****

**专业型硕士学位论文开题报告书**

|  |  |  |
| --- | --- | --- |
| **姓名** | **：** | **焦樵** |
| **专业** | **：** | **应用统计** |
| **研究方向** | **：** | **金融统计方向** |
| **录取类别** | **：** | **非定向全日制** |
| **入学时间** | **：** | **2020.09** |
| **校内指导教师** | **：** | **魏金龙** |
| **校外指导教师** | **：** |  |
| **学院（中心）**  **名称** | **：** | **统计与数学学院** |

**中南财经政法大学研究生院**

一、基本情况

|  |  |  |  |
| --- | --- | --- | --- |
| 姓 名 | 焦樵 | 所在学院（中心） | 统计与数学学院 |
| 专 业 | 应用统计专硕 | 研究方向 | 金融统计方向 |
| 年 级 | 2020级 | 指导教师 | 魏金龙 |
| 论文题目 | 基于深度学习的金融领域命名实体识别方法研究 | | |

二、选题背景、研究意义、文献综述与实践现状

|  |
| --- |
| **（一）选题背景及研究意义**  **1）选题背景：**  随着互联网技术的进步以及世界金融的高速发展，现如今互联网每年产生的数据规模呈指数级增长，据IDC发布《数据时代2025》的报告显示，全球每年产生的数据将从2018年的33ZB增长到2025年的175ZB，相当于每天产生491EB的数据。在这其中用户社交、金融消息等信息是信息增长的重点。在数据爆炸式增长的同时，其内容分散，数据稀疏，无结构化信息等情况也逐渐凸显。因此，巨量的数据在为企业的投资者和决策者提供更多参考信息的同时，也使得如何高效地寻找有价值信息成为一个难题，比如当中国人民银行发布了一项新的政策举措时，这对金融领域相关公司的股票价格会有怎样的影响；或者一家投资公司在拟定投资方案时，如何了解拟投资行业的近期行情，有价值的独角兽公司又有哪些等。为了发现这些信息，知识发现成为一个重要的研究方向，信息抽取则是知识发现的核心。  命名实体识别（Named Entity Recognition，NER），作为信息抽取的基础，主要作用是将文本信息中特定的实体提取出来，自从在MUC-6会议（MUC-6, the Sixth Message Understanding Conferences）上首次被提出以来，一直在自然语言处理研究领域中有着不可或缺的地位。NER 技术是信息抽取、信息检索、机器翻译、问答系统等多种自然语言处理技术必不可少的组成部分。  命名实体（Named Entity，简称 NE）是NER的研究主体，通常可以将实体划分为3大类（实体类、时间类和数字类）和7小类（人名、地名、机构名、时间、日期、货币和百分比）命名实体。而在传统的金融领域，虽然也包含上述实体，但更侧重于人名，组织机构名称和项目条约等实体的识别，其中金融领域的组织机构命名规则变化多样，没有统一的约束格式，其识别的复杂度也相应得比较高。在传统金融实体的基础上，网络上被广泛使用的互联网金融实体数量也急剧增长，大量金融项目，概念和产品如井喷般涌现，将互联网实体内容分散，数据稀疏，非结构化等特点一同带入到金融领域之中，如何高效地找到我们所需的金融实体信息面临着越来越多的难题。  **2）研究意义：**  基于上述考虑，对现有模型进行改进从而提高面向金融领域文本识别效果至关重要。传统的命名实体识别方法是基于规则和现存知识库，而面对更新速度极快的互联网金融实体，不断更新现存库是件极其费时费力的事情。研究新的命名实体识别模型结构对于金融实体识别有着重要意义。本文提出一种改进的命名实体识别模型结构，可以更好的适应互联网金融实体特点，将有效提高互联网金融实体识别效果，从而更好的为有需求的相关机构和个体提供信息支撑。  除了信息提取的这一模型本职功能以外，寻找更加高效的命名实体方法对于金融领域的监管同样有着重要意义，互联网金融实体激增，与之伴随的是很多不规则的、无规律的实体的命名方式。其中混杂着很多金融诈骗团队创造的虚假金融概念，使得互联网金融实体的可信度和权威性受到广泛质疑。网络上频繁出现的非法贷款、违规项目为监管带来了很大难度，其中相当程度上的监管难处都在于识别实体技术手段的缺乏。因此金融实体识别方法的性能提升对于净化互联网金融环境，促进金融监管效率提高，进而预防金融诈骗等非法活动都是不可或缺的。  除此之外，金融实体识别在问答系统领域、机器翻译领域、数据标注领域等方面应用前景广泛，在本领域应用成功的方法同样可以为其他领域提供参考。因此寻找更加高效的金融实体识别算法对NER乃至自然语言处理相关领域都有着重要意义。  **（二）国内外文献综述**  关于命名实体识别的研究，从技术方案上可以分为两大类，一类为基于条件概率计算的传统命名实体识别，另一类则是基于深度学习的命名实体识别，目前较为主流的是基于深度学习的全监督命名实体识别研究，而关于命名实体所针对的领域，虽然大部分的研究仍停留在语料充分的领域，但也有不少学者扩展了新领域研究。  **1）传统命名实体识别研究**  关于NER的研究最早出现在 1991 年举办的第七届 IEEE人工智能应用会议，在会议上 Rau 发表了一篇关于抽取与识别公司名称的论文，其采用的是基于规则的方法。基于规则方法的NER一般需要收集相应的字典，以及一些常用的人名、地名、机构名等，然后人为制定相应的规则，利用文本的统计信息和语言规则在文本中进行识别，该方法十分依赖于领域知识与制定的规则，具有很强的局限性。  为了克服规则方法对规则的依赖，同时避免复杂的规则制定，在之后的研究中，基于机器学习的NER方法开始逐步出现。常用的包括隐马尔可夫模型（HMM），最大熵模型（MEM），支持向量机（SVM），条件随机场（CRF）。 在基于统计学的机器学习研究中，Zhou等人将NER看作是对文档中的词汇分类问题。基于这个想法，他们尝试使用隐马尔可夫来解决这个问题，并且取得了较好的效果。Collins 等人提出使用最大熵模型来解决 NER 任务，并且针对最大熵参数估计困难的问题，给出了自己的训练方法。Mccallum等人使用条件随机场模型来解决NER任务，并且在CoNLL-2003NER 的任务中取得了不错的效果。Finkel等人也使用条件随机场模型来解决NER任务，不同是，他们通过吉布斯采样（Gibbs sampling）进行训练模型，目的是在训练过程中可以添加一些非局部的特征到模型中，通过这种方法使得模型性能有了很大提升。Krishnan等人认为大部分NER的研究都是将这一问题看作序列标注问题，使用到的解决方法一般都需要基于马尔科夫假设，因而只能使用到有限的局部特征，为此，他们提出了一个简单的两层条件随机场方法，其中第二层条件随机场的特征来自第一层产生的输出。  总结上述的研究工作可以发现，虽然基于统计的NER方法相比于基于规则的有了很大的提升，但是其所带来的特征选取是一个新产生的问题，不同的领域实体的特征大相径庭。所以，当训练好的模型迁移到其他领域时，模型的性能往往严重下降。  **2）基于深度学习的命名实体识别研究**  深度学习是机器学习的一种，通过搭建人工神经网络来建立模型，这种方式在获得高性能的同时又具有相当的灵活性，最近几年，随着硬件性能和人工智能的发展，基于神经网络的深度学习在机器学习领域掀起了一股热潮。在自然语言处理方面，也有许多相关的应用，在 NER 研究领域，使用较为广泛的有BiLSTM（Bidirectional LongShort-Term Memory）-CRF、卷积神经网络（CNN）等，且效果均不错。基于深度学习的方法借助于神经网络的学习能力，不再依赖人为特征选取，能够更进一步理解语义信息。Feng等针对传统的NER方法需要构建特征工程和获取相关领域的知识的问题，提出了一种基于BiLSTM的神经网络结构的NER方法。该方法利用基于上下文的词向量和基于字的字向量，前者表达命名实体的上下文信息，后者表达构成命名实体的前缀、后缀和领域信息；同时，利用标注序列中标签之间的相关性对BiLSTM的代价函数进行约束,并将领域知识嵌入模型的代价函数中，进一步增强模型的识别能力。实验表明，所提方法的识别效果优于传统方法。李丽双等人在生物医学领域进行了NER任务研究，提出了一种基于CNN-BLSTM-CRF 的神经网络模型。首先利用CNN训练出单词的具有形态特征的字符级向量，并从大规模背景语料训练得到具有语义特征信息的词向量，然后将二者进行组合作为输入，再构建适合生物医学NER的BLSTM-CRF深层神经网络模型。在Biocreative和Jnlpba生物医学语料上测试，实验结果的 F-Score 分别为89.09%和74.40%。2018年，Shen等人提出了利用深度主动学习进行 NER 任务。将主动学习与深度学习相结合，可以利用少量的标注数据获得较高的学习准确度。由于主动学习的计算成本很高，因此他们提出了一个基于CNN-CNN-LSTM结构的轻量级模型。提出深度主动学习方法以便于减少标注量，降低数据标注的成本。实验表明，该模型能够迅速地对样本进行预测和评估不确定度。Bharadwaj 等提出了一种注意力神经模型（Attentional Neural Model）。该模型在原始的BiLSTM-CRF模型上加入了音韵特征，并在字符向量上使用注意力机制来关注并学习更有效的字符。该模型可以快速地应用于有少量数据或没有数据的新语言领域，从而实现了跨语言的迁移学习。  总结深度学习方面的研究工作，相比于前一阶段使用的基于统计的方法，深度学习使用词向量来表示词语，字向量来表示字，这些向量在其领域空间上具有特殊的意义，如相近的词语其词向量间的余弦相似度较高。通过这种方法解决了基于统计方法所面临的特征构建问题。 本文也将延续前人的工作针对深度学习NER这一领域展开研究。  **3）不同领域的命名实体识别研究**  在过去几年中，NER 在许多领域都有研究。在生物医学领域已经做了大量的工作。NER的大量研究属于一般领域，诸如中医，科学出版物，维基百科文章，临床笔记，网络文本等。另外，NER 对英语、中文、德语、日语、西班牙语、葡萄牙语等语言领域都有相应研究。  相比于新闻、医学等领域，命名实体识别在金融领域的应用虽然备受关注，但目前公开的研究成果却不多，原因在于金融领域公开的文本数据相对较少，标注数据则更为稀有且标注数据过程耗时耗力。Wang等提出了一种识别金融新闻文本中命名实体的新方法，该方法首先结合领域词典与条件随机场来识别金融实体（如股票名称等），之后利用互信息、边界熵以及上下文特征来识别缩写的金融实体，并在中文金融数据集上达到了91%的识别精度。另外有文献采用将条件随机场与协同训练结合的方式，将训练过程部署在 Hadoop框架之上，在满足识别精度的同时缩短了训练时间。Ei等人提出了实体关系提取与实体识别相结合的方法，通过引入规则与语法特征提取的方式来获取实体间的关系，并以获取的关系辅助金融实体识别从而提升识别精度。  **4）中文命名实体识别研究**  在中文领域的命名实体识别中，对中文语料进行基于深度学习模型训练时，输入可分为词向量与字向量。如果将语句分词后转换成词向量输入，分词效果的好坏将会直接影响 NER 模型的精度，但如果将词换成基于字向量输入，虽然避免了分词误差，却丢失了语义中词语之间的内在信息。对于上述问题，Zhang等人提出了将基于字和词的LSTM 模型结合的方法，将字的输出再加一层基于词的预测作为最终输出的Lattice-LSTM。对于包含噪音的数据，Yang 等人受到对抗网络学习的启发，他们在模型中使用BiLSTM 模型来分别学习标准标注数据和带噪音的标注数据，并通过对抗学习的方式优化模块的学习质量，使之收敛于真实数据。与之类似地，Zhang等人采用BiLSTM-Attention作为生成式对抗网络的生成器模型,并以CNN作为判别器模型，从众包标注数据集中整合出与专家标注数据分布一致的正样本标注数据来解决领域内标注数据缺乏的问题。Yang等人提出了一种基于注意力机制的化学药物命名实体识别方法，该方法先从海量生物文本中学习词向量，然后利用BiLSTM模型学习字符向量，之后再将词向量与字向量作为另一个BiLSTM 模型的输入并结合注意力机制来获取词在全文范围内的上下文表示，从而提高了同一篇文章中实体识别的一致性。  **（三）研究主题实践现状分析**  金融领域的命名实体识别也同样包括3大类7小类这些通用实体，但更侧重于公司名称、项目条约等实体的识别。相比于新闻、医学等领域，命名实体识别在金融领域的应用虽然备受关注，但目前公开的研究成果却不多，原因在于金融领域公开的文本数据相对较少，标注数据则更为稀有且标注数据过程耗时耗力。Wang等提出了一种识别金融新闻文本中命名实体的新方法，该方法首先结合领域词典与条件随机场来识别金融实体（如股票名称等），之后利用互信息、边界熵以及上下文特征来识别缩写的金融实体，并在中文金融数据集上达到了91%的识别精度。另外有文献采用将条件随机场与协同训练结合的方式，将训练过程部署在 Hadoop框架之上，在满足识别精度的同时缩短了训练时间。Ei等人提出了实体关系提取与实体识别相结合的方法，通过引入规则与语法特征提取的方式来获取实体间的关系，并以获取的关系辅助金融实体识别从而提升识别精度。而随着现在互大量互联网金融实体的出现，这些互联网金融实体具有名称长度差别大，内容分散，表达方式多种多样等诸多特点。传统基于规则的方法越发力不从心。没有专门面向该领域命名实体识别方案，只能对现有的命名实体识别方案的学习与归纳，也有不少前辈尝试用深度学习的方法来解决此类问题，但大多使用的方法都比较简单，均是从特征选择网络构建或者词向量等单方面入手，研究新的，更加复杂的命名实体识别模型结构对于金融实体识别有着重要意义。本文提出一种改进的命名实体识别模型结构，可以更好的适应当前金融实体特点。 |

三、研究方案

|  |
| --- |
| **（一）研究目标、研究思路、研究内容和拟解决的实际问题**  本研究的目标是实现面向金融实体识别模型的改进，提高开放环境下系统的准确率、召回率和F1分数值。并将其应用于金融领域实体识别之中，实现在公开数据集上准确识别命名实体。  研究思路：  研究思路  文本表示方法  实验参数设定  本文的主要内容包括以下几部分：  首先是绪论。该部分主要介绍选题背景和研究意义，国内外研究现状，主要研究内容、研究思路和研究方法，以及创新点。  第一章 相关理论基础。本部分主要介绍与本研究相关的理论和方法，包括命名实体识别的常用方法，文本表示方法等的相关理论，为后面的研究以及模型建立提供相应的理论支撑。   1. 金融实体文本数据集的构建。可分为模型的训练集和测试集两部分。其中训练集部分的内容包括收集数据的方式，清洗数据并对数据集中句子进行去除多余标点符号、将英文转化为小写、分词、剔除停用词和词性标注等预处理，将预处理好的语料转化为可标注的数据集，制作标注工具并对数据进行标注，最后再生成词向量，完成数据的结构化，提高后续预测流程的效率。测试集预计使用公开数据集比如BosonNLP，并结合手工构建的高识别难度语料，因此主要描述其构建过程。 2. BERT-BiLSTM-IDCNN-CRF模型设计。首先对模型做总体概述，然后从基本原理，注意力机制，预训练过程等方面对BERT预训练模型做描述。接着进入 BiLSTM-IDCNN-CRF模型设计阶段，分别论述BiLSTM神经网络模型和IDCNN神经网络模型的基础理论，搭建基于BiLSTM融合IDCNN的特征提取层，CRF推理层的模型网络结构。最后则是将以上三部分连接搭建起来，训练模型并在模型迭代过程中调整超参数和网络结构，提取句子上下文特征向量。搭建多个不同隐层单元构造的网络进行迭代测试，选择在测试集上表现最好的作为最终模型构造。分别训练多个batch\_size对应的模型，调整学习率和加入防止过拟合的Dropout层。   第四章 金融领域实体识别实验与分析。在交代实验准备工作与设定之后，对实验过程进行描述。对输入句子预测实体标签序列，针对测试结果优化算法。使用上文整理好的测试集来对模型的识别准确率进行测试。并对识别结果进行展示，分析命名实体识别结果不正确的语料，优化算法中各功能模块，迭代模型。并将改进完成的模型识别结果同BiLSTM-CRF等经典模型做对比以证明改进算法的优越性。  最后部分是结论与建议。结合前面的研究内容和实验结果，总结出本研究模型方法的创新之处，相对于传统命名实体识别方法的优势，对研究过程中遇到的问题以及存在的不足做一一论述，提出改进建议。  **（二）拟采取的研究方法及可行性分析**  **1）研究方法**  本课题拟采用定性与定量分析相结合、主观理论分析与客观实验分析相结合的分析方法。具体包括以下几种研究方法：  （1）文献研究法  文献研究法是根据研究目的，透过调查文献来获得资料，从而全面地、正确地了解掌握所要研究问题的一种方法。文献研究法被广泛用于各种学科研究中。其作用有：①能了解有关问题的历史和现状，帮确定研究课题。②能构成关于研究对象的一般印象，有助于观察和访问。③能得到现实资料的比较资料，有助于了解事物的全貌。在本研究中，通过广泛查阅研究国内外相关文献和研究资料，了解各类型命名实体识别技术的发展现状，从已有的模型之中总结出优点与存在的缺陷，并在其基础上加以完善和改进，整理相关文献，利用理论基础确定研究内容及思路。  （2）实验法  实验法是通过主动变革、挖制研究对象来发现与确认事物间的因果联系的一种科研方法。其主要特点是：第一、主动变革性。观察与调查都是在不干预研究对象的前提下去认识研究对象，发现其中的问题。而实验却要求主动操纵实验条件，人为地改变对象的存在方式、变化过程，使它服从于科学认识的需要。第二，控制性。科学实验要求根据研究的需要。借助各种方法技术，减少或消除各种可能影响科学的无关因素的干扰，在简化、纯化的状态下认识研究对象。第三，因果性。实验是发现确认事物之间的因果联系的有效工具和必要途径。在本研究中具体表现为：通过一整套设定好的实验条件与参数，在同一实验环境下使用同一公开数据集来检测不同模型的预测精度，以形成对比证明新模型的优越性。  （3）BERT  BERT的全称为Bidirectional Encoder Representation from Transformers，是一个预训练的语言表征模型。于2018年由Devlin等提出，并在 11 项自然语言处理任务上取得优异成绩。该模型采用表义能力较强的双向 Transformer 网络结构来预训练语言模型。Transformer是一种基于 Attention 机制叠加而成的深度网络，不但可以捕捉长距离特征，且具有较好的并行计算能力。  （4）迭代膨胀卷积神经网络CNN  普通的卷积神经网络在序列标注中存在一个劣势，即在经过卷积之后，末层神经元可能只是得到了原始输入数据中一小部分的信息。而对命名实体来讲，整个句子的每个字都有可能都会对当前需要标注的字造成影响。  针对CNN这一缺点，[Fisher Yu等人在2015](https://arxiv.org/abs/1511.07122) 提出了膨胀卷积神经网络模型（Dilated Convolutional Neural Network，DCNN）。其思路是在正常CNN的卷积核中增加空洞（dilation width），当作用在输入矩阵的时候，会直接跳过位于dilation width中的输入数据；这样就在保持卷积核矩阵本身大小不变的情况下，达到了扩大模型的感受野的目的。而迭代膨胀卷积神经网络（IDCNN）一般是由多个具有相同结构的膨胀卷积块拼接在一起，而每个膨胀卷积块里面则使用多层的膨胀卷积层，每层卷积的输入都是上一层的输出，并且每层的rate可以不同。  （5）双向长短期记忆网络BiLSTM：  LSTM 虽然可以很好的缓解 RNN 带来的“梯度消失”问题，但是在做自然语言处理时表现的效果并不是十分理想，因为 LSTM 只能利用前文的信息，并没有考虑后文对前文的影响，而且对于序列标注任务没有充分利用上下文信息，那么进行预测的结果也会有影响。  针对上述问题，学者们又提出了双向长短期记忆模型（Bidirectional Long Short-Term Memory，BiLSTM），按照两个方向去计算输入序列，使其可以同时使用上下文信息，也就是整个输入序列的信息。BiLSTM 由一个正向的 LSTM 和一个反向的 LSTM 构成，同时读取输入序列进行计算可以充分利用上下文信息，最后将计算结果同时传给输出层进行输出。   1. **可行性分析：**   关于本研究的可行性，学生将从数据获取，理论支撑，软硬件支持和自身能力四个方面进行论证：  首先本研究有两类的主要数据来源，一部分选自诸如公开数据集，另一部分由各类金融相关网站爬虫得来，此类数据可以使用Python中的BeautifulSoup4等三方库编写爬虫脚本，再将程序导入到Web Scraper等浏览器插件中来实现大批量的语料获取。因此在数据获取方面基本不存在困难。  理论技术方面，根据学生对国内外文献的阅读整理，从2016年Zhang等人将LSTM应用到中文领域的NER开始，已经有很多前辈们尝试在用不同的方式提升金融领域命名实体识别模型的识别能力，既有基于BILSTM-CRF和IDCNN-CRF进行网络构架方面的改进，也有从词向量的角度着手对模型的输入内容加以完善，BERT这类较新且表现良好的预训练模型也被很多学者拿来做改进尝试。因此学生提出的模型创新有着众多前人的理论基础和实践检验，并且我院大数据方向老师在神经网络和都有扎实的理论知识和丰富的建模经验，学生相信在各位优秀老师的指导下可以顺利完成研究的理论支撑。  软件方面，本研究的编程语言选择了3.6.13版本的Python，具有良好的第三方库支持性，在此基础上采用2.4.1版本的TensorFlow深度学习框架，完全可以满足模型的搭建需求；硬件方面，本研究的操作系统为Linux 3.10.0，CPU采用Intel(R) Xeon(R) Silver 4116，内存为128G，GPU采用Quadro P5000 + Quadro P4000，显存为16GB/8GB，算力完全满足本研究的神经网络数据训练需求。  学生自身在过去一年的时间里已经学习过机器学习和深度学习的相关课程，且熟悉卷积神经网络CNN类的构造，在数据文本爬虫以及自然语言处理方面拥有实战经验并撰写过文章，熟练掌握Python相关的第三方库以及深度学习框架的使用方法，具备在前人基础上进行模型改进的能力。  **（三）本研究的特色与创新之处**  目前对命名实体识别算法的研究有很多，基于前人的研究成果，本文从深度学习方法上开展了研究，探索了金融实体识别的新模型，主要的创新之处有以下几个方面：   1. 针对基于传统词向量表示不能表示字的多义性问题，引入BERT这一具有表征一词多义能力的模型，通过联合调节所有层中的上下文，使用双向Transformer作为编码器动态生成字的语义表征，具备传统的词向量所不具备的表示语句特征的特点。 2. 针对BiLSTM忽略掉局部特征的缺点，引入 IDCNN 模型，建立BiLSTM与IDCNN的融和特征层，使其能够兼顾局部特征的优势。具体方法是将由BERT模型预训练获取的词向量输入到BiLSTM与IDCNN中进行特征提取，并将两者提取到的特征向量进行拼接，最后通过CRF进行修正后输出。 3. 为了捕捉对于分类起关键作用的特征，本文在所提方法的基础上引入了注意力机制，即在BiLSTM与IDCNN提取特征后，通过注意力机制对所提取到的特征进行权重分配以强化关键特征的作用。最后通过CRF对金融领域中的各类实体进行标注。   通过比较改进模型与已有模型的训练效果，预计新模型的识别效果相对于其它几种主流命名实体识别方法会有一定的提升。 |

四、论文大纲（需列举至三级标题）

|  |
| --- |
| 绪论  一、研究意义与背景  二、国内外文献综述  三、研究思路与方法  第一章 相关理论基础  第一节 命名实体识别常用方法介绍  一、基于词典和规则的方法  二、基于统计机器学习的方法  三、基于深度学习的方法  第二节 文本表示方法  一、词嵌入  二、Word2Vec模型  第二章 金融实体文本数据集的构建  第一节 模型训练集  一、数据来源  二、数据预处理  三、词向量方法  第二节 模型测试集  一、数据来源  二、数据描述  第三章 BERT-BiLSTM-IDCNN-CRF金融命名实体识别模型  第一节 模型输入层  一、BERT介绍  二、Transformer模型与注意力机制  三、预训练过程  第二节 模型特征层  一、BiLSTM神经网络模型  二、IDCNN神经网络模型  三、BilSTM-IDCNN融合特征层  第三节 模型输出层  一、注意力机制  二、CRF推理层  第四章 金融领域实体识别实验与分析  第一节 实验准备工作  一、实验评估标准  二、实验数据与环境  第二节 实验设定  一、对比模型设定  二、对比参数设定  第三节 实验过程  一、模型参数实验对比与分析  二、不同模型间的实验对比与分析  三、 实验小结  结论及展望  一、研究结论  二、对策建议  三、研究展望  参考文献  附录  致谢 |

五、主要参考文献书目

|  |
| --- |
| [1] Rau L F. Extracting Company Names from Text[C]. Proceedings of the 7th IEEE Conference on Artificial Intelligence Applications,1991: 29-32.  [2] Zhang X, Li Z. Named entity recognition in Chinese electronic medical records based on multi-feature integration [J]. Software Guide, 2017,16 (2): 128-131.  [3] Mccallum A, Freitag D, Pereira F C. Maximum Entropy Markov Models for Information Extraction and Segmentation[C]. ICML, 2000: 591-598.  [4] Joachims T. Making large-scale support vector machine learning practical[C]. Advances in Kernel Methods, 1999.  [5] Lafferty J D, Mccallum A, Pereira F C N. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[J]. Proceedings of Icml, 2001, 3(2): 282-289.  [6] Su J, Su J. Named entity recognition using an HMM-based chunk tagger[C]. 40th ACL, 2002: 473-480.  [7] Martí, N, Prodinger H. Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms[C]. Acl-02 Conference on Empirical Methods in Natural Language Processing, 2002: 1-8.  [8] Cucchiarelli A, Velardi P. Unsupervised named entity recognition using syntactic and semantic contextual evidence[M]. MIT Press, 2001.  [9] Tsuboi Y, Kashima H, Mori S, et al. Training Conditional Random Fields Using Incomplete Annotations[C]. International Conference on Computational Linguistics, 2008: 897-904.  [10]Xiong R, Nicholsep, Shen Y. Deep Learning Stock Volatility with Google Domestic Trends[J].Papers, 2016.  [11]Greff K, Srivastava R K, J Koutník, et al. LSTM: A Search Space Odyssey[J]. IEEE Transactions on Neural Networks & Learning Systems,2017, 28(10):2222-2232.  [12]俞鸿魁,张华平,刘群等.基于层叠隐马尔可夫模型的中文命名实体识别[J]. 通信学报, 2006,27(2):87-94.  [13]李丽双,何红磊,刘珊珊等.基于词表示方法的生物医学命名实体识别[J]. 小型微型计算机系统, 2016,37(2):302-307.  [14]尹存燕,黄书剑,戴新宇,等.中英命名实体识别及对齐中的中文分词优化[J].电子学报, 2015,43(8):1481-1487.  [15]易黎,黄鹏,彭艳兵,等.基于联合模型的中文社交媒体命名实体识别[J].计算机与数字工程, 2017,45(12):2402-2406.  [16]李波.基于自主推理的中文命名实体识别方法研究[D].西安:西北大学,2012,30-35.  [17]鲁亚楠,孙锐,姬东鸿.基于位置敏感 Embedding 的中文命名实体识别[J].计算机应用研究, 2017,34(2):365-368.  [18]金留可.基于递归神经网络的生物医学命名实体识别[D].大连:大连理工大学,2016,21-30.  [19]张雪英,叶鹏,王曙,等.基于深度信念网络的地质实体识别方法[J].岩石学报,2018,34(2):343-351.  [20]张晓艳,王挺,陈火旺.基于混合统计模型的汉语命名实体识别方法[J].计算机工程与科学, 2006,28(6):135-139.  [21]王红斌,沈强,线岩团.融合迁移学习的中文命名实体识别.小型微型计算机系统, 2017,38(2):346-351.  [22]俞鸿魁,张华平,刘群,等.基于层叠隐马尔可夫模型的中文命名实体识别[J].通信学报, 2006(2): 87-94.  [23]尹存燕,黄书剑,戴新宇,等.中英命名实体识别及对齐中的中文分词优化[J].电子学报, 2015,43(8):1481-1487.  [24]易黎,黄鹏,彭艳兵,等.基于联合模型的中文社交媒体命名实体识别[J].计算机与数字工程, 2017, 45(12):2402-2406.  [25]李波.基于自主推理的中文命名实体识别方法研究[D].西安:西北大学, 2012,30-35.  [26]吕雅娟,赵铁军,杨沐昀,等.基于分解与动态规划策略的汉语未登录词识别[J].中文信息学报, 2001(1): 28-33.  [27]刘浏,王东波.命名实体识别研究综述[J]. 情报学报, 2018, 37(3): 329-340.  [28]邱泉清,苗夺谦,张志飞.中文微博命名实体识别[J]. 计算机科学, 2013, 40(6): 196-198.  [29]曹文斌,武卓峰,杨涛,等.基于文本语料的涉恐事件实体属性抽取[J].工程科学学报, 2020, 42(4): 500-508.  [30]杨锦锋,关毅中文电子病历命名实体和实体关系语料库构建[J].软件学报, 2016, 27(11): 2725-2746. |

六、论文写作计划、进行方式和采取主要措施

|  |
| --- |
| 第一阶段 开题准备阶段：  1.2021.03-2021.05：阅读大量文献，确定选题方向，并与指导老师交流讨论，确定选题，并认真撰写开题报告。  2.2021.05-2021.06.11：写出开题报告，制作开题PPT，与指导老师充分沟通，做好开题报告答辩准备  3.2021.06.12：开题报告论证答辩  第二阶段 论文研究阶段  1.2021.06.12-2021.07：收集相关数据来形成训练数据集。阅读金融领域命名实体识别的相关文献，对已有的经典方法进行深入了解，并在总结常用的实体命名方法的创新的切入点与仍存在不足，在此基础上提出改进和创新的想法，为后文的模型的细节改进做铺垫。  2.2021.07-2021.09：对收集到的数据进行清洗和处理，生成词向量，构建初步的命名实体模型，并使用测试数据对其进行训练。  3.2021.09-2021.10：在前文得到的初步模型基础上，对参数做估计、检验及分析以及进一步的调试。与此同时完成测试数据集的选取，将优化过的模型同其他已有方案进行评价指标对比。根据评价指标对比结果，撰写论文报告得出结论以及相关的建议。  第三阶段 论文撰写与修改阶段  1.2020.11 结合前面所做的研究成果，整理完善研究内容，开始撰写完整的学位论文，同时与指导老师交流沟通，不断的修改完善论文。  2.2020.12 提交论文初稿，准备论文预答辩。 |

七、指导教师意见

|  |
| --- |
| 焦樵同学的选题为“基于深度学习的金融领域命名实体识别方法研究”。本选题新颖，具有一定的实用价值。该课题研究内容详细、充实，主要观点突出，并且时效性强。该课题硏究方法和硏究计划合理，研究内容难度适中，学生能够在预定时间内完成该课题的书写。同意该课题开题。    导师签名：  年 月 日 |

八、开题报告会组成人员及意见记录

|  |  |  |
| --- | --- | --- |
| 姓 名 | 职 称 | 工作单位 |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
| 开题报告记录人签名：  年 月 日 | | |

九、开题报告评价

|  |
| --- |
| 1.综合评价意见。  2.是否通过  （1）通过：按照答辩专家意见修改完善。 [ ]  （2）不予通过：按照答辩专家意见修改后，重新开题。 [ ]  3.答辩成绩  （1）优秀[ ] （2）良好[ ] （3）合格[ ] （4）不合格[ ]  （请在括号内打“√”选择）  4.成员签名： **、 、 、**  、  年 月 日 |