Inline Text Wrapping Picture

北京邮电大学

硕士研究生学位论文开题报告

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: 大数据与智能信息处理

导师姓名: 于艳华

攻 读 学 位: 工学硕士

2019年11月18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 | | |
| 选题来源 | 其他 | 论文类型 | 应用研究 |
| 开题日期 | 2019-11-01 | 开题地点 | 教三615 |
| **一、立题依据（包括研究目的、意义、国内外研究现状和发展趋势，需结合科学研究发展趋势来论述科学意义；或结合国民经济和社会发展中迫切需要解决的关键科技问题来论述其应用前景。附主要参考文献目录）（不少于800字）**  一、研究背景  近些年来互联网飞速发展，网络社交媒体、电子商务等新型应用随之产生，随着智能手机和个人PC普及，各类APP如雨后春笋般出现，网络用户呈爆炸式增长，同时也产生了海量的网络数据。据统计，2019年中国的网民数量已经接近9亿，而伴随网民数量增长的是互联网数据，据互联网络数据中心报告，预计到2020年全球数据总量将超过40ZB。随着互联网技术的不断发展，互联网数据从量和形式都发生了巨大改变。  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。  二、研究意义  文本数据是当今互联网数据中非常重要的一种数据，它有着与图像、视频等其他网络数据不同的特点，体现在文本数据包含信息丰富，实时性强、多语言化、适用范围广等。人们的日常交流、新闻媒体报道都离不开文本数据，如何从这些数据中挖掘出有价值的信息具有重大意义。实体关系抽取技术可以从这些非结构化文本数据中抽取到蕴含语义知识的三元组数据，这些数据去除了原始文本中的大量冗余信息，在许多领域都有极大的应用价值，下面将简要介绍该技术的应用场景。  2012年Google率先提出了知识图谱（Knowledge Graph，简称KG）这一概念，表示将在搜索结果中加入知识图谱的功能。其目的在于优化搜索引擎，提高搜索反馈结果质量。知识图谱是一种语义网络，它由实体关系三元组（实体-关系-实体）构成，其中实体作为结点，实体间的语义关系作为边。实体关系抽取为构建知识图谱提供了大量三元组数据，通过从海量非结构化文本中挖掘具有语义价值的三元组数据，加入到知识图谱，不断迭代构建大型知识图谱。Google知识图谱已经包含了5亿多个实体，超过35亿条三元组知识信息，为返回优质搜索结果提供了坚实基础。而随着知识图谱技术的不断发展与演进，推荐系统、自动问答、对话机器人、机器阅读理解等诸多领域中都可以应用知识图谱这一技术。例如商品推荐系统里，通过挖掘用户与商品之间的关系，可以知道用户的喜好进而推荐相似产品。  实体关系抽取技术作为构建知识图谱的上游技术，为知识图谱的构建提供了大量实体关系三元组数据，随着知识图谱技术的广泛应用，实体关系抽取技术发挥着重要作用并具有重大意义和价值。目前针对中英文语料的实体关系抽取研究工作非常丰富，但是这些语料的数据来源复杂且广泛，没有针对特定领域进行抽取的数据集。不同领域数据包含的关系种类和类型都不相同，针对特定领域数据进行关系抽取对该领域知识图谱的构建具有重大意义。本课题针对新闻事件数据进行关系抽取，旨在针对互联网上海量新闻文本里的关键人物和关键事件进行关系抽取。  三、国内外研究现状和发展趋势  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。  早期的关系抽取方法主要基于手工设计的规则和模板，该方法需要针对特定数据集，由领域专家来编写抽取规则，利用定义好的抽取模板与文本匹配做抽取任务，这种方法虽然在特定数据集上准确率高，但是费时费力且不具备良好的泛化性。  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。  四、主要参考文献  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.  33. Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018a. Deep contextualized word representations. In NAACL.  34. Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In NIPS.  35. Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In Empirical Methods in Natural Language Processing (EMNLP), pages 1532– 1543. | | | |

|  |
| --- |
| **二、研究内容和目标（说明课题的具体研究内容，研究目标和效果，以及拟解决的关键科学问题。此部分为重点阐述内容）（不少于2500字）**  一、主要研究内容  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。原始 GAN 理论中，并不要求 G 和 D 都是神经网络，只需要是能拟合相应生成和判别的函数即可。由于神经网络拟合函数的优越性能，一般均使用深度神经网络作为 G 和 D 。一个优秀的GAN应用需要有良好的训练方法，否则可能由于神经网络模型的自由性而导致输出不理想。  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u∈U，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、基于注意力的文本特征表示  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。  二、研究目标和效果  我们在关系抽取的工作希望能够达到以下目标：1）基于生成对抗网络模型，构建生成模型和判别模型。2）实现对抗训练算法，完成生成模型和判别模型的训练，使得生成模型能够生成训练数据，解决训练数据量少的问题，判别模型能够对文本进行关系抽取。3）针对关系抽取模型，通过引入更丰富的词嵌入向量表示、考虑更多的语法结构信息，实现更好的关系抽取效果。4）引入注意力机制到神经网络中，使得关系抽取模型关注与任务相关的信息而避开无关信息的影响。  对于关系抽取模型的效果，以Precision，Recall，F1-score，P-R曲线作为评价标准。对于生成模型，通过对抗训练实现扩充训练数据集的功能，最终将生成对抗网络集成于关系抽取系统中。  三、拟解决的关键科学问题  1）解决关系抽取任务训练数据量少的问题。现存有监督学习的关系抽取任务缺乏大量有标注数据，远程监督关系抽取任务包含大量噪声数据，通过构建生成对抗网络，基于少量有标注数据和关系种子，从Wiki百科等语料库发现噪声数据自动标注，训练生成网络模型，实现生成有标注训练数据的功能。  2）解决关系抽取模型考虑特征少的问题。现有关系抽取模型考虑特征较少，在有监督学习领域还没有模型考虑引入注意力机制。在文本特征表示方面，综合考虑更多文本特征，利用目前主流的word2vec和BERT等模型进行词向量嵌入，利用加权位置向量等特征，结合注意力机制，通过注意力机制使得最终表示的文本向量更关注与实体关系相关的信息，避免其他无关信息的影响，进而提高关系抽取模型的性能。 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **三、研究方案设计及可行性分析（包括：研究方法，技术路线，理论分析、计算、实验方法和步骤及其可行性等）（不少于800字）**  **一、研究方案设计**  1.研究方法  针对本课题的研究内容和特性，达到研究目标和完成毕业论文，采用如下的研究方法：  （1）理论知识和技术准备：采用文献调查法，利用学校图书馆、网上相关学术数据库等资源，了解当前最新的关系抽取相关的研究进展。学习自然语言处理，深度学习以及相关技术实现，选择部分论文进行复现。  （2）设计技术方案：在理论和技术准备的基础上，结合自己的创新点，提出新的算法，设计出本课题详细的研究方案；在对设计研究方法结果分析的过程中，比较不同算法，证明其有效性和可行性。  （3）实现及测试：实现自己的算法并在公开公认的数据集上进行测试。  （4）发表一篇小论文。  （5）完成毕业论文：整理文献资料、代码和数据等，完成论文。  2、技术路线  实验步骤初步设计如下：  （1）数据采集与预处理：根据调研的论文，选择公开认可度高的数据集作为基本训练数据，在此基础上选取关系种子，在Wiki百科或者纽约时报等大型语料集上发现噪声数据集。对所得到的文本数据进行预处理，需要对文本数据进行分词、构建词表等操作，获得实体位置、词-实体相对位置等信息，对不同长度的文本数据作对齐处理。  （2）文本编码器：编码器主要包括文本特征表示、句子特征抽取和注意力层。编码器结构如图1所示。    图 1文本编码器结构图  文本特征主要包括词向量特征、位置向量特征等特征。词向量方面，拟采用两种方式进行词向量表示：1）使用基于word2vec模型训练得到的词向量。2）使用基于BERT预训练得到的词向量。位置向量方基于每个词分别对两个实体的相对位置。句子特征抽取基于PCNN、RNN等模型搭建，最后引入注意力机制完成整个编码器的搭建。  （3）生成对抗网络模型构建：构建生成模型和判别模型，定义模型损失函数，设计对抗训练过程。 生成模型对从Wiki百科等语料库中发现的噪声数据进行打分，其目标函数为最大化判别模型输出概率，因此得到其损失函数计算公式如3-1所示：   |  |  |  | | --- | --- | --- | |  |  | （3-1） |   其中为生成模型的输出。  判别模型的损失函数由两部分构成，一部分为判别真实训练数据得到，一部分为判别自动标注的噪声数据得到，计算公式如3-2所示：   |  |  |  | | --- | --- | --- | |  |  | (3-2） |   生成对抗网络结构如图2所示：    图 2生成对抗网络流程图  （4）模型效果评价：采用Precision，Recall，F1-score作为评价标准。  （5）关系抽取原型系统：在得到模型后，将其集成到关系抽取分析系统中，完成可视化结果展示。原型系统架构如图3所示。    图3 关系抽取系统架构图  **二、可行性分析**  1）数据预处理。数据预处理工作可以利用NLP开源工具，或者pytorch提供的Torchtext等工具，可以快速完成文本的分词、构建词表。  2）生成对抗网络搭建。生成模型和判别模型的结构较为简单，其损失函数公式定义明确，利用Pytorch可以完成两个模型搭建。  3）模型评估。采用准确率、召回率、F1值、P-R曲线等指标进行评估，这些指标计算公式明确，且可以借助matplotlib工具进行绘制相关图表。  4）系统搭建。基于Django框架可以快速完成系统的搭建，其官方网站有详细的指导手册便于学习，可以与Neo4j数据库完成后端的连接，实现可视化效果。 |

|  |
| --- |
| **四、本研究课题可能的创新之处（不少于500字）**  1、提供了一种扩充数据集的方法。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。  2、将对抗训练方法用于关系抽取模型中。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。  3、语义信息更加丰富的文本特征表示。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。 |
| **五、研究基础与工作条件（1.与本项目相关的研究工作积累基础 2.包括已具备的实验条件，尚缺少的实验条件和拟解决途径）（不少于500字）**  一、研究工作积累基础  在研究生期间学习了机器学习，统计机器学习，深度学习的内容，以及计算语言学的基础知识。对基本的NLP任务，包括句法分析，依存分析，分词等有一定了解，对深度学习在自然语言处理中的应用有一定的了解，包括词向量的构建，CNN，RNN在NLP中的典型应用，阅读调研了大量关系抽取论文和生成对抗网络论文。在实验室有其他同学对自然语言处理进行研究，对于目前研究的前沿问题进行讨论可以开阔思路。组内有做自然语言处理的导师，在遇到问题时请教他们，获取一些建议。本人在使用Pytorch框架进行深度学习模型搭建有一定的经验，能够熟练使用Pytorch搭建如CNN、RNN、BiLSTM等模型。  二、已具备的实验条件  目标所需的实验条件已具备如下所示：  硬件资源：联想P910工作站，配置（处理器：Intel(R) Xeon(R) CPU E5-2609V4/内存：16G DDR4 ECC/硬盘：1TB SATA GPU：1080Ti)  软件资源：基于Python开源框架的自然语言处理工具。本地开发环境和硬件设备，可以进行系统的开发、部署和测试。  三、缺少的实验条件和拟解决途径  为了验证已有神经网络模型在实体关系抽取的效果，需要复现一些神经网络模型，有些论文的代码并没有开源，复现工作比较困难复杂，可以通过向论文作者请求给予论文源代码或者和同学合作完成论文的复现。在系统开发上，由于开发工作量化经验不足，在具体内容上可能需要多次调整。系统开发需要熟悉Django框架和Neo4j数据库，目前还缺少这两方面的相关知识，可以通过在官网和相关技术博客学习。 |

**学位论文工作计划**

|  |  |  |
| --- | --- | --- |
| 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 | 生成对抗训练模型设计，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 | 毕业论文撰写 | 完成毕业论文撰写 |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 评  定  小  组  成  员 | 姓 名 | 职 称 | 单位名称 | 职务 |
| 吴斌 | 教授 | 北京邮电大学 | 组长 |
| 石川 | 教授 | 北京邮电大学 | 成员 |
| 王柏 | 教授 | 北京邮电大学 | 成员 |
| 于艳华 | 副教授 | 北京邮电大学 | 成员 |
|  |  |  |  |
|  |  |  |  |
| 导师意见： | | | | |
| 论文研究自然语言领域的热点之一——实体关系抽取，主要研究基于有监督学习的实体关系抽取算法，算法主要基于生成对抗网络。开题报告检索了实体关系抽取算法的相关文献，对实体关系研究现状进行了充分调研。有监督学习需要大量人工标注的语料，而远程监督学习需要远程知识库且引入了大量噪声数据，因此论文提出的利用生成对抗网络实现训练数据集扩充并得到一个具有抗噪声的实体关系抽取模型具有很好的理论和实际意义。  关于论文课题研究内容的技术可行性：报告详细阐述了生成对抗网络的模型，对网络中的生成模型和判别模型的功能、架构进行了细致的描述，生成模型作为选择器是可行的且具有一定的创新性。在文本特征提取方面，目前主流的方法包括word2vec、BERT等语言模型进行词向量提取，注意力机制对于过滤特征信息里的无关信息起到了关键作用，这些方法在实践中具有可行性。  关于论文成果要求和考核指标：选择公开，权威的数据集，测试模型的效果，以precision，recall，F1值来评价模型的效果。在数据集上具有较高的准确性和召回率。搭建实体关系抽取系统模型，将结果可视化展示。发表一篇EI检索以上学术论文。 | | | | |
| 导师（签名）：  日期： 年 月 日 | | | | |
| 开题报告小组意见： | | | | |
| 组长（签名）：  日期： 年 月 日 | | | | |
| 学院意见（签章）： | | | | |
| 负责人：  日期： 年 月 日 | | | | |