****



**研 究 生 毕 业 论 文**

**（申请工程硕士学位）**

|  |  |
| --- | --- |
| **论文题目** | 基于深度学习的信息抽取技术的设  计与实现 |
| **作者姓名** | 安 磊 |
| **学科、专业名称** | 工程硕士(软件工程方向) |
| **研究方向** | 软件工程 |
| **指导教师** | 伏晓　副教授 |

**2019年 04 月 11 日**

**学 号： MF1732001**

**论文答辩日期： 年 月 日**

**指 导 教 师：**

**基于深度学习的信息抽取技术的**

**设计与实现**

|  |  |
| --- | --- |
| **作 者:** | **安磊** |
| 指导教师: | **伏晓　副教授** |

|  |
| --- |
| **南京大学研究生毕业论文** |
| **(申请工程硕士学位)** |

|  |
| --- |
| **南京大学软件学院** |
| **2019年04月** |

**The Design and Implementation of Information Extraction Technology Based on Deep Learning**

**An, Lei**

**Submitted in partial fulfillment of the requirements for the degree of Master of Engineering**

Supervised by

Associate Professor **Fu, Xiao**

Software Institute

**NANJING UNIVERSITY**

Nanjing, China

Apr, 2019

# 摘 要

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系统就显得十分重要。本文基于从非结构化文本中进行信息抽取的背景，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。

本文基于针对非结构化文本的信息抽取系统的具体项目，针对金融上市公司领域进行命名实体识别、实体间关系抽取。由于基于深度学习的命名实体识别和关系抽取在不需要很多额外特征的基础上相对于传统机器学习就能得到更好的效果，所以本文从模型设计和实现的角度提出了基于深度学习框架Tensorflow实现模型、搭建信息抽取系统这一解决方案。本文的主要工作如下：

1. 针对命名实体识别的问题，本文基于深度学习框架Tensorflow采用双向长短时记忆神经网络（Bi-LSTM）+条件随机场（CRF）实现命名实体识别模型，对金融领域中的六个类别的实体进行命名实体识别。
2. 针对关系抽取问题，本文基于深度学习框架Tensorflow采用残差网络（ResNet）+卷积神经网络（CNN）实现关系抽取模型，对金融领域中的各种类别实体间的各种关系进行识别与抽取。
3. 针对关系抽取中文标注语料数量较少、标注困难的问题，文本采用远程监督的方法来大量扩充标注数据，从一定程度上减少模型对人工标注数据的依赖。

**关键词**：信息抽取，命名实体识别，知识图谱，关系抽取，深度学习

# Abstract

With the rapid development of artificial intelligence, more and more algorithm were applied.Natural Language Processing has always been the focus of many people as an important branch in the field of artificial intelligence,and Knowledge Graph is an important research direction in Natural Language Processing.The data sources for constructing Knowledge Graph include structured text and unstructured text.As an unstructured text that is one of the important data sources for constructing Knowledge Graphs,how to extract high quality information from it becomes a research hot spot.

In the face of the problems described above,an information extraction system that provides high quality data is very important.Based on the background of information extraction from unstructured texts,this thesis aims to construct data graphs form the unstructured data to the Knowledge Graph triples.

Based on the specific project of information extraction system for unstructured texts,this thesis proposes named entity recognition and inter-entity Relation extraction for financial listed companies.Due to Named Entity Recognition and Relation Extraction based on Deep Learning can achieve better results than traditional Machine Learning without much additional features,from the perspective of model design and implementation,this thesis proposes a model based on a Deep Learning framework Tensorflow and builds an information extraction system as the solution.The main work of this thesis is as follows:

(1) For the problem of Named Entity Recognition,this thesis based on the deep learning framework Tensorflow uses the Bidirectional long-short-term memory neural network(Bi-LSTM) + conditional random field(CRF) to implement the Named Entity Recognition model,and recognizes the six categories of entities in the financial field.

(2) For the problem of Relation Extraction,this thesis based on the deep learning framework Tensorflow uses the Residual Network (ResNet) + Convolutional Neural Network (CNN) to implement the Relation Extraction model to identify and extract various Relations among various categories of entities in the financial field.

(3) For the problem of Chinese Relation Extraction with small corpus and difficult labeling,this thesis uses Distance Supervision to expand the annotation data in a large amount,which reduces the dependence of the model on manual annotation data to a certain extent.

**Keywords**：Information Extraction,Named Entity Recognition,Knowledge Graph,

Relation Extraction,Deep Learning

**目 录**

[摘 要 I](#_Toc514163438)

[Abstract III](#_Toc514163439)

[图目录 VIII](#_Toc514163440)

[表目录 IX](#_Toc514163441)

[第一章 绪论 1](#_Toc514163442)

[1.1 项目背景 1](#_Toc514163443)

[1.2 国内外研究现状 2](#_Toc514163444)

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 本文主要的工作 4](#_Toc514163445)

[1.4 本文的组织结构 4](#_Toc514163446)

[第二章 技术综述 6](#_Toc514163447)

[2.1 Socket 6](#_Toc514163448)

[2.2 Netty 7](#_Toc514163449)

[2.3 ZooKeeper 9](#_Toc514163450)

[2.4 本章小结 11](#_Toc514163451)

[第三章 网关服务器架构的分析与设计 12](#_Toc514163452)

[3.1 项目总体规划 12](#_Toc514163453)

[3.2 系统需求分析 12](#_Toc514163454)

[3.2.1 通信模块需求分析 13](#_Toc514163455)

[3.2.2 网关服务器模块需求分析 14](#_Toc514163456)

[3.2.3 服务注册中心模块需求分析 15](#_Toc514163457)

[3.2.4 游戏服务器模块需求分析 16](#_Toc514163458)

[3.2.5 非功能性需求分析 17](#_Toc514163459)

[3.3 系统总体设计 18](#_Toc514163460)

[3.3.1 总体结构 18](#_Toc514163461)

[3.3.2 总体流程 19](#_Toc514163462)

[3.4 通信模块设计 20](#_Toc514163463)

[3.4.1 自定义通信协议 20](#_Toc514163464)

[3.4.2 心跳检测设计 22](#_Toc514163465)

[3.5 网关服务器模块设计 23](#_Toc514163466)

[3.5.1 服务管理设计 23](#_Toc514163467)

[3.5.2 消息路由设计 24](#_Toc514163468)

[3.5.3 单播与多播设计 24](#_Toc514163469)

[3.6 服务注册中心模块设计 25](#_Toc514163470)

[3.6.1 服务注册设计 25](#_Toc514163471)

[3.6.2 服务发现设计 25](#_Toc514163472)

[3.6.3 服务监听设计 26](#_Toc514163473)

[3.6.4 节点管理设计 26](#_Toc514163474)

[3.6.5 负载均衡设计 27](#_Toc514163475)

[3.7 游戏服务器模块设计 27](#_Toc514163476)

[3.7.1 单播与多播设计 28](#_Toc514163477)

[3.7.2 扩展接口设计 28](#_Toc514163478)

[3.8 本章小结 29](#_Toc514163479)

[第四章 网关服务器架构的实现 30](#_Toc514163480)

[4.1 通信模块的实现 30](#_Toc514163481)

[4.1.1 通信协议的实现 30](#_Toc514163482)

[4.1.2 心跳检测的实现 33](#_Toc514163483)

[4.2网关服务器模块的实现 34](#_Toc514163484)

[4.2.1 服务管理的实现 34](#_Toc514163485)

[4.2.2 消息路由的实现 35](#_Toc514163486)

[4.2.3 单播与多播的实现 37](#_Toc514163487)

[4.3 服务注册中心模块的实现 38](#_Toc514163488)

[4.3.1 服务注册的实现 38](#_Toc514163489)

[4.3.2 服务发现的实现 39](#_Toc514163490)

[4.3.3 服务监听的实现 40](#_Toc514163491)

[4.3.4 节点管理的实现 41](#_Toc514163492)

[4.3.5 负载均衡的实现 42](#_Toc514163493)

[4.4 游戏服务器模块的实现 43](#_Toc514163494)

[4.4.1 单播与多播的实现 43](#_Toc514163495)

[4.4.2 扩展接口的实现 44](#_Toc514163496)

[4.5 系统测试 45](#_Toc514163497)

[4.6 本章小结 48](#_Toc514163498)

[第五章 总结与展望 49](#_Toc514163499)

[5.1 总结 49](#_Toc514163500)

[5.2 进一步工作展望 50](#_Toc514163501)

[参 考 文 献 51](#_Toc514163502)

[致 谢 54](#_Toc514163503)

[版权及论文原创性说明 55](#_Toc514163504)

# 图目录

[图2.1 Socket模型 6](#_Toc514163394)

[图2.2单线程模型 8](#_Toc514163395)

[图2.3多线程模型 8](#_Toc514163396)

[图2.4 ZooKeeper系统角色 10](#_Toc514163397)

[图3.1 通信模块用例图 13](#_Toc514163398)

[图3.2 网关服务器用例图 14](#_Toc514163399)

[图3.3 服务注册中心用例图 15](#_Toc514163400)

[图3.4 游戏服务器用例图 17](#_Toc514163401)

[图3.5 系统总体结构 18](#_Toc514163402)

[图3.6 系统时序图 19](#_Toc514163403)

[图3.7 游戏协议 20](#_Toc514163404)

[图3.8 HTTP协议 21](#_Toc514163405)

[图3.9 二进制格式协议 22](#_Toc514163406)

[图3.10 节点管理 23](#_Toc514163407)

[图3.11 节点命名空间 26](#_Toc514163408)

[图4.1 自定义二进制协议 30](#_Toc514163409)

[图4.2 消息编码代码 31](#_Toc514163410)

[图4.3 消息解码代码 32](#_Toc514163411)

[图4.4 消息解包代码 33](#_Toc514163412)

[图4.5心跳检测代码 34](#_Toc514163413)

[图4.6管理游戏服务器代码 35](#_Toc514163414)

[图4.7随机路由算法代码 36](#_Toc514163415)

[图4.8指定路由算法代码 37](#_Toc514163416)

[图4.9单播与多播代码 38](#_Toc514163417)

[图4.10服务注册代码 39](#_Toc514163418)

[图4.11服务发现代码 40](#_Toc514163419)

[图4.12服务监听代码 41](#_Toc514163420)

[图4.13节点管理代码 41](#_Toc514163421)

[图4.14随机算法代码 42](#_Toc514163422)

[图4.15一致性哈希算法代码 42](#_Toc514163423)

[图4.16单播、组播与广播代码 43](#_Toc514163424)

[图4.17扩展接口代码 44](#_Toc514163425)

[图4.18测试部署图 45](#_Toc514163426)

[图4.19 ZooKeeper服务器1启动 46](#_Toc514163427)

[图4.20 ZooKeeper服务器2启动 46](#_Toc514163428)

[图4.21 ZooKeeper服务器3启动 46](#_Toc514163429)

[图4.22 客户端1发送单播消息 47](#_Toc514163430)

[图4.23 客户端1发送分组消息 47](#_Toc514163431)

[图4.24 客户端2发送分组消息 47](#_Toc514163432)

[图4.25 客户端3发送广播消息 47](#_Toc514163433)

# 表目录

[表 1.1命名实体识别方法汇总表 13](#_Toc514163434)

[表 1.2关系抽取方法汇总表 15](#_Toc514163435)

[表 3.3服务注册中心用例表 16](#_Toc514163436)

[表 3.4游戏服务器用例表 17](#_Toc514163437)

# 第一章 绪论

## 1.1 项目背景

自然语言处理（Nature Language Processing，简称NLP）是用计算机来处理、理解以及运用人类语言（如中文、英文等），它属于人工智能（Artificial Intelligence，简称AI）的一个发展方向，是计算机科学和语言学的交叉学科，又常被称为计算机语言学。由于自然语言是人类与其他动物的区别的主要标志，没有语言，人类的思维也就无从可谈，所以自然语言处理体现了人工智能的最高任务与境界。也就是说，只有当计算机具备了处理自然语言的能力时，机器才算实现了真正的智能。近十年，随着计算机设备和硬件的发展、计算设备的算力极大提升，自然语言处理的研究得到了很大的发展，并在各个领域都有了广泛应用。

信息抽取作为自然语言处理中的一项基础任务，在许多上层应用中都有着十分重要的影响。在上市公司新闻舆情分析中的关键信息就是新闻文本中的实体，如：公司、人、行业等实体信息以及实体间的关系信息，利用信息抽取中的命名实体识别和关系抽取就可以从新闻文本中抽出这些重要信息，进而对新闻舆情进行后续分析；在搜索引擎上进行搜索时的关键信息就是输入的搜索信息中的实体信息，但只通过分词进行输入文本的切分很容易把一个多词组成的命名实体切分成多个词，造成搜索不准确甚至搜索的结果完全不符合预期。如果能对输入文本进行命名实体识别就可以准确识别搜索的内容，再建立索引进行搜索，可以提高搜索的准确率也可以提升搜索的效率；在利用非结构化文本信息构建知识图谱的任务中关键信息就是非结构化文本中的关联实体以及实体间的关系组成的三元组，如：从“阿里巴巴集团的创始人是马云”这段描述中可以提取出“阿里巴巴，创始人，马云”这个三元组，可以表示“阿里巴巴的创始人是马云”这条信息。这里即用到了命名实体识别识别出了阿里巴巴和马云这两个实体，又用到了关系抽取来对阿里巴巴和马云这两个实体的关系进行识别和抽取。最后从海量非结构化文本中提取出大量三元组，经过实体消歧、知识融合后构建成一张知识图谱。

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[MUC-6, 1996]。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[Milleret al. 2000]。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是构建一个高质量、高精度的知识图谱的关键所在。

## 1.2 国内外研究现状

## 1.2.1 命名实体识别研究现状

自从在MUC-6 [MUC-6, 1996]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[Bikel et al. 1999]、最大熵模型（Maximum Entropy Models，简称MEM）[Tsai et al. 2004]、条件随机场（Conditional Random Fields，简称CRFs）[McCallum et al. 2003]。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[张祝玉等, 2008]。

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。表1.1为目前命名实体识别的一些方法汇总。

## 1.2.2 关系抽取研究现状

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[Milleret al. 2000]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[Kambhatla et al. 2004]、基于核函数的方法[Zhao and Grishman 2005]、基于条件随机场的方法[Culotta et al. 2006]。在有监督学习中针对需要大量人工标注的情况下，Mintz等人[Mintz et al. 2009]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。[Socher et al. 2012]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 et al. 2014]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 et al. 2015]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 er al. 2016]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 et al. 2016]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。

关系抽取的一些方法汇总如表1.2所示。

## 1.3 本文主要的工作

本文在总结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，将训练好的模型封装成api接口，在利用非结构化文本构建金融领域知识图谱时调用命名实体识别和关系抽取的api来进行命名实体识别和关系抽取任务。

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. 研究了传统机器学习算法和深度学习算法实现的命名实体识别和关系抽取模型，并与本文实现的模型进行对比比较。
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. 将本文实现的模型封装成api接口，供对非结构化文本进行信息抽取时调用。

## 1.4 本文的组织结构

本文的各章节内容安排如下：

第一章 绪论。本章介绍了论文的项目背景，信息抽取任务中的两个子任务命名实体识别和关系抽取，国内外在这两个方向上的研究成果，本文所做的主要工作和论文的组织结构。

第二章 技术综述。本章详细介绍了论文涉及的相关技术，包括词向量、双向长短时记忆模型、深度残差网络、注意力机制（attention）、远程监督等。

第三章 基于深度学习的命名实体识别模型。本章详细介绍了本文选择的命名实体识别模型的各层网络结构和参数设定的对比实验和分析。

第四章 基于深度学习的关系抽取模型。本章详细介绍了本文选择的关系抽取模型的各层网络结构和参数设定的对比实验和分析。

第五章 待定。

第六章 论文的总结与展望。总结在论文期间做了哪些工作，在信息抽取任务中未来的可改进的地方以及对接下来的需要做的工作的进一步展望。

# 第二章 技术综述

## 2.1 引言

本章主要介绍了本文中实现的基于深度学习的命名实体识别和关系抽取所涉及的相关技术，包括词向量、双向长短时记忆模型、深度残差网络和注意力机制。针对这些模型和算法的各自特点和实际使用过程，分析了它们的优缺点。

## 2.2 词向量

人类可以轻松理解所有的自然语言，而计算机是不能直接识别自然语言的，要想让计算机能够读懂自然语言，必须要对自然语言进行进一步处理才行，因此词向量（Word Embedding）的概念应运而生。计算机的通用语言是二进制数字，所以词向量就是一些包含了自然语言中词汇和语义信息的计算机能够识别的数字。

最早的词表示方法是One-Hot编码，即用N个状态编码器对N个状态进行编码，这种方法会造成一篇篇幅有n个字的文本，会用一个n\*n的矩阵来表示，每个n维向量中只有它对应的那个字的位置上是1，其余位置都是0。这种词表示方法在传统机器学习中，如条件随机场、逻辑回归、最大熵、支持向量机等模型中可以较好的完成一些自然语言处理的任务。但是这种词表示方法中默认每个单词之间是独立存在的，忽略了文本中单词间的语义关联，且通常向量的维度会很高，在用深度学习处理自然语言处理问题时如果使用One-Hot编码容易造成“维度灾难”[Bengio et al. 2002]。

除了忽略了单词间语义信息的One-Hot编码，[Hinton, 1986]提出了一种新的词表示方法：分布式表示法（Distributed Reputation）。基于[Harris, 1954]提出的分布假说：上下文相似的词，其语义也相似。词的分布式表示主要分为2类：基于矩阵的分布表示和基于神经网络的分布表示。

## 2.2.1 基于矩阵的分布表示

基于矩阵的分布表示主要是构建“词-上下文”矩阵，通过某种技术从矩阵中获取词的分布表示。其中，矩阵的每一行都表示一个词，列表示上下文信息。常见的上下文有：

1. 文档，即“词-文档”矩阵。
2. 上下文的每个词，即“词-词”矩阵。
3. n-元词组，即“词-n-元词组”矩阵。

矩阵中的每个元素表示词和上下文共现的次数，可以利用词频-逆文本频率指数（TF-IDF）、取对数等技巧进行加权和平滑。矩阵维度较高且较稀疏的情况还可以利用奇异值分解（SVD）和非负矩阵分解（NMF）等方法进行分解降维使之变成低维稠密矩阵。基于矩阵的分布表示的典型代表是Global Vector模型（GloVe模型）[Pennington et al. 2014]。

## 2.2.2 基于神经网络的分布表示

基于神经网络的分布表示一般通过神经网络训练语言模型得到的产物，[Bengio et al. 2006]提出了一种神经网络语言模型（Neural Network Language Model。简称NNLM），考虑对语言模型进行建模。但是NNLM中词向量知识语言模型训练得到的副产物，并没有指出哪一套向量作为词向量的效果更好，所以[Mikolov et al. 2013a; Mikolov et al. 2013b]在NNLM等模型的基础之上，提出了word2vec模型，在word2vec模型中设计了CBOW（Continuous Bag-of-Words）和Skip-gram两个模型，高效地训练获取词向量。

CBOW和Skip-gram的图2.1

CBOW模型是根据上下文的信息来预测中间的目标词。CBOW模型对NNLM模型做了简化：隐藏层不再是上文各词词向量的拼接，而是使用上下文各词词向量的平均值，减少了计算量；去掉了tanh隐藏层，提升了模型的训练速度。CBOW模型包含3层：输入层、投影层和输出层。输入层共有n-1个词的one-hot表示作为词向量，组成上下文的表示：公式1。投影层将输入层的n-1个向量做均值计算：公式2。输出层一共V个节点，第i个节点表示中心词是词wi的概率。CBOW模型根据上下文的表示，直接对目标词进行预测：公式3。对于整个语料库而言，CBOW的优化目标为最大化：公式4

Skip-gram模型与CBOW模型刚好相反，它是根据一个单词来预测它的上下文信息。Skip-gram模型训练一个双层神经网络，模型输入为一个词，来预测这个词的前n个词和后n个词出现的概率。Skip-gram和CBOW一样没有隐藏层。Skip-gram的优化目标为最大化：公式5

## 2.2.3 词向量的使用

目前可供使用的训练词向量的方法有很多，因为本文所使用的编程语言是Python，Python编程环境中有Gensim库可以直接调用models.word2vec来训练词向量。[Radford et al. 2017]提出使用任务相关的语料训练得到的词向量相对于别的词向量在特定任务中会有更好的效果。由于本文所涉及的项目是针对金融领域的文本进行信息抽取且命名实体识别的序列标注任务是针对每个字来进行标注的，所以使用金融相关文本作为训练字向量的文本语料，训练了一个包含50102个字的100维的字向量模型。

## 2.3 双向长短时记忆模型

## 2.3.1 循环神经网络

循环神经网络（Recurrent Neural Networks，简称RNN）[Rumelhart et al. 1986]指的是一个序列当前的输出和之前的输出存在某种关联的神经网络。具体的表现形式为网络会对前面的信息进行记忆，保存在网络的内部状态中，并应用于当前输出的计算中。所以循环神经网络常常被用于对序列数据进行建模。循环神经网络的结构图如图2.n所示。

理论上循环神经网络能够对任何长度的序列数据进行处理，但是在实践中发现”梯度消失”会影响循环神经网络的训练，随着神经网络层数的增加，“梯度消失”的风险就会越大，当层数很大时，底层的神经元将接收不到返回的信号，出现“梯度消失”的问题。而长短时记忆模型（Long-Short Term Model，简称LSTM）[Hochreiter et al. 1997]的出现缓解了这个问题。

## 2.3.2 长短时记忆模型

长短时记忆模型的提出是为了解决循环神经网络的“梯度消失”问题，它之所以能够缓解这个问题主要取决于LSTM特有的门型结构，如图2.n所示。

可以看到，LSTM主要包括遗忘门、输入门和输出门等结构，其中：

遗忘门：公式7

输入门：公式8

输出门：公式9

新记忆单元：公式10

最终记忆单元：公式11

隐藏层：公式12

三个不同的门型结构和记忆单元是LSTM的核心所在，记忆单元负责保存历史和当前的重要信息，遗忘门负责控制选择性丢弃之前的记忆单元中的一些信息，输入门负责控制当前的输入中的某些信息加入到记忆单元中，输出门负责控制记忆单元中的哪些信息可以进行输出。这些门控结构可以帮助LSTM在长时和短时的记忆中都能处理并解决数据依赖以及选择是否丢弃信息（保留重要信息并丢弃不重要的信息）。三个门相互合作是LSTM能够解决RNN梯度消失/梯度爆炸的关键所在。

## 2.3.3 双向长短时记忆模型

一般我们所说的RNN、LSTM都是单向的，也就是说在一个方向上进行隐藏状态的传递，然而在某些任务中需要考虑的不仅仅是当前输入单词前面的文本信息，还需要考虑当前输入单词后面的文本信息。双向长短时记忆模型就可以解决这种类型的问题，它不仅可以从前向后传递隐藏状态的信息，还可以从向前传递隐藏状态的信息，这样在计算输出时考虑的就是完整的上下文信息。可以把双向长短时记忆模型看成是两个方向相反的LSTM的叠加，在分别做计算后把两个状态向量拼接得到最后的输出。

## 2.4 深度残差网络

深度学习中对于网络深度加深遇到的主要问题是梯度消失和梯度爆炸，传统对应的解决方案是数据的初始化（normalized initialization）和（batch normalization）正则化。

深度残差网络（Deep Residual Network，简称DRN）[He K M. et al. 2015]的提出是源于一个“反常”的现象：在训练深度神经网络的时候，训练误差和测试误差都随着网络深度的增加而增加（Degradation）。在自然语言处理中，利用网络结构训练文本数据，浅层的网络更容易学习到文本中的词汇信息，而深层的网络更容易学到文本中的语义信息，但是由于层数的增加会导致训练误差的累积，残差网络将浅层的网络与深层的网络进行级联再输出，很好地控制了训练误差的累积，在自然语言处理的各项任务中取得了很好的效果。

假设深度网络中的某隐藏层为公式13，如果多个非线性层组合可以近似于一个复杂函数，那么同样可以认为隐藏层的残差近似于某个复杂函数，那么可以将隐藏层表示为公式13。这样一来就可以得到一种全新的残差结构单元，如图2.n所示。

图

残差单元的输出由多个卷积层级联的输出和输入元素相加（维度相同的情况下），再经过ReLU激活后得到。将多个这样的结构级联起来，就得到了残差网络。

可以注意到残差网络有这样几个特点：1.网络较瘦，控制了参数数量；2.存在明显层级，特征图个数逐层递进，保证输出特征表达能力；3.使用了较少的池化层，大量使用下采样，提高传播效率；4.没有使用Dropout，利用BN和全局平均池化进行正则化，加快了训练速度。

## 2.5 注意力机制

深度学习中的注意力机制从本质上讲和人类的选择性视觉注意力机制类似，核心目标是从众多信息中选择出对当前任务的目标更关键的信息。

## 2.5.1 Encoder-Decoder框架

要深入理解深度学习中的注意力机制，就必须先介绍一下Encoder-Decoder框架，因为目前大多数的注意力机制的实现主要依靠Encoder-Decoder框架。

Encoder-Decoder框架是深度学习的一种研究模式，应用在很多场景之中。图2.n是文本处理中Encoder-Decoder框架的一种简单表示。

文本处理任务中可以把Encoder-Decoder框架看作一个由一篇文本生成另一篇文本的通用的模型。对于文本对[Source, Target]，目标是输入Source，通过模型最后生成Target。Encoder的作用是对输入文本Source=(x1,x2,...,xm)进行编码，通过非线性变化转化为中间语义表示C=F(x1,x2,...,xm)。再通过Decoder对中间语义表示C和已经生成的历史信息y1,y2,...,yt-1来生成t时刻要生成的单词yt：yt=G(C,y1,y2,...,yt-1)。每个yt依次生成，最后看上去整个模型就是输入Source最后生成了目标文本Target。

## 2.5.2 注意力模型

在Encoder-Decoder模型中，在生成目标句子的单词时，无论生成哪个单词，它们使用的中间语义表示C都是一样的没有任何区别。而中间语义表示C是由Source经过Encoder编码而成，这意味着无论生成哪个单词，Source中的任意单词对生成它们时所产生的影响大小都是相同的。当输入文本较短时这样的缺点暴露的还不够明显，一旦输入文本很长，此时所有的语义完全通过一个中间语义向量来表示将会丢失很多重要信息。这也是为什么要引入注意力模型：让模型更多地把注意力集中在对生成输出更重要的输入单词上。引入注意力机制后的Encoder-Decoder模型示意图如图2.n所示。

其中Ci的计算公式为：公式6 。其中Lx代表输入Source的长度，aij代表在Target输入第i个单词时Source中第j个单词的注意力分配系数，hj是Source中第j个词的语义编码。中间语义表示的形成过程类似于图2.n所示。

图2.n Attention的形成过程

本质上来说，注意力机制就是计算每一个输入单词对所有输入单词的相似度，得到一个分数即权重，然后根据这个权重对所有的输入进行加权求和。图2.n说明了注意力分配概率分布值的通用计算过程。

图2.n 注意力分配概率计算图

## 2.6 远程监督

远程监督（Distant Supervision）是目前关系抽取中比较常见的一种做法。它既不是传统意义上的监督学习，也不是无监督学习，它属于一种用KB（Knowledge Base）去对齐朴素文本的标注方法：1.使用NET（Named Entity Tagger）标注；2.对已标注的金融领域文本中出现的三元组提取特征（从所有出现该三元组的句子中），构造训练数据；3.采用多类别逻辑回归进行分类。

在测试使用时，先使用命名实体识别模型对目标文本中的命名实体进行标注，抽取其中的命名实体对和特征。如果多个句子的命名实体对一样，则把它们的特征合并在同一个特征向量中，再利用逻辑回归分类器对关系名称进行识别。

简单来说，如果训练语料中的句子所包含的实体对在已有的KB中有关系的体现，则认为语料库中所包含相同实体对的句子都表达相应关系。虽然这样做可以减少关系抽取任务对人工标注数据的依赖，但是会引入很多噪声数据。因为两个实体之间可能有别的关系（不同于KB中已有的关系）或者是没有关系，这样的训练数据就会对我们的关系抽取任务产生一定影响。

近些年来，也有很多国内外学者在研究如何降低远程监督中的噪声数据对关系抽取造成的影响，如使用Multi-instance从训练集中抽取置信度较高的训练样例来训练模型、利用Attention机制对数据进行全方位的权重计算从而得到全面而不失“选择”的训练数据等方法。

本文在利用远程监督处理标注数据时，结合Attention机制和残差网络（ResNet）共同作用减少噪声数据对关系抽取模型带来的影响。

## 2.7 深度学习框架Tensorflow

Tensorflow既是一个实现机器学习算法的接口，也是执行机器学习算法的框架。它前端支持Python、C++、Go、Java等多种开发语言，后端使用C++、CUDA等写成。除了执行深度学习算法，Tensorflow还可以用来实现很多传统机器学习算法，如逻辑回归、随机森林等。

Tensorflow使用数据流式图来规划计算流程，可以把计算映射到不同的硬件甚至是操作系统。Tensorflow的计算可以表示为有状态的数据流式图，可以让用户简单地实现并行计算，同时使用不用的硬件资源进行训练，同步或异步地更新全局共享的模型参数和状态。

## 2.9 本章小结

本章主要介绍了本文对信息抽取中的命名实体识别和关系抽取的研究中涉及的相关技术。先是介绍了作为模型输入的深度学习自然语言问题的基础：词向量；之后又分别介绍了双向长短时记忆模型、深度残差网络、注意力机制、远程监督等，并介绍了本文实现模型使用的深度学习框架Tensorflow。

# 

# 第三章 基于深度学习的命名实体识别模型

在总结和归纳了各种已发表的命名实体识别技术相关的论文后，本章基于已有的相关技术与模型，选择了双向长短时记忆模型+条件随机场来解决命名实体识别问题。本章讲详细介绍模型的每一部分具体的网络结构和各种参数的对比实验。

## 3.1 模型的整体架构

本文使用了一个端到端的模型来解决命名实体识别问题。本章采用了Bi-LSTM+CRF的模型结构来实现命名实体识别，其原理图如图3.1所示。

图3.1

先把输入文本按字组成一个输入文本的字典，字典中每个字都有一个唯一的ID（类似于唯一标识符），再把输入文本对照字典得到每个字在字典上的相对位置和字向量进行嵌入（embedding）操作，得到每一个字的向量。本文使用的字向量模型是基于金融领域上市公司新闻、金融公告、行业信息等中文语料，使用Gensim库中的word2vec模型训练得到的50102个字的100维字向量。把输入文本的字向量和词汇特征作为模型的输入，在中间网络层上使用Bi-LSTM，通过前向隐藏状态传递和后向隐藏状态传递分别对输入序列进行编码操作（即进行特征提取），有效地利用了前向和后向的所有信息（即上下文信息），最后利用CRF层从这些特征中计算出序列中每一个元素的标签。图3.2所示为模型训练的具体步骤。

图3.2

## 3.2 模型的标注策略

由于本文所研究的是金融领域下的信息抽取任务，所以标注语料都是来自于金融领域的新闻文本、上市公司信息、金融公告等。在做金融领域的命名实体识别时，本文将实体的类型分为公司（C）、人（P）、组织机构（INS）、行业（IND）、地点（L）、主营业务（PRO）六大类。本文对中文语料采用的是BIO标注方式，具体格式如图3.3所示：

（标注文本图）

B-标签1作为命名实体第一个字的标签，如果命名实体的字数大于1，则后面的字的标签都为I-标签1；如果命名实体只有一个字的话，则这个字的标签为B-标签2（中文中一个字作为命名实体的情况较少）；如果文本中的字不为命名实体的组成部分，则标签为O。所以最终得到的标签有：B-C、B-P、B-L、B-IND、B-INS、B-PRO、I-C、I-P、I-L、I-IND、I-INS、I-PRO、O共13种标签类型。

这样做的好处是能够把命名实体和非命名实体的字词有效地分隔开来，有利于之后的输出层进行标签分类。

## 3.3 输入层

模型的输入层是将文本输入的单词序列进行处理转化成单词的特征向量序列汇总后传输进网络层进行处理，而单词的特征向量是由输入文本的字向量和额外特征向量拼接组成的。如图3.4所示。

图

### 3.3.1 字向量

### 在利用深度学习模型解决中文自然语言处理问题时，通常将一个单词转化成词向量的形式作为模型的输入，由于本文的任务是中文命名实体识别，它属于一项序列标注任务，标注的对象是文本中的每个字，得到的输出也是针对每个字进行一个预测标签的计算，所以本文使用的是字向量。具体的详细设计如图3.5所示：

图3.5

使用预训练好的字向量能够使输入序列包含了字符的语义信息，所以能够提高模型的效果。

### 3.3.2 额外特征

对于输入文本，除了可以用文本的字向量作为其特征向量外，还可以引入额外的特征向量。本文在选取额外的特征向量时，把输入文本的词汇信息作为额外特征。本文采用jieba分词（jieba分词为Pyhton编程环境下的一个分词工具包），并结合金融领域的自定义词典对输入文本进行分词，对分词后的文本进行标记：若单词的字符长度大于等于3，则单词的第一个字符被标记为1，最后一个字符被标记为3，中间的所有字符被标记为2；如果单词的字符长度等于2，则单词的第一个字符被标记为1，第二个字符被标记为2；如果单词的字符长度为1（即1个字），则这个字符被标记为0。具体实现和示例如图3.6所示：

图3.6

最后将输入文本的字向量和分词后的词汇特征向量做拼接后作为模型的输入层结果输入到网络层中进行编码操作，如图3.7所示。

图3.7

由于深度学习模型中模型的参数很多，而标注好的训练样本又比较少，所以训练深度学习模型的时候最容易出现的问题就是模型的过拟合问题，模型的过拟合具体表现在：模型在训练数据上的损失函数值较小，预测的准确率较高。但在测试数据上的损失函数值较大，预测的准确率较低。本文在输入层和网络层之间加了Dropout来缓解神经网络过拟合的发生[Hinton et al. 2012]，在一定程度上起到了正则化的作用，具体实现如图3.11所示。

图3.11

Dropout之所以可以缓解神经网络过拟合的发生，是因为在前向传播时，Dropout让某个神经元的激活值以一定的概率p停止工作，这样可以使得模型的泛化能力变强，因为这样可以让模型不会太依赖某些局部的特征，如图3.12所示。

图3.12

## 3.4 网络层

一般来说同一个任务的不同模型主要的差异就是网络层选择的编码器不同，本文采用的是Bi-LSTM作为编码器对模型输入的特征向量进行处理，得到每个字符的预测特征向量，再传输给CRF层计算得到每个字符最终的预测标签。

图3.8

如图3.8是一个基于双向长短时记忆模型的编码层的结构图。假设图中维度为n的向量xt代表当前时刻输入的一个字对应的总特征向量。先将xt输入到网络层的双向长短时记忆模型中，分别与前向层、后向层中传递过来的t时刻的前、后的状态信息进行计算，得到t时刻前、后时刻的状态对当前状态的影响。再把前向层和后向层的当前时刻的输出做拼接，得到长度为2\*n的向量作为网络层的输出，图3.9为网络层的具体实现：

图3.9

## 3.5 输出层

输出层的作用是把网络层编码后的输出进行处理后得到最终的每个字符对应的预测标签。本文的输出层是由Softmax全连接层和CRF层联合构成。

### 3.5.1 全连接层

假设输出层得到的网络层输出为X=（x1, x2,...,xm），其中xt为当前时刻的字符对应的模型预测向量，xt会经过如公式3.5的计算得到输入对应的状态分数矩阵，原理图如图3.10所示。

图3.10

其中W的维度为(标签类别数量, 字向量的维度)，偏置项b的维度为标签类别数量。对于输出层来说需要训练迭代更新的参数就是W和b。

### 3.5.2 CRF层

在一般的命名实体识别模型中，即使不加入CRF层，即用Softmax层计算得到的状态分数矩阵并取其中的最大值作为预测标签的依据也可以得到最终的分类标签结果，但是那样的结果会存在一些问题导致最终的分类效果不好，如图3.13所示。

显然图中的分类结果并不准确，这也是为什么需要引入CRF层：CRF层可以学习到句子的约束条件，CRF层在训练数据时可以自动学习到这些约束，保证最终预测结果是有效的。

其中可能的约束条件有：

1. 句子的开头应该是“B-”或者“O”，而不是“I-”。
2. “B-标签1 I-标签2 I-标签3...”这样的模式中，标签1、标签2、标签3应该是同一种实体类别。比如，“B-P I-P I-P”这样的序列是对的，而“B-P I-C I-C”这样的序列是错误的。

有了这些约束条件，模型预测出错的概率将会大大降低。

CRF层接收到全连接层计算得到的状态分数矩阵为SC，sci,j代表序列中第i个字符预测为标签j的得分。CRF层中除了状态分数矩阵S，还有一个非常重要的分数---转移分数t。转移分数指的是序列从标签1转移到标签2的分数，如“B-C”转移到“I-C”的分数。为了使转移分数矩阵具有更好的鲁棒性，本文在每个句子的开头加上START标签作为句子开始的标记（所以此时的标签类别个数为14个），在转移分数矩阵中START的分数为0。表3.1为转移分数矩阵的示例。

表3.1

从表中可以看出转移矩阵已经学到了一些有用的约束信息，如：从“START”到“I-P”或者“I-C”的分数很低，但到“B-P”和“B-C”的分数很高，可以看出句子的第一个单词的标签应该是“B-”或者“O”，而不是“I-”；“O”后面接“I-标签”的分数很低，可以看出命名实体的开头应该是“B-”而不是“I-”。

设公式3.6为CRF层的转移分数矩阵，ti,j代表从标签i转移到标签j的分数。转移分数矩阵作为Bi-LSTM+CRF模型中的一个参数，在训练模型之前可以随机初始化得到这样一个矩阵T，T中的分数会随着模型训练的更新迭代不断更新，也就是说CRF层可以自己学到这些约束，CRF层中转移分数矩阵T的定义如图3.14所示。

图3.14

假设输入序列X的可能预测序列类别为Y=（y1,y2,...,yn）,每种可能预测序列标签的分数为pi，共有n条路径，那么路径的总分为：公式3.7

其中公式3.8，

那么得到正确的预测序列标签的条件概率为P(y|X):公式3.10

随着模型训练的进行，模型的参数不断更新，使得真实的预测序列标签占所有预测路径的比例越来越大。取Ptotal的负对数作为损失函数Loss，训练的目标为最小化损失函数以此来调整模型参数，即：公式3.9。训练完成后，使用viterbi\_decode()调用维特比算法计算出概率最高的序列标签作为预测序列标签。

## 3.6 实验

宋体，小四

### 3.6.1 实验数据

宋体，小四

### 3.6.2 实验评价指标

宋体，小四

### 3.6.3 实验

宋体，小四

## 3.7 游戏服务器模块设计

### 3.7.1 单播与多播设计

### 3.7.2 扩展接口设计

## 3.8 本章小结

# 第四章 基于深度学习的关系抽取模型

在总结和归纳了各种已发表的命名实体识别技术相关的论文后，本章基于已有的相关技术与模型，选择了残差网络+卷积神经网络来解决关系抽取问题。本章讲详细介绍模型的每一部分具体的网络结构和各种参数的对比实验。

## 4.1 模型的整体架构

深度残差网络作为一种比较新的深度神经网络模型常被用于图像领域相关任务，因为它在层数较深时的表现十分优异。本文使用了一种新的CNN模型结合残差网络的方法用于关系抽取问题的研究，不同于以往残差网络的结构较深的特点，本文采用的卷积神经网络结构只有9层，并且得到了不错的分类效果。实验中还发现使用身份映射（identity mapping）能够显著提升远程监督关系抽取问题的性能。

ResCNN主要的结构分为输入层、卷积层、残差卷积块、最大池化层和输出层几部分。模型的结构图如图4.1所示。

图4.1

输入层包括词向量层和位置向量层；卷积层是对输入层输出的文本特征向量做卷积操作来提取特征；残差卷积块利用残差学习结合卷积神经网络对卷积层的输出进行进一步处理；最大池化层对残差卷积块的输出进行最大池化操作并取其中的最大值；最后把这些特征传入Softmax层进行计算得到最终的预测结果。本章将会对每层的网络结构和设计细节做详细的介绍。

## 4.2 输入层

### 模型的输入层是将文本输入的单词序列进行处理转化成单词的特征向量序列汇总后传输进后续的卷积层进行卷积处理，而单词的特征向量是由输入文本的词向量和相对位置向量拼接组成的。

假设xi是输入文本的第i个词，(e1, e2)是文本中的实体对，将xi在预训练好的字向量表中对应的字向量设为WFi，相对位置向量设为PFi，最终把WFi和PFi拼接起来变成词xi的总特征向量vi=[WFi, PFi]。

### 4.2.1 字向量

由于本章关系抽取的任务是接着上一章的命名实体识别之后继续对金融领域的文本进行实体间关系抽取，所以本章所采用的关系抽取模型使用的字向量也是基于之前的金融领域上市公司新闻、金融公告、行业信息等中文语料，利用Gensim库中的word2vec方法进行预训练得到的字向量模型。

每个xi对应的总特征向量vi都是一个实值向量，所有的向量组成了一个特征向量矩阵公式4.1，其中V是一个输入文本的所有单词构成的词表。

### 4.2.2 相对位置向量

在对输入文本做特征提取时，通常的做法是对输入文本中的单词做一个词向量转换来学习到文本中的语义信息。然而在关系抽取任务中，要想最终获得两个实体间的关系，不仅可以从词汇信息和语义信息的角度出发，还可以从当前单词与实体之间的相对位置中获取一些有用的信息[Zeng et al., 2014]。

相对位置向量PFi结合了当前字符xi分别与两个实体e1和e2的相对距离，比如在句子“x0x1e1x2x3x4e2x5.”中，x2与e1和e2的相对距离分别为1和-3。之后通过一个随机初始化的相对位置矩阵公式4.2将相对位置距离转化成实值向量。如果当前字符距离命名实体的距离过远，那么可能会被认为与实体间的关系没有关联而在最后分类的时候被忽略。

最后把每个字的字向量和相对位置向量做拼接，再把句子中的每个字的总特征向量拼接成一个向量：公式4.3

## 4.3 卷积层

卷积层的作用是把输入层输出的总特征向量进行卷积操作后学习到输入文本的潜在特征。

用vi:i+j来表示特征向量vi,vi+1,...,vi+j拼接得到的向量。卷积操作就是用一个大小为h的窗口在卷积层的输入特征向量上进行滑动，对窗口中的所有字符对应的特征向量做矩阵计算，得到新的特征c。特征ci是由一个滑动窗口中的字符对应的特征向量vi:i+h-1通过公式公式4.4计算得到的。其中b为偏置项，f是一个非线性函数。

最终通过滑动窗口使所有的特征向量经过卷积操作得到了特征c=[c1, c2, ..., cn-h+1]作为卷积层的输出。

## 4.4 残差卷积块

残差学习将低层的信息直接与高层的信息进行联接来消除深层网络中的梯度消失问题，本文中的残差卷积块就是利用了快捷连接（shortcut connections）对卷积层的输出进行处理。

每个残差卷积块是一个由2层卷积神经网络组成的序列，所有卷积神经网络中的卷积核大小都为h，每层卷积神经网络后还接了一个ReLU激活函数。ReLU激活函数常在训练神经网络时作为中间隐藏层的一种激活函数，其表达式为公式4.9。相比于最常用的Sigmoid激活函数，由于优化参数时一般会用到误差反向传播算法，要对激活函数进行求导，而Sigmoid激活函数的导数表达式为公式4.10，对应的图形如图4.5所示。

图4.5

由图可以看出Sigmoid激活函数的导数从0开始后很快就会趋近于0，容易产生“梯度消失”现象，而ReLU激活函数就不会存在这样的情况，它的导数表达式为公式4.11，对应的图形如图4.6所示。

图4.6

使用ReLU作为激活函数的好处有：

1. 单侧抑制。即使所有的负值都变为0，而正值的数值不变。正因为ReLU的单侧抑制才使得神经网络中的神经元具有了稀疏激活性，也就是说当模型的层数增加了N倍后，ReLU神经元的激活率将降低2的N次方倍。
2. 对于线性函数而言，ReLU的表达能力更强，尤其在网络很深的情况下。
3. 对于非线性函数而言，ReLU由于非负区间的梯度为常数，所以不存在梯度消失的问题，而且模型的收敛速度维持在一个比较稳定的状态。

设残差卷积块中的2层卷积神经网络的卷积过滤器分别为w1和w2，公式4.5。第一个卷积层：公式4.6。第二个卷积层：公式4.7.

其中b1和b2都是偏置项。最后再通过残差学习操作：公式4.8。具体的残差卷积块的实现代码如图4.4所示。

图4.4

本文实现的模型中使用了9个这样的残差卷积块，进行了一系列的卷积+残差连接的操作后最后得到的特征矩阵仍记为C。

## 4.5 最大池化层

卷积神经网络中

## 4.6 输出层

宋体，小四

## 4.7 实验

宋体，小四

# 第五章 总结与展望

## 5.1 总结

本文以上海某大数据人工智能公司的知识图谱产品中的信息抽取模块为背景。随着互联网的普及与发展，网上的数据量呈爆炸式增长，而其中非结构化文本数据一直都是占比最高的数据类型之一，所以在知识图谱的构建和更新过程中如何从非结构化文本中提取有用信息一直是大家关注的重点。

## 5.2 进一步工作展望

宋体，小四

# 

# 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.

# 致 谢

转眼之间我的研究生生涯就要结束了，对南大有一丝丝不舍，对未来有一丝丝憧憬和期待，在此我要向南大的研究生老师们和同学们表达最真挚的感谢。

首先我要感谢我的毕业论文指导老师申富饶教授，感谢申老师对我毕业论文的指导和修正。毕业论文开题前，申老师与我就毕业论文事宜进行了详谈，给我毕业论文的选题和方向做出了指导和建议，在论文的写作过程中，申老师一直耐心地给我写作意见，申老师对毕业论文负责任的态度，给了我莫大的帮助。

其次我要感谢南京大学软件学院的所有老师们，感谢老师们严谨的教学态度和学术态度，帮助我培养了良好的学习习惯和学习态度。还要感谢周围同学们在我学习生活中给予的帮助，给了我继续前进的动力。

再次我要感谢我的家人，他们在我最困难的时候默默地支持和鼓励着我，让我战胜了困难，最终进入南京大学就读研究生，实现了自己的梦想。

最后，由衷地感谢在百忙之中评阅论文的各位专家和教授。

# 版权及论文原创性说明

任何收存和保管本论文的单位和个人，未经作者本人授权，不得将本论文转借他人并复印、抄录、拍照或以任何方式传播，否则，引起有碍作者著作权益的问题，将可能承担法律责任。

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不包含其他个人或集体已经发表或撰写的作品成果。本文所引用的重要文献，均已在文中以明确方式标明。本声明的法律结果由本人承担。

作者签名：

日期： 年 月