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

插图清单

附表清单

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）对命名实体方法的发展脉络进行梳理，通过文献综述等方式比较

了利用深度学习解决中文命名实体识别时，各模块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，对该模型详细介绍并描述了本文对该模型的实现。

（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，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。

（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。

（4）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 2-1 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
B-PER	人名开头
I-PER	人名内部
B-LOC	地名开头
I-LOC	地名内部
B-ORG	组织机构名开头
I-ORG	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。

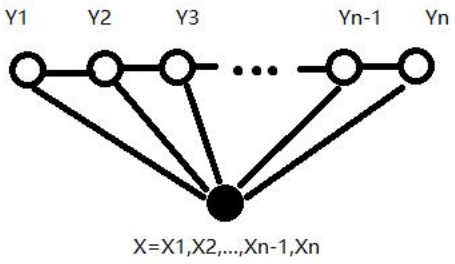

图 2-1: 线性链条件随机场

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/。”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

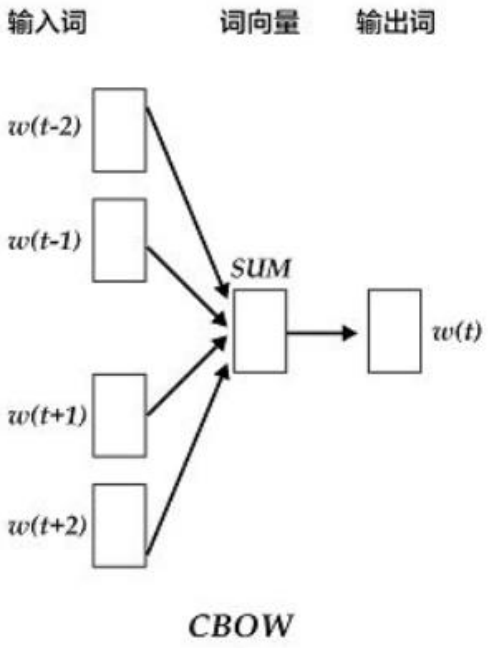

图 2-2: CBOW 模型

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，反而输出的是上下文的词向量，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。

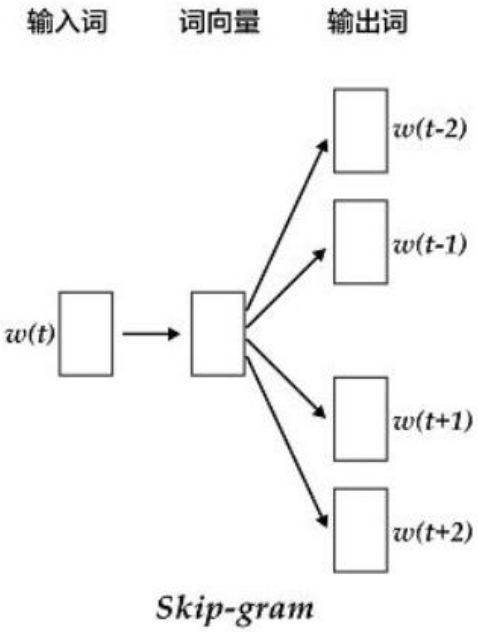

图 2-3: Skip-gram 模型

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

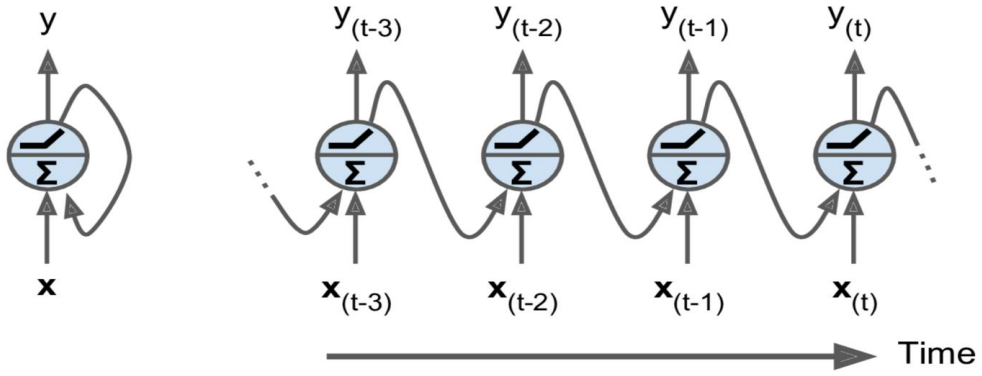

图 2-4: 简单的 RNN 模型

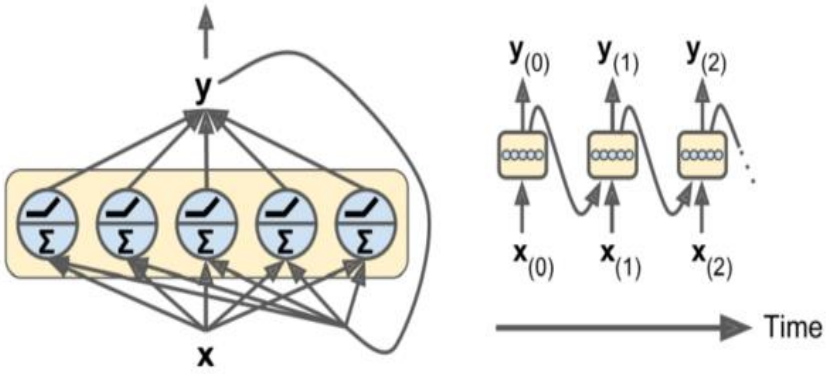

图 2-5: 一个 RNN 层和展开结构

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。

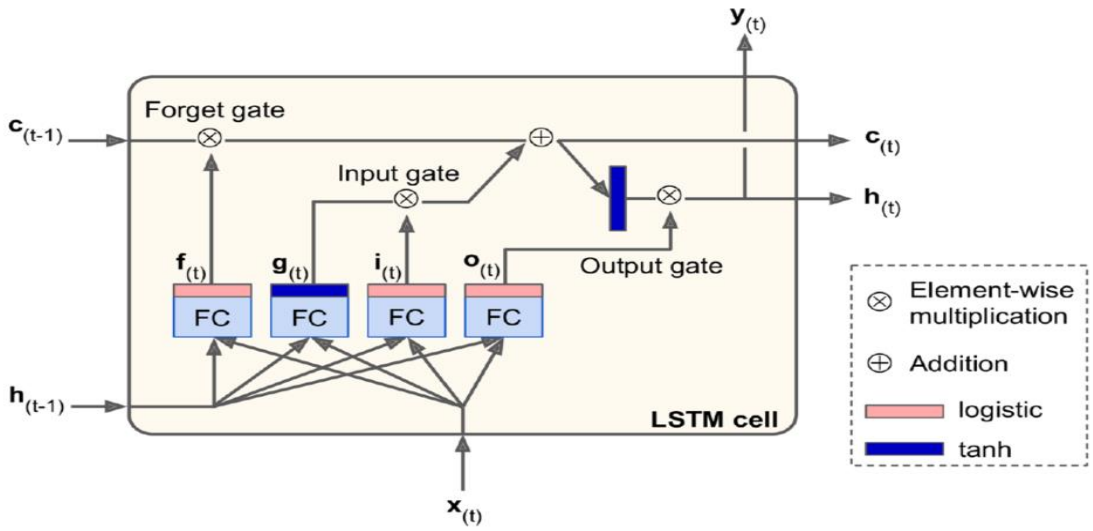

图 2-6: LSTM 单元

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。

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）时会产生一个严重的问题，误差梯度随时间长

度成指数级别增长式得消实。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，直到数值被与之过滤掉。因此含有 LSTM 的单元的 RNN 可以被有效训练并记住长时间的信息。

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，同事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。

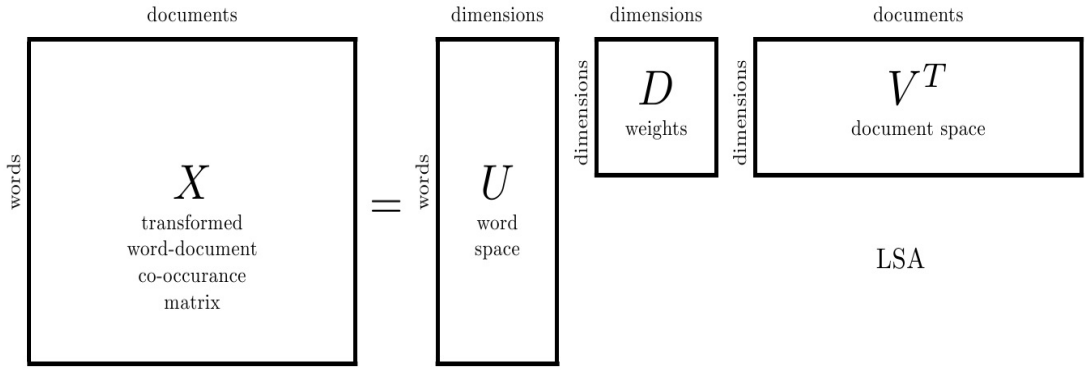

图 2-7: LSA 模型

$$X = UDV^T$$

LSA 模型如上图所示，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。

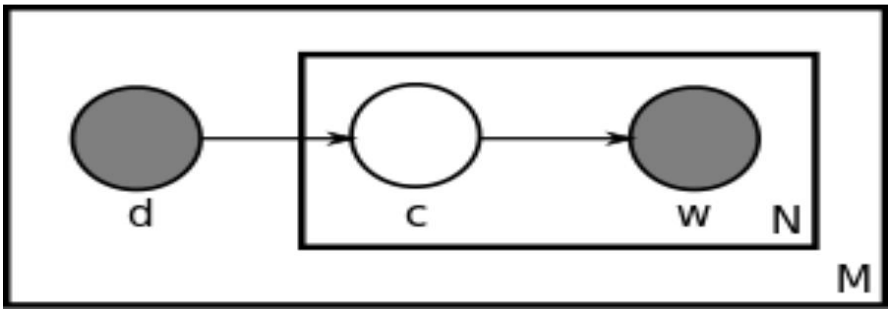

图 2-8: PLSA 模型

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

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 像比如 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）是有 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。

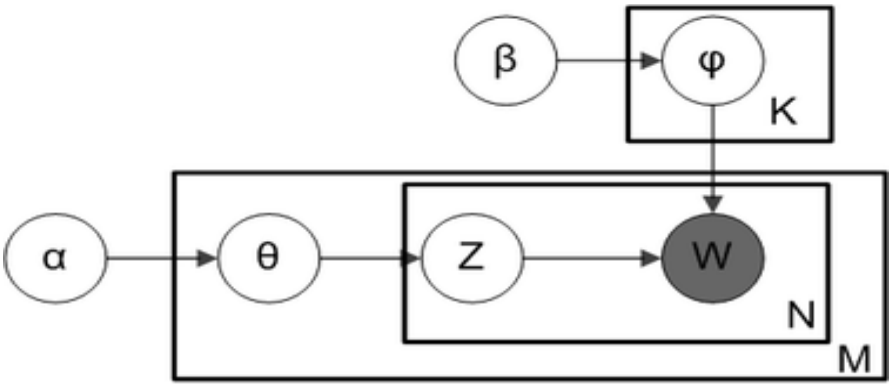

图 2-9: LDA 模型

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 2-9，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：

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

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

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

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

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。主要的求解方式是采用吉布斯采样算法、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方法，这种方法基于马尔科夫链的蒙特卡洛方法。这种方法收敛较快、容易并行化，效果也不错，因为被广泛使用。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

本文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任务中的重要任务之一。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。命名实体识别是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的成熟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字符向量嵌入 -BiLSTM-CRF 模型进行命名实体识别，以该模型为基础解决中文复杂命名实体识别问题。

相比于英文命名实体识别，中文命名实体识别要困难的多，因为英文天然的空格隔开词语，不同词性的词语词缀词根也不同，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，但是中文词语间没有间隔，一词多义，在特征表示上相交英文来说效果不够好。

本章的工作就是面向中文文本的特征表示，针对中文命名实体问题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难以区分的问题，提出基于位置信息丰富单个字符的向量优化方法；（2）针对字符向量训练时受制与上下文窗口大小的限制，字符向量信息量少，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基于字符向量的 BiLSTM-CRF 命名实体识别模型；3.4 节介绍面向位置信息的字符向量优化方法；3.5 介绍面向主题信息的字符向量构造方法；3.6 节描述实验方法及实验结果和分析；3.7 节是本章小结。

3.2 相关理论与工作

本节介绍命名实体识别领域在特征表示方面的工作，这些工作是本文的研究基础，也是本文工作的理论支持。深度学习这种能力强大的模型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的优点就是该模型可以在训练过程中自动寻找数据的特征，因而采用深度学习完成命名实体识别任务的第一个任务环节就是要将文本进行分布式表示。

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。[30] 提出一种基于迭代扩张卷积神经网络（ID-CNN）的标记方案，通过 skip-ngram 对 SENNA 语料库进行 100 维嵌入式训练。[42] 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。[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，之后用 LSTM 抽取字符级别的表示，抽取命名实体。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][43][31] 在命名实体识别模型上也有一定的功效。近期还有 Devlin 等 [44] 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 [45] 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，对词典构造有一定依赖。Zhao 等 [46] 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单词，达到了与注意力机制相似的效果。林泽斐等 [47] 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 [48] 利用 LSTM 中文分词技术优化分词模块，在中文微博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，我们可以得出结论，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）字词的分布式表示（2）深度学习网络有监督训练模型（3）对序列中的每个字词进行标签标注。本论文使用的基准模型是 CharEmbedding-BiLSTM-CRF 命名实体模型，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字典的，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。因而选取直接字符嵌入作为基模型研究。

（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，信息见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 的一个特例，在善于对序列问题建模的同时，还有着易于求解，能够长期保存重要信息。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 同样利用反向信息，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，实验证明 BiLSTM 往往比 LSTM 有着更好的表现 [49]，特别实在序列标注模型上 [29] 可以同时学习过去的序列和未来的序列有着更好的效果。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）和接全连接层用 softmax 激活。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。因为本文序列标

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词向量的训练相似，对于中文语料集首先去除非字符，按每个字符作为一个输入按空格隔开。本文训练语料采用的是人民日报 2014 数据集，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。

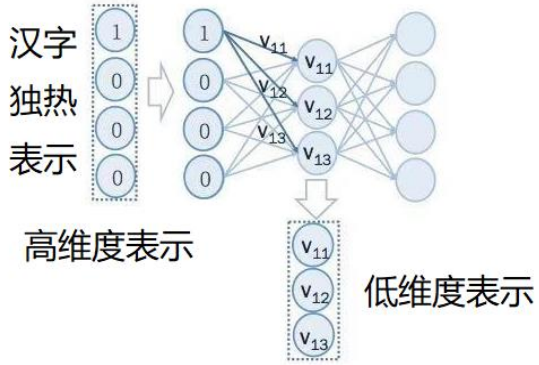

图 3-1: 中文字符向量表示

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，结果如表 3-1 所示。

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	谮 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

word2vec 这样的模型不仅将词汇用我们设定的维度分布式表达，同时也能够保留字与字之间的语义联系，很多文献 [50] 与工作表明这样分布式的表现对后续神经网络处理自然语言问题有着很好的效果。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（见图 3-2），可见语义相近的字向量降维后的坐标相互靠近，语义无关的字向量距离较远，说明字向量的训练达到了一个较好的效果。

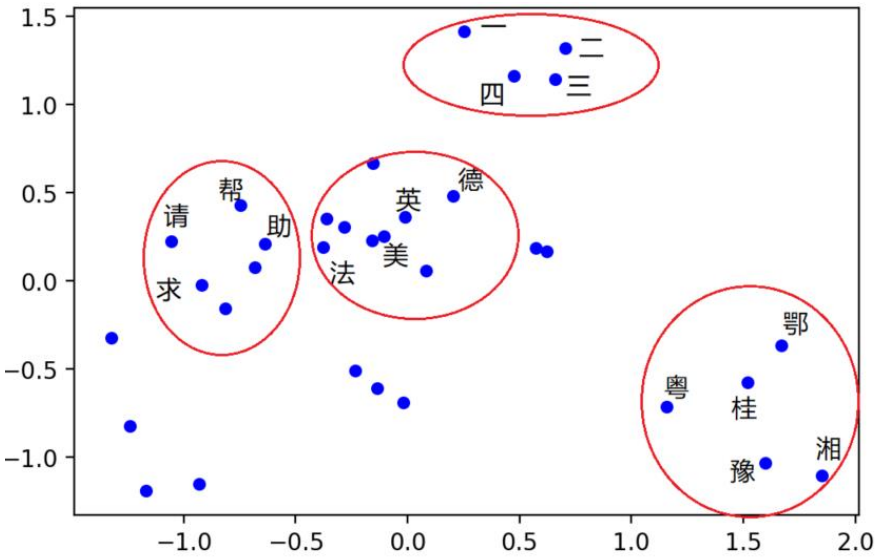

图 3-2: 训练出的字向量降维效果图

（2）神经网络构建

神经网络模型选取双向长短时记忆网络（Bidirectional LSTM Networks），传统的 RNN 模型在训练中会遇到梯度爆炸和梯度消失等难以求解的问题。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。如图 3-3，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。

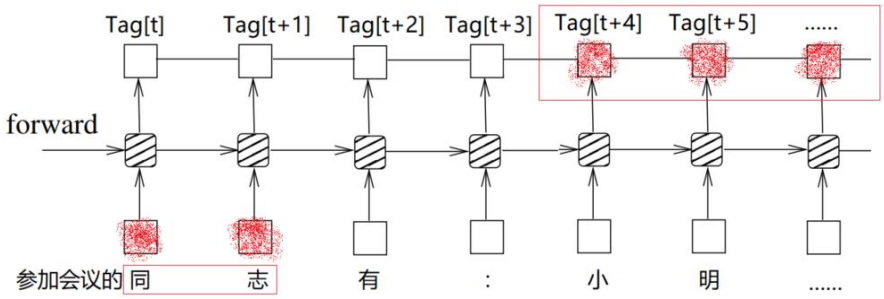

图 3-3: 命名实体识别中的长期信息

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”，通过这些语言片段，我们可以发现我们人类识别出人名、地名、组织名的过程中，往往是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。如图所示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。

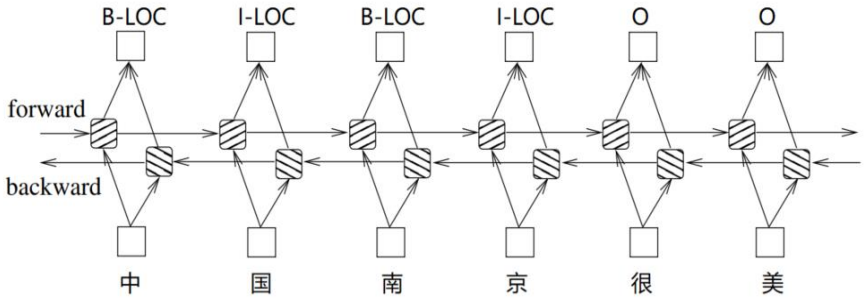

图 3-4: 双向 LSTM 示意图

（3）条件随机场与序列标注

条件随机场模型（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准确率。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、组织名标签不会混搭，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。

设我们输入的序列是 $X = (x_1, x_2, \ldots, x_n)$，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

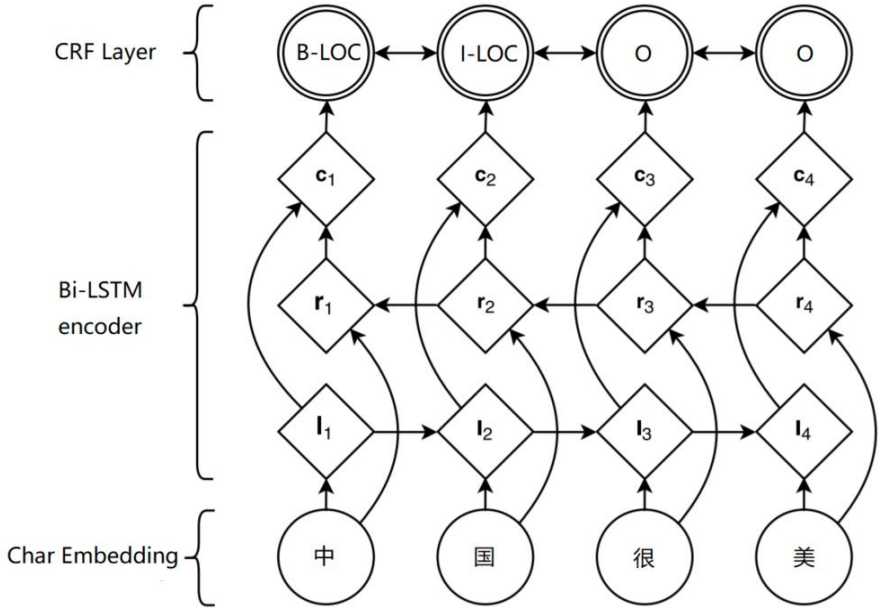

图 3-5: CharEmbedding-BiLSTM-CRF 模型

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，在一个条件随机场里计算标签序列的对数似然值，以此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工作。

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已经成为一个经典成熟的方法，很多工作证明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。

然而当前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抵触，这类实体的命名往往碎片性、象征性、随机性更强和组词关系较小。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“平常”的错误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，这样的目标也与命名实体的目的相违背。另一类工作是寻找中文字符更小的粒度，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频繁，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，可以发现影响命名实体识别准确度召回度的很重要原因就是一字多义问题。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源多于中文字符向量的，也就是说

汉字一字多义、一字多性的现象十分普遍。

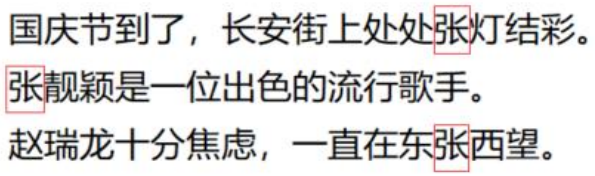

图 3-6: 一字多义、一字多性现象

以上三句话中都含有“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，将字词鸿沟现象更好的解决，将会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，我们同样可以得知，判断一个字符的含义如何，关键还是靠其周围的文字信息。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，具体过程如下：

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 $W = (w(x_1), w(n), \cdots w(x_t), \cdots x_n)$，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，还本句文本中其余的字符含义相关，为方便计算我们将训练的字符向量作为新模型中，本句其他字符的表示。为了统一形式，我们将原始 x_t 的向量提出，这样函数 S=0 时，字符向量即为初始的字符向量。

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

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

这里为了简化模型，本文将函数 S 建模成一个线性函数，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，我们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。

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

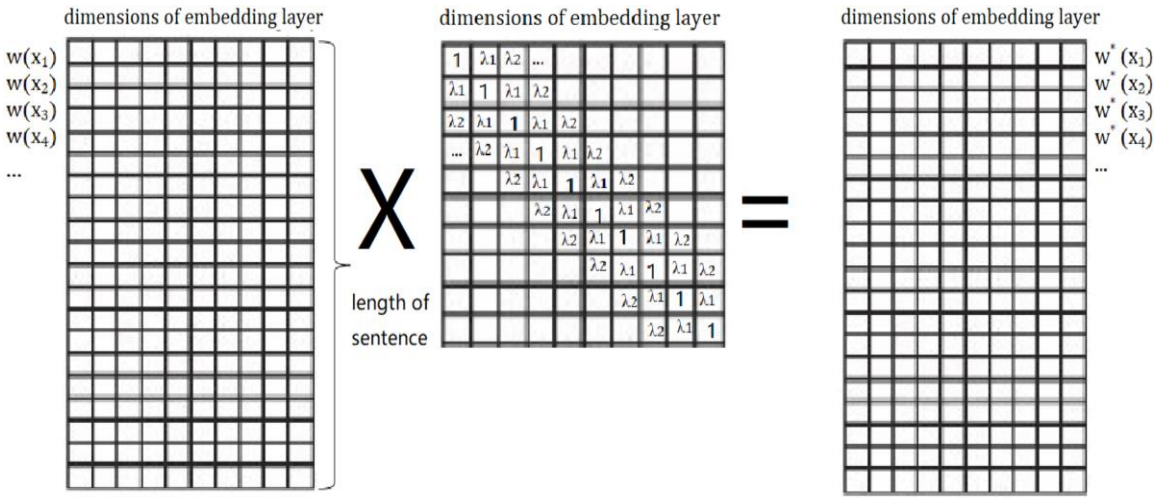

图 3-7: 基于位置信息优化初始字符向量

本文对于该模型的参数求解的方法采用的是网格搜索法，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，这中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，我们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，数据量的大小，对于数据特征的抽取和表示对整体机器学习模型的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，而在 Skip-gram 模型是按当前词的向量预测上下文的词向量。这样的学习关联主要还是基于窗口内的信息，而缺少对于全局信息的把握。总体而言 word2vec 对于短文本的训练学习有很好的语义表征效果，但是如果训练集是新闻正文这样的长文本的话，就会丧失对于同一篇文档属于同一个主题这样有用的信息。

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”矩阵也是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。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 [51] 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 [52] 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。

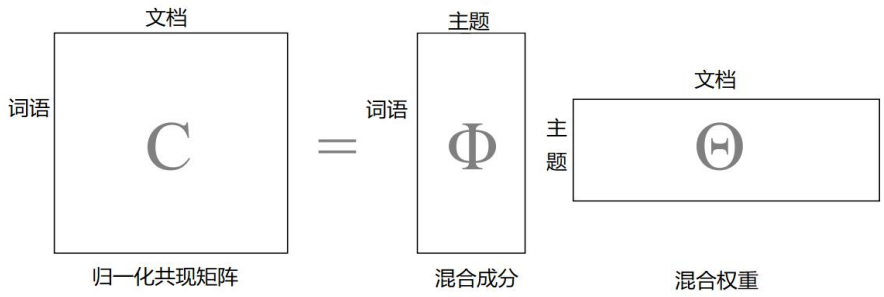

图 3-8: 主题模型

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。主题模型这类非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），隐含表达着一个主题，而这个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（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，通过隐狄利克雷分配模型（LDA）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。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。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。我们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: 主题 -字符参数矩阵 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: 迭代完成后输出主题 -字符参数矩阵 Φ，该矩阵即为不同的词语在不同主题下的概率情况

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，训练完成后，我们观察距离每个主题概率值最大的字符，效果如下：

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。我们将字符主题向量与字符 word2vec 训练出的向量进行比较，由于很多词语在各主题上概率值较小，我们对主题向量做适当放大，在整个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

与图 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想过也证明了我们上文思路的合理性，加入字符级主题信息可以提高表达字符分布式的效果，提高整体命名实体识别系统的效果。下一节的将会用实验结果的形式来说明主题向量在中文字符级分布式特征表示中的作用。

为了在主题特征表示时同事用上 word2vec 训练出的中文字符向量和 LDA 模型训练出的字符主题向量，我们将两类向量拼接。由于两类向量的均值不同，我们在所有主题向量的值前乘以系数 μ，使得两类向量的总体均值相同，避免 LDA 主题向量概率值太小计算下溢的问题出现。其中 μ 的值为：

$$\mu = \frac{\sum^{Dic} \sum_{i}^{dim(W)} w(i)}{\sum^{Dic} \sum_{i}^{dim(Z)} z(i)} \tag{3-8}$$

其中 Dic 表示字典中的所有字符，dim(W) 是 word2vec 字向量的维度，dim(Z) 是主题向量的维度，即 μ 使得主题向量的均值与 word2vec 向量均值相同。将两类字符向量拼接嵌入成为一种改进的中文字符特征分布式表示方法，

示意如下图：

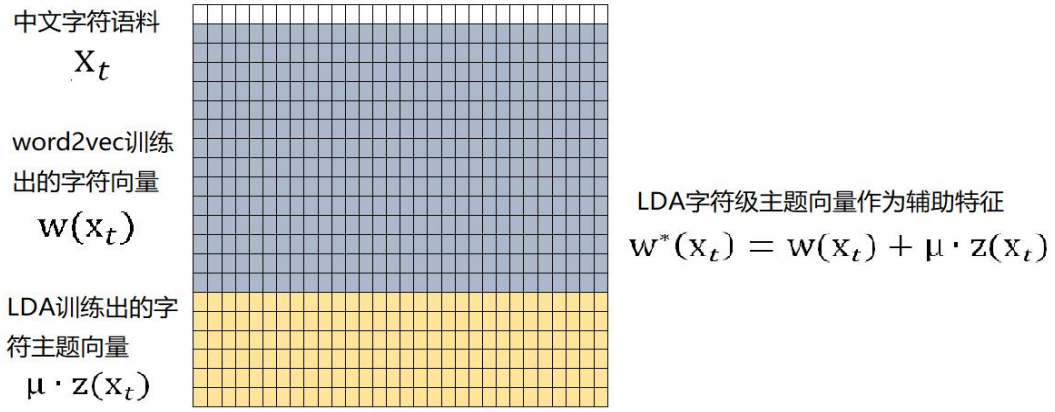

图 3-9: LDA 主题向量作为辅助特征

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，该方法及整体基于深度学习的命名实体识别模型流程图如下：

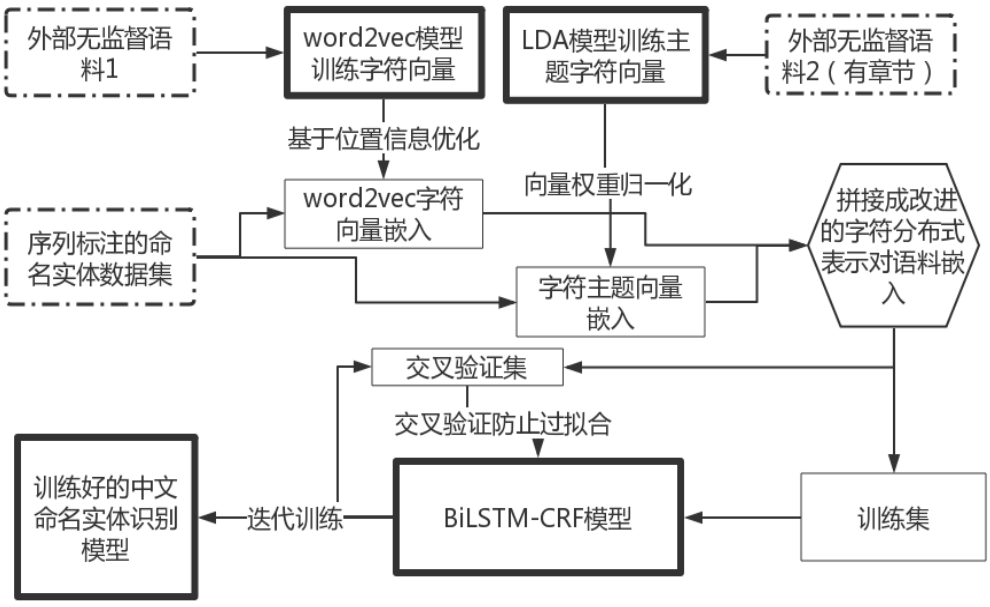

图 3-10: 模型流程图

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 2.20GHz，16GB 内存，处理器四核 i5 的个人计算机，操作系统为 window，还有一台主频为 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，Gensim 是一个可用于无监督主题建模和自然语言处理的开放元法库，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、LDA、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[53] 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

（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 word2ve 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：批处理数（batch_size）为 64；最大迭代轮次为 50；隐层神经元个数为 300；优化算法采用 Adam[54]（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

目标函数设置为交叉熵函数，交叉熵函数起源于信息论中，用来衡量两种概率分布的差异，交叉熵的定义为：概率分布 p 和概率分布 q 的交叉熵为：

$$H(p,q) = E_p[-\log q] = H(p) + D_{KL}(p||q) \tag{3-9}$$

其中 $H(p)$ 是 p 的熵，$D_{KL}(p||q)$ 是从 p 到 q 的 KL 散度。对于本文的离散情况，四散分布 p 和 q 的交叉熵为：

$$H(p,q) = -\sum_{x} p(x)\log q(x) \tag{3-10}$$

在训练的过程中，用该次迭代的批处理数据通过网络计算出预测结果，计算预测结果与真实结果间的交叉熵，做为本轮迭代的损失，根据反向传播算法对神经网络中的参数进行调整。

（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 [**?**] 提出，思路是将初始化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，对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 实验结果与分析

（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 3-4: CharEmbedding-BiLSTM-CRF 模型实验结果

	准确率	召回率	F 值
LOC	92.60%	88.67%	90.59
ORG	83.78%	84.22%	84.00
PER	86.29%	82.46%	84.33
整体	88.63%	85.72%	87.15

表 3-4 是 CharEmbedding-BiLSTM-CRF 模型的实验结果，该模型对于中文命名识别有较优的效果。同时观察不同类别实体的识别准确度召回度，也能

看出模型对于地名的识别较为准确，对于人名、组织名的识别较差。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，我们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，效果如下：

根据上图我们可以看出 λ 大约取当只考虑 x_t 前后 1 个字符的信息时，取不同的 λ 值时，效果如下：

（3）基于主题信息的构造方法实验与分析

（4）整体改进中文字符特征方法的实验与分析

3.7 本章小结

第四章　面向复杂中文命名实体识别的层次深度神经网络模型

4.1　引言

在上一章节，我们在目前成熟鲁棒的中文命名实体识别的模型基础上，利用中文构词逻辑和辅助语义信息，提出了一种改进的中文字符级特征表示方法，通过实验表明该方法在标准数据集上提升了命名实体识别任务的效果。

实际的工程领域中，中文命名实体识别技术还有很多可以提高的方面。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不会遇到的问题：（1）实际中应用中会出现长度很长，地名人名组织名嵌套的命名实体，识别这样的实体模型的准确率会下降。（2）互联网文本信息结构杂乱，信息表示的结构各不相同，直接交给中文命名实体识别系统的效果不好，需要采用一些方法对文本进行合理的切割来提高效果。（3）在实际应用中，命名实体识别系统的召回率往往比准确率重要的多，而在上一章的模型中，召回率相较于准确率而言比较低，需要在两者间找到一个更好的平衡点。

基于这些原因，本文在众多工作的基础上提出了一个层次深度神经网络模型来解决复杂中文命名实体在实际工程中遇到的种种问题。设计多层 CharEmbedding-BiLSTM-CRF 结构，上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，对于识别出的命名实体送交下一层网络实现文本分割的工作，下层网络再精准的判断语言片段中有哪些命名实体。

本章的结构如下：4.2 节介绍复杂命名实体识别的相关工作研究；4.3 节介绍本章提出的层次深度神经网络命名实体识别模型；4.4 节是针对该模型的实验以及结果分析；4.5 节是本章小结。

4.2 复杂命名实体概述与相关工作

英文：

中文：中文嵌套命名实体关系抽取研究许浩亮

4.3 层次标签与层次深度神经网络模型构建

4.4 实验及结果分析

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

基于位置信息不足，因字符调整比如 tf-idf 从词频角度出发，非对称模型

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. [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, .

[42] J. LI A S, JOTY S. Segbot: A generic neural text segmentation model with pointer network[J], .

[43] ANON. Disease named entity recognition by combining conditional random fields and bidirectional recurrent neural networks[J], .

[44] DEVLIN J. Bert:Pre-traing of deep bidirectional transformers for language understanding[J], .

[45] ZHANG Y, YANG J. Chinese NER Using Lattice LSTM[J], .

[46] ANON. Chinese Name Entity Recognition Using Highway-LSTM-CRF[J], .

[47] 林泽斐. 多特征融合的中文命名实体链接方法研究 [J]. 情报学报, .

[48] 王超. 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, .

[49] Alex; Schmidhuber GRAVES J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures[J], .

[50] KIM Y. Convolutional neural networks for sentence classification[J], .

[51] 康宁. 基于主题模型和卷积神经网络的命名实体识别研究 [J]. 南京大学硕士毕业论文, .

[52] 龚凌晖. 中文命名实体识别与歧义消解研究 [J]. 复旦大学硕士毕业论文, .

[53] Jeff; Monga DEAN R E A. TensorFlow: Large-scale machine learning on heterogeneous systems[J], .

[54] DIEDERIK KINGMA J B. Adam: A Method for Stochastic Optimization[J], .

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。