



兰州大学

## 本科毕业论文（设计）

论文题目（中文） 基于上下文特征的药物命名实体识别

论文题目（英文） Drug Named Entity Recognition  
Based on Context-Level Features

学生姓名 方 啸

指导教师 赵志立

学 院 信息科学与工程学院

专 业 计算机科学与技术

年 级 2019 级

兰州大学教务处

## 诚信责任书

本人郑重声明：本人所呈交的毕业论文（设计），是在导师的指导下独立进行研究所取得的成果。毕业论文（设计）中凡引用他人已经发表或未发表的成果、数据、观点等，均已明确注明出处。除文中已经注明引用的内容外，不包含任何其他个人、集体已经发表或未发表的论文。

本声明的法律责任由本人承担。

论文作者签名： 方啸

日 期： 2023 年 5 月 21 日

## 关于毕业论文（设计）使用授权的声明

本人在导师指导下所完成的论文及相关的职务作品，知识产权归属兰州大学。本人完全了解兰州大学有关保存、使用毕业论文（设计）的规定，同意学校保存或向国家有关部门或机构送交论文的纸质版和电子版，允许论文被查阅和借阅；本人授权兰州大学可以将本毕业论文（设计）的全部或部分内容编入有关数据库进行检索，可以采用任何复制手段保存和汇编本毕业论文（设计）。本人离校后发表、使用毕业论文（设计）或与该毕业论文（设计）直接相关的学术论文或成果时，第一署名单位仍然为兰州大学。

本毕业论文（设计）研究内容：

☒ 可以公开

☐ 不宜公开，已在学位办公室办理保密申请，解密后适用本授权书。

（请在以上选项内选择其中一项打“√”）

论文作者签名： 方啸

导师签名： 赵志立

日 期： 2023 年 5 月 21 日

日 期： 2023 年 5 月 25 日

# 基于上下文特征的药物命名实体识别

## 中文摘要

命名实体识别 (Named Entity Recognition, NER) 是对文本进行数据挖掘的重要子任务之一。在医学研究领域, 药物命名实体识别可以帮助研究者对病历数据进行统计分析并构建知识图谱。然而, 临床工作环境中的终端计算资源有限, 这意味着一些基于神经网络的 NER 方法由于训练成本高而不适用于该场景, 即使它们在实验中表现优异。为了找到适合实际工作环境的药物命名实体识别方法, 本文主要开展了以下工作:

(1) 比较了常见的计算资源友好的药物命名实体识别方法。药物命名实体识别任务中常见的低计算成本方法有三类: 词典匹配, 基于规则的 (Rule-Based) 方法, 条件随机场 (Conditional Random Field, CRF) 模型。本文从实体提取的精确水平比较该三类方法。实验结果表明, CRF 在准确性 (Precision) 和召回率 (Recall) 上相对于另外两类方法都具有一定优势。

(2) 通过改进特征工程提高了基础 CRF 模型的性能。本文从单词的上下文级别引入了基于形态学、基于语言学和基于句法结构的三大类若干特征, 通过提供更全面的上下文信息来优化 CRF 模型在药物命名实体识别任务中的表现。实验结果显示, 在这三类特征中, 基于语言学的特征能够显著地提升识别结果对药名实体的召回率, 较基础模型提高了约 25%, 对增强 CRF 模型在该任务中表现的有效性最强。

**关键词:** 命名实体识别; 条件随机场; 特征工程

# Drug Named Entity Recognition

## Based on Context-Level Features

### Abstract

Named Entity Recognition (NER) is one of the important subtasks of text mining. In the field of medical research, NER of drugs can help researchers perform statistical analysis on medical records and build knowledge graphs. However, the limited computing resources in clinical work environments mean that some NER methods based on neural networks with high training costs are not suitable for this scenario, even if they perform well in experiments. To find a drug NER method that is suitable for practical working environments, this paper mainly conducted the following work:

(1) This paper compares common computing resource-friendly drug NER methods. Three low-cost methods commonly used in drug NER tasks are dictionary matching, rule-based methods, and Conditional Random Field (CRF) models. This paper compares the three methods based on the F1-Score of entity extraction. The experimental results shows that the CRF model had certain advantages in precision, accuracy, and recall compared to the other two methods.

(2) This paper introduces several features based on morphology, linguistics, and syntax at the context level of words, which optimizes the performance of the CRF model in drug NER tasks by providing more comprehensive contextual information. Experimental data shows that among the three types of features, linguistics-based features significantly improve the recall rate of entity recognition for drug names, which is about 25% compared to the basic model, and were the most effective in enhancing the performance of the CRF model in this task.

**Keywords:** named entity recognition; conditional random field; feature engineering

# 目 录

中文摘要 .....	I
Abstract .....	II
第一章 绪 论 .....	1
1.1 研究背景、目的及意义 .....	1
1.2 研究现状分析 .....	2
1.2.1 命名实体识别 .....	2
1.2.2 特征工程 .....	2
1.3 本文的主要工作 .....	3
1.4 本文的组织结构 .....	3
第二章 相关理论与技术 .....	5
2.1 序列标注 .....	5
2.2 马尔可夫随机过程 .....	6
2.3 隐马尔可夫链 .....	6
2.4 条件随机场 .....	7
2.5 BFGS 优化算法 .....	8
2.6 本章小结 .....	8
第三章 常见计算资源友好型方法的性能对比 .....	9
3.1 数据收集与预处理 .....	9
3.1.1 数据来源与结构 .....	9
3.1.2 数据预处理 .....	10
3.2 实验设计 .....	11
3.2.1 问题定义 .....	11
3.2.2 评估标准 .....	11
3.3 构建基础模型 .....	12
3.3.1 词典匹配法 .....	12
3.3.2 基于规则和贝叶斯统计的方法 .....	13
3.3.3 条件随机场 .....	15
3.4 不同模型的性能对比 .....	16
3.5 本章小结 .....	17

第四章 使用上下文级别特征优化 CRF 模型 .....	18
4.1 上下文级别特征 .....	18
4.1.1 形态学特征 .....	18
4.1.2 语言学特征 .....	19
4.1.3 句法结构特征 .....	20
4.2 对比不同特征的可用性 .....	22
4.3 本章小结 .....	22
第五章 总结与展望 .....	23
5.1 总结 .....	23
5.2 展望 .....	23
参考文献 .....	25
附    录 .....	28
致    谢 .....	31

## 图 目 录

图 2.1 序列标注任务示意图 .....	5
图 2.2 马尔可夫随机过程示意图 .....	6
图 2.3 隐马尔可夫链模型示意图 .....	7
图 2.4 不同条件随机场结构示意图 .....	8
图 3.1 病历文本数据结构示例 .....	9
图 3.2 药物命名实体占单词总体的比重示意图 .....	10
图 3.3 按词频降序排列的药物命名实体的频率密度分布 .....	13
图 3.4 不同特征和药名实体标签内部间的线性相关性 .....	14
图 3.5 不同基础模型在训练集和测试集上的表现对比 .....	17
图 4.1 使用形态学类特征的 CRF 模型与基础模型对比 .....	19
图 4.2 使用语言学类特征的 CRF 模型与基础模型对比 .....	20
图 4.3 使用三类特征的 CRF 模型与基础模型对比 .....	21
图 4.4 使用不同特征的 CRF 模型间的横向对比 .....	22

## 表 目 录

表 3.1 二元分类混淆矩阵的相关定义 .....	11
表 3.2 词典法的实验结果 .....	13
表 3.3 基于规则和贝叶斯统计方法的实验结果 .....	15
表 3.4 条件随机场基础模型的实验结果 .....	16
表 4.1 使用形态学特征的 CRF 模型的实验结果 .....	18
表 4.2 使用语言学特征的 CRF 模型的实验结果 .....	20
表 4.3 同时使用三类特征的 CRF 模型的实验结果 .....	21



# 第一章 绪 论

## 1.1 研究背景、目的及意义

随着互联网技术和信息科学的发展，数据化已经成为当前各行各业的一致趋势。海量的数据不断扩张，为了处理、管理和分析这些数据，数据科学应运而生。为了对这些数据进行管理归档，挖掘其蕴藏的价值，使用自然语言处理技术对数据展开定量分析是数据科学的一个研究重点。在医疗产业中，不同类型的多模态非结构化数据每时每刻都在源源不断地被推送进数据库中等待分析<sup>[1]</sup>。其中最常见无疑是病例文本数据。然而，病例文本除了可以被人类阅读之外，因其结构化程度低的特性，难以直接被用于数据分析。与工业领域的设备日志分析问题类似，如何对病历文本进行数据价值的挖掘无疑也是充满潜力和挑战的一个问题<sup>[2]</sup>。

命名实体识别（Named entity recognition, NER）是自然语言处理领域的一项基础任务，它的目的是识别文本中的命名实体，例如人名、地名、组织机构名等等。同时，命名实体识别还是许多下游自然语言处理任务的重要基础。如果将命名实体识别应用于处理病历文本，则可以药物名称，处方时间等实体种类为目标，从病历中提取出对应的实体信息。随后，即可以这些实体为节点，通过关系提取技术获取节点之间的语义连接，从而构建知识图谱，最终将扁平的病历文本转换成高度结构化的知识图谱。从本质上来说，命名实体识别有助于对病历文本中的信息进行结构化提取，并传递给其他下游自然语言处理任务，从而实现对病历文本的数据价值挖掘。

在将这项任务部署到实际生产环境中时，还应考虑到医疗环境中可用设备的计算性能。近年来，大规模预训练语言模型如 BERT（Bidirectional Encoder Representations from Transformers），以其高性能和多任务学习能力在自然语言处理的研究领域饱受青睐。然而大规模预训练语言模型在训练的过程中需要消耗大量的计算资源，且对优质语料库的规模要求极高<sup>[3]</sup>。在快速变化的临床工作环境下，无论是计算资源还是病历文本的数据量，都难以满足大规模预训练语言模型的需求。相反，一些规模较小的传统方法则更适合在临床医疗环境中工作，例如条件随机场（Conditional Random Fields, CRF）<sup>[4]</sup>。它们可以被轻量级地部署，学习数据并投入工作，从而允许工作人员快速地跟随数据变化调整模型。

为了有效地利用病历文本，自然语言处理技术非常关键。为了顺利展开下游自然语言处理任务，以药物名称等实体为目标的命名实体识别任务是必要的。在快速变化的临床环境中，规模较小的传统方法比大规模预训练语言模型更符合轻量级的工作需求。因此，探究使用计算资源友好型方法对病历文本进行药物命名实体识别，对挖掘海量医疗数据中的潜在价值具有重要意义。

## 1.2 研究现状分析

### 1.2.1 命名实体识别

药物名称实体提取是命名实体提取的一个特殊用例。命名实体提取则是自然语言处理中信息提取的一个子任务。一般来说，命名实体提取的目标是从非结构化文本中提取出实体。给定的实体类别随着需求和上下文有多种，常见的实体类型有人名、组织、位置、时间表达式等等<sup>[5]</sup>。

在命名实体提取任务发展的初期阶段，常用方法有建立词典进行匹配<sup>[6]</sup>，或从一般的文本中寻找规则，从而建立基于规则的模型<sup>[7]</sup>。这两种方法在特定的任务上可以有很好的表现。例如在提取国家名称实体时，可以建立一个包含所有国家名称的列表，对目标文本进行分词后，将构建的列表中的实体进行匹配。如果需要提取时间表达式，常见的一种做法是构建规则模式，通过匹配文档中形式为年月日的词组，并验证其是否为时间表达式。

随着技术的发展，概率模型逐渐被引入实体提取的解决方案之一，常见的概率模型有隐马尔可夫、最大熵和条件随机场模型<sup>[8][9]</sup>。本质上，这些模型的解决方案都是对文本进行序列标注，即将文本看作词组构成的序列，对每一个序列进行成分标注。成分可以分为若干类。当具体到实体提取任务时，成分也可以被解读为非特定类实体和特定类实体这两类。当命名实体成分被标注出来时，提取任务就完成了。对于这一类模型最大的挑战，是提取有效的特征帮助模型进行预测。

近年来，基于深度学习的命名实体识别方法也取得了新的进展。例如，Greenberg 等人提出了一种基于条件随机场和双向长短记忆循环神经网络（Bidirectional Long Short Term Memory Recurrent Neural Network）模型进行命名实体识别的新技术，可以在不相交的多个生物医学数据集中训练单个模型，从而最大程度地利用数据集中蕴含的可帮助识别实体的信息<sup>[10]</sup>。

### 1.2.2 特征工程

特征工程是机器学习领域中的一项重要任务，它指的是根据领域知识或数据分析的经验来设计和选择合适的特征，以提高机器学习模型的性能。在过去的几十年中，手动特征工程曾是机器学习研究的重点之一，尤其是在自然语言处理领域<sup>[11]</sup>。然而，由于手动特征工程需要研究者具有充分的领域知识和经验，因此对于某些任务，它可能非常耗时且具有挑战性。

随着深度学习技术的发展，自动特征工程逐渐成为研究热点。自动特征工程的目标是通过算法自动从原始数据中提取特征，以减轻手动特征工程的负担。目前，深度学习中的卷积神经网络（Convolutional Neural Network）和循环神经网络（Recurrent Neural Network）等模型已经广泛应用于自动特征工程的研究中<sup>[12][13]</sup>。然而，手动特征工程仍然是一项必要的任务，手动的构造特征可以一定程度上提高模型的可解释性，帮助用户理解模型的预测

结果。

近年来，一些研究也试图将手动特征工程与自动特征工程相结合，以提高机器学习模型的性能。例如，一些研究尝试使用深度学习模型自动提取特征，然后使用人工设计的特征来扩展这些自动提取的特征。这些方法又被称为“深度特征学习”，它们可以将自动特征工程的优势与手动特征工程的优势结合起来，从而提高机器学习模型的性能<sup>[14]</sup>。

总的来说，手动特征工程和自动特征工程各有其优缺点，它们可以相互补充，以提高机器学习模型的性能。未来的研究方向可能是将手动特征工程和自动特征工程更好地结合起来，以实现更高效特征工程方法。

### 1.3 本文的主要工作

药物命名实体识别是医学研究领域下的一类 NER 任务，可帮助研究者提取病历数据中的药名，从而进行定量分析。与普通的 NER 任务不同的是，药物命名实体识别需要考虑到临床环境中的终端计算资源有限的问题。为了找到适合临床工作中药物命名实体识别的低计算成本方法，本文主要开展了以下工作：

#### (1) 比较常见低计算成本方法在药物命名实体识别任务上的表现

针对实际临床环境中的药物命名实体识别任务，相较于基于神经网络的方法，一些低计算成本方法拥有更好的可解释性和计算资源友好性。药物命名实体识别任务中常见的低计算成本方法有三类：词典匹配，Rule-Base 方法，CRF 模型。本文将从实体提取的精确水平比较 CRF 模型、词典匹配和 Rule-Base 方法，分析并探讨 CRF 模型的表现该任务场景下存在一定优越性的原因。

#### (2) 探索上下文特征在药物命名实体识别任务中的可用性

对于命名实体识别任务，传统模型普遍把文本视为扁平的词串，仅关注单词级别的信息，句段之间的层次关系信息则会被丢失。本文拟通过使用 CRF 模型，并优化特征工程，从上下文级别引入基于形态学、基于语言学和基于句法结构的三大类若干特征，从而拓展输入信息维度并提高模型在药物命名实体识别任务中的表现。

### 1.4 本文的组织结构

本文旨在比较常见计算资源友好型方法在药物命名实体提取任务中的性能差异，选择最佳方法后进行特征工程，通过从上下文级别引入新特征对模型进行优化。具体的章节安排如下：

**第一章 绪论。**本章主要介绍了药物命名实体提取的研究背景、目的、研究现状和意义，并引出本文的研究思路 and 主要研究内容。

**第二章 相关理论知识与技术。**本章详细介绍了与药物命名实体提取相关的概念和方法，主要包括序列标注，马尔科夫链，条件随机场等。

**第三章 常见计算资源友好型方法的性能对比。**本章详细地比较和分析了词典匹配，结合基于规则和统计回归的方法，以及条件随机场，这三种对计算资源友好的命名实体提取方法在药物命名实体识别任务上的性能。

**第四章 基于上下文级别特征 CRF 模型优化。**本章从单词的上下文级别引入了形态学（Morphological），语言学（Linguistic），句法结构（Syntactic Structure）三大类特征帮助提高 CRF 模型在药物命名实体识别任务上的表现，比较了这三类特征的差异，并分析了这些差异的来源。

**第五章 总结与展望。**本章总结了本文的主要工作及分析实验结果获得的结论，并对未来相关研究方向下的挑战进行了归纳和展望。

## 第二章 相关理论与技术

本文旨在探究条件随机场模型在提取病例中的药物名称实体的表现，以及如何对基本模型进行改进。因此，本章将命名实体识别的定义开始介绍，随后由马尔可夫随机过程引出隐马尔可夫链，并阐述条件随机场模型是如何被建立起来并应用于命名实体提取的。最后，本章将介绍实验中用于优化模型的 BFGS 算法。

### 2.1 序列标注

序列标注是一种重要的自然语言处理任务，其目标是对自然语言文本中的标记序列进行识别和分类。序列标注通常用于识别实体、分词、词性标注、命名实体识别等任务。序列标注技术在信息抽取、文本分类、机器翻译等领域具有广泛的应用<sup>[15]</sup>。在序列标注中，输入文本被表示为一个序列，每个位置有一个标记，标记通常是一个预定义的标记集合中的元素。标记集合的选择通常取决于任务的性质和需求。例如，在分词任务中，标记集合通常是单词和标点符号；在词性标注任务中，标记集合通常是词性标签；在命名实体识别任务中，标记集合通常是实体类型标签。如图 2.1 所示，在这句话中，“John” 作为一个人名实体的第一个单词，“Smith” 则作为该人名实体的一个非首单词。而“New” 和“York” 则构成了一个地理位置名称实体。

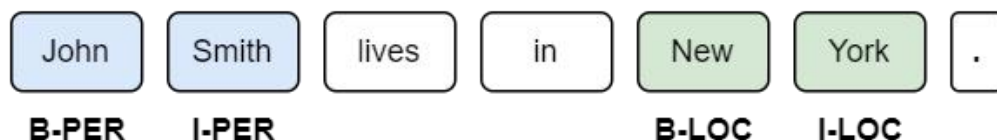


图 2.1 序列标注任务示意图

序列标注方法可以分为基于规则的方法和基于机器学习的方法两种。基于规则的方法通常使用手工编写的规则来识别标记，而基于机器学习的方法则使用机器学习算法从已标注的训练数据中学习如何识别标记。目前，基于机器学习的序列标注方法已成为主流<sup>[16]</sup>。

在基于机器学习的序列标注方法中，最常用的方法是条件随机场和循环神经网络等。其中，条件随机场是一种基于统计学习的序列标注方法，通过将标记之间的依赖关系建模为一个图来实现标记之间的约束，而循环神经网络则通过一个递归的神经网络结构来实现对序列的建模。

近年来，随着深度学习的发展，基于深度学习的序列标注方法也取得了很大的进展。例如，BiLSTM-CRF 模型将双向循环神经网络和条件随机场结合起来，有效地解决了标记之间依赖关系的建模问题，BERT 模型则使用预训练的语言模型来提取上下文信息，即使在 Zero-Shot 学习的情况下也能取得较好的结果<sup>[17]</sup>。

## 2.2 马尔可夫随机过程

马尔可夫随机过程是一种描述在时间上连续的随机演化过程的数学模型。其特点是具有马尔可夫性质，即未来状态的发展仅与当前状态相关，而与过去状态无关<sup>[18]</sup>。以图 2.2 为例，假设某人在吃饭、睡觉和玩耍三个状态间的切换遵循马尔可夫性，则只需要知道该人的当前状态就可以得到其下一个状态分别为吃饭、睡觉或玩耍的概率，而无需了解该人之前的任何状态信息。

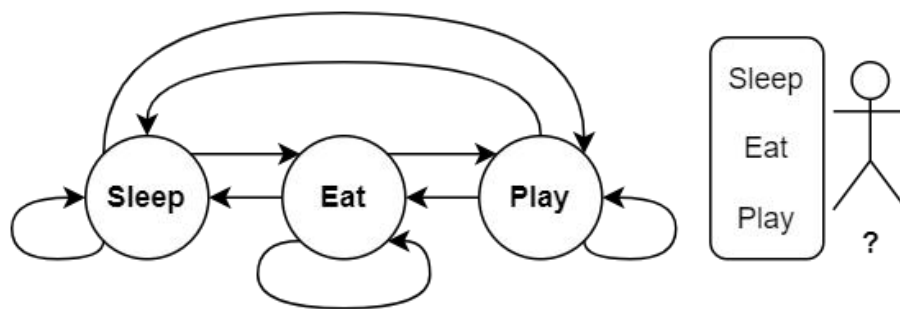


图 2.2 马尔可夫随机过程示意图

马尔可夫随机过程的基本概念是状态空间、转移概率和时间。状态空间是指随机过程所处的状态的集合。转移概率是指在一个时刻处于某一状态时，下一时刻转移到其他状态的概率。时间是指随机过程在不同状态之间转移的时间。同时，马尔可夫随机过程还可分为多种不同的类型，如连续时间马尔可夫过程和离散时间马尔可夫过程等。其中连续时间马尔可夫过程是指状态空间和时间都是连续的，通常用随机微分方程描述。而离散时间马尔可夫过程是指状态空间和时间都是离散的，通常用状态转移矩阵描述。在实际应用中，离散时间马尔可夫过程更为常见<sup>[19]</sup>。

马尔可夫随机过程在许多领域中得到了广泛应用，例如人工智能中的强化学习<sup>[20]</sup>，物理学中的布朗运动和随机漫步问题，化学中的化学反应动力学，统计学中的马尔可夫链蒙特卡罗方法等等。除了基本的马尔可夫随机过程，还有许多相关的研究课题，例如马尔可夫决策过程、隐马尔可夫模型等。这些模型都是在基本马尔可夫随机过程的基础上进行扩展和推广的。

## 2.3 隐马尔可夫链

隐马尔可夫模型，或称隐性马尔可夫模型，是一类用于描述含有隐含未知参数的马尔可夫过程的统计模型。其难点是从可观察的参数中确定该过程的隐含参数，然后利用这些参数来作进一步的分析，例如模式识别<sup>[21]</sup>。因为隐马尔可夫模型遵循马尔可夫假设，即当前时刻的状态只与其前一时刻的状态有关。而在序列标注任务中，当前时刻的状态也应该

同该时刻的前后的状态均相关。于是，很多序列标注任务中也引入了隐马尔可夫模型或其衍生方法，例如条件随机场。

在一般的马尔可夫模型中，状态对于观察者来说是直接可见的。这样状态的转换概率便是全部的参数。而如图 2.3 所示，在隐马尔可夫模型中，状态并不是直接可见的，但受状态影响的某些变量则是可见的。每一个状态在可能输出的符号上都有一概率分布。因此输出符号的序列能够透露出状态序列的一些信息。

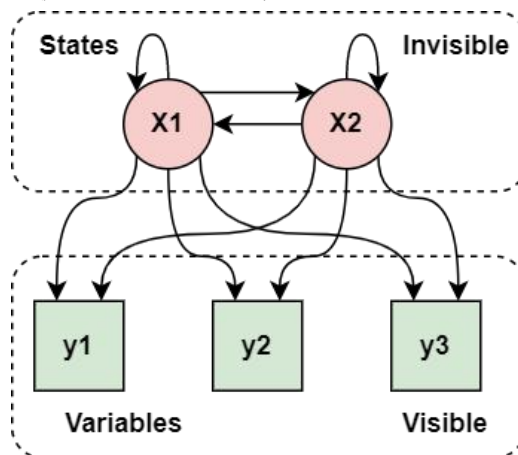


图 2.3 隐马尔可夫链模型示意图

隐马尔可夫模型有三个典型问题。预测：已知模型参数和某一特定输出序列，求最后时刻各个隐含状态的概率分布。平滑：已知模型参数和某一特定输出序列，求中间时刻各个隐含状态的概率分布<sup>[22]</sup>。解码：已知模型参数，寻找最可能的能产生某一特定输出序列的隐含状态的序列<sup>[23]</sup>。使用隐马尔可夫模型对词序列进行序列标注，本质上就属于解码任务的一种。

## 2.4 条件随机场

条件随机场，是一种鉴别式概率模型，是隐马尔可夫链模型的一种，常用于标注或分析序列资料，如自然语言文字或是生物序列。条件随机场模型又被称为概率无向图模型，它的结构被视为一张图，而图中的顶点代表随机变量，顶点间的连线代表随机变量间的依赖关系<sup>[24]</sup>。

如图 2.4 所示，条件随机场的图模型布局是可以任意给定的。但在实际应用中，一般使用链式架构。因为链式架构不论在训练、预测、或是解码上，都存在高效的计算方法可供求解问题，而任意构建的图结构可能会存在求解困难的问题<sup>[25]</sup>。

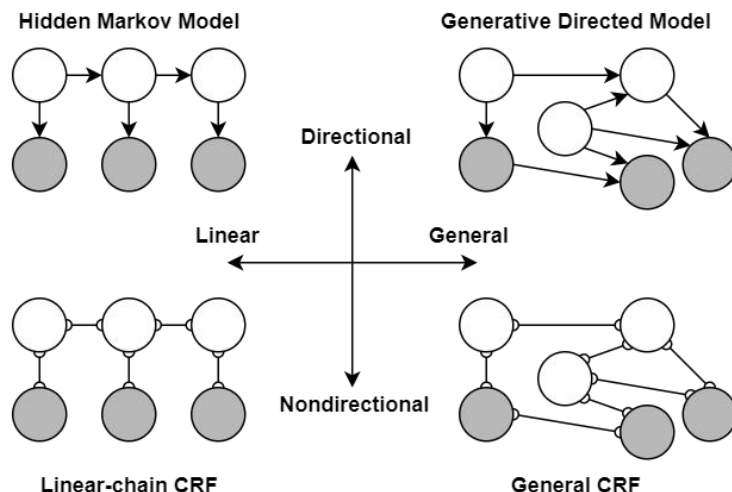


图 2.4 不同条件随机场结构示意图

在 CRF 模型中，标记序列的概率由两部分组成：特征函数和概率分布。特征函数是一个指示函数，用于表示输入序列的各种特征，例如当前单词的词性、上下文单词等。概率分布则是 CRF 模型的核心，它定义了从输入序列到标记序列的条件概率分布。而 CRF 模型的训练通常采用最大似然估计方法。该方法旨在寻找能够最大化标注序列概率的参数，从而让模型能够更准确地预测命名实体。

## 2.5 BFGS 优化算法

BFGS 算法是一种用于非线性最优化问题的凸优化算法，是由 Broyden、Fletcher、Goldfarb 和 Shanno 四个人分别提出的，故称为 BFGS<sup>[26]</sup>。它是基于拟牛顿法的一种迭代算法，优点是可以自适应地调整步长和方向，使得算法能够更快地收敛到全局最优解。BFGS 算法已经被广泛应用于各种领域，如机器学习、图像处理和信号处理等。

除了在传统的优化问题中，BFGS 算法还被广泛应用于机器学习和深度学习中。例如，L-BFGS（Low Memory BFGS）算法在深度学习中被广泛使用，以优化神经网络的损失函数。相对于随机梯度下降算法，BFGS 算法在收敛时间上具有一定优势<sup>[26]</sup>。

此外，为了提高 BFGS 算法的收敛速度，研究者们提出了一些加速 BFGS 算法的方法。例如，Chang 等人提出了一种通过引入动量加速线性收敛的 L-BFGS 算法<sup>[27]</sup>。

## 2.6 本章小结

本章按照需求-理论-实现的线索，先抛出了命名实体识别任务的概念。随后说明了马尔可夫随机过程，并引出了隐马尔可夫链和条件随机场模型，即对需求的解决方案。在后面的实验中，本文将使用基于最后介绍的 BFGS 算法的 L-BFGS 算法对模型进行优化。



### 第三章 常见计算资源友好型方法的性能对比

命名实体识别的常见计算资源友好型方法可分为几大类：基于规则和词典的方法、基于统计的方法、基于机器学习的方法等。其中，基于机器学习的方法最为先进，主要包括隐马尔可夫模型、最大熵、支持向量机、条件随机场等。这四种方法中，条件随机场为命名实体识别提供了一个特征灵活、全局最优的标注框架，但同时存在收敛速度慢、训练时间长的问题，并且对特征选取的要求较高。本章选择了词典匹配，结合基于规则和统计回归的方法，以及条件随机场三种对计算资源友好的命名实体提取方法，试验并比较它们在完成提取病历中的药物命名实体这一任务的性能效果。本章将从介绍数据集开始，并分析实际需求，确定评估标准。随后设计实验，阐述用于比较的技术方法是如何构造的。最后对这三种方法在实际任务中的表现进行比较分析和分析。

#### 3.1 数据收集与预处理

本文的实验的目标是从非结构化的病历文本数据中提取所含的药物名称。鉴于任务的领域特殊性，需要尽可能具有权威性的已标注数据集。经调研，本文选择了情境化药物事件数据集（Contextualized Medication Event Dataset, CMED）<sup>[28]</sup>。

##### 3.1.1 数据来源与结构

情境化药物事件数据集是一个用于捕获临床笔记中记录的药物变化相关背景的数据集，它是使用一种新颖的概念框架开发的，该框架将临床事件的背景组织成各种正交维度。在此过程中，作者定义了与药物变化事件相关的特定上下文方面，表征数据集，并报告初步实验的结果。CMED 注释了 500 多个临床笔记，并包含 9013 个药物实体。CMED 以“病历文件-标注文件”对的形式给出数据，并预先划分好了用于训练和评估的数据集。给定一个病历文本，通过其 ID 可以找到对应的标注文件。

```
data > train > 101-02.ann
1   T2 NoDisposition 830 838  atenolol
2   E2 NoDisposition:T2
3   T3 NoDisposition 1389 1408 hydrochlorothiazide
4   E3 NoDisposition:T3
5   T4 NoDisposition 1414 1422 atenolol
6   E4 NoDisposition:T4
7   T5 NoDisposition 758 777  Hydrochlorothiazide
8   E5 NoDisposition:T5
```

图 3.1 病历文本数据结构示例

如图 3.1 所示，标注文件中包含关于病历的多维度数据，但本实验仅关注药物实体的信息。为了标示出实体的不同，仅提供药物名称本身是不够的，因为同一单词在某些词句

中可能代表着药物，但在其他语句中却不是。此外，在同一病历文本中，同一药名还可能出现多次。因此，CMED 数据集对每一个药物命名实体还提供了其偏移量（Offset），即指明了实体在文本中出现的位置。

### 3.1.2 数据预处理

#### (1) 从标注文件中提取实体和偏移量

CMED 虽然包含实验所需的关于药物实体的信息，但其给出数据的格式依然是非结构化的，因此无法将实验结果直接与标注文件进行有意义的比较。本实验需要一个显式的评估标准用于计算评估函数。为完成这一目标，首先需要展开所有的病历文本，经过分词和停用词清理，将每一个单词看作样本。随后，根据标注文件中的信息，给定样本的分类，即“药物”与“非药物”。最后，读取所有的标注文件，提取并转化其中的药名信息为若干“病历 ID-起始偏移量-终止偏移量-药物名称”关系对。这些数据将在下一阶段帮助对病历文本的分词和构造评估标准。

#### (2) 构造评估标准

构造评估标准需要含有正确分类标签的样本集，也就是来自病历的词集。本文将病历中的每一个单词都视为样本，每个样本必为“药物”或“非药物”中的一类。注意，出现在不同位置的相同单词也应当被视为不同样本。

本文使用了来自 Spacy 的自动分词器。Spacy 是一款的工业级自然语言处理工具，可用于多种自然语言处理工作。自动分词器会先基于空格和特殊字符进行分词。随后基于英语的特殊分词逻辑进行更细致的划分，例如 “don't” 将会被划分为 “do” 和 “n't”，从而提高特征细粒度。此外，本文还去除了样本中的特殊字符，例如分词后得到的 “?” 或 “!”，从而避免进行无意义的预测。需要注意的是，CMED 并没有提供一套默认的分词机制，所以分词后的结果并不能完全覆盖所有的药物实体。经过多次探索，最终分词后的结果对来自标注数据中的药物实体达到了 99.8% 的覆盖率。

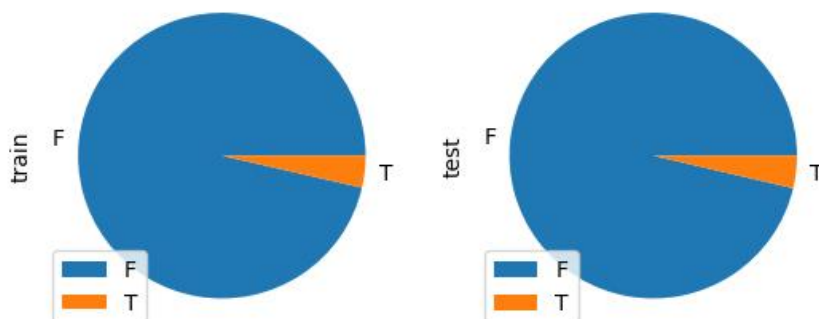


图 3.2 药物命名实体占单词总体的比重示意图

最终，本文分别得到了用于训练和评估的标准数据集。如图 3.2 所示，F 指代非药名实体的单词，T 指代药名实体单词，数据表明药名实体在总体中仅占约 3.6%。这说明数据

本身是非常不平衡的，使用传统的二分类方法将很难达到良好的预测效果。

## 3.2 实验设计

### 3.2.1 问题定义

药物命名实体识别是命名实体识别下的子任务，其应用场景的特殊性对本文解决问题的方法提出了算力限制。本实验集中关注词典匹配，结合基于规则和统计回归的方法，以及条件随机场这三类计算资源友好的方法，拟构建基础模型，测试并比较它们在 CMED 上的表现。本文将病历文本视为词序列的集合，对每一个样本，即单词进行二元分类。候选的类别为“药物”和“非药物”。如此即可将抽象的实体识别任务转化为具体且可进行评估的分类任务。在获得三类基础模型在标准数据集上关于该分类任务的表现后，本文将进一步探讨和分析不同模型表现差异的原因。

### 3.2.2 评估标准

本文将使用 Macro-F1 作为该实验的主要评估指标，并使用 Micro-F1，即精确度作为次评估指标帮助参考。为引入 Macro-F1 和 Accuracy 的定义，首先需引入混淆矩阵（Confusion Matrix）的相关定义：

表 3.1 二元分类混淆矩阵的相关定义

Confusion Matrix		Actual	
		Positive	Negative
Prediction	Positive	TP	FP
	Negative	FN	TN

如表 3.1 所示，在分类任务中，基于一特定目标类别：给定样本的真实（Actual）标签，定义为该类别的样本为阳性（Positive），非该类别的样本为阴性（Negative）；给定样本的预测（Prediction）结果，定义结果为目标类别的样本为阳性（Positive），非目标类别的样本为阴性（Negative）。TP 代表 True Positive，代表真实值为阳性的样本被正确地预测为阳性的总数。FP 代表 False Positive，代表真实值为阳性的样本被错误地预测为阴性的总数。FN 代表 False Negative，代表真实值为阴性的样本被错误地预测为阳性的总数。TN 代表 True Negative，代表真实值为阴性的样本被正确地预测为阴性的总数。

基于以上定义，可定义 Macro-F1 指标的计算公式为：

$$Macro\ F1 = \frac{2PR}{P + R} = \frac{2TP}{2TP + FP + FN} \quad (1)$$

Accuracy（Micro-F1）指标的计算公式则为：

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (2)$$

Micro-F1 和 Accuracy 都兼顾了 Precision 和 Recall 的表现。在各类样本量一致的情况下，Micro-F1 和 Accuracy 相等。而当样本量不平衡时，Accuracy 会更容易受到模型关于拥有更多样本的类别的表现的影响<sup>[29]</sup>。

在本文构建标准数据集的过程中，已经发现该数据集是极度不平衡的，药名实体仅占总体的 4%。但本文的最终目标是构建模型帮助实体提取，药名实体本身才是关注的重点。所以，Macro-F1 在该实验中较 Accuracy 具有更好的参考价值。

此外，为了进一步更具体地比较不同方法在该实验中的表现差异，并探讨这些差异的来源，本文还会引用准确率（Precision）和召回率（Recall）帮助展开分析。

Precision 和 Recall 的公式为：

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

直观来看，对于从病历文本中提取药物命名实体这一任务，Recall 描述了模型提取出的结果对所有应该被标注为药名的实体的覆盖情况，而 Precision 则表现了模型提取结果的正确率。

### 3.3 构建基础模型

#### 3.3.1 词典匹配法

词典法是一种比较常用的命名实体识别方法，其基本思想是利用预先构建的词典来匹配文本中的实体。具体来说，词典法通过构建包含实体名称的词典，然后在文本中扫描词典中的实体，匹配文本中的实体名称，最终识别出命名实体。该方法的优点在于准确率较高，尤其是对于已知领域的命名实体识别效果更佳。然而，该方法也具有无法改善的缺陷：

(1) 无法识别没有出现在词典中的实体。

(2) 无法利用上下文信息，对于一词多义的情况会出现错误匹配。

(3) 当词典本身出现噪声，即标注数据有误时，匹配也会随之出现错误，方法的鲁棒性较差。

遵循着词典法的基本思路，本文基于训练集构建了一个基础词集。基础词集中的每个词都是一个药名，对于由多个单词构成的药名，本文采用一个折中的方案，将其拆解成若干个单独的字词。

考虑到不同的大小写会影响同一拼写的语义，自然语言处理中常用的 uncased 化处理，

即小写化处理,可能并不适合该任务,本文在基础词集的基础上分别保留(cased)和去除(uncased)了单词的大小写,构建了两个不同的词典。cased 词典含有单词 1534 个,uncased 词典含有单词 1030 个。如下是两个词典的词频统计:

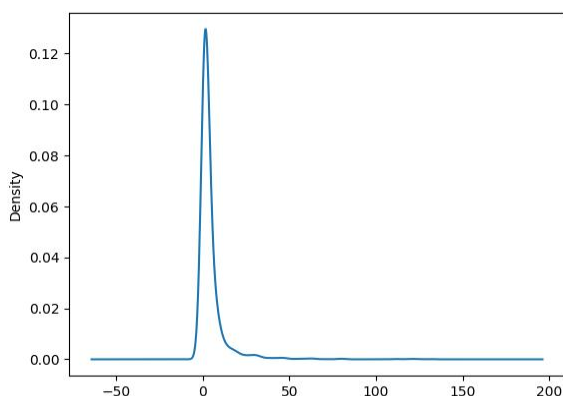


图 3.3 按词频降序排列的药物命名实体的频率密度分布

如图 3.3 所示,可以看到药名词汇的总体分布极端右偏,出现频率最高的前 20 个词占据了约总体的 95%,这同时也意味着其他药名仅占总体的 5%,生僻药名实体非常多。

使用两个词典分别对训练集和测试集进行了匹配,结果如下:

表 3.2 词典法的实验结果

data	dict	Macro-F1	Accuracy	Recall (is drug)	Precision (is drug)
Train	uncased	0.72	0.92	0.98	0.31
	cased	0.76	0.94	0.98	0.39
Dev	uncased	0.69	0.92	0.85	0.29
	cased	0.72	0.94	0.79	0.35

从表 3.2 中数据可以看出, cased 词典在总体表现上较 uncased 词典都具有优势,只有在关于训练集的 Recall 上逊于 uncased 词典,这可能是因为在训练集中的部分药名实体以不同的大小写拼写出现在了测试集中。比较各项指标,显然词典法在 Precision 指标上的表现过低。这可能是因为在同一单词在不同语境下会表现出不同的意思,词典法不区别上下文,所以额外将一些在特定语境下不是药物名称的词语也识别为了目标实体。

总体而言,词典法无需训练模型,处理速度快。在只需要关注大类下的特定实体且掌握了充足的预标注数据时,词典法可作为一种恰当的选择。

### 3.3.2 基于规则和贝叶斯统计的方法

基于规则的命名实体识别方法指的是基于人工设计的规则进行实体识别的方法。以人

名实体为例，一个人名通常由一个或多个名字组成，而地名通常由一个或多个位置词组成，这些规则可以通过人工制定并应用于自然语言文本中。这种方法的优点在于可解释性强，对于特定领域的命名实体举有很好的适应性，但其同时也需要大量的先验知识。

对于药物名称的构造，本章假设模型不具有充足的先验知识，但可以使用机器学习的方法自动生成“规则”。规则的本质是样本类别关于其各类属性特征的概率分布。对于药名，具体的分布是未知的，但可以构造特征，使用机器学习的方法近似地拟合这些分布。

贝叶斯模型是一类用于概率推断和机器学习的常见方法，本文将使用朴素贝叶斯方法拟合这些概率分布。朴素贝叶斯是基于贝叶斯定理和特征条件独立假设的分类器方法。朴素贝叶斯模型假设它使用的每个特征在给定某个类的情况下有条件地相互独立。贝叶斯定理的定义为：

若事件  $A$ ,  $B$  相互独立，则

$$P(A | B) = \frac{P(B | A) \cdot P(A)}{P(B)} \quad (5)$$

其中， $P(A | B)$  为事件  $B$  发生前提下事件  $A$  发生的概率， $P(B | A)$  为事件  $A$  发生前提下事件  $B$  发生的概率， $P(A)$  为事件  $A$  的独立概率， $P(B)$  为事件  $B$  的独立概率。

本文假设各项特征总体服从正态分布，选择了高斯朴素贝叶斯模型进行训练和预测该模型使用贝叶斯定理来计算在已知数据下某个假设的概率。本文还从单词级别的细粒度入手，简单的设计了 4 个规则，经过规则运算后得到的逻辑值即可作为特征属性: [is\_digit, is\_alpha, is\_upper, is\_title]。经皮尔森相关系数分析，本文使用了热图将这些特征与标签本身，即单词是否为药名，以及特征之间的线性关联性进行了表示。

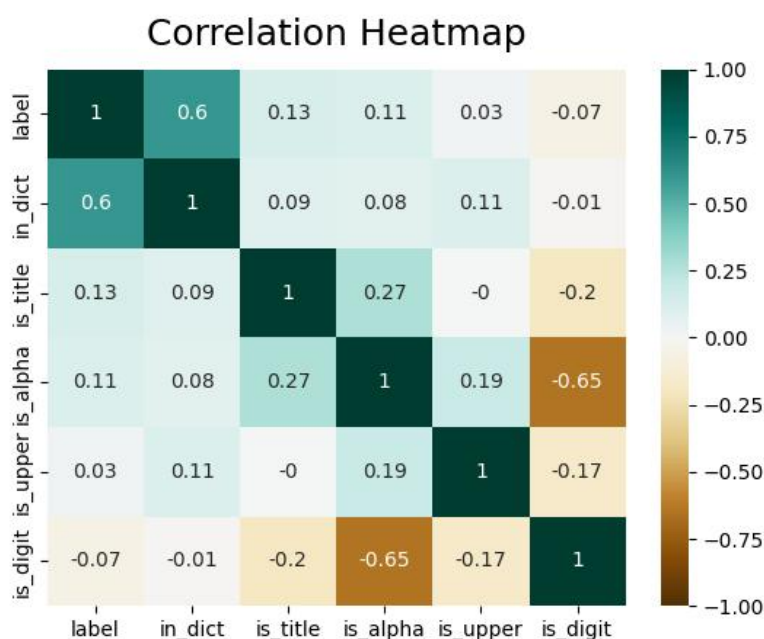


图 3.4 不同特征和药名实体标签内部间的线性相关性



如图 3.4 所示, 来自的词典信息 (`in_dict`) 显然与标签正相关性极强, 这也印证了词典法的合理性。此外, `is_title` 特征与 `is_alpha` 特征与标签也具有一定的正相关性, 这说明在训练集中的很多药名可能都是仅包含字母的单词, 不包含数字和符号, 且以大写字母开头。

本实验将高斯朴素贝叶斯模型的平滑系数设置为  $1e-9$ 。在训练集上训练后, 模型分别在训练集和测试集上的预测性能表现如下。

表 3.3 基于规则和贝叶斯统计方法的实验结果

data	Macro-F1	Accuracy	Recall (is drug)	Precision (is drug)
Train	0.80	0.96	0.97	0.46
Dev	0.77	0.95	0.79	0.43

观察表 3.3 中 Macro-F1 列可知, 基于规则和贝叶斯统计的方法相对于词典法在总体上存在优势。在药物实体提取的任务下, 该方法较词典法在对实体的召回率上较词典法并无优势, 但在准确率上有显著提高, 这是因为贝叶斯模型的假设相比于词典法的假设更加合理: 词典法假设词典中包含的词汇在任何情况下都指代药物实体, 而基于贝叶斯定理的模型认为词汇作为药物实体出现的事件遵循了一个概率分布, 事件发生的概率由多项已知条件共同影响, 而词典信息只是其中的一项。

基于规则和贝叶斯统计的方法总体上优于词典法, 但是本质上仍然没有解决单词在不同语境下的语义分歧。任意两个完全相同的单词, 在不同的上下文关系中仍然会被该模型施以相同的分类。对于该问题, 条件随机场将给出一种解决方案, 本文将在下一节引入它。

### 3.3.3 条件随机场

条件随机场是一种基于概率的序列标注方法。CRF 模型的输入通常是一段文本序列, 输出则是对应的标注序列, 例如对于一段包含人名、地名、组织机构名的文本序列, CRF 模型的输出序列将对应于每个实体的标记。对于药名实体识别这种单类实体标记, 给定样本, 即单词和其上下文, CRF 模型可选的标记为[B, I, O]。B 指该单词是药名实体的第一个词。I 指该单词是对应药名实体的非首词。O 指该单词不属于任何药名实体。

CRF 的主要思想是给定输入序列, 建立一个能够对输出序列进行联合概率建模的条件概率分布模型, 这与贝叶斯模型相似。但在贝叶斯模型中, 样本之间是独立的, 模型学习的概率分布只与样本自身的特征有关。因为 CRF 模型假设样本之间的转移遵循马尔可夫性质, 所以它可以利用上下文, 学习邻接样本的特征进行建模。显然, 对病历文本进行药名实体提取是强上下文相关的任务。例如, 给定一个未知词性的单词及其上下文, 已知该词

的前一个单词为副词时，则该单词不太可能为药名，因为副词后面一般跟的是动词，而药名应该是名词。上下文能够提供超出单词本身的信息，这也是为什么在序列标注任务中，CRF 模型往往能够取得比贝叶斯模型更好的效果。

因为本文要比较的是不同方法的基本模型性能差异，并证明 CRF 模型相对于另外两种方法的优越性，所以本实验仅使用最基础的词典信息进行预测。本实验将使用 `cased` 词典为测试集的每个一个样本构造“`in_dict`”特征。也就是说，对于一个单词，CRF 模型将根据前一个单词是否出现在词典中，该单词是否出现在词典中，后一个单词是否出现在词典中，这三类信息进行建模。本实验将 L1 正则化系数设定为 1，L2 正则化系数设定为  $1e-3$ ，迭代次数设置为 200，使用 L-BFGS 方法对训练集数据进行了拟合。并分别在训练集和测试集上进行了预测，最终的结果如下。

表 3.4 条件随机场基础模型的实验结果

data	Macro-F1	Accuracy	Recall	Precision
Train	0.84	0.98	0.62	0.77
Dev	0.78	0.97	0.47	0.75

观察表 3.4 中数据，可以发现 CRF 模型在预测的准确率上相对于另外两种方法有了显著的提高。CRF 模型在识别药物命名实体的任务上表现得更加谨慎，明显降低了非药名实体被识别为药名的比率。然而，基础模型在对药名实体的召回率上表现一般，这意味着基础模型可能需要更多的特征信息来进行学习和决策。

### 3.4 不同模型的性能对比

本文分别构建了基于词典，基于规则和贝叶斯统计，以及基于条件随机场的三类基础模型。在分别对这三者进行实验和分析后，本文将对实验数据进行横向比较。为保证对比的直观性，这里采用折线图，按照训练集结果、测试集结果进行了分组对比。



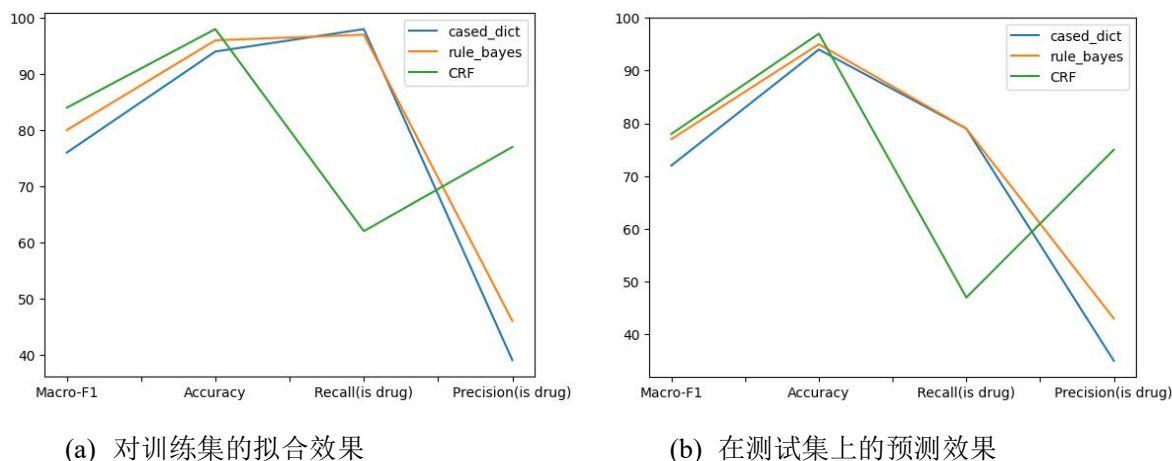


图 3.5 不同基础模型在训练集和测试集上的表现对比

如图 3.5 所示, 从 Macro-F1 指标上看, CRF 模型较另外两种方法更优, 但是 Recall 指标还有待提高。基于规则和贝叶斯统计的方法在召回率上与词典法相近, 并且在准确率上有明显的提高, 但其无法利用上下文信息, 因而正确率仍然低于 CRF 模型。总体来看, CRF 模型的表现最优, 其次是基于规则和贝叶斯统计的方法, 词典法最差。这暗示上下文信息在识别药名实体的过程中能够有效地帮助模型完成任务。

要提高词典法的表现, 只能通过扩大词典规模, 但这依然无法解决其识别正确率低的缺陷。而对于改进概率模型, 例如贝叶斯模型或 CRF 模型, 本文将通过引入更多的特征, 调整训练时的超参数以及先验概率来达到效果。词典法的改进空间有限, 而另外两种方法的表现还可以进一步提高。

### 3.5 本章小结

本章分别构建了基于词典, 基于规则和贝叶斯统计, 以及基于条件随机场的三类基础模型并进行了实验和分析。基于对实验数据的横向比较, CRF 模型在总体性能上存在优越性。这也暗示了上下文信息在命名实体识别任务中的重要性。但本实验构建的基础 CRF 模型在召回率上表现不佳。为提高基础 CRF 模型的性能, 本文将在第四章通过引入更多的特征从而为其提供更充分的上下文信息。

## 第四章 使用上下文级别特征优化 CRF 模型

经过此前的实验对比,本文验证了三类计算资源友好型方法中 CRF 模型的优越性。为了进一步挖掘 CRF 模型在药物命名实体提取任务中的潜力,接下来本文将尝试引入不同来自不同领域的特征,比较这些信息在上下文级别中帮助建模的可行性。

### 4.1 上下文级别特征

本章拟引入形态学、语言学和句法结构三大特征,每一类下含有若干类具体的子特征。

#### 4.1.1 形态学特征

形态学特征是指词语或语言单位的形态特征,包括词形、词根、词缀、屈折和构词方式等方面的特征。形态学是语言学的分支之一,研究词的构成和变化,以及词的形态特征在不同语言中的表现和规律。

研究表明,形态学特征在 NER 任务中的应用能够显著提高模型的识别准确率。在一项针对土耳其语的 NER 任务的研究中,研究人员通过探索使用形态学特征从土耳其语文本中提取信息,利用输入文本的不同特征来识别土耳其语文本中的不同命名实体,最终在测试集上取得了 91.08% 的平均 F1 得分<sup>[30]</sup>。这表明形态学特征的应用可以提高命名实体的识别准确率。此外,形态学特征在多种字母系统中都可以发挥重要作用。在一项针对希伯来语 NER 任务的研究中,研究人员使用了包括语义标注和形态学特征的模型,在测试集上达到了仅 8% 的错误率<sup>[31]</sup>。

本文从形态学的角度,构造了 9 个特征,其中部分特征已应用于基础模型: [len, is\_stop, is\_alpha, is\_digit, is\_title, is\_upper, is\_punct, contain\_upper] (详见附录源程序)。实验定义 L1 正则化系数为 1, L2 正则化系数为  $1e-3$ , 迭代次数为 200, 使用 L-BFGS 方法进行梯度下降,对训练集进行拟合。并分别在训练集和测试集上进行了预测。在仅使用形态学类特征的情况下,CRF 模型的表现如表 4.1 所示。

表 4.1 使用形态学特征的 CRF 模型的实验结果

data	Macro-F1	Accuracy	Recall	Precision
Train	0.97	1.00	0.89	0.98
Dev	0.91	0.99	0.71	0.97

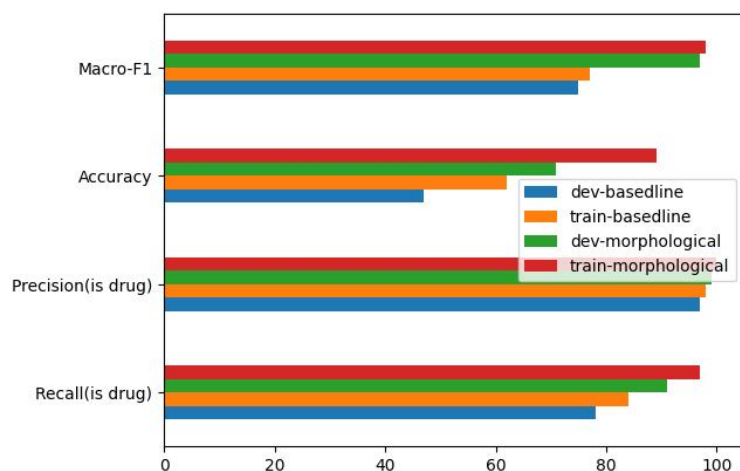


图 4.1 使用形态学类特征的 CRF 模型与基础模型对比

观察图 4.1 可知，对比基础模型，引入形态学特征后的 CRF 模型在各方面都有了明显的提高。在测试集上，改进后的模型对比基础模型在 Recall 指标上提高了 24%，在 Precision 指标上提高了 22%。这说明从上下文级别引入形态学特征可以帮助 CRF 模型更好地完成药物命名实体识别任务。

#### 4.1.2 语言学特征

在自然语言处理领域，语言学特征重点关注了单词的语义、语用、词汇等方面。语义特征包括词义、语义角色、语义关系等，词性标注就属于常见的语义特征。而语用特征包括话语功能、话语意图、话语逻辑、话语联想等，通常使用主题建模或情感分析去创建此类特征。词汇特征则包括词性、词义、词根、词缀、构词法等。据研究表明，引入关于语言学的词汇特征可以显著提高 NER 的准确率和召回率，特别是对于一些稀有实体和不常见的词汇<sup>[32]</sup>。

在设计特征的过程中，本文重点关注了词根和词缀：词根指单词的基础部分，通常是不能再分解成更小的部分，例如“booking”的词根是“book”。而词缀则是可以附加到词根或词干上的字母或单词，用来改变词的意义、词性或时态等。例如，“-s”可以添加到名词“book”后，表示复数形式的“books”。

本文从语言学的角度，构造了 10 个特征: [pref\_2, pref\_3, pref\_4, suff\_2, suff\_3, suff\_4, pos, tag, lemma, shape]（详见附录源程序）。其中，“pref\_n”类特征和“suff\_n”类特征引入了词根和词缀的思想：当两个单词具有相同的词根或词缀，它们大概率会具有相近的语义。“pref\_n”即指代单词的前 n 个字母，而“suff\_n”则指代单词的后 n 个字母，当单词长度不足 n 时，引用整个单词本身。而在构造 pos, tag, lemma, shape 四个特征的过程中，本文引入了语义特征的思想。实验定义 L1 正则化系数为 1，L2 正则化系数为 1e-3，迭代次数为 200，使用 L-BFGS 方法进行梯度下降。并分别在训练集和测试集上进行了预测。在仅使用语言学类特征的情况下，实验结果如表 4.2 所示。

表 4.2 使用语言学特征的 CRF 模型的实验结果

data	Macro-F1	Accuracy	Recall	Precision
Train	0.98	1.00	0.94	0.99
Dev	0.94	0.99	0.83	0.96

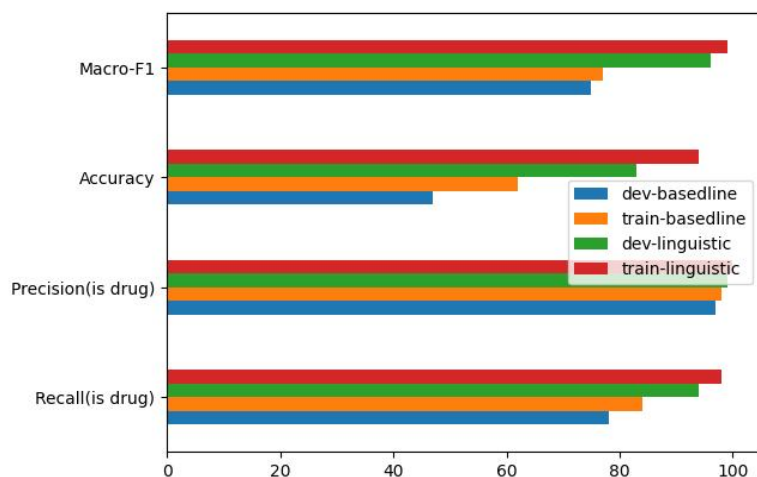


图 4.2 使用语言学类特征的 CRF 模型与基础模型对比

观察图 4.2 可知，对比基础模型和引入形态学特征后的 CRF 模型，引入了语言学特征后的 CRF 模型在召回率上有了惊人的提高，甚至强于使用 *cased* 词典的词典方法。这可能是因为药名实体间存在很多语言学相似性，例如某些主成分相同的药名会使用相同的词根或词缀。实验证明，基于上下文的语言学特征可以帮助 CRF 模型有效地学习药物命名实体的共性，从而挖掘出更多的药名实体。

#### 4.1.3 句法结构特征

句法结构特征来自句子成分之间的依赖关系，可使用句法解析技术进行提取。句法解析是一种自然语言处理技术，主要用于分析一个句子中单词之间的语法关系。在句法依存树中，每个单词都被视为一个节点，并且单词之间的语法关系被表示为树形结构，其中每个节点都有一个父节点和可能有多个子节点。句法解析的一个主要优点是它可以提供额外的信息表示句子中单词之间的关系，从而帮助模型更好地理解整个句子的含义<sup>[33]</sup>。

在命名实体识别任务中，句法解析的特征可以帮助识别句子中的命名实体，例如人名、地名、组织机构等。这是因为特定种类的实体通常与句子中的其他成分具有特定的语法关系，例如，人名通常是一个句子的主语，而地名通常是一个动词的宾语。

为了使用构造基于句法解析的特征，还需要引入句法依存树的概念：句法依存树是自然语言处理中的一种表示句子语法结构的方法。它通过表示句子中单词之间的依赖关系来

描述句子的语法结构。句法依存关系是指一个单词作为句子中其他单词的中心或修饰语，也就是它们之间的语法关系。在依存树中，每个单词都是一个节点，而单词之间的依赖关系则是节点之间的边。根据不同的句法理论和算法，句法依存树的形式可以有所不同，但通常包含一些常见的依存关系类型，比如主语、宾语、定语、状语等。

基于句法依存树的定义，本文构建了 10 个与句法分析有关的特征：[weight, position, has\_dep, n\_ancestors, n\_lefts, n\_rights, n\_conjunc, n\_child, bos, eos]（详见附录源程序）。然而，在仅使用形态学类特征的情况下，CRF 模型的表现非常糟糕，模型将所有的样本都标记为非药名实体。这意味着仅靠句法解析类特征提供的信息不足以让模型进行充分的学习，句法解析应当作为其他特征的补充。所以本文重新进行了实验，并结合了句法分析类特征与之前构建的各类特征，包括形态学特征，语义学特征，以及词典匹配，尝试将它们集成起来一同供模型学习。实验定义 L1 正则化系数为 1，L2 正则化系数为  $1e-3$ ，迭代次数为 200，使用 L-BFGS 方法进行梯度下降。并分别在训练集和测试集上进行了预测。实验结果如表 4.3 所示。

表 4.3 同时使用三类特征的 CRF 模型的实验结果

data	Macro-F1	Accuracy	Recall	Precision
Train	0.99	1.00	0.99	0.98
Dev	0.93	0.99	0.78	0.97

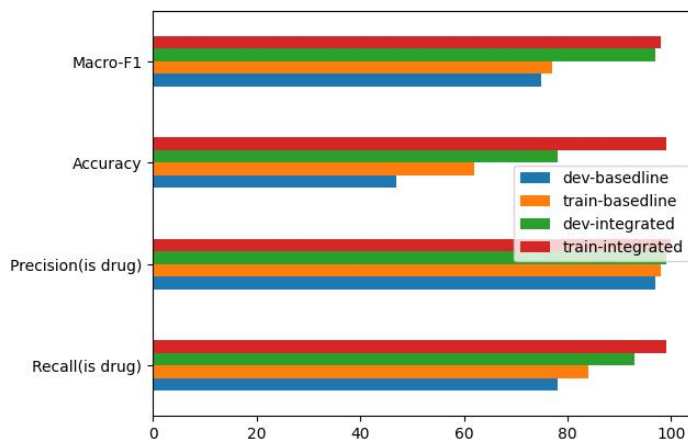


图 4.3 使用三类特征的 CRF 模型与基础模型对比

观察图 4.3 可知，在结合了基于句法解析的特征和其他各类特征后，CRF 模型在训练集上的 Recall 指标达到了最高的 98%。但是在训练集上，Recall 指标跌落至 78%。这意味着句法解析确实能提供一定的信息帮助 CRF 模型进行药名实体识别，但是可用性有限。

## 4.2 对比不同特征的可用性

为了提高 CRF 基础模型的性能表现, 本文分别引入了形态学, 语言学, 句法解析三大类特征。对于每一大类特征, 本文设计了若干具体的子特征, 并尝试只引用该类特征进行建模实验。其中, 语言学类特征对 CRF 模型的帮助最大, 单独使用句法解析的实验没有构造出有效的模型。本文还将各类特征整合, 训练了一个高度集成的 CRF 模型加入对比。如下, 本文选择了基础模型、只使用形态学特征的模型、只使用语言学特征的模型以及集成各类特征的模型, 将这 4 个模型的实验数据进行了分组对比。

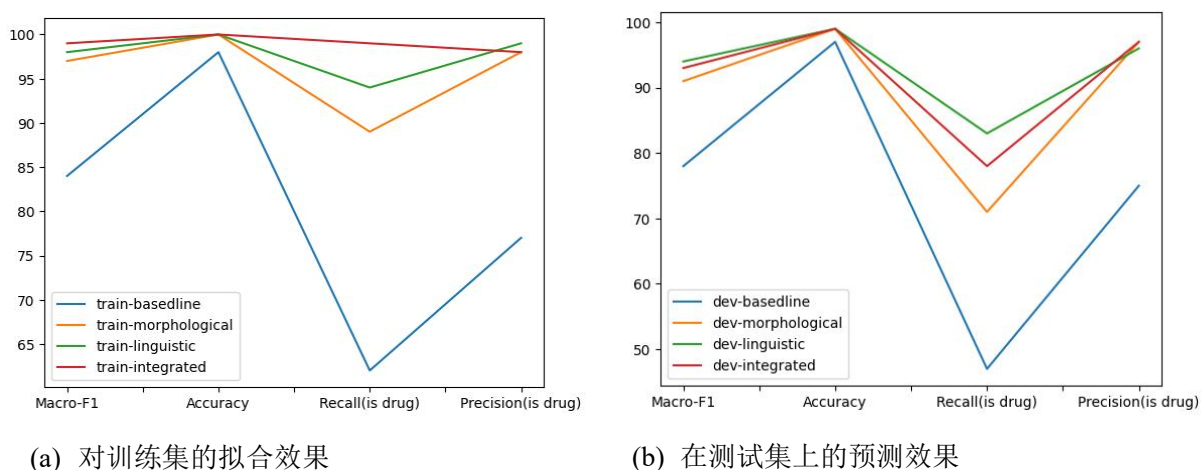


图 4.4 使用不同特征的 CRF 模型间的横向对比

观察图 4.4 可知, 上下文级别的新特征引入有效地解决了基础 CRF 模型在处理药物命名实体识别任务中召回率不足的问题。其中, 语言学类特征对提高 Recall 指标的有效性最强。这说明英文单词的语义与其语言学特征, 例如词根、词缀, 有着紧密的联系。同时, 本文发现集合了所有特征的 CRF 模型在训练集和测试集上的表现差距较大。这意味着该模型可能引入了过多的特征, 且特征之间提供的信息存在重叠, 出现了过拟合的问题。

## 4.3 本章小结

本章分别引入了形态学, 语言学, 句法解析三大类特征用于提高 CRF 模型在药物命名实体识别任务中的表现。据实验数据显示, 三类特征中, 语言学类特征能够显著地提高 CRF 模型识别药名实体的召回率, 对提高模型表现有效性最强。另外, 数据表明, 引入过多的特征可能会导致输入信息的冗余和模型的过拟合。如何进行特征选择, 在提高模型表现的同时兼顾特征的可解释性, 将会成为新的挑战。



## 第五章 总结与展望

### 5.1 总结

药物命名实体识别是分析并利用医疗数据中的重要任务，具体为从病历文本数据中提取出药物实体。该任务有助于从海量非结构化的病历文本中提取出结构化信息，为随后的下游工作，例如构建医疗知识图谱，打下坚实的基础。该任务属于命名实体识别下的子任务。

药物命名实体识别因其应用场景的特殊性，能够使用的终端计算资源有限。考虑到这一点，本文没有选择神经网络类模型，而是尝试通过一些计算资源友好的方法实现该任务。经前期调查，本文首先选择了词典匹配，结合基于规则和统计回归的方法，以及条件随机场三种方法进行比较实验，比较并分析了三者间的优缺点。

随后，本文通过在上下文级别引入了形态学，语言学，句法解析三大类特征，探究了使用上下文信息帮助条件随机场进行实体提取的可用性。根据实验结果，本文得出以下结论：

- (1) 在药物命名实体识别中，条件随机场相较于词典匹配与结合基于规则和统计回归的方法，可以有效地利用上下文语境信息，具有独特的优势。
- (2) 从单词的上下文级别引入语言学类特征，能够显著地提高 CRF 模型识别药名实体的召回率，从而有效地提高模型表现。
- (3) 对于特征维度过多的情况，模型需要合适的特征选择策略来降低其输入信息的冗余，并减少过拟合的概率。

### 5.2 展望

海量的病历数据不仅能提供关于药物名称的数据，还蕴藏着关于病人本身和诊断治疗的信息价值。对病例数据进行信息挖掘，可以帮助构建丰富的医疗领域知识图谱。在进一步使用药名数据构建和完善知识图谱的过程中，新的挑战在不断涌现：

- (1) 药物实体往往与处方，给药相关。这些事件都是时序相关的。如何才能将这些时序信息嵌入知识图谱，从而构建时序知识图谱<sup>[34]</sup>？
- (2) 不同的药物名称可能存在实体共指。当下，全球范围内有若干组织都在维护不同的开源数据库用于提供标准化的药物实体标识，例如美国国立卫生研究院维护的 RxNorm 数据库<sup>[35]</sup>。应当如何利用这些资源对提取出的药名实体进行标准化？
- (3) 在医疗领域，知识图谱主要被用于学术研究，但对临床工作的帮助有限。这是因为操作 Graph-Based 数据库有一定的技术门槛。是否可以使用语言模型帮助临床工作者们

进行图查询，降低知识图谱的使用门槛<sup>[36]</sup>？

尽管在探索和利用医疗数据的过程中仍存在众多挑战，但每克服一个关卡都意味着相关研究在利用医疗数据促进医学发展方面迈出了更大的一步。困难的前方，等待我们的将是更好的未来。



## 参考文献

- [1] Kruer, R. M., Jarrell, A. S., & Latif, A. (2014). Reducing medication errors in critical care: a multimodal approach. *Clinical pharmacology: advances and applications*, 117-126.
- [2] Spasic, I., & Nenadic, G. (2020). Clinical text data in machine learning: systematic review. *JMIR medical informatics*, 8(3), e17984.
- [3] Beam, A. L., & Kohane, I. S. (2018). Big data and machine learning in health care. *Jama*, 319(13), 1317-1318.
- [4] Chalapathy, R., Borzeshi, E. Z., & Piccardi, M. (2016). Bidirectional LSTM-CRF for clinical concept extraction. *arXiv preprint arXiv:1611.08373*.
- [5] Sharnagat, R. (2014). Named entity recognition: A literature survey. *Center For Indian Language Technology*, 1-27.
- [6] Quimbaya, A. P., Múnera, A. S., Rivera, R. A. G., Rodríguez, J. C. D., Velandia, O. M. M., Peña, A. A. G., & Labbé, C. (2016). Named entity recognition over electronic health records through a combined dictionary-based approach. *Procedia Computer Science*, 100, 55-61.
- [7] Chiticariu, L., Krishnamurthy, R., Li, Y., Reiss, F., & Vaithyanathan, S. (2010, October). Domain adaptation of rule-based annotators for named-entity recognition tasks. In *Proceedings of the 2010 conference on empirical methods in natural language processing* (pp. 1002-1012).
- [8] Morwal, S., Jahan, N., & Chopra, D. (2012). Named entity recognition using hidden Markov model (HMM). *International Journal on Natural Language Computing (IJNLC) Vol, 1*.
- [9] Ahmed, I., & Sathyaraj, R. (2015). Named entity recognition by using maximum entropy. *International journal of database theory and application*, 8(2), 43-50.
- [10] Greenberg, N., Bansal, T., Verga, P., & McCallum, A. (2018). Marginal likelihood training of BiLSTM-CRF for biomedical named entity recognition from disjoint label sets. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 2824-2829).
- [11] Scott, S., & Matwin, S. (1999, June). Feature engineering for text classification. In *ICML* (Vol. 99, pp. 379-388).
- [12] Li, M. A., Han, J. F., & Yang, J. F. (2021). Automatic feature extraction and fusion recognition of motor imagery EEG using multilevel multiscale CNN. *Medical & Biological Engineering & Computing*, 59(10), 2037-2050.
- [13] Jun, K., Lee, D. W., Lee, K., Lee, S., & Kim, M. S. (2020). Feature extraction using an RNN autoencoder for skeleton-based abnormal gait recognition. *IEEE Access*, 8, 19196-19207.
- [14] Ding, S., Lin, L., Wang, G., & Chao, H. (2015). Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 48(10), 2993-3003.

- [15] Nguyen, N., & Guo, Y. (2007, June). Comparisons of sequence labeling algorithms and extensions. In *Proceedings of the 24th international conference on Machine learning* (pp. 681-688).
- [16] He, Z., Wang, Z., Wei, W., Feng, S., Mao, X., & Jiang, S. (2020). A survey on recent advances in sequence labeling from deep learning models. *arXiv preprint arXiv:2011.06727*.
- [17] Tsai, H., Riesa, J., Johnson, M., Arivazhagan, N., Li, X., & Archer, A. (2019). Small and practical BERT models for sequence labeling. *arXiv preprint arXiv:1909.00100*.
- [18] Stroock, D. W. (2013). An introduction to Markov processes (Vol. 230). Springer Science & Business Media.
- [19] Dynkin, E. B. (1969). Boundary theory of Markov processes (the discrete case). *Russian Mathematical Surveys*, 24(2), 1.
- [20] Ding, Z., Huang, Y., Yuan, H., & Dong, H. (2020). Introduction to reinforcement learning. *Deep Reinforcement Learning: Fundamentals, Research and Applications*, 47-123.
- [21] Eddy, S. R. (2004). What is a hidden Markov model?. *Nature biotechnology*, 22(10), 1315-1316.
- [22] Samanta, O., Bhattacharya, U., & Parui, S. K. (2014). Smoothing of HMM parameters for efficient recognition of online handwriting. *Pattern Recognition*, 47(11), 3614-3629.
- [23] Käll, L., Krogh, A., & Sonnhammer, E. L. (2005). An HMM posterior decoder for sequence feature prediction that includes homology information. *Bioinformatics*, 21(suppl\_1), i251-i257.
- [24] Wallach, H. M. (2004). Conditional random fields: An introduction. *Technical Reports (CIS)*, 22.
- [25] Sutton, C., & McCallum, A. (2012). An introduction to conditional random fields. *Foundations and Trends® in Machine Learning*, 4(4), 267-373.
- [26] Mokhtari, A., & Ribeiro, A. (2015). Global convergence of online limited memory BFGS. *The Journal of Machine Learning Research*, 16(1), 3151-3181.
- [27] Chang, D., Sun, S., & Zhang, C. (2019). An accelerated linearly convergent stochastic L-BFGS algorithm. *IEEE transactions on neural networks and learning systems*, 30(11), 3338-3346.
- [28] Mahajan, D., Liang, J. J., & Tsou, C. H. (2021). Toward understanding clinical context of medication change events in clinical narratives. In *AMIA Annual Symposium Proceedings* (Vol. 2021, p. 833). American Medical Informatics Association.
- [29] Opitz, J., & Burst, S. (2019). Macro fl and macro fl. *arXiv preprint arXiv:1911.03347*.
- [30] Tatar, S., & Cicekli, I. (2011). Automatic rule learning exploiting morphological features for named entity recognition in Turkish. *Journal of Information Science*, 37(2), 137-151.
- [31] Mordecai, N. B., & Elhadad, M. (2005). Hebrew named entity recognition. *MONEY*, 81(83.93), 82-49.
- [32] Nguyen, D. B., Theobald, M., & Weikum, G. (2016). J-NERD: joint named entity recognition and disambiguation with rich linguistic features. *Transactions of the Association for Computational Linguistics*, 4, 215-229.

- [33] Van Gompel, R. P., & Pickering, M. J. (2007). Syntactic parsing. *The Oxford handbook of psycholinguistics*, 289-307.
- [34] Leblay, J., & Chekol, M. W. (2018, April). Deriving validity time in knowledge graph. In *Companion proceedings of the the web conference 2018* (pp. 1771-1776).
- [35] Liu, S., Ma, W., Moore, R., Ganesan, V., & Nelson, S. (2005). RxNorm: prescription for electronic drug information exchange. *IT professional*, 7(5), 17-23.
- [36] Hains, G. J., Khmelevsky, Y., & Tachon, T. (2019, May). From natural language to graph queries. In *2019 IEEE Canadian Conference of Electrical and Computer Engineering (CCECE)* (pp. 1-4). IEEE.

## 附 录

以下是用于提取形态学、语言学以及句法结构三类若干特征的 Python 源程序：

```
import spacy
import re

features = {
    "meta": ["sent", "label"],
    "drug_dict": ["in_dict"],
    "baseline": ["is_digit", "is_alpha", "is_upper", "is_title", "in_dict"],
    "morphological": ["token", "lower", "len", "is_stop", "is_alpha", "is_digit", "is_title", "is_upper",
    "is_punct", "contain_upper"],
    "linguistic": ["pos", "tag", "lemma", "shape", "pref_2", "pref_3", "pref_4", "suff_2", "suff_3", "suff_4"],
    "syntactic": ["position", "has_dep", "n_ancestors", "n_lefts", "n_rights", "n_conjunc", "n_child", "bos",
    "eos"]
}

features["integrated"] = list(set(features["drug_dict"] + features["morphological"] + features["linguistic"] +
features["syntactic"]))

nlp = spacy.load("en_core_web_sm")

def extract_features(file_names):
    info = []
    for f_name in file_names:
        file_id = f_name[-10:-4]
        doc = open(f_name, "r").read()
        sentences = nlp(
            re.sub(
                r"[\n\s\t]{1}|[\*|\(|\)|\~|\_|\#|^|/|\\]{1}",
                " ",
                doc
            )
        )
```

```
)
).sents

for i, sent in enumerate(sentences):
    sent_info = []
    sent_id = f"{file_id}-{i}"
    n_token = len(list(sent))
    sent_len = len(str(sent))

    for token in sent:
        lower_token = token.text.lower()
        token_len = len(lower_token)
        start = token.idx
        end = start + token_len
        if str(token).isalpha():
            assert doc[start: end] == str(token), (doc[start: end], str(token))

    ## if don't do data cleaning
    # assert doc[start: end] == token.text, (doc[start: end], token.text)
    token_info = {
        # meta
        "file": file_id,
        "sent": sent_id,
        "ref_1": f"{file_id}-{start}",
        "ref_2": f"{file_id}-{end}",
        "start": start,
        "end": end,

        # morphological features,
        "token": token.text,
        "lower": lower_token,
        "len": int((token_len / sent_len) // 0.1),
        "is_stop": int(token.is_stop),
        "is_alpha": int(token.is_alpha),
        "is_digit": int(token.is_digit),
```

```

    "is_title": int(token.is_title),
    "is_upper": int(token.is_upper),
    "is_punct": int(token.is_punct),
    "contain_upper": int(any(char.isupper() for char in token.text)),

    # linguistic features
    "pos": token.pos_,
    "tag": token.tag_,
    "lemma": token.lemma_,
    "shape": token.shape_,
    "pref_2": lower_token[:2] if len(lower_token) > 1 else lower_token,
    "pref_3": lower_token[:3] if len(lower_token) > 2 else lower_token,
    "pref_4": lower_token[:4] if len(lower_token) > 3 else lower_token,
    "suff_2": lower_token[-2:] if len(lower_token) > 1 else lower_token,
    "suff_3": lower_token[-3:] if len(lower_token) > 2 else lower_token,
    "suff_4": lower_token[-4:] if len(lower_token) > 3 else lower_token,

    # dependency & syntactic features
    "position": int((token.i / n_token) // 0.1),
    "has_dep": int(token.has_dep()),
    "n_ancestors": int((len(list(token.ancestors)) / sent_len) // 0.1),
    "n_lefts": int((token.n_lefts / sent_len) // 0.1),
    "n_rights": int((token.n_rights / sent_len) // 0.1),
    "n_conjunc": int((len(token.conjuncts) / sent_len) // 0.1),
    "n_child": int((len(list(token.children)) / sent_len) // 0.1),
    "bos": int(token.is_sent_start),
    "eos": int(token.is_sent_end),
}

if not (token.is_space or token.is_stop) \
    or token_info["contain_upper"]: # Ca is not a stopword, while ca is
    sent_info.append(token_info)

info.append(sent_info)

return info

```

## 致 谢

感谢我的家人，在我读书期间给予了我最多的支持、包容与鼓励。感谢兰州大学，给我在这里求学的机会，在这里度过的四年光阴我将永远怀念；感谢这四年间指导我的所有老师，成长路上遇到诸位老师是我莫大的荣幸。

这里我要特别感谢赵志立老师！赵老师不仅给予了我学业上的指导，也是我人生道路上的导师。他一直以身作则地教导我们，什么是不忘初心。

四年梦一场，前程漫漫，韶华赠我花草茵茵。

沿路绽放的邂逅化作漫天的星，长夜尽头将至。


挚友们啊，咱们后会有期。

毕业论文（设计）成绩表

导师评语

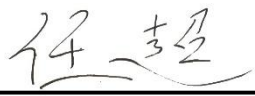
药物命名实体识别可以帮助研究者对病历数据进行统计分析并构建知识图谱。方啸同学的论文《基于上下文特征的药物命名实体识别》比较了常见的计算资源友好的药物命名实体识别方法，同时通过改进特征工程提高了基础条件随机场模型的性能。本论文结构完整，写作规范，达到了本科毕业论文的要求，同意参加学位论文答辩。

建议成绩      良

指导教师（签字）  


答辩委员会意见

经答辩小组一致讨论，该论文通过答辩，成绩为良。

答辩委员会负责人（签字）  


成绩

学院（盖章）

年    月    日



## 毕业论文（设计）成绩表

## 导师评语

药物命名实体识别可以帮助研究者对病历数据进行统计分析并构建知识图谱。方啸同学的论文《基于上下文特征的药物命名实体识别》比较了常见的计算资源友好的药物命名实体识别方法，同时通过改进特征工程提高了基础条件随机场模型的性能。本论文结构完整，写作规范，达到了本科毕业论文的要求，同意参加学位论文答辩。

建议成绩 良指导教师（签字）赵志立

## 答辩委员会意见

经答辩小组一致讨论，该论文通过答辩，成绩为良。

答辩委员会负责人（签字）任超成绩 良好

学院（盖章）

2023年5月27日

