**南京大学软件学院研究生学位论文中期检查报告格式**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **南京大学软件工程硕士学位论文中期检查报告** | | | | | |
| 导师1姓名 | 伏晓 | 研究生姓名  （学号） | 安磊  MF1732001 | 方向 | 软件工程 |
| 导师2姓名 |  |
| 论文题目 | 基于深度学习的信息抽取服务的设计与实现 | | | | |
| 论文选题来源及研究的目的和意义（500字左右）：  随着人工智能技术的快速发展，越来越多的算法得到应用。自然语言处理作为人工智能领域中的一个重要分支一直是许多人关注的焦点，而知识图谱是自然语言处理中一个重要的研究方向。构建知识图谱的数据来源包括结构化文本和非结构化文本，作为构建知识图谱的重要数据来源之一的非结构化文本，如何从中提取高质量的信息成为一个研究热点。  面对上文所述的问题，一个能够提供高质量数据的信息抽取服务就显得十分重要。本文以从非结构化文本中进行信息抽取为背景，以构建知识图谱时所需的三元组作为目标，实现从非结构化数据到知识图谱三元组的数据处理与转换。  命名实体识别（Named Entity Recognition，简称NER）作为信息抽取中的一个基本任务，最早是在MUC-6（the Sixth of the Message Understanding Conference）上将其作为信息抽取的一项子任务引入测评中。关系抽取（Relation Extraction，简称RE）作为信息抽取中的另一个基本任务，早在2000年就有基于句法解析增强的方法来实现关系抽取任务。而本文同时结合命名实体识别和关系抽取两项任务作为利用非结构化文本中的三元组信息来构建知识图谱的关键环节，如何确保实体识别和关系抽取的准确性，提供高效稳定的信息抽取服务是构建高质量、高精度的知识图谱的关键所在。  本文基于构建非结构化文本的信息抽取服务的具体项目。从模型设计和具体的使用场景出发，本文选择了两种不同的命名实体识别模型和一种关系抽取模型进行组合，并将模型封装在Docker容器中，部署到Kubernetes集群上构建一个信息抽取服务系统。 | | | | | |
| 该方向的研究现状或技术进展综述（2000字左右）  自从在MUC-6上将命名实体识别作为信息抽取的一个子任务之后，随着数据时代的到来，命名实体在各个领域的信息抽取任务中都得到了充分的利用和发展。在国外命名实体识别的研究中，考虑到英文本身的特性，英文的命名实体识别只需要考虑词本身的特征而不需要关注别的问题例如分词等问题的影响，所以实现的难度相对较小，准确率也较高，  中文的命名实体识别由于中文文字内在的特殊性导致了在做命名实体识别之前要先做分词、句法分析等步骤，导致中文命名实体相对英文而言难度大大提高。在MUC-6中国内专家学者们把“人名”、“地名”、“组织机构名”作为实体的名称表达式，将这三类实体作为主要研究目标。在随后的研究中，逐渐地有人将“时间”、“货币”、“百分数”等也作为各种实体的表达式加入到命名实体识别的任务中来。  最早在命名实体识别任务中，主要有基于专家知识构建规则的方法以及基于数理统计的方法两类。基于词典和规则的方法主要是靠人工建立规则体系，在构建了大量基于专家知识的词法和语义规则之后，系统会根据规则对输入文本进行解析，基于人工构建的有限规则集合对文本中可能的命名实体进行推理和识别。虽然在特定的数据集上基于专家知识构建规则的方法相对于基于统计方法的命名实体识别而言准确度更高，但这种方法在数据量逐渐增大、数据内容逐渐复杂之后会变得不再可行，因为基于某一小部分语料构建的规则体系在别的语料上讲不再适用，人们无法去扩充和维护一个十分庞大的规则体系，且随着近些年机器学习理论的不断完善和计算机计算性能的提高，大部分学者开始转而研究基于统计方法的命名实体识别。  基于统计方法的命名实体识别主要有有监督（Supervised）、半监督（Semi-Supervised）的机器学习方法，其中有监督的机器学习方法在拥有大批量的标注语料的前提下在不同领域的文本中都拥有更高的识别准确率和精度，从而被更多的学者、从业者们所广泛使用。有监督学习方法中在命名实体识别任务中表现比较出色的方法主要有：基于隐马尔科夫模型（Hidden Markov Models，简称HMM）的命名实体识别模型、最大熵模型（Maximum Entropy Models，简称MEM）、条件随机场（Conditional Random Fields，简称CRFs）。基于统计方法的命名实体识别对特征选取的要求比较高，需要从文本中挖掘出对于实体识别有用的单词信息、上下文信息、句法信息、语义信息等作为特征。随着研究的不断深入，大量实验结果表明条件随机场结合了HMM和MEM的有点，成为中文命名实体识别任务中表现更优秀的统计学习方法。有学者针对条件随机场的特征选择与交叉组合进行了研究，通过实验得出不同的特征以及特征组合在训练时的贡献大小。  中文命名实体识别（Chinese Named Entity Recognition，简称CNER）相对于英文命名实体识别任务仍存在许多不同，由于中文本身的特点造成许多特殊的难点：   1. 中文单词的边界相比于英文较模糊   英文中的单词之间都有分隔符来标识边界，命名实体如人名、地名等单词的首字母为大写，这些信息能很好的为命名实体的识别做边界和位置标识，而中文单词没有这些信息。   1. 中文命名实体的词语结构更加复杂   有的类型的命名实体单词长度没有限制；不同实体有不同的组成结构（嵌套、简称、别名等）；音译词没有统一的构词规范；组织机构名和人名、地名会有很多交叉重复的地方（如：星环信息科技（上海）有限公司）。   1. 中文命名实体的一词多义与一义多词   随着互联网的极速发展，越来越多的网络词汇不断出现，很多原本的命名实体在现在都有了新的称呼（如称淘宝为某宝），这些词汇的新称呼的出现大大增加了命名实体识别的难度。  随着深度学习的影响不断增加，用深度学习方法解决命名实体识别任务也获得的一些显著的成果。[Wu et al. 2015]提出使用深度神经网络（Deep Neural Network，简称DNN）从语料中训练词向量，再输入到另一个深度神经网络中进行命名实体识别，实验结果好于传统机器学习中效果最好的CRF模型。[Z Huang et al. ]提出了使用双向长短时记忆模型(Bidirectional Long-Short Term Memory Model，简称Bi-LSTM)和条件随机场进行命名实体识别。还有一系列深度学习方法对命名实体识别任务做了很多尝试，在不同的数据集上相较于传统方法都有不小的提升。  关系抽取的方法主要有：   1. 有监督的学习方法   把关系抽取任务当做一个分类问题，从训练数据中设计并提取有效的特征并训练学习各种分类模型进行关系预测。需要大量的人工标注语料，比较耗时耗力，但预测效果较好。   1. 半监督的学习方法   先人工设定若干种子实例，再迭代地从文本中抽取相应的关系模板和更多关系实例。   1. 无监督的学习方法   假设有相同语义的命名实体同样拥有相似的上下文信息，可根据每个命名实体对的上下文信息来代表它们的语义关系，并对所有实体对的语义进行聚类。  与另外两种方法相比，有监督的学习方法在抽取的准确率和召回率上都有着更好的表现。  自从2000年Miller等提出基于句法解析增强的方法来实现关系抽取后，越来越多的实体间关系抽取方法被提出：基于逻辑回归的方法、基于核函数的方法、基于条件随机场的方法。在有监督学习中针对需要大量人工标注的情况下，Mintz等人提出了使用远程监督（Distant Supervision）的方法来扩充标注语料，可以有效解决关系抽取的标注数据规模问题。  随着计算机硬件和计算能力的极速发展，深度学习开始逐渐被学术界和业界所关注，很多学者也开始将深度学习运用到关系抽取任务中。提出使用递归神经网络（Recurrent Neural Network，简称RNN）来解决关系抽取问题，通过递归神经网络学习到句子的词汇特征、句法特征、语义特征再用于关系分类和抽取。Zeng提出使用卷积神经网络（Convolutional Neural Network，简称CNN）来解决关系抽取问题，采用词向量和词的相对位置作为卷积神经网络的输入，通过卷积、池化、非线性计算等操作得到句子表示并用于关系抽取。Santos提出了一种新的卷积神经网络结构用于解决关系抽取问题，在这个新的结构中采用了新的Ranking损失函数并得到了更好的效果。Miwa提出一种基于端到端（End to End）神经网络的关系抽取模型，使用双向长短时记忆模型和树形长短时记忆模型同时对实体和句子进行建模。Lin提出基于句子级注意力机制的神经网络模型来解决关系抽取问题，文中的方法可以根据不同关系为每个实体对分配不同的权重。 | | | | | |
| 论文的主要技术路线、研究思路和实现方法；相关项目应用前景：（重点说明变更部分）：  论文主要采用深度学习框架Tensorflow开发和训练模型，结合Flask框架、Docker容器技术、Kubernetes集群来实现信息抽取服务系统。  可以满足客户对各种文本进行信息抽取的需求。 | | | | | |
| 本人在相关项目中的扮演的角色和承担的工作（重点说明变更部分）：  本人负责完成了项目中的模型所需的数据集采集和标注，数据清洗，模型开发和训练，信息抽取服务部署的全部工作。 | | | | | |
| 论文的主要工作（500字左右）：  本文在总结信息抽取现有的一些研究成果的基础上，结合命名实体识别和关系抽取两项信息抽取的基本任务，从多个深度学习模型中选择合适的模型，结合自己人工标注的部分金融领域新闻文本语料结合远程监督进行模型训练，将训练好的模型固化成pb模型，把模型进行组合封装到Docker中，并部署在集群上来提供信息抽取服务。  本文的主要工作有：   1. 对命名实体识别和关系抽取的研究现状和相关技术进行综述，分析了命名实体识别和关系抽取在构建知识图谱前的信息抽取环节的重要性。 2. 对本文的信息抽取服务系统进行了详细设计，包括需求分析、产品功能分析、本文选择的信息抽取模型的详细设计以及系统中每个模块的详细设计。 3. 利用远程监督扩充人工标注数据，缓解人工标注数据费时费力的缺点。 4. 将训练好的模型封装成ModelServer部署到集群上，构建一套信息抽取服务系统。 | | | | | |
| 实现论文三级大纲：   1. 绪论   1.1 项目背景  1.2国内外研究现状  1.2.1 命名实体识别研究现状  1.2.2 关系抽取研究现状  1.3 本文的主要工作  1.4 本文的组织结构   1. 技术综述   2.1 引言  2.2 预训练模型  2.2.1 词向量  2.2.2 BERT模型  2.3 双向长短时记忆模型  2.3.1 循环神经网络  2.3.2 长短时记忆模型  2.3.3 双向长短时记忆模型  2.4 深度残差网络  2.5 远程监督  2.6 其他相关技术  2.6.1 深度学习框架Tensorflow  2.6.2 Docker容器技术  2.6.3 Kubernetes集群  2.7 本章小结   1. 信息抽取服务分析与设计   3.1 信息抽取服务系统总体规划  3.2 信息抽取服务需求分析  3.2.1 功能性需求  3.2.2 用例描述  3.3 信息抽取服务设计与模块设计  3.3.1 数据预处理模块设计  3.3.2 命名实体识别模块设计  3.3.3 关系抽取模块设计  3.3.4 信息抽取服务模块设计  3.4 基于传统词向量的命名实体识别模型设计  3.4.1 输入层  3.4.2 网络层  3.4.3 输出层  3.5 基于BERT的命名实体识别模型设计  3.5.1 输入层  3.5.2 网络层  3.5.3 输出层  3.6 关系抽取模型设计  3.6.1 输入层  3.6.2 卷积层  3.6.3 残差卷积块  3.6.4 池化层  3.6.5 输出层   1. 信息抽取服务的实现   4.1 数据预处理模块的实现  4.1.1 数据采集  4.1.2 数据预处理  4.2 基于传统词向量的命名实体识别模型的实现  4.3 基于BERT的命名实体识别模型的实现  4.4 关系抽取模型的实现  4.5 实验评估  4.5.1 评估标准  4.5.2 对比实验与结果分析  4.6 信息抽取服务的实现与部署  4.6.1 构建Docker镜像  4.6.2 部署到Kubernetes集群  4.7 本章小结   1. 总结与展望   5.1 总结  5.2 进一步工作展望 | | | | | |
| 论文和相关项目的当前进度：  项目已完成，论文初稿已完成，正在进行格式修改和内容的修饰。 | | | | | |
| 论文和相关项目进展过程中遇到的困难和问题，以及解决的措施：  遇到的困难：BERT模型训练对硬件要求较高；论文编写时对论文的格式规范了解不够。  解决措施：借助公司的集群进行模型训练；咨询老师和往届同学的论文参考来熟悉论文书写规范。 | | | | | |
| 主要参考文献：  **[MUC-6, 1996] MUC-6, the Sixth in a Series of Message Understanding Conferences, was held in November 1996[OL].**<http://cs.nyu.edu/cs/faculty/grishman/muc6.html> **.**  **[Milleret al., 2000] Miller, Scott, Heidi Fox, Lance Ramshaw, and Ralph Weischedel. “A novel use of statistical parsing to extract information from text.” *In Proceedings of NAACL*, 2000.**  **[Bikel et al., 1999] Bikel D M,Schwarta R,Weischedel R M.An Algorithm that Learns What`s in a Name[J].*Machine Learning Journal Special Issue on Natural Language Learning*, 1999, 34(1-3): 211-231.**  **[Tsai et al., 2004] Tsai T,WU S,Lee C, et al. Mencius: A Chinese Named Entity Recognizer Using the Maximum Entropy based Hybrid Model[J]. *International Journal of Computational Linguistics & Chinese Language Processing*, 2004, 9(1):65-81.**  **[McCallum et al., 2003] McCallum A,Li W.Early Results for Named Entity Recognition with Conditional Random Fields, Features Induction and Web-enhanced Lexicons[C]. *In Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL*,2003: 188-191.**  **[张祝玉等, 2008] 张祝玉，任飞亮，朱靖波. 基于条件随机场的中文命名实体识别特征比较研究[C]. 见: *第4届全国信息检索与内容安全学术会议论文集*.2008.**  [Wu et al., 2015] Yonghui Wu, Min Jiang, Jianbo Lei,Hua Xu. Named Entity Recognition in Chinese Clinical Text Using Deep Neural Network. Stud Health Technol Inform. 2015;216:624-8.  [Z Huang et al., 2015] Zhiheng Huang, Wei Xu, Kai Yu. Bidirectional LSTM-CRF Models for Sequence Tagging. *arXiv*, 2015, 1508.01991 [cs.CL]  **[Kambhatla, 2004] Kambhatla, Nanda.”Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations.” *In Proceedings of ACL*, 2004.**  **[Zhao and Grishman, 2005] Zhao, Shubin, and RalphGrishman. Extracting relations with integrated information using kernel methods. *In Proceedings of ACL*, 2005.**  [Culotta et al., 2006] Culotta, Aron, Andrew McCallum, and Jonathan Betz. Integrating probabilistic extraction models and datamining to discover relations and patterns in text. *In Proceedings of HLT-NAACL*, 2006.  [Mintz et al., 2009] Mintz, Mike, Steven Bill, RionSnow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. *In Proceedings of ACL-IJCNLP*, 2009.  [Socher et al. 2012] Socher , Richard, et al. Semantic compositionality through recursive matrix-vectorspaces. *In Proceedings of EMNLP-CoNLL*, 2012.  [Zeng et al., 2014] Daojian Zeng, Kang Liu, et al. Relation classification via Convolutional Deep Neural Network. *In Proceedings of COLING*, 2014  [Santos et al., 2015] Cicero Nogueira dos Santos, Bing Xiang, Bowen Zhou. Classifying Relations by Ranking with Convolutional Neural Networks. *In Proceedings of ACL*, 2015.  [Miwa er al., 2016] Makoto Miwa, Mohit Bansal. End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures. *In Proceedings of ACL*, 2016.  [Lin et al., 2016] Yankai Lin, Shiqi Shen, Zhiyuan Liu, et al. Neural Relation Extraction with Selective Attention over Instances. *In Proceedings of ACL*, 2016.  [Bengio et al., 2002] Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult[J]. *IEEE Trans Neural Netw*, 2002, 5(2):157-166.  [Hinton, 1986] Hinton G E. Learning distributed representations of concepts[C]. *Eighth Conference of the Cognitive Science Society*. 1986.  [Harris, 1954] Harris Z S.Distributional structure[J]. *Word*, 1954, 10(2-3): 146-162.  [Pennington et al., 2014] Jeffrey Pennington, Richard Socher, and Christopher D. GloVe: Global Vectors for Word Representation. *Manning*. 2014.  [Bengio et al., 2006] Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model[J]. *Journal of Machine Learning Research*, 2006, 3(6):1137-1155.  [Mikolov et al., 2013a] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space[J]. *Computer Science*, 2013.  [Mikolov et al., 2013b] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]. *International Conference on Neural Information Processing Systems*. Curran Associates Inc. 2013:3111-3119.  [Radford et al., 2017] Radford A, Jozefowicz R, Sutskever I. Learning to Generate Reviews and Discovering Sentiment[J]. 2017.  [Rumelhart et al., 1986] David E. Rumelhart, Geoffrey E. Hinton & Ronald J. Williams. Learning representations by back-propagating errors[J]. *Nature*, 1986, 323(3):533-536.  [Hochreiter et al., 1997] Hochreiter S, Schmidhuber J. Long short-term memory[J]. *Neural computation*, 1997, 9(8):1735-1780.  [He K M et al., 2015] Kaiming He, Xiangyu Zhang, Shaoqing Ren, et al. Deep Residual Learning for Image Recognition[J]. *arXiv*, 2015, 1512.03385.  **[Hinton et al., 2012]** Hinton G E, Srivastava N, Krizhevsky A, et al. Improving neural networks by preventing co-adaptation of feature detectors[J]. *arXiv preprint arXiv*:1207.0580, 2012. | | | | | |
| 导师意见： | | | | | |
| 学院备案意见：        年 月 日 | | | | | |