 预训练模型对文档进行编码，通过对 token 和 mention 的平均操作获得实体的向量表示，接下来进行初始无向图的构建，主要是提出合适的规则去进行实体之间的连接，获得合适的邻接矩阵。通过使用图卷积网络更新节点表示，使实体的表示中含有部分与他相邻实体的信息，通过多次图卷积操作，使实体间的本地信息和全局信息充分交互，最终使用头尾实体的向量表示，通过全连接神经网络进行分类。

（2） 融合了知识表示的用于文档级别关系抽取任务的图卷积网络模型

根据文本信息，抽取知识库中相应的结构化信息，得到知识三元组信息。利用知识表示学习方法，获得低维、稠密的知识表示，一般得到的是实体向量和关系向量表示。通过注意力机制，将知识表示引入，融合知识表示和文本信息，进行图卷积操作并更新实体表示，构建高性能的文档级别关系抽取模型。

5.2 技术路线：

基于图卷积的文档级别关系抽取任务主要是对文档进行编码，获得实体表示，通过在构建的图上进行图卷积操作，更新节点表示用以关系分类。还考虑通过抽取结构化知识库中的三元组，通过知识表示学习方法获得实体和关系的向量表示，通过注意力机制将其融入到文本信息中，提升文档级别关系抽取模型的性能。如下图 5.1 是融合了知识表示的用于文档级别关系抽取任务的图卷积网络模型框架图。

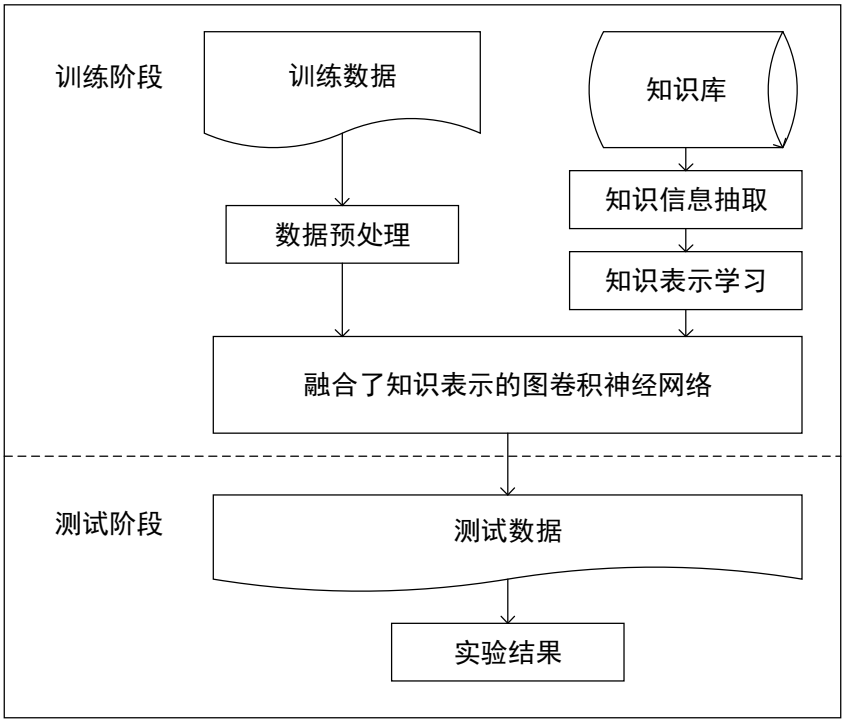

图 5.1 融合知识表示的用于文档级别关系抽取任务的图卷积网络模型框架图

以下内容是具体每个阶段的技术路线：

文档级别关系抽取可以看成是一个分类任务。即一直两个实体表示，可以通过训练好的分类器判断他们之间的语义关系，具体研究内容包含两部分。

（1）基于图卷积网络的文档级别关系抽取模型

首先通过 Glove 模型编码文本，并通过 LSTM 对上下文信息进行交互，或直接使用 Bert 预训练模型对文档进行编码，通过对 token 和 mention 的平均操作获得实体表示，接下来进行初始无向图的构建，主要是提出合适的规则去进行实体时间的连接，获得合适的邻接矩阵。通过使用图卷积网络更新节点表示，使实体的表示中含有部分与他相邻实体的信息，通过多次图卷积操作，使实体间的本地信息和全局信息充分交互，最终使用头尾实体的向量表示，通过全连接神经网络进行分类。

（2） 融合了知识表示的用于文档级别关系抽取任务的图卷积网络模型

根据文本信息，抽取知识库中相应的结构化信息，得到知识三元组信息。利用知识表示学习方法，获得低维、稠密的知识表示，一般得到的是实体向量和关系向量表示。通过注意力机制，将知识表示引入，融合知识表示和文本信息，进行图卷积操作并更新实体表示，构建高性能的文档级别关系抽取模型。

5.3 试验手段

实验所用语料库是 BioCreative V task 的评测语料库，该语料库主要收录了生物医学领域 PubMed 的文献，共计 1500 篇文章。其中 500 篇用于训练，500 篇用于验证，最后 500 篇用于测试。这些数据均通过人工的手段，将药物和疾病的实体，以及他们之间的关系标注出来用于进行实验。其中，训练数据共 1038 对疾病药物关系（chemical disease relation, CDR），验证集 1012 对关系，测试集 1066 对关系。评测采用了官方给予的评测工具，通过计算文档之中标注出来的疾病药物关系和系统预测出来的做比较。最终结果用精确率（*Precision*）、召回率（*Recall*）和 F 值（*F-score*）来进行表示。

精确率表示正确检测的 CDR 的比例，用公式 5.1 表示。

$$P = \frac{TP}{TP + FP} \tag{5.1}$$

其中 P 表示精确率，TP 表示正确检测的 CDR 的数目，FP 表示错误检测的 CDR 的数目。

（2）召回率表示正确检测的 CDR 的数目占评测语料中全部 CDR 的比例，用公式 5.2 表示。

$$R = \frac{TP}{TP + FN} \tag{5.2}$$

其中 R 表示召回率，TP 同公式 5.1，FN 表示系统未检测出的 CDR 数目。

（3）F 值（精确率和召回率的调和平均数），用公式 5.3 表示。

$$F = \frac{2PR}{P + R} \tag{5.3}$$

本文将基于深度神经网络的实体关系抽取看成是一个二分类问题。所有的训练、验证、测试数据均使用标准标注的实体来进行实验。

5.4 关键技术

基于图卷积网络的文档级别关系抽取模型，旨在通过利用文本编码器对文本信息进行编码获得实体向量表示，并通过在构建图上进行图卷积操作更新节点表示，将节点表示用于实体对之间关系分类。考虑抽取知识库的结构化信息用以进行知识表示学习，通过融合文本信息和知识表示，进一步提升文档级别关系抽取模型的性能。

（1） 基于图卷积网络的文档级别关系抽取模型

利用文本编码器编码文本信息，获得实体表示，并在构建图上使用图卷积操作不断更新节点信息，使用分类器对实体对之间的关系进行分类。

（2） 融合了知识表示的用于文档级别关系抽取任务的图卷积网络模型

知识库中的结构化信息以三元组形式来表示，知识表示学习将三元组中的实体和关系映射到低维向量空间中，获得相应的实体和关系的稠密向量表示。然后利用注意力机制等方法将知识表示与文本信息融合，构建高性能的基于知识表示和图卷积网络的文档级别实体关系抽取模型。

6. 年度研究计划

（1） 2020 年 7 月到 2020 年 8 月完成基本语料的预处理工作。并收集大量无标注的语料作为深度学习学习词向量的原始数据。

（2）2020 年 9 月到 2020 年 10 月，学习深度学习的理论知识，了解基本深度神经网络模型，如卷积神经网络（CNN）、循环神经网络（RNN）、长短时记忆网络（LSTM）、注意力机制（Attention）。研究如何深入使用神经网络抽取相关特征以及编码文本信息。

（3） 2020 年 10 月到 2020 年 11 月，熟悉预训练模型 Bert 的原理和使用方法，尝试用 Bert 模型编码文本并获得实体表示。

（4） 2020 年 12 月到 2021 年 1 月寻找相关的知识库，并完成知识库结构化信息的抽取。熟悉知识表示学习工具如 TransE 的使用方法，利用知识表示学习构建实体与关系的知识表示。

（4） 2021 年 2 月，学习图卷积网络的理论知识，研究利用图卷积网络对实体之间的复杂关系进行建模。探索如何解决图卷积网络的过渡平滑问题。

（5） 2021 年 3 月，学习并深入研究多源信息融合的方法，通过引入注意力机制等各类引入外部信息的方法，有效地融合知识表示信息和文本信息。尝试构建不同结构的图卷积神经网络模型。

（6）2021 年 4 月到毕业整理实验数据、图表与看过的相关论文，撰写和修改毕业论文，并参加毕业答辩。

7. 校内外实习实践条件及时间安排

（1） 校内实习。

（2） 在实验室进行实习实践。

（3） 校内实习实践时间安排：

① 2020 年 6 月至 10 月，深入研究自然语言处理，完成相关任务。

8. 现有的研究基础

（1） 已阅读一些关系抽取和关系分类的中英文文献，掌握一些自然语言处理的基本知识，并学会如何对语料进行预处理，如何抽取特征，以及如何训练模型，最后测试等一系列自然语言处理的基本流程。掌握一些自然语言处理工具包的使用方法。例如分词、词性标注、句法解析等常用的自然语言处理工具。

（2）了解一些生物医学相关的知识库，例如 Uniprot，CTD 等。

（3）学习并掌握一些知识表示学习的方法和对应工具，例如 TransE 等。

（4）阅读关于深度学习和神经网络的论文，正在研究深度学习的模型主要包括卷积神经网络（CNN）、递归神经网络（RNN）、长短时记忆网络（LSTM）和图卷积网络（GCN）。

（5） 了解一些常用的深度学习框架和工具，例如 Pytorch 框架。

9. 现有研究条件及可能遇到的困难和问题分析

通过阅读相关文献，预测实验中可能会遇到以下几个方面的问题：

（1）关系抽取中，尤其是生物医学领域的语料复杂多样，使用工具来进行词性标注和句法解析等难免会出现一定的误差。如何有效地减小这些误差对后续的实验来说十分关键。

（2）各种不同知识库中信息的抽取与融合具有一定难度，对于不同的生物医学知识库，同一种药物、疾病有不同的 ID 编号体系，需要进行识别与统一。

（4）如何将知识表示学习获得的知识库中知识信息与文本信息有效地融合，也是本文面临的难点之一。

参 考 文 献

[1] Chris Q and Hoifung P. Distant Supervision for Relation Extraction beyond the Sentence Bounda-ry [C]// Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, 2017:1171-1182.

[2] Wentau Y, Mingwei C, Xiaodong H and Jianfeng G. Semantic Parsing via Staged Query Graph Generation: Question Answering with Knowledge Base [C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, 2015:1321–1331, Beijing, China.

[3] Mo Y, Wenpeng Y, KaziSaidul H, Cicero D S, Bing X, and Bowen Z. Improved Neural Relation Detection for Knowledge Base Question Answering [C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017:571-581, Vancouver, Canada.

[4] Wang C and Fan J. Medical relation extraction with manifold models [C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014:828-838.

[5] Xu Y, Li G, Mou L L and Lu Y Y. Learning non-taxonomic relations on demand for ontology extension [J]//International Journal of Software Engineering and Knowledge Engineering, 2014, 24(08): 1159-1175.

[6] Segurabedmar I, Onez P M, Herrerozazo M O, et al. SemEval-2013 Task 9 : Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013)[C]// Proceedings of the Seventh International Workshop on Semantic Evaluation, 2013:341–50.

[7] Doğan, R. I., Kim S., Chatr-aryamontri, A., Wei, C.H., Comeau, D. C., and Lu, Z. Overview of the BioCreative VI Precision Medicine Track [C]// Proceedings of the 2017 Workshop on BioCreatice VI, Washington. 2017:83-87.

[8] Wei C, Peng Y, Leaman R, Davis A, Mattingly C, Li J, Wiegers T, Lu Z. Assessing the state of the art in biomedical relation extraction: overview of the BioCreative V chemical-disease relation (CDR) task [J]//Database, 2016:baw032.

[9] Davis A, Grondin C, Johnson R, Sciaky D, King B, McMorran R, Wiegers J, Wiegers T, Mattingly C. The comparative toxicogenomics.[J]//Database: Nucleic Acids Research. 2016, 45(D1):D972-D978.

[10] Lowe D M, O’Boyle N M and Sayle R A. Efficient chemical-disease identification and relationship extraction using Wikipedia to improve recall [J]//Database, 2016, doi:10.1093 /baw039.

[11] Gu J H, Qian L H and Zhou G D. Chemical-induced disease relation extraction with various linguistic features [J]//Database, 2016, doi:10.1093/database/baw042.

[12] Xu J, Wu Y H, Zhang Y Y, et al. CDREST: a system for extracting chemical-induced disease relation in literature [J]//Database, 2016,doi:10.1093/database/baw036.
[13] Chang C C and Lin C J. LIBSVM: A library for support vector machines [J//ACM Transactions on Intelligent Systems & Technology, 2011, 2(3): 389-396.
[14] Zhou H W, Deng H J and He J. Chemical-disease relations extraction based on the shortest dependency path tree [C]// Proceedings of the fifth BioCreative Challenge Evaluation Workshop, 2015:214-219.
[15] Panyam N C, Verspoor K, Cohn T, et al. Exploiting graph kernels for high performance biomedical relation extraction:[J]// Journal of Biomedical Semantics, 2018, 9(1):7.
[16] Zeng D J, Liu K, Lai S W, Zhou G Y and Zhao J. Relation classification via convolutional deep neural network [C]//Proceedings of the 25th International Conference on Computational Linguistics, 2014:2335-2344.
[17] Gu J, Sun F, Qian L, et al. Chemical-induced disease relation extraction via convolutional neural network:[J]//Database the Journal of Biological Databases & Curation, 2017, 2017(1).
[18] Hochreiter S. The vanishing gradient problem during learning recurrent neural nets and problem solutions [J]//International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 1998, 6(02): 107-116.
[19] Sunil K S, Ashish A. Drug-Drug Interaction Extraction from Biomedical Text Using Long Short Term Memory Network [J]//arXiv, 2017, 1701.08303.
[20] Yi Z, Li S, Yu J, Wu Q. Drug-drug Interaction Extraction via Recurrent Neural Network with Multiple Attention Layers [J]//arXiv, 2017, 1705.03261.
[21] Bordes A, Weston J, Collobert R, et al. Learning Structured Embeddings of Knowledge Bases[J]//2011.
[22] Bordes A, Usunier N, Garciaduran A, et al. Translating Embeddings for Modeling Multi-relational Data[C]//Neural information processing systems, 2013: 2787-2795.
[23] Wang Z, Zhang J, Feng J, et al. Knowledge graph embedding by translating on hyperplanes[C]//National conference on artificial intelligence, 2014: 1112-1119.
[24] Lin Y, Liu Z, Sun M, et al. Learning entity and relation embeddings for knowledge graph completion[C]. national conference on artificial intelligence, 2015: 2181-2187.
[25] Thomas N. K and Max W. Semi-Supervised Classification With Graph Convolutional Networks [C]//Proceedings of ICLR 2017.
[26] Yuhao Z, Peng Q, and Christopher D. M. 2018. Graph Convolution over Pruned Dependency Trees Improves Relation Extraction [C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Lan-guage Processing, 2018:2205-2215, Brussels, Belgium.

[27] Hao Z, Yankai L, Zhiyuan L, Jie F, Tat-seng C, and Maosong S. Graph Neural Networks with Generated Parameters for Relation Extraction [C]// In Proceedings of the 57th Annual Meeting of the Asso-ciation for Computational Linguistics, 2019:1331-1339, Florence, Italy.

实践导师考核意见（对学位论文工作及开题考核报告撰写情况、企业实践实习情况及计划、学习和工作态度等进行考查，给出考核成绩和具体改进意见和建议）：

1） 考核成绩：☑ 优秀，☐ 良好，☐ 中等，☐ 及格，☐ 不及格

2） 是否通过：☑ 通过，☐ 不通过

3） 关于开题考核报告撰写质量及学位论文工作的具体意见（可加页）：

徐奕斌同学在过去的一年里积极参加专业实践活动，学习认真努力，实践态度端正，具有良好的工作习惯，目前该同学主要研究文档级别的关系抽取任务，考虑使用较为流行的图卷积网络对文档间的实体关系进行建模，具有一定的研究价值。该报告研究任务合适，方法合情合理，同意中期答辩。

导师签字：周惠斌

年　　月　　日

大连理工大学专业学位硕士研究生开题考核报告评审意见表

学　号	31909185	学生姓名	徐奕斌	专业/领域	计算机技术
第一次开题 ☑			第二次开题 ☐		
实 践 导 师 信 息					
姓　名	周惠巍	性　别	女	职称/职务	副教授
专　业	计算机科学与技术	所在单位	大连理工大学	联系电话	15542679503
通讯地址	大连市凌工路 2 号大连理工大学创新园大厦			E_mail	zhouhuiwei@dlut.edu.cn

校内导师考核意见（对课程学习情况、校内实践实习情况、参加学术活动情况、学位论文工作及开题考核报告撰写情况、企业实践实习情况及计划、学生的学习和工作态度等进行考查，给出考核成绩和具体改进意见和建议）：

1） 考核成绩：☑ 优秀，☐ 良好，☐ 中等，☐ 及格，☐ 不及格

2） 是否通过：☑ 通过，☐ 不通过

3） 关于开题考核报告撰写质量及学位论文工作的具体意见（可加页）：

该同学在过去的一年间积极参与实验室的科研活动，学习认真努力，科研态度端正，具有良好的工作习惯，目前该同学主要研究文档级别的实体关系抽取任务，考虑使用目前较为流行的图卷积网络对文档中的实体之间交互进行建模，具有一定的研究价值，该同学对语料预处理的任务已经完成，正在进行模型的搭建。开题报告的撰写已经完成，主要内容有两方面，一是基于图卷积网络的文档级别关系抽取研究，二是融合知识表示用于文档级别关系抽取的图卷积网络研究，该报告的研究方法合情合理，同意进行中期答辩。

导师签字：姚卫记

年　　月　　日

		姓名	职称	学科专业	是否博导	签字
评议专家组	组长	孟军	教授	计算机科学与技术	是	
	成员	丁男	副教授	计算机科学与技术	是	
		孙亮	副教授	计算机科学与技术	否	
		任健康	副教授	计算机科学与技术	否	

专家组评审意见（对课程学习情况、校内实践实习情况、参加学术活动情况、学位论文工作及开题考核报告撰写情况、企业实践实习情况及计划、学生的学习和工作态度等进行考查，给出考核成绩和具体改进意见和建议）：

1） 选题是否属于本学科领域（含交叉学科）：☑是，☐不是（须重新开题）

2） 选题是否符合专业学位论文要求：☑是，☐不是（须重新开题）

3） 考核成绩：☐优秀，☐良好，☑中等，☐及格，☐不及格

4） 是否通过：☑通过，☐ 不通过

5） 关于开题考核报告撰写质量及学位论文工作的具体意见（可加页）：

徐奕同学的硕士学位论文选题恰当，具有一定的理论意义和实际价值，研究方法和研究计划基本合理，难度合适，学生能够在预定时间内完成该论文的设计，开题报告条理清晰，撰写基本规范。在后续工作中要特别注意按时完成预定计划，在实验实施方面需要进一步加强。

组长签字：

年　　月　　日