|  |  |
| --- | --- |
| 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 | 其他 |
| 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912 | 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. 基于BERT+PCNN的文本编码器   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。编码器结构如图1所示。    图1 BERT+PCNN编码器  **基于BERT的词嵌入层**：BERT模型是一种双向Transformer的编码器，通过Transformer的多头注意力机制和Masked的训练方法获得了一个较为完整的语言模型，解决了Word2vec模型中的一词多义问题，由BERT语言模型获得词向量完整的挖掘了其左右两边的语义信息。  **位置特征向量：**位置特征向量由句子中每个单词与两个实体的相对位置获得。对于每个单词我们计算出其距离两个实体的相对位置，如图2所示，单词moved相对实体Boss和Office的距离为1和-3。    图2 位置向量计算  最终句子中每个单词的特征向量表示由词向量、两层位置向量表示得到：  基于BERT获得句子表示矩阵，其中n是句子里包含的单词个数，d是特征向量的维度，为词向量维度，为位置向量维度。  **PCNN卷积层**：普通的CNN网络难以捕捉到实体之间的结构信息，由于实体关系抽取任务的特性，在给定句子和两个实体的条件下，我们可以通过实体位置分段卷积，利用卷积网络的特性挖掘文本中实体上下文之间的语义信息，PCNN模型结构如图3所示。    图3 分层卷积网络  根据两个实体，我们将句子分为三个部分，通过将这三部分的卷积向量拼接最终可以得到句子的向量表示。该过程分为两个步骤，第一步为卷积，将句子向量矩阵与卷积核卷积相乘得到卷积向量，对于一个句子获得的卷积向量如公式2所示。  卷积层输入词特征向量维度为n\*d，过滤器维度为w\*d，得到卷积向量维度为n+w-1。通常需要多个过滤器来捕捉实体上下文信息，本文采用了k过滤器拼接获取卷积向量，因此最终获得的向量矩阵。  第二步为分段池化，对于每个过滤器获得的卷积向量，可以将其分为三个部分，通过对这三部分分别采用最大化池化进行维度压缩，再将这三部分拼接起来获得一个3维向量，如公式3所示。  最终经过一个tanh激活函数获得句子的最终表示。   1. 生成对抗网络   生成对抗网络模型如图4所示，其中网络模型主要包括三个部分：编码器、选择器、判别器。  D:\djangohkh\Conference-LaTeX-template_10-17-19\architecture.png  图4 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。生成对抗网络中生成器和判别器编码器的结构一致。  如图4所示，数据分为有监督数据和噪声数据，有监督数据为已经标注好的文本数据（Supervised Data，简称SD），而噪声数据是由一种发现策略自动化标注的不可靠数据（Noisy Data，简称ND）。本文提出了一种启发式算法，根据已有的标注数据进行自动化标注。通过该方法可以获取大量标注的噪声数据，这部分数据中有一部分为标注正确的数据，通过利用该噪声数据训练GAN网络。选择器的训练目标为选择噪声数据里的数据迷惑判别器，判别器的训练目标为对于给定句子s，最大化，和。由此可以得到生成对抗网络的训练目标如公式4所示。  判别器：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。  给定一个包含两个实体的句子，判别器需要预测该句子属于关系r的概率。通过编码器获得句子表示后，该特征向量输入到一个softmax层获得概率如公式5所示。  其中为转置矩阵参数，，为关系类型种类，由训练数据给定，为句子特征向量维度。r为关系的one-hot编码。判别器将有监督数据给出高概率而对于噪声数据中的实例给出低概率。由此得到判别器的损失函数如公式6所示。  其中由选择器对于句子的预测概率获得。  选择器：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。给定句子s经过编码器后获得句子特征向量，选择器计算ND数据中关系标注准确的概率，如公式7所示。  其中是转置矩阵的参数，选择器的目标为使得的概率最大化，由此可得选择器的损失函数如公式8所示。  其中P(r|s)由判别器计算得到。判别器和选择器对抗训练后，可以训练出一个选择器，能够将噪声数据里的有效数据筛选出来，一个判别器能够对给定的句子进行关系抽取，预测其所属的正确关系。   1. 基于生成对抗网络的关系抽取系统研究     图5 关系抽取系统架构图  基于以上算法，本文搭建关系抽取系统如图5所示，该系统能够对数据库中的关系数据进行查询和可视化展示，能够展现实体节点属性和实体节点之间的关系。可以对用户的文本输入进行关系抽取，系统加载生成对抗网络中的判别模型对输入进行关系抽取，并返回实体关系。能够利用选择器判断用户给定的标注化文本数据的可靠性。   * 论文计划   表1 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。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**  本论文的主要工作包括：研究生成对抗网络模型在实体关系抽取领域的应用。针对传统文本特征表示一词多义的问题，本文提出了使用BERT和分层卷积网络PCNN进行解决。针对常规的文本特征单一等问题，本文提出了基于实体相对位置的特征向量表示。鉴于有监督关系抽取任务数据的样本较少，本文提出了启发式算法自动标注化数据，并利用生成对抗网络对自动标注的数据进行筛选，从而扩充训练数据，提高在SemEval 2010 Task8数据集上的效果。融合以上两种算法，本文实现了基于生成对抗网络的实体关系抽取系统，系统可以接受用户输入的句子，返回实体关系，能够对给定关系的句子判别其正确性。具体工作计划和安排如表2所示：  表2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3所示：  表3 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   **1、已完成学位论文工作的内容**  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进行关系抽取任务的研究。  本文对BERT模型，PCNN神经网络，生成对抗网络等进行深入研究，提出了基于BERT和PCNN的生成对抗网络模型，并基于Pytorch进行模型搭建。本文提出了启发式算法，根据已有训练数据发现大量噪声数据进行自动化标注，从而训练生成对抗网络，并利用GAN中的生成网络进行噪声数据的筛选，筛选出有效数据后完成对训练数据的扩充。GAN中的判别模型能够有效的对给定句子的两个实体进行关系抽取，通过噪声数据的对抗训练，具有一定的抗噪能力。  **已完成的工作如下：**   1. 环境搭建，数据处理如下表所示   表4环境配置和选取的数据集   |  |  | | --- | --- | | 环境 | 版本号 | | Pytorch | 1. 0 | | CUDA | 9.0 | | GPU | 1080Ti，11G | | 数据集 | 语言 | | Semeval 2010 Task 8 | 英文 | | NewYork Times | 英文 |  1. 完成BERT、PCNN神经网络、生成对抗网络的研究。 2. 完成基于BERT和PCNN的生成对抗网络模型的搭建和训练。   **2、取得的阶段性成果**  **1）数据处理和评估方法研究**  本文采用在实体关系抽取领域广泛使用的Semeval 2010 Task8数据集，该数据集包含8000训练数据和2717测试数据，每个数据实例包含一对实体和标注好的关系，关系种类包含9种，且具有方向属性，分别为CauseEffect, Instrument-Agency, Product-Producer, ContentContainer, Entity-Origin, Entity-Destination, ComponentWhole, Member-Collection 和Message-Topic，其他类型关系定义为Other。本文通过NewYork Times的语料进行噪声数据的自动化标注，为了解决原始训练数据样本不均衡的问题，本文根据原始样本分布对每种关系进行了不同程度的噪声数据扩充，如图6所示。    图6 数据集关系种类实例分布  其中蓝色为SemEval数据集，红色部分为扩充的噪声数据，经过启发式算法发现，使得各种关系类型数据趋近于平衡。  **②评估方法选择：**本文遵循官方的评价标准分别为微平均（Micro-averaged）和宏平均（Macro-averaged）的精度（P）、召回率（R）、F1值，其中P、R、F1值计算公式如下：  对于多标签分类指标，Micro-averaged是对数据集中的每一个实例不分类别进行统计建立全局混淆矩阵，然后计算相应指标。Macro-averaged先对每一个类统计指标值，然后在对所有类求算术平均值。  **2）基于生成对抗网络的关系抽取任务实验结果**  本文通过网格搜索设置不同参数训练模型，最终模型参数设置如图7所示。其中PCNN模型基于Zeng[18]论文中的设置不变。    图7 模型参数设置  各种神经网络在该数据集上的表现如表5所示，本文与近几年的SVM、RNN、CNN、CR-CNN、Attetnion-CNN和Attention-BiLSTM模型进行了对比，效果均优于以上模型。  表5关系抽取效果对比   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | Method | Macro-averaged | | | | | | P | R | | F1 | | | SVM (Rink and Harabagiu, 2010) | 82.25 | | 82.28 | | 82.2 | | RNN (Socher et al., 2012) | N/A | N/A | | 77.6 | | | CNN (Zeng et al., 2014) | N/A | N/A | | 82.7 | | | CR-CNN (Santos et al., 2015) | 83.7 | 84.7 | | 84.1 | | | Attention CNN (Shen and Huang, 2016) | N/A | N/A | | 85.9 | | | Attention BiLSTM (Lee et al., 2019) | N/A | N/A | | 85.2 | | | **PCNN-AD (ours)** | **87.2** | **88.1** | | **87.6** | |   为了证明生成对抗网络的选择器的效果，本文将原始CNN和PCNN模型分别在噪声数据和经过筛选的数据两种情况下分别训练，得出结果如表6所示。  表6 在不同训练数据下的训练效果   |  |  |  |  | | --- | --- | --- | --- | | Method | Dataset | | | | SemEval | +Noisy | +Filtered | | CNN | 82.7 | 81.8 | 83.3 | | PCNN | 83.1 | 82.9 | 84.5 |   由表6可以看到，在噪声数据的影响下原有模型的关系抽取效果有着明显的下降，因为训练集混入了大量不准确数据，而经过GAN网络中选择器筛选后得到的训练数据有着明显的效果提升，这说明了选择器能够有效的选择出用于训练的可靠数据。  **3、已发表的与学位论文相关的学术论文等**  Yanhua Yu, Kanghao He, Jie Li.2020. Adversarial Training for Supervised Relation Extraction. Tsinghua Science and Technology.（在投）。   * 主要创新点   （1）实现了生成对抗网络模型。  生成对抗网络中的生成模型具有生成训练数据的能力，针对目前关系抽取的方法，有监督学习方法难以获得大量标注数据集，远程监督学习通过对齐知识库获得的标注数据集噪声大。基于生成对抗网络训练得到的生成模型，能够从大量噪声数据中选择合适的数据当作训练数据，从而实现扩充训练数据的效果。这种方法能够针对现有数据针对性的扩充数据集且具有对抗噪声的效果。现有的对抗训练方法通常是在嵌入层添加一段随机噪声来影响关系抽取模型。本课题通过在现实语料集上发现大量噪声数据，通过生成模型进行选择与关系抽取模型进行对抗训练，可以有效提高关系抽取模型针对真实数据的抗噪声能力。  （2）实现基于BERT+PCNN文本特征向量表示模型。  传统词嵌入表示法如word2vec训练得到的词向量具有一定的缺点，同一个词在不同文本和句子中表达的意思可能不同，而word2vec训练的词向量是固定不变的，基于BERT模型训练的词向量考虑了文本多义性表达。在位置特征表示方面，通常用两组相对位置向量进行表达，本课题考虑加权获得的相对位置向量，这种方法既简化了特征维度，也考虑到了相对两个实体的位置关系。引入注意力机制，使得文本特征最终表示更关注与实体关系相关的信息上，得到更具有语义信息的表达。利用PCNN网络分段卷积提取实体上下文语义信息。 |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **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 | | 吴斌 | 教授 | 成员 | 北京邮电大学 | | 石川 | 教授 | 成员 | 北京邮电大学 | | 王柏 | 教授 | 成员 | 北京邮电大学 | | 于艳华 | 副教授 | 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） |