

2020 届硕士专业学位论文

分类号: _____

学校代码: 10269

密 级: _____

学 号: 71174500171



華東師範大學

East China Normal University

硕士专业学位论文

MASTER' S DISSERTATION

论文题目：融入辅助信息的中文医学命名实体识别技术研究

院 系: 软件工程学院

专业学位类别: 工程硕士

专业学位领域: 软件工程

论文指导教师: 王晓玲教授

论 文 作 者: 许明司

2020 年 10 月 12 日

Dissertation for master degree in 2020

Student ID:

University code:10269

East China Normal University

**Title: Research on Chinese Medical Named Entity
Recognition Technology Incorporating Auxiliary
Information**

Department:	<u>School of Software Engineering</u>
Type:	<u>Master of Engineering</u>
Domain:	<u>Software Engineering</u>
Supervisor:	<u>Prof.Wang Xiaoling</u>
Candidate:	<u>Xu Mingsi</u>

Oct , 2020

华东师范大学学位论文原创性声明

郑重声明：本人呈交的学位论文《融入辅助信息的中文医学命名实体识别技术研究》，是在华东师范大学攻读硕士/博士（请勾选）学位期间，在导师的指导下进行的研究工作及取得的研究成果。除文中已经注明引用的内容外，本论文不包含其他个人已经发表或撰写过的研究成果。对本文的研究做出重要贡献的个人和集体，均已在文中作了明确说明并表示谢意。

作者签名： 许明

日期：2020年11月15日

华东师范大学学位论文著作权使用声明

《融入辅助信息的中文医学命名实体识别技术研究》系本人在华东师范大学攻读学位期间在导师指导下完成的硕士/博士（请勾选）学位论文，本论文的研究成果归华东师范大学所有。本人同意华东师范大学根据相关规定保留和使用此学位论文，并向主管部门和相关机构如国家图书馆、中信所和“知网”送交学位论文的印刷版和电子版；允许学位论文进入华东师范大学图书馆及数据库被查阅、借阅；同意学校将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于（请勾选）

（ ） 1. 经华东师范大学相关部门审查核定的“内部”或“涉密”学位论文*，
于 年 月 日解密，解密后适用上述授权。

（☒） 2. 不保密，适用上述授权。

导师签名 王曉玲

本人签名 许明

2020年11月15日

* “涉密”学位论文应是已经华东师范大学学位评定委员会办公室或保密委员会审定过的学位论文（需附获批的《华东师范大学研究生申请学位论文“涉密”审批表》方为有效），未经上述部门审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权）。

许明司 硕士学位论文答辩委员会成员名单

姓名	职称	单位	备注
曹奇英	教授	东华大学	主席
朱明华	教授	华东师范大学	
张新宇	副教授	华东师范大学	

摘要

医学命名实体识别是医学文本挖掘的一项基础性任务，是其诸多自然语言处理下游应用的前提。目前中文医学文本存在大量医学专用术语，其标注成本通常较高，如何在有限的标注信息下更好地利用辅助信息提高中文医学实体识别的精度，目前是该领域的一大挑战。

本文利用了医学电子病历数据集和医学线上问答数据集进行探索。首先使用开放域实体识别的两个经典模型 **Linear-Chain-CRF** 和 **BiLSTM-CRF**，分别用添加特征模板和超参数调优的方式对于两类模型进行优化。实验发现，在两类数据集上均是基于字粒度的 **BiLSTM-CRF** 模型的 **Micro-F1** 结果最好。与此同时，**医学词典特征与医学知识特征**以特征模板形式加入对于 **Linear-Chain-CRF** 模型也能够有较大提升，甚至对于某些医学实体类的识别精度超越了深度学习模型的效果。由此，本文开始探索如何深度学习模型中融入医学词典信息与领域知识信息，以进一步提高中文医学命名实体识别的效果。

医学词典信息融入：本文从大众健康网站，医学权威期刊等不同风格文本中爬取得到医学词典与医学文本语料作为辅助信息，构建**上下文词典特征**和**词界特征**两种医学词典特征融入字向量，以词典分词的形式将信息融入词向量。初步实验后发现更新后的字向量与词向量在两个数据集中识别精度均有提高。为了进一步探索字与词的信息融合方式，根据词是作为字的信息补充或者词与字的地位对等两种先验假设的不同设计了**字词串联法模型**和**字词并联法模型**。在此基础上进一步细分为**字词直接串联法**、**字词间接串联法**与**字词后置并联法**、**字词前置并联法**四种模型，验证了不同字词融合方式对于医学词典信息融入效果的差异。

知识图谱领域信息融入：本文使用 **BERT** 预训练模型在两类医学数据集上进行微调，并验证了添加医学词典特征对于 **BERT-CRF** 模型效果的提高。以知识图谱为载体进行领域知识融入的尝试，本文借鉴了北京大学与腾讯公司于 **AAAI-2020** 会议上提出的 **K-BERT** 模型中**软位置编码**与**可见矩阵**的思路，在此基础上设计了适用于中文医学实体识别数据的 **CMK-BERT** 模型。该模型在融入医学词典

特征基础上的 BERT 模型中分别融入 CN-DBpedia、知网(HowNet)和医学知识图谱(MedicalKG)三种不同类型的知识图谱,利用知识图谱领域信息使得医学实体识别效果取得进一步的提高。

关键词: 中文医学命名实体识别, 医学词典特征, 字词融合方式, CMK-BERT

ABSTRACT

Medical Named Entity Recognition is a basic task of medical text mining and a prerequisite for many downstream applications of natural language processing. Medical named entity recognition is a basic task of medical text mining and other many downstream applications of natural language processing. At present, there are a large number of medical-specific terms in Chinese medical texts, and the cost of labeling is very high. How to make better use of auxiliary information to improve the accuracy of Chinese medical entity recognition with limited annotation information is a major challenge in this field.

This article uses the medical electronic medical record data and the medical online question and answer data to explore. Firstly, two classic models called Linear-Chain-CRF and BiLSTM-CRF of open domain entity recognition are used to optimize the two types of models by adding feature templates and tuning hyperparameter respectively. It is found that the Micro-F1 results of the BiLSTM-CRF model based on character granularity are the best on both types of data sets. The addition of **medical dictionary features** and **medical knowledge features** in the form of feature templates can also greatly improve the Linear-Chain-CRF model, and even the recognition accuracy of certain medical entities exceeds the effect of deep learning models.

Information integration based on medical dictionary: This article crawls from public health websites, medical authoritative journals and other different texts to get medical dictionaries and medical text corpus as auxiliary information. It constructs **context dictionary features** and **word boundary features**, and integrates two medical dictionary features into word vectors. Based on it, there are four models: direct series connection method, indirect series connection method, word post-parallel parallel method, and word pre-parallel method, verifying the effect of different character and word fusion methods on the integration of medical dictionary information.

Domain information integration based on Knowledge Graph: This article uses the BERT pre-training model to fine-tune on two types of medical data sets, and verifies that adding medical dictionary features improves the effect of the BERT-CRF model. In an attempt to integrate domain knowledge with the knowledge graph as a carrier, this article draws on the idea of **soft-position embedding** and **visible matrix** in the K-BERT model, and design CMK-BERT model which is suitable for Chinese medical entity recognition data. Our model integrates three different types of knowledge graphs including CN-DBpedia, HowNet and MedicalKG into the BERT model based on the features of medical dictionaries. The knowledge graph domain information is used to achieve better results in medical entity recognition .

Keywords: *Chinese medical named entity recognition, medical dictionary features, character and word method, CMK-BERT*

目录

第一章 绪论	1
1.1 研究背景与意义	1
1.2 本文主要工作与创新点	3
1.3 本文组织结构与行文思路	5
1.4 国内外研究现状	6
1.4.1 基于词典和启发式规则的方法	6
1.4.2 基于统计学习的方法	7
1.4.3 基于深度学习的方法	8
1.4.4 引入外部信息辅助的方法	10
第二章 背景信息介绍	11
2.1 统计学习模型 Linear-Chain-CRF	11
2.2 深度学习模型 BiLSTM	12
2.3 基于 BERT 模型的预训练	12
2.4 知识图谱信息的利用	14
第三章 基于开放域实体识别模型实验	16
3.1 命名实体识别实验流程与评价指标	16
3.1.1 问题描述	16
3.1.2 两种解决流程	17
3.1.3 严格指标与松弛指标	18
3.1.4 多类实体的微平均指标	20
3.2 中文医学实体识别数据集	20

3.3 基于 Linear-Chain-CRF 的实验设计与分析	22
3.3.1 数据预处理	22
3.3.2 医学特征模板设计	24
3.3.3 实验结果分析	26
3.4 基于 BiLSTM-CRF 的实验设计与分析	28
3.4.1 模型结构	28
3.4.2 模型超参数选择	30
3.4.3 结果分析	35
3.5 本章小结	37
第四章 融入医学词典与语料信息	38
4.1 词典信息的获取与使用	38
4.1.1 医学词典与语料的获取	38
4.1.2 字粒度中融入词典信息	39
4.1.3 词粒度中融入词典信息	43
4.1.4 字向量与词向量的训练	45
4.2 字词融合的深度学习模型	47
4.2.1 单粒度增强实验	47
4.2.2 字词串联法模型设计	50
4.2.3 两种字词串联法的实验结果及分析	55
4.2.4 字词并联法模型设计	55
4.2.5 两种字词并联法的实验结果及分析	60
4.3 本章小结	62
第五章 BERT 模型融入知识信息	63

5.1 基于 BERT-CRF 的实体识别模型	63
5.1.2 BERT-CRF 模型融入词典特征实验	65
5.2 BERT 模型融合知识图谱	67
5.2.1 软位置编码与可见矩阵	67
5.2.2 CMK-BERT 模型总体设计	69
5.2.3 基于 CMK-BERT 模型的数据转换	71
5.2.4 融入知识图谱的实验分析	73
5.3 本章小结	80
第六章 总结与展望	81
6.1 本文总结	81
6.2 未来展望	82
参考文献	83

插图

图 1.1 HMM、HEMM、CRF 示意图	8
图 2.1 Linear-Chain-CRF 模型结构图	11
图 2.2 BERT 模型示意图	13
图 3.1 中文语料实体标注图	16
图 3.2 基于统计机器学习方法的实验流程图	17
图 3.3 基于深度学习方法的实验流程图	18
图 3.4 CCKSNER 数据集源文件图	21
图 3.5 CMQANER 数据集源文件图	22
图 3.6 CCKSER 数据集标注结果图	22
图 3.7 CMQANER 数据集标注结果图	23
图 3.8 BiLSTM-CRF 结构图	29
图 3.9 CCKSNER 数据集下字/词向量维度对效果的影响图	31
图 3.10 CMQANER 数据集下字/词向量维度对效果的影响图	31
图 3.11 CCKSNER 数据集下优化器和学习速率对效果的影响图	32
图 3.12 CMQANER 数据集下优化器和学习速率对效果的影响图	32
图 3.13 LSTM 层数对效果的影响图	33
图 3.14 隐藏层维度对于模型的 F1 值影响图	33
图 3.15 Dropout rate 对于模型的 F1 值影响图	34
图 4.1 词典来源门类图	38
图 4.2 医学词典图	39
图 4.3 辅助医疗信息样例图	39
图 4.4 不使用医学词典的 jieba 分词图	44
图 4.5 使用医学词典的 jieba 分词图	44
图 4.6 CBOW 与 Skip-gram 模型图	45
图 4.7 Skip-gram 模型样例图	46
图 4.8 字粒度分割的医学语料图	46

图 4.9 词粒度分割的医学语料图	47
图 4.10 所训练的医学字向量图	47
图 4.11 所训练的医学词向量图	47
图 4.12 直接串联法模型图	51
图 4.13 间接串联法模型图	53
图 4.14 字词后置并联法模型图	56
图 4.15 字词前置并联法模型图	58
图 5.1 BERT 遮罩语言模型示意图	63
图 5.2 BERT 模型输入示意图	64
图 5.3 融入词典特征的 BERT+CRF 模型图	65
图 5.4 句子树软位置与硬位置编码图	68
图 5.5 带有软位置信息的 BERT 模型输入图	69
图 5.6 CMK-BERT 模型图	70
图 5.7 CMK-BERT 模型输入变换图	72
图 5.8 CMK-BERT 模型文本输入图	73
图 5.9 CN-DBpedia 图谱	73
图 5.10 HowNet 图谱	74
图 5.11 MedicalKG 图谱	74
图 5.12 CCKSNER 数据收敛速度图	79
图 5.13 CMQANER 数据收敛速度图	79

表格

表 3.1 CCKSNER 数据集数据实体分布表	21
表 3.2 CMQANER 数据集数据实体分布表	21
表 3.3 CCKSNER 数据集标注表	23
表 3.4 CMQANER 数据集标注表	24
表 3.5 不同特征模板对 CCKSNER 数据集实体识别效果的影响表	25
表 3.6 不同特征模板对 CMQANER 数据集实体识别效果的影响表	26
表 3.7 CRF 模型下个 CCKSNER 数据集各实体识别精度表	27
表 3.8 CRF 模型下个 CMQANER 数据集各实体识别精度表	27
表 3.9 CCKSNER 数据集在 BiLSTM-CRF 模型的参数选择表	34
表 3.10 Linear-Chain-CRF 与 BiLSTM-CRF 模型效果对比表	35
表 3.11 CCKSNER 数据集在 BiLSTM-CRF 模型下的各个实体识别效果对比 表	35
表 3.12 CMQANER 数据集在 BiLSTM-CRF 模型下的各个实体识别效果对比 表	36
表 4.1 N-gram 特征模板表	40
表 4.2 词典特征向量表	41
表 4.3 分词结果表	42
表 4.4 BiLSTM-CRF 模型下不同字向量输入对 CCKSNER 数据集识别效果比 较表	48
表 4.5 BiLSTM-CRF 模型不同字向量输入对 CMQANER 数据集识别效果比较 表	49
表 4.6 不同词向量输入对 CCKSNER 数据集实体识别效果的影响表	50
表 4.7 不同词向量输入对 CMQANER 数据集实体识别效果的影响表	50
表 4.8 直接串联法模型的参数选择表	52
表 4.9 间接串联法模型的参数选择表	54

表 4.10 字词串联法模型对于 CCKSNER 数据集的效果表	55
表 4.11 字词串联法模型对于 CMQANER 数据集的效果表	55
表 4.12 后置并联法模型的参数选择表	57
表 4.13 前置并联法模型的参数选择表	59
表 4.16 CCKSNER 数据集在字词前置并联法模型下各个实体上识别效果对比 表	61
表 4.17 CMQANER 数据集在字词前置并联法模型下各个实体上识别效果对比 表	61
表 5.1 BERT 不同版本参数表	64
表 5.2 BERT-CRF 模型对于 CCKSNER 数据集的效果表	66
表 5.3 BERT-CRF 模型对于 CMQANER 数据集的效果表	66
表 5.4 CMK-BERT 模型融入不同图谱在 CCKSNER 数据集的效果表	75
表 5.5 CMK-BERT 模型融入不同图谱在 CMQANER 数据集的效果	75
表 5.6 融入 CnDbpedia 图谱之后在 CCKSNER 数据集的效果表	76
表 5.7 融入 CnDbpedia 图谱之后在 CMQANER 数据集的效果表	77
表 5.8 融入 Medical 图谱之后在 CMQANER 数据集的效果表	78

第一章 绪论

1.1 研究背景与意义

随着近年来文本数据规模的爆炸式增长,以及大规模知识库的构建需求。命名实体识别(Named Entity Recognition, NER)成为了自然语言理解任务中的热点问题,其对特定领域的知识图谱构建起到了非常关键的作用。

命名实体指的是在一些特定文本中具有某些特定的意义或者指代性非常强的实体。而命名实体识别任务则是从相关的文本语料中解析并识别出具有特定含义的实体词语,例如人名、地名、机构名等并将其进行归类划分。命名实体识别的评测任务于1995年的MUC-6会议中首次提出,并且该会议最先把命名实体作为一个明确的概念[1]。自1999年起的自动文本抽取(简称ACE)项目将实体类别进行了更多的扩展,引入了人名,地名,机构名,地缘政治实体等。在CoNLL-2003会议上进行了人名、地名、组织机构名(ORG)和其他杂项实体的命名实体识别,该会议的目标是建立一个可以识别两种不同语言的实体识别系统。参赛人员主要采用的是马尔可夫模型(HMM)、最大熵模型(MEMM)、支持条件随机场(CRF)等统计机器学习模型,相比于之前无疑是巨大的技术进步[2]。这些测评会议快速推动了命名实体识别技术的发展。

在命名实体识别领域,评价一个特定名词是否为命名实体的正确与否通常包括两个方面:模型不仅要正确地识别实体边界,并且还要正确地识别出实体类型。即使实体类型标注类别正确,其边界可能仍然是错误的。当实体边界正确时,可能标注了错误的实体的类型。对于英文实体来说,其形态特征较为明显,例如人名等首字母往往是大写的。这使得模型识别命名实体相对较为容易。而汉语受到分词精度的影响,实体识别任务相对更难。总体来说,命名实体识别任务的困难体现在如下的方面[3]:

- 1)不同文本下的命名实体类型差异很大:命名实体很难用几个类型完全概括。

同一类实体下不断有新的实体出现,测试数据出现训练数据中所没有的未登录词,很难建立一个大而全的命名实体库。

2)命名实体的构成较为复杂:在特定领域的一些命名实体对于长度没有限制。不同类型的实体结构差异大,例如,医疗领域中的组织名一级特定医学术语中往往存在大量缩略词,并且还存在于医学方面特有的名称嵌套问题等。

3)在不同领域下,命名实体的分类标准不同:地名和组织名中可以包含人名,医学领域病毒名称可以包含细胞名称。这些名词及容易和常规名词混淆,容易出现标注失误,影响识别效率。为了正确标注,往往需要能够理解上下文语境的行业专家进行标注。

4)中文命名实体识别的环节往往与分词、浅层语法分析等过程处于同一任务的不同环节的关系:分词,语法分析的准确性直接影响了实体识别的准确性。这使得中文实体识别具有一定的困难。

在医学领域中,医学文本中有大量真实的个人案例信息与病理知识信息,潜藏着丰富的医学价值。而医学命名实体识别(Medical Named Entity Recognition, MNER)则是医学文本信息抽取任务的基础,该工作的好坏则对其他任务有着非常重要的影响,比如医学实体关系抽取,医学事件抽取等任务,医学命名实体识别是医学文献信息检索以及医学知识图谱构建的重要前提。开放域实体识别任务主要研究对象是地名、机构名等,如北京大学发表的中文命名实体“人民日报语料”。而医学命名实体识别则主要针对症状、手术等医疗实体。这些实体存在着大量的简写、缩写或模糊的表达及专业术语一词多义或多词同义的情况,例如“N-乙酰半胱氨酸”和“N-乙酰基 - 半胱氨酸”都指的是同一个命名实体;“TCF”可以指“T 细胞因子”,也可以指“组织培养液”。

此外,由于大量医学术语的存在,中文语料相比于英文语料还存在实体边界识别困难的问题,缺乏针对中文医学语料开源的高效分词工具。同时考虑到现有中文医学实体的标注数据较为稀缺的情况,仅仅利用传统开放域实体识别模型直接对医疗命名实体语料进行训练很难取得较好的效果,相对英文的医学实体识别任务,中文医学实体的识别具有更大的挑战性。在此当前条件的限制下,很难直

接获取更多带标注医学语料，因此考虑是否能够利用其他辅助去提高中文医学实体识别的精度，能够获得的语料信息包括不带标注的医学语料，线上医学词典，甚至一些包含丰富语义信息知识图谱三元组。如果能高效利用这些辅助信息去提高识别的精度，不但可以提高下游医学文本挖掘的精度，对于其他垂直领域的信息抽取效率的提高也有极大的促进作用。

1.2 本文主要工作与创新点

基于上述的研究背景，本文主要针对中文医学语料的命名实体识别问题进行探索。对于中文语言来说，最小的构成单位是字，而英文语言最小构成单位的是词，鉴于分词错误所导致的错误传播的影响，在中文实体识别最常用的输入单位是字。本文使用了医学电子病历数据集和医学线上问答数据集两份不同文本风格的中文医学语料，尝试解决以下中文医学命名实体的问题。

1)在传统的统计学习方法中，提高模型实体识别的精度主要依赖于手工设置的特征模板的好坏。在深度学习模型兴起之后，通过调节网络超参数的方法很大程度上取代了以往的设置特征模板方法。在不同中文医学语料中，是否基于深度学习方法依旧能够超过传统的统计学习方法？传统的特征模板设置对于深度模型是否具备一定的借鉴作用？

2)深度学习模型的效果受到训练语料规模的限制，在语料规模既定的情况下能否利用辅助信息进行模型效果的提升。医学词典作为一种容易想到的辅助信息，以何种方式融入中文医学文本输入的字或词当中能够较好的保存词典的信息？对于中文的字与词两种输入粒度形式，倘若字与词同时融合了医学词典的辅助信息，字与词之间需要如何融合会使得辅助信息的效果得到最大的发挥？

3)近期以 BERT 为代表的预训练语言模型体现出了强大的上下文语义表征能力，在不同的中文医学文本的实体识别任务上是否也能够取得较好的效果？如果词典特征能够对于基于 LSTM 模型的效果有较大的提升，是否对于基于 BERT 微调的模型也依旧能够有提升？此外，知识图谱以三元组的形式包含了丰富的领域

知识信息，能否以一种方式将此种信息注入到基于 BERT 的模型上，以进一步提高模型表现？

基于以上三方面的问题，本文提出了相应的解决方案，主要工作与创新点如下：

1)探索开放域实体识别模型在医学这个垂直领域上的实体识别效果：本文使用了医学电子病历数据集和医学线上问答数据集两种不同风格的中文医学文本进行实验，利用 Linear-Chain-CRF 模型和 BiLSTM-CRF 模型两种开放域实体识别的经典方法进行实验，通过添加特征模板和优化模型超参数的方法下发现基于神经网络的方法会取得更好的效果，但同时发现加入医学词典特征与医学知识特征的 Linear-Chain-CRF 模型在某些医学实体上的识别效果可以超过深度学习模型的效果，于是开始进一步探索如何在深度学习模型中更好地融入医学词典信息与领域知识信息。

2)探索如何在医学实体识别模型中更好地融入医学词典辅助信息：本文从不同大众健康网站，医学期刊等不同风格文本中爬取得到医学文本语料与医学词典作为辅助信息，在此基础上设计了上下文词典特征与词界特征用于融入字向量，并利用医学词典进行分词得到词向量，实验发现更新的词向量与字向量均取得更好的效果。为了进一步融入词典特征，本文基于词是作为字的信息补充或者词与字的地位对等两种先验假设的不同设计了字词串联法模型和字词并联法模型，包含直接串联法、间接串联法与后置并联法、前置并联法四种模型，验证了不同字词融合方式对于医学词典信息融入效果的差异，探索出在当前任务下较好的医学词典信息融入方式。

3)探索在医学实体识别问题中使用预训练语言模型以及领域知识信息融入的效果：本文使用 BERT 预训练模型分别在两类医学数据集上进行微调，并验证了添加医学词典特征的方式对于 BERT-CRF 模型也能有效果上的提高。并以知识图谱为载体进行领域知识融入，借鉴了 K-BERT 模型中软位置编码与可见矩阵的思路，设计了适用于中文医学数据的 CMK-BERT 模型，在对数据进行一定转换之后，模型在融入医学词典特征基础上分别融入 CN-DBpedia、知网(HowNet)和医

学知识图谱(MedicalKG)三种不同类型的知识图谱,利用知识图谱领域信息使得医学实体识别效果取得进一步的提高。

1.3 本文组织结构与行文思路

第一章部分为绪论,介绍了自然语言处理中命名实体识别任务研究的现状,和这个领域研究所面临的困难。接着针对医学领域这个垂直领域的命名实体识别问题提出其所面临的特定的问题和这些问题的解决所能带来的现实意义。并继续介绍了国内外针对医学命名实体识别相关的研究方法。

第二章介绍了文本模型的背景知识,包括统计学习模型 **Linear-Chain-CRF**,深度学习模型 **BiLSTM**,基于 **BERT** 模型的预训练和目前将知识图谱融入 **BERT** 的最新工作。

第三章首先介绍了命名实体识别问题基于统计学习与深度学习方法的两种实验流程,单个实体识别与多实体识别的测评指标以及本文实验所用到的医学数据集,分别使用 **Linear-Chain-CRF** 模型和 **BiLSTM-CRF** 模型进行实验的优化与结果分析。

第四章主要介绍了医学词典信息等外部信息的获取和使用并将医学词典信息以不同方式融入到模型之中。另外,还提出基于不同字词融合方式的深度学习模型,包括字词串联法模型以及字词并联法模型,详细分析了不同模型在不同数据集上的效果和模型之间的比较。

第五章利用预训练好的 **BERT** 模型进行在医学命名实体识别领域的应用探索,并添加了继续添加医学词典来提升模型效果。为了更好的融入知识图谱领域信息,本文借鉴了软位置编码与可见矩阵的思想,提出了 **CMK-BERT** 模型,并进行了充分的实验与分析。

第六章部分为全文总结,并提出目前工作的不足,并且对于后续工作提出了期待和展望。

1.4 国内外研究现状

1.4.1 基于词典和启发式规则的方法

和开放域的实体识别问题相似，在医学领域的实体识别中最早使用的也是字典的方法。Proux 等于 1998 年使用英语词典，将属于词典中的医学名词为实体，不属于医学词典的名词不是实体[7]。这种方法虽然简单，而且准确率往往非常高，但医学领域命名实体的词库在不断更新，而且很多命名实体存在变体和多种写法，想要简历一个词典去包含住所有的命名实体是不可能的，这导致了一个较低的召回率。

但词典作为最早期的方法并非完全没有借鉴意义，很多文献会将词典特征融入机器学习方法[8]，甚至去融入神经网络中。如 Long 等利用在线医疗资源将语义信息融入了医学词典中，将条件随机场方法融入医学词典中，得到的 F 值为 0.8372，显著超越了之前纯粹使用条件随机场方法的表现[9]。但如果想要进一步提高精度。要考虑到词典本身词汇覆盖率，更新速度的局限。

启发式规则方法也在早期广泛被应用在生物医学领域的实体识别研究当中，Fukuda 等人开始利用规则，如固定的前缀后缀去判断蛋白质的名称[10]。Tsuruoka 等人第一次在医学领域将启发式规则融入词典，对于医学术语的歧义性和变化，利用规则进行规范化以提高查找效率。但设计这些规则需要专家的领域知识和多年的经验，随着新的实体不断涌现，需要耗费专家和工程人员巨大精力去维护和扩展领域规则。

尽管如此，耗费早期专家大量精力的方法也被融入到后期的机器学习方法[11]，甚至深度学习方法中。如 Wei 等在条件随机场的模型的基础上加入了一系列基于规则的逻辑进行处理[12]。启发式规则的方法主要考虑边界、中心词、词性等特征来整理出一系列规则库。但是规则库的规模随十分庞大且难以总结地完全充分。

有研究者使用 Bootstrapping 方法去自动生成规则[13]，去弥补人工整理方式的不足。

1.4.2 基于统计学习的方法

统计学习的方法自上世纪 90 年代开始活跃于自然语言处理任务的各个领域，同时带动了医学文本中的命名实体识别技术的发展。统计学习方法可以从样本数据中估计出相关参数和使用的特征而建立一个识别的模型，具有较强可移植性，但一个拥有较多参数的复杂模型往往需要较多的训练数据。统计学习方法根据对于标注语料的依赖程度，可以分为完全需要语料的监督学习，部分需要语料的半监督学习以及理论上无需语料的无监督学习三种类型[14]。监督学习是在有 label 的数据集上进行模型训练，再去训练好的模型去预测没有 label 的数据。半监督学习往往其使用场景较为缺少标注数据，先使用一部分的标注数据进行训练，利用未标注数据的分布信息去辅助区分两个类的分类超平面。无监督学习是指对于无类标数据进行直接建模，通常在自然语言处理领域使用相关聚类算法去利用上下文语料。在医学领域中的实体识别主要还是使用监督学习的方法。

命名实体识别的问题可以看作是对于词的分类问题，因此可以使用一些经典分类模型支持向量机(Support vector machine, SVM)，最大熵(Maximum entropy, ME)模型等。如 Wei 等使用了支持向量机分类器去构建疾病命名实体识别系统和进行实体标准化[15]。Roberts 等人使用支持向量机方法对于影响心脏病的因素进行了识别[16]。

此外，带实体类标的字符串序列可以看作是标签序列，实体识别又可以被看作是序列标注的问题，其思想是按照序列化的方法对字符串进行标注，最终选择联合概率最大的标注序列[17]。主流的序列标注模型包括最大熵马尔可夫模型、隐马尔可夫模型和条件随机场等,如图 1.1 所示。Settles 等使用语言形态学特征，语法特征，正交特征构建了名为 ABNER 的系统，是最早将 CRF 模型用于实体识别的系统，在 JNLPBA 2004 任务中取得了接近 70%的 F1 值，成为了当时的冠军

[18]。Liu 等在中文电子病历的命名实体识别中也使用了 CRF 方法，并在其中加入四种新特征，其中 F 值达到了 89.152%，相对于传统的分类模型算法，基于序列标注的模型往往可以取得更好的效果[19]。

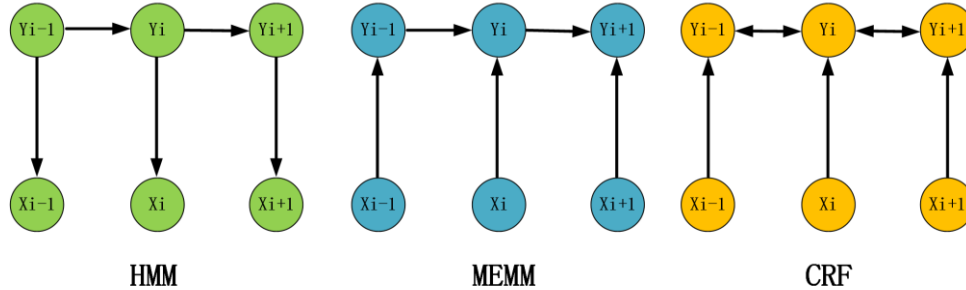


图 1.1 HMM、HEMM、CRF 示意图

单一的统计学习方法以及词典，启发式规则的方法都有各自的优缺点。因此只使用单一的某个模型很难在多个生物医学领域的实体识别场景中表现出较好的效果。因此学者考虑将多个模型进行集成去提升总体的识别效果。Wang 等人使用了仲裁规则法等方法去组合支持向量机，条件随机场等模型并应用在英文语料 JNLPBA2004 上，最终 F 值达到了 77.57%，优于任意单一模型的效果[20]。Li 等人在 BioCreative II 数据集使用了多种特征组合以及统计学习方法中的若干模型并使用了集合操作、投票策略以及双层叠加的方法进行对比，最终的 F 值达到 88.42%[21]。

在本文统计学习方法有关实验中，主要结合医学文本的特点设计不同的特征组合在经典的线性链条件随机场模型上进行效果的对比。

1.4.3 基于深度学习的方法

深度学习是机器学习的重点研究领域。该方法试图模仿大脑神经元的传递，将模型表示为层次结构[25]。CNN 网络 Alexnet 模型获得了 2012 年 ImageNet 的图像识别比赛冠军，从此引发了深度学习的研究热潮，并逐渐在自然语言处理领域产生较大的影响。

传统上的机器学习方法应用命名实体识别时，往往依赖于手工设计的特征工

程对于文本进行数字向量的表示，根据每个单词得到的向量表示再将单词分为特定的实体类别。尽管这样的处理方法在处理医学的实体识别问题上非常有效，也在很长一段时间内是主流的命名实体识别方法，但其项目的主要精力落在构造特征工程上面，并且构造好的特征工程在不同风格不同领域的文本上没有很强的移植性。

而基于深度学习的方法将神经网络作为特征抽取器，避开了手动设置特征工程的步骤，对于大规模的训练集比起传统的统计学习方法具有一定的优势。在深度学习领域中，卷积神经网络，循环神经网络，与其改进版本长短时记忆网络以及最近兴起的 Transformer 模型，这些神经网络结构都可以作为特征抽取器去替代手工特征的设计。

循环神经网络输入序列数据，以序列开头到序列结束的方式进行运算。通过每个时间步的权值共享和定义时间步的长短去模拟任意长度的输入序列[26]。Hu 等在 CCKS2017 的比赛中，使用了一个基于规则和 RNN+CRF 的系统，在更为严格的标准下，依然可以达到 94.26% 的 F1 值。但 RNN 往往伴随着梯度消失或者梯度爆炸的问题，对于梯度爆炸可以采用梯度裁剪等方法进行缓解，而梯度消失的问题可以通过使用 RNN 上的变体 LSTM 去缓解，在过长的序列建模时比 RNN 取得的更好的效果。此外，CNN 通过共享卷积核，移动滑动窗口的方式对上下文信息进行建模，在医学命名实体中也是比较常用的模型。Zhao 等在 NCBI 语料库上利用多标签卷积神经网络(MCNN)进行疾病的实体识别，利用多重标签策略去替代 CRF 层，得到的 F1 值为 85.17%，高于当时其他方法的效果[27]。

Huang 等首次提出 BILSTM-CRF 的模型，并在模型中融入了自定义的拼写特征，上下文 n-gram 的特征和词典特征，在词性标注、命名实体识别等多个序列标注任务中达到了当时最好的效果[28]，BILSTM-CRF 由于在多个数据集上具有较强的模型稳定性，可以作为整个命名实体识别领域的经典模型。其后的大多模型都在 BILSTM-CRF 之下，它们的不同主要在于编码端的不同处理方式。如 Peter 等人基于预训练语言模型去丰富词向量的特征，并加入到 BILSTM-CRF 网络中[29]。Devlin 等使用著名模型 BERT 的 Transformer 编码端进行预训练，改进了原

先 LSTM 不能并行的缺点，再在 NER 的标注数据上进行微调，提高了原先 BILSTM-CRF 框架的效果[30]。

1.4.4 引入外部信息辅助的方法

对于医学领域，一个较为明显的外部信息就是医学词典。龙光宇等在条件随机场模型中将词典作为 n-gram 的特征加入特征模板中，设计了不同的词典特征模板[9]。Xu 设计了直接输入法和间接输入法在神经网络中融入大众医疗词典信息[31]。

除了字典这样的典型外部信息，模型还可以进一步习得语义层面的知识，例如知识图谱就包含了丰富的语义信息。百度近期提出了基于知识增强的语义表示模型 ERNIE[32]，先利用 TransE 算法这样的表示学习方法去获取对应实体的向量，再并入到 BERT 中，在包括命名实体识别等多个任务中的模型效果全面超越 BERT 模型，训练语料引入了部分外部知识，对其他类别文本进行学习以增强语义表示能力。

无独有偶，为了将知识图谱的中的语义信息与语言模型进行结合。北大与腾讯联合推出了 K-BERT 模型则是对 BERT 模型的改进[33]，该模型使用统一的向量空间将知识注入到模型中，并提出了软位置和可见矩阵技术解决了三元组信息处理输入序列模型中的信息损失问题，在 8 个开放领域测评任务和 4 个依赖背景知识的特定领域中超过了 BERT 模型的结果。

本文将利用最新的 K-BERT 模型的领域信息融合方法对于医疗领域实体识别数据集进行进一步的探索与分析。

第二章 背景信息介绍

2.1 统计学习模型 Linear-Chain-CRF

条件随机场(Conditional Random Fields, CRF)是一种结合隐马尔可夫模型与最大熵模型优点的判别式模型[35]。本身是无向图，对结构是没有要求的，而大多数自然语言处理的问题大多数为序列问题，对应的模型则属于线性链条件随机场模型(Linear-Chain Conditional Random Field, CRF)，其结构如图 2.1 所示。

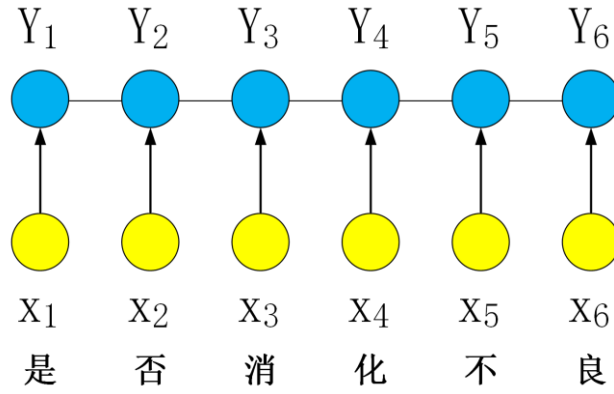


图 2.1 Linear-Chain-CRF 模型结构图

设 $X = (X_1, X_2, \dots, X_n)$, $Y = (Y_1, Y_2, \dots, Y_n)$, 在 Linear-Chain CRF 中马尔科夫性形式化定义如 2.1:

$$P(Y_i | X, Y_1, \dots, Y_{i-1}, Y_{i+1}, \dots, Y_n) = P(Y_i | X, Y_{i-1}, Y_{i+1}) \quad (2.1)$$

基于 Hammersley-Clifford 定理，可以将无向图的联合概率分布表示成所有最大团上势函数乘积的形式，形式化定义如 2.2:

$$P(Y | X) = \frac{1}{Z} \exp\left(\sum_{i,k} \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} u_l s_l(y_i, x, i)\right) \quad (2.2)$$

式子中 t_k 和 s_l 是特征函数， t_k 代表转移特征，受当前状态和上一个状态的影响。 s_l 是定义在节点上的特征函数，收当前状态的影响。而 λ_k 和 u_l 是特征函数所对应的权值。特征函数 t_k 和 s_l 取值只能为 1 或 0，满足某个特征条件时取 1，不然

就取 0。

2.2 深度学习模型 BiLSTM

LSTM:循环神经网络(Recurrent Neural Networks, RNNs)是一种处理长文本序列元素数据的神经网络,其被称为循环网络是因为它对于序列的每个元素执行相同的任务。其变体模型长短期记忆网络(Long Short Term Memory networks, LSTM)极大地缓解了长距离依赖的问题,其被广泛的应用在了包括实体识别等自然语言处理的领域[36],其结构包括三态门:输入门,遗忘门和输出门这三个门来控制细胞状态(Cell State),第 t 时刻 LSTM 单元的主要计算公式如 2.3:

$$\begin{aligned}
 i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \\
 f_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \\
 c_t &= f_t c_{t-1} + i_t \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \\
 o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \\
 h_t &= o_t \tanh(c_t)
 \end{aligned} \tag{2.3}$$

其中 i, f, o 和 c 分别表示输入门,遗忘门,输出门和记忆单元。 σ 表示 sigmoid 激活函数, W 表示权重矩阵,如 W_{hi} 表示隐层与输入门的权重矩阵。

本文为了有效建模上文和下文的信息,使用了双向 LSTM(BiLSTM)网络,其可以展开为两个单独的 LSTM 网络,分别对于输入的序列任意第 i 个字符采用正向和逆向顺序得到两个单独的隐层表示 \vec{h}_i 和 \overleftarrow{h}_i , 将其进行拼接或者相加得到一个最终的表示 h_i 。

2.3 基于 BERT 模型的预训练

出于从大量语料中得到获取外部知识,进行迁移学习的目的,人们尝试在未标注语料上使用预训练的语言模型。如 Peter 等人使用多层的双向 LSTM 语言模型得到的 Elmo 模型[37]和 Radford 等基于 Transformer 与 LM 模型得到的 GPT 网

络[38]。但都存在各自的缺点，如 GPT 网络只使用了单向信息进行建模，Elmo 模型仅仅是将双向的信息拼接在一起，并没有做到真正的语义融合。而 BERT 将双向的 Transformer 运用到语言模型，并使用了基于遮罩的语言模型，第一次进行了完整的双向语言模型的训练。由于 Transformer 的编码器解决了 LSTM 模型递归问题，可以同时基于单词的左右两侧进行学习，实现了真正双向。

模型输入：在 BERT 模型中，其输入表示可以由三个 Embedding 相加得到。输入可以分成两个部分，如图 4.16 所示。其中是[CLS]是特殊的符号，“my dog is cute”是完整的话，中间插了一个分隔符[SEP]，包括 mask 总共三种特殊符号。第一个向量 Token Embeddings 是词向量，在中文语料中可以是字向量。该表示将单词划分成一组公共子词单元，如“playing”被拆分成了“play”和“ing”。CLS 标志可以用于下游分类任务，若是非分类任务可以忽略[CLS]标志。第二个向量是 Segment Embeddings，其用来区别两种句子，因为 BERT 预训练不光进行语言模型，还要做两个句子的分类任务。第三个向量 Position Embeddings 是位置编码，和之前 Transformer 中的位置编码不同，不是三角函数而是学习出来的，将位置信息编码成特征向量，使用相对位置编码。然后对三个 Embedding 进行相加，BERT 的输入表示如图 2.2 所示。

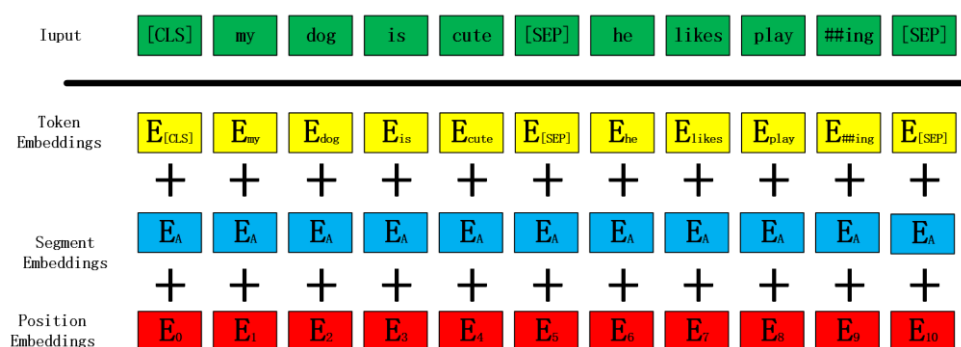


图 2.2 BERT 模型示意图

模型预训练：BERT 模型为了能够基于大规模语料进行无监督训练，巧妙的设计了两个无监督预测任务：遮罩语言模型(Masked Language Model,MLM)和下一句预测(Next Sentence Prediction,NSP)。

MLM：BERT 模型为了训练双向 Transformer,随机遮罩(mask)了句子中 15%

的 token，利用上下文去预测这些被遮罩的 Token，当编码器不知哪些词会被替换时，它需要对每个单词给予上下文的表示。此时 Masked LM 模型学到的词表示是真正融合了词的上下文。

下一句话预测：由于 BERT 之后的下游任务可能包括问答(Question Answering, QA)，需要句子级别信息向量。因此添加了两个句子的连续性预测任务，即预测输入 BERT 的文本是否是连续的上下文。训练时，模型第二部分以一定的概率从全部文本中抽取(负例)，剩下则为正例，进行 sentence-level 二分类问题，判断句子是否真的是下一句。

在命名实体识别的应用场景下，可以使用 BERT 模型替换了原来网络得到训练词向量的部分，从而构成新的 Embedding 层，或者直接使用 BERT 进行序列预测。

2.4 知识图谱信息的利用

在各种辅助信息中，知识图谱是一种新的存储方式，包括结点：代表实体(entity)，以及边：代表实体之间的各种语义之间的关系(relation)。一般来说，一个知识图谱由若干个三元组 (h, r, t) 组成，其中 h 和 t 代表一条关系的头结点和尾节点， r 代表关系，在有些文献中称为(主语，谓语，宾语)的 RDF 三元组，其本质是连接实体间关系的图，即揭示实体之间关系的语义网络[39]。目前开源知识图谱既有类似以百科知识数据为基础构建，也有以领域知识或语言学知识为基础构建，如 Wikidata、Freebase、NELL、YAGO 等。

知识图谱中包含的丰富的语义信息如何有效转化为先验知识，辅助模型的训练，是目前图谱应用的一大挑战。

目前以 BERT 为代表的预训练语言模型，尽管在许多任务中取得了很好的效果，但模型本身不能够将知识信息融入到语言理解中。将外部信息融于语言表示的预训练模型有两个挑战。

1)结构化知识编码：对于给定的文字，如何为语言表征模型有效地抽取和编码与 KG 对应的信息实体。

2)异构信息融合: 基于语言模型预训练过程与图表示学习过程略有不同, 这会产生两个独立的向量空间, 需要设置特殊的预训练任务来融合词汇、语义和知识信息。

清华大学在 ACL 2019 会议中提出的 ERNIE 模型, 其利用了 TransE 算法在知识图谱中进行了特征学习, 并完成了文本与图谱的实体对齐[32]。在 MSRA-NER 数据集上对开放域的命名实体进行了识别, 超过了 BERT 模型的表现。

针对 ERNIE 模型中将实体向量引入语言模型中存在的知识噪音问题, 北京大学与腾讯在 AAAI-2020 会议中提出的 K-BERT 模型对于实体向量和词向量使用了统一的向量空间, 并提出了软位置和可见矩阵技术解决了三元组信息处理输入序列模型中的信息损失问题, 在 8 个开放领域测评任务和 4 个依赖背景知识的特定领域中超过了 BERT 模型的结果[33]。

文本利用最新的 K-BERT 模型探索在融入知识图谱的语义信息后, 对于中文医学文本命名实体的模型效果增强。

第三章 基于开放域实体识别模型实验

区别于医学这样的垂直领域命名实体识别，开放域的命名实体识别更多关注的是人名、地名、机构名等在大多数文本中出现的实体，这样的实体很多可以在常规的词典中找到。基于统计学习模型的 Linear-Chain-CRF 和基于深度学习的 BiLSTM-CRF 模型是开放域命名实体识别的代表性模型。

McCallum 等[42]使用条件随机场模型在 CoNLL-2003 新闻数据集上识别人名、地名、机构名等实体，灵活设计特征模板的方法受到了学界广泛的关注，直到现在也是实体识别领域的常用方法之一。而双向长短期记忆网络与条件随机场的模型可以直接利用神经网络进行文本特征学习，不依赖其他条件。在标注信息充足的条件下可以在大量实体识别任务上取得较好的效果[43]。本节采用开放域的这两个经典模型在医学电子病历和医学在线问答的两个医学领域数据集上进行探索。

3.1 命名实体识别实验流程与评价指标

3.1.1 问题描述

命名实体识别问题，可以将其理解成单个词的多分类问题(每个词单元的多个标签可以理解为多个类别)，也可以看作是一种寻找全局最优的序列标注任务。命名实体识别任务要求可以检测出实体的边界并将实体归类到预先定义好的类别当中。对于不同语言的文本，其常用的输入粒度与数据预处理方法不同。对于英文而言，由于单词之间基本以空格为切分，少有分词错误的问题，因此输入的格式可以单词(word)分割为主；而对于中文来说，由于分词错误会导致实体识别的错误累加，因此主要的输入粒度为字(Character)。目前中文医学实体识别数据集相对较少，且相较英文数据集上识别的正确率较低，相较于与英文医学的实体识别问题是更大的挑战，本文主要研究中文医学命名实体识别的问题。

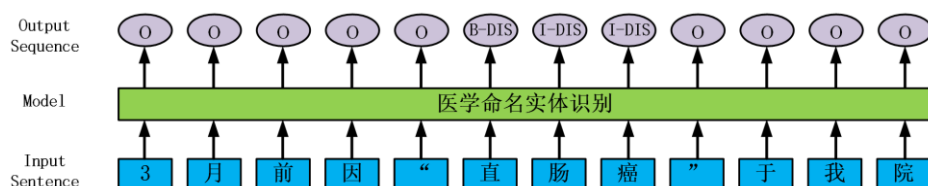


图 3.1 中文语料实体标注图

对于标注系统方面，常见的标注方法有 BIO、BIOS 和 BIOES，本文采用 BIO 标注体系，B 代表实体的开始，而 I 代表实体的中间或者结束，O 则表示其他非实体。如图 3.1 所示，“直肠癌”是一个表示疾病的命名实体，“DIS”表示实体类型是一种疾病，分别用 B、I 表示了直肠癌的开始位置和非开始位置。

3.1.2 两种解决流程

无论是使用深度学习的方法还是统计学习方法，前期对于文本数据的预处理都是相同的，都要进行词汇的切分，将文本整理成便于训练的 BIO 实体标注的格式。

对于统计学习的模型而言，模型需要人工设计特征模板，特征工程的好坏很大程度上会决定模型效果的优劣。当数据预处理阶段与训练数据与验证、测试数据的切分完毕时，若手工添加的特征工程训练好的模型在验证数据上可以较好的效果则可以完成整个流程，若达不到理想的效果则需要根据分析模型错判的样本去更改特征模板或者添加一些后处理的方法，整个流程绘制如图 3.2 所示。

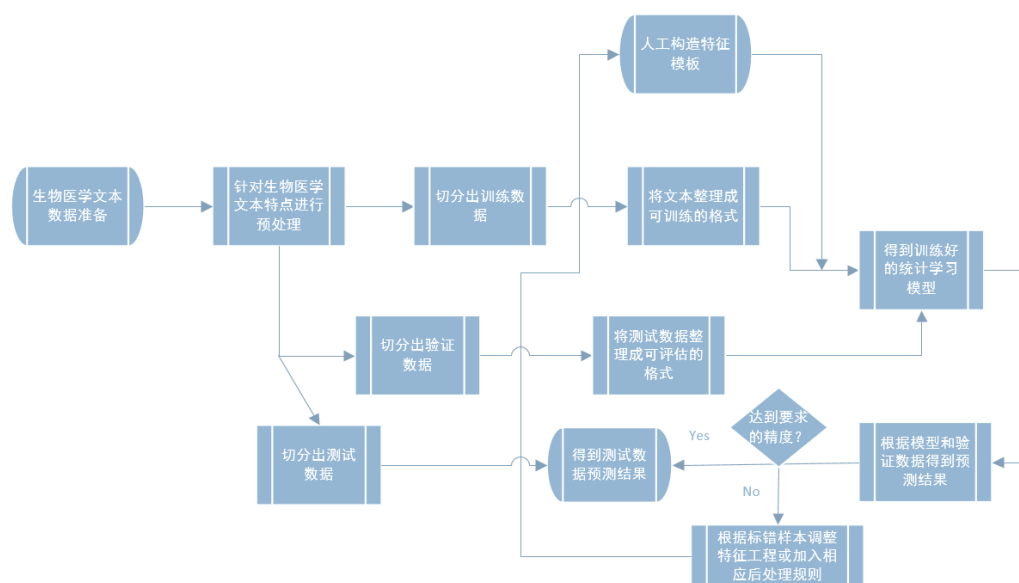


图 3.2 基于统计机器学习方法的实验流程图

而在深度学习的方法中，由于采用了神经网络的结构作为特征的抽取器，避开了纯人工设计的特征模板，影响结果的主要因素是模型参数的调整和模型架构的设计。模型架构往往可以分为嵌入层、编码层与解码层，在深度学习的方案中可以尝试基于模型标错的样本，在嵌入层中在词向量或句向量的基础上加入新的

文本特征融入神经网络中，也可以进一步调整超参数与架构，流程图如图 3.3 所示。

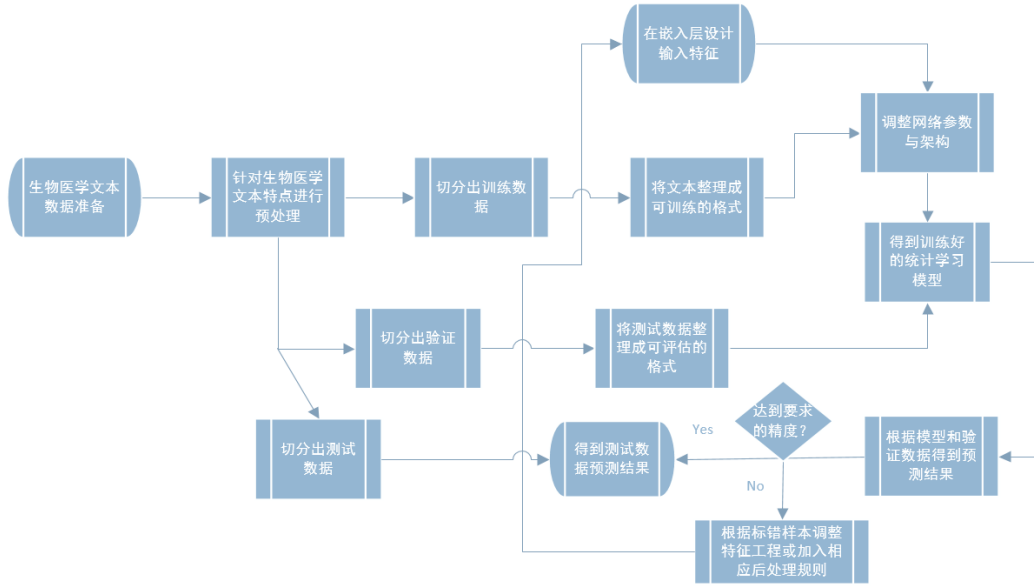


图 3.3 基于深度学习方法的实验流程图

3.1.3 严格指标与松弛指标

由于命名实体识别问题也可以看作是对于词的多分类问题，多分类问题从集成学习的思路往往也可以看作若干二分类问题的集成。可以使用二分类问题的方法来解释测评指标，机器学习常用的测评指标一般包含四个，分别是正确率(Precision, P)，召回率(Recall, R)和 F1 值(F1-Score)。这些测评指标由如下四个统计量计算得到：

真阳性指标(True Positive, TP)：预测为正确，实际也为正确的实体个数

假阳性指标(False Positive, FP)：预测为正确，实际却为错误的实体个数

假阴性指标(False Negative, FN)：预测为错误，实际却为正确的实体个数

真阴性指标(True Negative, TN)：预测为错误，实际也为错误的实体个数

$$P = \frac{TP}{TP + FP} \quad (3.1)$$

$$R = \frac{TP}{TP + FN} \quad (3.2)$$

$$F_{\beta} = \frac{(\beta^2 + 1)P \cdot R}{\beta^2 P + R} \quad (3.3)$$

当 β 取 1 的时候的 F 值即为 F_1 ，该情况下均衡的考虑了召回率和正确率的大小。

在命名实体识别问题中，将模型输入的结果集合记为 $S = \{s_1, s_2, \dots, s_m\}$ ，人工标注的识别类别集合为 $G = \{g_1, g_2, \dots, g_n\}$ ，每个集合元素为一个一个实体提及，可以表示成四元组 $\langle d, pos_b, pos_e, c \rangle$ ，其中 d 表示文档， pos_b 和 pos_e 分别对应实体提及的起始和终止下标， c 表示提及所属的类别，可以有两种层面的评价标准。

严格指标：定义 $s_i \in S$ 和 $g_j \in G$ 严格等价当且仅当：1. $s_i.d = g_j.d$ 2. $s_i.pos_b = g_j.pos_b$ 3. $s_i.pos_e = g_j.pos_e$ 4. $s_i.c = g_j.c$

基于以上的等价关系，可以定义集合 S 与 G 的严格交集为 \cap_s ，可以得到严格测评指标为：

$$P_s = \frac{|S \cap_s G|}{|S|} \quad (3.4)$$

$$R_s = \frac{|S \cap_s G|}{|G|} \quad (3.5)$$

$$F_{1s} = \frac{2PR}{P + R} \quad (3.6)$$

松弛指标：定义 $s_i \in S$ 和 $g_j \in G$ 严格等价当且仅当：1. $s_i.d = g_j.d$ 2. $\max(s_i.pos_b, g_j.pos_b) \leq \min(s_i.pos_e, g_j.pos_e)$ 3. $s_i.c = g_j.c$

基于以上等价关系，定义集合 S 与 G 的松弛交集为 \cap_r ，可以得到松弛测评标准为：

$$P_r = \frac{|S \cap_r G|}{|S|} \quad (3.7)$$

$$R_r = \frac{|S \cap_r G|}{|G|} \quad (3.8)$$

$$F_{1r} = \frac{|2PR|}{|P + R|} \quad (3.9)$$

可以看出，松弛指标与严格指标的差距主要在于实体边界上的识别上，如果符

合严格指标一定符合松弛指标，因此本文对于特定某一类的医学实体识别的效果采用严格指标进行评估，这也是学术界最常采用的评价标准。

3.1.4 多类实体的微平均指标

由于本文采用的两份医学数据集包括了多种实体，从下面的 2.2 小节也能看出不同类别的实体数据量并不均衡。若对于多类实体的总体评估指标仅仅是对于每一类实体的评估指标进行一个平均显然没有考虑到两份数据集上类别不均衡的问题。因此本文采用的是微平均指标，即对于每一个样本不分类别去建立一个全局的混淆矩阵。假设数据集有 N 个类别，那么相对于二分类的问题，可以得出 N 分类下的微平均指标如下：

$$\begin{aligned}
 Micro-P &= \frac{\sum_{i=1}^N TP_i}{\sum_{i=1}^N TP_i + \sum_{i=1}^N FP_i} \\
 Micro-R &= \frac{\sum_{i=1}^N TP_i}{\sum_{i=1}^N TP_i + \sum_{i=1}^N FN_i} \\
 Micro-F_1 &= \frac{2 \times Micro-P \times Micro-R}{Micro-P + Micro-R}
 \end{aligned} \tag{3.10}$$

在文本中对于多个实体的总指标在没有特殊说明的情况下使用的都是微平均指标。

3.2 中文医学实体识别数据集

本次用于模型训练的数据集主要包括了由医渡云发布，用于 CCKS2019 实体识别任务测评的中文电子病历数据集。要求从医疗文本中识别实验室检验、影像检查、疾病和诊断、解剖部位、手术和药物 6 种命名实体。阿里巴巴研究团队在原始文本的基础上进行了一定程度上的扩充，又标注了第 7 种症状实体，本文使用的是经过扩展的数据集[34]。经统计后，训练集验证集与测试集的实体分布如表 3.1 所示。

表 3.1 CCKSNER 数据集数据实体分布表

实体类别	疾病和 诊断	影 像 检 查	实验室 检验	手术	药物	解剖部 位	症状
训练集	3824	222	318	946	1646	5623	2095
验证集	173	55	55	52	84	252	78
测试集	149	29	31	43	72	220	88

由于医学电子病历是由医生亲自书写,内容较为规范,较少出现错误的文本,且包含大量医学专业用语,因此作为本次实验最为主要的实验数据,以下简称 CCKSNER 数据集。

原始文本以 json 文件的格式保存,其对象属性分别保存了原始文本,与文本中每一个实体提及的真实标签与位置信息,如图 3.4 所示。

```

"text": "患者3月前因“直肠癌”于我院行全麻上行直肠癌根治术(dixon术),手术过程顺利,术后给予抗感染及营养支持治疗,患者恢复良好,切口愈合良好。",
"mention_data": [{"mention": "直肠癌", "label": "疾病和诊断", "offset": "8"}, {"mention": "直肠癌根治术(dixon术)", "label": "手术",
"text": "患者因罹患“胃癌”于2013-10-29在我院行全麻上胃癌根治术,术中见:腹腔内腹水,腹膜无转移,肝脏未触及明显转移灶,肿瘤位于胃体、胃底",
"mention_data": [{"mention": "胃癌", "label": "疾病和诊断", "offset": "7"}, {"mention": "胃癌根治术", "label": "手术", "offset":
"text": "患者3月余前于我院诊断为“直肠癌”,于2015-10-26在全麻上行腹腔镜直肠癌根治术,术后病理示:201518502:(直肠)腺癌(中度分化),浸润深",
"mention_data": [{"mention": "直肠癌", "label": "疾病和诊断", "offset": "14"}, {"mention": "腹腔镜直肠癌根治术", "label": "手术",

```

图 3.4 CCKSNER 数据集源文件图

除了专业的电子病历数据,本文也采用了阿里巴巴研究团队在 2020 年 8 月开源的用于中文医学语言理解评估基准的一份开源的医学社区问答数据集[34],经统计后,训练集验证集与测试集的实体分布如表 3.2 所示。以下简称为 CMQANER 数据集。经统计,其包含的 11 种实体的分布如表所示。

表 3.2 CMQANER 数据集数据实体分布表

实体类别	疾病	人群	症状	身体部位	治疗方法	时间	药物	范围	生理机能	检测	科室
训练集	3908	736	2299	2525	1071	212	546	312	384	486	146
验证集	332	48	130	213	107	18	45	27	32	70	13
测试集	432	78	231	238	145	32	62	28	45	49	10

```
{
  "mention_data": [{"mention": "便秘", "guid": "04660884-51b6-46f6-bb3f-f86f4f857a2e", "type": "disease", "offset": "0"}, {"mention": "感冒", "guid": "eab76346-9b40-49f8-8a11-4a6e64e3-6e60-4a01-b4be-c0d899cd5e63", "type": "disease", "offset": "3"}],
  "text": "便秘两个多月不清楚。饮食跟以前一样。换了一次钙以后就便秘了。后来不吃那种钙。也没见好。",
  "mention_data": [{"mention": "前列腺癌", "guid": "69e124f8-c620-4b47-8624-2af6e3b9e323", "type": "disease", "offset": "7"}],
  "text": "治疗费用。治疗前列腺癌的费用？",
  "mention_data": [{"mention": "孩子", "guid": "None", "type": "crowd", "offset": "4"}, {"mention": "感冒", "guid": "eab76346-9b40-49f8-8a11-4a6e64e3-6e60-4a01-b4be-c0d899cd5e63", "type": "disease", "offset": "3"}],
  "text": "请问你。孩子是否经常感冒发炎？鼻塞流涕喷嚏？咳嗽咳痰？哮喘发作？从资料分析看。初步考虑为鼻-鼻窦炎。咽扁桃体炎",
  "mention_data": [{"mention": "疹", "guid": "e430b3df-1367-4583-adc8-ba68bf564f3a", "type": "disease", "offset": "12"}],
  "text": "一两年身上突然长了很多红疹。",
  "mention_data": [{"mention": "iga肾病", "guid": "415d8d2-294b-42b6-a0de-c007d7b137c9", "type": "disease", "offset": "0"}],
  "text": "iga肾病尿酸高怎么办？",
  "mention_data": [{"mention": "附睾炎", "guid": "4ae664e3-6e60-4a01-b4be-c0d899cd5e63", "type": "disease", "offset": "3"}, {"mention": "前列腺炎", "guid": "4ae664e3-6e60-4a01-b4be-c0d899cd5e63", "type": "disease", "offset": "3"}],
  "text": "上海治附睾炎花多少钱我经非淋治疗后引起咽喉炎了和前列腺炎了？我想知道咽喉炎该怎么处理。前列腺炎已经去看医生了。我害怕也引起附睾炎症。精囊炎。"}]
```

图 3.5 CMQANER 数据集源文件图

该数据集提供的原始格式同样以 json 文件存储，保存原始文本与实体提及的信息。可以看出 CQANER 的每份数据均采用问句与答句的形式，相较于电子病历数据更为口语化，文本风格与实体类型较 CCKSNER 数据集有较大不同。

3.3 基于 Linear-Chain-CRF 的实验设计与分析

3.3.1 数据预处理

如 3.2 小节所示，数据集的初始格式是由原始文本以及实体提及组成的 json 文件，无法直接输入模型，需要将其转化为的可以被 CRF 模型可以直接训练或测试的格式。本文采用 BIOES 的标注格式，CRF 模型的输入为字粒度，两个数据集转化后的数据格式如图所示。

```
宫 B_解剖部位
颈 I_解剖部位
活 O
检 O
： O
低 B_疾病和诊断
分 I_疾病和诊断
化 I_疾病和诊断
鳞 I_疾病和诊断
癌 I_疾病和诊断
， O
宫 B_解剖部位
腔 I_解剖部位
少 O
许 O
低 B_疾病和诊断
分 I_疾病和诊断
化 I_疾病和诊断
鳞 I_疾病和诊断
癌 I_疾病和诊断
， O
```

图 3.6 CCKSNER 数据集标注结果图

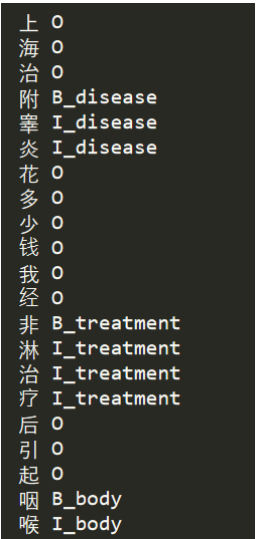


图 3.7 CMQANER 数据集标注结果图

如表 3.1、表 3.2 所示，CCKSNER 数据集为电子病历数据，包括了 7 种实体。CMQANER 数据集为医学线上问答数据集，包含了 11 种实体。两份数据集的实体类型，示例实体，与标注方法如表 3.3 和表 3.4 所示。

表 3.3 CCKSNER 数据集标注表

实体类型	文本中示例	实体开始	实体中部或结尾
解剖部位	盆腔	B_解剖部位	I_解剖部位
影像检查	腹部 b 超	B_影像检查	I_影像检查
实验室检验	血小板	B_实验室检验	I_实验室检验
疾病和诊断	腺癌	B_疾病和诊断	I_疾病和诊断
药物	奥曲肽	B_药物	I_药物
手术	胆管支架植入术	B_手术	I_手术
症状	头晕	B_症状	I_症状

表 3.4 CMQNER 数据集标注表

实体类型	文本中示例	实体开始	实体中部或结尾
疾病(disease)	前列腺炎	B_disease	I_disease
人群(crowd)	宝宝	B_crowd	I_crowd
症状 (symptom)	气喘	B_symptom	I_symptom
身体部位 (body)	喉咙	B_body	I_body
治疗方法 (treatment)	抗生素	B_treatment	I_treatment
时间(time)	三个月	B_time	I_time
药物(drug)	乌鸡白凤丸	B_drug	I_drug
范围(feature)	严重	B_feature	I_feature
生理机能 (physiology)	分泌物	B_physiology	I_physiology
检测(test)	基因检测	B_test	I_test
科室 (department)	口腔外科	B_department	I_department

3.3.2 医学特征模板设计

特征模板设计：Linear-Chain-CRF 模型通过在训练中调整特征的权重来调节模型的精度，因此手工特征模板的选择是最重要的因素，本文将选用的特征分成如下几类。

第一类特征是**字符上下文特征**：包括了 N-gram 的特征，包含了本身的字符，由文本中连续的 N 个字符所组成，本文使用 uni-grams, bi-grams 和 tri-gram 的特征。

N-gram 的特征包含了单向的信息,此外还可以将上文特征与下文特征融合成双向上下文信息,如“是否是有消化不良的可能。”消化不良是疾病命名实体,“是否是”和“有”是上文特征,“的”和“可能”是下文特征。相比于单向的 N-gram 信息,双向上下文特征可以提取更多语义信息。

本文中的字符上下文特征也包括了是否含有特殊字符,医学数据中的实体常常包括了英文字母(如“CT”可以表示诊疗手段,“dixon”可以表示一种手术名称)和一些特殊符号(如“阳性(+)”),此类信息有助于判断实体的类别与边界信息。

第二类特征是**医学词典辅助特征**,包括了词典上下文特征(构建方法见第四章),其类似于 N-gram 的特征,判断的是 N-gram 是否存在于医学词典中。还包括了医学词典分词之后的词边界特征,通过词的边界信息去辅助判断词典的边界信息。由于分词工具往往具备了词性标注(Part of Speech tags, POS)的功能,命名实体大多数以名词的形式存在,名词的边界在一定程度上就是命名实体的边界,于是将分词之后的词性标注特征作为辅助特征之一。

第三类特征是**医学常识特征**:文本手工整理一些医学实体常用的字作为常识特征,如收集了病字头的字如“病”,“疾”,“症”“癌”标记为疾病特征和以月字旁“脏”,“肠”,“腹”的字作为身体部分特征,将“切”,“术”,“疗”等词作为手术特征。将“液”,“片”,“素”等词作为药物特征。通过显式引入医学常识的方式进行了特征的融入。不同特征模板在两个数据集上的综合表现如表 3.5 和表 3.6 所示。

表 3.5 不同特征模板对 CCKSNER 数据集实体识别效果的影响表

特征列表	Precision	Recall	F1 值
字符上下文特征	0.706	0.698	0.702
字符上下文特征+医学词典辅助特征	0.723	0.715	0.719
字符上下文特征+医学词典辅助特征+医学常识特征	0.733	0.726	0.729

表 3.6 不同特征模板对 CMQANER 数据集实体识别效果的影响表

特征列表	Precision	Recall	F1 值
字符上下文特征	0.724	0.617	0.665
字符上下文特征+医学词典辅助特征	0.747	0.653	0.699
字符上下文特征+医学词典辅助特征+医学常识特征	0.75	0.666	0.705

3.3.3 实验结果分析

通过表 3.5、表 3.6 可以发现，在两个数据集上均是使用了三类特征的模型表现最佳，对于 CCKSNER 的数据集，其医疗术语较多，且用词严谨，在使用了医学词典特征的特征之后 F1 值出现了 1.7% 的增幅，说明医学词典作为辅助信息在电子病历中可以覆盖到相当比例的医疗实体，在加入之后医学常识特征之后 F1 值出现了 1.0% 的涨幅。说明电子病历中实体用语的规律性较强，常识信息的加入可以弥补人工收集的医学词典覆盖率的不足。

对于 CMQANER 的医疗问答数据，由于文本风格和实体种类的不同，其总体的实体识别难度高于电子病历数据，在加入医学词典辅助特征之后，F1 值有了显著的提高(3.4%)，说明词典在该数据集上实体的概率较高，而且词典的词边界和文本上的实体边界存在更多的重合。在加入医学常识特征之后 F1 值出现了 0.6% 的提高，则说明医学常识的辅助信息对于问答文本数据也起到了一定的辅助作用。

为了探索两个文本在不同实体上分别的识别精度，将使用三类特征的结果整理成表格形式如表 3.7、表 3.8 所示。

表 3.7 CRF 模型下个 CCKSNER 数据集各实体识别精度表

测试指标	Precision	Recall	F1 值
解剖部位	0.702	0.691	0.696
影像检查	0.793	0.762	0.777
实验室检验	0.800	0.771	0.785
疾病和诊断	0.658	0.658	0.658
药物	0.820	0.858	0.839
手术	0.621	0.621	0.621
症状	0.876	0.843	0.860
总指标(Micro)	0.733	0.726	0.729

表 3.8 CRF 模型下个 CMQANER 数据集各实体识别精度表

测试指标	Precision	Recall	F1 值
疾病(disease)	0.828	0.791	0.809
人群(crowd)	0.878	0.923	0.900
症状(symptom)	0.808	0.734	0.769
身体部位(body)	0.760	0.662	0.708
治疗方法 (treatment)	0.680	0.469	0.555
时间(time)	0.900	0.563	0.692
药物(drug)	0.681	0.525	0.593
范围(feature)	1.000	0.964	0.982
生理机能 (physiology)	1.000	0.956	0.977
检测(test)	0.722	0.531	0.612
科室(department)	0.700	0.875	0.778
总指标(Micro)	0.750	0.666	0.705

对于 CCKSNER 数据集, 其药物与症状的实体的 F1 值较高, 说明收集的医学词典信息和医学常识的对于这两种实体的覆盖率较高。而疾病和诊断与手术的实体识别的精度较低, 说明目前提供的辅助信息对这两种实体的覆盖率不高, 或者在专业医学用语中, 疾病或者手术实体名称的形式多样化程度较高, “病”或者“术”这样的简单常识特征不易捕获到一些专业用语的实体边界。

对于 CMQANER 数据集, 其各个实体识别的精度分布差距较大, 对于生理机能(physiology)与范围(feature)这样的实体能够有极高的精度, 通过分析测试集与训练集的数据发现, 两者数据的实体分布有着较大重合度, 如都经常出现“出汗”, “白带”, “消化”等生理机能实体, 都经常出现“局部”, “严重”, “剧烈”等范围实体, 即测试集的数据大量存在训练语料中, 降低了识别的难度。

另外, 两个数据集上共同存在的疾病实体, 在 CCKSNER 数据集仅有 65.8% 的 F1 值, 而在 CMQANER 数据集有着 80.9% 的 F1 值, 经过观察, 问答数据上的疾病名称和日常用语较为接近, 在医学词典中的覆盖率较高, 而且用语规律性较强, 大多数是以“病”, “癌”等医学常识的结尾的一致, 而电子病历数据集中一些生僻不规则的疾病较多, 因此识别精度的差距主要来自两个数据集本身疾病实体的分布不一致所致。

而 CMQANER 数据集上治疗方法与药物实体的 F1 值最低, 甚至远远低于 CCKSNER 数据集上药物实体的 F1 值(83.9%)。通过观察数据, 发现问答数据中出现了较多与品牌相关的药物, 如“乌鸡白凤丸”, “复方鲜竹沥液”, “小儿麻甘颗粒”等, 在识别实体边界时, 出现较多错误。而治疗方法实体的准确率较低和用语多样化原因相关, 如存在较多“中医妙方”, “打吊针”, “辨证调理”等一些简单医学常识和常用医学词典难以覆盖的词汇。

3.4 基于 BiLSTM-CRF 的实验设计与分析

3.4.1 模型结构

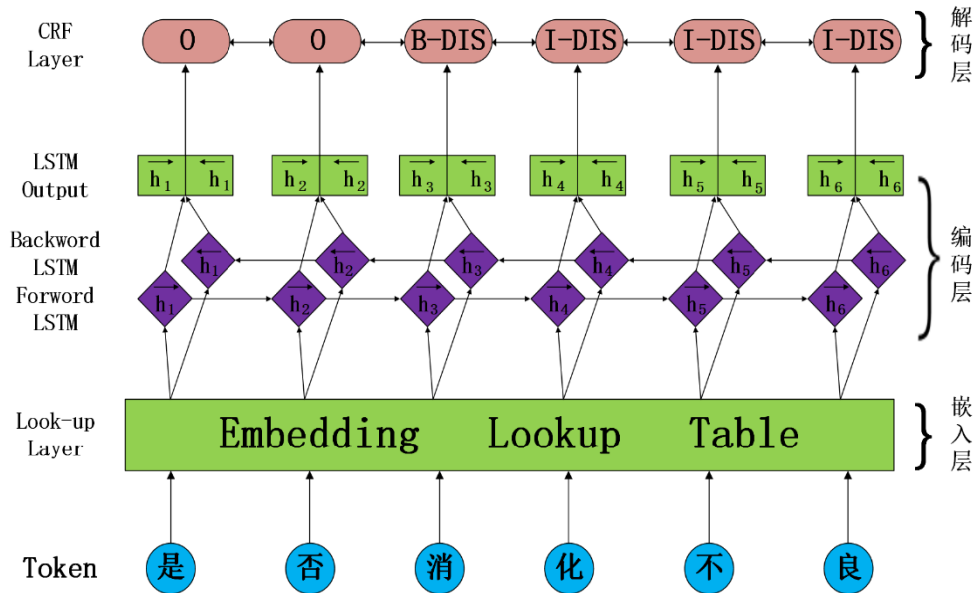
不同于 Linear-Chain-CRF 需要手工设置大量特征模板的方式, BiLSTM-CRF 利用了 BiLSTM 结构作为编码层对上下文进行了自动编码, 去自动习得语义相关的上下文特征, 其具体结构如下。

BiLSTM-CRF 以句子为输入单位, 一个包含 n 个字符的句子可以记做 $S = \{s_1, s_2, \dots, s_n\}$, 如图 3.8 所示, 可以将框架分为输入层 (Input Layer), 编码层

(Encoder Layer) 和解码层 (Decoder Layer)。

第一层是输入层，输入的是分布式表示的字向量或者词向量。常见的输入是可以 skip-gram 或者 glove 模型所训练好的字向量或词向量。

第二层是编码层，这里采用的是 BiLSTM,将句子经过正向 LSTM 得到的各个位置的隐层输出 \vec{h}_i 和经过负向 LSTM 得到的各个位置的隐层输出拼接起来得到



的

图 3.8 BiLSTM-CRF 结构图

$[\vec{h}_i, \vec{h}_i]$ 。有时需要在之后接入一层 dropout 来防止过拟合。如果每个隐藏层输出的是 m 维向量，需要把 m 维向量利用全连接层将其映射为 k 维， k 的个数即实体识别中标签个数。可以得到一个标签矩阵 $P \in R^{n \times k}$ ，标签矩阵中的第 t 列代表将第 t 个时间步的词分类到第 j 个标签的概率。

第三层是解码层，Softmax 函数是一个较为常见的解码层，但在实体识别任务中，每个标注的顺序存在一定的逻辑关系。此时往往会以 CRF 为解码层，通过加入约束保证最终预测的结果必定是符合逻辑的，这些约束会在训练过程中自动习得(CRF 自身模型中包含转移特征，使得它考虑输出类标之间的顺序)。

需要学习一个标签的概率转移矩阵 A_{ij} 表示从第 i 个标注到第 j 个标注的转移概率，考虑到句首和句尾的两个状态，得到 $A_{ij} \in R^{k+2}$ 。在模型输入层得到输入序列 X ，得到标签矩阵 P 的情况下，得到模型若将标签的序列预测为 $y = \{y_1, y_2, \dots, y_n\}$ ，因此对于模型的预测的某一个标签路径可以得到分值：

$$S(X, y) = \sum_{i=1}^{n+1} A_{y_{i-1}, y_i} + \sum_{i=1}^n P_{i, y_i} \quad (3.11)$$

由于模型最后需要得到概率化的形式，则需要使用 softmax 对结果进行归一化处理，如下所示：

$$P(y | X) = \frac{\exp(S(X, y))}{\sum_{y'} \exp(S(X, y'))} \quad (3.12)$$

实际训练过程利用最大似然估计((Maximum Likelihood Estimation, MLE))进行化简：

$$L = \log(P(y | X)) \quad (3.13)$$

由于所有路径的个数分值的计算是指数级的时间复杂度，这里采用维特比算法(Viterbi)进行优化，去计算出最优的标记序列 y^* ：

$$y^* = \arg \max_{y'} S(X, y') \quad (3.14)$$

在进去编码层之前，输入嵌入层的往往是已经训练好的字向量或者词向量。训练数据包括了维基百科的公共领域和利用爬虫工具爬取的医疗领域数据，其训练方法和语料的获取将在下一章医疗字向量和词向量的训练中会进行具体介绍。

3.4.2 模型超参数选择

在以 BiLSTM-CRF 为代表的深度学习模型中，影响实体识别的精确度的因素主要在于超参数的选择。本节对字/词向量长度，优化器的和初始学习率的选择，LSTM 层数，LSTM 单元中隐向量的个数，是否 dropout 等因素进行讨论，后文深度学习模型中选择超参数的方法与本节类似，因此不再赘述。

字/词向量维度：其大小反映了字/词所承载的语义信息的丰富程度。考虑到本文所采用的医疗语料的大小，字/词向量维度不宜设置太大，本次实验采取 50,100,150,200 这 4 个维度进行实验，两份数据集上的实验结果如图 3.9 和图 3.10 所示。

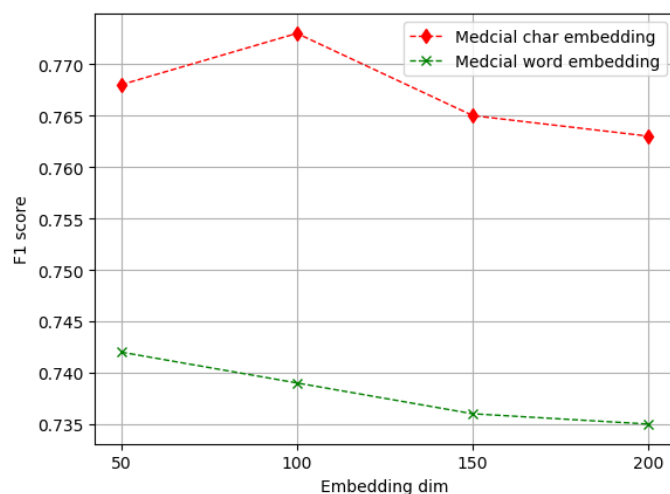


图 3.9 CCKSNER 数据集下字/词向量维度对效果的影响图

CCKSNER 数据集下，词向量随着维度的增加效果降低，而字向量的效果是随着维度先升高后下降，因此本次字向量维度为 100，词向量维度为 50

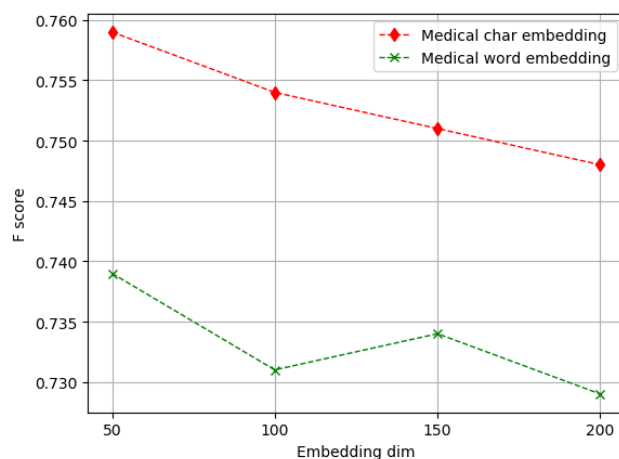


图 3.10 CMQANER 数据集下字/词向量维度对效果的影响图

CMQANER 数据集中，对于词向量，随着维度的增大，模型效果呈现先减少后上升的趋势。而字向量随着维度的上升，模型效果逐渐下降，综合来看字向量维度取 50，词向量维度取 50 模型 F_1 取得最好的效果。

最佳字/词向量的维度与语料的大小有关，从实验结果来看，本次爬取的医学语料随着字/词向量维度的上升而产生相对于语料的冗余信息，反而降低了模型的效果，在后续实验中，两个数据集上向量维度固定为 50 维。

优化器与初始学习速率：优化器的选择将影响神经网络的收敛过程，在实体识别领域 SGD (Stochastic Gradient Descent) 和 Adam 是两种广泛使用的优化算法，SGD 是一种固定学习率的优化算法，后期可能收敛到局部最优解而不是全局最优。而 Adam 结合了 Adagrad 和 RMSprop 两种自适应优化算法的优点，收敛速度较快，但在训练后期可能出现学习率震荡的问题。不存在一个在任何数据上都能表现最好的优化算法，文本使用 SGD 与 Adam 两个优化算法结合学习率进行模型的比较，所得到模型效果如图 3.11，图 3.12，对于两个数据集分别选择最优的优化器与初始学习速率，从结果来看 Adam 优化器与 0.001 的初始学习率对于两个医学数据集均取得了最好的效果。

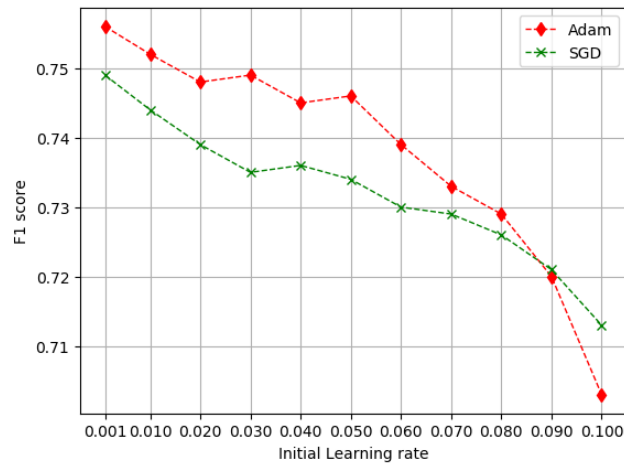


图 3.11 CCKSNER 数据集下优化器和学习速率对效果的影响图

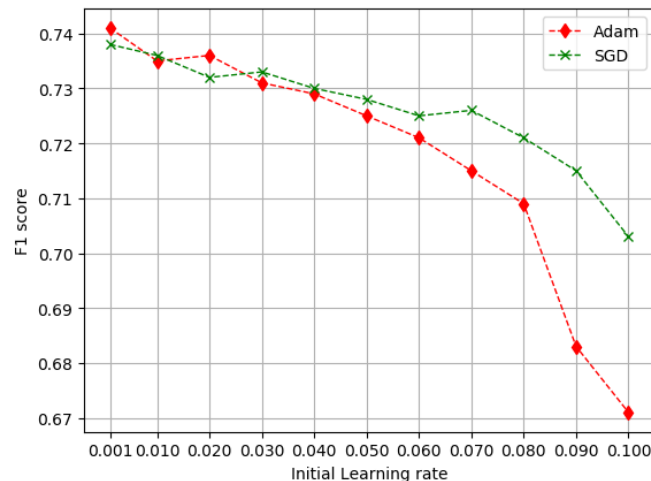


图 3.12 CMQANER 数据集下优化器和学习速率对效果的影响图

LSTM 层数：理论上来说 LSTM 层数越高，模型可以提取出更丰富的语义信息。但 LSTM 层数越多意味着参数量的上升，如果训练数据的规模不能满足

模型的要求便很容易发生过拟合的现象。本实验的基础编码单元是 BiLSTM,因此是前向 LSTM 与后向的 LSTM 层数同时增加,结果如图 3.13 所示,发现随着 LSTM 层数的上升,在两个数据集上的模型效果均发生不同程度的下降,因此 LSTM 层数选择为 1。

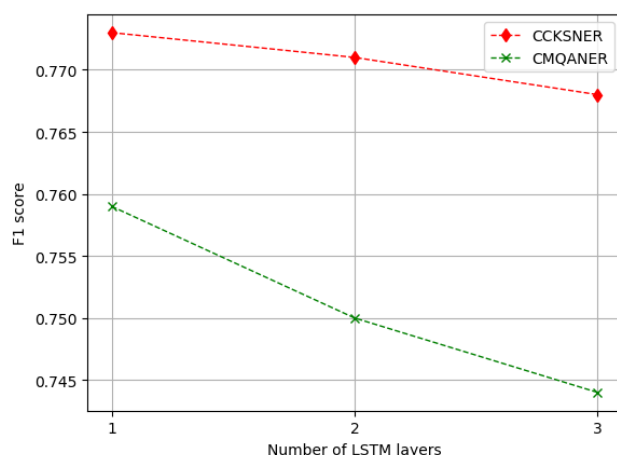


图 3.13 LSTM 层数对效果的影响图

LSTM 隐层节点个数: LSTM 隐层节点可以存储大量上文语义信息,理论上隐层节点个数越多,模型拟合能力越强,但与之同时参数规模也随之增大,提高了过拟合的风险,如图 3.14 所示,随着 LSTM 隐层节点个数的个数增加,两个数据集上的效果均呈现先增加后减少的趋势,说明隐层单元数超过 200 左右开始有过拟合的现象,因此在两个数据上均取 200 的隐层单元数。

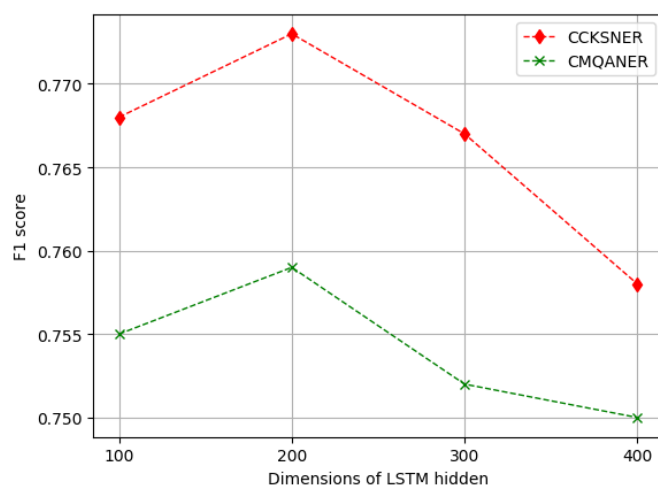


图 3.14 隐藏层维度对于模型的 F1 值影响图

Dropout 比例: Dropout 是一种防止过拟合的有效手段,在模型训练的过程中随机使得网络的一定比例的节点失效,则在做测试的时候,使得所有节点都有效。

从集成学习的观点看，每次的节点失效相当于训练了一个子网络，测试阶段所有节点进行预测，可以理解为多个子网络的集成学习。根据本节训练数据的规模，如果会使得失效的节点比例过多，会使得训练效果较差，因此选择了 0.1, 0.2, 0.3 的 Dropout 比例进行实验。本文在两处加入 Dropout 层，一处是嵌入层的输出之后，另一处是 BiLSTM 编码层输出之后。对于 CCKSNER 数据选择了 0.2 的

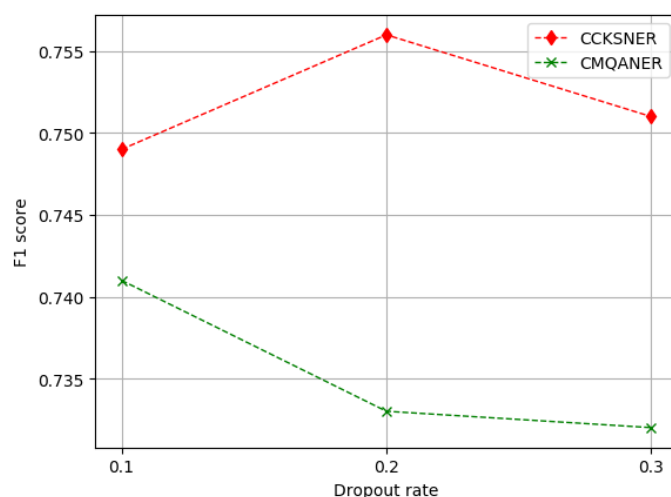


图 3.15 Dropout rate 对于模型的 F1 值影响图

Dropout 比例，对于 CMQANER 数据集，使用 0.1 的 Dropout 比例。CCKSNER 数据集超参数调优表 3.9 所示。

表 3.9 CCKSNER 数据集在 BiLSTM-CRF 模型的参数选择表

参数	取值	参数含义
Char_emb_size	100	字符级词嵌入层维数
Char_dropout	0.2	嵌入层随即丢弃概率
LSTM_layer	1	LSTM 层数
Learning_rate	0.001	学习率
LSTM_hidden	200	隐藏层维数
Weight_decay	0.0001	权重衰减
Optimizers	Adam	优化器
Batch_size	64	每次训练输入的句子数

3.4.3 结果分析

两个数据集使用医疗语料的得到的字向量输入 BiLSTM-CRF 模型，其中 CCKSNER 数据集在参数调优后结果与如表 3.10 所示，对比表中 Linear-Chain-CRF 模型的识别结果可以得出结论：深度学习模型基于稠密向量表示的输入，更好的表征了医学上下文，基于深度学习模型进行超参数调优的得到微平均指标结果在两个数据集上都优于统计学习模型引入特征模板方法得到的微平均指标。

表 3.10 Linear-Chain-CRF 与 BiLSTM-CRF 模型效果对比表

数据集	Linear-Chain-CRF			BiLSTM-CRF		
	Precision	Recall	F1 值	Precision	Recall	F1 值
CCKSNER	0.733	0.726	0.729	0.776	0.744	0.759
CMQANER	0.75	0.666	0.705	0.781	0.701	0.741

表 3.11 CCKSNER 数据集在 BiLSTM-CRF 模型下的各个实体识别效果对比表

测试指标	Precision	Recall	F1 值	F1 值 (CRF)
解剖部位	0.721	0.662	0.690	0.696
影像检查	0.826	0.679	0.745	0.777
实验室检验	0.593	0.667	0.628	0.785
疾病和诊断	0.723	0.759	0.741	0.658
药物	0.952	0.922	0.937	0.839
手术	0.675	0.659	0.667	0.621
症状	1.000	0.889	0.941	0.860
总指标(Micro)	0.776	0.744	0.759	0.729

对于 CCKSNER 数据集，尽管在 Micro-F1 指标上 BiLSTM-CRF 模型超过可以 Linear-Chain-CRF 模型的结果，但是在“解剖部位”，“影像检查”和“实验室检验”这三类实体上，CRF 的模型取得了更好的效果，尤其是“实验室检验”

这类实体, 其 F1 值达到了 78.5%, 远远超过了深度学习模型 62.8% 指标。经分析发现, “实验室检验” 实体中存在较多由特殊符号组成的词, 有些可以在词典中发现, 如 “ast”, “ca199” 等。

表 3.12 CMQANER 数据集在 BiLSTM-CRF 模型下的各个实体识别效果对比表

测试指标	Precision	Recall	F1 值	F1 值 (CRF)
疾病(disease)	0.784	0.784	0.784	0.809
人群(crowd)	0.877	0.821	0.848	0.900
症状(symptom)	0.797	0.738	0.766	0.769
身体部位(body)	0.705	0.684	0.694	0.708
治疗方法 (treatment)	0.755	0.531	0.623	0.555
时间(time)	0.765	0.406	0.531	0.692
药物(drug)	0.694	0.410	0.515	0.593
范围(feature)	1.000	0.964	0.982	0.982
生理机能 (physiology)	0.947	0.800	0.868	0.977
检测(test)	0.718	0.571	0.636	0.612
科室(department)	0.800	1.000	0.889	0.778
总指标(Micro)	0.781	0.706	0.741	0.705

对于 CMQANER 数据集, 尽管在 Micro-F1 指标上 BiLSTM-CRF 模型存在较大优势, 但在 “人群”, “症状”, “身体部位”, “时间”, “生理机能” 这些单个实体上的效果都没有超过融入医学词典特征与医学知识特征的特征模板的效果, 如 “生理机能” 实体, 在深度学习模型上的 F1 值为 86.8%, 而在统计学习模型中可以达到 97.7%。经过观察发现, 收集的医学词典对 “生理机能” 类实体有较高的覆盖率, 证明了这些辅助信息可以在 CRF 模型对上下文表征能力不足的条件 下提高模型的效果。

3.5 本章小结

本章首先描述了处理命名实体识别问题的在统计学习方法和深度学习方法上的处理流程，对于明确了在文本的实验中，测评指标选用采用严格指标，对于多类实体的综合指标选取微观指标。并分别利用基于 **Linear-Chain-CRF** 模型进行添加特征模板的实验，基于 **BiLSTM-CRF** 模型进行超参数调优的实验，并通过微平均指标的对比得出 **BiLSTM-CRF** 模型综合的模型效果在两个数据集上都达到了超越 **Linear-Chain-CRF** 模型的效果。但在某些单个医学实体识别上，添加了医学词典特征与医学知识特征的 **Linear-Chain-CRF** 模型可以得到超越神经网络进行参数学习的效果，经过分析证明了医学词典信息与医学知识信息的重要性，为下面两个章节进行进一步融入医学词典信息与融入领域知识信息的实验探索提供了前提理由与铺垫。

第四章 融入医学词典与语料信息

4.1 词典信息的获取与使用

4.1.1 医学词典与语料的获取

词典作为一种重要的辅助信息，其质量直接影响到了分词的效果，在基于词粒度作为输入的实体识别模型中，可以直接影响模型最后的精度。而医学行业专业术语数量庞大，但现有的词典主要是通用词典，很难涵盖众多专业术语。为了提高医学分词及其下游任务的精度，本文从主要从两方面获取词典，一是从搜狗输入法医疗版中 22 种门类下爬取 40 万余医疗词条，具体门类如图 4.1 所示。

基础医学(26)	西药学(30)	中医(66)	中药(50)	针灸(1)	疾病(13)	超声医学(6)
耳鼻喉科(2)	法医学(2)	护理学(2)	解剖学(4)	口腔医学(7)	美容外科(4)	皮肤科(6)
兽医(3)	医疗器械(14)	医学影像学(4)	肿瘤形态学(1)	医学检验(1)	医疗(19)	外科(6)
其它(37)						

图 4.1 词典来源门类图

另一方面也使用了北京大学北大开放数据平台所开源的一份医学分词词典[40]，其包括了疾病、症状、检查、药物、手术 5 种医学实体类别，其原始来源包括了权威词表、官方网站、大众健康网站。

对收集的词典进行如下步骤预处理：

- 1)为医学词语标注类别和数据来源
- 2)删除圆括号的补充信息，删除英文别名
- 3)将词语方括号中的补充同义词替换成新词
- 4)过滤出字符长度大于 15 的词语进行人工审核或

"纵火行为"	"Disease"	"MeSH"
"足底"	"Disease"	"MeSH"
"足畸形"	"Disease"	"MeSH"
"足疾病"	"Disease"	"MeSH"
"阻塞性"	"Disease"	"MeSH"
"阻生"	"Disease"	"MeSH"
"组织胞浆菌病"	"Disease"	"MeSH"
"组织细胞病"	"Disease"	"MeSH"
"左右转位"	"Disease"	"MeSH"
"阿片制剂药物反应"	"Disease"	"搜狗-ICD10疾病编码"
"阿斯匹林药物反应"	"Disease"	"搜狗-ICD10疾病编码"
"阿斯匹林中毒"	"Disease"	"搜狗-ICD10疾病编码"
"阿托品的意外中毒"	"Disease"	"搜狗-ICD10疾病编码"
"阿托品药物反应"	"Disease"	"搜狗-ICD10疾病编码"
"癌病"	"Disease"	"搜狗-ICD10疾病编码"
"癌前色素沉着病"	"Disease"	"搜狗-ICD10疾病编码"
"癌前色素沉着病内的恶性黑瘤"	"Disease"	"搜狗-ICD10疾病编码"
"安眠酮药物反应"	"Disease"	"搜狗-ICD10疾病编码"
"安眠药药物反应"	"Disease"	"搜狗-ICD10疾病编码"
"氨卞青霉素药物反应"	"Disease"	"搜狗-ICD10疾病编码"
"氨基苯酚衍生物反应"	"Disease"	"搜狗-ICD10疾病编码"

图 4.2 医学词典图

最后得到的词典形式如图 4.2 所示，第一列为医学名词，第二列为实体类型，第三列为获取来源。由于在有的获取来源中并不包含实体类型标注，因此部分第二列为空。最后得到的医学词典共包含 534983 个单词。

除了医疗词典，本文也爬取了大众健康网站如 39 医学教育网中的社区问答，精华病例等线上语料与《中国实用医药》，《中国兽医杂志》，《中华现代儿科学杂志》等权威期刊杂志中的文本，得到了约 0.8G 的医学训练语料，如图 4.3 所示。由于语料风格中既包括了大众问答偏口语化的语料，也包括了权威杂志期刊等包含大量专业用语的语料，可以同时作为 CCKSNER 和 CQANER 的辅助语料信息。

原发性肾上腺皮质功能减退症又称阿狄森氏病，是由肾上腺皮质本身的病变所致，1855年首先由英国医生Addison氏所描述。其主要病因是结核、癌瘤及特发性萎缩，在我国及日本主要是结核造成的肾上腺组织破坏，约占全部病例的68%。近年来由于结核感染之病例减少，本病的发生也相应下降而较罕见。基于肾上腺储备能力很大，一般须待肾上腺组织破坏达到80%~90%以上时，临床才出现明显的肾上腺皮质功能低下的症状。现代西医尚缺乏特效疗法。代谢性酸中毒是最常见的一种酸碱平衡紊乱，是细胞外液H⁺增加或HCO₃⁻丢失而引起的以原发性HCO₃⁻降低 (<17.35mmol/L) 和PH值降低 (<7.35) 为特征。在代谢性酸中毒的临床判断中，阴离子间隙 (AG) 有重要的临床价值。按不同的AG值可分为高AG正常型及正常AG高型代谢性酸中毒。血中尿素、肌酐、尿酸等非蛋白氮 (NPN) 含量显著升高，称氮质血症 (azotemia)。正常人血中NPN为25~35mg%，其中尿素氮为10~15mg%，氮质血症是一个生化名词，有广义和狭义的两方面概念。广义的概念是只要血中的尿素氮或肌酐等非蛋白氮超出正常范围，均可称为氮质血症。各种肾脏病迁延不愈，晚期可发生肾功能损害，这样血中氮质排泄障碍，遂蓄积于血液中，这是肾衰的结果。但是正常人在一个较短的时间里大量进食高蛋白食物，如过年过节或平时参加宴会过多，虽然肾功能正常，但短时间内不能迅速地排出过多的氮质，则会出现一过性的氮质血症。大便失禁即肛门失禁是指粪便及气体不能随意控制，不自主地流出肛门外，为排便功能紊乱的一种症状，亦称大便失禁。肛门失禁的发病率不高，但非罕见。虽不直接威胁生命，但造成病人身体和精神上的痛苦，严重地干扰正常生活和工作。

图 4.3 辅助医疗信息样例图

4.1.2 字粒度中融入词典信息

目前使用字典已经被证明在多个基于词粒度的命名实体识别的任务中效果显著[41]，但由于医学领域目前不存在某一字典能够包含医学词汇，医学领域的中文分词的精度还有较大提升空间，字粒度的输入相对于词粒度在医学实体识别的

任务中目前的应用相对更多。词典除了直接用于词粒度的分词，对于字粒度的输入是否也能起到一定的辅助作用？本节将从两种输入粒度分别讨论医学词典辅助信息的作用。

上下文词典特征：在多种分词主流的分词工具中，N-gram 的特征往往会作为分词算法的重要理论基础之一。本文将 N-gram 的特征结合词典信息命名为上下文词典特征，该特征将被作为特征模板之一应用于 CRF 统计学习模型以及作为字向量的辅助信息融入 BiLSTM-CRF 的深度学习模型中。

N-gram 信息的提取过程是利用一个长度为 N 的滑动窗口去获取当前字符的上下文信息的过程，对于每一个字符去判断上下文是否存在于词典集合中。N 的范围如果太大则会增加计算复杂度，造成一定的信息冗余；而 N 的范围如果太小则会减小辅助信息量，降低下游任务的模型精度，在 Wang 等人工作的基础上 [42]，本文使用了 5 种 N-gram 的特征，如表 4.1 所示。

表 4.1 N-gram 特征模板表

类型	上下文词典特征模板
2-gram	$w_{i-1}w_i, w_iw_{i+1}$
3-gram	$w_{i-2}w_{i-1}w_i, w_iw_{i+1}w_{i+2}$
4-gram	$w_{i-3}w_{i-2}w_{i-1}w_i, w_iw_{i+1}w_{i+2}w_{i+3}$
5-gram	$w_{i-4}w_{i-3}w_{i-2}w_{i-1}w_i, w_iw_{i+1}w_{i+2}w_{i+3}w_{i+4}$

在统计学习模型 CRF 中，这 8 个特征模板可以直接使用，将在后文的 CRF 实验部分进行模型对比。而在深度学习的模型中，常规方法是将其转换成向量的形式去融入字向量。

本文中涉及到两个数据集，其实体类型与个数并不相同。CCKSNER 数据集包含了 7 类实体，每一个上下文特征模板可以用一个 8 维的 one-hot 向量表示，如果在医疗词典中，其标注的类型是 7 种实体对应的类型之一，则相应的位置会置

1, 如果不属于这 7 种类型, 则最后一维将置为 1, 如果上下文对应的词不在词典中则直接置成全 0 向量, 如表所示。由于总共用到 8 个特征模板, 每个字符总的上下文词典特征为 64 维。

表 4.2 词典特征向量表

查找到的词典类型	上下文词典特征向量
疾病和诊断	[1,0,0,0,0,0,0,0]
影像检查	[0,1,0,0,0,0,0,0]
实验室检验	[0,0,1,0,0,0,0,0]
手术	[0,0,0,1,0,0,0,0]
药物	[0,0,0,0,1,0,0,0]
解剖部位	[0,0,0,0,0,1,0,0]
症状	[0,0,0,0,0,0,1,0]
其他类型	[0,0,0,0,0,0,0,1]
未在词典中找到	[0,0,0,0,0,0,0,0]

当输入序列为“是否有消化不良的可能。”, 其对应一个长度为 12 的字符串序列 $\{w_1, w_2 \dots w_{11}, w_{12}\}$, 经过语料训练会得到字向量矩阵 $X = \{x_1, x_2 \dots x_{11}, x_{12}\}$, 融入上下文词典特征矩阵 $C = \{c_1, c_2 \dots c_{11}, c_{12}\}$, 在 BiLSTM-CRF 的输入层中直接拼接可以得到任意第 i 个字符的输入为:

$$z_i = x_i \oplus c_i \quad (4.1)$$

此时模型输入就成为一个维度为 $(N+64) \times 12$ 的字向量矩阵, N 为训练好的字向量维度。同理对于包含 11 种实体的 CQANER 数据集, 模型输入则变成 $(N+96) \times 12$ 字向量矩阵, 特征补充的方法是一致的, 因此过程不再赘述。

词界特征: 本文构造的上下文信息, 主要提供给模型的信息是当前的字符上

下文属于词典中的哪一个词和与其相对应的类别。分词之后的结果对于每个字符会产生词边界信息。

如表 4.3 所示，由于“消化不良”实体出现在了医学词典中，在医学词典中记录了其对应的实体信息，并且是数据集中的 11 种实体类别之一，于是将“消化不良”词首和非词首的字符分别标记为 B-DIS 和 I-DIS，DIS 是英文单词疾病 Disease 的简写。而“是否是”，和“可能”不在医学词典中出现。对于不在医学词典中出现的分词结果，或者医学词典中没有标记类型的词，又或者医学词典中标记的类型，但这个类型不在数据集的 11 种实体类别中，则不能直接判断为非医学实体，需要将信息输入模型结合上下文进行判断。因此将词首和非词首的部分标记为 B-O 和 I-O。尽管没有区分词中心位置和词尾位置，词尾特征作为序列输入的简单规律容易被 BiLSTM 这样的序列模型所学到，同时减少了分词特征的种类数，使得字向量不会过于稀疏，便于模型的训练。

表 4.3 分词结果表

句子	是	否	是	有	消	化	不	良	的	可	能	。
分词	是否是			有	消化不良				的	可能		。
词界	B-O	I-O	I-O	O	B-	I-	I-	I-	B-O	I-O	I-O	O
标记					DIS	DI	DI	DI				
						S	S	S				

词界特征同样以类似于上下文词典特征，同样以进行表示 one-hot 向量的表示，例句中“是否有消化不良的可能”来自 CMQANER 数据集，其包含了 11 种实体，增加一种其他实体的标记，每个字符有实体词的首位和不在实体词的首位两种可能，因此将词界特征表示成 24 维的 one-hot 向量。同理 CCKSNER 数据集包含了 7 种实体，将其词界特征表示成 16 位的 one-hot 向量。

当输入序列为“是否有消化不良的可能”，其对应一个长度为 12 的字符串序列 $\{w_1, w_2 \dots w_{11}, w_{12}\}$ ，经过语料训练会得到字向量矩阵 $X = \{x_1, x_2 \dots x_{11}, x_{12}\}$ ，融入词界特征矩阵 $C = \{b_1, b_2 \dots b_{11}, b_{12}\}$ ，在 BiLSTM-CRF 的输入层中直接拼接可以得

到任意第*i*个字符的输入为:

$$z_i = x_i \oplus b_i \quad (4.2)$$

4.1.3 词粒度中融入词典信息

在 4.1.2 上下文词典特征融入字粒度的过程中,只要在上下文中出现在词典中就会被引入模型,难免会引入误差。另外使用字作为输入粒度,尽管在融入词界特征之后,会减少词边界信息的损失。词粒度的输入在分词效果较好的情况下可以携带字粒度的输入所不具备的语义信息,在模型中既可以作为单独的输入,也可以结合字粒度的输入进行信息补充。

分词: 分词是自然语言数据预处理的一步,将字符序列组成的句子以一定的规则去重新组合词的集合。在中文分词领域开源了许多分词工具,如 jieba, SnowNLP, THULAC 等。jieba 中文分词工具是其中广泛使用且分词效果较好的一款,且支持在分词过程中加入医疗词典,其分词算法包括了三个部分。

1. 基于字典树结构实现高效词图扫描,并基于医学词典生成句子中汉字所有可能出现的词的组合的有向无环图(DAG)。

2. 动态规划查找最大概率路径,基于词频的概率最大的切分组合

3. 对于未登录词,采用了基于汉字成词的 HMM 模型,使用了 Viterbi 算法去计算最优状态

为了进行对比,本文分别进行了使用医学词典分词 和 不使用医学词典分词的时候,其转换成便于训练模型格式效果如图 4.4、4.5 所

```

除 O
头痛 S_symptom
外 O
鼻窦炎 S_disease
的 O
典型 O
的 O
急性 B_disease
鼻窦炎 I_disease
表现 O
还 O
包括 O
: O
鼻塞 S_symptom
, O
流脓涕 S_symptom
, O
暂时 O
性 O
嗅觉 B_disease
障碍 I_disease
, O
畏 B_symptom
寒 I_symptom
、 O
发热 S_symptom
、 O
食欲 B_symptom

```

图 4.4 不使用医学词典的 jieba 分词图

```

除 O
头痛 S_symptom
外 O
鼻窦炎 S_disease
的 O
典型 O
的 O
急性鼻窦炎 S_disease
表现 O
还 O
包括 O
: O
鼻塞 S_symptom
, O
流脓涕 S_symptom
, O
暂时 O
性 O
嗅觉障碍 S_disease
, O
畏寒 S_symptom
、 O
发热 S_symptom
、 O
食欲 B_symptom
不振 I_symptom
、 O

```

图 4.5 使用医学词典的 jieba 分词图

可以发现，使用了医学词典的 jieba 分词工具将例句中的“急性鼻窦炎”，“嗅觉障碍”，“畏寒”分成了一个独立的词，而未使用医学词典的 jieba 分词工具将这三个医学术语分成了“急性”，“鼻窦炎”，“嗅觉”，“障碍”，“畏”，“寒”这样的独立的词，在一定程度上丢失了部分词边界信息和语义信息。从另

一种角度来说,使用医学词典的 jieba 分词结果已经融入了医学词典信息。

4.1.4 字向量与词向量的训练

Mikolov 等人在 2013 年同时提出了连续词袋(CBOW, Continuous Bag of-Words)模型和跳字(Skip-gram)模型[44],其结构如图 4.6 所示。

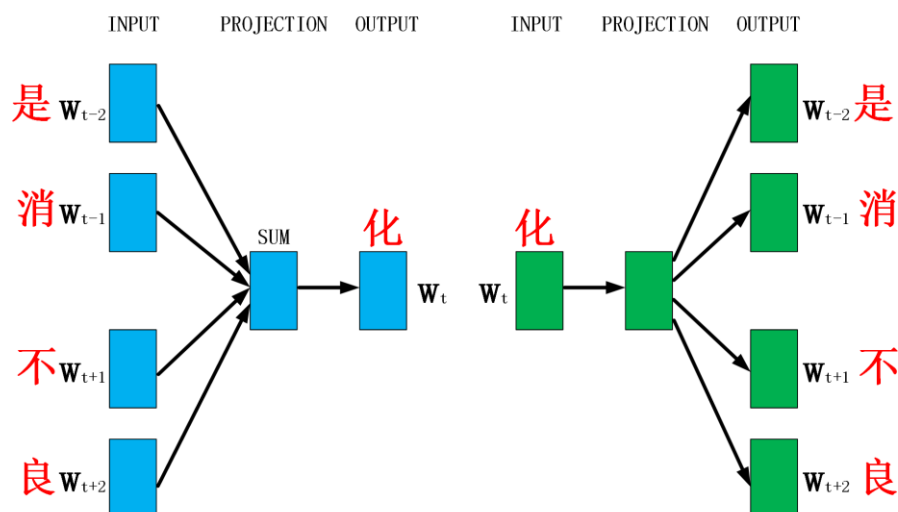


图 4.6 CBOW 与 Skip-gram 模型图

CBOW 是在中心词的上下文已知的前提下,对于中心词进行预测,训练的目标函数是使得 $P(w|\text{context}(w))$ 最大化;而 Skip-gram 是在中心词已知的前提下预测该中心词的上下文,训练的目标函数使得 $P(\text{context}(w)|w)$ 最大化。上下文采用滑动窗口的方式去定义,即只选取中心词窗口范围内的单词为上下文。由于 CBOW 模型在计算时去掉了隐藏层(Hidden Layer),对于输入层的向量进行了直接求和(工程代码中使用了加权平均表示上下文),这使得训练速度变的更快。

至于 CBOW 模型和 Skip-gram 模型训练速度的差别,是由于 CBOW 的输入是每个单词的全体上下文的表示,输出是中心词的表示,而 Skip-gram 模型输入是中心词,输出是上下文的某一个。因此 CBOW 模型的训练次数是和文本的词数成正比,而 Skip-gram 模型训练的次数是和文本的词数与两倍滑动窗口大小的乘积成正比,恰恰是因为 Skip-gram 模型的训练次数较多,往往 Skip-gram 得到的训练结果更为精准,本文进行字向量与词向量的训练采用的是 Skip-gram 模型。

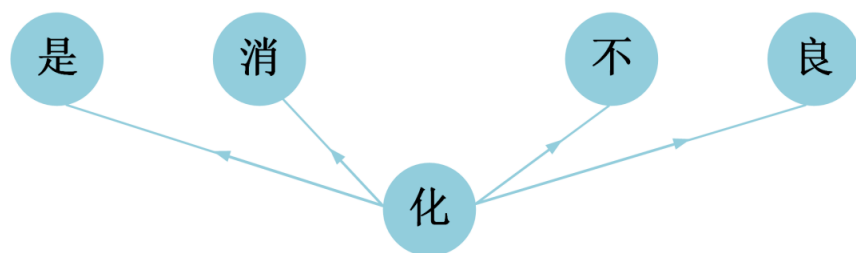


图 4.7 Skip-gram 模型样例图

词向量训练的过程区别仅仅在于切分粒度的不同，因此本文以字向量为例分析训练过程。Skip-gram 模型假设当中心词既定的情况下，生成每个上下文单词的概率相互独立，设置 skip_window 参数为 2，如图 4.7 所示，给定中心词“化”生成上下文“是”，“消”，“不”，“良”的概率如下：

$$P("是","消","不","良"|"化")=P("是"|"化")\times P("消"|"化")\times P("不"|"化")\times P("良"|"化")$$

在 Skip-gram 模型中，每个字均承担两种不同的角色，中心词和上下文的单词，因此每个字分别使用两种 d 维的向量进行表示，假如字的集合长度为 N ，其字索引的取值为 0 到 $N-1$ ，则第 i 个字作为中心词时对应的字向量为 $v_i \in R^d$ ，第 i 个字作为上下文词时对应的字向量为 $u_i \in R^d$ 。若在词典中索引为 i 的词 w_i 为中心词，索引为 c 的词 w_c 为上下文词，则在中心词出现后上下文词出现的概率如公式 4.3 所示。

$$P(w_c | w_i) = \frac{\exp(u_c \cdot v_i)}{\sum_{j \in V} \exp(u_j \cdot v_i)} \quad (4.3)$$

转化成便于优化的对数损失的形式，如式 4.4 所示。

$$\log P(w_c | w_i) = u_c \cdot v_i - \log(\sum_{j \in V} \exp(u_j \cdot v_i)) \quad (4.4)$$

实际的计算的过程中，对于输入的每一个中心词与上下文词组成的字对，式子的后半部分需要对整个字的集合进行计算。因此在工程上往往会使用负采样或者 Hierarchical Softmax 的技术进行算法优化。

实验部分利用 gensim 框架实现，字向量和词向量生成的模型输入分别为经过空格分割的字与词，如图 4.8 与图 4.9 所示。

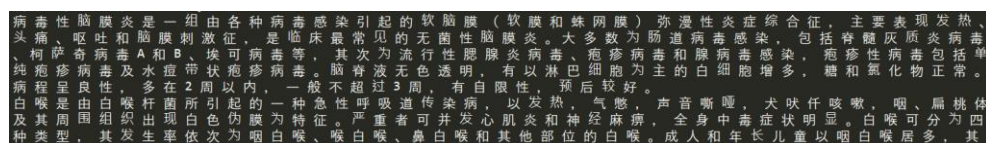


图 4.8 字粒度分割的医学语料图

动脉瘤是动脉病变或损伤所造成局限性动脉节段的持久性扩张。常见于主动脉和下肢主干动脉。分为由动脉全层构成的真性动脉瘤和由动脉周围结缔组织构成的假性动脉瘤两类。具有动脉壁病变或动脉损伤史。局部搏动性肿物是典型表现，搏动与心跳一致，可伴收缩期震颤可有疼痛和压痛。压迫周围组织出现相应症状。CT、B超确定肿物存在，动脉造影可以确诊。肺动脉高压是一种极度严重的疾病。75%患者集中于20~40岁年龄段，15%患者年龄在20岁以下。肺动脉高压的症状包括呼吸短促、易于疲劳、晕厥、胸痛以及腿部和踝部水肿。此外，心脏听诊可闻P2亢进。如果不及时治疗，患者的肺动脉高压会逐步加重，甚至使寿命缩短。多数肺动脉高压相关的症状源自右心衰竭。在上世纪90年代以前，医学界对这种疾病确实缺少治疗手段。但此后一些新的药物陆续被研发出来，患者5年或10年平均生存率可提高数倍。药物之外，近几年基因治疗、活体肺移植、房间隔造瘘等新疗法也不断出现，现已有多种治疗手段。

图 4.9 词粒度分割的医学语料图

徽 1.3077921 -4.1113024 2.7438204 2.999122 -7.913698 4.113898 -1.3870864 -7.81228 3
3.4599385 -1.3887235 -1.575589 -8.86941 -1.4861044 -0.5442215 -7.0680737 -0.9092146
-5.542662 4.8670983 -4.081442 0.81990135 -2.0287628 6.035518 -4.177359 -1.0640726
-1.1341127 2.3243783 1.7625751 -4.066773 -5.9427495 -0.3415615 2.6857667 0.6169147
朋 -1.0110643 1.2862842 1.8171657 -2.0908573 -3.812488 2.093698 -0.015400407 7.9309
1.2657714 3.5931756 -2.2239165 8.640834 -3.8673503 0.13390304 0.6623845 3.0017586
-1.6276894 -4.0949636 -2.1715405 -4.871446 7.104668 -5.3421264 -1.4615512 2.3377814
-3.2550766 2.0755832 -2.1474452 -6.536591 3.3468833 -3.354196 2.141773 2.072265 -3
癌 9.8040495 -5.258762 -1.8288074 4.159516 4.4290237 1.6329651 3.9146059 -0.0938534
4.6634455 1.3300039 5.8372626 -4.3983974 -0.31681797 -2.1213017 3.2905643 2.9737306
-4.2065783 -3.4091628 -1.6444683 0.48081833 -3.0361652 0.46164665 4.29324 -4.776898
5.3049417 -0.7215297 2.2397308 0.11068427 0.19125064 -1.177572 2.952148 -4.6549153
隶 -6.657692 -3.426576 -0.28249878 5.410695 -0.8603508 9.433398 -2.7251923 -5.34819
1.4135962 3.3394237 8.018143 -7.1258273 -1.0380712 -4.0650635 -0.3783661 3.932507
-13.528329 4.402171 4.4435153 -0.3295521 -2.358273 0.8386803 -1.2012386 0.4739641
-3.0712006 -2.3499904 -6.490969 0.8964844 -12.859635 -2.6711192 1.443105 3.2977161

图 4.10 所训练的医学字向量图

脾虚泻泄 -0.010446816 -0.026690707 -0.20333777 -0.17256035 0.20136419 -0.6
-0.23110028 0.09872324 -0.071501374 0.028918745 0.29413265 0.014395309 -0
0.028147874 0.07184924 0.002620896 -0.12758592 -0.028466638 -0.050261497
-0.08575902 0.13343793 0.034634367 0.06677306 -0.07466548 -0.3688744 0.05
-0.020951742 -0.28747687 -0.06507366 -0.14004672 0.13926992
益气 -0.079051025 -0.048334986 0.0680419 -0.02243474 0.029003073 0.1022774
-0.010673002 -0.06811166 0.15816514 0.1798641 -0.26302272 -0.20307618 -0.
-0.025106609 -0.09045726 -0.0031709117 0.082762934 -0.1731082 0.14830463
-0.17709704 0.0009543783 -0.0999136 -0.20413871 0.08325931 -0.2650377 -0.
-0.18990177 -0.0060033686 0.3177356 0.00011496512 0.015764078 -0.03408190
五淋丸 0.019312922 -0.16361052 -0.1968306 -0.17417087 0.14628893 -0.052559
0.024481809 -0.035558954 -0.030114776 0.22788899 -0.070439205 0.06633427
0.04689671 0.025854612 -0.03750888 -0.054827686 -0.08911452 -0.017287556
0.18884504 0.20891973 -0.039099436 -0.05023534 -0.036846843 -0.09996397 -
0.08629096 0.11130362 -0.21135588 -0.21833007 -0.17207421 0.15293781
五苓片 0.03964508 -0.04977902 0.075364016 -0.1844771 0.08263598 -0.0225282
-0.055272553 -0.043328352 0.011296414 -0.05535392 -0.06635794 0.087761305
0.0028218557 -0.106996395 -0.12662031 -0.00064369844 -0.08353507 0.055632
0.055223014 0.12149137 -0.15375748 -0.019182261 -0.23377445 -0.11165377 -
0.119731694 0.062436424 0.010211062 0.10435587 0.022142483 0.008317471
五妙水仙膏 -0.0015954825 -0.0391801 -0.16426817 0.11616918 -0.03506986 0.1
-0.047265273 0.005515723 -0.06344248 -0.17629272 0.08124845 -0.13649358 -
-0.101157136 0.07465488 0.092203565 -0.06708053 -0.0734034 0.15342933 -0.
-0.04639552 0.0012191507 -0.061290827 0.09732657 -0.057397164 -0.12816384
0.062824875 0.026319735 0.035046875 0.057607036 0.07795921 0.12120454
参芪膏 0.04583827 -0.14111555 -0.2236239 -0.1001315 -0.07329026 0.06321966

图 4.11 所训练的医学词向量图

基于 Skip-gram 模型通过医学语料训练得到的医学字向量与医学词向量如图 4.10 与 4.11 所示。

4.2 字词融合的深度学习模型

4.2.1 单粒度增强实验

为了在字粒度上比较引入医疗语料与医疗词典对模型结果的影响，对于两份数据集以经典的 BiLSTM-CRF 框架，分别采用基于大众语料训练得到的字向量，

基于医疗语料训练得到的字向量，与基于医疗语料训练得到的字向量拼接上词典上下文特征和词界特征得到的新的字向量作为模型嵌入层的输入。得到的结果如表 4.4 和表 4.5 所示。

对于 CCKSNER 数据集，可以发现在使用医疗语料去替换大众语料进行字向量的训练时，模型的 F1 值从 0.746 提升到了 0.759，说明电子病历的文本上下文信息与大众文本的上下文信息存在较大区别。当加入了上下文词典特征与词界特征到字向量之后，模型的 F1 值提高到了 0.773，说明词典信息融入神经网络之后仍旧可以起到一个较强的提升效果，且上下文词典特征和词界特征的形式是一种较好的方式去描述上下文中是否有词典中的词出现，以及词典中的词的边界信息。

对于 CMQANER 数据集，当加入医疗语料时，其提升幅度不如 CCKSNER 数据集，说明其线上问答的用语风格与大众文本的用语风格有一定的相近之处，大众文本的用语风格在一定程度上可以表示线上问答数据集的分布情况。而当加入了上下文词典特征与词界特征之后，其 F1 值从 0.741 飙升到了 0.759，说明本次收集的医学词典对于线上问答数据集相较于电子病历可能有更大的覆盖率。

表 4.4 BiLSTM-CRF 模型下不同字向量输入对 CCKSNER 数据集识别效果比较表

模型输入	Precision	Recall	F1 值
基于大众语料训练的字向量输入	0.761	0.733	0.746
基于医疗语料训练的字向量输入	0.776	0.744	0.759
基于医疗语料训练的字向量输入+上下文 词典特征+词界特征	0.786	0.761	0.773

表 4.5 BiLSTM-CRF 模型不同字向量输入对 CMQANER 数据集识别效果比较表

模型输入	Precision	Recall	F1 值
基于大众语料训练的字向量输入	0.772	0.697	0.733
基于医疗语料训练的字向量输入	0.781	0.701	0.741
基于医疗语料训练的字向量输入+上下文 词典特征+词界特征	0.796	0.725	0.759

基于医学词典所支撑的分词结果，除了作为字粒度的输入的一个补充，也能够单独进行基于词粒度的单独的输入，在文本处理上，对于大众语料使用 jieba 分词工具原生的词典，对于医学语料使用收集的医学词典。发现对于两个数据集，召回率的提升相对于准确率更高，尤其是 CMQANER 数据集，其召回率由原先的 63.3% 提高到了 68.6%，推断是由于分词效率的提高，显著提高了词的边界与医学实体边界的重合度，促进了提高预测的医学实体在真正的医学实体上的比例。

还有一个现象通过和基于字向量输入的结果可以横向比较得到，即使通过经过医学词典分词的医疗语料训练的粒度输入的结果也不如基于大众语料训练的字向量的结果，在 CCKSNER 数据词粒度最好的 F1 值是 0.741，而字粒度中最差的结果也达到了 0.746；在 CMQANER 数据集上词粒度最好的 F1 是 0.720，在字粒度上最差的结果达到了 0.733，造成这样的结果的原因可能三个。

1) 由于词语的个数远远超过了汉字的个数，训练词语所要求的特征空间也远远大于字的特征空间，当前医学语料规模使得词向量很难充分学习到其特征空间。

2) 尽管有医学词典的存在，也能很难完全避免分词错误的可能，当医学词典的规模并不充分大的时候，分词错误所带来的负向效果可能会超过词所提供的更充分的语义所带来的正向效果。

3) 在基于分词的词粒度的模型中，经统计，词典中的高频词往往占比例不到 10%，绝大部分词在语料中出现的频率不到 5，存在着严重的长尾问题，不利于模型的训练。

因此可以基本得出结果，在单粒度的输入下，医学中文命名实体识别的问题中往往更加倾向于使用字粒度的输入，并且词典信息可以作为字粒度的一个直接补充。

表 4.6 不同词向量输入对 CCKSNER 数据集实体识别效果的影响表

特征列表	Precision	Recall	F1 值
基于大众语料训练的词向量输入(原生词典分词)	0.73	0.708	0.719
基于医疗语料训练的词向量输入(医学词典分词)	0.742	0.739	0.741

表 4.7 不同词向量输入对 CMQANER 数据集实体识别效果的影响表

特征列表	Precision	Recall	F1 值
基于大众语料训练的词向量输入(原生词典分词)	0.741	0.633	0.683
基于医疗语料训练的词向量输入(医学词典分词)	0.760	0.686	0.720

4.2.2 字词串联法模型设计

为了词粒度信息对于字粒度信息的辅助作用，本节开始将讨论基于字词混合模型的实验。依据词是作为字的补充还是字与词是地位等价的输入这样先验假设的不同，本文分别将其称为字词串联法与字词并联法。本节主要讨论字词串联法，本节设计了两种字词串联法，字和词的信息按照一定顺序进行拼接，因此称为串联法，按照字与词信息拼接位置的不同分为直接串联法与间接串联法。

直接串联法：如图 4.12 所示，当输入序列为“是否消化不良”，其对应一个长度为 6 的字符串序列 $S = \{s_1, s_2 \dots s_5, s_6\}$ ，经过按字分割的医学语料训练会得到字向量矩阵 $X = \{x_1, x_2 \dots x_5, x_6\}$ ，经过按词分割的医学语料训练会得到词向量矩阵

$W = \{w_1, w_2 \dots w_5, w_6\}$ ，并将上下文词典特征矩阵 $C = \{c_1, c_2 \dots c_5, c_6\}$ 与词界特征矩阵 $B = \{b_1, b_2 \dots b_5, b_6\}$ 拼接成词典特征矩阵 $D = \{d_1, d_2 \dots d_5, d_6\}$, $d_i = c_i \oplus b_i$ ，在 BiLSTM-CRF 的输入层中将 X, W 和 D 三个矩阵进行拼接，可以得到任意第 i 个字符的输入如公式 4.5 所示：

$$z_i = x_i \oplus d_i \oplus w_i \quad (4.5)$$

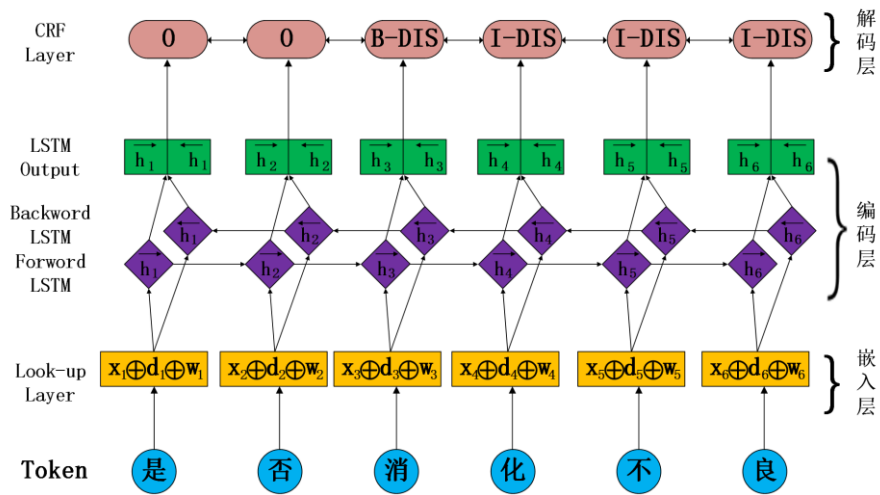


图 4.12 直接串联法模型图

在输入序列“是否消化不良”中，“是”对应的字向量 x_1 与“否”对应的字向量 x_2 是两个不同的字向量，“是”对应的词典特征 d_1 与“否”所对应的词典特征 d_2 也是两个不同的词典特征，而“是”对应的词向量 w_1 与“否”对应的词向量 w_2 相同的词向量，即“是否”对应的向量。

表 4.8 直接串联法模型的参数选择表

参数	取值	参数含义
Char_emb_size	50	字符级词嵌入层维数
Char_dropout	0.2	嵌入层随即丢弃概率
Word_emb_size	50	词级别词嵌入层维数
LSTM_layer	1	LSTM 层数
Learning_rate	0.001	学习率
LSTM_hidden	200	隐藏层维数
Weight_decay	0.0001	权重衰减
Optimizers	Adam	优化器
Batch_size	64	每次训练输入的句子数

如表 4.8 所示，在直接串联法模型的超参数选择中，我们选择字符向量维度与词向量维度均设置为 50，LSTM 层数设置为 1，学习率设置为 0.001，优化器选择为 Adam 优化器，Weight_decay 设置为 0.0001。Dropout 设置为 0.2

间接串联法：如图 4.13 所示，字向量与词向量所用的训练机制较为相似，但是语料所用的基本单元不同，由于字与词特征空间的不同，相同语料下两者的训练的充分程度程度也有所不同。并且对于词典特征向量，不同于字与词训练后的稠密向量，词典特征向量十分稀疏，不同类型的向量使用直接串联法直接拼接后得到的向量给模型后续的训练增大了难度。

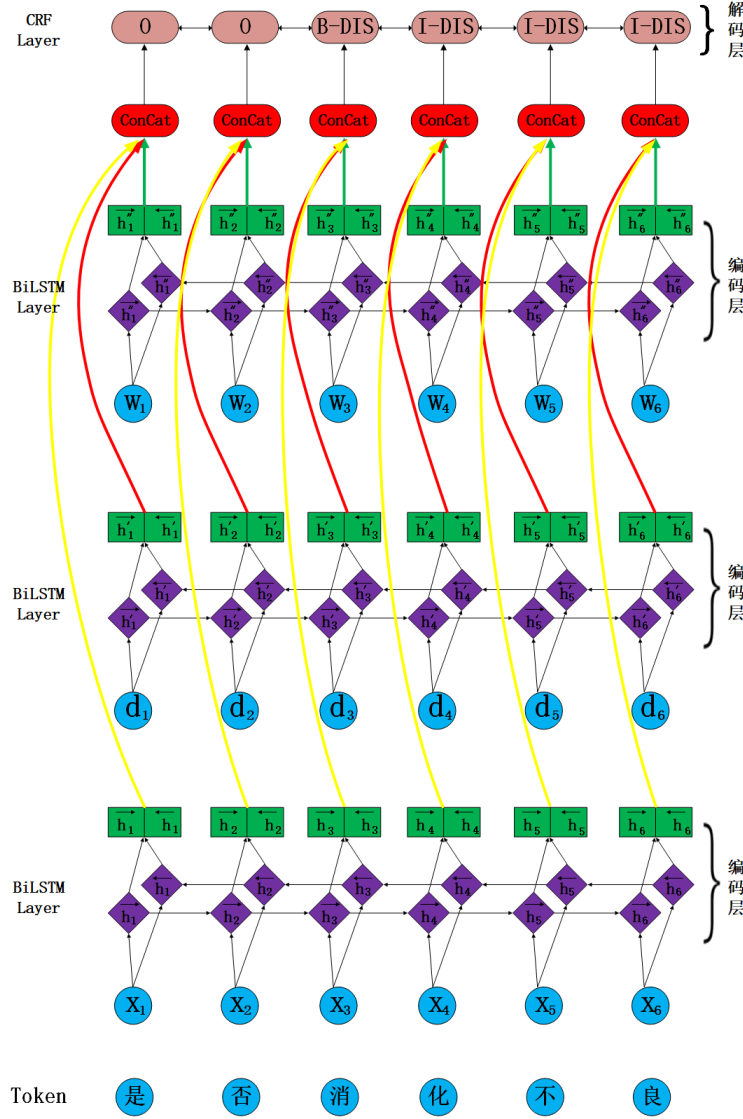


图 4.13 间接串联法模型图

考虑到上述问题，本节设置了三层 BiLSTM 网络，三者参数相互独立，分别对于第 i 个字符的输入 x_i, d_i 和 w_i 进行编码，将编码后的结果进行拼接，统一送入解码层，如公式 4.6 所示：

$$\begin{aligned}
 h_i &= [\vec{h}_i, \overleftarrow{h}_i] = BiLSTM(\vec{h}_{i-1}, \overleftarrow{h}_{i+1}, x_i) \\
 h'_i &= [\vec{h}'_i, \overleftarrow{h}'_i] = BiLSTM(\vec{h}'_{i-1}, \overleftarrow{h}'_{i+1}, d_i) \\
 h''_i &= [\vec{h}''_i, \overleftarrow{h}''_i] = BiLSTM(\vec{h}''_{i-1}, \overleftarrow{h}''_{i+1}, w_i)
 \end{aligned} \tag{4.6}$$

对于第 i 个字符的输入 x_i ，其经过 BiLSTM 网络的输入取决于当前时间步的输入 x_i ，上一个时间步的前向 LSTM 的隐向量 \vec{h}_{i-1} ，以及下一个时间的后向 LSTM 的隐向量 \overleftarrow{h}_{i+1} ，将前向 LSTM 的结果 \vec{h}_i 和后向 LSTM 的结果 \overleftarrow{h}_i 进行拼接可以得到

BiLSTM 网络网络的输出 h_i ，对于输入 d_i 与 w_i 也做相似的处理得到 h'_i 与 h''_i ，最后拼接得到的 h_i^c ，其作为 CRF 层的输入，如公式 4.7 所示。

$$h_i^c = h_i \oplus h'_i \oplus h''_i \quad (4.7)$$

模型在 x_i ， d_i ， w_i 的输出之后会分别经过三次 dropout 层，在拼接得到 h_i^c 之后又经过一层 dropout 层。

表 4.9 间接串联法模型的参数选择表

参数	取值	参数含义
Char_emb_size	50	字符级词嵌入层维数
Char_dropout	0.2	嵌入层随即丢弃概率
Word_emb_size	50	词级别词嵌入层维数
Word_dropout	0.2	嵌入层随即丢弃概率
D_emb_size	10	字典特征向量
D_dropout	0.1	嵌入层随即丢弃概率
Char_LSTM_layer	1	LSTM 层数
Word_LSTM_layer	1	LSTM 层数
D_LSTM_layer	1	LSTM 层数
Learning_rate	0.001	学习率
LSTM_hidden	200	隐藏层维数
Weight_decay	0.0001	权重衰减
Optimizers	Adam	优化器
Batch_size	64	每次训练输入的句子数

如表 4.9 所示，在间接串联法模型的超参数选择中，我们选择字符向量维度与词向量以及维度均设置为 50，D 向量维度设置为 10，LSTM 层数均设置为 1，学习率设置为 0.001，优化器选择为 Adam 优化器，Weight_decay 设置为 0.0001，Dropout 设置为 0.2。

4.2.3 两种字词串联法的实验结果及分析

表 4.10 字词串联法模型对于 CCKSNER 数据集的效果表

字词串联法模型	Precision	Recall	F1 值
字词直接串联法	0.821	0.764	0.791
字词间接串联法	0.842	0.785	0.812

表 4.11 字词串联法模型对于 CMQANER 数据集的效果表

字词串联法模型	Precision	Recall	F1 值
字词直接串联法	0.809	0.755	0.781
字词间接串联法	0.835	0.779	0.806

从表 4.10 和表 4.11 中可以看到，字词串联法模型在 CCKSNER 数据集以及 CMQANER 数据集的各项指标均取得了超越单粒度输入模型，字词直接串联法在 CCKSNER 数据集的 F1 值提升了 1.8%，在 CMQANER 数据集提升了 2.2%。而字词间接串联法则比直接串联法在 CCKSNER 数据集以及 CMQANER 数据集 F1 值各提升了 2.1% 和 2.5%。相对于直接串联法简单地将词向量、字符向量与词典特征向量进行拼接，间接串联法由于对于词向量与词典特征也做了进一步的特征抽取工作，使得原先粒度不同的词与字与稀疏的词典特征，在某种程度上做了一层统一的特征映射 使得字符融入了同一特征空间的信息，便于模型的训练，最终使得各个指标均有较大提升。

4.2.4 字词并联法模型设计

在字词串联法中使用字词拼接的方法，主要将词作为字的辅助去补充字的信息。因此在 LSTM 的每个时间步中，都提取词的信息与字的信息进行匹配。而在

本节的实验中，将词与字看作并列的两个成分，输入的时间步依旧以字为单位，词仅仅在字的位置到达某个词的词尾位置时才出现一次。文本将这种将字与词并列看待，而不仅仅将词简单的拼接字的方法叫做字词并联法。

字词后置并联法：本文在字词并联法中，设计了两种字词融合的方法，本文将在编码层进行字词融合的方法称为字词前置并联法，将在编码层之后进行字词融合的方法称为字词后置并联法。

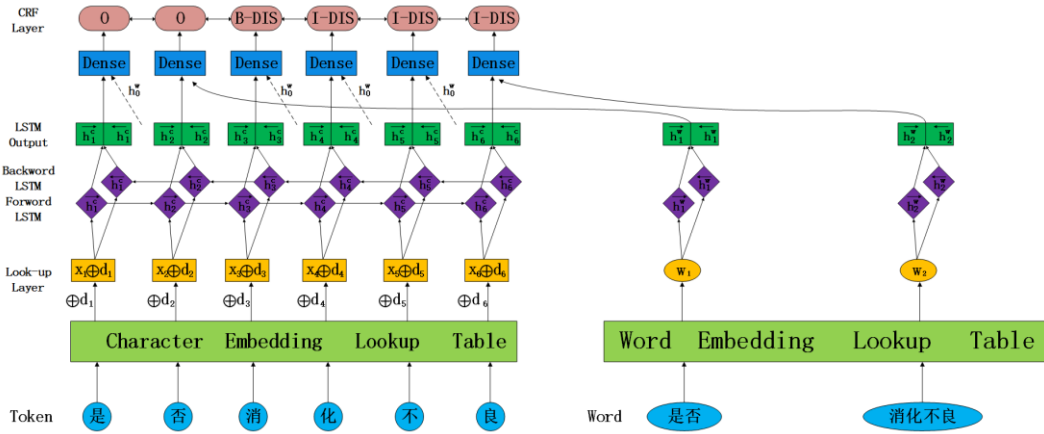


图 4.14 字词后置并联法模型图

如图 4.14 所示，当输入语料为“是否消化不良”时，其按字拆分的语料和按词拆分的语料分别会得到字向量矩阵 $X = \{x_1, x_2, \dots, x_5, x_6\}$ 和词向量矩阵 $W = \{w_1, w_2\}$ ，如果当前时间步对应着词尾的位置，如字符“否”所对应的时间步，其传入全连接的 Dense 层包含了“否”的字向量信息经过编码后的结果和“是否”的词向量信息经过编码后的结果，而字符“消”所对应的时间步，由于其字符所在位置不是词尾位置，其传入全连接 Dense 层的是“消”的字向量信息经过编码后的结果与默认的填充向量 h_0^w 。因此经过编码层处理之后，传入全连接层的中间结果矩阵 H 如公式 4.8 所示。

$$H = \{h_1^c \oplus h_0^w, h_2^c \oplus h_1^w, h_3^c \oplus h_0^w, h_4^c \oplus h_0^w, h_5^c \oplus h_0^w, h_6^c \oplus h_2^w\} \quad (4.8)$$

该模型与字词间接串联法最大的区别是：在字词串联法中，词作为字的辅助信息，在输入序列中词的个数比字的个数少的情况下，同样的词信息被多次的使用。而在本节的字词后置并联法中，词仅仅在字对应词尾的时间步才会出现，在

其他时间步均以固定向量进行填充的方式去保证 Dense 层统一的输入维度。此外，由于并联法中强调字与词的并列地位，将词典特征 d_i 作为补充信息，在输入编码层之前进行了直接的拼接。即说明 d_i 是 x_i 的补充，而 w_i 与 $x_i \oplus d_i$ 是地位相等的并列关系。

表 4.12 后置并联法模型的参数选择表

参数	取值	参数含义
Char_emb_size	100	字符级词嵌入层维数
Char_dropout	0.2	嵌入层随即丢弃概率
Word_emb_size	100	词级别词嵌入层维数
LSTM_layer	1	LSTM 层数
Learning_rate	0.001	学习率
LSTM_hidden	200	隐藏层维数
Weight_decay	0.0001	权重衰减
Optimizers	Adam	优化器
Batch_size	64	每次训练输入的句子数

如表 4.12 所示，在直接串联法模型的超参数选择中，我们选择字符向量维度与词向量维度均设置为 100，LSTM 层数设置为 1，学习率设置为 0.001，优化器选择为 Adam 优化器，Weight_decay 设置为 0.0001。Dropout 设置为 0.2。

字词前置并联法：在字词前置并联法中，字与词信息的融合发生在编码层中，本节的模型借鉴了发表在 ACL 2018 会议上的 Lattice LSTM 框架[45]。区别在于本文是融入医学信息的字向量与词向量，并且需要对于词典进行额外处理，如图 4.15 所示。

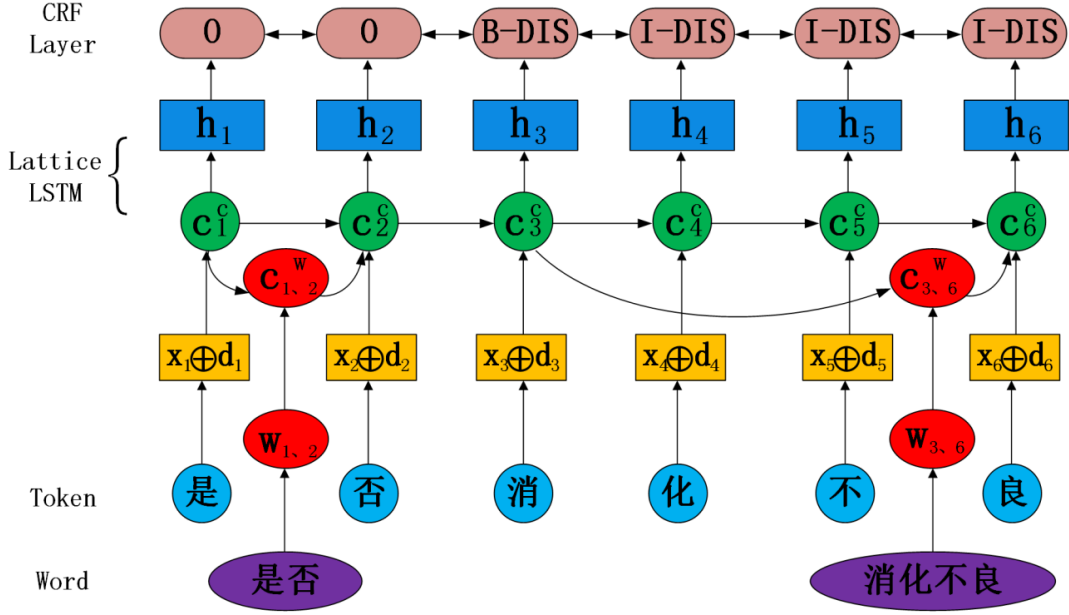


图 4.15 字词前置并联法模型图

当输入语料为“是否消化不良”时，对应的字向量矩阵 $X = \{x_1, x_2, \dots, x_5, x_6\}$ ，对应的词向量矩阵为 $W = \{w_{1,2}, w_{3,6}\}$ ，词向量 $w_{b,e}$ 的下标 b 和下标 e 表示了以字符为单位的开始位置索引和结束位置索引，因此在本例中 $w_{1,2}$ 表示“是否”所对应的词向量， $w_{3,6}$ 表示“消化不良”所对应的词向量。对于第 j 个字符的 x_j 与 d_j 的联合输入 $x_j \oplus d_j$ ，如公式 4.9 所示：

$$\begin{aligned}
 i_j^c &= \sigma(W_{xi}^c(x_j \oplus d_j) + W_{hi}^c h_{j-1}^c + b_i^c) \\
 o_j^c &= \sigma(W_{xo}^c(x_j \oplus d_j) + W_{ho}^c h_{j-1}^c + b_o^c) \\
 f_j^c &= \sigma(W_{xf}^c(x_j \oplus d_j) + W_{hf}^c h_{j-1}^c + b_f^c) \\
 C_j^c &= \tanh(W_{xc}^c(x_j \oplus d_j) + W_{hc}^c h_{j-1}^c + b_c^c)
 \end{aligned} \tag{4.9}$$

同样对于第 b 个字符开始，第 e 个字符结束的词所对应的向量 $w_{b,e}$ ，用 $C_{b,e}^w$ 表示词的记忆细胞状态，有公式：

$$\begin{aligned}
 i_{b,e}^w &= \sigma(W_{wi}^w w_{b,e} + W_{hi}^w h_b^c + b_i^w) \\
 f_{b,e}^w &= \sigma(W_{wf}^w w_{b,e} + W_{hf}^w h_b^c + b_f^w) \\
 C_{b,e}^w &= \tanh(W_{wc}^w w_{b,e} + W_{hc}^w h_b^c + b_c^w) \\
 C_{b,e}^w &= f_{b,e}^w \otimes C_b^c + i_{b,e}^w \otimes C_{b,e}^c
 \end{aligned} \tag{4.10}$$

由于采用了 Lattice-LSTM 的结构, 记忆单元的状态有了更多的输入源。如 C_6^c 的输入源包括了“不”对应的记忆单元 C_5^c 和词“消化不良”对应的记忆单元 $C_{3,6}^w$ 。因此对于任意 C_e^c , 其中 e 属于词尾字符序号, 均需要考虑其字符与词典特征所对应的联合向量 $x_e \oplus d_e$ 与对应词记忆单元 $C_{b,e}^w$ 的影响。通过为了控制词记忆单元的影响程度, 额外增加了一个输入门进行限制, 如公式 4.11 所示:

$$i_{b,e}^c = \sigma(W_x^c(x_e \oplus d_e) + W_C^c C_{b,e}^w) \quad (4.11)$$

表 4.13 前置并联法模型的参数选择表

参数	取值	参数含义
Char_emb_size	100	字符级词嵌入层维数
Char_dropout	0.2	嵌入层随即丢弃概率
Word_emb_size	50	词级别词嵌入层维数
LSTM_layer	1	LSTM 层数
Learning_rate	0.001	学习率
LSTM_hidden	200	隐藏层维数
Weight_decay	0.0001	权重衰减
Optimizers	Adam	优化器
Batch_size	1	每次训练输入的句子数

如表 4.13 所示, 在直接串联法模型的超参数选择中, 我们选择字符向量维度为 100, 词向量维度设置为 50, LSTM 层数设置为 1, 学习率设置为 0.001, 优化器选择为 Adam 优化器, Weight_decay 设置为 0.0001, Dropout 设置为 0.2。

4.2.5 两种字词并联法的实验结果及分析

表 4.14 字词并联法模型对于 CCKSNER 数据集的效果

字词并联法模型	Precision	Recall	F1 值
字词前置并联法	0.861	0.808	0.834
字词后置并联法	0.837	0.779	0.806

表 4.15 字词并联法模型对于 CMQANER 数据集的效果

字词并联法模型	Precision	Recall	F1 值
字词前置并联法	0.856	0.781	0.817
字词后置并联法	0.817	0.761	0.789

从表 4.14 和表 4.15 中可以看到，字词前置并联法在两个数据集中均取得了最好的效果，相比于串联法中效果最好的字词间接串联法，在 CCKSNER 以及 CMQANER 数据集中均有较大幅度的提升，分别提升了 2.2% 以及 0.9%，符合数据集特性，CCKSNER 数据集上有更多专业性较强的医学术语，与其边界识别的好坏程度强烈相关，通过在 LSTM 本身记忆单元的结构进行信息融合，能够更好地进行字与词的融合，使得对于实体边界更加敏感的 CCKSNER 数据集有了更大的效果提升。说明了即使同样都是中文医学语料，不同数据集对于不同的融合方式的敏感性也有所差异。但后置并联法相比于单字拼接特征词典特征的输入结果也有着较大地提升，说明了即使对字的输入中拼接了词的边界特征，一个较为精确的词粒度输入仍然可以对其有提升空间。在字词并联法与字词串联法的比较中，发现结果最好的是字词前置并联法，说明字与词独立的假设不失为一个较好的思路，但是在字的词尾时间步上对于融合方法进行较好的设计。

表 4.16 CCKSNER 数据集在字词前置并联法模型下各个实体上识别效果对比表

测试指标	Precision	Recall	F1 值
解剖部位	0.801	0.742	0.770
影像检查	0.942	0.911	0.926
实验室检验	0.733	0.84	0.782
疾病和诊断	0.823	0.771	0.796
药物	0.915	0.844	0.878
手术	0.685	0.627	0.653
症状	0.986	0.901	0.942
总指标(Micro)	0.861	0.808	0.834

表 4.17 CMQANER 数据集在字词前置并联法模型下各个实体上识别效果对比表

测试指标	Precision	Recall	F1 值
疾病(disease)	0.864	0.823	0.843
人群(crowd)	0.932	0.832	0.932
症状(symptom)	0.866	0.824	0.845
身体部位(body)	0.792	0.795	0.793
治疗方法 (treatment)	0.846	0.619	0.713
时间(time)	0.838	0.498	0.621
药物(drug)	0.734	0.535	0.618
范围(feature)	1.0	0.964	0.982
生理机能 (physiology)	0.981	0.838	0.903
检测(test)	0.817	0.631	0.711
科室(department)	0.8	1.0	0.889
总指标(Micro)	0.856	0.781	0.817

字词前置并联法模型在 CCKSNER 数据集以及 CMQANER 数据集中均取得了最优的效果，因此我们将模型在各个数据集中各个实体的指标列出对比，从表

4.16 中可知在 CCKSNER 数据集中,模型对“药物”、“影像检查”以及“症状”类别的实体识别效果较好,F 值均达到了 87% 以上,而对“实验室检验”、“手术”、“解剖部位”等类别识别效果较差,究其原因,是因为 CCKSNER 数据集属于电子病历数据集,对于检验和手术等专有名词较多,名词复杂,实体边界较为难以区分,而模型中基于医疗的词典信息大多来源于通用医疗语料,对于一些特定病症的特定名词,很难做到有效覆盖。

根据表 4.17,对于 CMQANER 数据集,该数据集本身来自于日常医院问答对话,专业名词较少,除了“药物”以及“时间”实体类别 F 值效果较差,仅为 0.618,和 0.621,其余类别 F 值均超过 70%,究其原因是因为“药物”类别专有名词较多且药物种类五花八门,很难有一种较为统一的特征,而“时间”类别较低则是因为经过分析,对话中“时间”种类较为复杂,除了阿拉伯数字的时间也有汉语的时间。对于其他种类,“生理机能”,“范围”,“人群”的 F 值均超过 90%,说明字词前置并联法模型对该类别实体识别效果较好。

4.3 本章小结

在本章中先获取了医学语料与医学词典,并设计了在字粒度级别利用医学词典的上下文词典特征与词界特征,在词粒度上进行了基于医学词典的分词,并利用获得的医学语料进行了词向量与字向量的训练。通过字粒度与词粒度的单粒度增强实验,表明了医学语料与医学词典不论对于何种粒度的输入都具有一定的提升作用。并基于词是字的补充信息或字与词是两种并列的输入单位这两种先验假设的不同设计了字词串联法与字词并联法,又将其细分为字词直接串联法与字词间接串联法,字词前置串联法与字词后置串联法。在两类医学数据集上的实验结果表明,不同的字词融合方式对于携带词典信息的字向量与词向量得到的最终模型效果有较大的区别,且不同类型的中文医学数据集对于不同字词融合方式的敏感度并不相同。字词前置并联法取得了最高的效果,说明字与词相互并列输入这样的先验假设具备一定的可取之处。

第五章 BERT 模型融入知识信息

5.1 基于 BERT-CRF 的实体识别模型

5.1.1 基于 BERT 模型的微调

上一章中主要采用 BiLSTM 网络作为编码层，其本质是将前向的 LSTM 网络与后向的 LSTM 网络得到的隐层表示进行拼接，对于单个 LSTM 网络，在预测某一个时间步的输出时仅仅使用了当前时间步的输入与上一个时间步的隐层输出。严格而言，没有真正做到融合上下文的语义表示。

因此近期研究人员和工程人员开始使用 BERT 作为预训练方法，使用了遮罩语言模型去预测一部分被遮罩(masked)的词汇,是首次在预训练中真正意义上的融合了上下文的词表示。并且在预训练中进行了句子级别的建模，更好的把握了上下文关系，在命名实体识别等自然语言处理领域效果极好。BERT 使用的是一种降噪自动编码器(Denoising Autoencoder,DAE)的思想，被遮罩的词相当于在输入端加入了噪音，可以自然的利用上下文去预测被遮罩的词，如图 5.1 所示。当输入文本“是否是消化不良”时，由于“消”被 BERT 模型遮罩，需要利用上下文信息进行预测，因此“消”字符得到信息做到了真正意义上融合了上下文的语义。

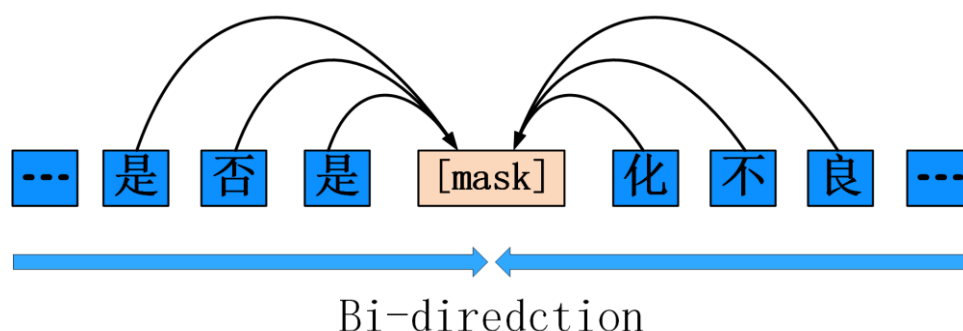


图 5.1 BERT 遮罩语言模型示意图

BERT 是一种多层双向 Transformer 编码器，需要大量的数据和 TPU 级别的计算力才适合进行训练，Google 目前开源了两个版本的 BERT,其参数如下：

表 5.1 BERT 不同版本参数表

版本	Transformer 层数	隐层大小	Attention head 个数	总参数量/Million
BERT Base	12	768	12	110
BERT Large	24	1024	16	340

考虑到已经预训练好的 BERT 所采用的语料大小远远超过本文收集到的医学语料的大小，因此不做 BERT 模型的重新训练，而是使用已经被中文语料训练好的 BERT Base 模型，分别在文本中的两份医学语料数据集中进行微调。

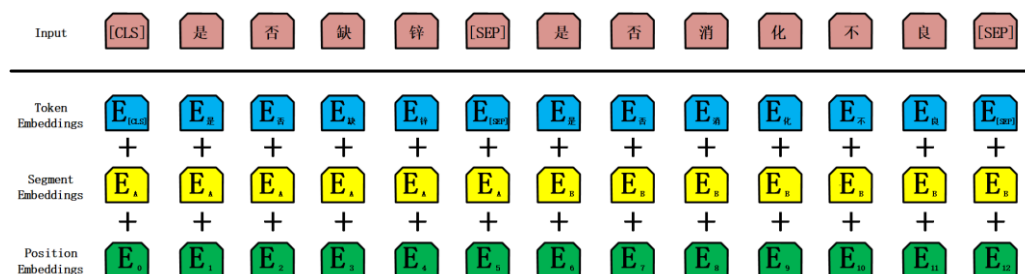


图 5.2 BERT 模型输入示意图

如图 5.2 所示，当输入 CMQANER 数据集两个连续的句子“是否缺锌”和“是否消化不良”时。BERT 模型会将输入文本每个最小单位的字符(在 BERT 模型中称为 WordPiece)转化成三个特征向量相加的形式，三个向量分别是 Token Embeddings, Segment Embeddings, Position Embeddings:

1.Token Embeddings: BERT 利用基于遮罩的双向语言模型进行预训练，微调过程对于每个 WordPiece 会输出一个 768 维的向量，在本例中形成一个 768 行，13 列的 Token Embeddings 矩阵。

2.Segment Embeddings:预训练的过程中也进行了句子对的分类任务,因此存在 Segment Embeddings 的向量对于不同的句子类标给予了不同的向量表示，本例中两个连续的句子字符分别得到两种向量表示。

3.Position Embedding: Transformers 构成了 BERT 的基本单元 [30]。而 Transformers 单元本身采用了多头注意力的机制抽取上下文的特征，不包含位置信息，因此需要进行位置编码，如公式 5.1 与公式 5.2 所示。

$$PE(p, 2i) = \sin\left(\frac{p}{1000^{\frac{2i}{d_{pos}}}}\right) \quad (5.1)$$

$$PE(p, 2i+1) = \cos\left(\frac{p}{1000^{\frac{2i}{d_{pos}}}}\right) \quad (5.2)$$

下标 pos 表示词的位置, i 表示词的维度, d_{pos} 为 512 维, 位置 i 可以映射成 d_{pos} 长度的位置向量。通过三角函数的转换, 位置是 $k+p$ 的位置向量可以表示为位置是 k 的位置向量的线性变化, 因此 PE_{pos+k} 可以使用 PE_{pos} 的线性组合进行表示。

在分别经过 CCKSNER 数据集和 CMQANER 数据集微调之后的 BERT 模型, 输入长度为 N 的文本序列 $S = \{s_1, s_2, \dots, s_{N-1}, s_N\}$ 的文本, 能够对于文本序列的每个单位输出一个经过医疗文本动态调整的 768 维的字符向量。

5.1.2 BERT-CRF 模型融入词典特征实验

可以从理论层面认为 BERT 模型比 BiLSTM 模型更好的提取了文本的上下文表征, 在本节中为了验证词典特征的效果, 实验中对 BERT 模型中的第 i 个字符输出 h_i 拼接上前文中提到的词典特征 d_i , 将 $h_i \oplus d_i$ 作为第 i 个字符的 CRF 层输入, 如图 所示。并与不拼接 d_i , 直接输入将 BERT 模型输出作为 CRF 层输入进行对比, 直接将 h_i 输入进 CRF 层的结果进行对比。添加词典特征与不添加词典特征的

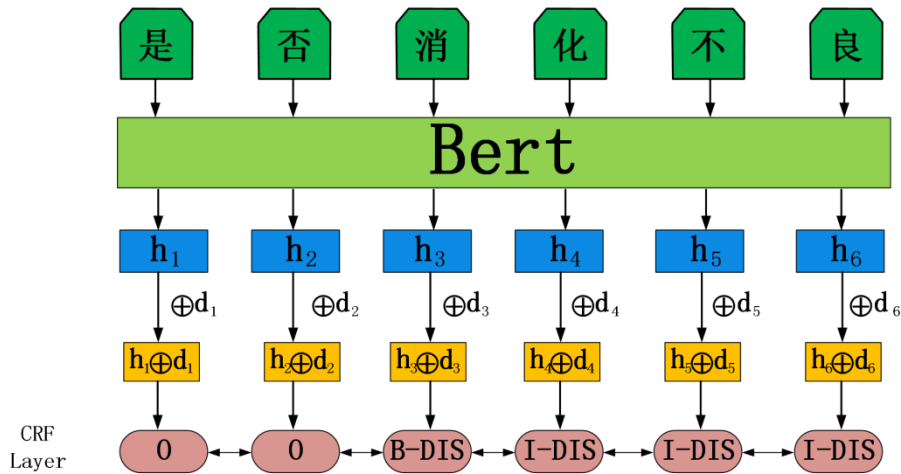


图 5.3 融入词典特征的 BERT+CRF 模型图

BERT-CRF 模型在 CCKSNER 数据集和 CMQANER 数据集的结果如表 5.2 和表 5.3 所示。

表 5.2 BERT-CRF 模型对于 CCKSNER 数据集的效果表

CRF 层输入	Precision	Recall	F1 值
拼接词典特征	0.818	0.828	0.823
不拼接词典特征	0.783	0.792	0.788

表 5.3 BERT-CRF 模型对于 CMQANER 数据集的效果表

CRF 层输入	Precision	Recall	F1 值
拼接词典特征	0.789	0.816	0.802
不拼接词典特征	0.773	0.796	0.784

对于 CCKSNER 数据集, 在使用词典特征之后 F1 值从 0.788 提升到 0.823, 而召回率从 79.2%提升到了 82.8%, 超过了此前所有的基于 BiLSTM 的模型, 通过分析得到主要原因是 BERT 模型进行更好的上下文表征, 如文本中存在一例标错的样本, 将“胆囊切除术后”标记成了手术, 在测试集中再次“胆囊切除术后”的文本时, 此前基于字的几种 LSTM 的模型中均将“胆囊切除术后”识别成手术实体, 而两种 BERT-CRF 的模型, 不论是否拼接上词典特征, 均成功地将“胆囊切除术”整体识别成手术实体。这体现了经过大规模语料预训练的 BERT 模型输出字符向量能够更好的捕获上下文的语义信息, 对于错误标注的样本具备一定容错率。

对于 CMQANER 数据集, 在使用词典特征之后 F1 值从 0.784 提高到了 0.802, 其召回率从 79.6%提升至了 81.6%, 相对于 CCKSNER 数据集, 其提升幅度较小。分析认为主要原因在于线上问答数据集的上下文语义与大众的文本的语义有较大的相似之处, 观察发现文本中的人群(crowd)实体如“妇女”, “宝宝”, “成人”等, 不论是在拼接词典特征之前还是拼接词典特征之后(拼接前后单实体 F1 值分别为 0.846 与 0.848)均有较高的识别精度, 且可以通过测试样本量得到, 前后指标并不具有显著性差异。

此外, 在与基于 LSTM 的字词混合模型进行横向比较中, 发现拼接词典特征

的 BERT-CRF 的模型在两个数据集上没有超过字词前置并联法的效果，但是超过了另外三个字词混合模型的效果。经过分析，原因主要有两点。

1)医学词典信息利用程度不同：在基于 LSTM 的字词混合模型中，医学词典特征除了体现在词典特征向量，还体现在了基于医学词典得到的词向量上。而在 BERT-CRF 模型中仅仅是对于词典特征向量做了简单的拼接，相对于此前基于 LSTM 的模型，对于词典信息利用并不充分。

2)未使用医疗语料重新预训练：本节考虑到收集到医疗语料的规模与 BERT 预训练需要的计算力成本，并未融合医学语料与大众文本语料进行重新预训练。而是使用已经用中文大众文本语料预训练的 BERT 模型在医学语料上进行微调，即当前医学语料对于 BERT 模型的训练程度并不完全充分。

尽管当前的 BERT-CRF 的模型对于医疗文本的训练程度成为达到最优，但通过本节的实验结果与之前的结果进行对比，仍然可以得到下面两个结论：

1)基于 BERT-CRF 的模型由于对于上下文的表征更强，在不考虑其他辅助特征的情况下，在两个不同文本风格的医疗实体识别问题上均会得到超过 BiLSTM-CRF 的效果。

2)医学词典信息的加入，不但在基于 LSTM 结构的深度学习模型上会有显著的效果提升，在基于 BERT 预训练语言模型的结构上也能够进一步提升效果。可以参考上文对于 LSTM 结构的词典信息融入方式，进行进一步的探索。

5.2 BERT 模型融合知识图谱

5.2.1 软位置编码与可见矩阵

在医学命名实体识别问题中，医学词典是一种常用的辅助信息。但是这种形势的辅助信息本身不包含领域知识。比如深度学习模型中无法从词典中获取“板蓝根颗粒可以治疗感冒”这样的领域知识。本节在上节的拼接上词典特征的 BERT 模型基础上讨论，如何进一步融入领域知识信息。

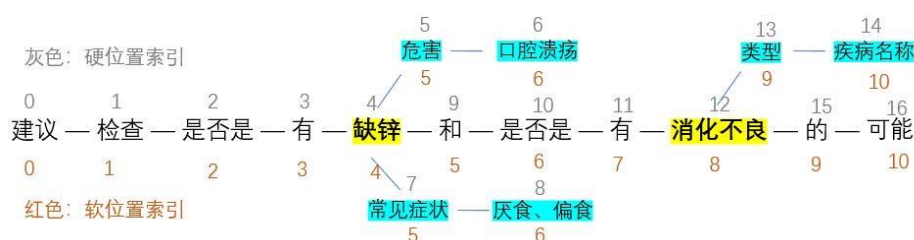


图 5.4 句子树软位置与硬位置编码图

对于 CCKSNER 数据集的文本序列“建议检查是否有缺锌和是否有消化不良的可能”，由于“缺锌”和“消化不良”都是医学实体，在文本中所采用的医学知识图谱(MedicalKG)可以检索到“缺锌-常见症状-厌食、偏食”，“缺锌-危害-口腔溃疡”，“消化不良-类型-疾病名称”，当利用知识图谱对于文本序列进行知识注入，结果如图 5.4 所示。此时文本序列由线性结构转换成了树结构，传统的 BERT 模型无法直接处理树形结构的输入。为了解决这个问题，本节模型借鉴了北京大学与腾讯在 AAAI-2020 会议中提出的 K-BERT 模型中软位置编码与可见矩阵两个思想[33]。

软位置编码：图 5.4 中灰色数字对应了硬编码的索引，硬编码的表示文本序列的输入 BERT 模型的顺序，即输入模型的线性结构是“建议检查是否有缺锌危害空腔溃疡常见症状厌食、偏食和是否有消化不良类型疾病名称的可能”，其软位置索引的功能在于能够使得通过硬位置进行编码的输入恢复原始的顺序，如在原始输入中“缺锌”处于第 4 个索引的位置，“危害”和“常见症状”这两个知识图谱由“缺锌”这个头实体所对应的关系都处于第 5 个索引的位置，其原始文本本身在“缺锌”后面的字符“和”在软编码中仍是对应第 5 个索引的位置。

对于输入文本在得到软编码与硬编码的索引信息之后，如图 5.5 所示，软位置编码将代替原先的顺序的位置编码，当原始医学文本注入知识图谱信息之后文本序列变成了“建议检查是否有缺锌危害空腔溃疡常见症状厌食偏食和”，其对应的 Token Embeddings 与 Segment Embeddings 并未发生改变，仅仅是将软位置索引信息替换原先的绝对索引信息，表明在本节的融入知识图谱的模型中不需要改变 BERT 原先的构造与预训练好的结果，仅仅是需要进行文本预处理以融入图谱信息，因此不需要重新去训练一个新的 BERT，大大减少了计算成本，具备了

较强的模型迁移能力。



图 5.5 带有软位置信息的 BERT 模型输入图

可见矩阵：对于任何通过添加三元组信息得到的句子树，尽管可以通过硬编码转化成线性结构，通过软编码方式保存原始顺序，但整合了知识图谱之后，尽管添加了一定的辅助信息，但可能会使得原始语义发生变化，即知识图谱提供辅助信息的同时也可能带来知识噪声(Knowledge Noise)的问题。在一个通过知识图谱转化成的句子树中，其附加的三元组信息不应该影响其他词汇的语义。在图 5.4 所示的例子中，“缺锌-常见症状-厌食、偏食”是头实体“缺锌”所对应的三元组，其作用仅仅是用于丰富主体词“缺锌”所对应的头实体信息，而不能跨越“缺锌”对于其他词汇比如“是否是”的语义产生影响。处于这样的考虑，本文也采用了可见矩阵的结构，如公式 5.3 所示：

$$M_{ij} = \begin{cases} 0 & w_i = w_j \\ -\infty & w_i \neq w_j \end{cases} \quad (5.3)$$

其中 $w_i = w_j$ 表示两个词单元处于同一分支， $w_i \neq w_j$ 表示两个词不处于同一个分支。处于同一个分支的词单元处于相互可见的状态，即在计算两个词单元的计算相关性(Attention 权重)的过程中需要参与计算，不增加额外的约束，可见矩阵对应的值为 0。若不处于同一分支，则两个词单元处于不可见的状态，在计算相关性时由于负无穷的作用，使得相关系数变成 0，变相地实现了相互不可见的功能。

5.2.2 CMK-BERT 模型总体设计

在使用了软位置编码信息之后，解决了树形结构无法输入 BERT 模型的问题。在使用了可见矩阵后，缓解了 BERT 模型的输出受到知识图谱产生的知识噪声影响的问题。在此基础本文设计了适用于中文医学命名实体识别的 CMK-BERT 模型，如图 5.6 所示，其可以分为四个部分：

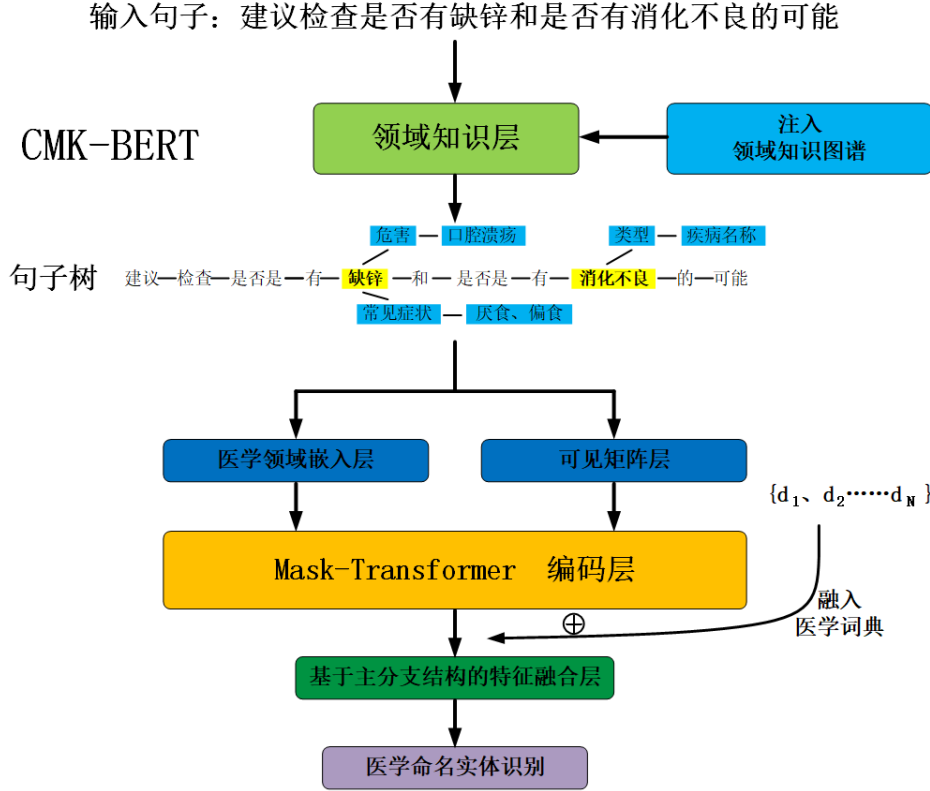


图 5.6 CMK-BERT 模型图

1)领域知识层: 对于任意给定的医学文本序列 $s = \{w_1, w_2 \dots w_N\}$, 与知识图谱 K , 会进行知识查询(Knowledge Query)与知识注入(Knowledge Inject)两个部分, 知识查询可以得到三元组集合 E , 如公式 5.4 所示。

$$E = \text{Knowledge_Query}(s, K) \quad (5.4)$$

得到的 $E = \{(w_i, r_{i0}, w_{i0}), \dots, (w_i, r_{ik}, w_{ik})\}$, 表示以第 i 个词单元为头实体在图谱 K 中进行检索所能够找到的所有三元组。

$$t = \text{Knowledge_Inject}(s, E) \quad (5.5)$$

对于第 i 个词单元进行知识查询之后, 需要将查询到的所有三元组注入到医学文本 s 中, 得到更新的句子树 t , 若有 k 个分支, 则会以 w_i 为树的根节点形成 k 个分支的句子树, 如公式 5.5 所示。本例中输入一个长度为 11 的文本序列(假设是以词进行分割, 以字符进行分割的步骤相似, 在词尾位置进行注入), 对于每一个进行词单元进行查询和注入, 最终得注入 3 个三元组。

2)医学领域嵌入层与可见矩阵层: 结合上节的内容,医学领域嵌入层与常规的 BERT 模型结构一致,经过医学文本训练集的微调,最大的区别在于利用硬位置编码进行输入,软位置编码保存原始文本和三元组信息。可见矩阵层通过可见矩阵来缓解知识噪声的影响。

3)Mask-Transformer 编码层: 在常规的自注意层的基础上添加了可见矩阵 M , 如公式 5.6 所示。

$$\begin{aligned} Q^{i+1}, K^{i+1}, V^{i+1} &= h^i W_q, h^i W_k, h^i W_v \\ S^{i+1} &= \text{soft max}(\frac{Q^{i+1} + K^{i+1T} + M}{\sqrt{d_k}}) \\ h^{i+1} &= S^{i+1} V^{i+1} \end{aligned} \quad (5.6)$$

通过二值的可见矩阵实现了不可见的功能,若词单元 w_k 对于词单元 w_j 是不可见的,由于可见矩阵 M_{jk} 取值为负无穷,利用 Attention 权重计算得到的相关系数 S_{jk}^{i+1} 会无趋于 0,即词单元 w_k 对于词单元 w_j 的编码向量的形成是没有贡献的。

4)基于主分支结构的特征融合层: 由于在模型中的前面三层涉及到知识图谱的信息融合与编码转换的问题。医学词典特征无法直接融入前面三层的任何一层,因此在 Mask-Transformer 编码层增加了特征融合层。可以注意到对于长度为 N 的文本输入,词典特征矩阵的元素个数同样为 N ,意味着在特征融合层中会对于主分支进行选择,如软编码位置信息为 $\{0,1,2,3,4,5,6,5,6,5,6,7,8,9,10,9,10\}$ 时,将会跳过三个不属于主分支的三元进行特征融合,去加入原始文本中真正的上下文信息。

5.2.3 基于 CMK-BERT 模型的数据转换

在使用 CMK-BERT 模型对医学语料进行微调时,原始的医学文本会经过领域知识层进行知识查询和知识注入,而在基于 LSTM 的模型中采用的数据格式是单字单行的标注方式,没有对于分句之间进行显性切分,不利于知识查询。在 CMK-BERT 模型中对于以字符为单位的输入,要求原始文本需要整理成如图 5.7 所示

格式，将一段独立语义的文本与其标注整理为一行，主要目的是便于在原始文本进行知识查询时具有文本查询的边界，且知识注入后便于同步标注信息的变化。

知识注入后的文本以硬位置索引顺序的线性结构展示，文本将医疗领域嵌入

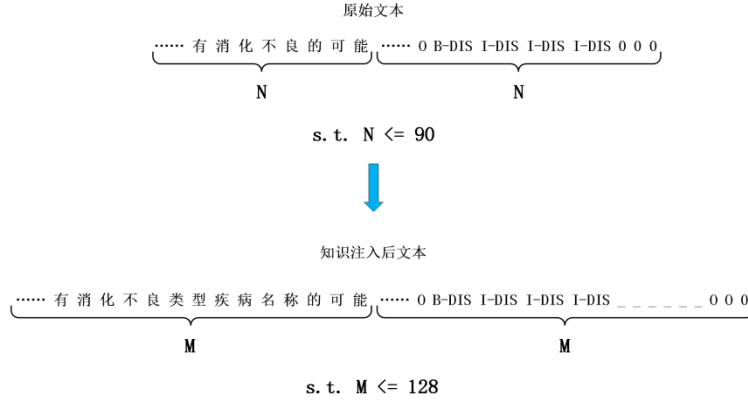


图 5.7 CMK-BERT 模型输入变换图

层能够处理的最大文本序列长度作为实验的超参数设置为 $2M$ ，随着 M 的增高，模型的训练时间与训练资源要求也随着增高。本文限制 M 为 128，由于原始文本经过领域知识层会发生长度的增加，因此在本文设置原始文本的最大长度 N 为 90，减少增加三元组信息之后数据溢出的可能性。因此在本文中需要基于 N 设置的大做出过如下步骤的数据转换：

1)切割：根据设定的策略对于单字单行的标注文本进行切割，具体策略包括有：将出现句号设为语义表达完整的依据，进行切分；针对有空格的段落先根据空格进行切分，再进行人工检验判断切分结果是否完整；对于文本长度超过 90 的单句，若其中包含了切分语义的标点符号如“，”或“；”，则按照标点符号进行迭代式的切分，直至每个切分单位长度都不超过 90。若仍存在长度超过 90 的切分单位，则以第 90 个字符为切割面进行强制切分，直至所有切割单元的长度都不超过 90。

2)转换与拼接：将切分单元的输入文本和文本标注分别按空格分割置于同一行，再以空格为连接符号拼接文本和文本标注。

3)融合不同的语义单元，并清理空行。

图 5.8 CMK-BERT 模型文本输入图

当进行完数据转换，文本转变为适用于 CMK-BERT 模型直接读取的形式，转换后文本如图 5.8 所示。

5.2.4 融入知识图谱的实验分析

本文使用的三个知识图谱分别为 CN-DBpedia、知网(HowNet)和医学知识图谱(MedicalKG)。其中 CN-DBpedia 是由复旦大学知识工场实验室研发的中文通用百科知识图谱，其包含了 900 万以上的百科实体[46]。HowNet 是在线常识知识库，将义原描述成词的最小单位，其三元组均为“头实体-义原-尾实体”的形式。而 MedicalKG 是医疗方向的知识图谱，包括了疾病、症状、临床检查等常见的医疗实体。

图 5.9 CN-DBpedia 图谱

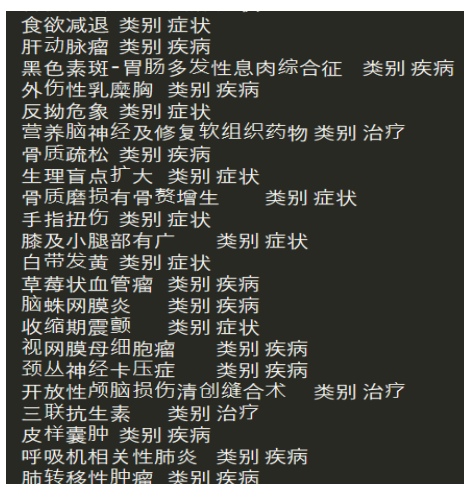


Figure 5.10 displays a HowNet graph, which is a hierarchical knowledge structure. It lists various medical terms and their relationships, categorized by type (类别) and treatment (治疗). The terms include:

- 食欲减退 类别 症状
- 肝动脉瘤 类别 疾病
- 黑色素斑-胃肠多发性息肉综合征 类别 疾病
- 外伤性乳腺瘤 类别 疾病
- 反拗危象 类别 症状
- 营养脑神经及修复软组织药物 类别 治疗
- 骨质疏松 类别 疾病
- 生理盲点扩大 类别 症状
- 骨质磨损有骨赘增生 类别 症状
- 手指扭伤 类别 症状
- 膝及小腿都有广 类别 症状
- 白带发黄 类别 症状
- 草莓状血管瘤 类别 疾病
- 脑蛛网膜炎 类别 疾病
- 收缩期震颤 类别 症状
- 视网膜母细胞瘤 类别 疾病
- 颈丛神经卡压症 类别 疾病
- 开放性颅脑损伤清创缝合术 类别 治疗
- 三联抗生素 类别 治疗
- 皮样囊肿 类别 疾病
- 呼吸机相关性肺炎 类别 疾病
- 肺转移性肿瘤 类别 疾病

图 5.10 HowNet 图谱

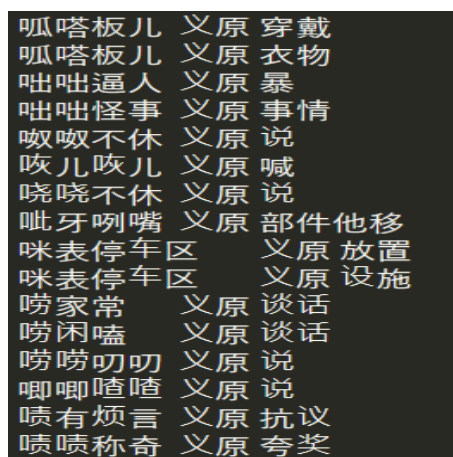


Figure 5.11 displays a MedicalKG graph, which is a hierarchical knowledge structure. It lists various medical terms and their relationships, categorized by type (类别) and treatment (治疗). The terms include:

- 呱嗒板儿 义原 穿戴
- 呱嗒板儿 义原 衣物
- 咄咄逼人 义原 暴
- 咄咄怪事 义原 事情
- 嗷嗷不休 义原 说
- 咳儿咳儿 义原 喊
- 唠唠不休 义原 说
- 毗牙咧嘴 义原 部件他移
- 咪表停车区 义原 放置
- 咪表停车区 义原 设施
- 唠家常 义原 谈话
- 唠闲嗑 义原 谈话
- 唠唠叨叨 义原 说
- 唧唧喳喳 义原 说
- 啧啧有烦言 义原 抗议
- 啧啧称奇 义原 夸奖

图 5.11 MedicalKG 图谱

三个图谱信息如图 5.9,5.10,5.11 所示,可以发现 CN-DBpedia 图谱中词汇的门类更广泛, HowNet 图谱采用义原的形式更加偏向发现词汇本身的属性信息,而且 MedicalKG 图谱对医学领域针对性较强,三类图谱可以反映出不同类型的领域信息。

本节基于上节设计的 CMK-BERT 模型框架,在两个数据集上分别融入了这三种不同的类型的图谱进行实验。

表 5.4 CMK-BERT 模型融入不同图谱在 CCKSNER 数据集的效果表

融入图谱种类	Precision	Recall	F1 值
Hownet 图谱	0.829	0.829	0.829
CnDbpedia 图谱	0.840	0.843	0.842
Medical 图谱	0.833	0.830	0.831

表 5.5 CMK-BERT 模型融入不同图谱在 CMQANER 数据集的效果

融入图谱种类	Precision	Recall	F1 值
Hownet 图谱	0.799	0.822	0.810
CnDbpedia 图谱	0.803	0.825	0.814
Medical 图谱	0.809	0.820	0.814

对于 CCKSNER 数据集，可以看到在领域知识层注入了 CnDbpedia 图谱取得了最好的模型效果，其 F1 值到了最高的 0.842。而融入了医学领域的 Medical 图谱之后的效果仅仅比融入 Hownet 图谱的 F1 值高 0.2 个点。经过分析发现，Medical 图谱中仅有 13864 份三元组，其覆盖的实体大多数疾病、症状、治疗方案等常规的实体类型，随机从 CCKSNER 的数据集中抽取出若干份实体，发现在 Medical 图谱上的覆盖率较低，如药物实体“替吉奥”，疾病实体“左上肺低分化癌”，手术实体“左半结肠切除术”等实体均不能在 Medical 图谱中找到。而 CnDbpedia 图谱中尽管是百科数据，也能找到相当比例的医学三元组信息，比如“气脖子病-别称-大脖子病”表明可以通过“气脖子病”的头实体找到别称关系的三元组。通过“病因”关系可以找到“气脖子病-病因-缺少碘元素”这样的三元组。说明了 CnDbpedia 这样的百科图谱对于医学实体识别这样垂直领域任务也具备了不可忽视的效果提升的作用。

而对于 CMQANER 数据集，发现融入 CnDbpedia 图谱与融入 Medical 图谱最后的 F1 值效果相当，融入 CnDbpedia 图谱取得的召回率为 82.5%，略高于融入

Medical 图谱取得的 82% 的召回率。融入 Medical 图谱取得的 80.9% 的准确率，略高于融入 CnDbpedia 图谱取得的 80.3%。从图谱数据的分析来看，通过 Medical 图谱的信息辅助之后，其医学实体中融入了更多三元组信息辅助了医学实体的判断，而 CnDbpedia 图谱通过加入了一些百科语料的知识信息，对于人群，时间这样的文本中非与强烈医学相关的实体起到了更多的辅助作用。

为了进一步验证上述推断的合理性，将 CCKSNER 数据中融入 CnDbpedia 图谱的模型与 CMQANER 数据集融合 CnDbpedia 图谱和 Medical 图谱的模型对于每个医学实体的识别精度进行单独的分析，并于基于字粒度的 BiLSTM-CRF 模型的结果进行单独的比较，结果如表 5.6, 5.7, 5.8 所示。

表 5.6 融入 CnDbpedia 图谱之后在 CCKSNER 数据集的效果表

测试指标	Precision	Recall	F1 值	F1 值 (BiLSTM_CRF)
解剖部位	0.842	0.820	0.831	0.690
影像检查	0.830	0.770	0.799	0.745
实验室检验	0.563	0.655	0.605	0.628
疾病和诊断	0.752	0.773	0.762	0.741
药物	0.948	0.920	0.933	0.937
手术	0.660	0.663	0.661	0.667
症状	0.948	0.967	0.957	0.941
总指标(Micro)	0.840	0.843	0.842	0.759

相比于单纯基于字的 BiLSTM-CRF 模型，CMK-BERT 既在上下文语义中进行了增强，也额外加入了领域知识信息，但从表 5.6 的结果中发现并非所有实体都发生了效果的提升。如手术实体、药物实体和实验室检查实体的识别就发生了下降，尤其是实验室检查的 F1 值从原先的 0.628 降低到了 0.605，针对这个问题去分析了 CnDbpedia 图谱的数据集，发现“实验室检查”出现在多出尾实体中，如“白细胞计数-检查分类-临床实验室检查”，“血红蛋白-检查分类-临床实验室检查”，但在原始文本里面这些头实体所对应的类型并非实验室检查，在模型中判断为实验室检查实体的错例中，也发现有一定比例有将“血红蛋白”错判为实验

室检查实体的错例，这在之前的 BiLSTM-CRF 模型中没有出现过。说明知识图谱所带来的知识噪声会导致原始文本一定程度上语义的改变，这样的改变在某些实体上反而会带来负面影响。

对于 BiLSTM-CRF 模型，CMK-BERT 模型在解剖部位实体的识别精度发生了显著的提高，从原先的 0.690 提升到了 0.831，观察到有较多“肾下腺”，“膀胱”，“肛门”一些容易在百科图谱中找到的实体，比如在三元组中包括较多“膀胱结石-就诊科室-泌尿外科”，“肛门反射-所属分类-神经电生理”这样的信息，说明百科知识图谱对该实体可以起到较大的信息补充作用，从而直接提高了识别的精度。

表 5.7 融入 CnDbpedia 图谱之后在 CMQANER 数据集的效果表

测试指标	Precision	Recall	F1 值	F1 值 (BiLSTM_CRF)
疾病(disease)	0.802	0.808	0.805	0.784
人群(crowd)	0.899	0.900	0.900	0.848
症状(symptom)	0.806	0.738	0.771	0.766
身体部位(body)	0.751	0.742	0.746	0.694
治疗方法 (treatment)	0.752	0.540	0.629	0.623
时间(time)	0.805	0.543	0.648	0.531
药物(drug)	0.703	0.516	0.595	0.515
范围(feature)	0.997	0.984	0.990	0.982
生理机能 (physiology)	0.950	0.779	0.856	0.868
检测(test)	0.723	0.601	0.656	0.636
科室 (department)	0.802	1.000	0.890	0.889
总指标(Micro)	0.803	0.825	0.814	0.741

表 5.8 融入 Medical 图谱之后在 CMQANER 数据集的效果表

测试指标	Precision	Recall	F1 值	F1 值 (BiLSTM_CRF)
疾病(disease)	0.800	0.810	0.805	0.784
人群(crowd)	0.866	0.823	0.844	0.848
症状(symptom)	0.802	0.740	0.770	0.766
身体部位(body)	0.699	0.690	0.694	0.694
治疗方法 (treatment)	0.740	0.581	0.651	0.623
时间(time)	0.735	0.484	0.584	0.531
药物(drug)	0.713	0.557	0.625	0.515
范围(feature)	0.998	0.966	0.981	0.982
生理机能 (physiology)	0.949	0.860	0.902	0.868
检测(test)	0.706	0.586	0.640	0.636
科室 (department)	0.811	0.997	0.894	0.889
总指标(Micro)	0.809	0.820	0.814	0.741

表 5.7 与表 5.8 是两个图谱在 CMQANER 数据集上取得的最好效果。可以发现当融入医疗图谱之后，疾病，药物，生理机能等实体的识别精度发生了显著的提高，尤其是药物的识别精度，直接从 0.595 上升到了 0.625，而融入 CnDbpedia 图谱的结果仅仅是提高到了 0.595，在分析图谱三元组数据中也能发现，在 CnDbpedia 仅仅有非常小的比例的药物实体，对于药物实体识别的增强作用不如使用 Medical 图谱明显。而使用 CnDbpedia 图谱效果提高较大的是人群，身体部位，时间，范围等实体，例如人群实体的 F1 值直接从 0.848 提升到了 0.900，而在加入 Medical 图谱之后对于人群实体的 F1 值没有显著影响，在 CnDbpedia 图谱也能发现较多与人物相关的实体，说明了该图谱的直接提升效果。

此外，对于两类数据融入不同知识图谱后的收敛速度进行观察，如图 5.12 与图 5.13 所示，对于 CCKSNER 数据发现使用 CnDbpedia 图谱和 Medical 图谱都第二轮训练时达到了 0.82 的 F1 值，而 Medical 图谱在接下来几轮训练中逐渐地趋

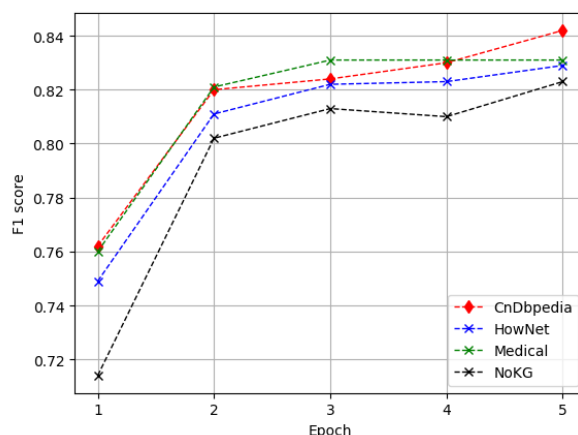


图 5.12 CCKSNER 数据收敛速度图

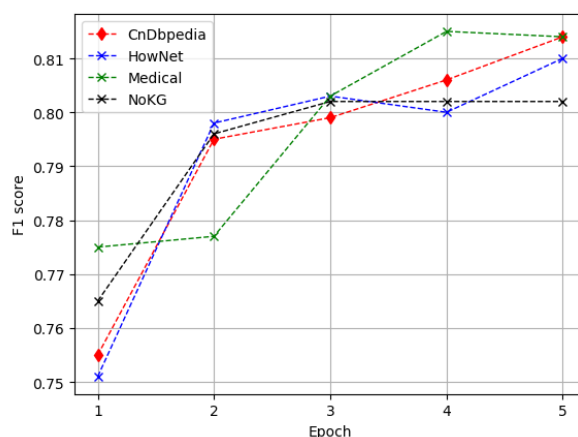


图 5.13 CMQANER 数据收敛速度图

于稳定，和融入 HowNet 图谱的效果持平，而融入 CnDbpedia 图谱的效果随着训练轮数的增加还在逐步的增长，这与 CnDbpedia 图谱本身的规模也有关，往往需要较多的训练轮数才能够训练充分。

而对于 CMQANER 数据集，在第 5 轮的训练中，融入 Medical 图谱的结果与融入 CnDbpedia 图谱的结果相当，融入 Medical 图谱的结果已经逐渐趋于稳定，但是融入 HowNet 图谱和融入 CnDbpedia 图谱还有继续上涨的趋势，这与 CMQANER 数据集特性有关，由于 CMQANER 数据集是线上医学问答数据，采用的用语风格和日常用语相对接近，其实体类别中也包含了日常用语会使用到的实体，比如时间实体，人群实体，对于这些实体识别精度的提高，CnDbpedia 图谱和 HowNet 图谱中的三元组信息会是很好的补充。

5.3 本章小结

文章使用了已经预训练好的 BERT 模型，在两类医学数据集上进行微调，并验证了医学词典在 BERT-CRF 模型上的提升效果。并以三类知识图谱为领域知识的载体，基于软位置编码和可见矩阵的思想，设计了适用于中文医学实体识别的 CMK-BERT 模型，在进行数据转换之后，通过实验分析了两类医学数据集在融入不同知识图谱之后得到不同表现与其原因，在单个医学实体上与之前基于字粒度的得到的 BiLSTM-CRF 进行对比，分析不同知识图谱融入后，在不同医学实体上识别效果提升和效果下降的原因，在融入图谱得到性能提升的同时也发现存在知识噪声的现象。

第六章 总结与展望

6.1 本文总结

命名实体识别是自然语言处理中的一项基础但极为重要任务，识别精度的好坏影响了下游任务如关系抽取，智能问答，知识图谱构建等模型效果的上限。此前更多的工作集中在了开放领域的实体识别，而医学这个垂直领域的语料不同于大众语料文本，存在着繁多且不规则的医学专业术语，在中文医学文本中还存在着词边界难以识别的问题，若要达到一个好的识别效果往往需要大量的标注语料进行模型的训练。但中文医学实体识别的数据集标注成本往往非常高，目前标注数据缺乏是一个客观存在的问题，如何在现有的标注语料下最大程度的利用辅助信息去提高中文医学实体识别的效果是目前巨大的挑战。

本文在医学电子病历数据集和医学线上问答数据集两种不同的中文医学文本上进行探索，首先使用开放域实体识别中两个经典模型 **Linear-Chain-CRF** 与 **BiLSTM-CRF**，发现 **Linear-Chain-CRF** 模型通过添加医学词典特征与医学常识特征的方法尽管在微平均指标上没有超过 **BiLSTM-CRF** 模型，但在某些实体的识别精度上可以达到超过深度学习模型的效果。这推动文本去探索如何高效去融入医学词典与领域知识这样的辅助信息去提高模型的效果。

本文通过大众健康网站，医学期刊等文本中爬取得到医学词典与医学文本语料。并基于医学词典设计了上下文词典特征与词界特征作为字的辅助信息，并基于医学词典进行分词将词典信息融入到词粒度中，并通过单粒度增强实验验证了这些辅助信息对于模型的效果提升。为了探索字词融合方法所带来的进一步效果提高，文本基于词是字的辅助信息或者字与词的地位并列两个不同的先验假设设计了字词串联法与字词并联法，包含了字词直接串联法，字词间接串联法与字词前置并联法，字词后置并联法。通过两类医学数据集的表现既验证了融合方法能够影响词典的融入效果，也得到了不同数据集对于融合方式差异的敏感性不同的结论。

本文也使用了目前较为流行的 BERT 预训练模型，在两份医学数据集上进行微调，并基于 BERT-CRF 验证了医学词典在预训练模型上的效果提升。此外，使用了三类知识图谱作为领域知识信息的载体，设计了一套适用于中文医学实体识别的 CMK-BERT 模型，在经过适当数据预处理后将不同的知识图谱分别融入两类医学数据集，发现融入领域知识信息之后，总体识别效果都发生了提升，但对于单个医学实体进行分析时，发现某些实体在融入知识后识别效果反而下降，即融入领域知识提高总体效果的同时也会引入一定程度的知识噪声，影响部分实体的识别。

本文融入辅助信息的方法具备一定的通用性，除了对于医学领域，在其他的实体识别领域甚至在自然语言处理的多个领域都会有标注信息缺失的问题，融入辅助信息以提高模型效果的思路在处理其他领域问题时有一定的借鉴意义。此外，选用的百科型的通用知识图谱也能对于医学这个垂直领域的任务有一定的提升，说明了该模型具有一定的借鉴意义，在其他垂直领域进行适当的模型改造也能在一定程度上提高特定领域模型的效果，具有较高的应用价值。

6.2 未来展望

1) 预训练模型在自然语言处理各个任务中均取得了极为突出的成果，本文提出的 CMK-BERT 模型是将知识图谱融入到中文医学命名实体识别的任务中，然而经过实验发现，知识图谱融入预训练模型，会改变原始文本结构，对某些疾病实体的识别反而带来知识噪音。如何减少知识噪声是需要探索的一个方向。除此之外，能否将词典信息更加高效融入 BERT 模型中，进一步提高模型精度是未来的探索方向。

2) 本文提出的融入辅助信息的中文命名实体识别模型虽然在识别实体精确度方面有所提升，但随之带来的是模型容量的提升和推理时间的增长，如在字词前置并联法中会遇到难以并行化的问题。如何在提升模型指标的基础上，有效地控制模型容量以及推理速度，也是需要解决的一个问题。

参考文献

- [1] GRISHMAN R, SUNDHEIM B. Message Understanding Conference-6: A Brief History[C] // Proceedings of 16th International Conference on Computational Linguistics. 1996 : 466 – 471.
- [2] SIGNLL.Language-Independent Named Entity Recognition [EB/OL]. [2018-11-7].<http://www.clips.uantwerpen.be/con112003/ner/>.
- [3] 第一届中国中文信息学会汉语处理评测(CIPS - CLPE)暨第四届国际中文自然语言处理 Bakeoff[EB /OL].[2010 -01 -11] . <http://www.china-language.gov.cn/bakeoff08/>.
- [4] 刘浏, 王东波.命名实体识别研究综述[J].情报学报, 2018, 37(3): 329-340.
- [5] 郑强, 刘齐军, 王正华, 等.生物医学命名实体识别的研究与进展[J].计算机应用研究, 2010, 27(3):811-815+ 832.
- [6] XIA Y, WANG Q. Clinical named entity recognition: ECUST in the CCKS-2017 shared task 2[C]//CEUR Workshop Proceedings. Chengdu, China: the Technical Committee on Language and Knowledge Computing of The Chinese Information Processing Society of China, 2017, 1976: 43-48.
- [7] ANDRADE, M. AND VALENCIA, A., Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families, Bioinformatics, 14 (7): 600-607, 1998.
- [8] MCDONALD R, PEREIRA F. Identifying gene and protein mentions in text using conditional random fields[J]. BMC Bioinformatics, 2005, 6 (1) :S6.
- [9] 龙光宇, 徐云. CRF 与词典相结合的疾病命名实体识别[J]. 微型机与应用, 2017 (21): 51-53.
- [10] FUKUDA K, TSUNODA T, TAMURA A, et al. Toward information extraction: identifying protein names from biological papers[C]//Pac symp biocomput. 1998,

- 707(18): 707-718.
- [11] LIN Y, TSAI T, CHOU W, et al. A maximum entropy approach to biomedical named entity recognition[C]//Proc of the 4th ACM SIGKDD Workshop on Data Mining in Bioinformatics.2004:56-61.
- [12] WEI Q, CHEN T, XU R, et al. Disease named entity recognition by combining conditional random fields and bidirectional recurrent neural networks[J]. Database, 2016, baw140: 1-8.
- [13] HU J, SHI X, LIU Z, et al. HITSZ CNER: A hybrid system for entity recognition from Chinese clinical text[C]//CEUR Workshop Proceedings. Chengdu, China: the Technical Committee on Language and Knowledge Computing of The Chinese Information Processing Society of China., 2017: 25-30.
- [14] 刘浏, 王东波. 命名实体识别研究综述 [J]. 情报学报, 2018, 37(3).
- [15] WEI Q, CHEN T, XU R, et al. Disease named entity recognition by combining conditional random fields and bidirectional recurrent neural networks[J]. Database, 2016, baw140: 1-8.
- [16] ROBERTS K, SHOOSHAN S E, RODRIGUEZ L, et al. The role of fine-grained annotations in supervised recognition of risk factors for heart disease from EHRs[J]. Journal of Biomedical Informatics, 2015, 58(S): S 111-S 119.
- [17] 杨锦锋, 于秋滨, 关毅, 等. 电子病历命名实体识别和实体关系抽取研究综述 [J]. 自动化学报, 2014, 40(8): 1537-1562.
- [18] SETTLES B. ABNER: An open source tool for automatically tagging genes, proteins and other entity names in text [J]. Bioinformatics, 2005, 21(14): 3191–3192.
- [19] LIU K, HU Q, LIU J, et al. Named Entity Recognition in Chinese Electronic Medical Records Based on CRF[C]//2017 14th Web Information Systems and Applications Conference (WISA). Piscataway, NJ: IEEE Press, 2017: 105-110.
- [20] WANG X, YANG C, GUAN R. A comparative study for biomedical named

- entity recognition[J]. International Journal of Machine Learning and Cybernetics, 2018, 9(3): 373-382.
- [21] LI L, FAN W, HUANG D, et al. Boosting performance of gene mention tagging system by hybrid methods[J]. Journal of biomedical informatics, 2012, 45(1): 156-164.
- [22] TORII M, HU Z, WU C H, et al. BioTagger-GM: a gene/protein name recognition system[J]. Journal of the American Medical Informatics Association, 2009, 16(2): 247-255.
- [23] SCHWARTZ A S, HEARSTMA. A simple algorithm for identifying abbreviation definitions in biomedical text[M]//Biocomputing 2003. 2002: 451-462.
- [24] CHAN S K, LAM W, YU X. A cascaded approach to biomedical named entity recognition using a unified model[C]//Seventh IEEE International Conference on Data Mining (ICDM 2007). IEEE, 2007: 93-102.
- [25] HINTON G E, SALAKHUTDINOV R R. Reducing the dimensionality of data with neural networks[J]. science, 2006, 313(5786): 504-507.
- [26] ALMGREN S, PAVLOV S, MOGREN O. Named entity recognition in swedish health records with character-based deep bidirectional lstms[C]//Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM2016). 2016: 30-39.
- [27] ZHAO Z, YANG Z, LUO L, et al. Disease named entity recognition from biomedical literature using a novel convolutional neural network[J]. BMC medical genomics, 2017, 10(5): 75-83.
- [28] HUANG Z, XU W, YU K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv preprint arXiv:1508.01991, 2015.
- [29] PETERS M E, NEUMANN M, IYYER M, et al. Deep contextualized word representations[J]. arXiv preprint arXiv:1802.05365, 2018.
- [30] DEVLIN J, CHANG M, LEE K, et al. BERT: Pre-training of Deep Bidirectional

- Transformers for Language Understanding[J]. 2018.
- [31] 徐国海.面向中文医疗文本的命名实体识别研究[D].华东师范大学, 2019.
- [32] SUN Y, WANG S, LI Y, et al. Ernie: Enhanced representation through knowledge integration[J]. arXiv preprint arXiv:1904.09223, 2019.
- [33] LIU W, ZHOU P, ZHAO Z, et al. K-BERT: Enabling Language Representation with Knowledge Graph[C]//AAAI. 2020: 2901-2908.
- [34] ZHANG N, JIA Q, YIN K, et al. Conceptualized Representation Learning for Chinese Biomedical Text Mining[J]. arXiv preprint arXiv:2008.10813, 2020.
- [35] LAFFERTY J, MCCALLUM A, Pereira F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[J]. 2001.
- [36] LAFFERTY J, MCCALLUM A, Pereira F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[J]. 2001.
- [37] PETERS M E, NEUMANN M, IYYER M, et al. Deep contextualized word representations[J]. arXiv preprint arXiv:1802.05365, 2018.
- [38] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training[J]. 2018.
- [39] 刘峤, 李杨, 段宏, 等. 知识图谱构建技术综述[J]. 计算机研究与发展, 2016, 53(3): 582.
- [40] 王若佳, 赵常煜, 王继民. 中文电子病历的分词及实体识别研究[J]. 图书情报工作, 2019, 63(2): 34-42.
- [41] ZHANG Q, LIU X, FU J. Neural Networks Incorporating Dictionaries for Chinese Word Segmentation[C]//AAAI. 2018: 5682-5689.
- [42] MCCALLUM A, LI W. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons[J]. 2003.
- [43] LI J, SUN A, HAN J, et al. A survey on deep learning for named entity recognition[J]. IEEE Transactions on Knowledge and Data Engineering, 2020.
- [44] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of

- words and phrases and their compositionality[C]//Advances in neural information processing systems. 2013: 3111-3119.
- [45] ZHANG Y, YANG J. Chinese ner using lattice lstm[J]. arXiv preprint arXiv:1805.02023, 2018.
- [46] XU B, XU Y, LIANG J, et al. CN-DBpedia: A never-ending Chinese knowledge extraction system[C]//International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. Springer, Cham, 2017: 428-438.