南京大學

研 究 生 毕 业 论 文

（申 请 硕 士 学 位）

论 文 题 目 (论文长标题第一行)

(论文长标题第二行)

作 者 姓 名 作者

学科、专业方向 计算机科学与技术

研 究 方 向 分布式计算

指 导 教 师 某 教授

2016 年 **6** 月 **8** 日

LaTeX NJU thesis template

by

Author

Supervised by

Professor

A dissertation submitted to
the graduate school of Nanjing University
in partial fulfilment of the requirements for the degree of
MASTER
in
Computer Science and Technology

Department of Computer Science and Technology
Nanjing University

May 20th, 2016

南京大学研究生毕业论文中文摘要首页用纸

毕业论文题目：南大本科毕业论文 LaTeX 模板

计算机科学与技术 专业 2012 级硕士生姓名：作者

指导教师（姓名、职称）：某 教授

摘　要

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

关键词： 关键词 1　关键词 2

南京大学研究生毕业论文英文摘要首页用纸

THESIS: englishabstracttitlea
nglishabstracttitleb

SPECIALIZATION: Computer Science and Technology

POSTGRADUATE: Author

MENTOR: Professor

Abstract

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

keywords: keyword1 keyword2

前　　言

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

作者

20xx 年夏于南京大学

目　　录

前　　言 ····· **iii**

目　　录 ····· **iv**

插图清单 ····· **vi**

附表清单 ····· **vii**

1　绪论 ····· **1**

1.1　研究背景及意义 ····· 1

1.2　国内外研究现状 ····· 2

1.2.1　命名实体识别问题的定义 ····· 2

1.2.2　传统方法研究发展现状 ····· 3

1.2.3　深度学习方法研究发展现状 ····· 5

1.3　研究内容及工作 ····· 7

1.4　论文的组织结构 ····· 8

2　相关技术介绍 ····· **9**

2.1　引言 ····· 9

2.2　命名实体识别问题建模 ····· 9

2.3　条件随机场 ····· 10

2.4　词向量 ····· 11

2.5　循环神经网络 ····· 14

2.5.1　循环神经网络概述 ····· 14

2.5.2　长短期记忆网络 ····· 15

2.6　概率隐语义分析技术 ····· 17

2.7　本章小结 ····· 17

3 改进的中文字符级特征表示方法 ····· 18
3.1 引言 ····· 18
3.2 相关理论与工作 ····· 18
3.3 面向位置信息的字符向量优化方法 ····· 18
3.4 面向主题信息的字符向量构造方法 ····· 18
3.5 结合位置信息与主题信息的中文字符级特征表示方法 ····· 18
3.6 本章小结 ····· 18

4 面向复杂命名实体识别的层次深度神经网络模型 ····· 19
4.1 引言 ····· 19
4.2 复杂命名实体概述与相关工作 ····· 19
4.3 层次标签与层次深度神经网络模型构建 ····· 19
4.4 实验结果及分析 ····· 19
4.5 本章小结 ····· 19

5 中文复杂命名实体识别在企业风险识别中的应用 ····· 20
5.1 引言 ····· 20
5.2 应用背景 ····· 20
5.3 数据爬虫 ····· 20
5.4 文本分类模块 ····· 20
5.5 命名实体识别模块 ····· 20
5.6 企业风险识别 ····· 20

6 总结与展望 ····· 21
6.1 工作总结 ····· 21
6.2 不足与展望 ····· 21

致　谢 ····· 22

参考文献 ····· 23

A MPTCP 内核源代码修改 ····· 26
A.1 函数 mptcp_v4_subflows() ····· 26

简历与科研成果 ····· 27

插图清单

2-1 线性链条件随机场 …… 11
2-2 CBOW 模型 …… 13
2-3 Skip-gram 模型 …… 14
2-4 简单的 RNN 模型 …… 15
2-5 一个 RNN 层和展开结构 …… 15
2-6 LSTM 单元 …… 16

附表清单

2-1 BIO 命名实体标签体系 …… 10

第一章　绪论

1.1　研究背景及意义

互联网自诞生于上个世纪以来，正一步一步的改变这地球上每个人的生活。特别是伴随移动通信技术的革新，结合互联网技术和移动通信技术的移动互联网正深刻地改变着生活的方方面面。饮食、购物、交通、居住、社交、娱乐等等方面，在当下的移动互联时代都有了新的运作生态。随着移动互联技术的成熟和在“摩尔定律”下硬件成本的降低，人们越来越容易地融入互联网时代，根据中国互联网络信息中心发布的数据，截至 2018 年 6 月，我国网民规模已经超过了 8 亿，渗透率近六成，而其中移动互联网用户比例高达 98%。海量的用户在使用互联网时也正有意无意地创造着海量的数据，而海量的数据中蕴藏着巨大的价值。这些数据的类型包括数值型数据、文本型数据、图片型数据、视频型数据、音频型数据等，利用好这些不同类型的数据可以创造出大量的经济价值和社会价值。

自然语言处理技术（Natural language processing）是计算机信息工程的一个子领域，目标便是处理和分析海量的文本数据，使得计算机程序可以利用词法、语法、语义等信息对自然语言文本完成识别、理解与输出等任务，例如词语分割、命名实体识别、关系抽取、机器翻译、自然语言生成、问答系统、情感分析等等。自然语言技术通过规则学习、统计学习等方法的研究与探索日臻成熟。近十年，表示学习、深度神经网络类机器学习技术给自然语言处理技术带来了新的探索与发展，在部分自然语言处理问题上可以达到良好而稳定的结果。自然语言处理技术在各行各业有着多种应用：社交媒体上的评论文本数据可以用来辅助监测舆情舆论的走向；财经新闻中包含诸多经济数据、公司运营情况，利用这些文本数据可以辅助量化交易的执行；利用新闻媒体中的海量文本数据，我们可以对用户兴趣话题进行建模，高效地为读者进行内容过滤和兴趣推荐；机器翻译技术可以将不同语言为载体的文献自动翻译，促进不同文化间的沟通和交流；知识图谱技术可以链接不同的人和组织，构造知识库，服务与多种商业应用。

命名实体识别（Named-entity recognition），又称实体抽取技术、实体分块技术，是自然语言处理技术的一个子问题。目标在于将非结构化文本中提及的命名实体抽取出来，例如人名，组织名，地点名，医疗术语法规术语，时间，数量，货币价值等等。例如在财经文章中需要准确地抽取企业名称、重要人物名称、货币价值等命名实体；在政治新闻中需要准确地抽取政治人物名称、国家地理名称、组织机构名称、事件名称等命名实体；在判决文书文本中，需要抽取出当事人名称、处罚条款、量刑情况、关联组织等信息。可以说，命名实体识别问题是自然语言处理最基础的任务之一，命名实体识别的准确率、召回率的高低直接影响着后续自然语言处理问题，例如信息抽取、文本分类、文本摘要、问答系统等等研究方向。

因而研究中文命名实体识别问题，对于中文自然语言处理技术的研究有着关键性的地位。通用命名实体技术对于中文命名实体识别有着不错的效果，然而中文与其他许多种语言的构词、语法有着诸多不同。特别是词语的边界模糊、无大小写和时态词型的变化、一词多性，一字多义、简称方法独特等等独特之处，因而相适应地根据中文语言的特点对通用命名实体技术进行优化有着很强的必要性。因而针对中文语言的各种特点，相适应地研究如何提高命名实体识别的效果就有着重要的意义。

1.2 国内外研究现状

1.2.1 命名实体识别问题的定义

命名实体识在信息抽取和自然语言处理工作中有着重要的地位。例如在一段文本中准确地识别人名、地名、组织名、时空表达等等要素，对后续其他的自然语言处理过程有着基石作用。人们对该问题的研究历程中使用过很多很多种方法，大体上这些方法可以分为知识工程方法和机器学习方法。一个典型的命名实体识别系统的输入应该是输入的自然语言文本，而输出应该是抽取出的信息并包括这些信息的边界，以及这些信息分别属于什么样的命名实体类型。例如“小明在小雨的陪同下，一起在南京看了中国国家队的比赛。”这样的一个文本输入，命名实体识别系统应该给出，“小明”、“小雨”【人物】，“南京”【地点】，“中国国家队”【组织名】这样的输出。

命名实体识别研究是一个较宽泛的范围，影响命名实体识别工作方法方案的有以下几个因素 [1]：

（1）语言因素。目前，大量的研究工作都是以英文为研究对象，但是这些方案并不一定可以推广到所有的语言。例如几个东方文明的语言中文日文韩文语并不像英文单词那样有着天然的空格作为隔断，而且也不存在字母的大小写，这些特性都会使得命名实体识别的方法上有着较大的差异。就算是同是西语概念下的法文德文西班牙文等也有自身语言的特殊性。除英文外目前中文日文法文意大利文希腊文已经有了大量研究，收集整理了大量的语料和工具。针对印度文丹麦文韩文土耳其文等语言的研究也在一直的进步中。这些工作很多都是限定在自己的语言边界内，研究出一个跨语言跨文化的命名实体识别模型是当下的一个目标。

（2）领域因素。最初命名实体识别的很多工作是面对半结构化的数据，例如病历、报名表、简历、申请材料这样的文本，这些工作有各自不同的特殊性方法中会使用很多先验知识，因而技术技巧的移植困难。除此之外，文本所属领域的不同对命名实体识别工作影响也很大，文本内容属于科学技术文章或是商业、体育、旅游等领域，这些不同的领域内容也会对最终命名实体识别的效果产生很大的影响。因此，最终一个效果良好稳定可靠的命名实体识别系统需要拥有尽可能多的不同领域的预料，找到并实时更新大量不同领域的语料库也是目前一个相当大的挑战。

（3）实体和标注方法。命名实体这样一个概念在不同的语境和不同的业务需求下是有区别的。大多数的命名实体识别的研究将实体的类别分为“人”、“地点”、“组织”三类。这种分类方法在 MUC-6（6th Message Understanding Conference，命名实体识别重要会议）中被确定为一个标准称为“Enamex”分类法。然而许多文章指出这种分类方法较为粗糙，类别还可以继续细分，比如“人”这个标签下可以分为“医生”、“政治人物”、“艺人”等等，“地点”也可以为分为“国家”、“州省”、“景区”等等。

1.2.2 传统方法研究发展现状

早期命名实体识别任务并不统一，主要目标是要从一堆文本数据中自动识别出命名实体。最早的相关研究论文是 1991 年 Rau[2] 在人工智能应用会议上发表的，论文介绍了一个自动识别公司名称的系统。主要的方法是采用启发式方法和手工规则。1996 年，命名实体识别这个术语在 MUC-6 会议上被 R. Grishman 和 Sundheim 正式提出，该领域被越来越多的人关注，进入快速发展时期。

然而早期的研究大多还是主要依靠手工规则等办法，后来用有监督的机器学习方法逐渐火热起来。规则的设计大多是基于特殊的领域知识。Kim[3] 用规则的的方法对于口语输入的文本进行自动的命名实体识别。在生物医学领域，Hanisch[4] 利用预处理的同义词点来识别生物医学文本中提到的潜在的蛋白质术语。Quimbaya[5] 等提出了一个基于词典的方法来提取电子医疗记录中的命名实体。实验结果表明这样的方法在提高召回率上很有效，但是在提高准确率上效果有限。大多数这样基于手工规则的方法都是利用启发式的语法语义特点，或是要用到领域内特殊的知识做成字典。这些系统在提高识别的精确的和召回度上都有很多局限。

很多学者在命名实体识别任务中引入了基于特征的有监督统计学习方法。命名实体识别问题被建模为一个多分类的任务或者一个序列标注任务。将有标注的样例数据交给模型训练，利用机器学习算法来识别其他数据中潜在的类似的命名实体模式。在这一类方法中，特征工程就会起到关键作用。单词的表示方法 [6]，单词的的特征（如形态、读音等）[7]、文本语料的特征（局部句法特征和出现次数等）[8] 等等自然语言固有的特性都被用来提高识别的效果。不同的机器学习模型也会带来不同的识别效果。经典的有监督机器学习模型被一一引入命名实体识别系统：隐马尔科夫模型（HMM）[9]、决策树模型（DT）[10]、最大熵模型（ME）[11]、支持向量机模型（SVM）[12]，条件随机场模型（CRF）[13]。这些有监督机器学习模型读入大量带标签的训练语料，学习出命名实体的特征，最终对新的语料中的命名实体进行预测。

针对标注训练语料费时费力，而大量未标注的数据容易获得的情况，半监督学习方法在许多学者的尝试下引入了命名实体识别问题。S. Brin[14] 利用词汇的特征，通过半监督学习方法，以一小部分单词作为种子通过词汇特征产生新的训练语料。J. Heng[15] 等指出 bootstrapping 方法是如何提高一个现有的命名实体识别系统的识别能力的，给出了如何去选择半监督学习中无标记数据的方法。

经过多年的发展，基于统计学习的命名实体识别方法逐渐成熟。在各种语料数据集上表现好的模型基本上是将命名实体识别任务建模成为序列标注的任务，利用大规模的标注语料进行学习，对于句子中的每个单词进行标签的分类。McName 和 Mayfield[16] 采用了 1000 个语言相关的特征和 258 个拼写和标点特征来训练 SVM 分类器，每个分类器将单词分类至 8 个类别标签实现命名实体识别。Isozaki 和 Kazawa[17] 发明了一种使得 SVM 分类器在命名实体识别

任务上训练更加快速的方法。Li[18] 等提出了一种基于 SVM 模型的从用不平坦分类超平面的命名实体识别方法，在一些数据集上有不错的表现。

不过 SVM 方法的缺陷是并不包含单词的“邻居”信息，而这一信息在判断命名实体边界时十分重要。因为训练的语料中，相同的单词（汉语的字）在不同的语境中对应的标签常常不同，十分依赖上下文的信息。而 HMM 和 CRF 这样的概率图模型在这一任务上表现颇佳。McCallum 和 Li[19] 提出了一个特征归纳方法通过 CRF 进行命名实体识别，在英文数据集 CoNLL03 上 F 值达到了 84.04%。何炎祥 [20] 等提出的基于 CRF 和规则相结合的地理命名实体识别方法取得了不错的中文命名实体识别效果。张素香 [21] 对于同样的中文语料对比了最大熵和条件随机场模型，得出了条件随机场模型的表现优于最大熵模型的结论。加入非局部特征，使用条件随机场模型进行命名实体识别成为了一种较为主流的方法。

1.2.3 深度学习方法研究发展现状

近年来深度神经网络模型在图像处理和语音识别任务上取得了巨大的成功，许多学者也迅速将这一模型引入到自然语言处理任务中来。2011 年，Collobert[22] 用神经网络模型来自动化命名实体识别任务中的特征抽取，从而减少特征工程的工作量。从此开始，神经网络在命名实体识别任务中的研究开始流行起来。

深度学习是机器学习的一个子领域，该类方法通过多层的抽象层来学习和表示数据。典型的层次结构是人工神经网络。这种方法采取端到端的机器学习概念，从原始数据中自动探索潜在的表示特征，并进行分类和识别。深度学习对于命名实体识别工作来说有三个关键优点：（1）相较于线性模型（典型的比如线性链条件随机场），深度学习可以找到更多非线性的联系，表示能力更强。（2）传统方法的命名实体识别花了大量的工作和技巧在构造数据特征上，而深度学习自动从原始数据中学习有用的数据表示（3）深度学习模型能够端到端地通过梯度下降求解，这个性质让我们可以设计一些更加复杂的命名实体识别系统。

在深度学习在自然语言处理和序列标注领域学者们做了一系列有成效的工作：

（1）输入的分布式表示：词语级别的表示 [23]，收集大量数据采用例如 CBOW 和 skip-gram 等无监督学习算法进行训练，得到预训练好的词语向

量，作为命名实体识别模型的输入，常用的词语向量有 Google 的 Word2Vec，Stanford 的 GloVe，Facebook 的 fastText 等。Yao 等 [24] 利用这种表示方法提出了一个基于词与向量表示的生物以学命名实体识别系统，该系统采用总大小 205924 的词典，训练出位数为 600 的词向量；字符级别的表示，词语的粒度还可以继续细分 [25]，对于中文等字词词素文字，还可以以单个字为粒度做分布式表示，对于英文等字母文字则可以有意义的单词子序列（比如前缀后缀等）作为要素做分布式表示。Ma 等 [26] 采用了一个 CNN 模型来提取字母级别词语的表示。Yang 等 [27] 提出了在卷积层设置一个固定大小窗口来提取字符级别的特征。Lample 等 [28] 采用了一个双向 LSTM 模型来抽取字符级分布式表示；混合分布式表示，除了词语级别和字符级别的表示外，还有很多工作利用了其他的信息。比如在基于深度学习的表示外联合基于特征的信息，合并这些表示一同放入神经网路的输入。Huang 等 [29] 构建了一个 BiLSTM-CRF 模型，共使用了四种类型的特征，分别是拼写特征，内容特征，词语向量和词典特征，他们的实验表明这些额外的特征可以提高标签的准确率。Strubell 等 [30] 训练了 100 维的词语嵌入式表示和 5 维的单词形状向量（例如全字母大写，小写，首字母大写，包含大写字母）作为输入。还有很多混合型方法用到了情感、语义 [31] 等特征。

（2）上下文编码器结构：深度学习处理命名实体任务的第二个环节就是给上下文选取合适的编码器结构，最常用的模型是卷积神经网络（CNN）、循环神经网络（RNN）。Wu 等 [32] 使用卷积层来生成全局特征；Zhou 等 [33] 发现 RNN 模型中后来的词语对于最终句子表示的影响大于之前词语的影响；Strubell[30] 等采用了迭代空洞卷积网络（ID-CNNs）来做命名实体识别，效果较传统 CNN 方法好。RNN 结构，包括他的变体 GRU、和 LSTM，被证明在序列数据上有着很好的表现。特别是双向 RNNs[29] 即利用了过去的信息和状态也可以利用未来的信息和状态，效果良好。这样的双向结构已经成为一个标准结构被广泛使用。Katiyar 和 Cardie[34] 提出了一个针对标准 LSTM 结构的修改来应对嵌套命名实体识别问题。Ju 等 [35] 提出一个动态栈来识别嵌套命名实体识别，对于探测出的实体进行下一步的嵌套命名实体识别。

（3）标签解码结构：标签解码是命名实体识别模型的最后一个环节。这个过程是要将以文本表示为输入，最终得出一系列标签关联到输入序列上去。主要采用的方法有以下几种，多层感知机 +Softmax 层输出 [30]、条件随机场 [29]、循环神经网络 [36] 几种。多层感知机是将问题建模为一个多分类问题，

每一个标签独立预测，没有参考周边“邻居”信息。条件随机场解码方法是最常用的解码方法，在 CoNLL03 数据集上有目前最好的结果 [37]。循环神经网络解码方法的优点是 [36] 当命名实体类别较多时，解码的速度较快，因为条件随机场的解码方法采用动态规划思想的维特比算法，在类别多时解码时间较长。

总而言之，用深度学习方法解决命名实体识别问题，RNNs 结构和 CRF 解码器的组合是现今使用最广泛的模型。尤其是 BiLSTM-CRF 结构是采用深度学习进行命名实体识别最常见的结构。而模型成功的关键也很依赖于输入的表示方法。

1.3 研究内容及工作

本文研究的问题是复杂中文命名实体识别问题。复杂中文命名实体识别指在中文环境下，具有命名实体名称长，标签混淆等特征的命名实体识别问题。该问题有以下困难和挑战：

（1）中文字符与字符之间没有英文那样的天然空格分隔语义，将词素直接转化为向量不可行。若是采用中文分词技术的话，分词的准确率就会极大地影响后续的命名实体识别工作，一旦分词错误，后续命名实体识别任务基本不能成功，除此以外命名实体中的人名组织名往往在含义上与词库中字词有很大的差别，分词工作执行困难。单纯的以字符为单位的话，汉字一字多意，一字多性十分普遍，信息混淆影响分布式表示效果，将字符向量化的过程有很多困难。

（2）中文以单字为最小单位嵌入分布式表达的情况下，中文字符较英文单词而言，总量小，词根前缀后缀等构词信息。在汉字演进简化的过程中，构词信息变化太大，把握困难。这些因素造成了直接将字符向量嵌入的方法，信息量不足，命名实体识别召回率不高。

（3）复杂中文命名实体构词长，可能由多个命名实体拼接而成，也有可能因为复杂度高被误识别为更长的命名实体。从命名实体识别模型的角度来看就是命名实体标签混淆，从而导致识别困难，准确度下降。

针对上述问题，本文提出了对应的解决方案，构建了面对复杂中文命名实体问题的实体识别框架，并将该框架应用到了企业风险识别的应用中。本文的主要工作如下：

（1）在分布式特征表示方面：针对中文字符向量缺少环境信息，一词多义

而产生的信息表达不准确问题，本文提出了一种面向位置信息的字符向量优化方法；针对中文字符向量信息量不足，缺少字词与字词间联系的问题，本文提出了加入字符级的中文主题特征向量，为神经网络层传递更多语境语义信息，增强分布式表示的表达效果。

（2）在网络结构方面：针对复杂中文命名实体长度长，标签含义混淆的难点，本文提出了一个多层次的深度神经网络模型，上层模型利用标签关联关系抽取模糊信息，定位目标文本，下层模型根据准确的标签信息对复杂中文命名实体识别进行精确抽取和识别。

（3）最终本文将提出的在分布式表示的优化和网络结构的优化集成在命名实体识别系统中，将基于这些改进方法的中文命名实体识别模型应用到企业产品风险监测系统中。

1.4 论文的组织结构

第 1 章：绪论。介绍研究的背景及意义，从命名实体识别问题的定义和研究子方向起，介绍了传统方法和深度学习的研究发展和研究现状。讨论了目前中文命名实体识别的困难和挑战，介绍了本文的工作以及论文的内容安排。

第 2 章：相关技术介绍。首先介绍了对命名实体识别问题的建模方法，之后对于本文涉及的相关技术一一介绍，包括条件随机场模型，词向量技术、深度神经网络技术、概率隐语义分析技术。

第 3 章：改进的中文字符级特征表示方法。提出了两种中文字符级分布式的改进，即面向位置信息的字符向量优化方法和面向主题信息的字符向量构造方法。结合这两个优化方法提出了一种改进的中文字符级特征表示方法。实验结果表明这种方法比直接嵌入字符向量有更好的表现。

第 4 章: 面向复杂命名实体识别的层次深度神经网络模型。针对复杂中文命名实体，提出了一种多层次深度神经网络模型，该模型上层网络提取粗粒度的实体，下层网络提取准确实体，实验表明该模型对于复杂型命名实体有较好的表现。

第 5 章：中文复杂命名实体识别在企业风险识别中的应用。

第 6 章：总结与展望。对本工作做出总结，指出目前工作不完善的方面，提出之后改进的方向。

第二章　相关技术介绍

2.1　引言

本文的工作是以目前表现成熟效果优良的命名实体识别模型作为基础，并在特征分布式表示和网络结构方面进行优化。本章将会介绍本文工作设计的相关理论与技术。

本章下面的章节结构如下：2.2 节介绍命名实体识别问题如何建模成序列标注模型，介绍标签体系；2.3 节介绍条件随机场模型；2.4 节介绍词向量技术；2.5 节介绍深度神经网络模型，包括卷积神经网络和循环神经网络；2.6 节介绍概率隐语义分析技术；2.7 节是本章小结。

2.2　命名实体识别问题建模

命名实体识别是自然语言处理中的一项很基础的任务，是指从文本中识别出特定命名只想的词，比如人名、地名和组织机构名等。目前最常用，最成功的建模方法是将这一问题建模成序列标注问题。即对于输入序列 $X=(x_1,x_2,x_3,\ldots,x_n)$, 给出对应标签序列 $Y=(y_1,y_2,y_3,\ldots,y_n)$。标签体系是两类标签的组合，一类标签是命名实体所属的类别，最常用的有人名 PER，地名 LOC，组织名 ORG，一类是该词在命名实体的位置信息。常用的是 BIO 标注体系和 BIOE 标注体系。BIO 标注体系即将标签分为非命名实体（O），命名实体开头（B），命名实体内部（I）三类，而 BIOE 标注体系中多了一种标注类型，即命名实体结尾（E）。命名实体的标签就是将类别标签和位置标签组合，一个典型的识别人名、地名、组织名的 BIO 标签体系共有 7 个类别标签，如图 2-1s 所示。

输入一个字符串：金正恩与特朗普将在越南首都河内会晤。

以字为单位的标注结果应该是：

金 \PER-B 正 \PER-I 恩 \PER-I 与 \O 特 \PER-B 朗 \PER-I 普 \PER-I 将 \O 在 \O 越 \LOC-B 南 \LOC-I 首 \O 都 \O 河 \LOC-B 内 \LOC-I 会 \O 晤 \O 。

表 2-1: BIO 命名实体标签体系

标签类别	标签说明
PER-B	人名开头
PER-I	人名内部
LOC-B	地名开头
LOC-I	地名内部
ORG-B	组织机构名开头
ORG-I	组织机构名内部
O	非命名实体

也有的方法是以分词后的词语作为标注单位，采用这种方法的标注结果应该是：

金正恩 \PER-B 与 \O 特朗普 \PER-B 将 \O 在 \O 越南 \LOC-B 首都 \O 河内 \LOC-B 会晤 \O 。

也很容易看出分词的效果将很大的影响该种方法命名实体识别的效果。一个显而易见的例子是“南京市长江大桥”，被正确分词为“南京市 \ 长江 \ 大桥”就很容易正确的识别出字符串中的命名实体，而如果分词结果是“南京 \ 市长 \ 江 \ 大桥”就不可能正确的识别出长江大桥这样的命名实体。因而对于中文而言，命名实体识别任务更常用、更具研究价值和潜力的方法应该是以字为单位的标注方案。

目前较成熟较先进的命名实体识别模型是面向字符级的分布式向量表示，再经由深度神经网络模型训练，提取特征，最终通过条件随机场模型预测每个字符所属类别。这种方法也是本文工作的基础方法，接下来将对这些涉及到的模型方法和本文用到的其他技术方法进行简要的介绍。

2.3 条件随机场

条件随机场模型解决给定一组输入的随机变量的情况下，另一组输出随机变量的条件分布模型，该模型的前提假设是随机变量构成马尔科夫随机场。线性链条件随机场由 Lafferty 等人 [13] 与 2011 年提出，线性链条件随机场模型是解决序列标注问题的最经典的方法。具体来说就是若 $\mathrm{X} = \{x_1, x_2, \cdots x_n\}$ 为观测序列，$\mathrm{Y} = \{y_1, y_2, \quad \cdots y_n\}$ 为与观测序列一一对应的标注序列，而条件随机场模

型就是要构建两者之间的条件概率 P(X|Y)。

条件随机场属于概率图模型中的概率无向图模型，即有联合概率分布 P(Y)，由无向图 $G = (V, E)$ 表示，其中图 G 的节点集合 V 表示 Y 的一系列随机变量，而边的集合 E 表示随机变量之间的依赖关系。如果联合概率分布 P(Y) 满足成对、局部或全局马尔可夫性（这三者等价，即对于图中每一个随机变量而言，在给定图中点与其不相邻的随机变量的条件下，和于其相邻的随机变量相互条件独立），则称这个联合概率为概率无向图模型，即条件随机场。概率无向图模型最大的特性就是联合概率便于因子分解，通过最大团上的势函数可以方便的将联合概率分解为概率相乘的形式，便于概率的计算。

线性链条件随机场是条件随机场在链式结构上的表示。假设 X 和 Y 有相同的结构并构成如下图所示的线性链结构，设 $X = (X_1, X_2, \cdots, X_n)$，$Y = (Y_1, Y_2, \cdots, Y_n)$ 均为线性链标识的随机变量序列，若在给定随机变量序列 X 的条件下，随机变量序列 Y 的条件概率分布 $P(Y|X)$ 满足马尔可夫性，$P(Y_i|X, Y_1, \cdots, Y_{i-1}, Y_{i+1}, \cdots, Y_n) = P(Y_i|X, Y_{i-1}, Y_{i+1})$，$i = 1, 2, \cdots, n$，则称 $P(Y|X)$ 为线性条件随机场。该模型是解决序列标注问题的经典模型，因为该模型充分考虑了 X_i 对应的标签 Y_i 与前后文标注的关系。

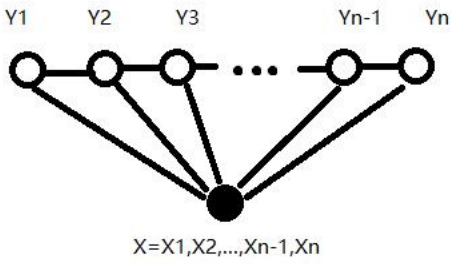

图 2-1: 线性链条件随机场

条件随机场实际上是定义在时序数据上的对数线性模型，具有表达长距离依赖性和交叠性特征的能力，能够较好地解决标注偏置等问题。

2.4 词向量

分布式词嵌入是自然语言处理中一组语言建和特征学习技术的总称。将词汇表中的单词或短语映射成预定好维数的实数向。从数学空间映射的角度来看，这个过程是将一个维数等于不同词语数量的空间映射到一个连续的低维向量空间。生成这样映射的方法有神经网络、基于词共现的维数约减等方法。将

词语等文本信息通过这样的分布式表现形式作为嵌入，这样的方法已经被证明可以提高自然语言处理任务的效果。

将文本表达成计算机可以理解的形式是自然语言处理任务的第一步。最早最朴素的表示方法是独热表示法（One-hot Representation），即用向量的每一维表示词库中的一个词。例如：

“中国”表示为 [1，0，0，0，……，0，0]

“美国”表示为 [0，1，0，0，……，0，0]

显然向量的维数是词语的总数，这样的表示方法简单易于理解，但是浪费极大的空间，而且并不能表现出词语与词语之间的关联。将单词表示为较低维度的向量的技术起源于 20 世纪 60 年代信息检索向量空间模型的发展。使用奇异值分解减少维度的数量，然后在 20 世纪 80 年代后起引入了潜在语义分析（LSA），2000 年 Bengio 等 [38] 在一些列论文中提出了神经概率语言模型，通过对学习单词的分布式表示来降低上下文中单词表示的高维性。该领域在 2010 年后逐渐发展真正成为一种热门的方法，一个重要的原因是在那时向量训练的质量和速度方面取得了重要的进展。许多研究组开始在单词嵌入上投入更多经历。2013 年，由 Tomas Mikolov 领导的谷歌团队创造了 Word2Vec 这样一个单词嵌入工具包，相较于之前的方法，该模型可以更快地训练出向量空间模型。现今大多说新的单词嵌入技术都是基于神经网络架构，而不是传统的 n-gram 模型和无监督学习。

训练 word2vec 有两种经典的模式 CBOW 和 Skip-gram 模型。

（1）CBOW

CBOW 模型训练的方法是通当前词语的上下文词语来预测改词的向量。因而 CBOW（Continuous Bag-of-Words）的输入是当前词上下文的词向量，输出就是当前词的词向量。比如下面这句话，“国家主席/习近平/在/进博会/上/宣布/设立/科创板/，并/于/板块/内/进行/注册制/试点/。”我们上下文窗口取值为 6 的话，特定词为“科创板”，即我们需要求出“科创板”的词向量，前后各有 6 个词共 12 个，这 12 个词是 CBOW 模型的输入，在最基本的 CBOW 模型中，采用的是词袋模式，即这 12 个词的权重一致，并不考虑每个词和目标词的距离。

CBOW 模型的网络结构如图所示，在 CBOW 模型中输入的是 12 个词向量，输出的是所有词的 softmax 概率，损失函数是期望 u 内联网本特定词对应的 softmax 概率最大。对应的 CBOW 模型输入层是 12 个神经元，输出层的个

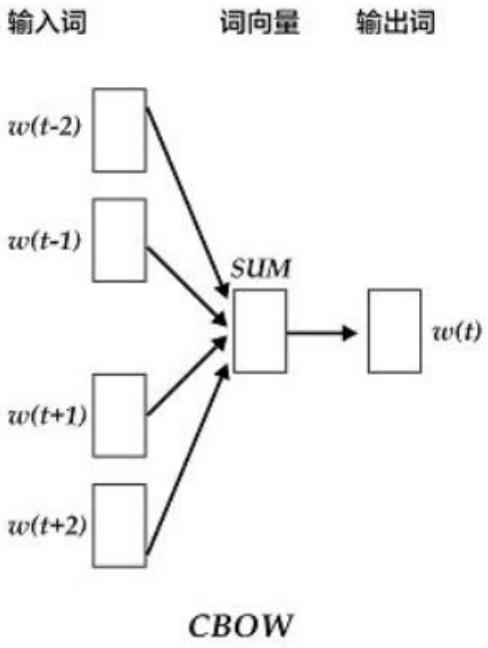

图 2-2: CBOW 模型

数和词汇表的总大小一致，有词汇表大小个神经元。隐藏层的神经元个数可以自行设定，对于神经网络的求解通过经典的反向传播算法求解，迭代完所有语料便可以求解出所有词汇表中的词向量。

（2）Skip-gram

Skip-gram 模型与 CBOW 模型相反，Skip-gram 是输入特定的一个词向量，反而输出的上下文的词向量，例如上文的例子的话输出的就是上下文 12 个词的词向量，同样用反向传播算法，求出概率排前 12 的 softmax 该列车对应的神经元对应的词即为所求。

word2vec 有许多重要的参数直接决定了训练的效果。上下文窗口决定了给订单词前后包含多少个单词作为上下文单词，在原始的 word2vec 模型中窗口内各个单词的权重一直，也有一些研究加权词向量的训练方法 [39] 来提高词向量的质量。维度也是影响词向量的因素之一，研究表明嵌入词的质量随着维数的增加而提高，但达到某一点后，边际效益将减少 [40]。在 word2vec 模型实际运作中，为了解决词汇表太大，训练时间太长的问题，哈会使用霍夫曼树等方法优化训练过程，节省训练的时间。半采样也是实际 word2vec 训练中常用的优化方法，因为高频词通常提供的信息很少，频率高于设定阈值的单词会被降采样来提高训练的速度和质量。

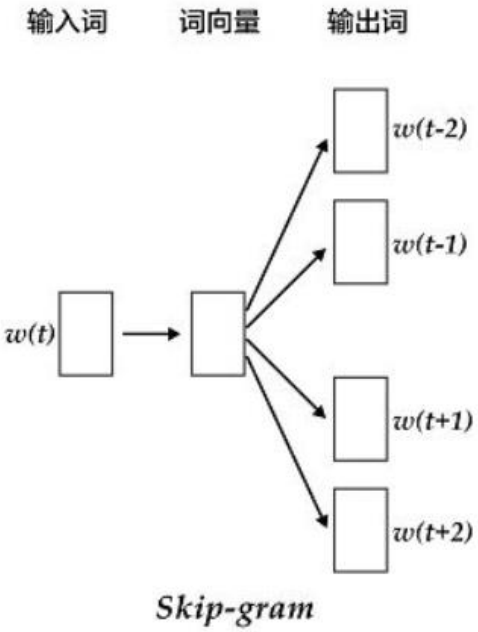

图 2-3: Skip-gram 模型

2.5 循环神经网络

2.5.1 循环神经网络概述

深度学习是基于学习数据表示的更广泛机器学习方法的一种，深度神经网络是深度学习中最重要，应用最广的模型。深度神经网络体系结构已应用与计算机视觉、语音识别、自然语言处理、音频识别、机器翻译、生物信息学、药物设计、医学图像分析等领域。递归神经网络是一类人工神经网络，这类网络善于预测未来。常用于分析时间序列数据，例如股票价格；在自动驾驶中可以预测汽车的行驶轨迹，避免事故。总的来说这样的网络可以输入文本、句子、语音等，使用到机器翻译、语音识别、语义分析等任务中。

不同于前馈神经网络，循环神经网络激活方向不仅仅只向一个方向流动。最简单的 RNN，只由一个神经元接受输入产生输出，然后将输出发回自身，如图 2-4 所示。每个时间步骤 t，每个神经都从上个时间步骤 $y_{(t-1)}$ 接收输入向量 x 和输出向量，一层 RNN 神经元如图所示。

由于在时间步骤 t 时刻循环回去的信息是上一个时刻的信息，这种网络结构使得输入可以是以前的时间步骤，因而在某种意义上 RNN 是一种具有“记忆”的神经网络结构。单个盛景园在时间步骤间保持某种状态，成为记忆单

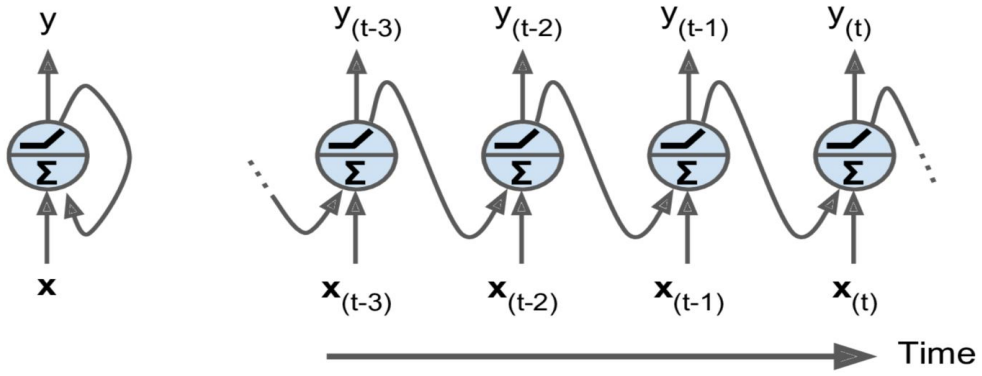

图 2-4: 简单的 RNN 模型

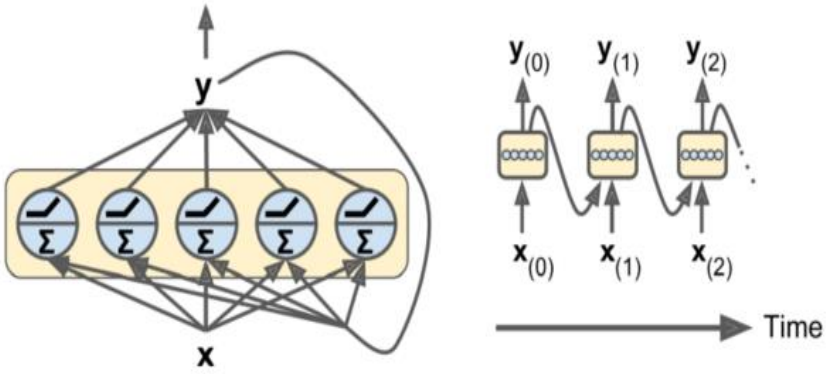

图 2-5: 一个 RNN 层和展开结构

元。单个循环神经元或是一层循环神经元是一个基本结构，有更多复杂的网络结构是以这样的基本结构堆叠出来的。RNN 可以同时接受一些列输入并同时产生一系列输出，如图所示。这样的类型的网络在处理序列数据上有着非常好的表现，可以用于预测时间序列等任务。

2.5.2 长短期记忆网络

长短期记忆网络（LSTM）细胞是由 Sepp Hochreit 和 Jurgen Schmidhuber 于 1997 年提出的 [41]，并在多个领域任务上创造了记录。在语音识别技术、连续手写识别领域都曾创造当时的最佳纪录。2014 年后，在深度神经网络模型中收到了广泛的应用，如果将 LSTM 单元视为一个黑盒，那么它可以非常像一个基本单元，并且性能会更好，训练收敛也更加容易，并且这种结构可以检测数据中的中长期依赖性。如果不看 LSTM 单元内的结构的话，LSTM 单元看起来和一个普通单元一样，知识它的状态被分成两个向量:$h_{(t)}$ 和 $c_{(t)}$，可以将 $h_{(t)}$ 看做短期状态，$c_{(t)}$ 看作长期状态。LSTM 单元的结构如图所示。

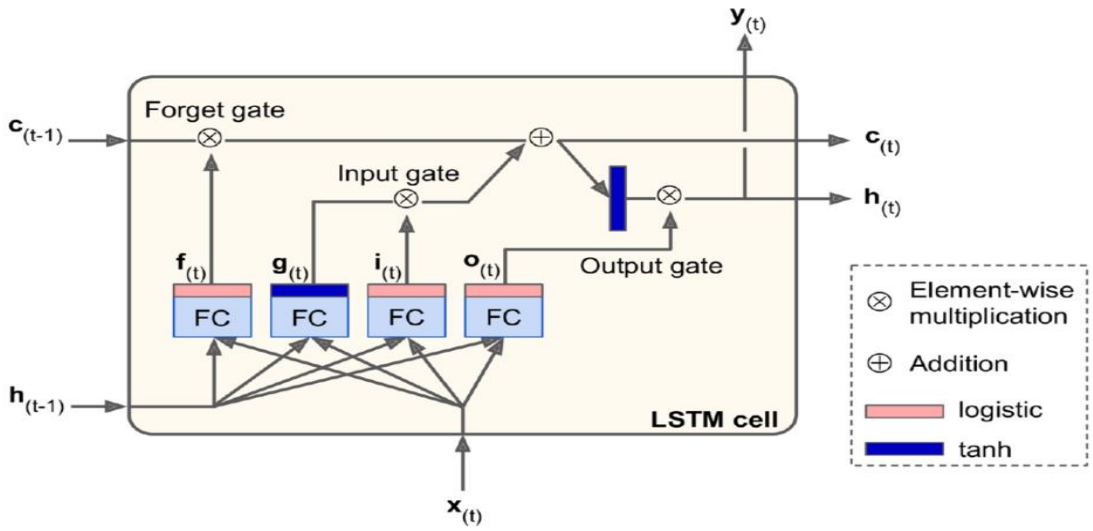

图 2-6: LSTM 单元

长短时记忆网络的思路很简单，相较原始 RNN 结构比，原始 RNN 隐藏层只有一个状态 h，这个状态对短期的输入敏感对长期的输入不敏感。因而 LSTM 单元中，增加一个状态 c，用这个状态来保存长期状态。LSTM 有能力向单元状态中添加或是一处信息，通过结构来管理，这种结构称为门限。LSTM 关键思想是让网络可以学习长期状态下选择性地存储的内容、丢弃内容以及从中读取内容。当长期状态 $c_{(t)}$ 从左向右穿越网络时，可以看到数据流首先通过一个起始门，丢弃一些记忆，然后通过加法运算（添加由输入门选择的记忆）添加一些新的内存，最终 $c_{(t)}$ 不做更多的转换被直接发送出去。因而在这个模型中，在每一个时间步中，一些记忆被丢弃，一些记忆被添加进来。除此以外，$h_{(t)}$ 在执行了第二步操作后，对长期状态进行复制并通过 tanh 函数传递，然后通过输出门对结果进行过滤。这就产生了短期状态（相当于在这一个单元步骤中的输出）。

简要而言，一个 LSTM 单元在输入门，可以学习去识别一个重要的输入，并将输入存储在长期状态中。只要没有被遗忘门作用的话，这个状态会一直被保存，直到需要的时候抽取走这个状态。这就解释了为何 LSTM 模型可以成功的原因，LSTM 模型的特点就是捕获长期模式，因而在长时间录音、文本、音频等任务上有很好的发挥。

2.6 概率隐语义分析技术

2.7 本章小结

第三章　改进的中文字符级特征表示方法

3.1　引言

3.2　相关理论与工作

3.3　面向位置信息的字符向量优化方法

3.4　面向主题信息的字符向量构造方法

3.5　结合位置信息与主题信息的中文字符级特征表示方法

3.6　本章小结

第四章　面向复杂命名实体识别的层次深度神经网络模型

4.1　引言

4.2　复杂命名实体概述与相关工作

4.3　层次标签与层次深度神经网络模型构建

4.4　实验结果及分析

4.5　本章小结

第五章　中文复杂命名实体识别在企业风险识别中的应用

5.1　引言

5.2　应用背景

5.3　数据爬虫

5.4　文本分类模块

5.5　命名实体识别模块

5.6　企业风险识别

第六章　总结与展望

6.1　工作总结

6.2　不足与展望

致　　谢

时光荏苒，在南京大学的本科的学习生活即将结束，四年的时间转瞬即逝，这几年的经历必将成为我人生宝贵的财富。在此论文完成之际，谨向这几年来帮助我的老师和同学表达最衷心的感

参考文献

[1] ANON. A study on the approaches of developing a named entity recognition tool[J], .

[2] ANON. A study on the approaches of developing a named entity recognition tool[J]. IEEE Conference on Artificial Intelligence Application, 1991.

[3] KIM, WOODLAND. A rule-based named entity recognition system for speech input[J]. ICSLP, 2000.

[4] HANISCH F. Prominer rule-based protein and gene entity recognition[J], .

[5] QUIMBAYA. Named entity recognition over electronic health records through a combined dictionary-based approach[J], .

[6] NADEAU, SEKINE. A survey of named entity recognition and classification[J], .

[7] SETTLES. Biomedical named entity recognition using conditional random fields and rich feature set[J], .

[8] RAVIN, MOTTA. Espotter: Adaptive named entity recognition for web browsing[J], .

[9] EDDY. Hidden marov models[J]. current opinion in structural biology, .

[10] QUINLAN. Induction of decision trees[J]. Machine learning, .

[11] KAPUR. Maximum-entropy models in science and engineering[J], .

[12] HEARST D O P S. Suport vector machines[J], .

[13] McCallum LAFFERTY P. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[J], .

[14] BRIN S. [J], 1998.

[15] HENG J, GRISHMAN. [J], 2006.

[16] MCNAME, MAYFILED. Entity extaction without language-specific resources[J]. 6th conference on Natural language learning, .

[17] ISOZKI, KAZAWA. Efficient support vector classifiers for named entity recognition[J], .

[18] LI, BONTCHEVA C. Svm based learning system for information extraction[J], .

[19] MCCALLUM, LI. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons”[J], .

[20] ANON. 基于 CRF 和规则相结合的地理命名实体识别方法 [J], .

[21] 张素香. 信息抽取中关键技术的研究 [J], .

[22] COLLOBERT. [J], 2011.

[23] Sil NGUYEN D F. Toward mention detection robustness with recurrent neural networks[J], .

[24] HLiu L. YAO Y L X, ANWAR. Biomedical named entity recognition based on deep neutral network[J], .

[25] KURU O A C, YURET. Charner: Character-level named entity recognition[J], .

[26] MA X, HOVY E. End-to-end sequence labeling via bidirectional lstm-cnns-crf[J], .

[27] ZHANG J Y, DONG F. Neural reranking for named entity recognition[J], .

[28] M Baesteros G. LAMPLE S. Neural architectures for named entity recognition[J], .

[29] Z. HUANG W X, YU K. Bidirectional lstm-crf models for sequence tagging[J], .

[30] P Verga STRUBELL D B. Fast and accurate entity recognition with iterated dilated convolutions[J], .

[31] JANSSON, LIU. Distributed represent, lda topic modelling and deep learning for emerging named entity recognition from social media[J], .

[32] M Jiang Y. WU J L, XU H. Named entity recognition in Chinese clinical text using deep neural network[J], .

[33] S Zheng P. ZHOU J X Z Q H B, XU B. Joint extraction of multiple relations and entities by using a hybrid neural network[J], .

[34] KATIYAR A, CARDIE C. Nested named entity recognition revisited[J], .

[35] M. JU M M. A neural layered model for nested named entity recognition[J], .

[36] H Yun Y. SHEN Z C L Y K, ANADKUMAR A. Deep active learning for named entity recognition[J], .

[37] A. AKBIK D B, VOLLGRAF R. Contextual string embeddings for sequence labeling[J], .

[38] YOSHUA BENGIO R D, VINCENT P. A Neural Probabilistic Language Model[J]. Département d'informatique et recherche opérationnelle, Université de Montréal, number 1178, 2000.

[39] 马远浩，曾卫明，石玉虎，徐鹏. 基于加权词向量和 LSTM-CNN 的微博文本分类研究 [J], .

[40] MIKOLOV T. Efficient Estimation of Word Representations in Vector Space[J], .

[41] Sepp; Schmidhuber HOCHREITER J. Long Short-Term Memory[J]. Neural Computation, .

附录 A　MPTCP 内核源代码修改

A.1　函数 mptcp_v4_subflows()

```
static void mptcp_v4_subflows(struct sock *meta_sk, const struct mptcp_loc4
    *loc, struct mptcp_rem4 *rem)
{
    int i;
    int num;
    printk(KERN_INFO "******** Entering mptcp_v4_subflows ********\n");

    initial_my_global_var();
    switch(my_counter)
    {
         case 1 : num = Fir; break;
         case 2 : num = Sec; break;
         case 3 : num = Thi; break;
         default : num = Fir;
     }

     for (i = 1; i < num; i++)
     {
         printk(KERN_INFO "******** in mptcp_v4_subflows i = %d num = %d
             ********\n",i,num);
         mptcp_init4_subsockets(meta_sk, loc, rem);
     }
     printk(KERN_INFO "******** Leaving mptcp_v4_subflows ********\n");
}
```

简历与科研成果

基本信息

韦小宝，男，汉族，1985 年 11 月出生，江苏省扬州人。

教育背景

2007 年 9 月 — 2010 年 6 月 南京大学计算机科学与技术系 硕士
2003 年 9 月 — 2007 年 6 月 南京大学计算机科学与技术系 本科

攻读硕士学位期间完成的学术成果

1. Xiaobao Wei, Jinnan Chen, “Voting-on-Grid Clustering for Secure Localization in Wireless Sensor Networks,” in Proc. IEEE International Conference on Communications (ICC) 2010, May. 2010.
2. Xiaobao Wei, Shiba Mao, Jinnan Chen, “Protecting Source Location Privacy in Wireless Sensor Networks with Data Aggregation,” in Proc. 6th International Conference on Ubiquitous Intelligence and Computing (UIC) 2009, Oct. 2009.

攻读硕士学位期间参与的科研课题

1. 国家自然科学基金面上项目 “无线传感器网络在知识获取过程中的若干安全问题研究”（课题年限 2010 年 1 月 — 2012 年 12 月），负责位置相关安全问题的研究。
2. 江苏省知识创新工程重要方向项目下属课题 “下一代移动通信安全机制研究”（课题年限 2010 年 1 月 — 2010 年 12 月），负责 LTE/SAE 认证相关的安全问题研究。