|  |  |
| --- | --- |
| Inline Text Wrapping Picture |  |

硕士研究生学位论文阶段报告

学 号: 2018110744

姓 名: 何康镐

学 院: 计算机学院

专业(领域): 计算机科学与技术

研究方向: 大数据技术与智能信息处理

导师姓名: 于艳华

北京邮电大学

2020年9月24日

|  |  |  |  |
| --- | --- | --- | --- |
| 论文题目 | 基于生成对抗网络的实体关系抽取研究与实现 | | |
| 论文类型 | 应用研究 | 选题来源 | 中央、国家各部门项目 |
| 开题日期 | 2019-11-01 | 是否开题题目 | 是 |
| 论文开始日期 | 2019-11-01 | 报告日期 | 2020-9-24 |
| 报告地点 | 教三楼612 | 报告时间 | 上午8:30 |
| **研究内容简介**   * 选题背景   互联网从早期的信息互联已经发展到了信息爆炸的阶段，网络数据的形式在不再单一化，如今网络数据的呈现形式丰富多样，包括文本、视频、音频、图像等，这些数据包含着大量信息且蕴含着巨大的价值。其中文本数据非常常见且数据量巨大，各种网络媒体、自媒体、公众号、个人社交账号每天都在产生大量文本数据，这些数据以互联网为媒介传递，影响到每一个人的生活。为了更好的获取网络文本信息，搜索引擎技术应运而生，早期的搜索引擎通过检索用户输入的关键词，返回网页数据。随着计算机科学技术的不断发展以及用户需求的不断改变，搜索引擎技术也逐步更新。时至今日，网络文本数据越来越复杂，获取非结构化数据已经不再能满足用户的需求，例如，用户想要获取某新闻信息，用户更希望获得结构化数据，如该新闻的主要事件信息、主要人物信息等，这些信息是无法直接从非结构化文本中获得的，需要从非结构化文本中获取结构化信息。如何快速有效的从海量互联网新闻文本中获得我们需要的结构化信息成为了一个新的挑战和需求。  为了应对这一问题，近些年来工业界和学术界的广大学者对信息抽取技术展开大量研究，其中实体关系抽取(Entity Relation Extraction,简称RE)技术是信息抽取技术中的一个子任务，该任务定义为：给定文本，从文本中识别出实体，抽取实体间的语义关系，输出实体关系三元组（实体-关系-实体）。比如，给定一个句子‘日本地震引发了海啸’，‘地震’和‘海啸’是该句子的两个实体，它们之间存在因果关系，从该句子中可以抽取出‘地震-因果-海啸’三元组。通过关系抽取，可以将非结构化的文本信息转化成结构化的三元组信息，这种结构化信息过滤掉了原始非结构化文本中的大量冗余重复信息，为上层应用提供了数据基础。  实体关系抽取任务在1998年消息理解会议[1]（Message Understanding Conference，MUC）被首次提出。1999年，美国标准技术研究院( Nationalinstitutof Standards and Technology， NIST) 组织了自动内容抽取[2]( Automatic Content Extraction，ACE)评测，将实体关系抽取任务作为一项重要评测任务，且扩充了语料集。ACE 2005提供了多语种语料库包含英语、阿拉伯语和汉语，数据由实体、关系、事件组成，其中定义了7大类实体关系，每个大类又分为若干子类，共计17种关系。  SemEval[3](semantic evaluation)会议是继 MUC、ACE 后信息抽取领域又一重要评测会议， SemEval-2010 评测任务 8对关系抽取语料进行了丰富和完善，MUC、ACE 评测会议语料包含的关系种类具有局限性，SemEval-2010 Task 8将实体关系类型扩充到 9 种，分别是: Component-Whole、Instrument-Agency、Member-Collection、Cause-Effect、Entity-Destination、Content-Container、Message-Topic、Product-Producer 和 Entity-Origin，并且考虑到句子实例中实体对的先后顺序问题，引入“Other”类描述不属于前述关系类型的实体关系，共包含19种实体关系。SemEval-2010 Task 8对关系抽取任务的推进和发展有着重大影响，至今仍有大量学者在使用该数据集研究关系抽取模型。  ACE 2005和SemEval 2010 Task 8都是有监督训练数据集，需要大量人工标注数据。随着知识库的发展，为了减少标注数据带来的巨大人力消耗，Mintz et al[4](2009)提出了远程监督学习（Distant Supervised Learning）的方法用于生成大量数据。利用远程知识库如Freebase，如果知识库中的实体对存在某种关系，则认为包含该实体对的所有文本都存在这种关系。由于这种方法所带来的噪声过大，Riedel et al.[5]将问题建模为多示例学习问题，构造了NYT数据集，该数据集基于纽约时报语料和Freebase数据库，包含53种关系。这种方法虽然能获得大量训练数据，但是显然基于这种方法得到的数据含有大量噪声，如何降低噪声的影响成为了关系抽取的一个难点。  随着机器学习技术和自然语言处理技术的发展，这些方法也应用在实体关系抽取技术中。有监督机器学习方法主要包括基于特征和基于核函数的方法。Kambhatla,Suchanek[6,7]等主要采用了基于特征的方法，利用文本中的词法、语法信息（如语法树）作为特征，构造特征向量进行关系抽取模型训练。基于特征的方法受限于文本的语法、语义特征信息，这些特征选择的好坏直接影响模型的性能。Qian, Giuliano[8,9]等主要采用了基于核函数的方法，这种方法受到核函数的形式和参数影响，且模型训练时间较长。  由于有监督学习需要大量有人工标注的训练数据，Brin[10]率先提出了基于Bootstrapping的半监督实体关系抽取方法。该方法需要先筛选一些关系三元组作为种子，通过包含关系种子的上下文发现实体关系三元组，通过质量评估后加入到关系种子集合，不断迭代更新，从而达到扩充数据集的效果。这种方法需要人为筛选出初始种子，且需要高质量、覆盖范围广的种子。随着Freebase等知识库的建立和远程监督学习方法的提出，研究者将注意力转移到了如何做好远程监督学习的方向上。  近些年来，随着深度学习技术的成熟，神经网络模型被广泛应用到关系抽取领域。对有监督学习和远程监督学习的关系抽取技术都产生了深远影响。在有监督学习方面，Liu[11]率先使用了卷积神经网络（CNN）模型来进行关系抽取，利用CNN得到的词法特征进行分类。Zeng[12]在CNN的基础上提出利用实体位置特征信息，将词法级别信息和句子级别信息结合进行关系抽取。Santos[13]在Zeng的基础上提出了ranking-loss，不再使用传统的交叉熵作为损失函数。Zhang[14]考虑到对长依赖句子信息抽取不适用CNN，提出了使用循环神经网络（RNN）用于关系抽取，Zhou[15]在此基础上提出了BiLSTM结合注意力机制，取得了较大的提升。随后Wang[16]将注意力机制应用到CNN模型上，提出了基于CNN模型的多层级注意力机制。Cai[17]受RNN启发提出了循环卷积神经网络RCNN。  在远程监督学习方面，由于远程监督学习受噪声影响，不适合直接使用有监督学习的方法，Zeng[18]提出了PCNN（Piecewise-CNN），将句子特征分为三部分，分段卷积提取特征。Lin[19]提出了PCNN-ATT模型，在PCNN的基础上引入了注意力机制用于过滤远程监督带来的噪声影响，随后Lin[20]又提出多语言关系抽取模型MNRE，利用多语言关系数据表达一致的特性，结合中英文两种语料进行关系抽取。Xiong[21]首次提出远程监督学习具有长尾效应，很多关系的三元组数据过少，现有模型对于这部分关系的学习不充分，提出了小样本学习（Few-shot learning），利用少量训练样本进行关系抽取。  生成对抗网络（Generative Adversarial Networks,GAN）也是近些年深度学习技术中常用的一种模型，该模型由Goodfellow[22]提出，包括一个生成模型和一个判别模型，生成模型具有生成数据的能力，判别模型具有对抗噪声的能力，在图像领域有广泛的应用，如图像生成、图像修复、视频还原等。近些年来，GAN网络和对抗训练的思想也被用于NLP领域，Miyato[23]提出了将GAN用于文本分类问题，Yu[24]将GAN用于文本序列生成问题上，利用GAN网络中的生成模型生成序列化文本，Wu[25]通过在词嵌入上添加随机噪声，提出了基于对抗训练的关系抽取模型，冯冲[26]等利用对抗训练针对因果关系的进行抽取。Qin[27]，Han[28]等将对抗训练用于远程监督学习的关系抽取模型中，利用对抗训练学习的方法提高了关系抽取模型抗噪声的能力。Wang[29]等将对抗训练用于事件检测任务，训练得到的生成模型能够从原始数据中发现有效的训练数据用于事件检测任务中，证明了生成对抗网络具有扩充训练数据集的能力。   * 研究内容   目前关于对抗训练用于关系抽取领域的研究还不够全面，主要存在以下问题：  1）Wu[25],Qin[27],Han[28]等提出的基于对抗训练的关系抽取模型旨在利用添加随机噪声提高关系抽取模型的抗噪声能力，主要提升了关系抽取模型的分类性能，并没有利用生成模型能够生成数据扩充训练数据集的能力。对于关系抽取任务，有监督学习受限于标注数据集样本过少，而远程监督学习存在长尾问题和关系样本不均衡等问题，通过GAN网络训练一个有效的生成模型，利用生成模型的特点，可以实现自动标注数据集的功能，从而达到扩充训练数据的效果。  2）在以往的对抗训练过程中，主要通过在词嵌入层添加随机噪声，没有利用真实文本噪声数据来进行对抗训练，在词嵌入层添加噪声主要针对词向量层进行抗噪声训练，没有对关系抽取模型整体进行抗噪声训练，可以通过直接使用真实文本噪声数据训练判别模型，进一步提高模型整体的抗噪声能力。  3）对于文本特征标识，现有的方法主要通过文本词向量和位置信息表示文本特征，进而利用神经网络深入挖掘文本语义信息，这种方法利用到的文本特征不够全面，文本的语法特征、词性特征、实体上下文特征等都可以利用，进而充分挖掘文本中实体之间关系。Zhou[15]、Wang[16]等通过注意力机制提升了关系抽取模型的性能，在对抗训练过程中，判别模型也可以利用注意力机制加强判别模型的关系抽取能力。  因此，本课题针对以上问题进行研究，主要针对生成对抗网络的构建、对抗训练过程的实现、生成算法以及文本特征表示这几个部分展开研究，期望训练得到一个有效的生成模型用于生成有效的训练数据和一个判别模型具有良好的关系抽取能力。  1、生成对抗网络及生成算法  生成对抗网络（Generative Adversarial Networks ，GAN）主要由两部分构成：生成模型（Generative Model）和判别模型（Discriminative Model），通过它们的互相博弈学习产生相当好的输出。  判别模型就是通常分类任务里的分类器，原始的GAN模型里判别模型用于判别一张图片是否是真实的，它的输入参数是图片编码，输出D(x)代表图片是真实的概率。生成模型通过输入一个随机噪声z，生成迷惑判别模型的输入数据G(z)，在图像领域里G(z)就是图像的编码。  对抗训练过程：生成模型G的目标是尽量生成真实的图片去迷惑判别模型D，而D的目标是尽量判别出图片是属于真实的图片数据还是G生成的图片数据。通过反复G和D的博弈最终达到平衡，在理想状态下，最终训练结果是G生成的图片数据足以欺骗D，使得D对于G生成的图片数据判别概率为0.5。  基于对抗训练的关系抽取训练算法主要包括两个任务：（1）生成模型和判别模型的定义。GAN在NLP领域的应用不同于在图像领域的应用，文本数据通常是离散的数据难以直接用生成模型生成，且对于多分类任务，判别模型不能仅仅只用来判别生成数据的真实性，因此需要针对具体的NLP任务来定义生成模型和判别模型。（2）对抗训练的过程。由于神经网络的复杂性，对抗训练过程很难达到平衡状态，如何训练得到有效的网络是一个难点，应当选择合适的时机来停止对抗训练。  针对第一个任务，由于文本数据的离散性，用生成模型直接生成文本具有很大的难度，Lanto Yu[24]提出的SeqGAN虽然可以直接生成文本序列，但是该方法生成的文本序列通常没有明确的主题和方向，生成的文本可能不包含足够的实体或者没有明确的关系倾向，因此不适用于关系抽取任务。本课题受Wang[29]利用生成对抗网络进行事件检测的启发，利用半监督学习中bootstrapping的策略，从原始训练数据集中抽取实体关系种子，对网络文本数据中进行自动标注从而获得大量含有噪声的有标注数据，将原始训练数据认为是真实数据集R，将自动标注的数据认为是虚假数据集U。将生成模型设计成一个选择器，对输入U打分，通过设置阈值，当分数超过阈值则认为U足以迷惑判别模型，可以将此数据加入到真实数据集R里。原始的GAN网络里的判别模型是一个二分类模型，用于判别输入数据的真假，这种模型不适用于关系抽取这一多分类任务，Springenberg[30]将GAN网络用于图像多分类任务中，提供了一种多分类判别模型的模型结构。因此本课题的判别模型设计为一个关系分类模型，输入文本数据，抽取实体之间关系，输出该文本中实体关系对应类别的概率。  针对第二个任务，需要分别设计生成模型和判别模型的损失函数，进而定义训练过程。判别模型的损失函数主要由两部分构成，一部分来自于判别真实数据的损失函数，一部分来自于判别虚假数据的损失函数，这部分损失函数可以由多分类任务的交叉熵得到。生成模型的目标是迷惑判别模型，原始GAN网络生成模型损失函数由判别模型得到，本课题的生成模型是一个选择器，需要考虑生成模型的输出，拟采用生成模型的输出结果与判别模型输出概率的乘积作为损失函数。对抗训练过程则采用半监督学习方法，首先用原始训练数据训练判别模型，使其具有分类性能，进而对生成模型和判别模型交替训练，每轮训练完成后都对判别模型进行测试，保留分类性能最好的判别模型则停止训练。  训练结束后便可以利用生成模型对虚假的噪声数据进行判断，通过设置阈值，对于输出结果大于阈值的数据，将其加入到真实训练数据集R中，便可以实现训练数据集的扩充，从而达到生成训练数据的效果。  2、基于分段卷积网络的文本特征表示  关系抽取模型主要由四个层级网络构成：词嵌入层、位置嵌入层 、句子编码层、分类层。其中词嵌入层用于表示输入文本数据，将稀疏的文本数据转化为向量表示，词嵌入方法主要利用word2vec模型在纽约时报数据集上训练得到的词向量；位置嵌入层对句子中每个词相对于实体的位置进行编码，对于文本中的两个实体，考虑每个词相对于实体的相对位置，分别可以得到对实体1和实体2的相对位置信息，通过词嵌入和位置嵌入可以较为充分的表示原始文本，句子编码层主要为神经网络层如PCNN、RNN、BiLSTM等，利用神经网络特性对嵌入表达的文本向量充分挖掘，得到深层语义信息。神经网络的好处在于不需要人为再设计特征抽取方法和手写规则便可以得到具有语义信息的句子编码，进而输入到分类层对实体关系分类实现关系抽取的功能。现有的模型对文本的特征表示集中于词嵌入和位置嵌入，对其它相关语法特征考虑较少，本课题拟针对更丰富的语法特征进行研究，如实体上下文信息、语法依赖结构等。对于词嵌入表示，随着近几年语言模型的发展，在大语料文本数据预训练词向量的应用越来越广泛，word2vec模型训练得到的词向量具有一定的缺点，得到的词向量是固定不变的，没有考虑到词语在不同语境下具有不同的语义信息，Peters[31]提出的ELMO和Devlin[32]提出的BERT都是在丰富语料文本上训练得到的语言模型，其中BERT模型在NLP各个任务上都表现出了其优越的特性，因此本课题将考虑使用多种词嵌入方法。对位置信息考虑更多种计算方法，考虑实体上下文位置信息、相对两个实体的加权位置信息等。对于句子编码层，传统有监督学习方法仅采用CNN、RNN等模型，并没有加入注意力机制，注意机制最初在图像领域中使用，通过引入注意力机制可以影响模型输出结果的倾向，使得模型更关注于任务相关的信息，受远程监督关系抽取中利用注意力机制过滤噪声的启发，本课题考虑将注意力机制引入有监督学习任务下，通过结合注意力机制和更丰富的语法特征信息学习得到性能更好的关系抽取模型。  3、实体关系抽取原型系统  基于上述研究内容，设计并实现实体关系抽取原型系统，系统主要包括数据层，模型层和可视化层。数据层基于图数据库Neo4j进行存储，Neo4j是一种图形数据库，能够存储关系抽取得到的实体-关系-实体三元组，并具有CQL查询语言提供查询功能。模型层由生成对抗网络构成，主要功能为关系文本生成功能和关系抽取功能，关系文本生成功能主要为给定句子文本和其实体对应关系，由生成模型给该文本打分，通过设置阈值决定文本是否能够加入训练集；关系抽取功能为给定文本和实体，抽取实体对之间关系。能够对实时输入文本进行实体关系抽取，并返回抽取结果。可视化层基于Django框架实现，对系统功能和数据库中存储的三元组数据查询功能做可视化展示。   * 关键技术  1. 生成对抗网络模型   生成对抗网络模型如图1所示，主要包括三个部分：编码器、选择器、判别器。  D:\djangohkh\Conference-LaTeX-template_10-17-19\architecture.png  图1 生成对抗网络模型  编码器的作用是将文本转化为低维向量，给定一个句子，编码器E将句子转变为向量x。本文采用BERT+PCNN的方式获取句子向量。其中BERT模型作为词嵌入层，PCNN作为卷积层。  基于BERT的词嵌入层：  Word2vec作为常用词向量模型难以解决一词多义的问题，而在实体关系抽取任务中，对于实体语义的准确表达非常关键，如“乔布斯创办了苹果公司”，其中“苹果”作为一个具有多种含义的实体词语，其对应的词向量应当是不同的，BERT模型很好的解决了一词多义的问题。BERT模型结构如图所示。  PCNN卷积层：  普通的CNN网络难以捕捉到实体之间的结构    选择器即为GAN中的生成网络，在编码器的基础上增加了一个输出层，给定一个含有两个实体的句子和实体间的关系，输出该句子中两个实体关系正确的概率。  判别器即为GAN中的判别网络，在编码器的基础上增加了一个多分类层，给定一个句子，输出该句子中两个实体的所属关系，其中关系类型由训练数据给定。   1. 对抗训练过程   编码器的作用是将文本转化为低维向量，给定一个句子，编码器E将句子转变为向量x。本文采用BERT+PCNN的方式获取句子向量。其中BERT模型作为词嵌入层，PCNN作为卷积层。   * 论文计划   表2 论文计划   |  |  | | --- | --- | | 2019.9－2019.12 | 阅读文献，进行前期调研，论文开题 | | 2020.1－2020.3 | 筛选出值得借鉴的论文并对相关模型进行复现 | | 2020.4－2020.6 | 对抗训练模型设计，并验证效果 | | 2020.7－2020.9 | 实体关系抽取模型设计，并验证效果 | | 2020.10－2021.1 | 整理实验数据，原型系统设计 | | 2021.2－2021.4 | 原型系统实现和论文初稿撰写 | | 2021.4－2021.6 | 论文撰写 |  * 论文进度及目标   截止目前，前四项任务已全部完成，第五项任务完成80%，完成开题计划中应该完成的工作。  本文旨在使用深度学习和传统自然语言处理任务相结合的方法，完成关系抽取算法的实现和系统的设计开发。针对关系抽取中实体一词多义的问题，本文基于BERT模型做词嵌入层，较好的解决了该问题，针对句子特征表示的问题，本文提取了句子中的词特征、实体特征、实体上下文特征以及位置特征，提出了分段卷积神经网络PCNN进行句子编码，充分挖掘了原始文本信息，鉴于如今有监督的关系抽取数据训练样本较少，本文提出了启发式算法，基于训练样本种子发现大量未标注的噪声数据，进行自动化标注，并通过训练生成对抗网络获得一个选择器，用于筛选出自动化标注的噪声数据，将标注正确的句子筛选出来用于扩充训练数据，并解决原始训练数据中样本不均衡的问题。结合以上算法，本文设计实现了基于生成对抗网络的实体关系抽取系统。  论文整体架构及内容如下：  第一章为绪论，主要介绍本论文的研究背景及意义，并通过阅读大量文献归纳国内外研究发展现状，进而提出本论文的研究内容和和组织架构。  第二章为详细介绍关系抽取算法和生成对抗网络，包括不同的关系抽取任务和算法，评价指标等。  第三章为相关技术介绍，主要介绍本论文需要用到的相关技术，包括基于Pytorch的组件研究、实体关系抽取算法的研究、BERT模型的研究，PCNN模型的研究，生成对抗网络模型的研究。  第四章为基于BERT和PCNN的编码器模型实现，以及实验设置和模型效果展示。  第五章为生成对抗网络的研究，以及实验设置和扩充数据集的效果展示。  第六章为基于生成对抗网络的实体关系抽取系统，包括系统需求分析，数据收集方式，各模块设计和系统展示。  第七章为总结与展望，客观地总结在论文的写作、算法实现和系统设计方面主要做的工作，并给出论文的写作、算法实现和系统设计和测试环节中发现的不足之处，并对以后的改进进行展望。  目前，本文已完成对基于生成对抗网络的实体关系抽取算法调研，模型搭建和调优，实验结果分析和系统设计与实现部分。  接下来的时间里，将进一步优化系统；预计在2021年3月前完成论文初稿。并广泛征求老师和同学的建议，对论文进行修改，使论文的整体结构更加合理，论述更加专业化，在2021年5月前完成论文的修改工作，产出最终的毕业论文，并完成毕业答辩各项准备工作。 | | | |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **论文进展情况**  **1、工作计划**  本文旨在：研究长文本下的阅读理解模型和多段落阅读理解中候选答案排序模型。对于长文本阅读理解问题，本文提出利用句法依存关系的注意力机制，解决噪声干扰问题。对于多段落阅读理解中候选答案的排序问题，本文提出基于文本蕴含特征和其它文本特征的候选答案排序模型，充分考虑了问题，答案和文章三者之间的关系。融合以上两种算法，本文实现了基于深度学习的多段落阅读理解系统，系统分为阅读理解分析模块和中英文开放域问答模块。具体工作计划和安排如表3所示：  表3具体工作计划和安排   |  |  |  | | --- | --- | --- | | **时间** | **工作内容** | **预期成果** | | 2019.9－2019.12 | 前期调研，确定研究方向与目标, 制订实施方案。 | 了解现有方法的不如，初步确定解决方案 | | 2020.1－2020.3 | 学习Pytorch框架，掌握Pytorch架构及工作原理，熟悉Pytorch的使用，可以搭建模型。 | 能熟练使用Pytorch框架，能够复现论文，搭建模型 | | 2020.4－2020.6 | 数据分析处理和搭建模型，完成整个模型的搭建和调优 | 完成模型编码，并撰写一篇小论文 | | 2020.7－2020.9 | 优化模型 | 优化模型，完成论文撰写 | | 2020.10－2021.1 | 搭建系统 | 发表一篇论文 | | 2021.2－2021.4 | 系统各模块整合测试，提升系统性能 | 完成系统测试 | | 2021.4－2021.6 | 论文终稿撰写，完成答辩材料的准备 | 完成论文终稿和答辩材料 |   **2、实际进展情况**  目前，实际进展情况如表4所示：  表4 实际进展情况   |  |  |  | | --- | --- | --- | | **时间** | **工作成果** | **完成情况** | | 2019.9－2019.12 | 了解现有方法的不如，初步确定解决方案 | **已完成** | | 2020.1－2020.3 | 能熟练使用Pytorch框架，能够复现论文，搭建模型 | **已完成** | | 2020.4－2020.6 | 完成模型编码，并撰写一篇小论文 | **已完成** | | 2020.7－2020.9 | 优化模型，完成论文撰写 | **已完成** | | 2020.10－2021.1 | 发表一篇论文 | 未完成 | | 2021.2－2021.4 | 完成系统测试 | 未完成 | | 2021.4－2021.6 | 完成论文终稿和答辩材料 | 未完成 | |
| **工作成果**   * 已完成的学位论文工作的内容和阶段性成果   Yanhua Yu, Kanghao He, Jie Li.2020. Adversarial Training for Supervised Relation Extraction. Tsinghua Science and Technology.（在投）。  **1、已完成学位论文工作的内容**  本文。  **2、取得的阶段性成果**  **1）数据处理和评估方法研究**  **①数据集描述和处理：**  本文采用SQuAD2.0[21][22], DuReader[23]和DRCD[24]三个数据集对基于BERT和句法依存关系注意力机制的阅读理解模型进行评估。  SQuAD2.0[22]是斯坦福提出的英文阅读理解数据集，数据集统计结果如表6所示。答案都给出了开始位置和结束位置，并且每个问题只有一个对应段落。对于没有答案的问题，数据集还给出了可疑答案。  表6 SQuAD2.0数据集统计结果   |  |  |  | | --- | --- | --- | |  | all | no-answer | | train | 130,319 | 43,498 | | dev | 11,873 | 5,945 | | test | 8,862 | 4,332 |   DuReader[23]是百度提出的中文多段落阅读理解数据集，数据集统计结果如表7所示。其中各个类型的问题和答案的比例如表8所示。在阅读理解实验中，文本只用到其中的golden paragraph和golden answer两个值。  表7 DuReader数据集统计结果   |  |  |  | | --- | --- | --- | | train | dev | test | | 271,574 | 10,000 | 20,000 |   表8 DuReader数据集各个类型的问题和答案的比例    DRCD[24]是一个中文的阅读理解数据集，数据形式和SQuAD1.0类似，即所有问题都是可回答的。数据集统计结果如表9所示。答案都给出了在文中的开始位置和结束位置，并且每个问题只对应一个段落。但是每个段落可能对应多个问题。  表9 DRCD数据集统计结果   |  |  |  | | --- | --- | --- | | train | dev | test | | 26,932 | 3,524 | 3,485 |   三个数据集的对比结果如图4,5所示：    图4 各数据集训练集测试集和验证集数据量对比    图5 各数据集段落平均长度对比  本文采用WebQuestions[25],Quasar-T[26], WikiMovies[27], SQuADopen[21][22]和CuratedTREC[28]五种数据集对基于候选答案对的融合文本蕴含特征和其它文本特征的候选答案排序算法进行验证，数据集对比结果如表10，图6所示。除去Quasa-T和SQuAD数据集外，WebQuestions, WikiMovies和CuratedTREC数据集都没有提供候选段落，因此本文使用[12]对数据集处理后的结果进行试验。处理后的数据集已经进行了文档检索，每个问题有对应的五个分数最高的段落，可以直接用于本文阅读理解和候选答案排序任务。  表10 各数据集训练集和测试集的数量   |  |  |  | | --- | --- | --- | |  | Train | Test | | Quasar-T | 28,496 | 3,000 | | SQuADopen | 71,231 | 10,570 | | WikiMovies | 36,301 | 9,952 | | CuratedTREC | 3,464 | 694 | | WebQuestions | 4,602 | 2,032 |     图6 各数据集训练集和测试集数据量对比  WebQuestions[25]是构建的问答数据集，是从Freebase KB中获取的数据，本文使用[12]对数据进行处理后的版本，即把Freebase-IDs替换成答案。  WikiMovies[27]数据集包含来自电影行业的数千个问答对。它的设计使得所有问题都可以由知识库（即OpenMovie Database）或全文内容（Wikipedia）来回答。  CuratedTREC[28]数据集是来自文本检索会议（TREC）的问答任务数据集。  SQuADopen[21][22]数据集是斯坦福的问答数据集，包含100000个问题-答案-段落三元组，本文用SQuADopen数据集，即忽略了段落信息。  Quasar-T[26]数据集由问答对构成，并通过搜索引擎从ClueWeb09中抓取相关段落。  **②评估方法选择：**阅读理解测评方法可分为两类，分别是针对抽取式阅读理解的评估方法和针对自由回答形式的阅读理解的评估方法。  针对抽取式的阅读理解评估方法有两种，分别是EM(Exact match)和F1 score。  EM评估方法：如果预测答案和正确答案完全相同则为1，否则为0。  F1 score评估方法：计算预测答案和正确答案的平均词语重合程度。预测答案和正确答案都以词袋的方式表示，词级别的F1 score计算公式为：   |  |  | | --- | --- | |  | (22) | |  | (23) | |  | (24) |   针对自由回答形式的阅读理解任务，评估方法类似机器翻译和摘要生成算法的评估方法，分别为计算ROUGE和BLEU两个指标。  ROUGE(Recall-Oriented Understudy for Gisting Evaluation)是一种基于召回率的相似性度量方法，ROUGE包含四种方法，GOUGE-N,ROUGE-L,ROUGE-W,ROUGE-S，对于阅读理解算法的评估，我们主要是用ROUGE-L评估方法，L即是LCS(longest common subsequence),最长公共子序列，ROUGE-L的计算公式为：   |  |  |  | | --- | --- | --- | |  | (25) | | |  | (26) | |  | (27) |   其中表示正确答案和预测答案的最长公共子序列长度，m,n分别表示正确答案和预测答案的长度。  BLEU(Bilingual Evaluation Understudy) BLEU 采用一种N-gram的匹配规则，计算公式为：   |  |  | | --- | --- | |  | (28) |   其中表示单词k(n-gram)在候选答案中出现的次数，表示单词k(n-gram)在正确答案中出现的最多次数，这是标准答案有多个情况下的计算公式。  为了防止随着候选答案长度变短而效果变好的情况,BLEU在最后的平分结果中引入了长度惩罚因子:   |  |  | | --- | --- | |  | (29) |   其中代表候选答案的长度，代表正确答案的长度。  最终的评价公式为：   |  |  | | --- | --- | |  | (30) |   BLEU的原型系统采用的是均匀加权，即。N的上限取值为4。  **2）基于BERT和句法依存关系注意力机制的阅读理解模型实验结果**  本文实现了基于TensorFlow的阅读理解模型，采用stanfordcorenlp接口获取句法依存关系树。采用长度为200，步长为150的滑动窗口进行阅读理解。在SQuAD2.0,DuReader和DRCD三个数据集上进行模型测试。针对SQuAD2.0和DRCD数据集，评测标准为EM和F1。针对DuReader数据集match-lstm模型和BiDAF模型的测评标准为BLEU和Rouge-L，BERT finetune模型和本文使用的模型的测评标准为EM和F1。针对SQuAD2.0,如果答案的开始位置和结束位置都为0，则认为答案不存在。针对DuReader数据集，本文仅使用golden paragraph和golden answer进行测评。模型结果如表11所示：  表11实验结果    实验结果显示，本文提出的模型在三个数据集上的效果均高于BERT finetune baseline模型。其中在DuReader数据集上的效果最明显，分析原因可能是因为DuReader数据集的平均长度较长，图7为各个数据集文章长度的对比。虽然DRCD的段落长度也很长，但是由于数据集简单，并且模型效果已经接近人工标注的效果，所以效果提升不明显。    图7 各数据集段落平均长度对比  通过横向对比，DuReader数据集的效果最差，是因为百度阅读理解数据集的答案类型比较复杂，图8为各种类型答案所占的比例，任务难度本身就比较大。    图8 DuReader数据集各种类型答案所占的比例  而SQuAD2.0数据集由于存在没有答案的问题，图9为有答案和没有答案的问题占的比例，所以导致模型效果没有达到最优。    图9 SQuAD2.0数据集有答案和没有答案的问题占的比例  **3）基于候选答案对的融合文本蕴含特征和其它文本特征的候选答案排序算法效果**  本文采用基于BERT句子对分类模型，构建基于（答案出自的句子，问题+答案）对的蕴含关系模型，将模型得到的分数作为一维特征。训练过程为，根据候选答案中的错误答案构建负样本，对正确答案采用远距离监督的方法构建正样本。在模型测试的过程中，为了便于与RankQA模型比较，本文首先使用RankQA模型使用的阅读理解模型就行测试。然后将阅读理解模型换成本文提出的基于BERT和句法依存关系注意力机制的阅读理解模型进行测试，测试结果如表12所示：  表12 实验结果，评估指标为EM    通过分析结果，本文提出的候选答案排序算法在WebQuestion，Quasar-T，WikiMovie和SQuADopen数据集上的效果优于RankQA模型。使用本文提出的阅读理解模型，在WebQuestion，Quasar-T和SQuADopen数据集中取得了几个对比模型中最好的效果。为了把模型应用到开放域的问答系统中，本文把所有数据集组合到一起进行模型验证，模型效果虽然低于最好的效果，但是效果相差不大，所以本模型是可以应用到开放域问答模块中的。  同时本文统计了只根据候选答案分数进行排序结果正确，而用本文提出的排序方法进行排序结果错误的数量，占所有采用本文排序方法结果错误的比例，统计结果如图10所示：    图10 模型效果对比  为了分析排序所使用特征的重要程度，本实验统计了针对每个数据集对结果影响最大的特征，统计结果如表13所示：  表13对候选答案排序影响最大的特征，评估指标为EM    通过分析，WebQuestion和WikiMovies数据集对文本蕴涵关系特征最敏感，而Quasar-T和CuratedTREC数据集对于候选答案分数特征最敏感，SQuADopen数据集对问题和段落的相似度最敏感。  文本同时对中文数据集做了测试，本文与百度阅读理解leadboard排行榜前九名对比，本文提出的算法效果达到了前九名的平均效果，结果如图11所示：    图11 DuReader数据集测试结果  **4）基于深度学习的多段落阅读理解系统**  结合上文提出的两种技术，本系统主要实现以下两个大的模块：一、阅读理解分析模块，由于深度学习模型的可解释性较差，所以需要对如attention矩阵，问题和文章中词的重要程度进行可视化分析，因此本文实现了阅读理解分析模块。二、多段落阅读理解可以应用在开放域的问答系统中，为了充分验证模型的效果和实用性，本文基于多段落阅读理解模型，实现了中英文开放域问答模块。  **①系统整体架构设计**  本文在结合网络爬虫数据和现有数据集的基础上，基于上文提出的两种模型，构建了一个基于深度学习的多段落阅读理解系统。系统整体架构如图12所示：    图12 系统架构图  系统自底向上主要包括基础支撑层，数据处理层，算法层和web展示层。  第一层是基础支撑层，包括GPU用于加速计算，TensorFlow框架用于搭建深度学习模型，scrapy爬虫用于扩充数据，本文主要抓取了百度知道的数据，Elasticsearch用于根据问题搜索相关文档。第二层是数据处理层，包括对数据获取，即检索到相关文档。候选段落获取，即获得用于回答问题的候选段落。以及数据预处理工作等。第三层是算法层，主要包括上文提出的两种算法和搜索引擎算法。第四层是web展示层，采用flask实现Python接口，用css，vue和js实现前端编程。  **②系统处理流程**  系统的整体处理流程如图13所示：    图13 系统流程图  系统分为模型离线训练和实时分析两个部分。在实时分析部分，会根据任务的不同调用不同的接口和模型，对于阅读理解分析模块，接口会返回问题的答案，attention矩阵，段落和问题中的重要词语。对于开放域问答模块，首先根据问题搜索相关段落，然后基于上文提出的多段落阅读理解算法生成最佳答案，并以对话框的形式返回问题的答案，同时还可以查询答案对应的段落和候选答案。   * 主要创新点   （1）实现基于生成对抗神经网络的关系抽取模型。  本文  （2）实现基于BERT的文本特征向量表示模型。  本文 |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **计划及进度安排**   |  |  |  | | --- | --- | --- | | 时间 | 研究内容 | 预期效果 | | 2019.9－2019.12 | 阅读文献，进行前期调研，论文开题 | 提出并确定研究方向 | | 2020.1－2020.3 | 筛选出值得借鉴的论文并对相关模型进行复现 | 复现模型，验证效果，发现不足 | | 2020.4－2020.6 | 对抗训练模型设计，并验证效果 | 完成对抗训练相关实验 | | 2020.7－2020.9 | 实体关系抽取模型设计，并验证效果 | 完成实体关系抽相关实验 | | 2020.10－2021.1 | 整理实验数据，原型系统设计 | 发表一篇学术论文 | | 2021.2－2021.4 | 原型系统实现和论文初稿撰写 | 完成原型系统部署和论文初稿撰写 | | 2021.4－2021.6 | 论文撰写 | 完成毕业论文撰写 | |
| **问题及整改方案**  1.论文后期工作中遇到的问题  （1）通过实验发现，仅通过生成对抗网络增强数据对模型效果的提升有待提高，生成模型生成的数据质量不高，需要探索更有效的方法进行训练。  （2）对文本的特征提取不够丰富，对关系抽取结果有一定的影响。  2.整改方案  （1）使用多种数据集进行对抗训练，提升生成器生成数据的质量。  （2）考虑提取句子上下文的语义特征，进一步丰富原始文本的特征向量表示。 |

|  |
| --- |
| **参考文献**  1.Sundheim B M,Chinchor N A.1993.Survey of the message understanding conferences[A].//Proceedings of the workshop on Human Language Technology[C].pages:56-60.  2. Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda.2006.ACE 2005 multilingual training corpus. Linguistic Data Consortium, Philadelphia.pages:57.  3. Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva,et al.2009.Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In Proceedings of the 5th International Workshop on Semantic Evaluation, pages 33–38.  4. Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In Proceedings of ACL-IJCNLP. pages 1003–1011.  5. Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. Machine learning and knowledge discovery in databases, pages 148–163.  6. Nanda Kambhatla. 2004. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics on Interactive poster and demonstration sessions.  7. Fabian M. Suchanek, Georgiana Ifrim, and Gerhard Weikum. 2006. Combining linguistic and statistical analysis to extract relations from web documents. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 712–717.  8. Longhua Qian, Guodong Zhou, Fang Kong, Qiaoming Zhu, and Peide Qian. 2008. Exploiting constituent dependencies for tree kernel-based semantic relation extraction. In Proceedings of the 22nd International Conference on Computational Linguistics, pages 697–704.  9. Giuliano C,Lavelli A,Pighin D,et al.FBK-IRST:Kernel methods for semantic relation extraction[A].2007.//Proceedings of the 4th International Workshop on Semantic Evaluations[C].pages:141-144.  10. Brin S.Extracting patterns and relations from the world wide web［M］.1999.Berlin: Springer Heidelberg,pages:172-183．  11. Liu CY,Sun WB,Chao WH,et al.Convolution Neural Network for Relation Extraction[C].2013.International Conference on Advanced Data Mining and Applications. Springer, Berlin, Heidelberg, pages:132-140.  12. Daojian Zeng,Kang Liu,Siwei Lai,Guangyou Zhou,Jun Zhao,et al.2014.Relation classification via convolutional deep neural network. In Proceedings of COLING, pages 2335–2344.  13. Santos CND,Xiang B,Zhou B.2015.Classifying Relations by Ranking with Convolutional Neural Networks[J]. Computer Science, pages:132-137.  14. Dongxu Zhang and Dong Wang. 2015. Relation classification via recurrent neural network. arXiv preprint arXiv:1508.01006.  15. Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attentionbased bidirectional long short-term memory networks for relation classification. In Proceedings of The 54th Annual Meeting of the Association for Computational Linguistics, page 207.  16. Linlin Wang, Zhu Cao, Gerard de Melo, and Zhiyuan Liu. 2016. Relation classification via multi-level attention cnns. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics.  17. Rui Cai, Xiaodong Zhang, and Houfeng Wang. 2016. Bidirectional recurrent convolutional neural network for relation classification. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics.  18. Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In Proceedings of EMNLP, pages 1753–1762.  19. Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In Proceedings of ACL, volume 1, pages 2124–2133.  20. Yankai Lin, Zhiyuan Liu, and Maosong Sun.2017.Neural Relation Extraction with Multi-lingual Attention. In Proceedings of ACL.  21. Wenhan Xiong, Mo Yu, Shiyu Chang, Xiaoxiao Guo, and William Yang Wang. 2018. One-shot relational learning for knowledge graphs. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages:1980-1.  22. Goodfellow, Ian, Pouget-Abadie, Jean, Mirza, Mehdi, Xu, Bing, Warde-Farley, David, Ozair, Sherjil, Courville, Aaron, and Bengio, Yoshua.2014.Generative adversarial nets. In Advances in Neural Information Processing Systems (NIPS) 27. Curran Associates,.  23. Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2016. Adversarial training methods for semi-supervised text classification. arXiv preprint arXiv:1605.07725.  24. Yu L , Zhang W , Wang J , et al. 2016.SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient[J].  25. Yi Wu, David Bamman, and Stuart Russell. 2017. Adversarial training for relation extraction. In Proceedings of EMNLP, pages 1778–1783.  26. 冯冲, 康丽琪, 石戈, 黄河燕. 融合对抗学习的因果关系抽取. 自动化学报, 2018, 44(5): 811−818.  27. Pengda Qin, Weiran Xu, and William Yang Wang. 2018. DSGAN: Generative adversarial training for distant supervision relation extraction. In Proceedings of ACL, pages 496–505.  28. Xu Han, Zhiyuan Liu, and Maosong Sun. 2018. Denoising distant supervision for relation extraction via instance-level adversarial training. arXiv preprint arXiv:1805.10959.  29. Xiaozhi,Wang,Xu Han,Zhiyuan Liu,Maosong Sun,Peng Li.2019.Adversarial Training for Weakly Supervised Event Detection. 998-1008. 10.18653/v1/N19-1105.  30. Springenberg J T. 2015.Unsupervised and Semi-supervised Learning with Categorical Generative Adversarial Networks[J]. Computer Science.  31. Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextualized word representations. arXiv preprint arXiv:1802.05365.  32. Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proc. NAACL-HLT. |
| |  |  |  |  | | --- | --- | --- | --- | | 姓 名 | 职 称 | 职务 | 工 作 单 位 | |  |  | 成员 | 北京邮电大学 | |  |  | 成员 | 北京邮电大学 | |  |  | 成员 | 北京邮电大学 | |  |  | 成员 | 北京邮电大学 |   **评审小组** |

|  |
| --- |
| **导师评语** |
| 导师：  日期： 年 月 日 |
| **阶段报告小组意见：** |
| 负责人：  日期： 年 月 日 |
| **学院意见：** |
| 负责人：  日期： 年 月 日 （签章） |