南京大學

研 究 生 毕 业 论 文

（申 请 硕 士 学 位）

论　文　题　目　复杂中文命名实体识别研究

作　者　姓　名　顾溢

学科、专业方向　计算机技术

研　究　方　向　数据挖掘

指　导　教　师　王崇骏教授

2019 年 **4** 月 **15** 日

Research of Complex Chinese Named Entity Recognition Based on Deep Learning

by

Gu Yi

Supervised by

Prof. Wang Chongjun

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

Department of Computer Science and Technology
Nanjing University

May 20th, 2019

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：复杂中文命名实体识别研究

计算机技术 专业 2016 级硕士生姓名：顾溢

指导教师（姓名、职称）：王崇骏教授

摘　要

近年来互联网社交媒体经历了飞速地发展，在这个过程中产生了海量的自然语言数据，这些数据中蕴藏着巨大的价值。命名实体识别技术研究如何从文本中识别具有特定意义的实体，是自然语言处理中最为基础重要的一环，为信息提取、问答系统、句法分析、机器翻译等应用服务，具有相当大的应用价值。但是目前中文命名实体识别方法还存在着一些不足：（1）目前流行的深度学习方法多以字符为粒度，而中文字符向量空间稀疏、信息量不足、含义混淆，制约着整个命名实体识别模型的性能。（2）嵌套命名实体、长文本、上下文错误关联等复杂命名实体问题影响着模型在应用中的效果，尤其是在当下网络文本数据的实际工程应用中，互联网数据规范性差、结构混淆，给命名实体识别工作带来了很大挑战。

本文针对这些问题进行了一系列研究，主要工作包括以下几个方面：

1）扼要综述了命名实体识别技术的发展历程，分析比较了常见模型的优劣，讨论了“字符向量分布式嵌入 -双向长短时记忆网络 -条件随机场”模型对于解决中文命名实体识别问题相比其他模型的优势，并对该模型加以实现验证。

2）在空间表示优化和全局信息表示的基础上，提出了一种改进的中文字符级特征表示方法。在空间表示优化方面，本文利用了中文构词造句的特征，基于位置信息对中文字符向量的表示进行优化；在信息量扩充方面，本文利用主题模型框架构造出中文字符主题概率向量作为辅助特征，补充特征表示中的全局信息。实验表明，本文改进的中文字符级特征表示方法相比于基础的 word2vec 训练出的字向量，提高了整体命名实体识别模型的效果。

3）分析了实际应用场景下影响中文命名实体识别效果的因素，提出了一个层叠的深度神经网络命名实体识别模型。该模型的低层网络通过改进损失函

数和解码方法，对粗粒度的命名实体进行识别，尽量不遗漏潜在的命名实体信息；高层网络接受被低层网络分割后的文本，通过加入卷积层、池化层的方法对命名实体边界精确识别，将最终结果作为整个命名实体识别模型的输出。实验表明该模型在标准数据集和真实网络文本数据下都取得了较好的效果。

关键词： 命名实体识别　双向长短时记忆网络　主题模型　层叠模型

南京大学研究生毕业论文英文摘要首页用纸

THESIS: Research of Complex Chinese Named Entity Recognition Based on Deep Learning

SPECIALIZATION: Computer Technology

POSTGRADUATE: Gu Yi

MENTOR: Prof. Wang Chongjun

Abstract

The rapid development of Internet social media has produced a huge amount of natural language data, which contains great value. Named entity recognition technology, as the most fundamental part of natural language processing, can excavate key entities from text data, and has considerable application value. However, some deficiencies remain in the current Chinese named entity recognition technology. Firstly, in the deep neural network model based on character-level embedding, which is popular today, sparse Chinese character vector space and insufficient information restrict the performance of the whole Chinese named entity recognition model. Secondly, some complex named entit problems such as nested named entities, long texts, and contextual errors negatively affect the performance of the model in application. Especially when we apply text data from Internet to practical projects, its poor standardization and confusing structures cause many challenges to named entity recognition. This paper conducts a series of studies on these issues, and the main conclusions include the following aspects.

1) This paper researches into the development course of named entity recognition technology, and the advantages and disadvantages of commonly used models. Then the advantages of BiLSTM-CRF Model based on character-level embedding over other models in solving the problem of Chinese named entity recognition are demonstrated and the model has been implemented.

2) In order to solve the problem of sparse Chinese character vector space and insufficient information in the deep neural network model based on character-level embedding, this paper proposes an improved method of Chinese character-level feature

representation. In terms of spatial representation, this paper uses the features of Chinese word formation to optimize the representation of Chinese character vectors based on position information. In terms of information supplement, this paper uses the topic model framework to construct the probability vector of Chinese character topics as an auxiliary feature to supplement the global information in character-level feature representation. Experiments indicate that, compared to the simple character vectors trained by word2vec, the improved method of Chinese character-level feature representation polishes the performance of the BiLSTM-CRF Model based on character-level embedding.

3) For the complex Chinese named entity problem, this paper analyzes the factors affecting the recognition effect of Chinese named entities in the application scenario, and proposes a cascaded deep neural network named entity recognition model. The low-level network of the model identifies coarse-grained named entities by modification of loss function and decoding method, and tries not to miss potential named entity information; the high-level network accepts the text segmented by the lower-layer network and joins the convolution layer and pooling layer to identify the named entity boundary more exactly. The model uses the final result of high-level network as the final output of the named entity recognition system. Experiments show that the model has achieved good results under both standard data sets and real network text data.

keywords: Named Entity Recognition Long Short Term Memory Network Topic Model Cascade Model

目　　录

图目录

表目录

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，正一步一步地改变着地球上每个人的生活。特别是伴随着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、娱乐等等情境在当下的移动互联时代都有了新的运作形态。随着移动互联技术的成熟和“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 12 月，我国网民规模已达 8.29 亿，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，目标是对海量文本数据处理分析，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。自然语言技术在规则学习、统计学习等方法的探索与研究下日臻成熟。近年来，表示学习、深度神经网络类机器学习方法给自然语言处理技术带来了新的方向与发展，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，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，服务于多种商业应用。

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领域。目标在于将非结构化文本中提及的命名实体抽取出来，包括人名，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。通用命名实体识别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、一字多性，一字多义、简称方法独特等等特殊之处。因而相适应地根据中文语言的特点对通用命名实体识别技术进行优化有着很重要的意义。

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类别的命名实体类型。例如“小明在小雨的陪同下，一起在南京看了场中国国家队的比赛。”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领域，影响命名实体识别工作方法的有以下几个因素 [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。例如几个东亚文明的语言中文、日文、韩文并不像英文那样单词有着天然的空格作为间隔，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、希腊文也已经有了大量研究，整理富集出大量的语料和工具。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该领域一个长远目标。

（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，因而技术方法移植困难。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，比如文本主题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语料知识，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。选择不同的标注方法，可以构造出不同的命名实体识别模型。

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，规则的设计大多是基于特殊的领域知识。Kim[3] 用规则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典来识别生物医学文本中提到的蛋白质相关术语。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度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说是一个序列标注任务。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（Hidden Markov Model，HMM）[9]、决策树模型（Decision Tree，DT）[10]、最大熵模型（Maximum Entropy，ME）[11]、支持向量机模型（Support Vector Machine，SVM）[12]，条件随机场模型（Conditional Random Field，CRF）[13] 等等都在命名实体识别任务上发挥着作用。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最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 模型利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%。何炎祥等 [20] 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，神经网络模型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特征表示，并进行分类和识别。深度学习相比于传统机器学习方法有三个关键优点：（1）相较于线性模型，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，而深度学习方法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，通过这个性质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：1、词语级别的表示，文献 [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，常用的词语向量训练方法有 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语向量表示的生物医学命名实体识别系统，该系统采用总大小 205924 的词典，训练出维数为 600 的词向量交给深度神经网络学习；2、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使用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] 采用了一个卷积神经网络（CNN）模型来提取字母级别词语的表示。Yang 等 [27] 提出在神经网络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；3、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）模型和循环神经网络（RNN）模型。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 等 [30] 采用了迭代空洞卷积神经网络（ID-CNNs）来做命名实体识别，效果好于传统 CNN 方法。RNN 结构，包括它的变体 GRU 和 LSTM，被验证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经典标准被广泛使用。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。这个过程是要将编码器输出的文本表示为输入，最终得出一系列命名实体标签关联到输入序列。主要采用的方法有以下几种，多层感知机和 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当命名实体类别较多时，解码的速度较快 [36]，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，命名实体嵌套，前后文冗长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较为困难。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可能成功，除此以外命名实体中的人名、组织名往往在含义上与分词词库中字词有很大的差别，分词工作执行困难。然而单纯的以字符为单位的话，汉字一字多义，一字多性十分普遍，信息混淆影响分布式表示效果，将字符更好地向量化表示有很多困难。

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，也缺少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）部分复杂中文命名实体构词复杂，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。实际工程应用中使用的互联网文本数据，结构混淆、规范性差，直接进行命名实体识别效果不佳，影响着命名实体识别系统在实际工程中的使用。

针对上述问题，本文提出了对应的解决方案，构建了面向复杂中文命名实体问题的命名实体识别算法框架，并将该框架应用到了企业风险识别的落地应

用中。本文的主要工作如下：

1）对命名实体方法的发展脉络进行梳理，通过文献综述等方法比较了利用深度学习模型解决中文命名实体识别时，各步骤采用不同模型不同方法的优劣。最终阐明了选择“字符向量嵌入 -双向长短时记忆网络 -条件随机场模型”（CharEmbedding-BiLSTM-CRF）作为工作基础的合理性，对该模型详细探究并加以实现。

2）在优化分布式特征表示方面：针对中文字符向量缺少环境信息、存在一字多义现象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字符主题概率向量作为辅助特征，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。

3）在优化网络结构方面：针对复杂中文命名实体长度长、标签含义混淆、上下文错误关联、实体相互嵌套等问题，本文提出了一个层叠深度神经网络模型，低层模型在优化损失函数和解码方法的手段下识别粗粒度的命名实体，尽量不丢失潜在命名实体的信息。高层模型接受低层模型的输入，在网络结构中加入卷积池化层，加强对命名实体边界的识别能力，识别出的命名实体作为结果输出。这样的模型经实验表明提高了中文命名实体识别系统在标准数据集和工程应用中的效果。

1.4 论文的组织结构

第 1 章：绪论。介绍研究的背景及意义，梳理命名实体识别问题的定义和研究子方向。介绍了传统方法和深度学习方法的研究发展和现状，探讨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、主题模型技术等。

第 3 章：改进的中文字符级特征表示方法。针对中文命名实体识别问题，本文提出了中文字符级分布式表示的改进方法。通过基于位置信息对字符向量表示进行优化、基于主题信息构造新的中文字符级分布式表示，结合这两个优化方法对中文字符级表示加以改进。实验结果表明这种方法比直接嵌入 word2vec 训练的字符向量在中文命名实体识别任务中有更好的表现。

第 4 章: 面向复杂中文命名实体识别的层叠深度神经网络模型。针对复杂中文命名实体，本文提出了一种层叠的深度神经网络模型，该模型低层网络提取粗粒度的实体传入高层网络，高层网络对实体边界精确识别得到模型的输出。实验表明该模型对于标准数据集和复杂命名实体较多的互联网文本数据有不错的表现。

第 5 章：总结与展望。对本文工作做出总结，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

上一章节介绍了命名实体识别问题研究的背景和意义以及前人的相关工作。命名实体识别模型作为自然语言处理技术中关键基础的一环，有着很强的研究价值。

命名实体识别历经数十年的研究发展，不论是建模方法还是求解方法，都有很多相关的研究。本文的工作以目前效果较好的命名实体识别模型作为基础，首先叙述了命名实体识别问题目前比较认可的建模方法，之后介绍一系列与基础模型相关的算法与技术。本文的改进优化工作也涉及到一些经典模型与算法，本章在此一并介绍。

本章的章节结构如下：2.2 节介绍命名实体识别问题如何建模成序列标注模型，介绍命名实体识别的标签体系；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，包括循环神经网络模型及其变体；2.6 节介绍主题模型相关技术；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指向的词，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 $\mathrm{X} = (x_1, x_2, x_3, \ldots, x_n)$, 给出对应标签序列 $\mathrm{Y} = (y_1, y_2, y_3, \ldots, y_n)$。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，最常用的有人名实体（PER），地名实体（LOC），组织名实体（ORG），一类是该词在命名实体的位置信息。位置信息最常用的标准有 BIO 标注体系和 BIOES 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 BIOES 标注体系中多了两种标注类型，即命名实体结尾（E）和单个实体（S）。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 2-1 所示。

表 2-1: BIO 命名实体标签体系

标签类别	标签说明
B-PER	人名实体开头
I-PER	人名实体内部
B-LOC	地名实体开头
I-LOC	地名实体内部
B-ORG	组织名实体开头
I-ORG	组织名实体内部
O	非命名实体

在序列标注建模方法和 BIO 标注体系下，对于中文文本的命名实体识别模型就是要为序列中的每个变量预测出所属的标签类别。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 。\O

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 。\O

从上面的例子也很容易看出分词的效果将很大的影响以词语为单位的命名实体识别方法的效果。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。

本文的工作是针对中文命名实体识别问题，以字符为粒度进行序列学习。BIO 标注体系已经可以很好的满足我们的要求，本文后续有标注的数据集都采用的是 BIO 标注体系。

目前较成熟先进的中文命名实体识别模型是面向字符级的分布式表示，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，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

条件随机场模型用于解决给定一组输入的随机变量的情况下，预测另一组输出随机变量的条件分布，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] 于 2011 年提出，线性链条件随机场模型是目前解决序列标注问题的最为经典的方法。具体来说，若 $\mathrm{X}=\{x_1, x_2, \cdots x_n\}$ 为观测序列，$\mathrm{Y}=\{y_1, y_2, \cdots y_n\}$ 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)，如图 2-1。

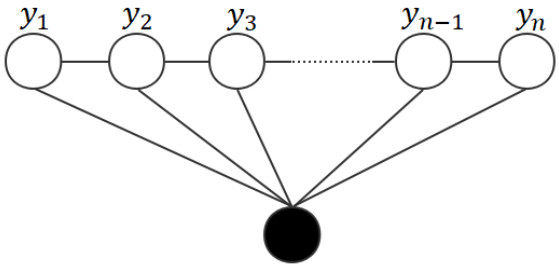

图 2-1: 线性链条件随机场

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地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，$P\left(Y_i|X, Y_1, \cdots, Y_{i-1}, Y_{i+1}, \cdots, Y_n\right)=P\left(Y_i|X, Y_{i-1}, Y_{i+1}\right)$，$i=1, 2, \cdots, n$，也就是概率 $P(Y_i|X)$ 只与序列 X 以及前后两个标签相关。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。

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。很多工作 [7] 都说明该模型对于命名实体识别技术有着很好的帮助。

2.4 词向量

分布式词嵌入是自然语言处理中一组语言建模和特征学习技术的总称。即将词汇表中的单词或短语映射成预定好维数的实数向量。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模型、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特征嵌入，该方法已经被验证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工作。最早也是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，存在着“词语鸿沟”问题。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）方法。2000 年 Bengio 等 [38] 在一系列论文中提出了神经概率语言模型，通过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 年后逐渐真正发展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方法研究上投入更多精力。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 [39]，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。现今绝大多数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 模型和 Skip-gram 模型（见

图2-2）。

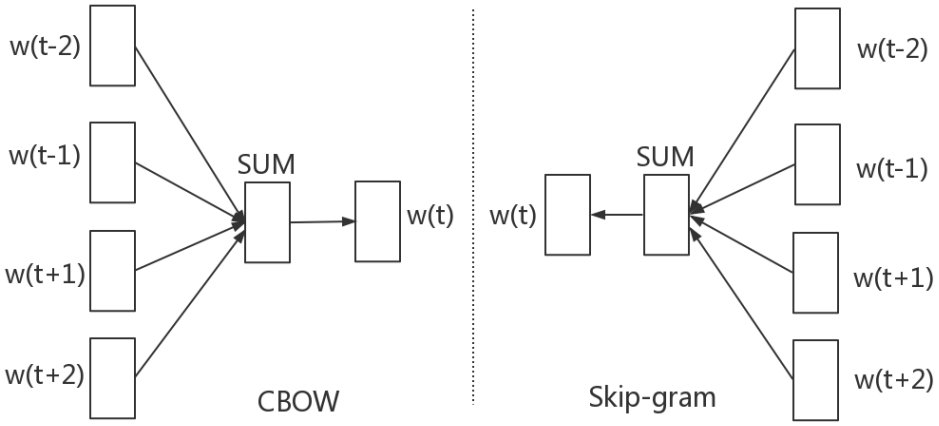

图 2-2: CBOW 模型与 Skip-gram 模型

（1）CBOW 模型

CBOW 模型训练的方法是通过当前词语的上下文词语来预测该词的向量。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/。”当上下文窗口取值为6，特定词为“科创板”时，即需要求出“科创板”的词向量。“科创板”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 模型的网络结构如下图所示，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，损失函数是使期望特定词对应的 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 模型

Skip-gram 模型与 CBOW 模型相反，Skip-gram 是输入特定的一个词向量，反而输出的是上下文的词向量。在上文的例子的中输出的就是上下文 12 个词的词向量，同样用反向传播算法，投影层即为词语到指定维数向量空间的映射。

两种方法比较而言，Skip-gram 方法中邻近上下文单词的权重大于较远上下文单词的权重 [40]，CBOW 方法速度较快，Skip-gram 方法速度较慢，但 Skip-gram 对不常见的单词效果更好，被更多地采用。

word2vec 有许多重要的参数直接决定了训练的效果。上下文窗口大小决定了给定单词前后包含多少个单词作为上下文训练单词，在原始的 word2vec 模型中窗口内各个单词的权重一致，也有一些研究加权词向量的训练方法 [41] 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 [42]。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。

2.5 循环神经网络

2.5.1 循环神经网络概述

深度学习（Deep Learning）是基于学习数据表示的更广泛的一类机器学习方法，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被应用于计算机视觉、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。循环神经网络（Recurrent Neural Network，RNN）是人工神经网络（Artificial Neural Network）中的一类，这类网络善于预测序列信息。常用于分析时间序列数据，例如在金融业中预测股票价格；在自动驾驶中可以预测汽车的行驶轨迹，避免事故等等。总的来说这样的网络可以输入文本、句子、语音等信息，应用到机器翻译、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 2-3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。

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。单个神经元在时间步骤间保持某种状态，成为记忆单

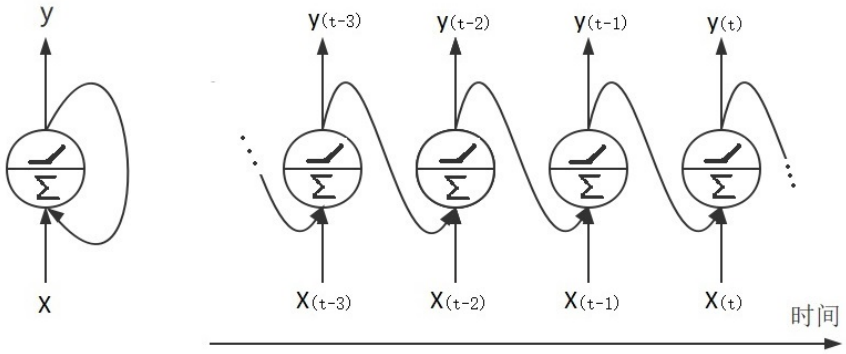

图 2-3: 单个 RNN 单元模型

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 可以同时接受一系列输入并同时产生一系列输出。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。

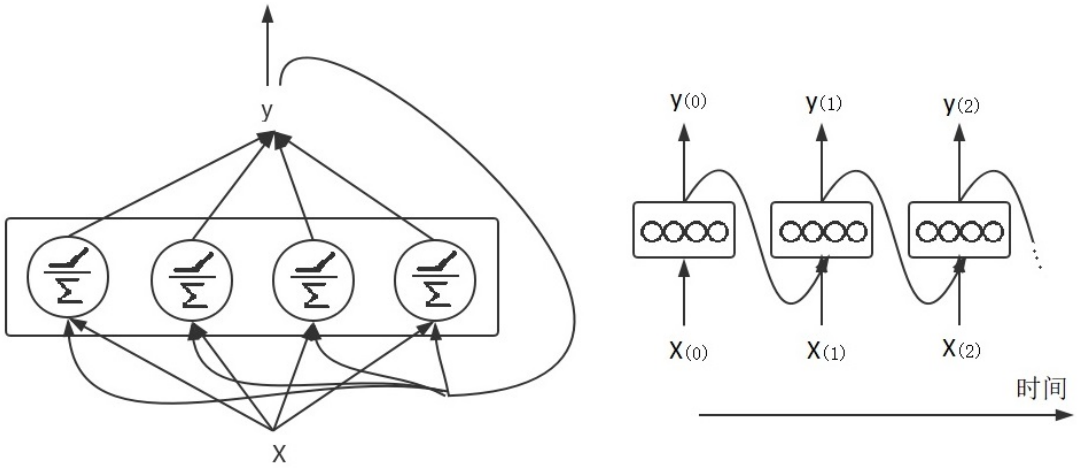

图 2-4: 一层 RNN 单元模型

一层 RNN 模型按时间展开如图 2-4所示，当一个输入序列送入 RNN 网络进行训练时，$X_{(0)}$ 时刻的信息可以被传递到 $X_{(1)}$ 在网络的训练过程，这是前馈神经网络所不具备的性质。也就是说在训练当前网络时，输入序列之前的内容仍然会被考虑到，因而 RNN 模型对于时序序列数据具有很强的处理能力。RNN 的训练方法常采用 BPTT（Back-propagation Through Time）算法，本质还是反向传播算法。RNN 处理的是时序数据，因而基于时间反向传播。

2.5.2 长短期记忆网络

长短时记忆网络（Long Short Term Memory Network，LSTM）是由 Sepp Hochreit 和 Jurgen Schmidhuber[43] 在 1997 年提出，并在多个领域任务上表现优异。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受到了广泛的应用，如果将 LSTM 单元视为一个黑盒，那么它可以非常像一个神经元基本单元，并且整个循环神经网络的性能会更好，训练收敛也更加容易，更重要的是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，只是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2-5。

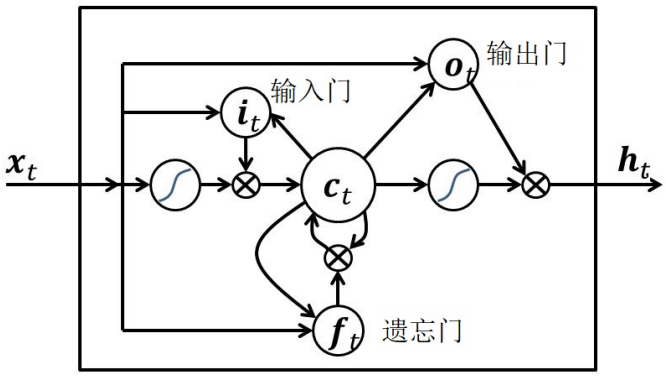

图 2-5: LSTM 单元

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而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丢失信息，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，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 单元在输入门模块，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表现。

LSTM 的训练方法也可以采用梯度下降法来最小化训练误差，一个经典的方法是应用时序性倒传递算法，这种算法依据错误修改每次的权重。用梯度下降法训练循环神经网络（RNN）时会遇到一个严重的问题，误差梯度随时间长度成指数级别增长式得消失。而当设置成 LSTM 单元时，误差可以被重新抽取倒回计算，从输出端回到输入端，再重新经过每一个门限，直到数值被阈值过滤掉。因此含有 LSTM 的单元的 RNN 模型可以被有效训练并记住长时间的信息，并可以通过梯度下降法较好地进行优化训练。

2.6 主题模型技术

2.6.1 潜语义分析

潜语义分析（LSA）是一种对文档主题进行建模的方法，通过产生一组与文档和词语相关的概念来分析文档与其包含的词语之间的关系。LSA 假设意义相近的单词出现在相似的文本片段中，通过语料集构建包含每个段落的词语的矩阵（行代表一个唯一的单词，列代表每个文档或段落），使用奇异值分解（SVD）技术来减少行数，同时可以保持列之间的相似性结构。取任意两行形成的两个向量之间的角度余弦来比较单词。接近 1 的值表示非常相似的单词，而接近 0 的值表示非常不同的单词。用任意两列向量之间角度的余弦来比较文档，接近 1 的值表示非常相似的文档，而接近 0 的值表示非常不同的文档。

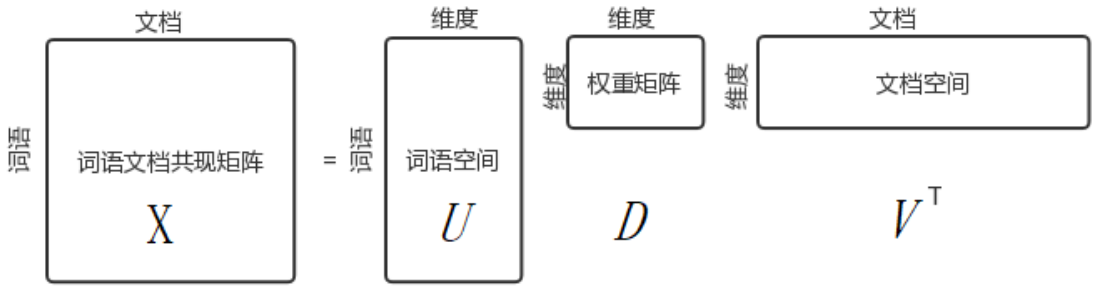

图 2-6: LSA 模型

LSA 模型如图2-6，其中 $X = UDV^T$。矩阵 U 中的每一行表示每个词在每个词类（主题）的值，每一列表示语义相近的词类，每行中的值表示每个词语在这个语义类中的重要程度。LSA 可以刻画同义词，因为词语相近的词对应着相同或相似的主题，SVD 的降维方法也可以去除部分噪声，使得特征鲁棒性更好。但是 LSA 无法处理一词多义的问题，LSA 将每个词映射到了潜语义空间中的一个点，即一个词对应的多个含义没有被区分。并且 LSA 的概率模型，是以文档和词的分布服从联合正态分布为前提假设的，但这个假设往往准确性不高。

2.6.2 概率潜语义分析

概率潜语义分析（PLSA）是根据观察到的变量与某些隐藏变量的相似性从而得出它们的低维表示，这种技术是潜语义分析技术的一个改进版本。与 LSA 是利用线性代数方法进行降维不同的是，PLSA 是基于估计一个概率模型参数达到降维效果。PLSA 模型将文档中的词看作来自混合模型的采样，假设每个词来自一个主题，同一个文档中不同词可能来自不同的主题。再将文档表示为多个主题的混合，每个主题具有不同的概率，每个主题又由多个词构成。PLSA 参数结构图如图2-7。

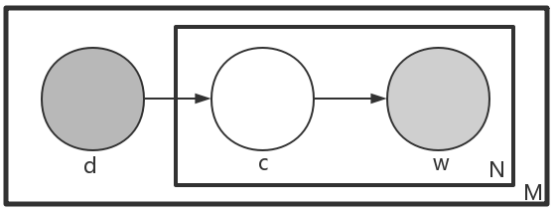

图 2-7: PLSA 模型

基于观察到的文档和单词的共现关系，PLSA 模型将共现的概率建模为条件独立的多项式分布。

$$P(w,d) = \sum_c P(c)P(d|c)P(w|c) = P(d)\sum_c P(c|d)P(w|c)$$

上式中 $P(d)$ 表示从文档集合选择文档 d 的概率，$P(c)$ 表示选择主题 c 的概率，主题个数是 PLSA 模型中的超参数需要提前设定，这里主题 c 是隐变量。在已知文档 $P(d)$ 和其对应词表 w 后，对主题进行推断。即求解

$$p(c|d,w) = \frac{P(c)P(d|c)P(w|c)}{\sum_{c' \in C} P(c')\,P(d|c')P(w|c')}$$

PLSA 对参数的求解通过 EM 算法进行学习。PLSA 比 LSA 可以较好的解决一词多义的问题，相比于 LSA，PLSA 使用多项式分布建模在真实数据上表现来说优于 LSA。然而 PLSA 的训练常常会遇到过拟合的问题，即训练出的数据效果在训练集上很好但是在其他语料上的表现差异较大。

2.6.3 隐狄利克雷分配模型

隐狄利克雷分配模型（LDA）是由 PLSA 的基础上发展而来。LDA 是一个生成式的统计模型，假定每个文档都是少量主题的混合体，并且每个单词的出现都归因于文档中的某个主题。

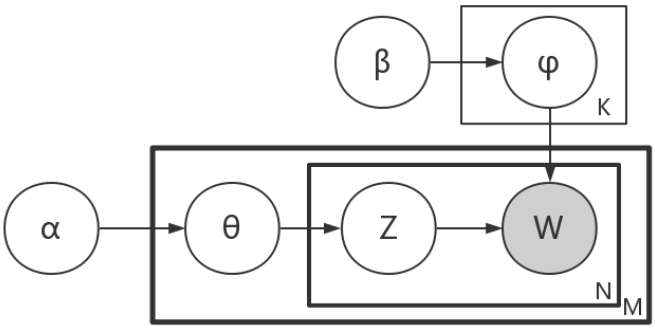

图 2-8: LDA 模型

PLSA 模型和 LDA 模型在建立文档和主题间的概率模型和建立主题与词的概率模型时都采用了多项式分布，为了计算的简便性并且让先验更有意义，LDA 在 PLSA 的基础上做了改进，选取了多项式分布的共轭先验分布狄利克雷分布作为概率分布的选择。如图 2-8，Z 是主题，θ 是决定文档主题的概率分布参数，α 是决定文档对主题的概率分布参数 θ 的超参数。W 是特定的单词，φ 是词语对主题的概率分布的参数，β 是决定词语对主题的概率分布参数 φ 的超参数。形式化来说模型如下：

$$\theta_i \sim \mathrm{Dir}(\alpha), i \in \{1,\ldots,M\}$$

$$\varphi_k \sim \mathrm{Dir}(\beta), k \in \{1,\ldots,K\}$$

$$z_{i,j} \sim \text{ Multinomial } (\theta_i), i \in \{1,\ldots,M\}, j \in \{1,\ldots,N_i\}$$

$$w_{i,j} \sim \text{Multinomial}\left(\varphi_{z_{i,j}}\right), i \in \{1,\ldots,M\}, j \in \{1,\ldots,N_i\}$$

LDA 模型求解是一个较为复杂的最优化问题，精确求解较为困难，一般采用近似求解方法。主要的求解方式有吉布斯采样算法、基于变分的 EM 算法和基于期望推进的方法。最主流的方法是吉布斯采样算法，这种算法基于马尔科夫链的蒙特卡洛方法，具有收敛较快、容易并行化，效果较好等优点，因而被广泛使用。LDA 模型在缺少海量数据的情况下表现明显超出 PLSA 模型，有较强的抗过拟合性，成为一种经典的主题模型方法。

2.7 本章小结

本章介绍了论文研究所涉及的模型与技术。先从命名实体建模方法开始介绍，引出命名实体的标注方法；接着介绍了条件随机场这种解决序列标注的经典模型；后文通过介绍词向量和循环神经网络介绍当下较为先进的命名实体识别问题的关键技术；最后介绍了本文工作中使用的主题模型方法。

第三章　改进的中文字符级特征表示方法

3.1　引言

命名实体识别任务是自然语言处理技术中的重要任务之一。命名实体指一个词语或者一个短语明确指称一个标识的实体，一般有人名、机构名、地名等。命名实体识别指将文本中的命名实体定位并分类为预定义的实体类别的过程。目前解决命名实体识别较成熟的方法是将命名实体问题建模成序列标注问题，首先将词语分布式表示，通过深度神经网络训练，最终通过条件随机场层输出。本文工作的基础模型是通过“字符向量嵌入 -双向长短时记忆网络 -条件随机场”模型进行命名实体识别，对该模型加以改进，并以此为基础解决中文复杂命名实体识别问题。

相比于英文为首的西语命名实体识别，中文命名实体识别要困难的多。因为英文天然存在空格隔开词语，不同词性词义的词语词缀词根也不同。总体而言，英文的词语构词方法与表现形式可以比中文体现出更多的语义信息。英文命名实体识别问题直接训练词向量做为特征的分布式表示就能达到不错的效果，但是中文词素间没有间隔，一字多义，一字多性，直接训练词向量作为特征表示，相较英文来说效果不够好。

本章的工作就是为了优化命名实体识别中中文文本的特征表示问题，针对中文命名实体存在的不足改进中文字符级特征表示方法。本章的主要创新点是：（1）针对字符向量一字多义、一字多性难以区分的问题，提出基于位置信息优化中文字符向量的方法；（2）针对字符向量训练时，受制于 word2vec 模型训练时上下文窗口大小的限制，字符向量信息量不足，缺少全局信息的问题，提出基于主题信息的字符向量构造方法。

本章章节内容安排如下：3.2 节介绍命名实体中特征分布式特征表示的相关理论与工作；3.3 节介绍本章工作的基础模型“字符分布式嵌入 -双向长短时记忆网络 -条件随机场”中文命名实体识别模型，并加以实现；3.4 节介绍本文提

出的基于位置信息的字符向量优化方法；3.5 介绍文本提出的基于主题信息的字符向量构造方法；3.6 节对本章的各个模型算法进行实验验证及实验结果分析；3.7 节是本章小结。

3.2 相关理论与工作

深度神经网络近年来在多个领域取得了进展和突破，在命名实体领域也有不错的表现。深度神经网络一个很大优点在于该模型可以在训练过程中自动寻找数据特征表示的模式。深度神经网络模型处理命名实体识别问题的第一个环节就是要将文本的特征进行分布式表示，以便深度神经网络模型更好地接收这些信息，不同的分布式表示方法对最终的模型效果有很大的影响。

在英文文本特征表示方法上，近年有许多启发性的研究成果。Strubell[30] 提出一种基于迭代扩张卷积神经网络（ID-CNN）的标记方案，使用 skip-ngram 模型，对 SENNA 语料库进行 100 维嵌入式训练。Li[44] 不仅仅局限在单词的维度，而是将字符拆解为词缀词根进行分布式表示。Kuru[25] 提出一个字符级的命名实体识别模型，并给出一个字符级的标注标签，该模型将一个句子看作是一个字母的序列，之后使用长短时记忆网络（LSTM）抽取字符级别的特征表示，以此来识别命名实体。混合单词信息、字符信息、词缀词根信息、单词主题信息的模型 [30, 31, 45] 在命名实体识别模型上也有一定的功效。近期还有 Devlin 等 [46] 提出一个新的语言表示模型 BERT，BERT 基于所有层中的左、右语境进行联合调整，来预训练深层双向表征，在很多任务上达到了很好的效果。

在中文文本特征表示方法上，Yue 等 [47] 等提出了一个词格模型来解决中文命名实体识别问题，不仅使用每个汉字字符的信息，再加上所有可能构成的所有单词的信息，这样的模型有效地减少了分词错误，但是效果对词典构造有一定依赖。Zhao 等 [48] 实现了一个高速 LSTM-CRF 模型，在高速层自动选择与当前字符更相关的字符，达到了与注意力机制相似的效果。林泽斐等 [49] 利用上下文信息对应的知识库知识，对于命名实体识别任务中的命名实体进行消歧。王超等 [50] 利用 LSTM 中文分词技术优化分词模块，在中文微博数据的命名实体识别上取得了较好的效果。

3.3 CharEmbedding-BiLSTM-CRF 中文命名实体识别模型

3.3.1 模型比较与优势

经过大量的文献综述和前沿方法总结，可以得出结论，目前主流成熟的深度学习中文命名实体识别方法大致流程是：（1）将字词进行分布式表示（2）使用深度学习网络有监督地训练模型（3）利用上下文信息对序列中的每个字词进行标签标注。本文使用的基准模型是“字符分布式嵌入 -双向长短时记忆网络 -条件随机场”（CharEmbedding-BiLSTM-CRF）命名实体识别模型，选择该基准模型是因为其有着建模合理，效果优良，优化潜力大等优点，理由如下：

1）字符嵌入方面。基于词语的嵌入方式的缺点是，中文词语切割是基于词典的，这样的分词技术一旦出现错误则后续的命名实体任务很难成功，因而基于词语的嵌入方式模型容量较低。对于各种基于分词的中文命名实体对于解决地名识别、组合型组织名（例如“中国国家男子足球队”）效果较好，但是对于人名、企业名等不是由词语构成的命名实体效果不好，因为这类命名实体组成方式和普通构词方式差异较大。因而本文选取直接字符嵌入作为基准模型研究。

2）神经网络方面。传统的前馈神经网络（例如 CNN）在分类任务上略有优势，然而对于信息序列来说，信息间彼此有着复杂的时间关联性，更重要的是对于命名实体识别任务来说信息长度各不相同，前馈神经网络建模困难，表现往往不好。因而对于序列任务反馈神经网络（即循环神经网络）将是优先的选择。而 LSTM 模型是 RNN 的一个变种，在善于对序列问题建模的同时，该模型还有着易于求解，能够长期保存重要信息的优点。而双向长短时记忆网络（BiLSTM）是 LSTM 模型的一个改进版本，传统的 RNN 输入是上文，输出是下文，根据上文推出下文，双向 RNN 同时利用反向信息，让模型从两个方向学习，这个概念也符合中文自然语言的构词遣句的思想。BiLSTM 便是 LSTM 的双向版本，实验表明 BiLSTM 往往比 LSTM 有着更好的表现 [51]，特别是在序列标注问题上，模型可以同时学习过去的序列和未来的序列的信息，有着更好的效果 [29]。因而神经网络模块本文选择 BiLSTM 模型。

3）序列标注模型方面。最常用的方法是条件随机场（CRF）模型或是采用接全连接层用 softmax 函数激活直接分类两种方法。CRF 将输出层面的关联性

分离出来，在预测标签时可以充分考虑上下文关联，更重要的是 CRF 的求解维特比算法是利用动态规划的方法求出概率最大的路径，这与命名实体识别的任务契合的更好，可以避免结果中出现“B-LOC”标签后接“I-ORG”标签这种非法序列的问题。因而本文序列标注上选择 CRF 模型。

3.3.2 模型流程与实现

（1）字向量训练

中文字符向量的训练与中文词向量的训练相似，对于中文语料集首先去除非字符，但是不需要像训练中文词向量那样通过字典匹配来进行分词过程。只需要将字符与字符间用空格隔开表示成独立的一个词素即可。本文训练语料采用的是人民日报 2014 新闻数据集，这与后文命名实体识别使用的语料类型一致都是属于新闻类文本数据。如3-1所示，中文字符向量训练网络由一个输入层、一个隐层、一个输出层构成。如 2.4 节介绍的 word2vec 模型根据需要设定相关网络参数，上下文滑动窗口窗口大小设置为 5，采用 Skip-gram 模式训练字向量。经统计不同字符有 4000 左右，按经验设置隐层神经元个数为 100，即训练 100 维度的字向量。

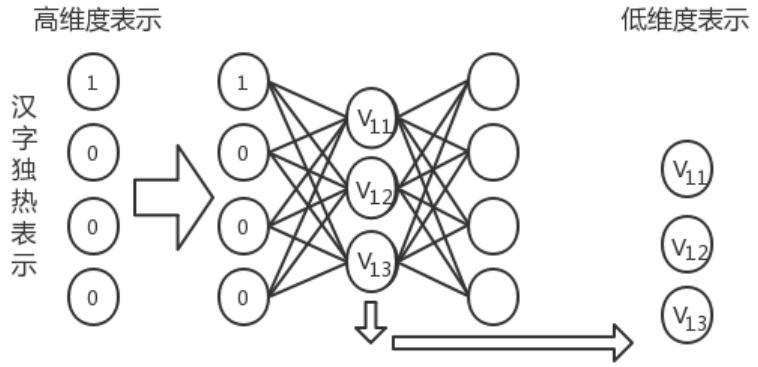

图 3-1: word2vecc 训练中文字符向量示意图

经过多轮迭代训练，训练出中文字符向量。这里举例“全”、“国”、“政”、“协”、“会”、“议”六个字符的字向量，在全部词典中找出与该字向量最相近的字向量，以观测字向量训练的效果，结果如3-1所示。

从表中可以看出，字符与字符间的语义存在一定联系，含义接近的字符向量相似度也更大。从模型训练的角度来解释，与该字符相似度越大，那么越可能出现在该字符滑动窗口内。同时 word2vec 模型会使得同一类语义的词具有相

表 3-1: 字向量训练效果

全	国	教	育	会	议
整 0.5757	华 0.5501	堂 0.6539	培 0.6382	协 0.6194	审 0.5820
排 0.5665	兰 0.5284	宗 0.6387	教 0.6091	届 0.5958	谘 0.5698
障 0.5405	洲 0.4830	徒 0.6241	体 0.5926	参 0.5499	选 0.5566
中 0.5101	暨 0.4752	皈 0.6185	训 0.5667	员 0.5341	协 0.5540
防 0.5080	侨 0.4752	育 0.6091	课 0.5374	暨 0.5175	党 0.5508
第 0.4943	央 0.4733	督 0.6034	养 0.5369	动 0.5121	遴 0.5474
并 0.4930	联 0.4723	仰 0.6017	健 0.5367	团 0.5109	席 0.5432
会 0.4913	盟 0.4606	圣 0.5998	学 0.5275	办 0.5061	宪 0.5429

近的向量，在字向量中这个现象同样有效，选取部分 100 维字向量用主成分分析法（PCA）降维至两个维度，在可视化的二维平面作图 (图 3-2)，可见语义相近的字向量降维后的坐标相互靠近，语义相关弱的字向量距离较远，说明字向量的训练达到了一个较好的效果，同时有着很好的可解释性。

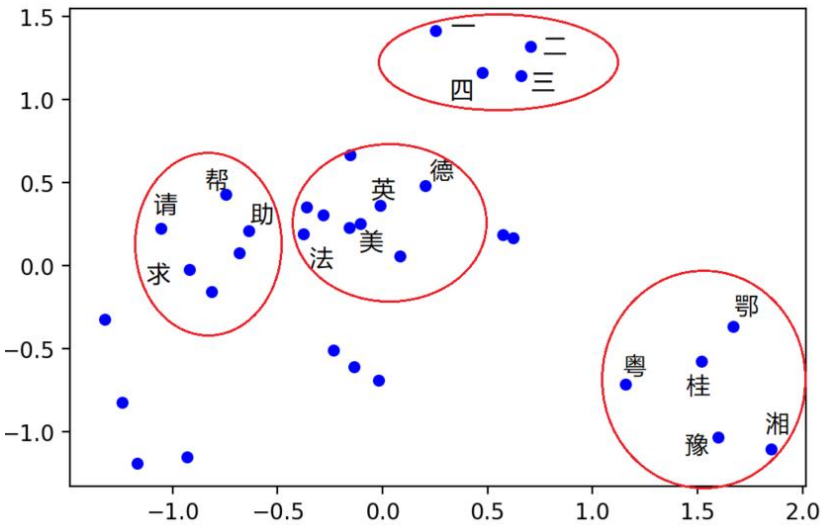

图 3-2: 训练出的字向量降维效果图

word2vec 这样的模型不仅将字符用设定的维度分布式表达，同时也能够保留字符与字符之间的语义联系，有文献 [52] 表明这样分布式表示对后续使用神经网络模型处理各类自然语言问题有着良好的效果。本章后续的实验也验证了这一点，使用 word2vec 模型训练好的字符向量比随机初始化字符向量在中文名实体识别任务中有着更好的表现。

（2）神经网络构建

本文神经网络模型选取的是双向长短时记忆网络（Bidirectional LSTM Networks，BiLSTM）。循环神经网络（RNN）可以在网络训练期间记住历史信息，对于序列问题有着强大的学习能力。传统的 RNN 模型在训练中会遇到梯度爆炸和梯度消失等问题而难以训练。所谓的梯度消失和梯度爆炸问题都是在通过反向传播训练计算时，梯度倾向于在每一时刻递增或者递减，经过一段时间后梯度就会发散到上限或是衰减到零。对于梯度发散的问题一般可以通过设定阈值使得梯度不能超过一个给定值来解决，但是对于梯度消失问题，简单的 RNN 模型就无法解决。在命名实体识别这样的序列标注问题中，梯度消失就表现为网络难以联接到远处的信息能力，从而降低最终模型的识别能力。如图 3-3，命名实体识别问题中，标签可以被较远的信息影响。简单的 RNN 模型对于传递较远的信息在训练时往往受梯度消失的影响而消失。如 2.5.2 节所述，LSTM 单元通过门限结构控制，重要的信息会被一直保存，不重要的信息会被遗忘门丢弃，从而能“记住”长期的信息。LSTM 模型中误差可以被重新抽取倒回计算，也解决了 RNN 网络训练时梯度消失的问题。

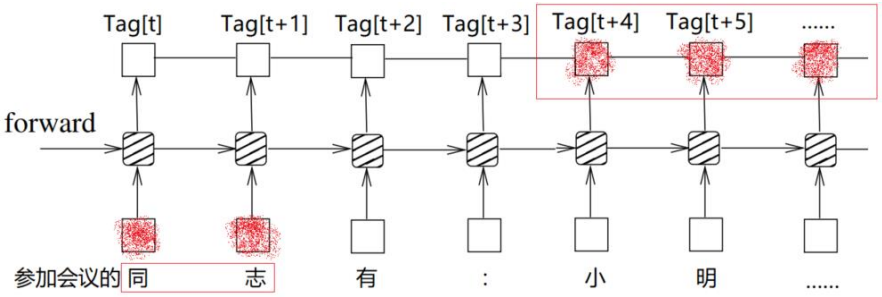

图 3-3: 命名实体识别中的长期信息

除此以外，命名实体问题与前向信息和后向信息都有很大的关联。例如“美国总统特朗普”、“百度（中国）有限公司”、“钱学森院士等参加座谈”，这些语言片段中出现的“总统”、“公司”、“院士”等文字虽然本不属于命名实体，但是对于命名实体的识别相当重要。人类识别出人名、地名、组织名的过程中，往往借助该实体前文的信息和后文的信息。正因如此，对于命名实体识别问题，构建向前和向后两个循环神经网络框架是一个更优的选择。

如图 3-4命名实体识别问题中，前向信息和后向信息对于命名实体的识别都有着相当大的贡献。如果命名实体识别时只有历史信息而没有未来信息的

话，“中国”两个字的标签仅与前文相关，“中国”可能是一个地名实体，也可能是一个组织名实体的开头，存在混淆。而训练时给与网络未来的信息，通过后文“南京”“很美”等信息，就更加容易判断这里“中”“国”的标签更有可能是“B-LOC”“I-LOC”。

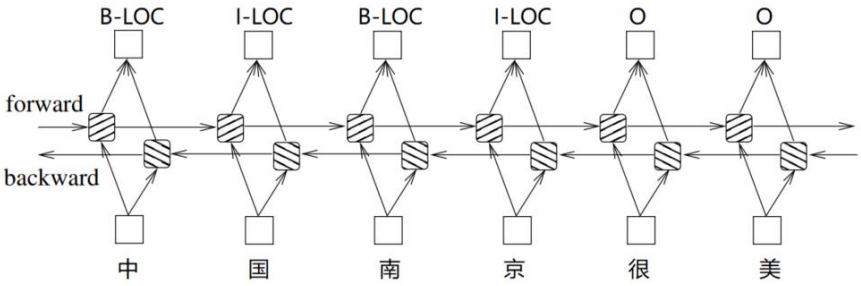

图 3-4: 双向 LSTM 示意图

（3）条件随机场与序列标注

条件随机场模型（Conditional Random Field，CRF）相较于传统分类模型而言，不仅仅关注个体标签的分类，还非常关注句子级别的信息。近些年的工作表明 CRF 模型在序列标注问题上有着很高的正确率，因而成为命名实体识别问题中一个经典的方案。正如 2.3 节所述，CRF 层可以通过训练语料学习得到一些基于全局的约束信息，比如句子中识别出的实体标签的起始应当是“B-”而不是“I-”；不同类的标签不会相互连接，识别出的人名、地名、组织名标签不可能混搭，从而能够识别出准确的命名实体。整体 CharEmbedding-LSTM-CRF 命名实体识别模型框架如图 3-5。

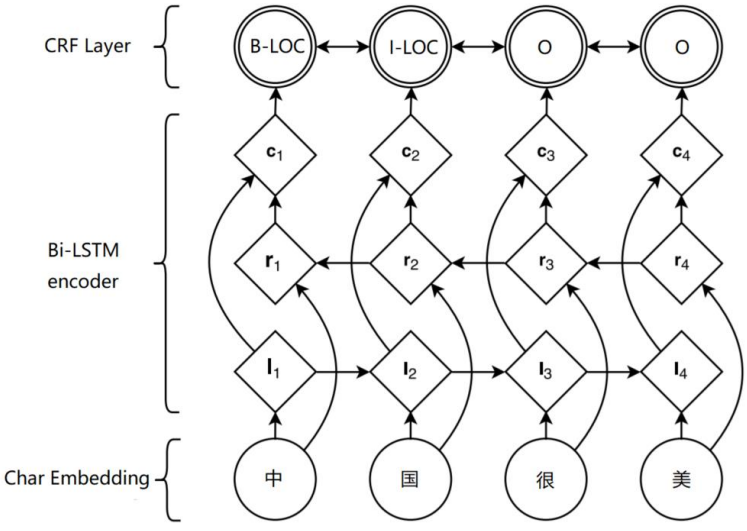

图 3-5: CharEmbedding-BiLSTM-CRF 模型

设输入的序列是 $X=(x_1,x_2,\ldots,x_n)$，经过分布式表示和 BiLSTM 模块后输出的概率矩阵为 P_{n*k}，其中 k 是标签的个数（例如在 BIO 标签系统内识别人名、地名、组织名，标签的个数为 7。$P_{i,j}$ 指 x_i 被标记为第 j 个标签的概率。$A_{i,j}$ 代表概率转移矩阵中第 i 个标签转移到第 j 个标签的概率。

对于将要输出的标签序列 $y=(y_1,y_2,\ldots,y_n)$，定义如下路径得分公式：

$$S(X,y)=\sum_{i=0}^{n}A_{y_i,y_{i+1}}+\sum_{i=0}^{n}P_{i,y_i} \tag{3-1}$$

$$y^*=\operatorname{argmax}S(X,y) \tag{3-2}$$

CRF 模型通过对输出标签二元组进行建模，使用动态规划算法找出得分最高的路径 y^* 作为最优路径进行序列标注。

在神经网络模型训练的过程中，每轮训练用当前 batch 作为测试样例，通过网络模型得出本轮次预测的标签，在条件随机场模型中计算标签序列的对数似然值，以该值的相反数作为神经网络的损失函数。通过误差逆传播算法，用一轮轮训练数据去优化神经网络的参数，直到模型参数稳定，训练误差长时间不再缩小为止。这样就将整个模型训练完毕，通过该模型可以对中文文本进行命名实体识别任务。

3.4 基于位置信息的中文字符向量优化方法

将文本信息进行分布式表示，是深度神经网络自动抽取特征，处理自然语言问题的关键步骤。这一过程将计算机程序不容易理解的字符串，转化为分布式的形式，便于神经网络去拟合复杂度更高的函数。使用 word2vec 训练词向量这种方法将文本的分布式表示与字词间的含义联系起来，消除了词语鸿沟的现象。使用预训练好的词向量作为深度学习处理自然语言问题的输入，已经成为一个经典成熟的方法。很多工作表明 [52] 使用预先训练好的词向量与随机嵌入相比，整个神经网络收敛速度更快；训练好的模型在准确度和召回度上都有较大的提升；特别是在数据量较小的情况下使用 word2vec 的方法优势更加明显。

然而在中文命名实体识别任务中，特征的分布式表示还有很多待解决的问题。由于中文和英文不同，没有空格间隔开词语，而单字对语义表达的能力不强，同样的中文词语都可能意义不同词性不同，同样的汉字含义词性更是千差万别。为了解决中文命名实体识别问题中的特征分布式表示问题，多种解决

方向也在被研究者们不断地探索：一类工作是在分布式表示时加入中文分词词典信息，比如中文分词后进行词向量的嵌入，这类方法的局限性是识别结果较依赖于构建的中文词典，而命名实体识别任务中较为关注的人名、组织名这样的实体他们的命名往往与分词技术相违背，这类实体的命名往往碎片性、象征性、随机性更强，而和组词关系较小。例如“邓小平常说教育要从娃娃抓起。”这句话分词时很容易发生分出词语“平常”，从而导致人名识别的失败，中文词典不可能收录无穷尽的人名组织名，这样的做法也与使用机器学习方法解决命名实体的目标相违背。另一类工作是寻找中文字符更小粒度的特征，就像英文工作中将单词粒度细分为词根词缀等一样，将中文汉字按偏旁部首继续细分，从更小的粒度寻找更细节的特征。汉字和英文单词发展有所区别，汉字起源于象形文字，每个偏旁部首带有一定的含义，在演化的过程中，汉字被不断简化，同音字合并，一字多音等现象十分普遍，这些导致偏旁部首的信息也更加复杂。这类方法对于寻找更细节的特征有助益，但对于解决命名实体识别问题中一字多义的问题还是较为乏力。

根据文献综述和对 CharEmbedding-LSTM-CRF 模型的实验结果分析，可以发现一字多义问题是影响命名实体识别准确度、召回度很重要的原因。中文常用汉字只有 3500 左右，而英文常用单词有 30000 个左右，从这个角度来看，在同样的语义空间内英文单词的词向量信息是多于中文字符向量的，也就是说汉字一字多义、一字多性的现象会更加普遍。

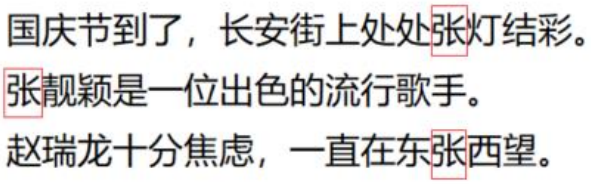

图 3-6: 一字多义、一字多性现象

图3-6中的三句话中都含有“张”字，显然这三个“张”字的含义并不相同，词性也不同。然而在基本的中文字符向量嵌入方法中，这三个张字的字向量嵌入的完全一致，都是通过大量语料中“张”字出现时，窗口内的其余信息训练出的。如果能够尽可能的将不同含义的“张”字表示成不同的字向量，扩充整个字向量空间，使得整体字向量的表达能力更强，字词鸿沟现象可能能够得到更好的解决，也许能对后续使用字向量训练神经网络来解决命名实体识别问题带来相当可观的帮助。

通过上面的例子，同样可以得知，判断一个字符的含义如何，关键要靠其周围的文字信息。这里期望能够在字向量的独立性和词向量的语义特征间取得一个平衡，因而本节提出一个字符向量基于位置信息的改进方法，如下所述。

对于一个句子输入 $X=(x_1,x_2,x_3,...,x_t,...x_n)$，对应嵌入中文字符级字向量 $W=(w(x_1),w(x_2),\cdots w(x_t),\cdots w(x_n))$，这里 w 表示字符转换为字符向量的函数。X 句子中第 t 个字符基于位置信息改进后的字向量表示为 $w^*(x_t)$，则其满足：

$$w^*(x_t)=s(x_t,x-\{x_t\}) \tag{3-3}$$

即优化后的字符向量不仅与其本身的字符有关，还与本句文本中其余的字符含义相关。为方便计算将通过 word2vec 预训练好的字符向量作为新模型中该句其他字符的表示。为了统一形式，将原始 x_t 的向量提取出，这样函数 S=0 时，字符向量即为初始的字符向量。

$$w^*(x_t)=S(w(x_t),w(x_1),w(x_2),\ldots,w(x_{t-1}),w(x_{t+1}),\ldots,w(x_n)) \tag{3-4}$$

$$w^*(x_t)=w(x_t)+S(w(x_1),w(x_2),\ldots,w(x_{t-1}),w(x_{t+1}),\ldots,w(x_n)) \tag{3-5}$$

这里为了简化模型，本文将函数 S 建模成一个线性函数（如式3-6），也就是说新的字符向量是原本的字符向量加上周边原字符向量的带权加和。这样的模型有其合理性，比如说“理发”、“发财”、“出发”，同样的“发”字，原本模型中向量相同，在改进后的模型中，字向量受到周边信息的影响，而周边字向量越相似的其修正后的字向量也就越相近，这也与现实中的词义关联一致，这样的方法可解释性较强。

$$w^*(x_t)=\lambda\cdot w；\ w\text{指 X 句子原始字符向量的矩阵} \tag{3-6}$$

对于模型参数的确定，基于计算方便和贴近现实的角度，这里对模型参数进行一些假设。首先是模型参数的值按 x_t 对称，也就是说假定基于位置信息的影响前后两个方向是相同的，这也符合汉字组词的常识。再者就是，按照信息传播的性质，传播的信息随着距离的增长，信息传递随之减弱。自然语言处理中经典的 N-Gram 模型也同样有类似的假设，即 x_t 前后 k 个字符的信息，而忽略更远的字符信息。优化后的向量可以通过句子的原始字符向量矩阵，乘以一

个多对角矩阵求得，具体方法如图3-7所示。

$$w^*(x_t) = \lambda_k w(x_{t-k}) + \ldots \lambda_2 w(x_{t-2}) + \lambda_1 w(x_{t-1}) + w(x_t) + \lambda_1 w(x_{t+1}) + \ldots + \lambda_k w(x_{t+k}) \tag{3-7}$$

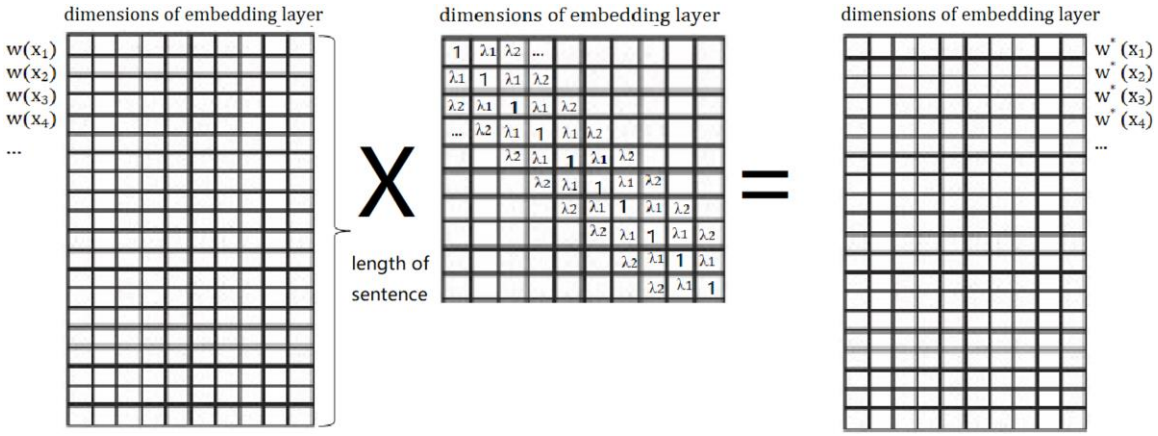

图 3-7: 基于位置信息优化字符向量计算方法

本文对于该模型的参数求解的方法采用的是网格搜索法，当 k 取 1 时，模型的效果较好，在 λ_1 取值在 0.2 左右时，基于此改进方法的中文命名实体识别模型效果达到最佳。具体实验过程与相关结果参见 3.6 节实验结果及分析。通过实验效果来看，这样的基于位置信息优化字符向量表示的方法提高了字符向量表示的效率。其他环节相同的情况下，在标准数据集上实验表明，这种优化方法提高了命名实体识别任务的准确率、召回率等指标。

3.5 基于主题信息的中文字符向量构造方法

在深度学习的分布式特征表示阶段，为了提高模型的效果，应该在特征表示阶段尽量加入足够多的信息以便于后续神经网络模型训练时自动对特征进行抽取，最终准确的对数据进行分类。在各类机器学习模型日臻成熟的今天，信息量的大小、对于数据特征的表示，对整体机器学习模型效果产生着越来越大的影响。

在 word2vec 训练词向量的过程中，是以一个窗口的区域来获得词与词之间的关系。对于 CBOW 方法是通过词语的上下文来预测当前词的向量，而 Skip-gram 方法是按当前词的向量预测上下文的词向量。这样的学习字符间的

关联性主要还是基于窗口内的信息，而缺少对于全局信息的把握。如前文描述（图3-3）指出的问题，可以看出长期信息对于命名实体识别问题也有可能有很大的贡献。在字符分布式表示时，仅仅将信息限制在的滑动窗口以内则有可能会丢失一些对命名实体识别有用的长距离的信息。总体而言 word2vec 对于窗口内部的字符向量的训练学习有很好的语义表征效果，但是同时没有去关注篇章级别的信息，就有可能丧失对于同一篇文档属于同一个主题这样有用的信息。

主题模型（Topic Model）对这样的问题有着一定解决方案，在训练文档的主题向量时，可以产生一个“词语 -主题”的概率矩阵。如图3-8，主题模型将原本的“词语 -文档”按照隐含的“主题”概念，分解为“词语 -主题”矩阵和“文档 -主题”矩阵。“文档 -主题”可以对文本进行聚类，而“词语 -主题”矩阵也可以看作一种词语的分布式表示。对于每一维隐含的主题，数值越高的越与该隐含主题相关性大，对于两个不同的词而言，主题向量越相近，他们背后的语义也越相近。更重要的是这样的“词语 -主题”概率矩阵的训练，是不考虑词序的前提下，用篇章级信息训练出的，或许可以对 word2vec 训练出的信息进行有效的补充。

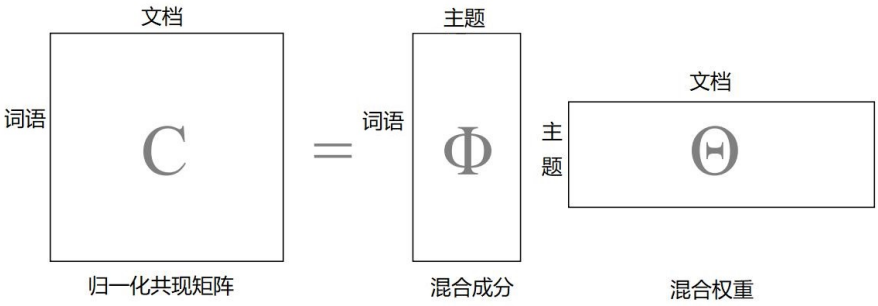

图 3-8: 主题模型概率矩阵

有过一些工作是将主题信息引入到命名实体识别工作中：Jansson 等 [31] 在英文命名实体任务中使用 250 个隐主题来加强词语的分布式表示；文献 [53] 在英文命名实体识别任务中联合卷积神经网络和词对主题模型对命名实体进行识别；文献 [54] 利用潜语义分析技术（LSA）对于条件随机场识别出的命名实体进行歧义消解。基于此，本文尝试训练中文字符级主题向量，将全局字符级信息引入到中文命名实体识别中来。

比如说，一篇科技类的文章中涉及到“百度”、“阿里巴巴”、“腾讯”等，同时也出现了“李彦宏”、“马云”、“马化腾”等信息，这样一篇文档

用主题模型训练，会让这些词的某个主题维度趋同于较大的概率。而只采用word2vec 这样词向量的方法的话，训练的信息更多是公司之间、公司与对应老板间、老板与老板间的关系，而缺少了这些词之间交叉的关联，缺少了这些词背后潜在的主题相关度，这是由于 word2vec 模型窗口的大小不可能无限制扩大决定的。因此，如果将训练集中的全局主题信息加入，可以使得字符级的中文分布式向量容纳的信息更多更全，强化向量表示能力以达到提高整体CharEmbeddeing-LSTM-CRF 命名实体识别模型效果的目的。

主题模型这类无监督文本聚类方法假设一个文本（可以是一篇文章也可以是一个段落），隐含表达着一系列主题，而这些主题可以由一个主题向量表示，主题向量越接近说明文本的相关性越强。不同于 word2vec 在大小为 10 左右的窗口内建模，主题模型基于篇章级别，只要在一篇文档内共现，主题模型都将计算词与词之间的联系。

基于前人工作和使用主题信息的合理性，本文将主题信息引入到中文字符级别的分布式表示中。采用隐狄利克雷分配模型（Latent Dirichlet Allocation，LDA），对字符级词语进行训练。将一个中文字符作为一个独立语义的词语，通过 LDA 算法训练模型。模型预先设置 K 个主题，每篇文档围绕这 K 个主题生成字。主题模型的建模方法是，文档按概率选择主题向量中的一个主题，这个主题再按概率选择一个该主题下的字，这样的方法生成一整篇文档。假设文档与主题符合多项式分布，字符与主题也符合多项式分布，而这两个多项式分布的参数符合具有先验参数的狄利克雷分布。本文的工作是需要训练这样的模型，得到字与主题之间的信息。

Algorithm 3.1 字符级主题向量训练算法

1: 输入：先验参数 α 和 β，主题数量 K，语料集 D
2: 输出: “字符 -主题”参数矩阵 ϕ，“文档 -主题”参数矩阵 θ
3: 对文档中的所有中文字符进行遍历，为其随机分配一个主题，
即 $z(m,n)=k\sim Mult(1/K)$，
4: 遍历文档，计算当前主题：
$P(z_i=k|z_{\neg i},w)=\frac{n^t_{k,\neg i}+\beta_t}{\sum_{t=1}^{V}n^t_{k,\neg i}+\beta_t}\cdot\frac{n^k_{m,\neg i}+\alpha_k}{\sum_{k=1}^{K}n^k_{m,\neg i}+\alpha_k}$
5: 迭代完成后输出“字符 -主题”参数矩阵 Φ，该矩阵即为不同的字符词语在不同主题下的概率情况

如前文 2.6.3 节所述，LDA 模型的参数求解采用最常用的方法是吉布斯采样算法。本文训练集采用腾讯新闻数据一万多篇不同类型的新闻文本，隐主题个数设为 50，迭代次数设为 100。训练完成后，观察距离每个主题概率值最大

的字符，效果如表 3-2：

表 3-2: 主题下字符分布情况

topic1	topic2	topic3	topic4	topic5	topic6	……
教 0.032	矿 0.175	酒 0.019	罪 0.015	狗 0.074	医 0.063	……
校 0.031	煤 0.090	驾 0.018	刑 0.011	犬 0.029	疗 0.048	……
聘 0.023	鹤 0.088	斑 0.012	贿 0.011	鼠 0.025	药 0.030	……
学 0.019	岗 0.060	肇 0.011	犯 0.009	樟 0.024	蚁 0.015	……
詹 0.019	井 0.055	乳 0.011	审 0.009	豚 0.018	诊 0.013	……
师 0.017	炸 0.036	车 0.011	案 0.009	礁 0.012	患 0.012	……
育 0.017	瓦 0.034	驶 0.011	判 0.008	猴 0.012	病 0.012	……
毕 0.016	难 0.032	飙 0.008	银 0.007	碚 0.012	痪 0.012	……

可见相同主题下的字词有着一定的联系，因为对于同一篇文档来说，文档中出现的字在主题信息上有一定的联系。将字符主题向量与字符 word2vec 训练出的向量进行比较，由于很多字词在各主题上概率值较小，对主题向量做适当放大，在整个字符词库中寻找与被比较字最相近的字符，效果如表 3-3：

表 3-3: 主题向量距离比较

全	国	教	育	会	议
施 18.412	点 83.003	育 16.681	教 16.681	议 13.521	共 12.872
各 20.236	推 89.117	考 25.913	试 17.997	十 14.062	会 13.521
加 20.858	商 91.716	校 26.159	培 21.305	共 16.252	协 14.504
应 21.162	网 94.371	科 30.933	科 23.932	协 16.611	作 17.373
度 21.946	际 102.126	试 32.029	府 24.139	领 16.873	问 17.684
措 23.018	内 104.328	府 32.148	干 24.844	题 16.917	式 18.287
提 23.213	闻 128.864	干 33.192	养 25.552	关 17.823	系 18.440
重 23.429	排 131.895	实 34.996	考 25.751	与 18.385	总 18.969

与通过 word2vec 算出的最近字符（表 3-1）相比，与“全”“国”“教”“育”“会”“议”几个字相似度高的词语，既有类似的部分也有不同的部分。总的来说 word2vec 生成的向量字与字之间可组词的较多，比较关注局部信息，而 LDA 生成的字符主题向量，相近的字符向量不仅有局部的信息，也纳入了潜在的主题关联。这与两种模型的设计、逻辑关联很大，word2vec 模型是根

据浅层神经网络在窗口内对目标词语和周边信息相互预测的结果，而 LDA 主题模型对于词序等局部信息关注不多，是基于整个文档而言，有更加丰富的全局信息，能够对 word2vec 向量进行更多信息的补充。这样有差异有趋同的效果也表明了上文思路的合理性，加入字符级主题信息可以提高字符分布式表示的信息量。

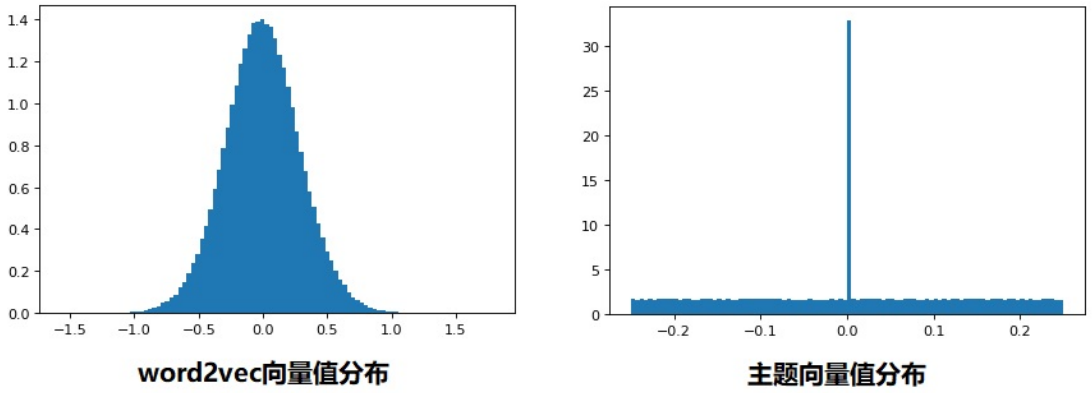

图 3-9: 两种类别字符向量分布直方图比较

图 3-9是取所有语料集中出现的字符，将对应使用 word2vec 训练的字符向量和 lda 模型训练出的主题概率向量做分布直方图比较。横坐标是两类向量每个维度值的大小，纵坐标是出现的频次。可以看出 word2vec 向量各维度值的分布大致符合正态分布，而主题向量值除了 0 附近的值，其余的值接近均匀分布。直接将两种分布差异较大的向量结合，训练效果并不好，因而需要对于主题向量采用归一化的方法，使得主题向量分布的方差，最值更为相似。本文是将主题向量每个维度减去其均值使得总体主题向量与 word2vec 字符向量一致均值在 0 附近，每个维度再除以该维度的方差。这样使得主题向量的最大值最小值在 1 与 -1 附近与 word2vec 相近，使得两类向量同时在神经网络中优化训练时更为迅速。将字符主题向量与 word2vec 训练出的字向量结合，构成新的字符分布式表示。这一表示方法经实验表明可以有效提高中文命名实体识别模型的效果，下一节将用实验说明。

本章改进的中文字符级特征表示方法，在初始 word2vec 中文字符向量的基础上，为了解决一字多义，字向量空间局限的问题，采用了基于位置信息对字符向量进行优化；为了解决原模型局限于局部信息的缺点加入了全文级别的字符主题特征，结合成为一种较优的中文字符级特征表示方法，该改进方法及整体基于深度学习的命名实体识别模型流程图如图 3-10所示。

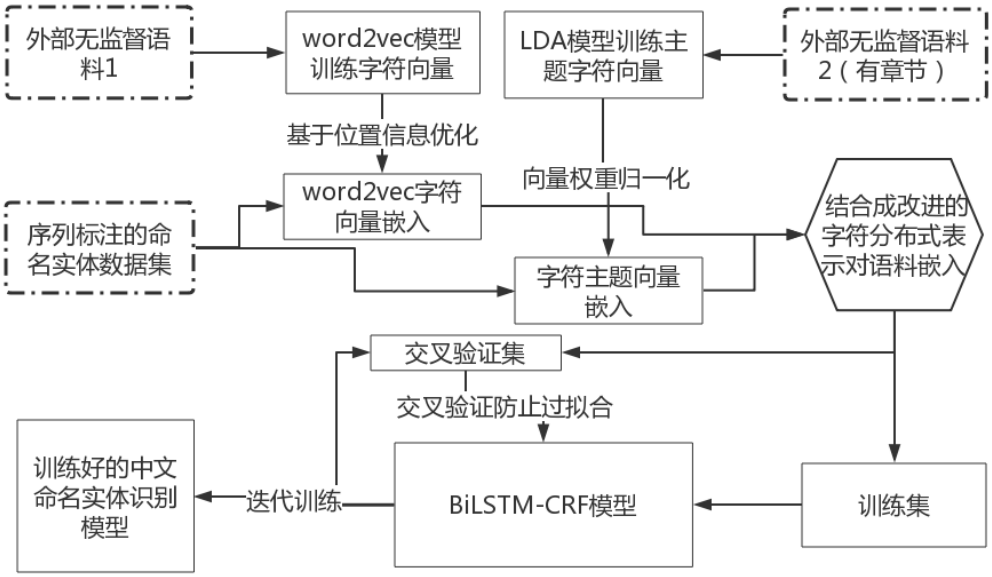

图 3-10: 改进的中文字符级特征表示方法模型流程图

3.6 实验

3.6.1 实验环境与设置

3.6.1.1 实验环境

本文的实验主要用到实验设备主要有一台主频 2.2GHz，16GB 内存，处理器四核 i5 的个人计算机，操作系统为 Windows；还有两台主频为 2.2GHz、处理器为 40 核的 Intel Xeon E5-2630，内存 128G，操作系统为 Ubun16.04 的服务器。

各项实验均在 Python3.6 环境下编写运行，其中词向量训练、主题向量训练采用 Gensim 开源软件库，版本为 3.0.1。Gensim 是一个可用于无监督主题建模和自然语言处理的开源软件库，利用现代统计机器学方法对语言建模，其对大规模语料和在线算法的支持性非常好。利用 Gensim 可以帮助实现 Fasttext、word2vec、doc2vec、LSA、LDA、NMF、TFIDF 等经典自然语言处理模型。

神经网络模型采用谷歌公司开发的开源软件库 TensorFlow[55] 实现。TensorFlow 是一个设计用于数据流可迭代计算的开源框架，由 Google Brain 团队开发，支持 CPU、GPU、TPU 等多种设备。TensorFlow 将计算表示为状态数

据流图，使得深度学习的各种模型可以方便地编写运行，常见的 CNN、RNN、LSTM 等网络模型可以方便的开发和实现。

3.6.1.2 实验数据

本实验的共要训练三个模型，分别是用词向量模型训练字符向量、用 LDA 主题模型训练字符主题向量和用 BiLSTM-CRF 训练命名实体是别模型。共涉及到三种数据集，具体如下：

（1）本文实验里中文命名实体识别数据集采用的是人民日报语料，按“BIO”标签方案标注，每个中文字符标注为{“O”，“B-PER”，“I-PER”，“B-LOC”，“I-LOC”，“B-ORG”，“I-ORG”}集合中的一个类别。将数据集分为训练集和测试集，训练集用来训练中文命名实体识别模型，测试集用来衡量模型最终的效果。训练集共有 46364 条语料，包括 17615 个人名实体，36517 个地名实体，20571 个组织名实体；测试集共有 4365 条语料，包括 1973 个人名实体，2877 个地名实体，1331 个组织名实体。

（2）本文实验中 word2vec 模型训练字向量的模型采用的是中文维基百科语料，共有 133 万余条词条，共有 438530048 字，涉及各类名词的定义，解释等等。

（3）本文实验中 LDA 模型字符主题向量训练采用的语料是腾讯新闻，共 15147 篇各主题的新闻，约 2000 万字。主题多样，包含体育、政治、娱乐、社会、财经等等话题。

3.6.1.3 网络参数与训练方法

本文实验采用的神经网络模型涉及多种参数以及各种网络训练优化方法，以下对这些内容一一介绍：

（1）网络参数

本实验 BiLSTM 网络模型中的参数设置如下：根据计算机内存情况，本文将实验的批处理数（batch_size）设为 64；经多次实验观察发现各类模型在整个数据集迭代 80 次内都能达到基本稳定，因而最大迭代轮次为 100；优化算法采用 Adam[56]（自适应时刻估计法）算法；clip 梯度设置为 5.0 来防止训练过程中遇到梯度爆炸的情况。

（2）目标函数

目标函数设置为条件随机场对数似然函数（crf_log_likelihood），该函数将

神经网络层输出的“字符 -标签”矩阵与实际标注标签比较，分别计算序列一元（发射概率）概率得分值与二元（转移概率）概率得分值，加和后得到序列得分值，将这个分值减去该条件随机场的对数归一化作为条件随机场对数似然。最后，将一个 batch 中的条件随机场对数似然和取负作为网络训练的目标函数，用于训练网络。

（3）Xavier 初始化

在深度学习过程中，各层神经元不能初始化为 0，不然将会遭遇梯度为 0 的情况。因而要对神经元初始化，Xavier 初始化由文献 [57] 提出，思路是参数初始化时应该使得各层的激活值和状态梯度的方差在传播过程中的方差保持一致，这样的初始化方式可以使得神经网络在向前传播时神经元输出值的方差不会不断增大，有较强的稳定性。

（4）随机失活（Dropout）

随机失活这种网络训练方法，借鉴正则化的思想，对随机失活作用下的一层节点，每个节点设置一个节点保留概率，取值在 0 到 1 之间，在训练时按概率超过设置阈值时失活。这样的方法使得神经网络不会过于偏向某一个节点，从而是的单个节点的权重不会过大。本实验中对所有神经网络层都采取这一优化训练的方法。

（5）shuffle

shuffle 方法是在每轮迭代前将训练集重新排序，使得这些例子能够被随机分配到不同的 batches 中。这样的方法可以使数据更加混乱，防止网络的过拟合，让模型训练的效果更好。本实验在每轮迭代前通过 shuffle 方法随机打乱训练集。

（6）早停（Earlystop）

模型迭代的次数太长的话会造成网络对于训练集过拟合，即在训练集上表现越来越好，但是在其他数据集上表现越来越差。训练模型时往往没有到达最大迭代次数时就可以提前结束训练数据，保存模型往往能达到在总体数据上较好的表现，这样的方法就叫做早停策略。利用这样的方法，每隔一段训练次数用交叉验证集测试当年模型的效果，当模型在交叉验证集上长时间效果不再提升甚至有下降的趋势时提前停止训练，保存模型。

3.6.2 实验评价指标

本文实验采用最常见的查准率和查全率作为实验的评价指标。将样例的真实类别与学习器预测类别组合分为真正例、假正例、假反例、真反例，令 TP、FN、FP、TN 分别等于它们的样例数。显然有 TP+FN+FP+TN= 总样例数，分类结果的“混淆矩阵”如表 3-4：

表 3-4: 分类结果混淆矩阵

真实情况	预测结果	
	正例	反例
正例	TP（真正例）	FN（假反例）
反例	FP（假正例）	TN（真反例）

查准率（precision）与查全率（recall）定义如下：

$$precision = \frac{TP}{TP + FP} \tag{3-8}$$

$$recall = \frac{TP}{TP + FN} \tag{3-9}$$

查准率和查全率是一对矛盾的度量，对于同一个模型来说，往往查准率越高时查全率就越低，用 F_1 值来兼顾这两个指标：

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{3-10}$$

3.6.3 实验结果与分析

本小节展现了本章涉及的各个模型的效果，包括基准模型的实验结果、基于位置信息优化方法的实验验证、基于主题信息的构造方法的实验验证以及本章提出的中文字符级特征表示的改进方法对整体命名实体识别模型效果的提升。

（1）基准模型效果

利用训练语料训练出 100 维中文字符向量，嵌入后通过双向长短时记忆网络和条件随机场进行训练，得到 CharEmbedding-BiLSTM-CRF 模型。

表 3-5 是 CharEmbedding-BiLSTM-CRF 模型的实验结果，该模型对于中文命名实体识别有较优的效果。同时观察不同类别实体的识别准确度召回度，也

表 3-5: CharEmbedding-BiLSTM-CRF 模型实验结果

	准确率	召回率	F 值
LOC	93.24%	88.25%	0.9068
ORG	82.31%	82.87%	0.8259
PER	88.77%	82.11%	0.8531
整体	89.37%	85.13%	0.8720

能看出该模型对于地名的识别较为准确，对于人名、组织名的识别能力较差。因为地名重复性强，而人名组织名重组复杂性高很多，这两类实体也是模型待提高的要点。

（2）基于位置信息优化方法实验与分析

根据 3.4 节内容，基于位置信息优化中文字符向量的方法涉及到偏移参数的寻找，采用网格搜索的方法找出较优的参数。当只考虑 x_t 前后 1 个字符的信息时，取不同的 λ 值时比较模型各类实体 F 值的变化，效果如图 3-11：

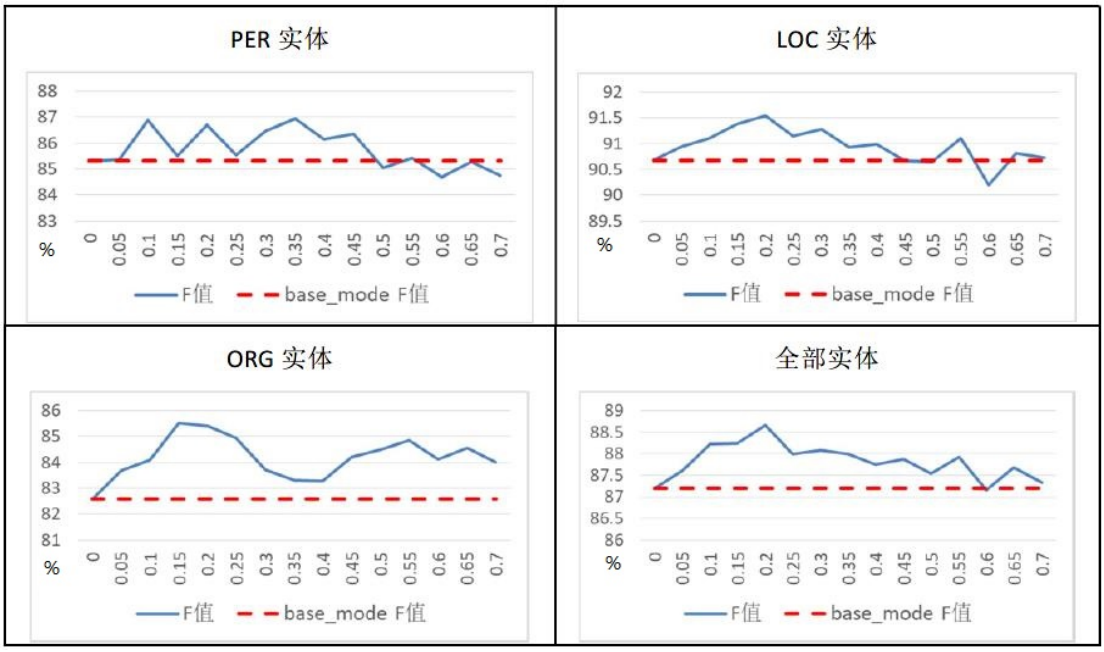

图 3-11: 加入周边一个字符信息模型 F 值随 λ 值变化情况

图中的虚线是基准模型“CharEmbedding-BiLSTM-CRF”模型在人名实体、地名实体、组织名实体的和全部实体下 F 值的情况，实线是 λ 取不同值时训练出的模型在各类实体下 F 值的情况。通过折线图可以看出，引入中文字符周边字符向量的信息对于提高最终命名实体识别模型的效果有这一定的帮助。

这种优化方法对于地名实体和组织名实体效果较好，对与人名实体的效果没有前两者明显，分析其原因这很可能是由于人名字符中，中文字符的含义不明，前后字符间联系没有那么紧密造成的。观察 λ 在 0 到 0.7 时 F 值的变化过程可以看出，λ 在 0.2 左右时模型各方面效果都比较好。

图 3-12: 加入周边两个字符信息模型效果图

接下来考虑周边两个字符，在 λ_1 为 0.2 的情况下，λ_2 取不同值观察这些模型的效果，如图 3-12。

图中虚线是仅有 λ_1 为 0.2 时模型对各类实体的 F 值，实线是 λ_2 由 0.01 到 0.15 间变化时对应模型的效果。从图中观察到周边距离为 2 的信息对模型效果的提升较为有限，表明距离为 2 的字符能够提供的语义信息较少。因而本文最终采用的中文字符向量嵌入的优化方法是在中文字符分布式表达时，在嵌入的向量中附加周边一个字符系数为 0.2 的信息。借此来解决中文字符向量一字多义，向量空间稀疏的问题。这样的方法与不做优化直接嵌入字符向量的方法相比，模型查准率提高了 0.74%，查全率提高了 2.72%，F 值提高了 1.75%。

每隔一千次迭代训练记录改进方法和原模型 loss 值和 F 值的变化情况如图 3-13所示，左坐标轴是训练 loss，右坐标轴是模型训练到此时在交叉验证集上的 F 值。可以看出改进的方法在收敛速度，训练效果上都有一定的优势。

（3）基于主题信息的构造方法实验与分析

根据 3.5 节内容，我们通过腾讯新闻数据集训练字符级 LDA 主题模型，其中主题数设为 50，训练迭代轮次设为 200。训练出模型后取出每个字符对应的主题概率向量，作为字符分布式表示的辅助特征。分别比较嵌入方法为 150 维字符向量、100 维字符向量结合 50 维随机向量和 100 维字向量结合 50 维字符

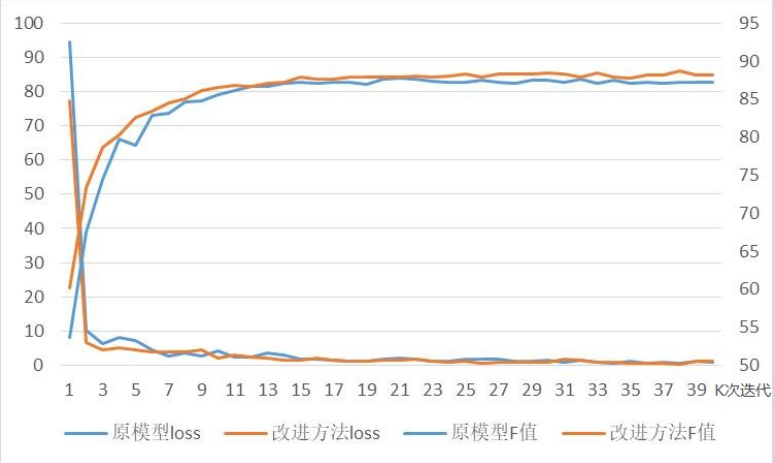

图 3-13: 训练过程比较

主题概率向量三种方法。结果如表 3-6：

表 3-6: 主题向量辅助特征效果

		150 维字向量	100 维字向量 50 维随机向量	100 维字向量 50 维主题向量
LOC 实体	准确率	0.9162	0.9201	**0.9404**
	召回率	0.8833	0.8843	**0.8947**
	F 值	0.9046	0.9018	**0.9170**
ORG 实体	准确率	**0.8544**	0.8085	0.8272
	召回率	0.8377	**0.8535**	0.8415
	F 值	**0.8460**	0.8304	0.8343
PER 实体	准确率	**0.8952**	0.8894	0.8739
	召回率	0.8443	0.8427	**0.8695**
	F 值	0.8690	0.8654	**0.8717**
全部实体	准确率	**0.8962**	0.8846	0.8935
	召回率	0.8656	0.8643	**0.8752**
	F 值	0.8806	0.8744	**0.8842**

从表中我们可以看出，加入字符的主题向量作为辅助特征相比于只使用 word2vec 训练出的字向量，提高了后续地名实体、人名实体的识别能力，但是在组织名实体识别上仍有欠缺。同时将该模型与 100 维字符向量结合 50 维随机向量作为嵌入方式相比，各项指标均有上升，这也说明加入的全局主题信息对

于命名实体识别是有帮助的。综合来看，在维数一致、利用 1 万多篇新闻训练出字符级主题模型的情况下，本文提出的基于主题信息的中文字符向量构造方法可以提高命名实体识别模型的 F 值在 0.4% 左右。

不过，在实验中发现，结合字符级主题向量的命名实体识别模型收敛较慢。在其余参数设定一致的情况下，仅使用 word2vec 模型训练出的字符向量，在本文实验中大约在所有训练集被训练 40-50 轮次时 F 值不再上升。加入主题向量后，训练完成需要更多的时间，往往要更多轮次的训练才能使得模型最优（大约 60-70 轮次）。这种现象很可能是两类结合的向量分布不同导致的，也是模型间比较时需要考虑的。

（4）整体改进中文字符特征方法的实验与分析

最后，按3-10所述流程，将基于位置信息的中文字符向量优化方法与基于主题信息的中文字符向量构造方法结合起来，训练出本章改进的中文字符级特征表示方法。在测试集中比较本文改进模型与其他模型的所有标签准确率、命名实体查准率、命名实体查全率、命名实体 F 值，结果如表3-7：

表 3-7: 不同模型结果比较

	准确率	查准率	查全率	F 值
CE-BiLSTM	0.9829	0.8065	0.8198	0.8131
RE-BiLSTM-CRF	0.9799	0.8638	0.8233	0.8431
CE-BiLSTM-CRF	0.9827	0.8937	0.8513	0.8720
modified-CE-BiLSTM-CRF	0.9851	**0.9035**	0.8710	0.8869
LDA-CE-BiLSTM-CRF	0.9855	0.8958	0.8729	0.8842
LDA-modified-CE-BiLSTM-CRF	**0.9859**	0.9017	**0.8815**	**0.8915**

其中“CE-BiLSTM”指采用 word2vec 训练的“char-embedding”字符向量经 BiLSTM 网络训练，最终用一层 softmax 激活函数对字符直接预测标签，可见该模型对于整体标签准确率尚可，但是没有联系标签之间的关联，对于命名实体识别能力较差。“RE-BiLSTM-CRF”指不使用预训练的字符向量，而是直接随机出字符向量嵌入的方法。“CE-BiLSTM-CRF”指使用预训练的字符向量嵌入，也是本章的基准模型，相比随机初始化，嵌入预训练的字符向量提高了模型的信息量，提升了命名实体识别模型的效果。“modified-CE-BiLSTM-CRF”指基于位置信息对字符向量嵌入时进行优化的方法。“LDA-CE-BiLSTM-CRF”指在字符向量表示中加入字符级主题向量信息作为辅助信息的方法。

“LDA-modified-CE-BiLSTM-CRF”指结合两种优化方法的模型。

实验结果表明，本文提出的改进的中文字符特征表示方法相比于基准模型，有效地提高了整体中文命名实体识别模型的效果。在标准中文命名实体数据集中，所有标签准确率提高了 0.33%, 命名实体准确率提高了 0.89%，命名实体召回率提高了 3.55%，F 值提高了 2.24%。

3.7 本章小结

本章针对目前的中文命名实体识别模型中，中文字符向量一字多义现象普遍、向量空间稀疏、信息量不足的问题，提出了一种改进的中文字符级特征表示方法。针对一字多义、空间稀疏的问题，采用基于位置信息的优化，在字符向量中加入周边字符的信息；针对信息量不足问题，引入中文字符级主题向量作为辅助特征，扩充字符级特征表示中的全局篇章级信息。实验表明，本章提出的改进的中文字符级特征表示方法有效地提高了整体中文命名实体识别模型的识别能力。

第四章　面向复杂中文命名实体识别的层叠模型

4.1　引言

实际的工程领域中，中文命名实体识别技术还有很多值得研究的问题。在项目工程中应用命名实体识别系统会遇到很多在标准数据集实验中很少遇到或不会遇到的问题：（1）实际应用中会出现许多地名人名组织名嵌套的命名实体，遇到这样的实体时，模型的准确率会下降；（2）互联网文本信息结构杂乱，形式多变，直接交给中文命名实体识别系统的效果不好；（3）当输入文本长度长时，命名实体识别模型的能力会明显下降，需要采用一些合理的方法对文本进行合理的切割来提高识别效果。我们对这些情况下的命名实体逐一分析：

（1）嵌套命名实体。如表4-1所示，该结果是单层中文命名实体识别模型对文本的实体识别结果图，目标识别的命名实体是“上海农商行”，但是该命名实体中还有子地名命名实体“上海”，通过 BiLSTM 识别出各类标签的概率后，通过条件随机场对各种序列的得分进行比较，最终系统并没有将“上海农商行”识别为一个整体。对该结果进行分析判断，“上”“海”两字的“B-LOC”“I-LOC”标签得分很高，即使加上后文非命名实体的标签得分仍然高过上海农商行“B-ORG”“I-ORG”“I-ORG”“I-ORG”“I-ORG”“I-ORG”的标签序列得分，影响了整体实体被识别为组织名的成功率。

表 4-1: 嵌套命名实体

单层模型输入：	谋求上市应该是上海农商行自带的话题光环，且目前已经有了实质性进展。
人名实体：	[]
地名实体：	[“上海”]
组织名实体：	[]

（2）句子长度长成分复杂。太长的文本对于命名实体识别模型来说难度更

大，特别是实际工程应用中，网络中文本的规则性，断句标点使用的规范性相比标准数据集差了很多，在文本长度很长时，条件随机场算法要通过维特比动态规划计算的最大得分路径作为最终的输出结果，从效果来看往往准确率下降很多。如表4-2所示，输入较长文本时出现了命名实体识别错误，而将文本切割输入较短文本，系统可以成功识别出正确的命名实体。这表明文本长度太长是影响命名实体识别系统在实际应用中较大的一个需要解决的问题。

表 4-2: 句子长成分复杂

单层模型输入：	新民晚报讯（记者叶薇）近日，有媒体发表报道称，“德国大众计划提高其在华合资企业的持股比例”，内容涉及上汽集团下属企业上汽大众。
人名实体：	[“叶薇”]
地名实体：	[”华”]
组织名实体：	[“德国大众计划提高其”，“上汽集团”，“上汽大众”]

（3）命名实体前后文错误关联。前文与后文有含义上的关联，被命名实体识别系统识别错了边界也是一种常见分类错误。如表4-3，“南京银行资金运营中心”命名实体与前文边界不明被错误标注。当模型分析文本较长时，由于前后文语义的关联，训练标注集数据不够多等因素，命名实体识别系统常常对于命名实体的边界把握不准。

表 4-3: 前后文错误关联

单层模型输入：	原上海银监局于 2017 年 12 月批复同意南京银行资金运营中心开业，同时核准董文昭南京银行资金运营中心副总经理的任职资格
人名实体：	[]
地名实体：	[]
组织名实体：	[“上海银监局”，“南京银行资金运营中心”，“同时核准董文昭南京银行资金运营中心”]

基于这些原因，本文在众多研究的基础上提出了一个层叠深度神经网络模型来解决实际应用中容易遇到的复杂中文命名实体识别问题。设计多层的字符级别嵌入的 BiLSTM-CRF 结构，低层网络在优先侧重召回率的情况下进行初步的命名实体识别，对于识别出的粗粒度命名实体送交下一层网络实现文本分割的工作，高层网络再精准地判断语言片段中有哪些命名实体。

本章的结构如下：4.2 节介绍多层模型结构在命名实体识别任务中的相关工作研究；4.3 节介绍本章提出的层叠深度神经网络命名实体识别模型；4.4 节

基于标准数据集和实际应用中的数据集，针对该模型进行实验验证以及结果分析；4.5 节是本章小结。

4.2 复杂命名实体及多层模型相关工作

复杂命名实体是在实际项目应用中对命名实体识别系统效果影响重要的一系列实体形式。这类问题的本质是潜在实体名之间相互影响，造成 BiLSTM-CRF 网络结构对于标签序列预测的准确率下降。这些复杂命名实体的典型特征主要表现在（1）句子长度长、成分复杂、识别其中命名实体困难。（2）命名实体名长度长、命名实体前后文关联强，造成命名实体识别系统不能够准确的识别命名实体名的边界。（3）实体识别中还含有子命名实体，实体名相互嵌套，标签混乱导致识别错误。

为了识别这类嵌套混淆类型文本中的复杂命名实体，很多学者在识别嵌套混淆的命名实体以及构建多层的学习网络开展了许多研究工作：

文献 [58] 在中文机构名命名实体识别任务中引入层叠条件随机场模型，低层条件随机场仅以观察序列为输入，识别较容易的人名、地名等命名实体，高层条件随机场的输入不仅包含观察值同时也包括低层条件随机场的识别结果，高层条件随机场用来识别复杂的机构名。文献 [59] 基于中文微博用词简明、流行用语多等特性设计了一个层叠的条件随机场模型来完成命名实体识别任务，低层条件随机场用一个滑动窗口来寻找句子的实体特征，高层随机条件场对句子成分进行实体标注。文献 [60] 在电子商务领域利用电子商务知识库，在 BiLSTM-CRF 模型的基础上，设计了一个层次的深度学习模型，多任务同时执行分词标注、命名实体识别和填槽任务，将该模型应用在商用的电商平台系统中。文献 [61] 利用动态栈和 LSTM-CRF 模型，一步一步抽取外部命名实体名，这种层次结构对于嵌套的英文生物医学命名实体识别有着很好的效果。文献 [62] 提出了一个简单的深度神经网络模型来识别嵌套的命名实体，该模型采用枚举法，列举所有可能的区域和跨度作为潜在的实体，利用神经网络模型对其进行分类，在生物医学领域的语料库有良好的表现。文献 [63] 提出了一种新的循环神经网络构建方法，该模型的特点是通过特征抽取构建嵌套实体间的超图表示，这种方法在英文命名实体抽取任务上表现优异。文献 [64] 提出了一种混合层叠的命名实体识别模型，该模型将人名实体识别、地名实体识别、机构名实体识别分别构成一层，高层模型同时利用低层模型识别出的结果，该方法在

中文论文数据文本中有较好的表现。文献 [65] 利用外部数据语料辅助中文嵌套命名实体识别任务，该文献利用中文维基百科数据，通过匹配、过滤和汇聚三个阶段，构建中文嵌套命名实体识别语料库，在嵌套命名实体识别的准确率上有着不错的表现，在召回率上有一定不足。

通过分析学者们的相关工作，可以发现，进一步解决复杂中文命名实体问题、提高识别效果，构建多层命名实体识别模型，层次性的处理输入文本是一个可行的方向。接下来将介绍本文提出的层叠深度神经网络模型，该模型可以针对性的优化复杂命名实体的识别问题。

4.3 层叠深度神经网络模型构建

4.3.1 层叠模型原理与构建

基于上述理论与经验，本文通过设计多层的命名实体结构来提升命名实体识别系统的效果。建立多层模型主要有两种不同的方法：一种多层模型是采用递归的方式建立模型，低层模型作为高层模型的一个子模型，在训练时一同优化，这类方法称为层次模型，例如文献 [60]；另一种多层模型是采用模型间的线性组合，低层模型与高层模型间相互叠加，不同层针对不同类型的问题，分别训练，这类方法称为层叠方法，例如文献 [59, 64]。

一般来说，层次模型的数学模型更为复杂，训练更为困难，模型的复杂度较大。而在层叠模型中，低层模型和高层模型间是一种松耦合的状态，各层模型独立训练，模型复杂度仅是线性加和关系。低层模型对命名实体识别问题初步解决，而该层识别过程中产生的错误，高层模型仍可以调整过滤，避免了错误的扩散，提高了最终的识别的准确率。本文采用层叠的方法设计低层网络和高层网络，层叠网络如图 4-1所示。

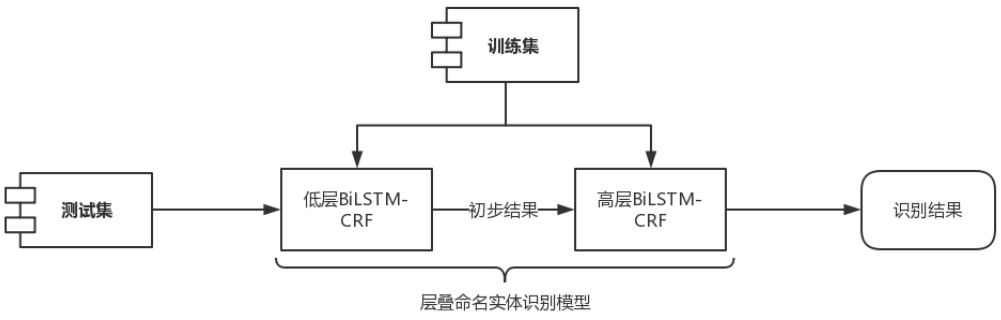

图 4-1: 层叠命名实体识别模型示意图

本文构建层叠结构的基本思路是低层网络在优先保证召回率的情况下，尽可能多地识别出命名实体，并将识别结果传递到高层网络。通过这样的识别后，将初步识别结果向后传递的过程也就完成了文本的合理分割，提升了后续识别的效率。高层网络对于低层传来的信息再次识别，若识别结果与原命名实体一致则说明层叠模型与单层模型判别一致，将结果输出即可。若识别结果与原命名实体不一致，则比较两个结果的可能性，选择较佳的结果作为最终的输出。

4.3.2 面向文本切割的低层网络构建

为了达到尽量不丢失潜在命名实体信息的目的，在低层 BiLSTM-CRF 的训练和识别阶段优化该模型。这里来看一下 BiLSTM-CRF 模型标签预测的流程。对于一个输入序列 $X=(x_1,x_2,\ldots,x_n)$，设该句子经过分布式嵌入，BiLSTM 网络计算后输出的矩阵为 P，P 矩阵维度为 $n\times k$，k 代表不同标签的个数，例如在本文“BIO”标签体系中 k 为 7。$P_{i,j}$ 即为第 i 个字符标记为第 j 个标签的得分，称为发射概率。对于潜在的一个预测序列 $y=(y_1,y_2,\ldots,y_n)$，定义这个序列的得分为：

$$s(\boldsymbol{X},\boldsymbol{y})=\sum_{i=0}^{n}A_{y_i,y_{i+1}}+\sum_{i=1}^{n}P_{i,y_i} \tag{4-1}$$

其中 A 是转移概率矩阵，大小为 $k\times k$，$A_{i,j}$ 表示标签 i 转移到标签 j 的转移概率。也就是序列得分由两部分构成，一部分得分来自序列中当前字符属于当前标签的概率，一部分来自于周边标签和当前标签的转移概率。当序列 X 给定时，根据所有可能标签序列的 softmax 函数值可得到序列 y 的概率：

$$p(\boldsymbol{y}|\boldsymbol{X})=\frac{e^{s(\boldsymbol{X},\boldsymbol{y})}}{\sum_{\tilde{\boldsymbol{y}}\in\boldsymbol{Y}_{\boldsymbol{X}}}e^{s^{s}(\boldsymbol{X},\tilde{\boldsymbol{y}})}} \tag{4-2}$$

在训练网络的过程中，需要最大化正确标记序列的对数概率函数：

$$\begin{aligned}\log(p(\boldsymbol{y}|\boldsymbol{X}))&=s(\boldsymbol{X},\boldsymbol{y})-\log\left(\sum_{\tilde{\boldsymbol{y}}\in\boldsymbol{Y}_{\boldsymbol{X}}}e^{s(\boldsymbol{X},\tilde{\boldsymbol{y}})}\right)\\&=s(\boldsymbol{X},\boldsymbol{y})-\underset{\tilde{\boldsymbol{y}}\in\boldsymbol{Y}_{\boldsymbol{X}}}{\operatorname{logadd}}\,s(\boldsymbol{X},\tilde{\boldsymbol{y}})\end{aligned} \tag{4-3}$$

这里 $\boldsymbol{Y}_{\boldsymbol{X}}$ 代表所有可能的标签序列，包括不满足真实标签排序的情况。通

过以上公式，希望网络可以生成一个有效的标签序列，在解码时通过维特比动态规划算法找出得分最高的序列：

$$\boldsymbol{y}^{*}=\underset{\boldsymbol{y}\in\boldsymbol{Y}_{\boldsymbol{X}}}{\operatorname{argmax}}s(\boldsymbol{X},\widetilde{\boldsymbol{y}}) \tag{4-4}$$

得到的最高分标签序列就是最终的输出序列，在通过遍历得到识别出的人名、地名、组织名等命名实体即可。整体标签预测的流程如下图4-2所示。

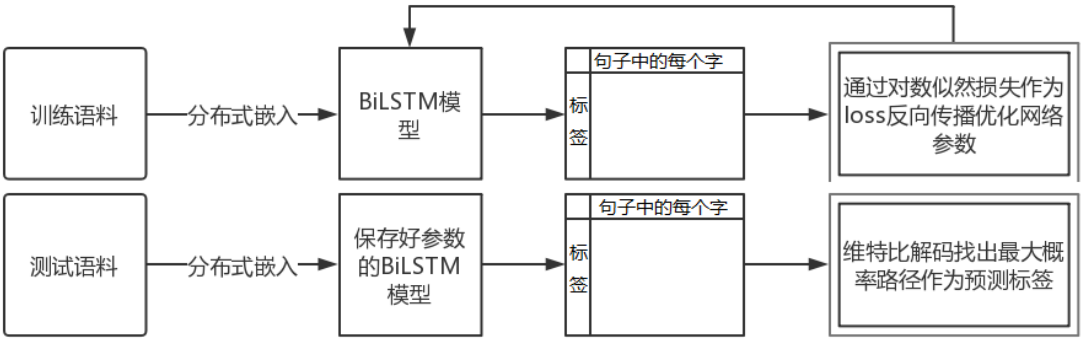

图 4-2: BiLSTM-CRF 模型标签预测流程

在训练过程中利用对数似然函数作为损失函数，计算当前模型下输出的“字符 -标签”发射概率矩阵的误差。而根据公式4-1，对数似然函数由一元损失值（发射概率矩阵误差）和二元损失值（转移概率矩阵误差）加和组成。将对数似然函数取负作为整个网络的目标函数，进行最小优化，反向传播训练整个网络的参数。

本文设计的低层网络的目标是优先考虑召回率的情况下尽可能多的识别出潜在的命名实体。因为有后续网络的存在，不必担心这里松弛算法对最终层叠模型整体准确率的影响。为了达到这个目标，在训练低层网络时对损失函数进行优化，公式4-1调整为：

$$\begin{aligned} s(\boldsymbol{X},\boldsymbol{y})&=\sum_{i=0}^{n}A_{y_i,y_{i+1}}+\sum_{i=1}^{n}r(P_{i,y_i})\\ r\left(P_{i,y_i}\right)&=\begin{cases}\lambda\cdot P_{i,y_i} & tag_i=''O''\\ P_{i,y_i} & \text{else}\end{cases}\end{aligned} \tag{4-5}$$

公式中 λ 是惩罚因子，取值在 0 到 1 之间。这样调整的含义是在计算标签序列路径得分时，当真实标记为“O”（不是命名实体）时，乘上一个惩罚系

数算入标签序列路径得分。因为现实中的语料集中，我们往往关注的命名实体相对整个数据集而言较小，使得模型偏向预测非命名实体标签，使得模型的损失值更小。可是这种偏好与我们希望找出所有命名实体的目标相违背。这里的惩罚因子使得真实标记为“O”的训练样例的权重降低，而真实标记不为“O”，即标签属于任何一类命名实体样例的权重相较而言得到提高。这样在计算 loss 时，真实标签为“B-PER”、“I-PER”、“B-ORG”等字符的预测结果对于网络训练影响更大。

如上文图4-1为例的嵌套命名实体问题将得到一定程度的改善，因为这样的优化改进会使得模型识别字符为“O”的意愿降低，能够被识别成更长的命名实体将会得到更高的分数。在图 4-1 的例子中，标签序列中含有的命名实体“上海农商行”（B-ORG，I-ORG，I-ORG，I-ORG，I-ORG）更容易得到比“上海”（B-LOC，I-LOC，O，O，O）更高的分数。

除此以外，训练好模型后对文本数据进行识别时，解码过程也可以使用惩罚因子提高低层模型识别的召回率。如 2.3 节所述，条件随机场是通过维特比算法进行解码的。维特比算法的核心思路是：（1）如果概率最大的路径经过概率图中的某点，则从开始点到该点的任意子路径也一定是从开始到该点路径中概率最大的，因为若该结论不成立的话则一定有一条概率更大的全局路径与题设矛盾。（2）假定第 i 时刻有 k 个状态，从开始到 i 时刻的 k 个状态有 k 条最短路径，而最终的最短路径必然经过其中的一条。（3）根据上述性质，在计算第 i+1 状态的最短路径时，只需要考虑从开始到当前的 k 个状态值的最短路径和当前状态值到第 i+1 状态值的最短路径即可。

为了使得低层网络在解码序列时更倾向于输出命名实体标签而不是输出非命名实体标签，本文在低层网络解码计算过程中将所有字符在属于标签“O”，即不为命名实体的概率乘以惩罚因子 μ，μ 取值在 0 到 1 之间，使得含有更多命名实体标签的序列更容易得到高的分数，被作为结果输出，解码算法流程见算法4.1。

这样低层命名实体识别模型的设计即可以有效解决嵌套命名实体的问题，对于文本做了有意义的切割，减小了问题的复杂度，与此同时也尽量避免了有效信息的丢失。而低层网络训练测试时产生的误差也可以靠下一章节的高层神经网络过滤，避免错误扩散到最终的结果中。

Algorithm 4.1 低层网络维特比动态规划解码算法

1: 输入：待预测文本字向量矩阵经模型计算后得出的发射概率矩阵
2: 输出：待预测文本的预测标签序列
3: 将当前发射概率矩阵标签为非命名实体的概率乘以惩罚因子 μ
4: 创建一个序列长度 × 标签个数的零矩阵 S 记录动态规划各子路径得分
5: 创建一个序列长度 × 标签个数的矩阵 B 记录 S 矩阵中的路径线索，用当前结点的上一个节点来记录路径
6: 从第一个节点到最后一个节点遍历：
通过发射概率矩阵和转移概率矩阵，在 S 矩阵中计算从开始点到每一个节点对应的每个标签的最大概率路径，同时在 B 中记录路径
7: 在 S 的最后一列找出最大概率路径的分值，并用回溯法遍历 B 矩阵，找出该最大概率路径的标签序列作为最终输出

4.3.3 基于卷积神经网络的高层网络构建

高层网络模型接受低层网络模型的输出，将接收的文本进一步处理，关键是要找准命名实体的边界。这里在训练高层 BiLSTM-CRF 模型时，在字符分布式嵌入后加入卷积神经网络模型 (Convolutional Neural Networks，CNN) 提高高层模型判断命名实体边界的能力。许多文献对该方向进行了研究 [26, 53]，不论是英文语料 [66] 还是中文语料 [67]，该方法都有一定提高效果。CNN 的目的是对字符分布式表示的特征进行更细致的特征抽取，使得局部信息产生更有效的连接，对实体边界的识别更为准确。

在本文模型设计中，字符向量表示方面采用第三章介绍的改进的中文字符向量改进方法训练。对于输入字符向量的分布式特征，用卷积层（convolution）进行局部信息的特征抽取，通过池化层（pooling）对特征进行降维压缩，减小过拟合风险的同时提高模型的容错性。最终将抽取出的特征与原始字符向量拼接一同传递给双向长短时记忆网络层。得到每个字符属于不同标签时的一元概率得分，最终送入条件随机场层，该层利用转移概率矩阵，计算当前情况下概率得分最大的路径，使用该路径对输入文本进行标注。高层模型结构图如4–3所示。

低层模型和高层模型训练完后，在测试使用阶段，输入的是一串新的文本，通过保存后的低层模型计算得出“字符 -标签”的概率得分矩阵（即发射概率矩阵）。在发射概率矩阵和转移概率矩阵的基础上通过维特比解码的方法动态规划地计算出概率得分值最大的路径作为最终的预测标签序列。将低层网络中识别出的合法命名实体序列作为粗粒度的识别结果，传入高层网络。对于高

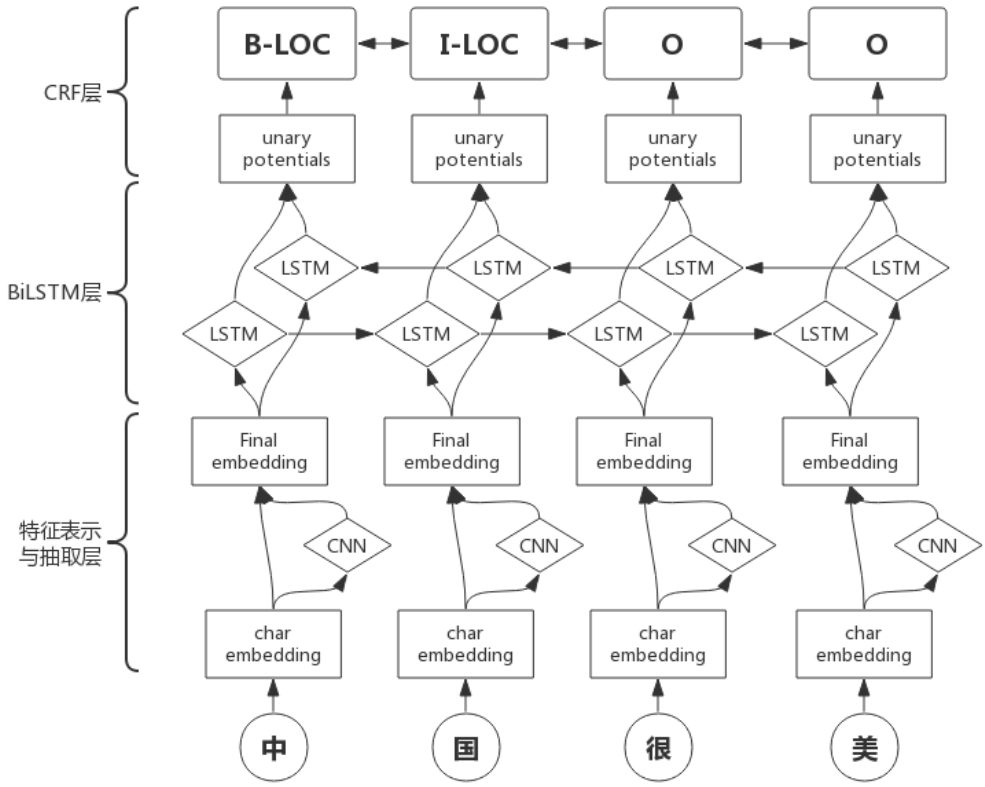

图 4-3: 高层网络模型结构图

层网络而言，维特比解码方式不变来保证结果的准确度和边界的严谨程度。高层网络接受低层网络输入的文本进行预测，预测结果有以下几种情况：

（1）高层网络识别的标签序列与输入实体及序列一致，认可该命名实体标签结果并输出。

（2）高层网络识别出单个实体，与输入实体标签序列不一致，以高层精确识别的实体作为最终的输出结果。

（3）高层网络识别出多个实体，分别将多个实体重新作为输入传入高层网络，重复以上步骤。

整体层叠深度神经网络命名实体识别模型流程如图4-4。

这样层叠的模型相比单层模型有许多优点，通过低层网络的将文本合理分割成更小的文本块，舍弃了没那么关注的非命名实体干扰信息，避免了句子长度长成分复杂、命名实体前后文错误关联等问题的影响。而通过惩罚因子 λ、μ 提高了低层网络的召回率，确保在经过低层网络时有效信息不会被错误放弃，同时也可以很好地提升对嵌套命名实体识别的识别率。高层网络接受低层网络的输出，防止了低层网络误差的扩散。简单来说，对于同样一个序列标注任务

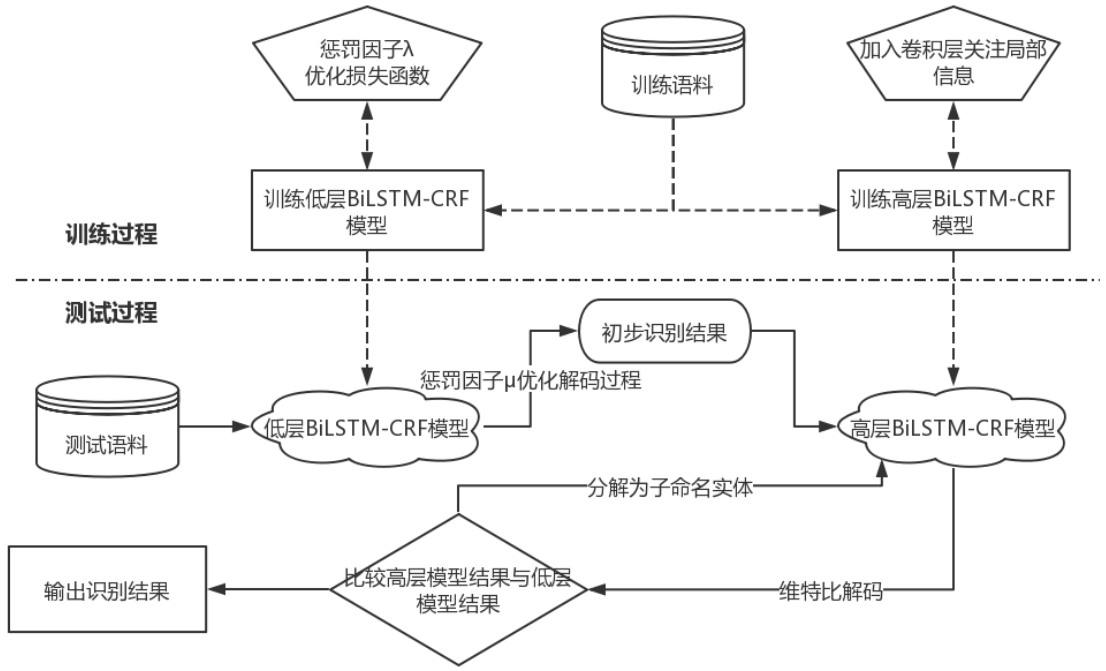

图 4-4: 层叠深度神经网络模型

而言，若序列长度为 n，则潜在的标注序列有 k^n 种，其中 k 是标签的种类数。那么当序列长度 n 越大时，潜在的标签序列数将以 k 的幂次速度增加，显然准确预测标签序列的难度就会加大。利用本文层叠的深度神经网络模型的方法，在损失一些不够重要的信息的代价下，较好的控制了问题的规模，降低了识别的难度。

本文的层叠深度神经网络命名实体识别模型在文本复杂，行文断句不规范的场景下有着更突出的优势。而层叠串行的结构，只需要分别训练两个深度神经网络模型，在复杂度上来看并没有增加太多，适合于实际的工程应用。对现今互联网信息时代，网络文本规范不一，却又蕴含海量信息的场景下本模型有着较好的适应性。实验表明（见下一节）本模型在标准数据集和实际工程应用中有着良好的效果。

4.4 实验

4.4.1 实验环境与设置

本章实验环境与设置与第三章实验相同，实验设备主要有一台主频 2.20GHz，16GB 内存，处理器四核 i5 的个人计算机，操作系统为 Windows；还

有两台主频为 2.2GHz、处理器为 40 核的 Intel Xeon E5-2630，内存 128G，操作系统为 Ubun16.04 的服务器。深度学习模型同样采用 Tensorflow 的编程框架，版本为 1.2.0。高层模型中为了更准确识别边界，采用了卷积神经网络对字符向量做特征抽取。有一个卷积层和一个池化层构成，卷积层的卷积核大小为 3×3，共 64 个，池化层窗口大小为 2×2，水平和垂直步长为 1。

本实验中使用了两个数据集，一个数据集是和第三章实验的数据相同的人民日报语料集，另一个数据集是本文作者实际工程项目中采集的企业风险识别数据，下面简要介绍一下该数据集。

作者在某项目中，为了通过互联网大数据技术识别出风险产品和存在风险的公司，通过爬虫技术爬取了大量网页信息。这些网站包括一系列质量监督的政府网站、生活类新闻门户网站还有利用关键词和搜索引擎查询的网站等等。项目的目标是识别出风险信息，并关联到相关风险公司以及相关组织。为了识别出这些风险公司和相关组织，采用基于深度学习的命名实体识别模型对组织名进行识别。这类数据包含很多复杂命名实体，反映了现实工程应用中对中文命名实体识别系统的需求。部分数据如图 4-5所示：

id	title	content	crawler_ti
1	省市场监管局关于注销江苏国安机动车检测有限公司检验检	根据《中华人民共和国行政许可法》及《检验检测机构资质认定管理办法》（原国家	2019-3-1
2	省市场监管局关于注销苏州市自动化仪器仪表研究所有限公	根据《中华人民共和国行政许可法》及《检验检测机构资质认定管理办法》（原国家	2019-3-1
3	江苏举办首届电梯检验人员职业技能竞赛	为贯彻落实国务院办公厅、省政府办公厅关于加强电梯质量安全工作有关文件精神，	2019-3-1
4	江苏省质监局通报冲锋衣、防晒服产品质量监督抽查结果	10月24日，江苏省质监局通报冲锋衣、防晒服产品质量监督抽查结果。抽查显示，冲	2019-3-1
5	徐州金源臭氧设备有限公司召回KJ-101型空气净化器	江苏省质监局根据市场购检线索，在风险评估的基础上，要求徐州金源臭氧设备有限	2019-3-1
6	江苏省质监局发布食品包装产品监督抽查结果	现如今，我们在市面上看到的食品包装多种多样，有塑料的、纸质的，还有我们直接	2019-3-1
7	马秋林副省长到省质监局进行工作调研	9月6日上午，马秋林副省长到省质监局进行工作调研。省政府副秘书长张乐夫，省质	2019-3-1
8	常州洪都电动车有限公司召回洪都牌（TDT06Z-1433-01）	江苏省质监局根据市场购检线索，在风险评估的基础上，要求常州洪都电动车有限公	2019-3-1
9	省质监局集中约谈20家肥料生产企业	7月12日，省质监局集中约谈20家肥料生产企业，督促被约谈企业正视问题，落实质	2019-3-1
10	我省2家企业入选国家重点用能行业能效"领跑者"企业名单	日前，工信部、国家市场监管总局对2017年高耗能行业能效"领跑者"予以公告，涉及	2019-3-1

图 4-5: 数据集构建

这些数据爬取于互联网，在断句、行文、结构上难免有很多不规范之处。该数据集公司企业名多，嵌套命名实体多，符合第四章提出的复杂命名实体特征。本实验中对该数据集的部分数据进行了标注，来验证本章层叠命名实体识别模型的效果。共标注了 2145 个组织名实体，396 个人名实体和 816 个地名实体，通过这些在实际工程项目中出现的互联网新闻复杂信息，对本章提出的层叠命名实体识别模型进行效果验证。

4.4.2 实验及结果分析

本实验分别用训练集训练低层网络和高层网络，训练低层网络时涉及到训练时损失函数的调整以及解码时权重的调整，依据模型在测试集上的表现设置相应参数。训练出模型后，对于测试数据先经由低层网络初步识别，得出粗粒度的实体，再传入高层网络得到最终结果。

表 4-4: 原单层模型识别结果

单层模型输入：	陷入三名高管“不能履职”风波的南京银行，公告中的一句“不涉及当前本行及鑫元基金的业务，本行及鑫元基金经营管理一切正常。”恐怕并不能消除公众的忧虑。
人名实体：	[]
地名实体：	[]
组织名实体：	[“南京银行”]

如表4-4所示，原单层模型中往往由于句子长度长，维特比解码时发生错误，对于命名实体的识别不够精准。组织命名实体“鑫元基金”被模型遗漏。

表 4-5: 层叠模型低层网络识别结果

低层模型输入：	陷入三名高管“不能履职”风波的南京银行，公告中的一句“不涉及当前本行及鑫元基金的业务，本行及鑫元基金经营管理一切正常。”恐怕并不能消除公众的忧虑。
人名实体：	[]
地名实体：	[]
组织名实体：	[“南京银行”，“鑫元基金”，“鑫元基金经营管理”]

表 4-6: 层叠模型高层网络识别结果

高层模型输入：	鑫元基金经营管理
人名实体：	[]
地名实体：	[]
组织名实体：	[“鑫元基金”]

表4-5和表4-6是该文本通过本文提出的层叠深度神经网络模型输出的结果。通过本章设计的层叠模型识别后，低层模型的训练方法提高了识别的召回率，尽量避免信息的丢失。而这样的过程也对文本进行了合理的切割，使得高

层识别较短的序列，提高高层识别的成功率。而高层神经网络加入了卷积池化层，对于实体边界有着更准确的分割，也过滤了低层模型识别的错误，使得低层识别的不准确信息不会扩散到最终的结果中。

表 4-7: 层叠模型低层网络识别结果

低层模型输入：	谋求上市应该是上海农商行自带的话题光环，且目前已经有了实质性进展。
人名实体：	[]
地名实体：	[]
组织名实体：	[“上海农商行”]

表 4-8: 层叠模型高层网络识别结果

高层模型输入：	上海农商行
人名实体：	[]
地名实体：	[]
组织名实体：	[“上海农商行”]

对于表4-1中为代表的嵌套命名实体问题，原单层模型由于标签权重判断不清，造成识别错误。如表4-7和表4-8所示，本章设计的层叠网络能够较好的解决这一问题。

用标准数据集人民日报语料对本章提出的层叠深度神经网络命名实体识别模型进行实验，字符嵌入分别使用直接使用 word2vec 训练出的字向量的简单表示和第三章提出的改进的中文字符级表示进行实验比较，结果如表4-9。

从实验结果可以看出，不论是采用简单 word2vec 训练出的字向量直接嵌入，还是使用本文第三章提出的改进方法进行嵌入，层叠模型较单层模型都有更好的表现。层叠模型相比于单层模型在组织名命名实体识别效果上提高了很多，在人名实体识别上有一定优势，在地名实体识别上与单层模型效果相当。这也与本章层叠模型设计的方法就是为了解决长度长、嵌套、混淆的命名实体，通过层叠网络分层处理，提高对复杂中文命名实体的识别能力。实验表明在人民日报数据集上，层叠深度神经网络模型相比单层模型准确率提高了 1.41%，召回率提高了 0.98%，F 值提高了 1.19%。

在企业风险识别数据集上对层叠的深度神经网络命名实体识别模型进行实验，结果如表4-10所示。

字符嵌入方法都按第三章改进后的中文字符表示进行嵌入，可以发现在企业风险识别数据集上，层叠模型的整体效果好于单层模型。除了人名实体识别

表 4-9: 人民日报数据集实验结果

		单层模型		层叠模型	
		简单表示	改进表示	简单表示	改进表示
LOC 实体	准确率	0.9324	0.9342	**0.9402**	0.9360
	召回率	0.8825	**0.9027**	0.8846	0.8992
	F 值	0.9068	**0.9182**	0.9115	0.9172
ORG 实体	准确率	0.8231	0.8502	0.8633	**0.8709**
	召回率	0.8287	0.8527	**0.8685**	0.8618
	F 值	0.8259	0.8515	0.8659	**0.8663**
PER 实体	准确率	0.8877	0.8906	0.8919	**0.9094**
	召回率	0.8211	**0.8700**	0.8528	0.8604
	F 值	0.8531	0.8802	0.8719	**0.8842**
全部实体	准确率	0.8937	0.9017	0.9075	**0.9132**
	召回率	0.8513	**0.8815**	0.8710	0.8787
	F 值	0.8720	0.8915	0.8888	**0.8956**

效果接近外，地名实体和组织名实体都有较大幅度地提高。这也与企业风险识别数据来源于互联网风险企业信息，包含复杂中文命名实体较多有关。对所有实体而言，层叠模型相比于单层模型准确率提高 3.83%，召回率提高 4.00%，F 值提高 3.95%。

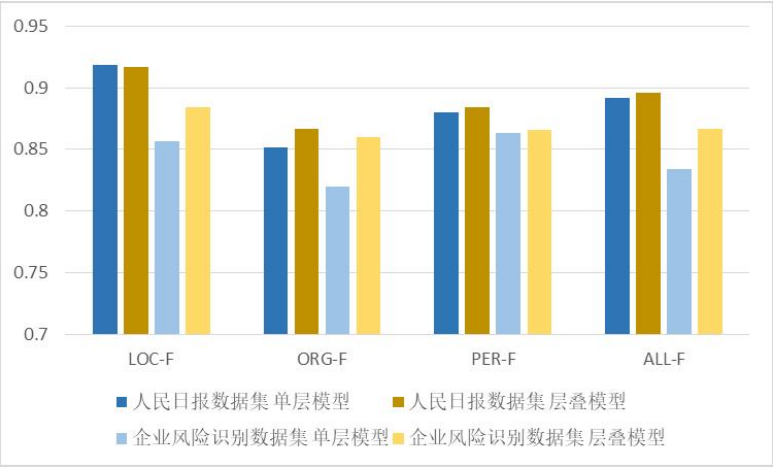

图 4-6: 各类别实体 F 值比较

将两个数据集各类实体识别的 F 值进行比较，如图 4-6。可以看出本章提

表 4-10: 企业风险识别数据实验结果

		单层模型	层叠模型
LOC 实体	准确率	0.8641	**0.8945**
	召回率	0.8493	**0.8730**
	F 值	0.8567	**0.8838**
ORG 实体	准确率	0.8298	**0.8695**
	召回率	0.8101	**0.8506**
	F 值	0.8197	**0.8600**
PER 实体	准确率	**0.8734**	0.8699
	召回率	0.8535	**0.8611**
	F 值	0.8633	**0.8654**
全部实体	准确率	0.8433	**0.8756**
	召回率	0.8241	**0.8571**
	F 值	0.8337	**0.8666**

出的层叠神经网络模型整体上对单层模型有优势，特别是对组织名实体的识别有较好的效果。不过层叠模型涉及到两层网络的训练，相比于单层网络而言，复杂度稍大，训练时间较长。如果数据集不具有本章 4.3.1 介绍的复杂命名实体特点的话，层叠模型的优势会减小。因而，数据杂乱、嵌套命名实体多、断句不清等情况下，可以优先使用该层叠模型。

4.5 本章小结

本章探究了中文命名实体识别中的复杂命名实体识别问题。本章首先分析了复杂命名实体的定义，总结了嵌套命名实体、长度长文本、上下文错误关联三类影响识别效果的因素。之后综述了前人对于复杂命名实体和多层次命名实体识别模型结构的研究，基于这些工作，叙述了本章层叠模型的构建思路。最终，针对复杂中文命名实体问题设计了一个层叠的深度神经网络模型来分层次识别。

本模型低层网络的目标是过滤无关命名实体识别的信息，通过优化损失函数和解码方法，提高潜在命名实体信息的召回率，同时切割了文本降低高层模型识别的难度。识别出粗粒度的命名实体后传入高层网络，高层网络在

BiLSTM-CRF 模型的基础上加入卷积池化层来提高对命名实体边界的判别能力。最终将文本信息划分为不可分割的实体后作为最终结果输出。实验表明，本层叠模型提高了中文命名实体识别的效果，特别是对于组织名实体的识别有了较大的提升。

第五章　总结与展望

5.1　工作总结

命名实体识别是自然语言处理中的基础任务之一，其目的是要从一段文本中识别出人名、地名、组织名等所要求的命名实体，提取出这些实体后可以为后续的自然语言处理任务服务，有着广阔的应用场景。目前中文命名实体识别任务取得了很多进展，尤其是近来基于深度学习的中文命名实体识别技术取得了一系列成果。但是目前的中文命名实体识别方法还存在着一下问题：（1）中文字符向量与分词技术间的矛盾，引入分词技术的话，识别结果某种程度上会局限于词典的质量，在部分复杂情况下效果一般；而不引入分词信息则中文字向量一字多义、一字多态、表达稀疏等问题制约着中文命名实体识别技术的效果。（2）现实中许多文本信息，特别是在互联网海量杂乱文本中，目前的中文命名实体识别模型没办法合理有效的对文本分割，从而处理嵌套命名实体、混淆命名实体等复杂命名实体。为了解处理复杂中文命名实体识别问题，本文主要做了以下工作：

1）本文在针对目前命名实体任务中中文字符向量表示所存在的问题，提出了改进的中文字符级特征表示方法。中文字符较少，向量空间稀疏，而直接使用分词信息会影响部分命名实体的识别效果。基于这些原因，本文首先对于中文字符向量基于位置信息进行优化，根据中文语言构词的逻辑，在中文字符向量中叠加周边字符的信息，以解决一字多义，向量稀疏的问题。而对于中文字符信息量不足的问题，本文提出了基于主题信息的中文字符向量构造方法。通过 LDA 主题模型，利用大量外部无监督语料，训练字符在不同主题上的概率，利用主题模型的全局信息，丰富 word2vec 训练字向量时的局部信息。实验表明，这两个通过不同角度对中文字符向量的优化方法有效地提高了基于深度学习的中文命名实体识别效果。

2）对于实际工程应用出现的复杂命名实体等问题，本文提出了层叠神经网络来解决这一系列问题。该层叠结构由低层高层两个模型组合而成。为了解决复杂命名实体中嵌套命名实体、关联命名实体、文本长度长等识别中的问题，

模型的低层模型训练时修改了损失函数的计算方法和解码时概率得分的计算方法，使得低层模型更偏向识别出命名实体标签，提高命名实体识别的召回率，尽量避免低层模型切割文本时信息的丢失。高层模型方面，为了使得其对于实体边界更好的识别，本文在中文字符向量嵌入的基础上加入卷积神经网络来抽取局部信息，识别出更精确的实体名。通过在标准数据集和工程应用中构造出的复杂中文命名实体数据上的实验表明，本文提出的层叠深度学习命名实体识别模型取得了较好的效果。

5.2 不足与展望

本文还有许多不足之处与需要改进的问题，需要在未来的工作中进一步去探究：

1）在中文命名实体识别问题中，首先在基于位置信息优化中文字符向量的过程中，目前采取的方法较为简单朴素，中文遣词造句的逻辑非常复杂，可以设计更复杂的模型来拟合字向量语义表示与周围信息的关联。

2）如何丰富字符向量的信息也是研究要点，字符向量信息少是制约目前中文命名实体识别模型效果的关键。今后会常识更多的方法，如基于词对的主题模型等方法来丰富分布式表示的信息量，提高字符向量的表示能力。

3）在实际工程应用中对于并行化，运算效率要求更高，第四章层叠深度学习命名实体识别模型中，引入了卷积神经网络，增加了模型的复杂度，使得模型训练预测的速度较慢。同时长短时记忆网络模型不利于并行化，如何将模型改进，让模型在实际应用中满足速度上的要求是下一个阶段研究的重点。

致　　谢

远东大道两侧的草地又渐渐换上了绿色的新衣，系楼前的樱花又迎来了盛开。转眼三年研究生生涯行将结束，我也要对八年南大生活学习说一声再见了。

在南大，我遇到了许许多多优秀的老师和同学。感谢恩师王崇骏教授多年来对我的指导和关心，您对学术的严谨，对工作的负责，对生活的热爱都一点点地影响着我，让我有勇气与能力面对一个又一个新的挑战。感谢 IIP 工作组的小伙伴们，三年来，我们一起学习成长，一起运动娱乐，点点滴滴都将是我青春的珍贵回忆。完成本论文工作过程中，感谢你们与我一起交流讨论，在每一个细节给我的帮助，祝你们今后的学习和工作顺利。感谢家人多年学生生涯的支持，让我可以无所顾虑的追求自己的理想，希望今后自己可以给你们带来幸福快乐。

最后再一次对母校南京大学说一声感谢，南大真诚、平实、儒雅、担当的内在人格将是我一生的财富。祝福从这里离开的、留在这里的、未来来到这里的人们一切都好。

参考文献

[1] MAHANTA H. A study on the approaches of developing a Named Entity Recognition tool[J]. International Journal of Research in Engineering and Technology, 2013, 2.

[2] RAU L F. Extracting company names from text[C] // [1991] Proceedings. The Seventh IEEE Conference on Artificial Intelligence Application: Vol 1. 1991: 29 – 32.

[3] KIM J-H, WOODLAND P C. A rule-based named entity recognition system for speech input[C] // Sixth International Conference on Spoken Language Processing. 2000.

[4] HANISCH D, FUNDEL K, MEVISSEN H-T, et al. ProMiner: rule-based protein and gene entity recognition[J]. BMC bioinformatics, 2005, 6(1): S14.

[5] QUIMBAYA A P, MÚNERA A S, RIVERA R A G, et al. Named entity recognition over electronic health records through a combined dictionary-based approach[J]. Procedia Computer Science, 2016, 100: 55 – 61.

[6] NADEAU D, SEKINE S. A survey of named entity recognition and classification[J]. Lingvisticae Investigationes, 2007, 30(1): 3 – 26.

[7] SETTLES B. Biomedical named entity recognition using conditional random fields and rich feature sets[C] // Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP). 2004.

[8] ZHU J, UREN V, MOTTA E. ESpotter: Adaptive named entity recognition for web browsing[C] // Biennial Conference on Professional Knowledge Management/Wissensmanagement. 2005: 518 – 529.

[9] EDDY S R. Hidden markov models[J]. Current opinion in structural biology, 1996, 6(3) : 361 – 365.

[10] QUINLAN J R. Induction of decision trees[J]. Machine learning, 1986, 1(1) : 81 – 106.

[11] KAPUR J N. Maximum-entropy models in science and engineering[M]. [S.l.] : John Wiley & Sons, 1989.

[12] HEARST M A, DUMAIS S T, OSUNA E, et al. Support vector machines[J]. IEEE Intelligent Systems and their applications, 1998, 13(4) : 18 – 28.

[13] LAFFERTY J, MCCALLUM A, PEREIRA F C. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[J], 2001.

[14] BRIN S. Extracting patterns and relations from the world wide web[C] // International Workshop on The World Wide Web and Databases. 1998 : 172 – 183.

[15] JI H, GRISHMAN R. Knowledge base population: Successful approaches and challenges[C] // Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1. 2011 : 1148 – 1158.

[16] MCNAMEE P, MAYFIELD J. Entity extraction without language-specific resources[C] // proceedings of the 6th conference on Natural language learning-Volume 20. 2002 : 1 – 4.

[17] ISOZAKI H, KAZAWA H. Efficient support vector classifiers for named entity recognition[C] // Proceedings of the 19th international conference on Computational linguistics-Volume 1. 2002 : 1 – 7.

[18] LI Y, BONTCHEVA K, CUNNINGHAM H. SVM based learning system for information extraction[C] // International Workshop on Deterministic and Statistical Methods in Machine Learning. 2004 : 319 – 339.

[19] MCCALLUM A, LI W. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons[C] // Proceedings

of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4. 2003 : 188 – 191.

[20] 何炎祥, 罗楚威, 胡彬尧, et al. 基于 CRF 和规则相结合的地理命名实体识别方法 [J]. 计算机应用与软件, 2015, 32(1) : 179r185 – 202.

[21] 张素香. 信息抽取中关键技术的研究 [D]. [S.l.] : 北京邮电大学, 2007.

[22] COLLOBERT R, WESTON J, BOTTOU L, et al. Natural language processing (almost) from scratch[J]. Journal of machine learning research, 2011, 12(Aug) : 2493 – 2537.

[23] NGUYEN T H, SIL A, DINU G, et al. Toward mention detection robustness with recurrent neural networks[J]. arXiv preprint arXiv:1602.07749, 2016.

[24] YAO L, LIU H, LIU Y, et al. Biomedical named entity recognition based on deep neutral network[J]. Int. J. Hybrid Inf. Technol, 2015, 8(8) : 279 – 288.

[25] KURU O, CAN O A, YURET D. Charner: Character-level named entity recognition[C] // Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. 2016 : 911 – 921.

[26] MA X, HOVY E. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF[C] // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) : Vol 1. 2016 : 1064 – 1074.

[27] YANG J, ZHANG Y, DONG F. Neural Reranking for Named Entity Recognition[C] // Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017. 2017 : 784 – 792.

[28] LAMPLE G, BALLESTEROS M, SUBRAMANIAN S, et al. Neural Architectures for Named Entity Recognition[C] // Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016 : 260 – 270.

[29] HUANG Z, XU W, YU K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv preprint arXiv:1508.01991, 2015.

[30] STRUBELL E, VERGA P, BELANGER D, et al. Fast and Accurate Entity Recognition with Iterated Dilated Convolutions[C] // Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017 : 2670 – 2680.

[31] JANSSON P, LIU S. Distributed representation, lda topic modelling and deep learning for emerging named entity recognition from social media[C] // Proceedings of the 3rd Workshop on Noisy User-generated Text. 2017 : 154 – 159.

[32] WU Y, JIANG M, LEI J, et al. Named entity recognition in Chinese clinical text using deep neural network[J]. Studies in health technology and informatics, 2015, 216 : 624.

[33] ZHOU P, ZHENG S, XU J, et al. Joint extraction of multiple relations and entities by using a hybrid neural network[G] // Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. [S.l.] : Springer, 2017 : 135 – 146.

[34] KATIYAR A, CARDIE C. Nested named entity recognition revisited[C] // Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers) : Vol 1. 2018 : 861 – 871.

[35] JU M, MIWA M, ANANIADOU S. A neural layered model for nested named entity recognition[C] // Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers) : Vol 1. 2018 : 1446 – 1459.

[36] SHEN Y, YUN H, LIPTON Z, et al. Deep Active Learning for Named Entity Recognition[C] // Proceedings of the 2nd Workshop on Representation Learning for NLP. 2017 : 252 – 256.

[37] AKBIK A, BLYTHE D, VOLLGRAF R. Contextual string embeddings for sequence labeling[C] // Proceedings of the 27th International Conference on Computational Linguistics. 2018 : 1638 – 1649.

[38] BENGIO Y, DUCHARME R, VINCENT P, et al. A neural probabilistic language model[J]. Journal of machine learning research, 2003, 3(Feb): 1137–1155.

[39] MIKOLOV T, CHEN K, CORRADO G S, et al. Computing numeric representations of words in a high-dimensional space[J], 2015.

[40] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C] // Advances in neural information processing systems. 2013: 3111–3119.

[41] 马远浩, 曾卫明, 石玉虎, et al. 基于加权词向量和 LSTM-CNN 的微博文本分类研究 [J]. 现代计算机 (专业版), 2018(25): 5.

[42] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.

[43] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735–1780.

[44] LI J, SUN A, JOTY S. SegBot: A Generic Neural Text Segmentation Model with Pointer Network.[C] // IJCAI. 2018: 4166–4172.

[45] WEI Q, CHEN T, XU R, et al. Disease named entity recognition by combining conditional random fields and bidirectional recurrent neural networks[J]. Database, 2016, 2016.

[46] DEVLIN J, CHANG M-W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[47] ZHANG Y, YANG J. Chinese NER Using Lattice LSTM[C] // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers): Vol 1. 2018: 1554–1564.

[48] ZHAO D, HUANG J, JIA Y. Chinese Name Entity Recognition Using Highway-LSTM-CRF[C] // Proceedings of the 2018 International Conference on Algorithms, Computing and Artificial Intelligence. 2018: 64.

[49] 林泽斐, 欧石燕. 多特征融合的中文命名实体链接方法研究 [J]. 情报学报, 2019, 38(1): 68–78.

[50] 王超, 王峥. 基于改进分词标注集的中文微博命名实体识别方法 [J]. 计算机与数字工程, 2019(1): 44.

[51] GRAVES A, SCHMIDHUBER J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures[J]. Neural Networks, 2005, 18(5-6): 602–610.

[52] KIM Y. Convolutional Neural Networks for Sentence Classification[C] // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014: 1746–1751.

[53] 康宁. 基于主题模型和卷积神经网络的命名实体识别研究 [D]. [S.l.]: 南京大学, 2018.

[54] 龚凌晖. 中文命名实体识别与歧义消解研究 [D]. [S.l.]: 复旦大学, 2011.

[55] GIRIJA S S. Tensorflow: Large-scale machine learning on heterogeneous distributed systems[J], .

[56] KINGMA D P, BA J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.

[57] GLOROT X, BENGIO Y. Understanding the difficulty of training deep feedforward neural networks[C] // Proceedings of the thirteenth international conference on artificial intelligence and statistics. 2010: 249–256.

[58] 周俊生, 戴新宇, 尹存燕, et al. 基于层叠条件随机场模型的中文机构名自动识别 [J]. 电子学报, 2006, 34(5): 804–809.

[59] XING Y, ZHU Y, ZHANG K, et al. Named Entity Recognition Among Chinese MicroBlog Based on Cascaded CRF[C] // 2018 International Conference on Audio, Language and Image Processing (ICALIP). 2018: 28–34.

[60] GONG Y, LUO X, ZHU Y, et al. Deep Cascade Multi-task Learning for Slot Filling in Online Shopping Assistant[J], 2019.

[61] JU M, MIWA M, ANANIADOU S. A neural layered model for nested named entity recognition[C] // Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers): Vol 1. 2018: 1446–1459.

[62] SOHRAB M G, MIWA M. Deep Exhaustive Model for Nested Named Entity Recognition[C] // Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 2843–2849.

[63] KATIYAR A, CARDIE C. Nested named entity recognition revisited[C] // Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers): Vol 1. 2018: 861–871.

[64] 贾大宇. 基于混合层叠模型的命名实体识别研究 [D]. [S.l.]: 东北大学, 2016.

[65] 李雁群, 何云琪, 钱龙华, et al. 中文嵌套命名实体识别语料库的构建 [J]. 中文信息学报, , 32(8): 19–26.

[66] ZHAI Z, NGUYEN D Q, VERSPOOR K. Comparing CNN and LSTM character-level embeddings in BiLSTM-CRF models for chemical and disease named entity recognition[C] // Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis. 2018: 38–43.

[67] JIA Y, XU X. Chinese Named Entity Recognition Based on CNN-BiLSTM-CRF[C] // 2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS). 2018: 1–4.

附录

攻读硕士学位期间完成的学术成果

1. QiaoY, **GU Y**, Wu J and etc, A Truthful Profit-oriented Mechanism for Mobile Crowdsensing. In Proceedings of ISPA2018, Melbourne, Australia, Dec. 11-13, 2018:64-71.

攻读硕士学位期间参与的科研课题

1. 科技部重点研发计划“跨时空异构数据的结构化描述和语义协同”，编号2016YFB1001102（课题年限2016年6月—），负责数据链接、语义融合的研究。
2. 江苏省质量和标准化研究院资助项目“缺陷产品案源分析系统”（课题年限2016年12月—2018年12月），负责命名实体识别、智能预警研究。
3. 江苏省质量和标准化研究院资助项目“电商平台缺陷消费品信息采集分析”（课题年限2018年12月—2019年12月），负责情感分析、缺陷产品召回研究。
4. 国家自然科学基金“复杂环境下众包机制设计问题研究”，编号61876080（课题年限2019年1月—2022年12月），负责众包、机制设计研究。