硕士学位论文

面向医疗领域的中文命名实体识别

RESEARCH ON CHINESE NAMED ENTITY RECOGNITION IN MEDICAL FIELD

薛天竹

哈尔滨工业大学

2017 年 6 月

国内图书分类号：TP391.2　　学校代码：10213
国际图书分类号：681.37　　密级：公开

工程硕士学位论文

面向医疗领域的中文命名实体识别

硕士研究生：薛天竹
导　　师：朱聪慧
申请学位：工程硕士
学　　科：计算机技术
所在单位：计算机科学与技术学院
答辩日期：2017 年 6 月
授予学位单位：哈尔滨工业大学

Classified Index: TP391.2
U.D.C: 681.37

Dissertation for the Master Degree in Engineering

RESEARCH ON CHINESE NAMED ENTITY RECOGNITION IN MEDICAL FIELD

Candidate: Xue Tianzhu

Supervisor: Zhu Conghui

Academic Degree Applied for: Master of Engineering

Specialty: Computer Technology

Affiliation: School of Computer Science and Technology

Date of Defence: June, 2017

Degree-Conferring-Institution: Harbin Institute of Technology

摘 要

随着近几年文本数据量的爆炸式增长、大规模知识库的建立和普及，命名实体识别研究已经逐渐成为自然语言处理领域的一大研究热点。然而，传统的基于有监督学习的方法，需要大规模的标注语料。在标注语料稀缺的医疗领域，传统的命名实体识别方法并不能够达到理想的效果。

随着深度学习的火热发展和普及，循环神经网络（RNN，Recurrent Neural Network），尤其是长短期存储单元 LSTM（Long-Short Term Memory）被广泛应用于自然语言处理领域，并在多个研究方向上取得显著高于传统方法的成绩。因此，我们首先利用 LSTM 模型进行医疗领域的命名实体识别的研究，并证明其无论是在研究效果评价还是实际应用层面，都能够达到比传统的条件随机场模型（CRF，Conditional Random Fields）更好的效果。

由于医疗领域的规范的标注语料相对稀少，我们在 LSTM 模型已经取得比 CRF 模型更好的效果的基础上，还希望它能够通过融合外部信息，同时学习到新闻领域的语言学特征和医疗领域的无监督语义信息，达到更好的效果。我们利用了深度学习中迁移学习和预训练的相关知识，对医疗领域的模型进行了参数融合和模型调优，使得模型的效果进一步提升。

最后，由于 LSTM 模型在实际应用中的缺陷，我们希望能够利用另一种方法进行领域自适应的命名实体识别。为了找寻不同知识域的领域差异，我们进行了多组混合不同领域语料的对比实验进行分析和探究。并通过 GBDT 模型集成领域差异和无监督的医疗领域的语义向量进行命名实体识别的研究，取得了较好的研究效果。

关键词：命名实体识别；LSTM；迁移学习；GBDT 模型；实际应用效果

Abstract

With the explosive growth of the amount of text data and the establis hment and popularization of large-scale knowledge base in recent years, th e research of named entity recognition has become a hot topic in the field of Natural Language Processing.However,the traditional methods of Named Entity Recognition are all based on supervised learning,and need large-sca le annotated corpora.Therefore,the traditional methods of Named Entity Rec ognition can not achieve good efforts in medical field where the annotated corpus is so scarce.

With the development and popularization of deep learning,RNN(Recurr ent Neural Network) model,especially LSTM(Long-Short Term Memory) uni ts are widely used in Natural Language Processing,and achieve remarkable results which is much better than the traditional method in many directions. So we first use the LSTM model to research in the Named Entity Recogni tion of medical filed.We prove that it can achieve a higher level both in r esearch evaluation and practical application than the traditional Conditional Random Field model as well.

Due to the scarcity of annotated corpus in medical field,we want our LSTM model to learn external information which include not only the ling uistic features in the general field but also the unsupervised semantic infor mation of the medical filed at the same time so that it can achieve a bett er result on the fact that it has do much better than CRF model. In this c ase, we make use of the pre-training method and transfer leaning in deep learning so that the effect of the model is further improved.

Finally,we hope to use another method to perform domain-adaptive Na med Entity Recognition due to the shortcomings of LSTM model in practic al applications. We search for the differences and effects of the two langu age domains by grouping entity identification experiments which the trainin g data is mixed by the annotated corpus in medical field and generic filed. We also use GBDT model to achieve a better practical application effort i n the Named Entity Recognition with the integration of unsupervised seme

ntic vector which is trained by unsupervised learning in the medical field and the domain difference.

Keywords： named entity recognition; LSTM; transfer learning;GBDT model; practical application effort

目 录

第1章 绪论

1.1 课题研究背景与意义

伴随着互联网技术日新月异的发展，大批量的结构形式复杂的文本信息出现在互联网上，种类繁多，具有很强的多样性，并且随着人们日复一日的使用互联网，其数量呈现爆炸式增长。利用自然语言处理技术进行文本数据的自动化处理、抽取和过滤，组织成结构化内容，是信息化时代的需求。通过自然语言处理技术中的信息抽取（Information Extraction，IE）技术，可以更快速的抽取文本中蕴含的想要获得的知识，分析文本，并为之后的自然语言处理工作做好铺垫。信息抽取希望通过对文本的分析，进而组织成半结构化或结构化文本，并利用相关技术进行文本信息的自动分析和理解，如命名实体识别、关系抽取等技术。信息抽取这一新兴领域也成为众多科研院所、大学、公司的研究热点。

命名实体是文本中包含特定语义的实体词汇或特殊知识域中的专有名词的统称。如新闻领域的地点名、机构名、人名等实体词，或是医疗领域的蛋白质名、疾病名等专有词汇。通过对文本进行命名实体识别(NER，Named Entity Recognition)，可以从文本中分析得到专有的实体名词与实体名词所属的类别，能够加快用户对无规则文本内容的理解以及对关键内容的把握。很多自然语言处理任务，如关系抽取、问答系统、自动文摘等都是在命名实体识别的结果的基础上进行工作。除此之外，命名实体识别也同样在非自然语言处理任务，如数据挖掘、机器学习等相关领域发挥着重要的作用[2]。

命名实体识别的研究实质是要同时完成找到文本中存在的命名实体和对找到的命名实体进行分类两个相关任务。早期的命名实体识别方法都包含一系列的人工抽取的规则、对应不同类别的实体类别关键词以及应用类别词和规则词的实体命名系统。规则集和命名实体的类别都是完全由人工制定的，对于规则集中包含的规则匹配到的语句效果相对精确，但是整体覆盖率很低，工作效率低。且规则集的扩充非常复杂，如果在已有的系统中加入新的知识都需要高昂的人力和物力代价。不适合广泛推广。相比之下，之后兴起的基于统计的命名实体识别方法能够更加灵活的结合丰富的语言学特征，借助机器学习模型强大的学习特征的能力，减少了对训练语料规模的依赖，达到更好的效果。因此目

前普遍实际应用的命名实体识别工具，以及开展的基础的命名实体识别相关的研究都是基于统计的方法。

近些年由于硬件设备的快速发展，尤其是 GPU(图形处理器)的计算速度和显存大幅度提升，深度学习的相关方法在自然语言处理领域的许多任务中逐渐受到关注并流行起来。深度学习这一概念是神经网络这一概念的延伸，实质上是利用神经网络提取输入数据的隐含特征，是一种深层次的特征表示工程。传统的机器学习方法通过人工分析和经验进行语言学的特征抽取，缺乏一定的全面性和深度，深度学习的优势，在于能够挖掘文本表层无法表示的深层语义信息和隐含信息，更贴近地对文本进行语义向量表示，大幅度的提高这些任务系统的性能。深度学习目前被广泛应用于自然语言处理领域中的多个方向，且在诸多任务如机器翻译、句法分析等任务上取得了比传统方法更加显著的成果，如何利用深度学习在表示层面的优势，并将其根据问题的特性恰当地运用到命名实体识别的研究中，也是目前相关方向的研究热点之一。

随着人们生活质量的逐渐提高，人们有更多的需求和目光放在了身体健康、医疗等医疗领域，希望能够通过网络获取一些常见的医疗相关信息，进而能够对自己的身体健康状况有更好的了解。随着大数据时代的来临，很多公司也把目光投向了医疗健康大数据这一方向，希望通过对已有的医疗领域的信息，尤其是文本信息，譬如医疗报告、医生诊断、电子病历等文档进行一定的处理，从中获取信息，并为下一步自动化的辅助分析做好准备。然而医疗领域的自然语言信息处理的关键问题在于对病人和医生有用的如药品名、机构名、疾病名等医疗专有名词，在新闻领域的带标注文本中出现次数极少，导致很难利用新闻领域的语料进行命名实体识别的模型构建，并将其应用到医疗领域中。而医疗领域公开的带有专有名词标注的数据极为稀缺，因此如何利用自然语言处理知识，解决医疗领域的命名实体识别是一个具有挑战性的任务。

目前网络上已经公开的医疗领域的命名实体识别工具或者研究，主要都是采用传统的统计模型并配合上一定的规则进行的，通过利用医疗领域的标注数据训练模型。然而由于医疗领域的标注语料过于稀少，使得训练得到的命名实体识别模型具有很大的局限性，且模型在实际应用中的可扩展性较差。除此之外，传统的统计模型不能够充分利用甚至无法利用文本的上下文信息、结构信息和语义信息等。因此如果能够利用充足的新闻领域语料中的这些知识信息，并融合到医疗领域，可能会对医疗领域的命名实体识别系统有着明显的提高。

面向医疗领域的命名实体识别研究，不仅能够为构建大规模医疗领域知识图谱打下重要的基础，还能够对日常生活中可能出现的如自动诊疗、导诊、问

答系统都起着重要的影响。无论是从自然语言处理的研究角度考虑，还是从日常的行医问诊的实际应用出发，本课题都有着极大的研究价值和实用价值。

1.2 相关工作研究现状

1.2.1 命名实体识别研究现状

命名实体识别最早是由 Lisa F. Rau[3]在一次从文本中识别公司及组织名的研究开始进入人们的视线。命名实体这一名词最早是在第 6 届信息理解研讨会（MUC-6[4]）上被提出的。当时的信息理解讨论会主要关注从报纸文章等非结构化文本和信息中，抽取出国防任务等相关内容的结构化信息，主要定义了如识别组织机构名、人名和地点名等相关任务即命名实体识别，并发现其对分析和整理非结构化文本的必要性。在这样的趋势和环境下，命名实体识别任务的相关研究得到了重视和快速的发展。

命名实体任务的相关研究是一个漫长而又逐步演变的过程。第一个在该领域的研究论文是 Lisa F. Rau[3]在 IEEE 人工智能应用会议上发表的基于启发式搜索和规则匹配的识别公司组织名的系统。人工撰写的规则一般都是通过对典型的文本中正面和负面的样例，根据人工经验总结而得到的。这样的规则对特定范围的文本的识别效果较高，然而人工成本较高，且覆盖率和平均工作效率都非常低。随着统计学习模型的发展以及计算机性能的逐步提高，基于统计学习中有监督学习的命名实体识别相关研究逐渐流行起来。通过自动发现和抽取标注语料库中正面和负面的样例的字、词、频次、位置等特征并建立机器学习模型，对无标注的测试文本进行标注，代替了人工抽取规则的流程。基于统计的命名实体识别方法大幅度的减轻了研究者的人工劳动，并且能够结合研究者人工补充规则更好地提升研究效果，常见的方法有最大熵马尔可夫模型[5]（ME，Maximum Entropy），隐马尔可夫模型[6]（HMM，Hidden Markov Model），条件随机场[7]（CRF，Conditional Random Fields），支持向量机[8]（SVM，Support Vector Machine）等模型。

条件随机场[7]（CRF，Conditional Random Fields）是由 Lafferty[9]在 2001 年提出的一种基于概率无向图的判别式模型，通过观测序列和待抽取的特征序列，建立对数似然模型进行特征学习。条件随机场模型既解决了隐马尔可夫模型[6]（HMM，Hidden Markov Model）无法根据整句的特征参数优化的缺点，

又解决了标记偏置，在解决与序列相关的自然语言处理任务中有很突出的效果，也一直是基于统计的命名实体识别方法中效果表现最好的模型之一，被广泛应用在学术研究和实际应用中。

Mccallum[10]在 CoNLL-2003 的命名实体识别的相关任务中，利用了条件随机场模型并取得了相对不错的成绩。与此同时，Finkel[11]利用了吉布斯采样（Gibbs sampling）的相关训练方法，并添加了部分非局部的特征，进行了条件随机场模型的训练，并将其应用到命名实体识别任务，也取得了较好的结果。Krishnan 等[12]进行了历史研究的总结和整理，对于模型不能够充分的利用到序列标注问题中的局部特征的情况，他们通过对条件随机场的输出进行再次的条件随机场的建模的方法，来进行更好的局部特征提取和利用。而关于如何更好的利用条件随机场模型，进行特征的抽取组合等相关的方向，张祝玉等[13]进行了充分的对条件随机场的比较实验，利用多种特征组合和抽取，通过多次的比较实验定义相应的特征的贡献度，并选择其中贡献度大的特征，同时还利用了集成模型的方法提升了条件随机场模型的效果。

综上所述，命名实体识别的研究在多个延伸方向都已基本完善。然而值得注意的是，绝大多数的研究工作都是基于标准语料，即带标注的新闻领域语料进行的，这使得在开放领域的规则文本和通用文本上，命名实体识别研究能够取得较好的结果和实用价值。而在缺乏标注语料的领域，如医疗等知识领域中，新闻领域的标注语料由于领域差异，不能起到很好的作用。因此目前更多研究者将目光集中到标注较少或者文本较不规范的领域。

1.2.2 深度学习在自然语言处理领域的研究现状

随着近几年计算机硬件的高速发展，尤其是图形处理器计算能力的大幅度进步，让深度学习逐步走入人们的视线并流行起来。尤其是近几年，深度学习在传统的机器学习领域，如图像识别，语音识别等领域取得了令人瞩目的成果。因此，众多的机器学习相关领域的研究人员都把目光投向了深度学习的方法。然而，深度学习看似在最近几年才逐渐兴起，实际上有着源远流长的发展历史。深度学习实际上起源于神经网络，随着 20 世纪 40 年代到 60 年代控制论的出现，人们开始尝试利用神经网络，诸如感知机[14]等类似神经元工作方式的模型出现，这些简单的学习算法开始逐渐影响机器学习然而又由于线性模型的局限性而被研究者放弃。神经网络的第二次浪潮是随着分布式并行处理技术的兴起，而其中一个重要成就就是反向传播算法[15]使得神经网络的训练非常成

功，然而由于神经网络的参数过多，并被硬件的计算能力所限制，且后续的核方法、支持向量机和图模型等的效果都要好于神经网络，这些因素导致了神经网络的第二次衰退。

神经网络最近的兴起和流行则源于 2006 年 Hinton 等[16]的研究。Hinton 经由多年对神经网络的深刻研究，提出了利用深度学习进行深度神经网络构建的观点。并且为了解决神经网络训练的繁琐，他使用了基于贪心的逐层预训练的策略，解决了陷入向量空间局部最小的问题，起到了非常好的作用，相关的训练方法也被 Bengio 和 LeCun[17]采用到其他工作上并提升‘深度’理论的性能和重要性上。同年，Hinton 等[18]提出了利用无监督学习的深度置信网络（DBN, Deep Belief Network），丰富了神经网络方法的种类。随着上述有代表性的工作发表后，研究者们逐渐把研究目光投向深度学习方法，各个领域的深度学习方法都涌现出诸多明显好于传统机器方法的成果。深度置信网络被分别用于语音识别[19]、图像识别[20]、人脸识别[21]、声学信号处理[22]等多个相关领域。在这些机器学习相关的领域中，深度学习都取得了显著的成果甚至是突破性的进展，无论是在准确率还是实用性上都要好于传统方法。

深度学习方法在自然语言处理领域的初步探索中，并没有像在其他领域一样，取得突飞猛进的成果。研究者们逐渐意识到这一问题的核心要点，在于如何对词进行分布式表示，并进行类似传统自然语言处理中的语言模型的模型构建。Bengio 等[23]创造性的提出用向量来表示词，提出了神经网络语言模型，结合上下文与词表进行后文的预测，通用优化带正则化的最大似然概率进行模型的训练，进而进行短语、句子等级别的表示。神经网络语言模型的出色效果，引起了很多自然语言处理领域研究者的注意。词袋模型的表示方法虽然对词进行了向量化，但是在大规模文本处理的过程中会出现维度灾难。除此之外，one-hot 表示得到的向量在语言学上并没有特定的含义。词向量的出现则在这两个问题上给出了很好的解决，词向量能够通过训练时指定维度，解决维度灾难；另外由于训练的时候主要关注上下文的信息，因此还可以通过向量之间的基础计算，如余弦相似度等计算两个词的相似度，词向量即是对词的语义的信息的一定程度的表示。由于词向量是通过无监督学习进行训练，对于不同知识域、不同要求的自然语言处理任务，可以利用最适合研究的语料进行词向量的训练，对后续的研究起到极大的帮助。因此词向量自一提出便得到广泛的应用，研究者们也都从不同的方向和方法入手，进行着词向量的研究和改进。

Collobert 等[24]在多个自然语言处理任务中进行了神经网络语言模型系统（SENNA）生成的词向量的试验，如在序列标注问题中经典地加入了全局代

价作为优化函数，并从神经网络语言模型的角度给出了如何在序列标注、语义角色标注等领域使用词向量解决相关问题，给出了相关模型构建和隐层组合的诸多方法，对自然语言处理领域的诸多任务的研究，尤其是序列标注问题的相关研究具有重大的影响作用。Huang 等[25]则在 Bengio 的神经网络语言模型的基础上，加入了一个子网络用来提取文本中的全局信息，取得了很好的效果。

另一些研究者希望对词向量训练的方法进行改进并做了很多尝试，如 Bengio[26]等人提出的层次神经网络语言模型，减少了网络中的参数使得训练速度大幅度提高。不过其中最为人所知的模型是由 Mikolov 等[27]提出的 word2vec，创造性地进行了结构上的改进和速度上的优化，使得词向量训练的速度大大加快，而且不仅能够通过上下文预测词向量，还可以通过词预测其上下文环境，word2vec 工具也迅速成为字或词向量化的最广泛使用的应用之一。除此之外，Mikolov 等[27]在这一基础上，提出了循环神经网络语言模型，能够更好的考虑到上下文信息，在处理序列数据，如文本数据等具有很大的优势，效果也得到了了大幅度的提升。

在词向量的基础上，很多研究者开始重视词向量等相关的能够表示一定语义信息的向量，在得到词向量的基础上进行更高层次的建模并得到诸如句向量[28]、文档向量[29]等。这些方法都在实际应用中发挥了其独有的效果，这也体现了自然语言处理领域的研究者已经开始接受文本的分布式表示这一深度学习的基础概念，并在这基础上开始进行后续的工作。

深度学习另一个影响自然语言处理领域的方面在于特定结构的神经网络，如 Mikolov 等[27]使用来训练语言模型的循环神经网络模型，循环神经网络（RNN, Recurrent Neural Network）是由 Jordan 等[30]在 80 年代提出的。循环神经网络最重要的特点就是输入的序列能够同时结合当前位置的输入与前一输入的隐含层进行计算，对于自然语言处理领域来说，通过循环神经网络可以训练得到的带有上下文信息的模型，更符合文本信息的语言学特征。而随着深度学习在自然语言处理领域的不断拓展，RNN 广泛吸引了研究者的注意，而随着 RNN 的大量使用，RNN 的缺点也被逐渐发现。由于深度神经网络传递残差时需要多次求导，如果网络的层数过高，经常会发现梯度消失的问题；除此之外，RNN 网络虽然能够保留历史信息，但是由于距离当前输入距离较远的历史信息，会在多次求导不断被稀疏，使得 RNN 网络更多的是保留距离当前输入较近的信息，并不能够真正保留所有历史信息，这在如句法分析、关系抽取等任务中会有较大的影响。

对于 RNN 模型中的如上问题如果改进，研究者们进行了许多方面的尝

试。1997 年 Hochreiter 等[31]创造性地提出了一种基于 RNN 模型的改进模型：长短期存储单元（Long-Short Term Memory, LSTM），之后的 Gers 等[32]人、Graves 等[33]人又在 LSTM 模型的基础上继续努力改进，相比于 RNN 模型仅仅简单使用前一位置的隐含层作为历史信息，不仅保存的信息较为简单，还容易造成梯度消失这一现状，LSTM 模型使用了多个类似于 RNN 隐含层的单元，即门（gate）来控制历史信息和输入信息如何进行输入、输出和更新。除此之外，LSTM 的关键在于使用了一个名为存储单元（Memory Cell）代替了传统 RNN 模型中的常规神经元，并建立了门机制。其中输入门（input gate）决定输入的信息如何流入到模型中，输出门用于将真正需要的信息流出到隐含层中，遗忘门则能够控制存储单元何时遗忘，遗忘多少历史信息。相比于 RNN 中仅有一个隐含层来处理历史数据的情况，LSTM 通过三个类似的门单元进行数据控制，能够更有效的保存更有价值的历史信息。

LSTM 模型的提出，立刻吸引了诸多自然语言研究者的目光，Bahdanau 等[34]将 LSTM 模型组合成端到端的结构，并加入了注意力机制，得到的机器翻译结果远好于最新的研究结果。Dyer[35]等在依存句法分析人物中使用了 LSTM 模型，在中文和英文两个测试集上都取得了显著的成果。Xu 等[36]利用文本的最短依赖路径信息进行关系分类问题的 LSTM 模型的建模，并加入词向量与语言学特征向量相结合的方式，在 SemEval 2010 数据集上获得了 83.7%的 F1 值的突出效果。Lample[37]等在命名实体识别任务上分别使用了带有 CRF 转移层的 LSTM 和栈式 LSTM，在多个语种的命名实体识别测试语料上都取得了最好成绩。

综上，深度学习在自然语言处理领域的诸多的公开的数据集和传统任务上都取得了显著的成绩。随着这些领域和任务的效果逐渐达到饱和，而如何在更多数据稀缺的领域，或是实用性较强的自然语言处理任务上达到更好的效果，将是近几年深度学习在自然语言处理领域的发展的一大重点。

1.2.3 医疗领域的命名实体识别研究现状

医疗领域的自然语言处理研究，旨在帮助医生对病情判断进行更好更有效率地诊断，对如电子病历、医疗结算单、医疗文献等文档利用自然语言处理技术进行自动化分析，医生在自动化分析的基础之上进行进一步的人工诊断和处理。由于医疗领域的命名实体识别，缺乏充足的标注数据，然而医疗领域的非结构化文本或半结构化文本，包含了丰富的医疗领域的命名实体。抽取出这些

命名实体并进行相关的自然语言处理工作具有极高的实用价值，引起了研究者的广泛兴趣。李刚[38]建立了一个识别医疗领域文献中的蛋白质名的贝叶斯模型，并对模型进行了词典的补充，取得了很好的效果。张金龙[39]则在传统的利用条件随机场的命名实体识别基础上，加入外部上下文特征与筛选规则，进行了中文医疗机构实体的识别。

随着研究的逐渐深入，医疗文本的命名实体识别，尤其是电子病历的命名实体识别任务，得到了各大组织机构和各个国家的高度重视。电子病历实质上是病人病情诊断的文本表述，富含医疗领域的特有名词与知识。对电子病历的自动化实体分析和识别，能够更快速地帮助医生诊断病情，提高工作效率。如何更好的通过评测，进行各种命名实体识别方法的比较，进而进行实际应用也成为了这一阶段医疗领域的命名实体研究更加关心的方面。美国集成生物与临床信息学研究中心（Informatics for Integrating Biology & the Bedside，I2B2）多次组织医疗领域的各种方向的自然语言处理相关任务，并发表规范的电子病历语料[40]，建立了相应的语料库。

规范化的语料库和流程，使得自然语言处理的研究者们对这一领域的研究热情得到了极大的提高。Jiang[41]分别采用了条件随机场模型和支持向量机进行了电子病历中的命名实体识别的对比实验，发现条件随机场模型略好于支持向量机模型。Jonnalagadda 等[42]进行了半监督的基于条件随机场的模型的训练，通过类似词向量训练的方式，从未标注语料中抽取出无监督的分布式语义特征向量表示，并在模型中加入了近邻的上下文信息，达到了 0.823 的 F 值。王鹏远[43]等在 Jonnalagadda 的语料上进行了多标签条件随机场模型的构建和测试，在数据中含有较多复合实体名称时大大提高了模型的效果，且不需要根据语料制定规则，系统的灵活性更好。诸如此类工作不胜枚举，其中最突出的是 Bruijn 等[44]使用了融入了大量上下文特征和外部特征的弱马尔科夫模型（semi-Markov HMM）进行序列标注，并采用了自训练的学习方法来应对训练语料缺乏的情况，增加了系统的扩展性和性能，获得了 I2B2 2010 实体抽取评测任务的最好成绩。

由于中文自然语言处理在分词等词法分析上相比于英文更加困难，且中文电子病历缺少统一的规范和评测标注，使得目前基于中文电子病历的命名实体识别研究进展仍然较为缓慢。且由于缺乏公开标注的权威性语料库的原因，很多模型和方法只能够作为研究使用，并不具备较好的实用价值，因此需要研究者在进行研究的同时进行相关语料库的建设。苏娅等[45]对 1000 篇医疗问答文本进行了共五类医疗实体的人工标注，在标注的基础上进行命名实体识别的实

验，通过利用条件随机场模型配合词典等外部特征得到了 81.26%准确率的不错结果。栗伟等[46]也进行了类似的工作，通过对 912 份电子病历进行标注和命名实体识别模型的构建，研究病历中不同属性如科室等特征如何影响命名实体识别的效果。曲春燕等[47]建立了基于专业人员标注的实体语料库，在语料库的基础上，进行了不同机器学习模型在中文电子病历的命名实体识别的效果比较，并达到了最好值为 92.71 的 F 值。

可以看出，随着自然语言处理技术的火热发展，以及人们对医疗领域信息的重视，医疗领域文本内蕴含的丰富的信息也逐渐被研究者们挖掘出来。如何更好地利用已有的知识库和语料进行医疗领域的命名实体识别研究，将是本文研究的重点。

1.3 本文主要研究内容

本文主要进行了医疗领域的命名实体识别研究，在单一领域的命名实体识别研究中通过深度学习中的 LSTM 在新闻领域进行了模型的基础构建，并在医疗领域进行了模型的训练，有效地提高了医疗领域的命名实体识别的效果。随后在 LSTM 模型的基础上，利用深度学习中的迁移学习和预训练等技巧融合了外部知识，进行了跨领域的知识的学习，使得模型的效果得到进一步的提高。最后，对于深度学习模型缺乏学习效率等问题的情况下，提出了利用 GBDT 模型进行领域自适应的弱监督的命名实体识别研究，进行了大量的探究性实验。

1.4 论文的组织结构

本文着眼于面向医疗领域的命名实体识别研究，通过大量对比实验和方法的实践，从研究和实践两个方面比较了本文提出的方法与传统方法的优劣。

本文主要的研究脉络是从单一领域的有监督学习方法入手，并通过跨领域多领域语料共同进行有监督学习的方法进行模型效果的提升，最后又根据已有方法的不足，提出了一种融合了无监督语义向量的领域自适应的集成方法进行了弱监督学习，具体的研究工作流程如图 1-1 所示。文章组织结构如下：

在第一章首先进行了面向医疗领域的命名实体研究的课题分析，通过命名实体识别、深度学习在自然语言处理中的研究前景和医疗领域的命名实体识别发展过程等三方面讲述了本课题的发展历史和研究背景，通过对研究现状的优缺点的分析进一步阐述了本课题的研究意义和必要性。随后简要介绍了本课题

的研究内容和本文的组织结构。

第二章主要介绍了如何利用深度学习知识，尤其是深度学习中的 LSTM 单元，进行单一领域的命名实体识别研究。在详细介绍了 RNN、LSTM 等深度学习模型的结构、原理后，对模型进行了实现，并详细说明了模型架构和实现细节，进行了多组对比实验验证了模型的效果。

第三章在第二章的模型基础上，进行了融合外部信息的跨领域的命名实体识别的相关研究，通过迁移学习和预训练等深度学习的技巧和方法进行参数融合和调参，进一步提升了医疗领域的 LSTM 的模型效果。

第四章针对深度学习模型在实际应用中存在的问题，提出和探究如何衡量新闻领域和医疗领域语料的差异度。在进行了多组对比实验并对实验结果进行分析后，尝试利用 GBDT 模型进行无监督语义向量与领域差异的集成学习，通过多种特征向量的组合方法和领域差异的衡量方法进行对比，取得了不错的效果。

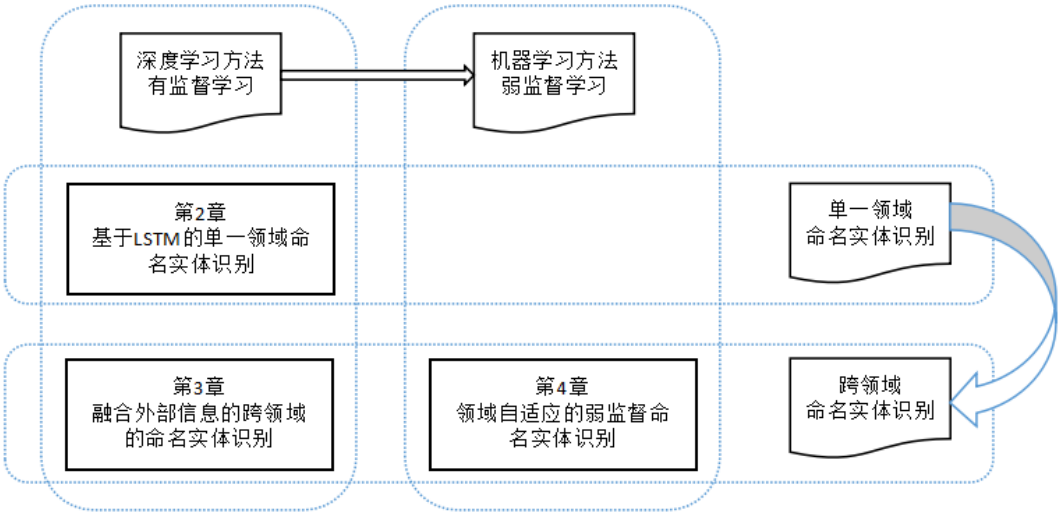

图 1-1 本文的研究章节示意图

第2章 基于 LSTM 的单一领域命名实体识别

在本章中，我们首先介绍了深度学习的 RNN 和 LSTM 等神经网络模型的相关基础知识和原理，并同时介绍了基线系统中使用的条件随机场模型。在进行了相关模型知识的铺垫后，介绍了 LSTM 模型的组成结构和训练技巧，利用 Theano 工具进行了模型的构建并在医疗领域进行了一系列不同参数配置的对比实验，充分说明了 LSTM 模型无论是在实验效果还是在实际应用的实际价值上都要优于传统的条件随机场模型。

2.1 深度学习的神经网络模型

2.1.1 前馈神经网络（FNN）

深度前馈网络（deep feedforword network）实质上是多层的具有更高模型深度的前馈神经网络（feedforward neural network），是一种二分类的机器学习分类模型，通过多层的非线性感知机模型的组合，近似的拟合某个分类函数 f^*，并进行分类。神经网络这一概念，借鉴于生物学神经系统中的概念。生物学中的神经网络的基础单元为神经元，神经元通过接受其他神经元传来的化学物质，改变自身的电位，当电位超过阈值时，就会进入激活状态。深度学习中的神经元的工作机制也与此类似，定义一个神经元有 N 个输入，神经元通过接受所有的输入，并经过激活函数产生神经元的输出。设神经元的 N 个输入中，序号为 i 的输入记做 x_i，记 x_i 的权值参数为 W_i，偏置参数为 b，则神经元的输出函数 $h(x)$ 可以定义为如下形式：

$$h(x) = f(\sum_{i=1}^{N} W_i x_i + b) \qquad (2\text{-}1)$$

$f(\mathrm{x})$ 被称为激活函数。非线性感知器的激活函数一般应用如图 2-1 中所示的双曲正切（tanh）函数和图 2-2 所示的 sigmoid 函数。

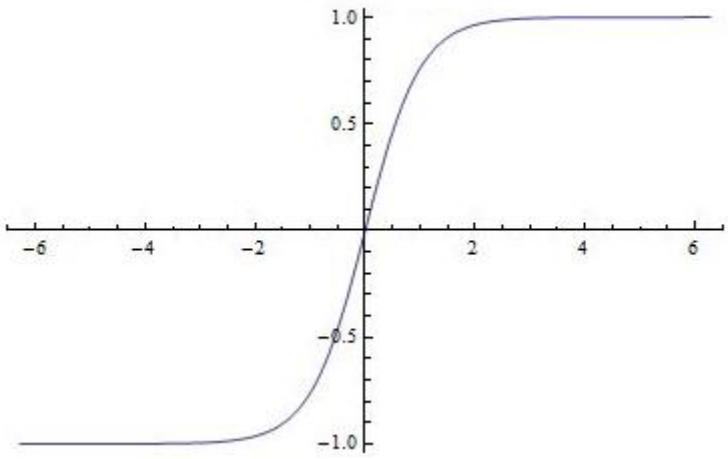

图 2-1 双曲正切（tanh）函数

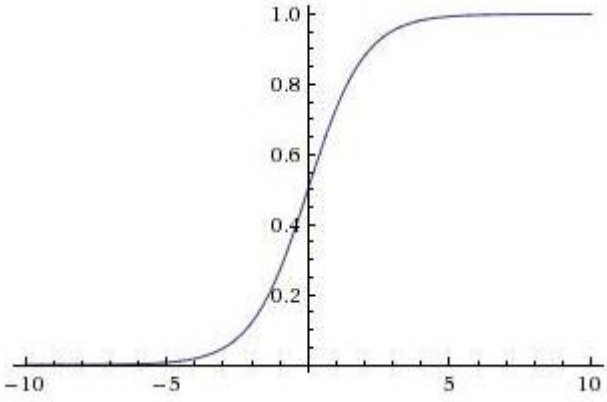

图 2-2 sigmoid 函数

许多个神经元，按照如图 2-3 的结构，将输入和输出按照一定的层次和顺序进行连接，就得到了一个简单的神经网络，整体神经网络的输出，是通过从最前端的输入层，逐层进行计算，图 2-3 总的 L1 层被称为输入层，输入层得到的输入通过 L2、L3 层两层隐含层的激活函数进行变换和加工，最终通过 L4 层输出层将神经网络的结果输出。

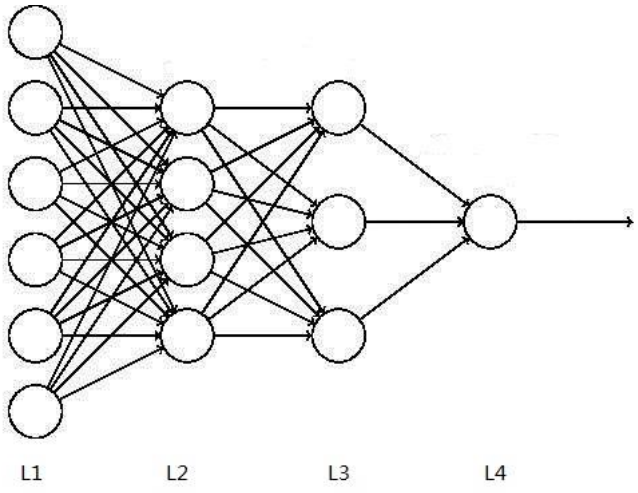

图 2-3 前馈神经网络

前馈神经网络的特点是随着网络层数的增加，参数数量过多，简单的感知机的训练方法很难适应这种参数复杂的情况。反向传播算法（error BackPropagation Algorithm，BP 算法）是目前为止最广泛应用的前馈神经网络的训练方法。其实质是最优化方法中的梯度下降法，即根据目标参数的负梯度方向进行参数调整进而寻找到最优解的算法。对给出的输入对(x,y)，BP 算法优化的函数如下：

$$\cos t(x;W,b)=\frac{1}{2}\|h(\mathrm{x})-\mathrm{y}\|^2 \tag{2-2}$$

对于每层网络的权重参数 W 和权重参数 b，根据梯度下降法计算其导数，更新公式如 2-3、2-4 所示：

$$W_{ij}=W_{ij}-\alpha\frac{\partial\cos t(x;W,b)}{\partial W_{ij}} \tag{2-3}$$

$$b_i=b_i-\alpha\frac{\partial\cos t(x;W,b)}{\partial b_i} \tag{2-4}$$

公式中的 W_{ij} 代表神经网络第 i 层中的第 j 个神经单元的权值参数，b_i 代表神经网络第 i 层的偏置参数，α 为梯度下降法的学习率。BP 算法由于参数过

多，求导过程比较繁琐，其核心思想简而言之，是通过前馈神经网络进行正向计算，求得每个神经元的输出值和激活值的差值，之后利用链式求导法则，将误差从该神经元反向传播至影响该神经元的每一个神经元，逐层反向对各个节点求残差，进而根据每个节点的残差计算导数。BP 算法通过前馈计算和误差反向传播的过程，得到节点的残差的导数，利用梯度下降法进行参数的更新。

2.1.2 循环神经网络（RNN）

循环神经网络（Recurrent Neural Network，RNN）在前馈神经网络的隐含层加入了一个环路，使得当前隐含层能够同时接收到当前的输入和上一个隐含层两个方向的信息进行计算。循环神经网络的简要结构和展开结构如下图 2-4 以及图 2-5 所示。其中图 2-4 为循环神经网络的一个神经元的节点的简单表示，图 2-5 是对于多个神经元排列在一起的神经网络的按照时间序列展开的循环神经网络结果的示意图。

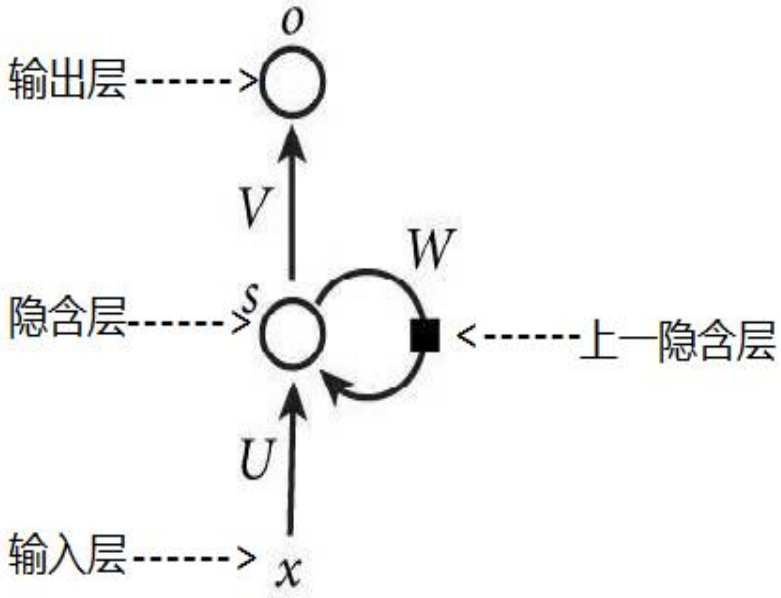

图 2-4 循环神经网络结构

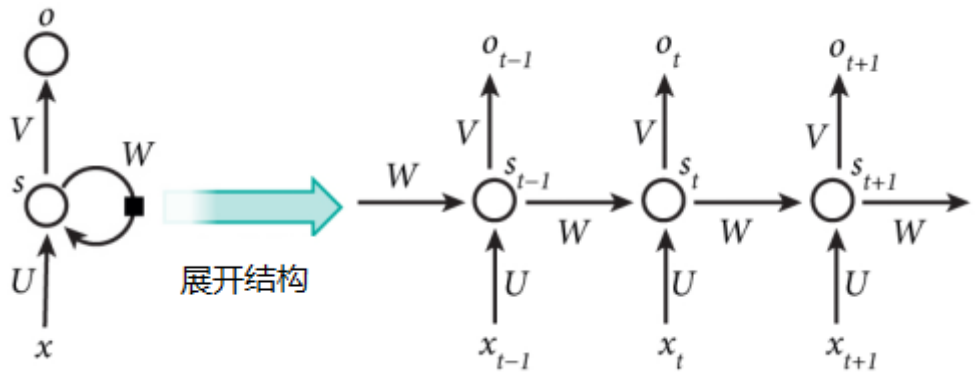

图 2-5 循环神经网络展开结构

循环神经网络相比较普通的前馈神经网络，对输入序列的各个时间节点对应的隐含层进行了联结，每个输入节点的隐含层都会从上一个输入的隐含层获取得到上一个输入的信息。图 2-5 是循环神经网络的展开结构，x 为不同时间的输入，o 为循环神经网络的输出，s 为隐含层的所有节点。与隐含层相连的权值参数中，上一节点隐含层到当前节点隐含层的权值为 W，输入层到隐含层的权值为 U，隐含层到输出层的权值为 V。根据以上权值参数，则对于 t 时刻的输出 x_t，计算该时刻的隐层时需要同时计算 x_t 和 s_{t-1}。隐含层的计算公式如公式 2-5 所示：

$$s_t = f(U \bullet x_t + W \bullet s_{t-1}) \tag{2-5}$$

RNN 模型能够在计算当前时刻的隐层时，一定程度上获取到历史信息，因此在处理序列的相关问题上有着更好的表现。

2.1.3 长短期存储单元（LSTM）

RNN 理论上在一定程度可以获取和使用当前时间节点前的所有历史信息，然而由于整体网络不断求导，导致 RNN 存在一定的梯度消失的问题（如图 2-6），距离当前节点位置较远的节点信息不断被稀疏，使得距离当前位置较远的节点信息在、实际应用中并不能被利用到。

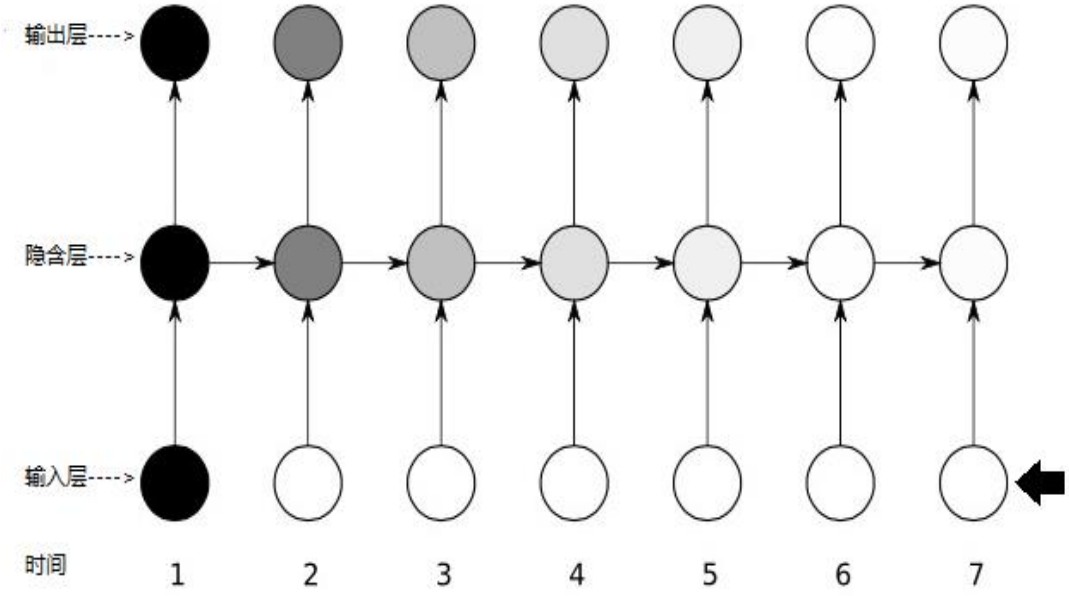

图 2-6 梯度消失问题

梯度消失实际上是指随着循环神经网络每一个时间节点的隐含层保存了历史隐含层的导数信息，然而随着时间的扩展，如图 2-6 中，颜色的深浅表示信息留存的多少，随着时间的推移，信息衰减越来越多。图中箭头指向时间序号为 7 的节点，如希望在该位置使用时间序号为 1 的节点的隐含层信息，中间会经过 6 次求导，sigmoid 函数的导数值域为[0,0.25]，利用链式求导法则将逐个隐含层单元的导数相乘，得到的乘积会越来越小趋近于 0，即梯度消失。Hochreiter 等[31]通过控制信息更为精细的长短期存储单元去代替 RNN 模型的隐含层来解决这个问题。LSTM 单元中存储单元（Memory Cell）与输入门（input gate）、输出门（output gate）、遗忘门（forget gate）相联结，进而控制和更新各个门单元的相关参数进行模型的学习和训练，即调整信息衰减、更新、去留的程度，使得存储单元能够有效的获得距离较远的历史信息。LSTM 单元的结构示意图和数据流向图如图 2-7 所示：

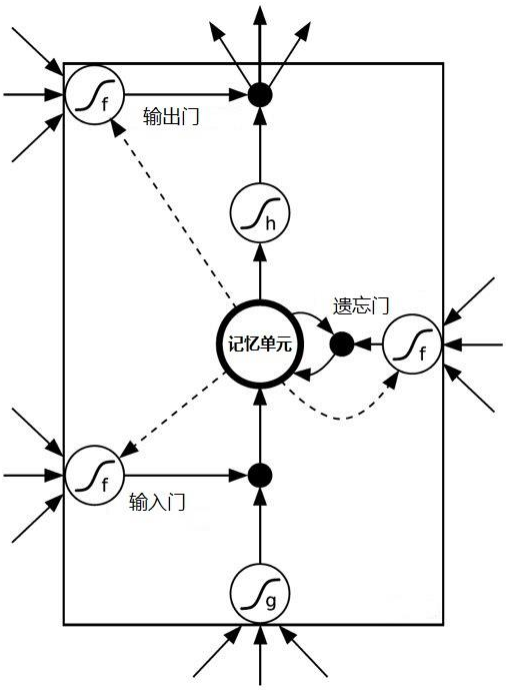

图 2-7 LSTM 单元结构

LSTM 单元的数据流入相对 RNN 模型较为复杂。对于时刻 t，LSTM 单元的输入包括输入 x_t，上一时刻的 LSTM 单元隐含层信息 h_{t-1} 和存储单元信息 c_{t-1}。并通过以下步骤计算隐含层信息 h_t 和存储单元信息 c_t。W 和 b 为权重参数，σ 为非线性变换函数，应用时常取 sigmoid 函数。

（1）首先利用输入信息 x_t 与前一个 LSTM 单元隐含层 h_{t-1} 进行如下三个门单元公式的计算：

输入门（input gate）公式：

$$i_t = \sigma(W_{x\_i}x_t + W_{h\_i}h_{t-1} + b_i) \tag{2-10}$$

输出门（output gate）公式：

$$o_t = \sigma(W_{x\_o}x_t + W_{h\_o}h_{t-1} + b_o) \tag{2-11}$$

遗忘门（forget gate）公式：

$$f_t = \sigma(W_{x\_f}x_t + W_{h\_f}h_{t-1} + b_f) \tag{2-12}$$

（2）分别计算不加入门信息和加入门信息后的存储单元值。不加入门信

息时的存储单元值如公式 2-13 所示：

$$\bar{c}_t = \tanh(W_{x\_c}x_t + W_{h\_c}h_{t-1} + b_c) \quad (2\text{-}13)$$

加入输入门和遗忘门的信息后的存储单元值利用 $\bar{c}_t$ 和前一个 LSTM 单元内存储单元计算得到：

$$c_t = f_t \bullet c_{t-1} + i_t \bullet \bar{c}_t \quad (2\text{-}14)$$

（3）通过输出门和存储单元值相计算得到该时刻 LSTM 单元隐藏层的值：

$$h_t = o_t \bullet \tanh(c_t) \quad (2\text{-}15)$$

通过上述一系列结构图和公式，总而言之：输入门通过与不加门信息的存储单元进行相乘来控制存储单元的输入信息的流入；遗忘门与上一时刻的 LSTM 单元的隐藏层的值相乘，控制上一个时刻的存储单元值的衰减程度；输出门则在 LSTM 单元运算的最后，乘以当前计算完成的存储单元值作为隐含层的输出，并对下一时刻 LSTM 单元中的门信息产生着影响。

LSTM 单元相比于 RNN，最大的优势是其通过存储单元，使得误差从输出层进行反向传播时，能够记忆下来相关的历史信息，解决了 RNN 中因为梯度消失导致的无法解决长期依赖的问题，使得 LSTM 单元在进行长序列的相关研究问题时，有着更好的表现。

2.1.4 双向 LSTM（BLSTM）

自然语言处理的研究任务，都会运用到上下文窗口信息，因为无论是字、词、短语，它们所处的上下文语境都对自然语言处理的研究尤其是序列标注问题有着很大帮助。对于经常使用的 LSTM 单元，一般是前向方向的，如果只使用前向的 LSTM 单元，模型在处理序列问题时，无法利用后文信息进行知识学习，导致对模型的效果造成负面影响。我们引入双向 LSTM（Bidirectional LSTM， BLSTM）模型，它能够联结了上文和下文两个方向的 LSTM 单元在同一时刻的输出并给出最终包含上下文信息的隐含层输出，进而提升整体模型的性能。图 2-8 是一个典型的双向 LSTM 的结构示意图。

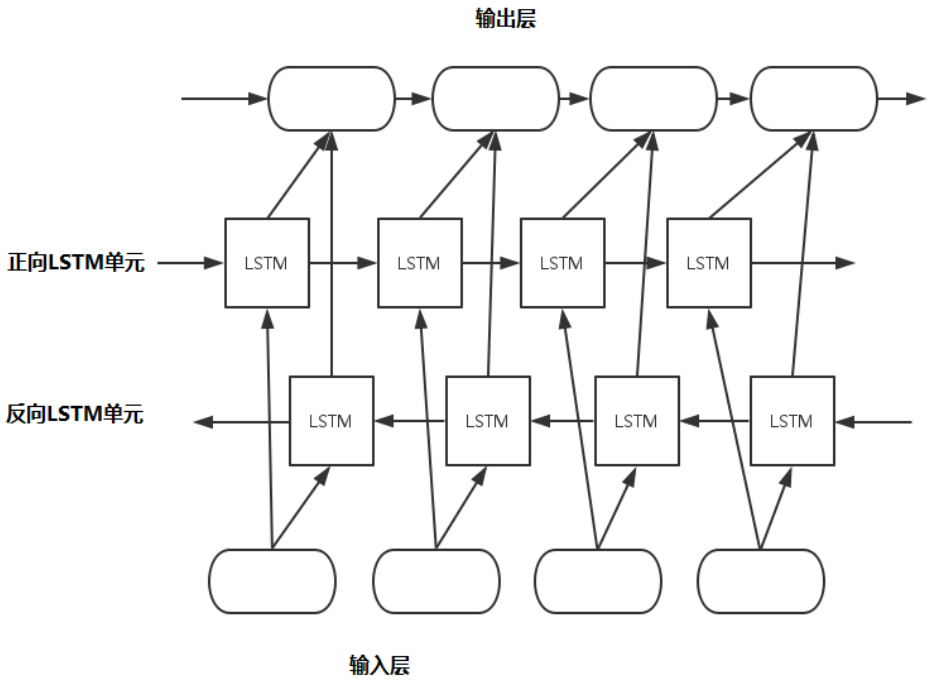

图 2-8 双向 LSTM 结构

2.2 利用条件随机场模型进行命名实体识别

条件随机场模型是在给定随机变量 X 的条件下，随机输出变量 Y 构成的马尔科夫随机场模型，是 Lafferty 等人[50]提出的概率无向图模型。由于它既能够进行序列概率的全局归一化，又能够自由设定序列的特征函数，因此被广泛使用于如分词、词性标记、命名实体识别等自然语言处理的相关应用中。

在命名实体识别任务中，经常使用的是条件随机场中的线性链条件随机场，即在给定随机变量序列 $X=(X_1,X_2,X_3....X_n)$和 $Y=(Y_1,Y_2,Y_3....Y_n)$情况下，若给定 X 的情况下，Y 的条件概率分布 $P(Y|X)$ 构成条件随机场，即满足马尔科夫性:

$$P(Y_i|X,Y_1,Y_2,...,Y_{i-1},Y_{i+1},...,Y_n)=P(Y_i|X,Y_{i-1},Y_{i+1}),\mathrm{i}=1,2,...,n \quad (2\text{-}16)$$

称 $P(Y|X)$ 为线性链条件随机场，其结构如图 2-9 所示。

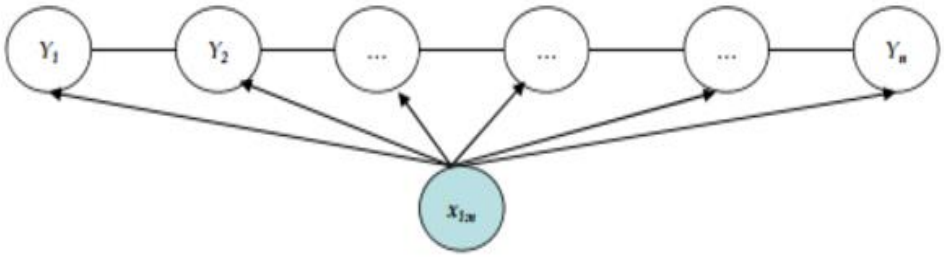

图 2-9 线性链条件随机场模型

对于给定的线性链条件随机场 $P(Y \mid X)$，可以按照下面的参数化形式进行条件概率的相关计算：

$$P(Y|X)=\frac{1}{Z(X)}\exp\sum_{t=1}^{T}\sum_{k=1}^{K}\lambda_k f_k(t,Y_t,Y_{t-1},X) \qquad (2\text{-}17)$$

其中，

$$Z(X)=\sum_{Y}\exp\sum_{t=1}^{T}\sum_{k=1}^{K}\lambda_k f_k(t,Y_t,Y_{t-1},X) \qquad (2\text{-}18)$$

公式 2-17 中 $f_k(t,Y_t,Y_{t-1},X)$ 表示了当给定输入序列中的位置 t 和输入 X，当前位置的标记 Y_t 和前一个位置的标记 Y_{t-1} 时的第 k 个特征值，λ_k 为特征权重，Z(X)为归一化因子。条件随机场模型利用前向-后向算法进行不同序列位置的条件概率和特征期望，使用拟牛顿法等极大化似然估计求解模型参数，利用维特比算法进行动态规划解码测试序列数据。

CRF 自从被提出以来，便被序列标注问题的相关研究者进行了广泛的研究和应用，也有诸多相关的开源 CRF 工具包。在本章以及之后的章节中，均使用了 CRF++0.58 工具包进行 CRF 模型训练及测试。由于面向医疗领域的中文命名实体识别的研究相对匮乏，且没有大规模公开的具有权威性的医疗语料库，因此在部分章节中将 CRF 得到的模型和训练结果作为实验的基线系统。

新闻领域的语料标签选择了 Nr，Ns，Nt 三种命名实体标记，分别代表：人名、地点名、机构名。对于某种实体构成的词，其第一个单子标记为-B，其余标记为 I，非实体标记为 O。例如：‘贾庆林在北京经济工作会上提出稳中求进’，在进行模型的训练时的输入数据为：

贾/Nr-B 庆/I 林/I 在/O 北/Ns-B 京/I 经/O 济/O 工/O 作/O 会/O 上/O 提/O 出/O 稳/O 中/O 求/O 进/O

医疗领域的语料标签，受到语料的限制只能选择 B,I,O 三种标签来标记一个词是否是一个医疗实体以及医疗实体开始和结束的边界。例如：‘查体：意识朦胧，不言语’，在进行模型的训练时的输入数据为：

查/O 体/O ：/O 意/B 识/I 朦/I 胧/I，不/B 言/I 语/I

CRF++可以根据语料和外部特征，自行选择如何加入特征。本章及之后进行的命名实体识别实验，如无特殊提及，均是基于字的不添加外部特征的命名实体识别，具体的特征规则如表 2-1 所示：

表 2-1 基础特征选择

特征	特征类别	个数
X[-2],X[-1],X[0],X[+1],X[+2]	Unigram	5
X[-2]X[-1],X[-1]X[0],X[0]X[+1],X[+1]X[+2]	Bigram	4
Y[-1]	Label Feature	1

其中 X 代表字特征，Y 代表标记特征，之后的数字代表位置相对于当前位置的偏移，即 X[0]代表当前字，X[1]代表下一个字作为特征。

2.3 基于 LSTM 的 NER 模型构建

本节主要讲述了利用不同的架构层与双向 LSTM 单元的组合，并通过模型构建时的各种细节与训练技巧，建立一个基于 LSTM 单元的完整的 NER 模型。

2.3.1 整体架构

图 2-10 展示的是基于 LSTM 单元的完整的 NER 模型的计算流程，主要分为以下 4 个步骤：

（1）完整的句子序列首先进入词向量层，词向量层训练维护了一个参数矩阵，称为词向量查找表，输入的句子能够通过这个矩阵转换为对应词的词向量的序列。查找表一般采用随机初始化得到一个初始值，并通过后续的不断训练和学习更新，形成最终的适合当前输入语料的词向量。

（2）句子序列根据词向量查找表变为词向量的序列后，根据设定的参数窗口大小将词向量进行连接，设窗口大小为 k，序列长度为 N，则得到长度为 N-k+1 的连接序列，作为 BLSTM 层的输入序列。

（3）利用随机初始化对 BLSTM 层的多个参数矩阵进行初始化，（2）步

得到的输入序列进入 BLSTM 层，即同时输入到正向 LSTM 层与反向 LSTM 层进行模型的计算和训练。为了防止出现过拟合的情况，可以在 LSTM 层的输入和输出部分加入 dropout 机制，最后拼接得到的两个方向的 LSTM 输出得到整个 BLSTM 层的输出序列作为隐含层的输出。

（4）隐含层的序列输出经过与参数矩阵相乘，得到转移概率的参数矩阵，维度为序列长度*输入标记种类个数，用来进行最终正确路径的搜索。模型训练时利用极大似然估计法进行概率计算，利用维特比算法进行测试时序列解码。

在接下来的几个小节中会对上述步骤中的细节进行详细具体的介绍。

2.3.2 窗口连接

使用单一方向的 LSTM 单元，整体模型在上下文信息的处理上，事实上是缺乏后文信息的。而双向 LSTM（Bidirectional LSTM， BLSTM）同时联结了上文和下文两个方向的 LSTM 单元，整个模型的计算量是单一方向的 LSTM 单元的双倍。而在 NER 等类似的词法分析任务中，有时候后文信息会更重要，因此使用了类似于 CRF 等模型的窗口方法，使得模型在不大幅度增加计算量的情况下，可以获取指定窗口大小的前后文信息。图 2-11 是窗口连接部分的示意图。

如指定 win 大小的前后文窗口，则位于第 i 个位置的词向量在窗口连接后的真实输入为：

$$x_i = word\_emb_{i-[\frac{win}{2}]} \oplus ... \oplus word\_emb_i \oplus ... \oplus word\_emb_{i+[\frac{win}{2}]} \quad (2\text{-}19)$$

如果窗口设置过大，导致拼接的向量过多，会出现 LSTM 模型训练极为缓慢的情况。为了防止这种情况，在实际应用中，将输入向量矩阵维度设置为序列长度乘以隐含层大小，并在训练的过程中添加一个额外的权重参数矩阵，权重参数矩阵与输入向量矩阵进行相乘，得到的结果就是 LSTM 单元真实输入的词向量序列。

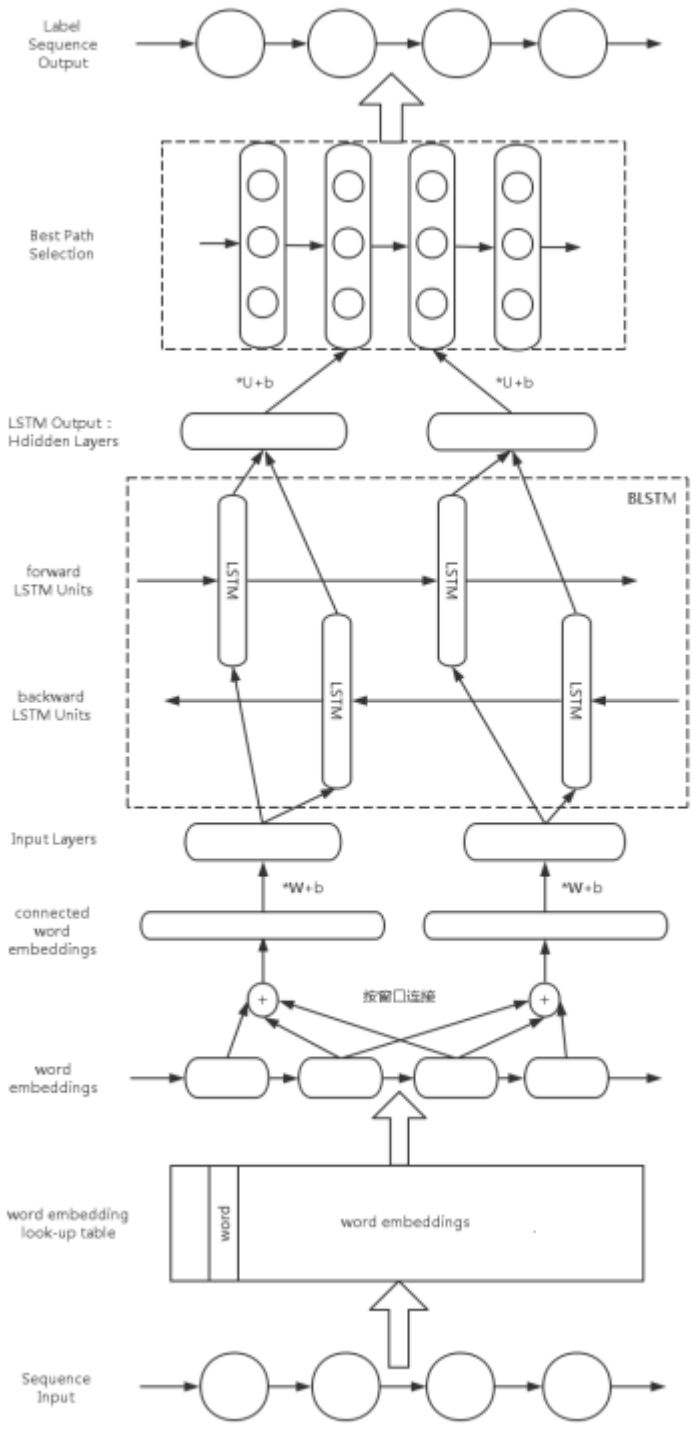

图 2-10 基于 LSTM 的 NER 模型整体框架

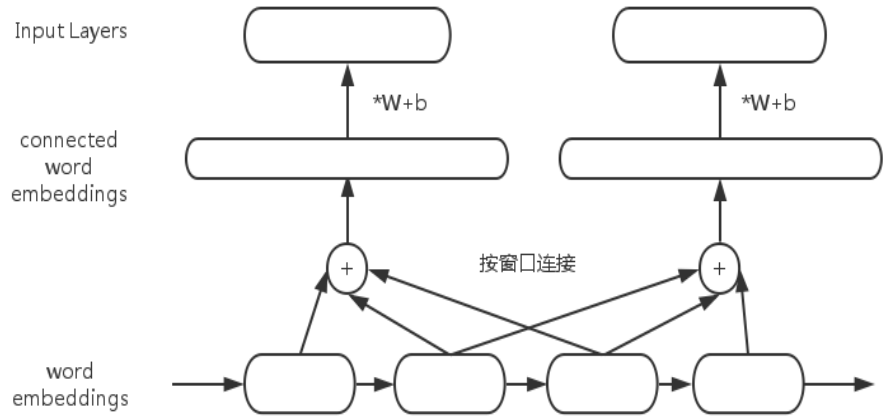

图 2-11 窗口连接示意图

2.3.3 Dropout 加入

在训练神经网络模型的时候，由于网络过深且参数过多，经常会出现过拟合的情况，为了解决神经网络的这种问题，Hinton 等[49]提出了 Dropout 技术，指在模型训练时，随机让网络中的某些隐层节点不工作，不工作的节点实质上仍会保留原有的权值，只是在本次更新权重时被跳过。

Dropout 技术已经被广泛应用到各种神经网络训练的过程中，也对防止过拟合起到了很好的作用，但是并没有一个有力的数学证明能够支撑起 Dropout。Hinton 本人在文章中给出了一个直观上的解释，他认为 Dropout 技术能够使权值的更新不依赖于固定组合的隐层节点，对于不同的输入模型都会产生共享部分节点权值局部结构不同的网络，最终的模型实际上是这些不同网络结构的 bagging 组合，使得模型能够更好地防止过拟合，并在测试集上得到更好的效果。

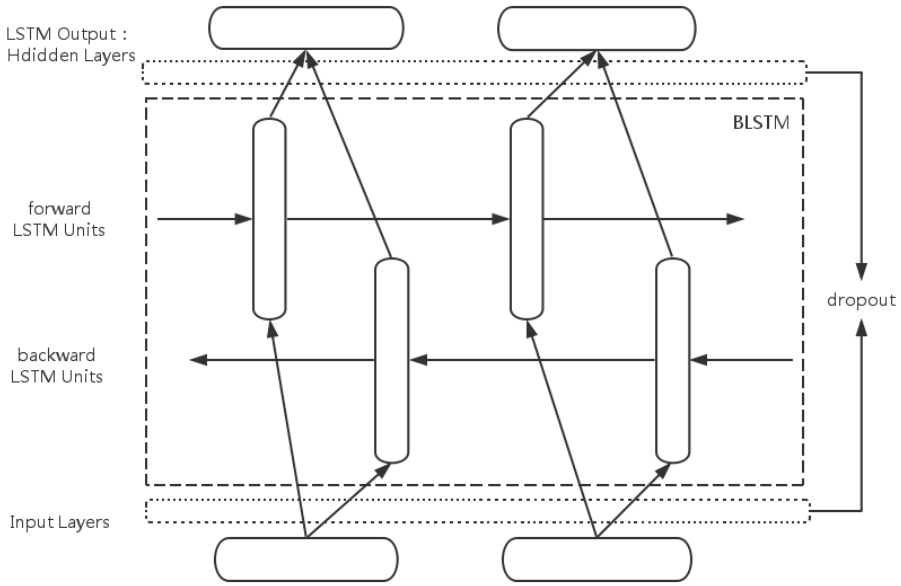

图 2-12　LSTM 的 Dropout 加入位置

如图 2-12 所示，在初始模型中，Dropout 被同时加在 LSTM 单元的输入层和 LSTM 单元的隐含层输出两端。实验证明，规模越小的语料，越容易出现过拟合的情况，Dropout 的提升效果就越明显。

2.3.4 引入转移代价的代价计算

LSTM 模型被普遍应用于如 NER，词性标注等词法分析任务之初，研究者们普遍将 LSTM 的输出利用 softmax 或最大似然函数进行所分类别的判断，这样的分类方式过少的考虑了上下文环境对于字词分类的影响。Collobert[24]等提出了模仿 CRF 等模型，计算路径的转移概率从而计算出正确路径的方法，之后被广泛应用。

转移概率的计算位于 LSTM 单元将隐含层得到序列输出之后，输出矩阵相乘维度大小为（隐含层大小*标注符号集大小）的参数矩阵后，得到了转移概率的代价矩阵，其维度大小为（序列长度*标注符号集大小）记作矩阵 F。这个矩阵的实质是输入序列中不同位置标注为不同待标注标记的概率代价矩阵。参照传统的如 CRF、HMM 等广泛应用于 NER 的模型，在模型的训练的过程中使用极大似然法进行优化，利用这个矩阵计算出标注正确结果的路径概率，

而在测试的过程中，会采用维特比算法进行最优路径的解码和选择。利用隐含层的输出进行概率计算和最优路径选择的过程如图 2-13 所示。

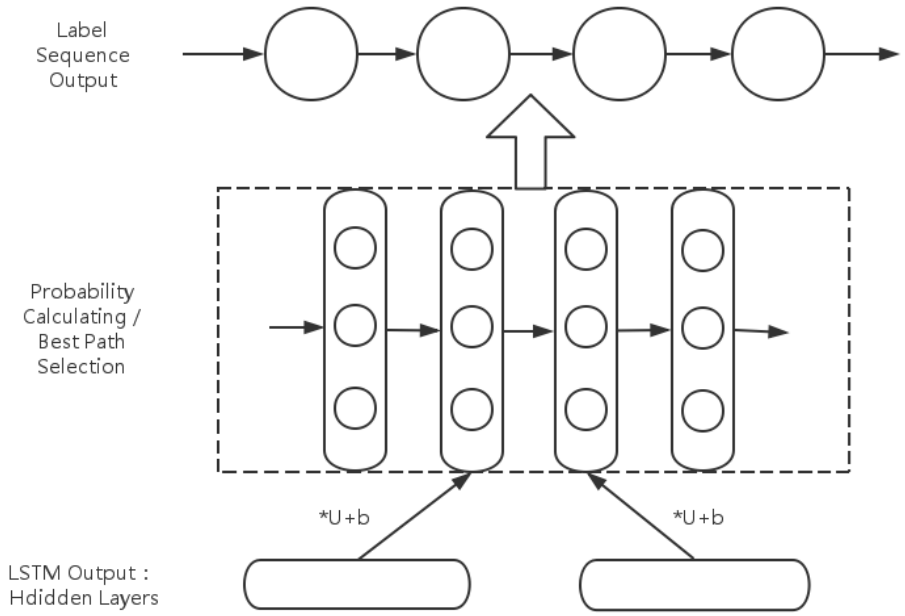

图 2-13 由隐含层输出路径示意图

在计算转移代价的时候，主要要计算路径中的不同节点，即代表的词和词之间，从一个标签转移到另一个标签的转移代价。我们通过维护和训练一个表示转移代价的矩阵 A 来进行相关的计算，A 是一个维度为标签数目的方阵，其中 A 中的元素 $A_{i,j}$ 代表了计算路径时从标签 i 到标签 j 的转移代价，对于给定的隐含层输出，设某条路径 L 的长度为 T，路径上的第 t 个标签为$L(\mathrm{t})$，则该条路径的转移代价为：

$$\mathrm{cost}(L)=\sum_{t=1}^{T}(F[t,L(t)]+\mathrm{A}_{L(\mathrm{t}\text{-}1),L(\mathrm{t})}) \qquad (2\text{-}20)$$

公式 2-20 中的$F[t,L(t)]$代表序列中第 t 个位置上标注为$L(\mathrm{t})$的代价。公式实质表示了路径 L 的总代价，为所有位置的标注代价之和，与相邻位置转移的代价之和的综合。转移代价在 NER 等任务中，对于刻画前后字词标记的关联性起到很重要的作用。如 NER 任务中采用的 BIO 标注规则，采用转移代价后可以学习到 I 只能在 B 后面出现等相关的规则，能够大幅度提高模型的效果。

在模型的训练过程中，可以利用正确路径的代价计算得到正确路径的概率：

$$p = \frac{\cos t(L_{right})}{\sum \cos t} \tag{2-21}$$

显而易见的是计算总路径代价是一个计算量庞大的任务，需要遍历矩阵中所有可能的路径。而利用动态规划算法可以在线性时间内得到正确路径的概率，极大似然化正确路径的概率后，便可以利用梯度下降法等进行问题的求解。

给定训练好的转移代价矩阵 A 和代价矩阵 F，在给定待标注语料的测试阶段，目标是求出概率最大的路径即最优路径。根据最优路径的定义，最优路径在部分的时间序列上也一定是最优的，因此可以利用维特比算法从时间的开始到结束逐步计算该序列上的最优路径，进而得到最优解，计算最优路径的同时还要标记最优路径的标注，在得到最优解后进行回溯即可得到最优路径的标注结果。对于时刻 t 的代价 v_t，定义 i 为时刻 t 前一时刻的标注，j 为当前时刻的标注，则有如下公式：

$$v_t = \max(v_{t-1} + F[t, \mathrm{j}] + A_{i,j}), \forall \mathrm{i}, \mathrm{j} \tag{2-22}$$

2.4 基于 LSTM 的 NER 实现

本节主要介绍利用 Theano 深度学习框架进行的基于 LSTM 的 NER 的实现过程和相关实验。

2.4.1 Theano 简介

Theano 作为基于 Python 的深度学习库，最大的好处是可以依靠 Numpy 数据库高效率的进行矩阵运算，这使得 Theano 可以帮助使用者充分利用 GPU 在高维矩阵运算中的强大运算能力。除此之外，Theano 编译成 C 语言代码进行运行，这使得 Theano 具有很好的稳定性。我鉴于如下几点原因，使用了 Theano 库进行了基于 LSTM 的 NER 模型的实现：

（1）方便的环境配置和 GPU 使用：Theano 无论是在 Linux，还是 Windows 系统下，都能够很容易的进行环境的安装和配置。通过对 Theano 进行简单

的参数设置就可以调用 GPU 进行计算，运算速度根据任务计算类型的不同最高可以达到 CPU 的几十倍。

（2）自动求导：Theano 最方便的一点就是可以根据程序编写的表达式自动进行梯度计算，大大的减轻了使用时在数学求导方面繁琐的工作量。

（3）相关可参考资源丰富：Theano 作为最早的深度学习的框架之一，通过 google,github 等能够收集到大量相关的深度学院源码、教程等资源。Theano 官方也公布了一系列的模型编写的教程。我在编写代码时，主要参考了 Theano 官方的 LSTM 源码和 CMU 的基于 LSTM 的 NER 的源码的编写结构，将不同的工作层进行了类的编写和封装。

虽然 Theano 作为一门函数式的编程语言，具有调试困难的缺点，且不支持并行，但是经过综合分析，还是选择 Theano 进行模型的编写。

2.4.2 Mini-batch 批量训练

Mini-batch 法更多的兼具 SGD（stochastic Gradient Descent），BGD（Batch Gradient Descent）两种最优化的方法的优点，既能节省整个批量的时间，同时可以进行多个样例的训练，同时相比于 SGD，又能够计算出更贴近的梯度收敛方向。在我们的模型中，虽然没有采用梯度下降法进行更新，但是也采用了 Mini-batch 的思想，并且由于 GPU 能够高效率的进行矩阵的计算，更能发挥 Mini-batch 的优势。

对于大小为 m 的训练集合，如果 Mini-batch 的批次样本量为 n，则共有 m/n 个 Mini-batch。实际训练中，每个 batch 的样本数不一定需要保持相同，可以根据训练情况和数据的实际情况进行动态调整，使得训练的时间和权重更新的速度达到相对最优的配比。

在序列标注的问题中，同一个 batch 内的不同序列长度并不一致，为了更高效地训练，引入了辅助的 mask 矩阵进行训练。利用 mask 矩阵进行 Mini-batch 的示意图如图 2-14。mask 矩阵的维度为同一 batch 内最长序列的长度，对于小于维度的序列，有序列输入的部分标记为 1，大于序列长度的部分为 0，这样利用 0 和 1 组成的矩阵对序列维度进行计数。训练更新某一序列时，如果 mask 矩阵该序列的某处为 1 就进行正常的更新，反之则跳过该处。

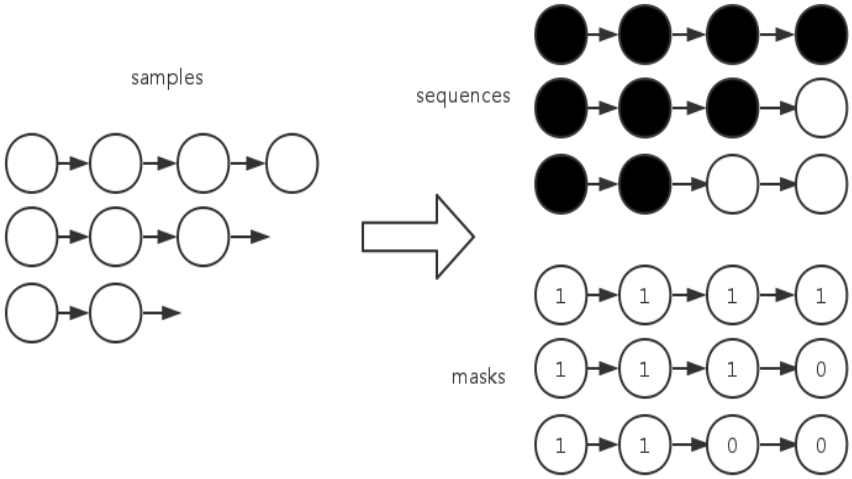

图 2-14 mini-batch 示意图

2.4.3 实验参数介绍

实验主要的可配置参数如表 2-2 所示：

表 2-2 工具包中常用的参数与默认值

参数	默认值	说明
dim_proj	128	隐含层维数
emb_dimension	100	词向量维数
win	3	前后文窗口大小
patience	10	几次 valid 测试无提升后 early stop
nepochs	5000	最多训练轮数
validFreq	50	几次 mini-batch 后进行 valid 测试
batch_size	16	mini-batch 的大小
valid_batch_size	64	交叉验证集的 mini-batch 的大小
loadModelTag	False	是否加载已训练模型
loadEmbTag	False	是否加载已训练模型的词向量参数
optimizer	adadelta	梯度拟合方式
emb_filename		加载词向量的路径
saveFreq	340	保存模型参数的频次
decay_c	0	权重衰减的参数

2.5 基于 LSTM 的单一领域 NER 实验

我们利用带有转移层的双向 LSTM 模型来进行模型的调参和有效性的验证，新闻领域的 LSTM 模型在后续章节的实验中还会进行使用。然后在医疗领域进行了 LSTM 模型的建立，并通过充分的调参和多组实验，对比证明 LSTM 模型要优于传统的条件随机场模型。

2.5.1 实验设置

2.5.1.1 实验环境

利用了 GPU 环境进行了模型的训练和调试，GPU 服务器的软硬件设置如表 2-3 所示：

表 2-3 实验所使用服务器的软硬件环境

项目	环境
GPU	Nvidia Geforce GTX Titan X
内存	16GB
硬盘	1TB
系统	Linux CentOS 6.0
Python 版本	2.7
Theano 版本	0.7

2.5.1.2 使用语料介绍

医疗领域的基于 LSTM 的命名实体识别实验的语料采用了电子病历命名实体识别语料[47]（github.com/WILAB-HIT/Resources），在后续的实验中简称‘电子病历语料’。电子病历语料是由哈尔滨工业大学网络智能实验室（WILAB）进行标注的，标注文本选取了近 1000 份电子病历的语料进行了自然语言处理相关标记的标注。语料标注的过程中，不仅邀请了专业人士进行指导，还进行了多轮的语料标注，语料整体标注质量很高。

2.5.1.3 评价标准

本章的命名实体识别的模型评价采用了传统命名实体识别中常用的精确率（Precision）、召回率（Recall）和 F 值（F-score）这三个指标。精确率（Precision）、召回率（Recall）和 F 值（F-score）的评价体系是命名实体识别相关研究主要采用的评价标注。精确率能够表示预测结果中预测出的正例样本有多少是真正的正例，召回率能够表示出标准答案中的正例样本有多少被正确预测。F 值即为精确率和召回率的调和平均值，能够平衡衡量精确率和召回率两个指

标。如果 *correct* 表示标注正确的实体个数，*recognized* 表示总共标注出的实体个数，*entities* 为标准答案总共包含的实体个数，则精确率、召回率、F 值的具体计算方法如下公式所示：

$$Precision = \frac{correct}{recognized} \tag{2-23}$$

$$Recall = \frac{correct}{entities} \tag{2-24}$$

$$F = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{2-25}$$

2.5.2 医疗领域的 LSTM 实验

在本节中通过控制变量法逐个参数进行医疗领域的 LSTM 模型的调参。通过对电子病历语料随机抽取 90%作为训练集，剩余的 10%作为测试集，并从测试集中随机抽取了 20%作为验证集的语料配比，进行了多次实验取平均值，得到如下结果。

2.5.2.1 词向量层大小调整实验

首先进行了 3 组不同大小的词向量层的对比实验，固定隐含层大小为 1024，窗口大小为 3，同时在输出层和输入层加入 dropout 机制，比较不同大小的词向量层对实验结果的影响。

表 2-4 词向量层大小调整实验结果

词向量层大小	Precision	Recall	F-score
100	88.25	87.24	**87.74**
200	88.08	87.09	87.58
300	87.09	87.43	87.26
基线系统：CRF	89.69	81.67	85.49

通过调整词向量层的大小，发现随着词向量的增大，整体效果呈现下降的趋势，这可能是因为扩展了词向量的维度以后，冗余的信息变多，降低了模型的效果，在后续的实验中，将固定词向量层大小固定为 100 维。

2.5.2.2 隐含层大小调整实验

在进行完词向量大小的实验后，进行隐含层的相关实验调参，固定词向量维度为 100，窗口大小为 3，同时在输入层和输入层加入 dropout 机制，比较不同大小的隐含层向量对实验结果的影响。

表 2-5 隐含层大小调整实验结果

隐含层大小	Precision	Recall	F-score
128	87.88	86.61	87.24
256	88.03	87.39	87.71
512	88.96	87.49	**88.22**
1024	88.25	87.24	87.74
1536	88.1	87.28	87.69
基线系统：CRF	89.69	81.67	85.49

通过实验结果可以看出，控制其他变量相同，隐含层大小在 512 的情况下，模型取得了最好的效果，随着隐含层的再次变高，模型效果变差，可能是因为隐层向量维度的增大，导致了特征过多，产生了一定程度的过拟合，在后续实验中，会利用加入多层 dropout 的方法来针对高维隐含层向量中出现的过拟合问题。

2.5.2.3 综合对比实验

表 2-6 双向 LSTM 结果

序号	词向量维度	隐层向量维度	窗口大小	Dropout 配置	F 值
1	100	128	3	输出层+输入层	87.24
2	100	256	3	输出层+输入层	87.71
3	100	1024	3	输出层+输入层	87.74
4	100	512	3	输出层+输入层	88.22
5	100	1024	3	输出层+输入层+输入层	**88.36**
6	100	512	5	输出层+输入层	87.71
7	200	1024	3	输出层+输入层	87.58
8	100	512	3	输出层+输入层+输入层	87.69
9	基线系统:CRF 模型				85.49

在前两小节通过比对实验进行了隐含层和词向量层大小维度的确定，本小节将在固定词向量维度为 100 的情况进行综合的比对，相比于前两小节的对比实验，在本节我们会将之前的实验和其它一些不同参数配置的实验进行比较，包括对不同参数组合和不同 dropout 设置时的结果如表 2-7 所示。

通过结果的比对可以看出，LSTM 模型的整体效果都要明显好于传统的 C

RF 模型，相比于 CRF 模型的基线系统效果在 F 值上最多提升达到了 2.87。由于基于字的命名实体任务，如果利用传统的机器学习模型进行研究，能够抽取的特征相对有限，而 LSTM 模型能够充分发挥深度学习在抽取深层次特征上的优势，并对深层次特征进行表达，达到更好的效果。同时，由于医疗领域语料的相对缺乏，能够看出当隐层向量维度过高如设置隐层向量维度为 1024 时，产生了一定程度的过拟合，添加 3 层的 dropout 后能够有效提高了模型的效果，而在隐层向量维度为 512 时的过拟合现象并不明显，添加 3 层的 dropout 反而会产生欠拟合的问题，同时增大窗口并没有对模型产生正面的效果，可能由于增大窗口后，特征向量维度增大，反而产生了过拟合。

2.5.3 医疗领域的实际应用测试

通过上一节的多次实验的对比和效果分析，已经能够明显看出在命名实体识别问题的研究上，LSTM 模型要显著好于传统的 CRF 模型。在本节将从实际应用的角度出发，通过在真实的无标注不规范的大规模语料上进行 LSTM 模型和 CRF 模型的比对，来证明 LSTM 模型在实际的应用中，也具有比 CRF 模型更好的对未知实体的识别能力和延展性。

在本节中利用随机抽取的 90%的电子病历语料训练了词向量维度为 100，隐层向量维度为 1024，窗口大小为 3，添加 3 层 dropout 的医疗领域的 LSTM 模型，与日常应用中广泛应用的基于词的 CRF 模型的命名实体识别模型进行对比。测试语料为天狗医疗网爬取的医疗文本数据、好大夫网的医疗数据以及医疗问答对数据，共约 200M。测试参考词典为电子病历语料抽取的医疗名词词典 6790 词，以及通过搜狗词表信息提取出的医疗相关词典共 359761 词。希望能够对比两个模型在测试语料上识别出的不在电子病历词典，而出现在搜狗医疗词典中的实体个数，来反应两个模型在实际应用中的识别未登录词的效果好坏。

在本节之后的测试结果分析中，我们定义新词表示不在电子病历词典，而出现在搜狗医疗词典中的实体词汇，定义待定义词表示既不在电子病历词典，也不在搜狗医疗词典中的实体词汇。基于上述定义，分别采用 LSTM 模型和基于 CRF 模型的词的命名实体识别模型进行了测试语料上的相关测试。

基于 CRF 模型的词的命名实体识别模型在测试语料上一共识别出 5851 个新词，同时还识别出 141934 个待定义词。基于 LSTM 的命名实体识别模型在测试语料上一共识别出 11157 个新词，同时还识别出 253401 个待定义词。通

过比对两个模型在新词和待定义词的发现能力，基于 LSTM 的命名实体识别在新词和待定义词的识别效果都明显好于 CRF 模型。图 2-15 的组合图能够更加直观的反应两个模型在这两方面的效果差别，其中主坐标轴代表了发现的新词个数，次坐标轴代表了待定义词个数。

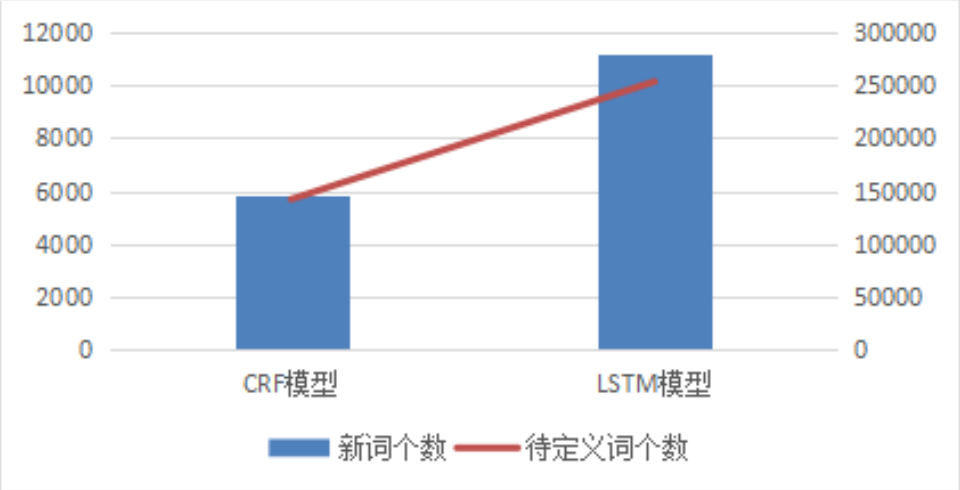

图 2-15 新词和待定义词的发现效果比较

为了能够更细致的分析 LSTM 模型在未登录词的扩展能力上的效果，我们将 LSTM 得到的 11157 个新词和 CRF 模型得到的 5851 个新词依据出现的词频进行了排序，并摘录了部分词频高的新词展示如表 2-8 所示：

对于两个模型发掘出的新词对比，能够分析得到如下几点结论：

（1）相比于 CRF 模型，LSTM 模型无论是标注出的实体个数还是每个实体的频次都要明显多于 CRF 模型，对于表中所列出的前 30 位的高频新词，LSTM 模型在每个相同排位的词的频次也都显著高于 CRF 模型，这从一定程度上证明了 LSTM 模型挖掘实体，根据训练语料发掘未登录词的能力要更强。

（2）利用加黑将测试结果中出现的不太具有医疗含义的词汇标注出来，可以看出 CRF 模型含有诸如‘工商银行’，‘建设银行’等明显不属于医疗领域的标注结果，而 LSTM 模型中‘补充’，‘保护’属于用途较为广泛且高频词出现在医疗语料中的词汇。相比于 CRF 模型，LSTM 模型对文本的语义理解能力更好，通过把字转化为向量后再进行标注，能够更好的学习到相关的语义知识，并将无医疗研究价值的标注结果排除，好于传统的 CRF 模型。

表 2-7　CRF 和 LSTM 发现新词的对比

CRF 模型的高频词	次数	LSTM 模型的高频词	次数
低盐饮食	1335	高压	22598
低脂低盐饮食	620	保持	4897
硝苯地平缓释片	499	低压	3989
手术后	456	低盐饮食	1668
术后	409	体育锻炼	1121
原发性高血压	327	湿热	849
卡托普利	312	调理	814
手术前	251	心痛	771
肝肾功能	221	清热	729
功能障碍	184	心跳	683
脑血管疾病	166	创伤	678
继发性高血压	166	**补充**	662
降压灵	151	损伤	657
病毒感染	148	白术	634
诊断标准	147	葡糖	613
颅内出血	135	**保护**	567
尼莫地平	128	肝肾功能	491
淋巴结肿大	124	护理	467
上消化道出血	123	甘草	464
血压下降	118	血肿	463
支持疗法	117	多喝水	448
工商银行	116	肝素	441
建设银行	115	食疗	419
中国银行	115	恶心呕吐	383
多吃	115	氢氯噻嗪	371
全身症状	114	保险	367
高血压疾病	110	泽泻	358
精神障碍	110	舒张压	355
图比灵	109	疱疹	351
诊断	94	辛伐他汀	340

2.6 本章小结

在本章中主要进行了利用深度学习的 LSTM 模型来进行单一领域的 NER 的相关实验。鉴于后续章节还要继续用到相关的模型和方法，因此在本章中，对 RNN、LSTM、CRF 等模型的相关原理、研究方法和应用都进行了详细介绍，并通过模型的分部构建的方式、整体的结构和模型每个部分的应用技巧来

逐步介绍如何根据已有的模型知识进行基于 LSTM 的 NER 模型的搭建，最后利用了 Theano 深度学习框架将方法和模型进行了实现。

在实现好的模型的基础上，通过多组不同参数设置的实验对比，证明了 LSTM 模型在经过细致调参后，在医疗领域对命名实体的识别效果相比 CRF 模型有了明显的提升。通过不同参数的控制变量的调节，也找出了如何进行参数配比，使得 LSTM 模型的效果能够更加出色。为了验证 LSTM 模型在无标注不规范的语料上的效果，又借助大批量的外部词典进行了定性实验，展示出了 LSTM 模型无论是在研究过程还是在实际的应用中，相比 CRF 模型都具有更出色的效果和延展性。

然而在实际的研究和使用中，医疗领域的规范的带标注语料极为缺少，很多时候并不能够支撑 LSTM 模型的训练，因此如何利用新闻领域的语料进行模型泛化语言学知识的学习，并在这一基础上，进行医疗领域的模型的训练，使得达到更出色的效果，将是下一节要关注和解决的问题。

第3章 融合外部信息的跨领域的命名实体识别

第二章主要从 LSTM 模型的发展、模型的实现、模型的效果比对等诸多方面展示了基于 LSTM 的命名实体识别模型。通过多次在单一领域语料上的实验，和医疗领域的大规模无标注的不规范的语料，来印证 LSTM 模型在命名实体识别任务上的出色效果。

然而在实际的研究任务和应用过程中，医疗领域的规范标注的语料极为稀缺，很多时候并不能支撑起一个无论在测试集还是实际效果都出色的模型。如果能够让模型通过利用数量庞大的新闻领域的语料学习到一定的泛化的语言学知识，再进行医疗领域知识的学习，则能够在一定程度上解决医疗领域语料规范的标注数据缺少的问题。

在本章中，我们尝试了如何利用新闻领域的 LSTM 模型，帮助医疗领域的 LSTM 模型更好的进行语言学知识的学习。通过深度学习领域中常见的预训练方法，能够在无标注的医疗领域语料上进行无监督学习得到不同字或词的语义向量表示，将这样的语义向量表示导入到医疗领域的模型中，对于资源稀缺领域的模型的学习十分重要。除此之外，还通过迁移学习的相关思想，将训练好的新闻领域的模型参数加载到医疗领域的模型中进行模型的学习和训练，并与没有进行迁移学习的实验进行效果比对，证明迁移学习的有效性。

3.1 深度学习中的迁移学习

随着深度学习的不断铺开，如何通过更好的训练方法，训练出效果更好且泛化能力更强的模型，成为诸多研究者不断尝试和学习的方向。Hinton 等[18]给出了先对置信网络的每一层预训练，然后连接整体网络进行精细调参的方法，是目前深度学习中的 Fine-Tuning 的方法的雏形。

Fine-Tuning 实质上是指对深度学习模型的参数，通过某种方法微调的统称，而迁移学习正是一种借鉴了 Fine-Tuning 思想的机器学习分支。对于在某些新兴领域中如果进行机器学习中的有监督学习模型的构建，很难得到足够支持起模型的数据量。然而由于任务和数据的共同性和互通性，可以通过迁移学习通过从其他外部环境进行有监督学习获得到的知识，应用于新的学习环境中。虽然原有的学习环境和新的环境基本不符合相同的数据分布，但是通过迁移学习学习到的知识却蕴含了不同学习任务中的共性知识，并在学习到的共性

知识的基础上，进行特定领域的知识学习，即知识的‘迁移’。

迁移学习的目的，实质上就是当源环境和目标环境的数据不符合独立同分布的概率学条件时，如何利用源环境的数据进行学习和模型的搭建，使得目标环境的模型具有更好的效果，更好的泛化能力。为此，需要通过不同的学习方式的迁移学习，使得模型能够学习到不同环境的共性知识，并在学习到共性知识后进行微调以达到更好的学习效果。迁移学习主要分为基于特征的迁移、基于模型的迁移和基于样本的迁移。基于特征的迁移学习，本质上就是将源环境和目标环境的解空间映射到相同的整体空间中，通过最小化两个解空间的距离，使得从源环境的空间学习到的模型迁移到新的目标环境之中。基于模型的迁移的目的，则是利用在源环境学习到的模型的参数，与目标环境的训练数据相结合，这样训练得到的模型既包含了源环境的先验知识，又能够有效地适应目标环境的数据分布，具有更好的泛化能力。而基于关系的迁移，则是通过一定程度的类别，将源环境中不同概念之间的关系，映射到相似的目标环境中的不同概念的关系上，进而完成不同环境的知识的迁移。综上来看，无论哪一种迁移学习的学习方式，实质都是通过一定程度的表示，将源环境和目标环境进行连接和对比，因此，如果源环境和目标环境，具有很高的相似性和共性，那么学习到的模型的能力就相对更好，这也是迁移学习中最重要的先决条件。

深度学习作为目前最流行的一种机器学习的方式，在之前的章节中已经大篇幅的介绍了相关的内容。然而深度神经网络发挥其强大的抽取深层隐含特征的前提是已标注的数据量能够足够驱动整体模型的构建，且由于深度学习隐层多，隐层节点多的特点，很容易在缺乏标注数据的领域中构建模型时，发生不同程度的过拟合现象，进而影响整个模型的泛化能力和实际应用的效果。而通过迁移学习，可以使整体模型在拥有充足语料的知识域中进行初步的知识学习和模型构建，并在目标领域中一定程度地规划了模型训练的方向，加速了整体模型的学习速率，并能够在一定程度上防止模型出现过拟合的情况。

综上来看，迁移学习和深度学习相结合的方式，能够在多个不同知识域中的知识学习和模型构建中发挥巨大的作用。自然语言处理的研究中，会经常出现不同知识领域的相同的问题的模型构建，如果能应用到相关的迁移学习的知识，将是大有裨益的。

3.2 深度学习中的预训练技巧

对于利用词袋模型表示每个词并进行组合成标量的输入序列，首先要利用

word embedding 查找表将输入序列转化为带有语义信息的词向量序列，具体的转化过程如图 3-1。输入的标量表示与 word embedding 查找表相乘，得到对应的词向量序列。word embedding 查找表实质上是一个表示词向量的矩阵，在整体模型训练的过程中，不断更新为更能表征任务和训练语料特征的词向量矩阵。查找表的初始化用两种方式，一种是采用一定值域上的随机初始化，还可以利用 Turian 等[48]提出的方法，利用神经网络语言模型或词向量工具，在大规模无标注语料上生成词向量，并进行加载。

在我们的模型中，首先使用了 word2vec[27]工具在与训练语料领域相同的大规模无标注语料上进行了词向量的训练，更完整地学习到词在大规模语料上的语义向量表示。设词表中的第 i 词的词向量为 vec_i。对于训练语料词表中的所有词，按照序号 i 的顺序，得到代表词向量表示的矩阵 vec。随机初始化的初始词向量矩阵，记为 $word\_emb$。且 vec 与 $word\_emb$ 维度相同，则有以下的赋值公式：

$$word\_emb_i = vec_i, \forall i \in input \cap voc \tag{3-1}$$

$input$ 表示训练语料得到的词表，voc 代表大规模预训练语料的词表。

预训练最主要的目的，是利用无监督学习的优势，在大规模语料上学习到目标词汇更完整和贴近真实分布的语义向量表示。除此之外，在标注数据稀缺的领域如医疗领域，预训练能够更好的引入大规模语料上的外部知识，更小概率陷入局部最优的问题中。且预训练得到词向量的过程较快，也能够使整体模型在训练时，沿着更容易得到最优解的方向进行迭代，既能够提高模型的效果，又加快了模型训练的时间。

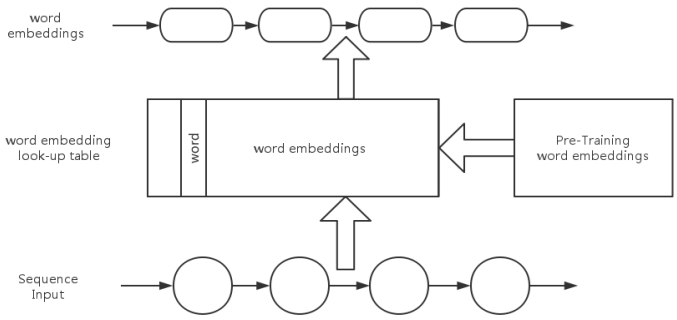

图 3-1 预训练示意图

3.3 融合外部信息的跨领域的命名实体识别

本节将通过迁移学习和预训练的方法，利用新闻领域的语料和无监督的医疗语料的信息，将新闻领域的 LSTM 模型的权重矩阵，融入到训练参数相同的医疗领域的 LSTM 模型中，并在其中加入利用无监督语料预训练得到的词向量，对医疗领域的命名实体识别模型的效果进行提升，实验的整体流程图如图 3-2 所示：

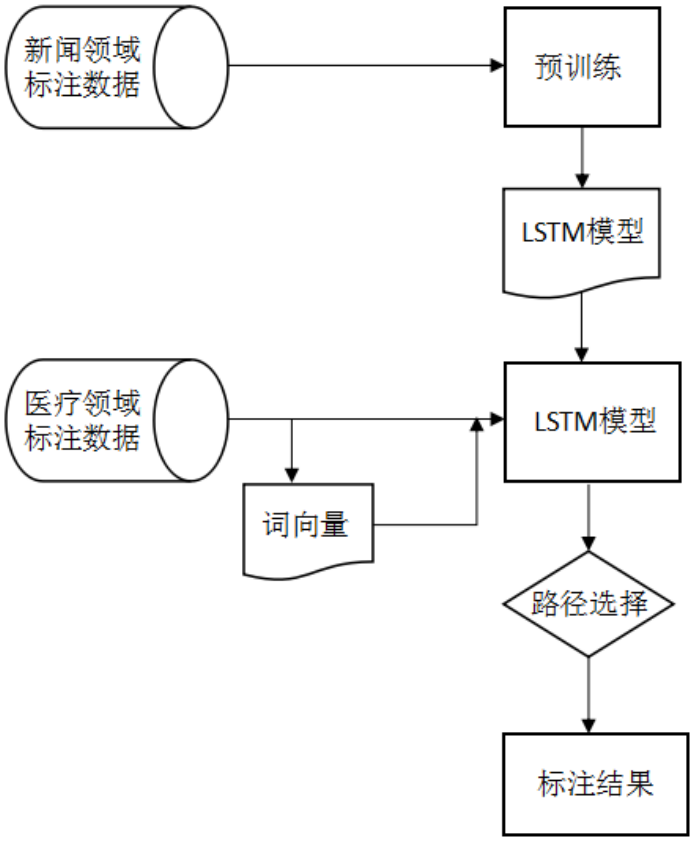

图 3-2 实验的整体流程图

在本节的实验中，设置了是否融入新闻领域模型、是否加入预训练以及不同的训练测试语料选择等不同的实验组合进行探究，来对比和验证我们的方法在医疗领域的命名实体识别的研究上的有效性。

3.3.1 实验设置

3.3.1.1 实验环境

本章实验的软硬件环境与第二章相同如表 3-1 所示：

表 3-1 实验所使用服务器的软硬件环境

项目	环境
GPU	Nvidia Geforce GTX Titan X
内存	16GB
硬盘	1TB
系统	Linux CentOS 6.0
Python 版本	2.7
Theano 版本	0.7

3.3.1.2 实验设置

本章实验分别在新闻领域的医疗领域两个领域进行了 LSTM 模型的构建，两个领域的标注语料均采用 BIO 标注规则。其中新闻领域实验的训练语料设置为《人民日报》2 月份语料，测试语料设置为《人民日报》1 月份语料，验证集语料随机抽取了测试语料的 5%进行验证。

医疗领域的实验仍然采用电子病历语料，进行了 3 种语料不同配比的实验，分别为：

（1）随机抽取 90%作为训练集，剩余 10%作为测试集，测试集中抽取全部语料的 2%作为验证集。

（2）随机抽取 30%作为训练集，剩余 70%作为测试集，测试集中抽取全部语料的 10%作为验证集。

（3）随机抽取 10%作为训练集，剩余 90%作为测试集，测试集中抽取全部语料的 10%作为验证集。

对于医疗领域和新闻领域的 LSTM 模型，均采用以下表 3-2 的参数配置进行模型的训练。本章及之后训练模型时用到的医疗领域词向量，是通过 word2vec[27]工具在电子病历命名实体识别语料[47]训练得到的 100 维词向量。

对于模型的评价，仍然采用了传统命名实体识别中常用的精确率（Precision）、召回率（Recall）和 F 值（F-score）这三个指标，已在 2.4.1.3 节中进行了相关介绍。

表 3-2 实验模型的参数设置

参数	默认值	说明
dim_proj	1024	隐含层维数
emb_dimension	100	词向量维数
win	3	前后文窗口大小
patience	10	几次 valid 测试无提升后 early stop
nepochs	5000	最多训练轮数
validFreq	50	几次 mini-batch 后进行 valid 测试
batch_size	16	mini-batch 的大小
valid_batch_size	64	交叉验证集的 mini-batch 的大小
optimizer	adadelta	梯度拟合方式
saveFreq	340	保存模型参数的频次
decay_c	0	权重衰减的参数

3.3.2 融合外部信息的跨领域的命名实体识别

3.3.2.1 加载新闻领域模型的比对实验

首先我利用已经训练好的新闻领域的模型的权重参数，加载到待训练的医疗领域的模型参数中，加载的模型参数包括 LSTM 隐层参数矩阵和转移层参数矩阵。实验具体设置和实验结果见表 3-3:：

表 3-3 加载模型参数的对比实验结果

序号	语料配比	是否加载新闻领域模型	训练集 F 值	验证集 F 值	测试集 F 值
1	随机 90%	是	96.78	88.16	87.92
2		否	94.51	88.72	**88.36**
3	随机 30%	是	97.52	86.6	**85.14**
4		否	94.63	87.3	85.05
5	随机 10%	是	97.51	79.3	78.82
6		否	96.74	81.09	**80.24**

通过表 3-3 可以发现，只单纯融入新闻领域的模型特征后，医疗领域的 LSTM 模型的整体效果是有所下滑的，这是因为医疗领域和新闻领域虽然都能够学习到语言学的泛化知识，但是不同的领域的知识分布差距很大，导致模型从新闻领域迁移到医疗领域时，不能够充分地学习到医疗领域的知识表示。除此

之外，还能够发现加载了新闻领域的模型，在训练集上达到的最好效果都比未加载效果要有明显提高，这可能是因为模型在通过加载大规模的新闻领域模型后，学习到了更完全的语言学的泛化知识，虽然在验证集和测试集上出现了一定程度的过拟合的现象，在后续的小节将会通过加入预训练词向量和综合调参的方式来改善这一问题。

3.3.2.2 加载预训练词向量的对比实验

上一小节进行了几组是否加载新闻领域的模型参数的对比实验，通过实验可以发现在加载模型的参数的情况下，产生了一定程度的过拟合情况，影响了模型的效果。本节将在不加载新闻领域的模型的情况下，进行了不同语料配比的是否加载预训练词向量的测试，来比对加入了词向量以后对模型效果产生的影响，具体实验设置及结果见表 3-4：

表 3-4 加载预训练词向量的对比实验结果

序号	训练语料配比	是否加载预训练词向量	训练集 F 值	验证集 F 值	测试集 F 值
1	随机 90%	是	94.65	88.8	**88.41**
2		否	94.51	88.72	88.36
3	随机 30%	是	93.52	87.55	**85.59**
4		否	94.63	87.3	85.05
5	随机 10%	是	94.27	82.61	**81.07**
6		否	96.74	81.09	80.24

通过表 3-4 的相同语料配比的不同加载词向量状态的对比实验可以发现，加入词向量以后，模型整体在测试集上有更好的效果。加入词向量之后模型的效果提升随着训练语料比例的增加而减少，这是因为预训练的词向量带有从训练语料中无监督学习得到的语义知识，当训练语料较少的时候能够有效地弥补训练语料由于数量不充足导致无法学习到完整的语义信息的问题。训练语料充足时，加入词向量提升效果有限。

3.3.2.3 混合加载词向量和新闻领域模型的对比实验

通过前两个小节的实验对比，加载新闻领域模型参数后，模型实际上出现了一定程度的过拟合现象，而加载预训练词向量后，模型在测试集上有着更好的效果。本节通过组合加载新闻领域模型和预训练词向量进行对比实验并综合不同的组合，分析这两种方式对模型的效果影响。

表 3-5 混合方法的对比实验结果

序号	训练语料配比	是否加载预训练词向量	是否加载新闻领域模型	训练集 F 值	验证集 F 值	测试集 F 值
1	随机 90%	否	否	94.51	88.72	88.36
2		是	否	94.65	88.8	88.41
3		否	是	96.78	88.16	87.92
4		是	是	96.26	89.03	**88.52**
5	随机 30%	否	否	94.63	87.3	85.05
6		是	否	93.52	87.55	85.59
7		否	是	97.52	86.6	85.14
8		是	是	95.36	87.97	**85.95**
9	随机 10%	否	否	96.74	81.09	80.24
10		是	否	94.27	82.61	81.07
11		否	是	97.51	79.3	78.82
12		是	是	96.0	83.15	**81.6**

从表 3-5 的实验对比可以发现，在所有的 3 组不同语料配比中，同时加载预训练词向量和新闻领域模型参数的效果在每组中都是效果最好的，对比只单独加载预训练词向量和新闻领域模型的对比实验的效果对比，可能是因为单独加入预训练词向量时，模型能够更好的抽取医疗领域的语义特征，加载新闻领域的模型时，模型能学习到新闻领域的泛化的语义表示和语言学特征。虽然单独加入新闻领域模型，效果会因为领域差异而下降，但是同时加入新闻领域模型和预训练词向量时，模型能够同时从上述两个方面学习到医疗领域的语义特征和泛化的语言学特征，进一步地提升模型的整体效果。

3.4 本章小结

在本章中进行了融合外部信息的跨领域的命名实体识别，即在拥有大规模新闻领域语料和小规模目标领域的情况下，如何通过融合跨领域和领域内的知识信息，进而提高命名实体识别模型的相关效果。

我们首先介绍了深度学习中的迁移学习方法，通过迁移学习能够从与目标领域不同的领域中学习到泛化的语言学知识并进行迁移，进而增强目标领域模型的泛化效果，一定程度弥补目标领域语料规模小，无法支撑模型训练的问

题。然后又对深度学习中的预训练技巧进行了概述，通过如 wor2vec 等无监督抽取目标领域语料语义的方式生成目标领域的词向量，并将词向量加载到相关深度学习模型的词向量层，能够加快模型的收敛并提升模型的效果。

在之后的小节中，对上述介绍的方法进行了相关的对比实验，对两种方法是否使用和不同语料配比等各种实验条件的组合进行设置，通过充分的实验数据进行对比分析各个方法加入时产生的效果优劣和可能原因，并通过最后的多条件的组合的实验证明关于已有实验结论的分析的正确性，并说明我们的方法能够在一定程度上，利用新闻领域的语料和医疗领域的两部分语料进行外部信息的学习和融合，进而对医疗领域的命名实体识别模型的效果进行提高。

第4章 领域自适应的弱监督命名实体识别研究

在第二章和第三章中通过深度学习中的 LSTM 模型进行了命名实体识别的相关研究，首先利用医疗领域的电子病历语料建立了基于 LSTM 的命名实体识别模型，并证明模型的效果显著好于传统的条件随机场模型。然后通过两种不同的利用新闻领域和医疗领域语料的方法增强了基于 LSTM 的命名实体识别模型的效果。

虽然我们的模型能够通过迁移学习的方式来将新闻领域的语言学特征以模型参数的形式融入到医疗领域的模型中，事实上单独融入新闻领域的模型参数对医疗领域的模型的效果是消极影响的。除此之外，LSTM 模型在实际应用中，相比于传统的条件随机场模型，存在诸如未登录词识别能力稍差，训练测试效率低等明显缺点。如果能够找寻到一种衡量不同领域间领域差异的方式，利用传统机器学习中的特征工程，将新闻领域的表示命名实体识别相关信息的特征向量，利用这种方式转化为医疗领域的特征向量，并融入医疗领域的无监督信息进行分类的话，能够在兼顾使用效率的同时，达到出色的实验效果。

在本章中将利用几组基于 CRF 模型的命名实体识别的对比实验，来探究新闻领域和医疗领域语料之间的差异，随后利用梯度优化决策树（GBDT，Gradient Boosted Decision Tree）模型进行新闻领域和医疗领域的跨领域的弱监督的命名实体识别研究的相关研究。

4.1 探究新闻领域和医疗领域语料的差异程度和影响

随着对命名实体识别的研究的逐渐深入，新闻领域由于规范的标注语料数量充足，很容易进行不同实验思路的设计和实现。然而对于标注语料稀缺的领域如医疗领域，无法直接使用新闻领域的模型进行实验。如何衡量新闻领域和语料稀缺领域的领域差异，量化领域差异并用于领域间的迁移学习，成为了目前命名实体识别研究的一大热点。相关的工作在英文命名实体识别中已经进行了一定程度的开展，Kulkarni 等[51]通过分析不同知识领域的的语言学差异，利用相同词在不同领域内的 word embedding 表示和概率分布来衡量不同知识领域的语义差异。Qu 等[52]在目前命名实体识别最流行的双向 LSTM 模型的基础上，利用相同词在不同领域内标注不同，学习不同领域内不同标注标签的差异，并加入类似自编码器的层设置，将源语言域的标签映射到目标领域中来学

习不同标签的差异。

然而基于英文的命名实体识别研究，中文的命名实体识别研究，尤其是稀缺领域的命名实体识别研究，相比于英文存在规范语料更少、标注不够规范、分词结果导致命名实体识别效果下降等诸多问题，因此如何通过类似的相关方法探究领域间差异，将是进行相关的命名实体识别研究的基础和重点。

4.1.1 实验设置

在随后的小节中，将通过几组训练数据和测试数据配置不同的基于 CRF 的命名实体识别实验，来进行新闻领域和医疗领域语料的差异程度和影响的探究实验。标准 CRF 的相关内容和工具设置在之前的 2.2 节中已经进行了介绍，除此之外还采用了带有部分标注功能的 CRF 模型[54]，部分标注信息即指通过词典对已确定标为实体的词汇进行数据的标记，未确定标记的数据标记为空。带有部分标注的 CRF 模型能够对只带有部分标注信息的训练数据进行 CRF 模型的训练和测试。

实验用语料采用了新闻领域的 2 月份人民日报语料和医疗领域的电子病历语料，在探究新闻领域和医疗领域差异的对比实验中，随机抽取了 10%的人民日报语料作为固定的测试集，并从 2 月份人民日报语料余下的部分抽取了整体的 10%和 30%作为不同实验的固定训练集；在探究领域差异的部分标注 CRF 模型的实验中，随机抽取了 70%的医疗语料与固定训练集组合成不同的训练集组合，余下的 30%语料与固定的 10%人民日报语料测试集作为不同的测试集。语料均采用了 BIO 标注规则。对于标准 CRF 模型的评价，仍然采用了传统命名实体识别中常用的精确率（Precision）、召回率（Recall）和 F 值（F-score）这三个指标已在 2.4.1.3 节中进行介绍。

4.1.2 探究新闻领域和医疗领域差异的对比实验

4.1.2.1 探究领域差异的标准 CRF 对比实验

首先利用标准的 CRF 模型，利用不同配比的人民日报语料进行训练，并通过不同的测试语料时模型的 F 值，来分析新闻领域和医疗领域的领域差异。由于语料规模的原因，本小节中的测试语料都只随机抽取了电子病历语料的 30%或 10%的人民日报语料，是因为两种抽取方式抽取的语料规模基本相同。具体的实验语料设置和结果见表 4-1 所示：

表 4-1 领域差异的 CRF 实验结果对照表

序号	训练数据	测试数据	Precision	Recall	F1 值
1	随机 10%人民日报	随机 10%人民日报	85.49	64.13	73.29
2	随机 10%人民日报	30%的电子病历语料	0	0	0
3	随机 30%人民日报	随机 10%人民日报	87.83	76.15	81.57
4	随机 30%人民日报	30%的电子病历语料	13.79	0.03	0.07

根据实验结果对照表，可以发现实验 1 和实验 2 进行对比时，对于相同的训练模型，利用医疗语料测试的结果为 0，从一定程度上说明了新闻领域与医疗领域有着较大的差异。实验 1 和实验 3 对比的结果则符合机器学习的一般规律和实验的心理预期：在训练语料不够充足时，加入相同知识域的语料能够有效提升模型的效果。然而还发现了实验 2 和实验 4 对比时，增加了 2 成人民日报语料后，在测试语料上的效果有了略微的提升，可以认为增加了训练的新闻领域的规模，促进了模型学习到了泛化的语言学知识，进而影响了在测试语料上的效果。

4.1.2.2 探究领域差异的部分标注 CRF 对比实验

在本小节中将标准 CRF 模型换成了带有部分标注功能的 CRF 模型，并通过混合不同比例的医疗领域和新闻领域的语料进行训练，对不同领域的测试语料进行测试来比对模型的效果并进行分析。

由于采用了非标准的带有部分标注的 CRF 模型，并配合通过电子病历语料提取的医疗词典进行相关实验，因为训练语料中含有部分语料不含有标注信息，无法利用传统命名实体识别的评价体系进行相关模型的评价，因此只引入 Accucracy 和 Precision 的评价指标，记 *correct* 为标对的字词个数，*total* 为总的字词个数，*correct entity* 为标注对的总的实体个数，*total entity* 为标注出的总的实体个数。具体的公式如下：

$$Accuracy = \frac{correct}{total} \tag{4-1}$$

$$\Pr ecision = \frac{correct\ entity}{total\ entity} \tag{4-2}$$

Accuracy 和 Precision 计算的方式和传统的命名实体识别评价指标相同，

其中 Accuracy 代单字的标注是否正确，Precision 代表是否标注正确词典中的实体。实验结果如表 4-2 所示：

表 4-2 部分标注的实验结果对照表

序号	训练数据	测试数据	Accuracy	Precision
1	70%电子病历语料+随机 10%人民日报	30%的电子病历语料	97.99	93.34
2	70%电子病历语料+随机 10%人民日报	随机 10%人民日报	97.40	91.39
3	70%电子病历语料+随机 30%人民日报	30%的电子病历语料	97.63	92.17
4	70%电子病历语料+随机 30%人民日报	随机 10%人民日报	97.62	92.14

通过实验 1 和实验 3 可以看出，增加了训练语料中新闻领域的语料的比例后，测试医疗领域的效果反而下降，由于部分标注的条件随机场方法通过医疗领域的词典进行部分标注，并在解码的部分利用领域词典对测试语句的解空间进行了限制，因此可以认为，医疗领域语料中出现的实体和新闻领域语料中出现的实体差异较大，导致增大了新闻领域的语料的比例后，测试医疗领域的效果出现了下降。而对比实验 3,4 可以看出，实验 3 和实验 4 的实验结果基本相同且训练集相同，但是训练集的语料的配比，医疗领域的语料约占人民日报语料的 3 分之 2，由此可见，医疗领域的语料领域特征明显且更容易被模型学习得到，如果能够抽取医疗领域的语言学特征，配合大规模新闻领域语料集上的语言学特征进行模型的研究，对医疗领域的命名实体识别的效果可能会具有更强的泛化能力。

4.2 基于 GBDT 模型的弱监督的命名实体识别研究

在本节中，首先介绍了梯度优化决策树（GBDT：gradient boosted decision tree）模型，GBDT 模型实质上是机器学习模型中的一种集成模型，属于集成学习（Boosting）的一种。随后我们利用 GBDT 模型进行对 CRF 转移概率组成的特征向量和 word2vec 生成的无监督语义向量的集成模型的学习，并对模型的结果进行了评价。

4.2.1 梯度优化决策树（GBDT）模型

GBDT（gradient boosted decision tree）是一种利用决策树进行集成学习的模型。GBDT 模型在进行分类时，利用对数似然函数对损失函数在负梯度方向进行拟合，并利用叶子结点进行梯度残差的拟合搜索，因此能够在模型上达到更好的训练效果。总体来说，GBDT 模型就是一个不断拟合残差并叠加拟合函数上的模型，在拟合的过程中，残差不断变小，损失函数也不断接近最小值。到相比于其他机器学习模型，GBDT 模型对数据格式和类型适应能力强，模型不经过太多调参就能够达到很高的准确率，具有较强的健壮性。

4.2.2 基于 GBDT 模型的弱监督的命名实体识别研究

本小姐将在探究新闻领域和医疗领域的领域差异基础上，探究能否利用转移矩阵描述领域之间差异，将 CRF 模型的标签特征向量进行领域差异的转化后，和无监督语义向量进行向量组合，并利用 GBDT 集成模型进行类别标签的训练。

4.2.2.1 实验设置

在本节及之后的实验中，训练语料采用了人民日报 2 月份语料及随机 70%电子病历语料，测试语料为剩余 30%的电子病历语料。实体标记均为 BIO 标注，对于基础单元为词时，利用 CRF 模型和相同的测试语料进行了分词模型的训练。基线系统选择了基于字的 CRF 模型。GBDT 模型使用了 xgboost 工具包[54]进行了 GBDT 模型的训练。

具体的实验流程如下：

（1）在全部的电子病历语料和人民日报语料上利用 word2vec 训练得到 100 维的无监督词向量。

（2）采用以下两种方式得到 2 个 100 维的方阵用以描述领域差异：

1.利用 word2vec 学习人民日报语料和医疗领域训练语料的实体标记差异，得到表示实体标记的 2 个 3*100 维的矩阵 *News*，*Med*，并求得 100*100 维度的转移矩阵 *Trans_tag*，使得 *News*Trans_tag=Med*，记该方法为 Tag。

2.选取电子病历语料和人民日报语料共现次数最多的前 100 词的在不同语料中的无监督词向量组成两个 100 维方阵，记人民日报语料得到的矩阵为 *News_mat*，电子病历语料得到的矩阵为 *Med_mat*，并求得 100*100 维度的转移矩

阵 *Trans_vec*，使得 *News_mat*Trans_vec=Med_mat*，记该方法为 Vec。

（3）利用人民日报语料进行 CRF 的 NER 模型训练，并进行测试语料的测试，对于每个单元（字或词），得到表示三种实体标记 BIO 的解码概率的 1*3 的向量 *Pro_vec*，将 *Pro_vec* 左乘 *News*，得到一个表示不同概率标记加权分布的维度为 1*100 维向量 *Weighted_vec*。

（4）将每个单元（字或词）得到的 *Weighted_vec* 左乘领域差异方阵，得到 1*100 维的特征向量，组合该单元（字或词）的特征向量和窗口大小为 5 的上下文无监督语义向量及该单元（字或词）的词袋表示，利用 GBDT 模型进行训练。

实验的过程示意图如图 4-1 所示：

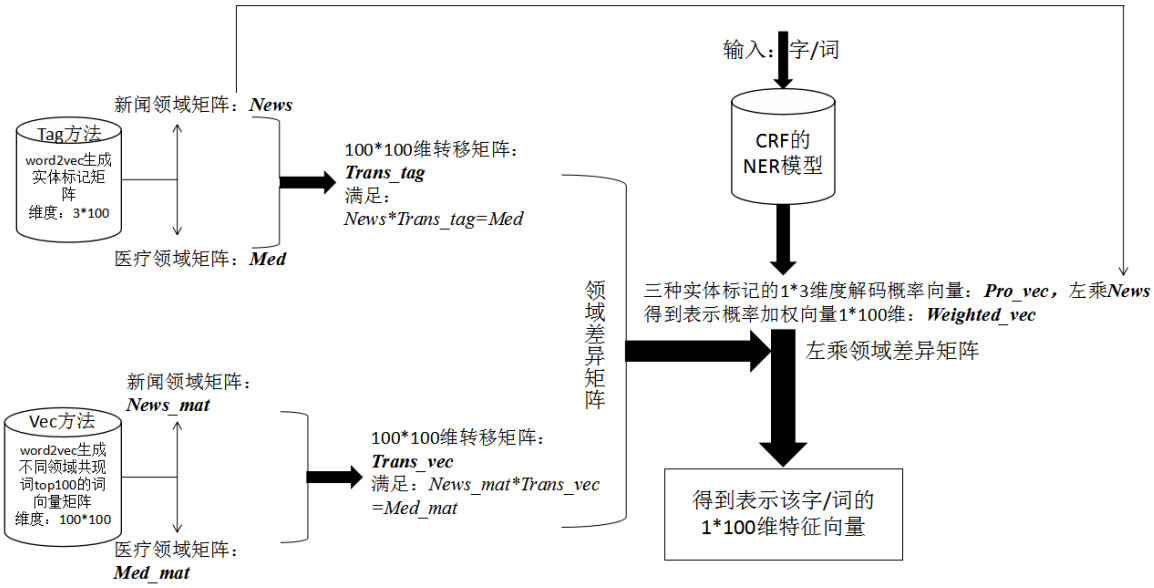

图 4-1 利用转移矩阵生成特征向量的示意图

4.2.2.2 实验结果及结论

在上述的实验流程中，进行了多组不同基础单元设置和衡量领域差异矩阵不同的向量组合的设置，实验结果如表 4-3 所示：

通过实验对比看出，基础单元选择为字时，模型的整体效果并不好，这是因为 GBDT 模型无法像传统的 CRF 模型，学习到字与字之间标记的连续关系，导致效果过差，而将基础单元选择为词后，不需要学习这种连续关系，模型的效果也得到了大幅度的提升。与不进行领域差异表示的实验相比，加入领域差异的实验都有了效果明显提升，这证明了加入领域差异向量来进行领域自

适应学习的方法的有效性。然而基于词的两种领域差异矩阵生成的向量表示的效果均相同，即使同时采用了两种方法增加了向量的维度，模型的效果也没有提升，这说明该方法在测试集上已达到目前的最好效果且没有产生太大的过拟合情况。相比于基线系统的 CRF 模型，我们的模型在融入无监督词向量后效果有了一定程度的提高，也证明了方法的有效性。

表 4-3 集成模型的实验结果对照表

序号	基础单元	领域差异矩阵	是否融入无监督词向量	Precision	Recall	F1 值
1	字	Tag	否	7.1	50.1	12.4
2	字	Tag	是	30.3	97.3	46.3
3	词	无	是	67.9	99.7	80.8
4	词	Tag	是	82.1	96.6	88.8
5	词	Vec	是	82.1	96.6	88.8
6	词	Tag+Vec	是	82.1	96.6	88.8
7	基线系统:CRF			87.8	84.2	85.9

4.3 本章小结

在前两章中主要介绍了基于 LSTM 的命名实体识别研究，虽然达到了不错的效果，但是深度学习模型受到计算能力等条件的限制，在实际应用中往往达不到理想的效果。因此在本章中，我们主要利用了机器学习中的 GBDT 模型进行了无监督语义和领域差异的集成学习，希望通过找寻和衡量领域差异的方式进行跨领域的命名实体识别研究。

首先利用多组不同领域语料混合作为训练数据的命名实体识别实验进行不同知识域领域差异的探索，希望通过这些对比实验来发掘新闻领域和医疗领域语料进行命名实体识别模型建模时是如何互相影响的。

通过这些对比实验，我们进而考虑使用 GBDT 模型集成新闻领域和医疗领域的领域差异，以及医疗领域的无监督语义向量，进行医疗领域的跨领域命名实体识别。GBDT 模型作为传统机器学习模型中目前分类和泛化效果最为突出的 GBDT 模型，实质上一种集成模型。随后借助 GBDT 模型进行了基于集成学习的命名实体识别研究，通过不同方法求得定量描述领域差异的矩阵，并证明加入描述领域差异的向量后，整体模型的效果得到了一定程度的提高。

结论

如何高效率高速地从无规则且不规范的文本数据中，获取和组织成结构化的格式数据，是自然语言处理领域最为关心的技术问题之一。命名实体识别作为词法分析中的基础任务，能够有目的性地对文本进行结构化和半结构化处理，是目前最为火热的研究热点之一。新闻领域的命名实体识别在研究领域的相关研究方向和方法已日渐成熟，然而在语料稀缺的医疗领域，没有足够规模的规范语料支撑模型的训练，导致命名实体识别的效果并不理想。

本文对面向医疗领域的命名实体识别进行了系统的研究，针对目前医疗领域命名实体识别效果差的问题，利用多领域的标注语料，从传统机器学习方法和深度学习方法两个角度，进行命名实体识别的模型构建，并从研究和实际应用两个层面对效果进行了评估，主要的突出性成果有以下几点：

（1）利用了深度学习中的 LSTM 模型，进行了单一领域命名实体识别模型的构建。借助于深度学习模型挖掘深层特征和表示能力强的优势进行文本序列的语义表示，学习文本序列中每个时间点的上下文窗口信息，完整地对整体序列进行模型建模。并利用 Theano 的深度学习框架进行了模型的实现和调优，利用如 Dropout 等多种深度学习中的技巧进行了调参，在医疗领域的语料上进行了测试，并进行了实际应用实验的对比，相比于传统的基于 CRF 的命名实体识别模型均有较大的提升。

（2）在单一领域的命名实体识别模型对传统方法进行了相对的提升后，我们通过迁移学习的思想，利用参数融合的方法，借助大规模的新闻领域的语料，希望学习到泛化的语言学知识用以改善医疗领域模型的效果，并利用深度学习中预训练的技巧，将无监督的医疗领域语义向量加载到命名实体识别模型中。借助以上两个方法使模型能够融合充分的外部信息，模型效果得到了进一步的提高。

（3）为了弥补深度学习模型在实际应用中的不足，通过多组命名实体识别实验进行了领域差异的探究，并借助传统机器学习中的 GBDT 模型建立了面向医疗领域的命名实体模型，将传统的 CRF 模型中的解码向量进行领域差异的转换，并结合无监督语义向量进行集成学习，取得了较好的研究效果。

本文从单一领域和跨领域两个角度进行了医疗领域的命名实体识别研究，取得了一定的阶段性成果，然而还有如下几点有待进一步深化和提高：

（1）利用 LSTM 模型进行大规模不规范的语料测试时的耗时较长，实际

应用中可能会出现无法及时给出反馈的情况，需要提高模型的训练和测试速度。

（2）在实际语料测试的实验中，仅利用从词典中获取的词的个数来进行结果分析，应进行语料的标注获取更准确的实验结果；同时对模型识别得到的词，应通过外部信息等建立相应规则进行二次过滤以提高词表的质量。

（3）通过预训练学习得到领域内的无监督语义向量的质量与训练语料的质量和风格有关，即使训练语料与测试语料的领域一致，不同的语言风格仍然会导致模型效果的下降。

（4）通过矩阵进行领域差异刻画的方式过于简单，可以在远期工作中利用深度学习模型加入一层转移层或自编码器模型进行领域差异的自动学习。

参考文献

[1] Doddington, George R., et al. "The Automatic Content Extraction (ACE) Program-Tasks, Data, and Evaluation." LREC. Vol. 2. 2004.

[2] 邱莎, 段玻, 申浩如, 等. 基于条件随机场的中文人名识别研究[J]. 昆明学院学报, 2012, 33(6): 64-66.

[3] Rau L F. Extracting company names from text[C]//Artificial Intelligence Applications, 1991. Proceedings., Seventh IEEE Conference on. IEEE, 1991, 1: 29-32.

[4] Grishman R, Sundheim B. Message Understanding Conference-6: A Brief History[C]//COLING. 1996, 96: 466-471.

[5] Klinger, Roman, and Katrin Tomanek. Classical probabilistic models and conditional random fields. TU, Algorithm Engineering, 2007.

[6] Bikel D M, Miller S, Schwartz R, et al. Nymble: a high-performance learning name-finder[C]//Proceedings of the fifth conference on Applied natural language processing. Association for Computational Linguistics, 1997: 194-201.

[7] 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. Association for Computational Linguistics, 2003: 188-191.

[8] Asahara M, Matsumoto Y. Japanese named entity extraction with redundant morphological analysis[C]//Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. Association for Computational Linguistics, 2003: 8-15.

[9] Lafferty J, McCallum A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[C]//Proceedings of the eighteenth international conference on machine learning, ICML. 2001, 1: 282-289.

[10] Andrew Mccallum and Wei Li. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In HLT-NAACL, pages 188–191.

[11] Finkel, Jenny Rose, Trond Grenager, and Christopher Manning. "Incorporating

non-local information into information extraction systems by gibbs sampling." Proceedings of the 43rd annual meeting on association for computational linguistics. Association for Computational Linguistics, 2005:363–370.

[12] Krishnan, Vijay, and Christopher D. Manning. "An effective two-stage model for exploiting non-local dependencies in named entity recognition." Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2006:1121–1128.

[13] 张祝玉, 任飞亮, 朱靖波. "基于条件随机场的中文命名实体识别特征比较研究 [C]." 见: 第 4 届全国信息检索与内容安全学术会议论文集. 2008.

[14] Rosenblatt F. The perceptron: A probabilistic model for information storage and organization in the brain[J]. Psychological review, 1958, 65(6): 386.

[15] Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors[J]. Cognitive modeling, 1988, 5(3): 1.

[16] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(5786): 504-507.

[17] Bengio Y, LeCun Y. Scaling learning algorithms towards AI[J]. Large-scale kernel machines, 2007, 34(5): 1-41.

[18] Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets[J]. Neural computation, 2006, 18(7): 1527-1554.

[19] Dahl G E, Yu D, Deng L, et al. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition[J]. Audio, Speech, and Language Processing, IEEE Transactions on, 2012, 20(1): 30-42.

[20] Bengio Y, Bastien F, Bergeron A, et al. Deep learners benefit more from out-of-distribution examples[C]//International Conference on Artificial Intelligence and Statistics. 2011: 164-172.

[21] Sun Y, Wang X, Tang X. Deeply learned face representations are sparse, selective, and robust[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 2892-2900.

[22] Lee H, Pham P, Largman Y, et al. Unsupervised feature learning for audio classification using convolutional deep belief networks[C]//Advances in neural information processing systems. 2009: 1096-1104.

[23] Bengio Y, Schwenk H, Senécal J S, et al. Neural probabilistic language models[M]//Innovations in Machine Learning. Springer Berlin Heidelberg, 2006:

137-186.

[24] Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch[J]. The Journal of Machine Learning Research, 2011, 12: 2493-2537.

[25] Eric H. Huang, Richard Socher, etc. ImprovingWord Representations via Global Context and MultipleWord Prototypes. ACL(2012). 2012.

[26] Morin F, Bengio Y. Hierarchical Probabilistic Neural Network Language Model[C]//Aistats. 2005, 5: 246-252.

[27] Mikolov T, Karafiát M, Burget L, et al. Recurrent neural network based language model[C]//INTERSPEECH. 2010, 2: 3.

[28] Huang P S, He X, Gao J, et al. Learning deep structured semantic models for web search using clickthrough data[C]//Proceedings of the 22nd ACM international conference on Conference on information & knowledge management. ACM, 2013: 2333-2338.

[29] Le Q V, Mikolov T. Distributed representations of sentences and documents[J]. arXiv preprint arXiv:1405.4053, 2014.

[30] Jordan M I. Serial order: a parallel distributed processing approach. Technical report, June 1985-March 1986[R]. California Univ., San Diego, La Jolla (USA). Inst. for Cognitive Science, 1986.

[31] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.

[32] Gers F A, Schmidhuber J, Cummins F. Learning to forget: Continual prediction with LSTM[J]. Neural computation, 2000, 12(10): 2451-2471.

[33] Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures[J]. Neural Networks, 2005, 18(5): 602-610.

[34] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv:1409.0473, 2014.

[35] Dyer C, Ballesteros M, Ling W, et al. Transition-based dependency parsing with stack long short-term memory[J]. arXiv preprint arXiv:1505.08075, 2015.

[36] Xu Y, Mou L, Li G, et al. Classifying relations via long short term memory networks along shortest dependency paths[C]//Proceedings of Conference on Empirical Methods in Natural Language Processing (to appear). 2015.

[37] Lample G, Ballesteros M, Subramanian S, et al. Neural architectures for named

entity recognition[J]. arXiv preprint arXiv:1603.01360, 2016.

[38]李刚. 生物医疗文献中的蛋白质名识别[D]. 大连理工大学, 2006.

[39]张金龙, 王石, 钱存发. 基于CRF和规则的中文医疗机构名称识别[J]. 计算机应用与软件, 2014(3):159-162.

[40]Uzuner Ö, South B R, Shen S, et al. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text[J]. Journal of the American Medical Informatics Association, 2011, 18(5): 552-556.

[41]Jiang M, Chen Y, Liu M, Rosenbloom S T, Mani S, Denny JC, Xu H. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. Journal of the American Medical Informatics Association, 2011, 18(5): 601¡606

[42]Jonnalagadda S, Cohen T, Wu S, et al. Enhancing clinical concept extraction with distributional semantics[J]. Journal of biomedical informatics, 2012, 45(1): 129-140.

[43]王鹏远, 姬东鸿. 基于多标签CRF的疾病名称抽取[J]. 计算机应用研究, 2017, 34(1):118-122.

[44]de Bruijn B, Cherry C, Kiritchenko S, et al. Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010[J]. Journal of the American Medical Informatics Association, 2011, 18(5): 557-562.

[45]苏娅, 刘杰, 黄亚楼. 在线医疗文本中的实体识别研究[J]. 北京大学学报(自然科学版), 2016, 52(1): 1-9.

[46]栗伟, 赵大哲, 李博, 等. CRF 与规则相结合的医疗病历实体识别[J]. 计算机应用研究, 2015, 32(4): 1082-1086.

[47]曲春燕, 关毅, 杨锦锋, 等. 中文电子病历命名实体标注语料库构建[J]. 高技术通讯, 2015 (2): 143-150.

[48]Turian J, Ratinov L, Bengio Y. Word representations: a simple and general method for semi-supervised learning[C]//Proceedings of the 48th annual meeting of the association for computational linguistics. Association for Computational Linguistics, 2010: 384-394.

[49]Hinton G E, Srivastava N, Krizhevsky A, et al. Improving neural networks by preventing co-adaptation of feature detectors[J]. arXiv preprint arXiv:1207.0580, 2012.

[50]Lafferty J, McCallum A, Pereira F C N. Conditional random fields: Probabilistic

models for segmenting and labeling sequence data[J]. 2001.

[51]Kulkarni, Vivek, Yashar Mehdad, and Troy Chevalier. "Domain Adaptation for Named EntityRecognition in Online Media with Word Embeddings." arXiv preprint arXiv:1612.00148(2016).

[52]Qu, Lizhen, et al. "Named Entity Recognition for Novel Types by Transfer Learning." arXiv preprint arXiv:1610.09914 (2016).

[53]段超群. 面向缺乏标注数据领域的命名实体识别的研究. 2015. Master's Thesis. 哈尔滨工业大学.

[54]Chen, Tianqi, and Carlos Guestrin. "XGBoost: Reliable large-scale tree boosting system." Proceedings of the 22nd SIGKDD Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA. 2016.

攻读硕士学位期间发表的学术论文及其他成果

哈尔滨工业大学学位论文原创性声明和使用权限

学位论文原创性声明

本人郑重声明：此处所提交的学位论文《面向医疗领域的中文命名实体识别》，是本人在导师指导下，在哈尔滨工业大学攻读学位期间独立进行研究工作所取得的成果，且学位论文中除已标注引用文献的部分外不包含他人完成或已发表的研究成果。对本学位论文的研究工作做出重要贡献的个人和集体，均已在文中以明确方式注明。

作者签名：薛秭竹　　日期：　　年　　月　　日

学位论文使用权限

学位论文是研究生在哈尔滨工业大学攻读学位期间完成的成果，知识产权归属哈尔滨工业大学。学位论文的使用权限如下：

（1）学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文，并向国家图书馆报送学位论文；（2）学校可以将学位论文部分或全部内容编入有关数据库进行检索和提供相应阅览服务；（3）研究生毕业后发表与此学位论文研究成果相关的学术论文和其他成果时，应征得导师同意，且第一署名单位为哈尔滨工业大学。

保密论文在保密期内遵守有关保密规定，解密后适用于此使用权限规定。

本人知悉学位论文的使用权限，并将遵守有关规定。

作者签名：薛秭竹　　日期：　　年　　月　　日

导师签名：朱聪慧　　日期：　　年　　月　　日

致谢

时间如白驹过隙，两年的硕士学习生活转瞬即逝，而我也将离开培育我多年的哈尔滨工业大学。硕士这两年的学习和研究生活，使我无论从研究能力、工程能力亦或是与人讨论交流的能力都有着极大的增长和提高。这些提高除去个人的努力之外，最重要的是有很多老师和同学们的帮助和支持。在毕业论文完成之际，我要向这些帮助过我的人表示我由衷的感谢。

首先我要特别感谢我的导师朱聪慧老师。朱老师为人宽厚，幽默风趣，治学严谨。我在学习生活中遇到的困难，朱老师都能够给出极具建设性的意见和前进方向，并在小的细节上严格要求，使我充分发现自己的不足并加以学习和改进。在日常的生活中，朱老师和同学们相处非常融洽，而且在我生活上遇到困难的时候，朱老师都能给予我及时的帮助，让我非常感动。

感谢实验室的赵铁军教授，正是有着赵老师不辞辛苦的努力和付出，实验室才能够越来越好，取得更多的成就。赵老师在学术上的严谨的态度，也是我努力学习的榜样。赵老师为人随和，不管自己多么忙碌，每次遇到我，都会亲切的问我学习以及生活中的进展以及遇到的困难。希望机器智能与翻译实验室在赵老师的带领下，能够发展地越来越好。

感谢李生老师、郑德权老师、杨沐昀老师、曹海龙老师和徐冰老师。感谢你们对实验室同学的关心帮助和无私付出，给大家营造了严格又不失和谐的学习氛围。

感谢实验室的段超群师兄，从我进入实验室开始给予我无私的帮助。在和段超群师兄的交谈和讨论中，我的很多学术问题一一得到解决。而段超群师兄在学术和科研上扎实努力的作风，深深的影响着我，也祝愿段超群师兄的博士生活一切顺利。

感谢李剑风师兄、陈科海师兄、史桦兴师兄、王宝鑫师兄和王晓雪师姐在我的学习和工作上遇到问题的时候，对我耐心的指导和帮助。

感谢同一届的刘笛、杨艳、王亚楠、李依尘、俞可、李冠林、吴芳颖和张红阳。正是你们在学习生活中的陪伴和帮助，使得我的研究生生活丰富多彩。希望大家都能够在以后的生活中一帆风顺。

感谢张元博、孙潇、张冠华、赵晶晶、赵玉坤等师弟师妹，你们让实验室的生活充满欢声笑语。

感谢 204 寝室的室友，王善策、朱海潮、曹雨，在这两年我们互相陪伴和

帮助，为了各自的未来共同努力和拼搏，希望我们都能在未来的日子里一帆风顺。

感谢一直默默为我付出，支持和鼓励我的双亲。正是你们的无私奉献、照料和支持，才让我有了今天的成就。希望你们在以后的日子里身体健康，工作顺利！

最后，感谢所有陪伴过我，帮助过我的朋友们。希望你们一切顺利！