中国人民大学

硕士学位论文

（中文题目） 养老领域词库的构建方法及应用

（英文题目） Research on the construction method and application of lexicon for senior care

学　　号：	2018104123
作者姓名：	银旭
所在学院：	信息学院
专业名称：	软件工程
导师姓名：	左美云

论文主题词：（3-5 个）	养老词库；嵌套术语识别；词典构建
论文提交日期：	2021 年 5 月 27 日

独创性声明

本人郑重声明：所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写的研究成果，也不包含为获得中国人民大学或其他教育机构的学位或证书所使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

论文作者（签名）：银旭 日期：2021.5.8

关于论文使用授权的说明

本人完全了解中国人民大学有关保留、使用学位论文的规定，即：学校有权保留送交论文的复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存论文。

论文作者（签名）：银旭 日期：2021.5.8

指导教师（签名）： 日期：2021.5.8

摘要

随着我国人口老龄化趋势越来越严峻以及互联网和通讯的迅猛发展，信息化养老、智慧养老等在国家政治层面和研究人员中越来越成为炙手可热的话题，越来越多不同领域的从业者投入到了养老事业中。而且，现在是网络大数据时代，通过网站浏览网页是人们日常的学习和生活中不可或缺的信息和知识来源方式之一。因此，网络文本有非常重要的研究价值，研究人员可以从中挖掘出有用的信息，应用到养老领域研究中。

基于以上背景，本文首先对国内外养老领域词库构建的研究现状展开了分析，发现了养老领域目前还未有统一的养老词库。针对这一缺口，本文着眼于养老领域文本挖掘及养老领域词库构建研究。首先，通过网络检索养老相关的词典作为基础词典。其次，爬取了养老领域的文章作为语料库，提出了基于成词概率（POW）的嵌套术语识别算法，使用该算法对养老领域的网络文章集合进行术语识别得到未登录词典的嵌套术语集合。然后，将新发现的嵌套术语与基础词典进行整合去重，得到最终的养老领域词库。最后将养老领域词库和嵌套术语识别算法应用到嵌套术语识别分词系统建设中。

基于上述研究内容，本文的理论贡献主要包括（1）本文提出的嵌套术语识别算法可为自然语言处理领域的术语识别、新词识别以及 NER 相关领域提供新的思路和方法，丰富了词库构建以及术语识别和新词识别的研究。（2）本文提出的养老领域词库可为养老领域的知识图谱研究和命名实体所识别研究提供参考。本文的实践贡献主要包括（1）本文提出的养老领域词典可对养老领域的相关主体对养老信息文本分析时提供分词词库，可提高分词准确率以及后续分析结果。（2）本文提出的养老领域词库可为百度百科、知乎等知识平台的词条以及话题创建提供一定的参考。（3）本文建设的嵌套术语识别分词系统可为相关人员提供非结构化文本的术语识别和分词功能。

关键词：养老词库；嵌套术语识别；词典构建

Abstract

With the aging trend of population in China becoming more and more severe and the rapid development of Internet and communication, informatization senior care and smart senior care have become more and more hot topics in the national political level and researchers, and more and more practitioners from different fields have invested in the cause of the senior care. Moreover, it is the age of big data on the Internet. Browsing web pages through web is one of the indispensable information and knowledge sources in people's daily learning and life. Therefore, the network text has very important research value. Researchers can dig out useful information from it and apply it to the research of the senior care field.

Based on the above background, this paper first analyzes the current situation of the construction of the lexicon for senior care at home and abroad, and finds that there is no unified lexicon in the field of senior care. In view of this gap, this paper focuses on the text mining and the construction of the lexicon in the field of senior care. First, the relevant dictionaries of the senior care are searched as the basic dictionary through the Internet. Secondly, the article in the field of senior care is crawled as corpus, and a nested term recognition algorithm based on probability of word (POW) is proposed. The nested term recognition algorithm is used to identify the net articles in the field of senior care to get the nested term set of unlisted dictionary. Then, the new nested terms and the basic dictionary are integrated to get the final lexicon of the senior care field. Finally, the word base and nested term recognition algorithm in the field of senior care are applied to the construction of nested term recognition segmentation system.

Based on the above research content, the theoretical contributions of this paper include: (1) The nested term recognition algorithm proposed in this paper can provide new ideas and methods for the field of natural language processing, new word recognition and natural language processing related fields, and enrich the research of lexicon construction and terminology identification and new word recognition. (2) The lexicon of the senior care field proposed in this paper can provide reference for the study of knowledge graph and the identification of named entities in the field of senior care. The practical contributions of this paper include: (1)The lexicon of the senior care field proposed in this paper can provide word segmentation database for the relevant subjects in the field of senior care information text analysis, and improve the accuracy of segmentation and the subsequent analysis results. (2) The lexicon of the senior care field proposed in this paper can provide some reference for Baidu Encyclopedia, Zhihu and other knowledge platforms and topic creation. (3)The Word

segmentation system for nested term recognition constructed in this paper can provide the related personnel with the function of unstructured text term recognition and word segmentation.

Key words: Lexicon for senior care; Nested term recognition; Lexicon construction

目录

图目录

表目录

第1章 绪论

随着我国的老年人人口比例日益上升，人口老龄化越来越严重，预计到2025年我国将进入中度老龄化，目前的养老方式已难以解决日益严峻的养老问题。对于养老问题，政府和相关研究人员们提出了新的解决方案：信息化养老、智慧养老等。为了将养老领域与计算机领域进行有机结合，我们需要充分利用网络大数据时代的特点。网络数据已经成为信息和知识的主要载体，尤其是网络上的文本数据，如网页文章。而且在网站上浏览网页文章已成为我们日常学习和生活中不可缺少的信息和知识获得方式。所以，网络文本数据有非常重要的研究价值，研究人员可以从网络文本中挖掘出有用的信息，然后可以将其应用到养老领域的各项研究中。但是目前还未有统一的养老领域词库，在词库构建的领域中，未登录词/新词识别以及术语识别一直是该领域需要解决的问题。

本文主要的研究工作包括：一、设计一个嵌套术语识别算法；二、基于嵌套术语识别算法，构建一个养老领域词库；三、将研究一的算法和研究二的词库应用到中文嵌套术语分词可视化系统建设中。

1.1 研究背景

预计到“十四五”（2021年-2025年）末期，我国年龄超过六十岁的老年人口数量将占全国总人口数量的20%以上，到那时，我国将进入到人口中度老龄化阶段。

因此，有关如何养老的这个热点民生问题引起了各领域从业者的关注，相关话题如“互联网+养老”、“智慧养老”等的热度越来越高。现在是大数据时代，互联网上的数据有非常重要的研究价值，根据第47次《中国互联网络发展状况统计报告》，如图1-1所示，截至统计时间：2020年12月，我国已有9.89亿网民，占全国总人口数量的70.4%。这些互联网使用者中，老年网民数

量超过了一亿，如图 1-2 所示，年龄在六十岁及以上的人数占比为 11.2%，互联网在向中老年群体渗透。

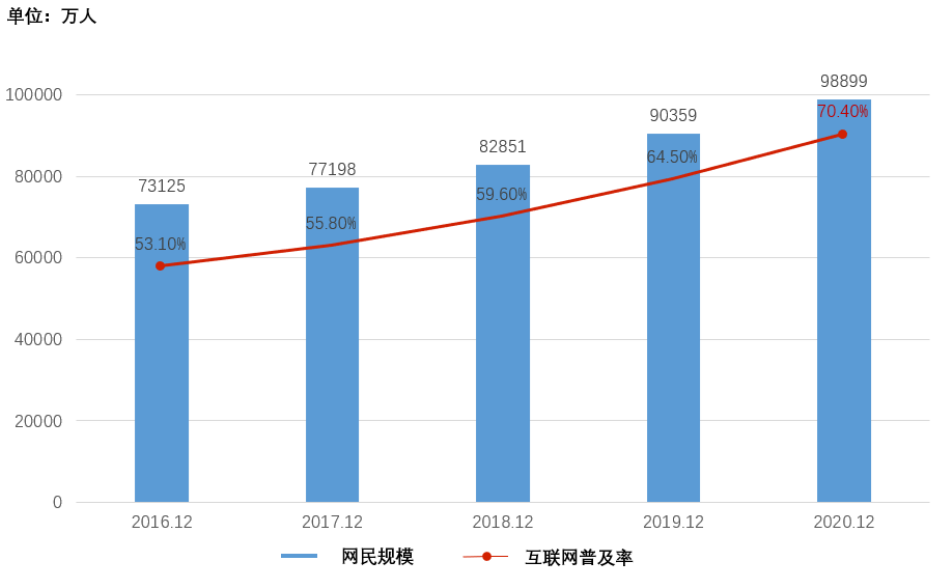

图 1-1 网民规模和互联网普及率

资料来源：第 47 次《中国互联网络发展状况统计报告》

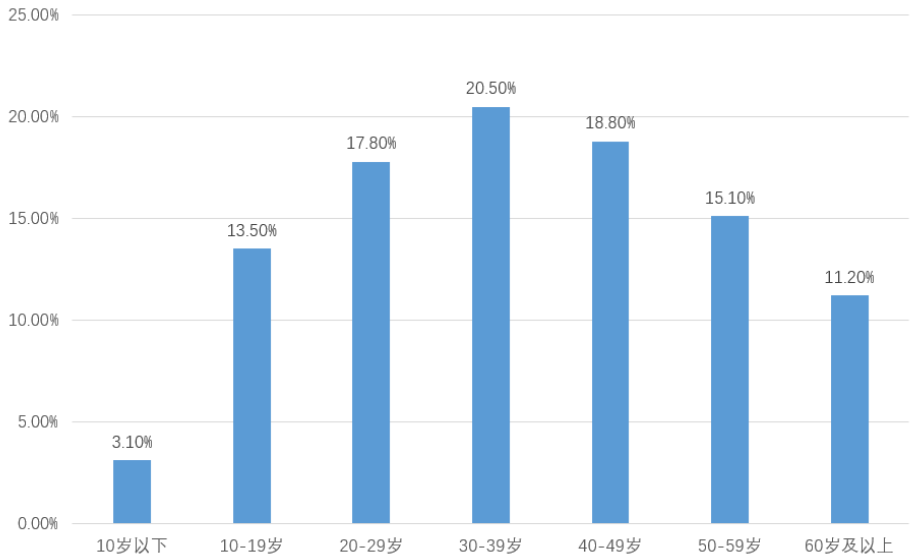

图 1-2 网民年龄结构

资料来源：第 47 次《中国互联网络发展状况统计报告》

以往关于养老领域的问题分析和老年人需求分析研究通常都是以结构化数据如调查问卷、访谈记录等作为数据源，网络文章这种非结构化数据很少有研究，所以有必要对其进行文本挖掘和分析，以期能识别出更多的养老需求及养老领域的热点话题。

在文本数据挖掘的研究中，基于词典匹配是现阶段比较有效的方法。其中，该方法的关键工作内容就是需要构建一个适用于该领域的领域词库。但是，目前还没有养老领域的专业词库，在做养老数据分析时难免会遇到无法识别的专业词语，影响文本分析的结果。采用人工的方式收集、整合领域词典是准确率比较高的一种方法，但是收集和整理词条有非常大的工作量，并且由于人的精力有限，难以确保领域词典能够完全覆盖整个领域。因此，我们需要一种工作量小且准确率高的方法来完成领域词典的自动构建工作。

在词典构建中，未登录词识别、新词识别、术语识别是领域内一直无法很好解决、需要长期解决的问题。

基于此，本文将重点研究：一、如何在领域语料中识别出未登录词典的嵌套术语？二、如何构建一个养老领域词库？三、养老领域词库如何应用？期望通过以上三个问题的解决，能为养老领域研究提供基础能力和参考价值。

1.2 研究意义

在当前我国信息化养老、智慧养老的大背景下，本文提出的领域术语识别算法和养老领域词库可为领域内学者提供一定的参考，同时养老领域词库将有利于研究人员展开对养老领域的文本分析与文本挖掘工作，具有一定的理论意义和实践意义。

1.2.1 理论意义

（1）术语识别是词库构建中非常重要的一部分工作，且各专业领域的词库通常含有较多的嵌套术语，但是目前中文嵌套术语识别的相关研究较少，本文提出的嵌套术语识别算法，可以从领域文章中识别出嵌套术语，对目前已有的基础词典进行扩展。本文提出的嵌套术语识别算法丰富了新词识别、未登录词识别、嵌套术语识别领域的研究方法，可以为领域内研究人员提供一定的参考。

（2）在文本挖掘和文本分析研究中，分词是一个非常重要的步骤之一。目

前广泛使用的是准确率和效率较高的基于词典的方法。但是目前还没有公开的养老领域词库，所以养老领域的文本挖掘工作会因为分词准确率不高的问题受到一定的影响。本文提出的养老领域公开词库可以为养老领域文本分析、文本挖掘工作提供一定的外部分词词典支持，可以提升养老领域文章分词的准确率，进一步提升该领域文本挖掘结果的可靠性。

1.2.2 实践意义

（1）我国老龄化问题越来越严峻，投身于养老领域的企业也随之增加，养老数据量越来越大，养老数据的分析和融合也越来越难。本文提出的养老领域公开词库可以为养老相关企业提供一定的数据参考，可为养老数据分析和养老数据融合做出一定贡献。

（2）目前常用的输入法如：百度、搜狗等还没有养老领域的词库，本文提出的养老领域公开词库上传到输入法词库后，可以丰富词库种类，提高打字时的编辑效率。

（3）本文建设的术语识别分词系统可为相关人员在做非结构化文本分析和挖掘时，提供便捷的术语识别和分词功能。

1.3 研究内容与方法

1.3.1 研究内容

本文共有三个研究内容，研究内容一是设计并实现一个嵌套术语识别算法，该算法可在给定的专业语料中识别出未登录词典的嵌套术语词汇。本文的研究内容二是为养老领域构建一个专业的中文养老词库。首先使用研究内容一提出的方法识别出养老领域文章中未登录词典的嵌套术语词汇，然后与基础词库整合，最后构建一个养老领域公开词库。本文的研究内容三是养老领域术语识别和分词系统的设计与实现，即将研究内容一提出的算法和研究内容二构建的养老领域词库应用到该术语识别分词系统中。

1.3.2 研究方法

本文的研究方法主要包括：文献分析法、网络数据抓取法、自然语言处理、统计机器学习以及专家调查法。

1.文献分析法。在文献库检索并阅读目前已有的国内外与养老词库构建领域相关的文献，将其分类并分析出相关领域的研究现状和值得进一步研究的空白或缺口，参考和借鉴领域内前沿的理论和方法，提出对构建养老领域词库相关工作的有启示的建议和想法。

2.网络数据抓取法：在编程环境中编写爬虫代码，使其可以自动地抓取网站上的内容。本文主要用于获取养老领域基础词典的网络词条和养老领域文章语料库。

3.自然语言处理法：能够使人类和计算机用语言进行有效通信的方法。在本文中主要用于养老领域词库构建过程中的格式转换、分词、词性标注和信息提取等。

4.统计机器学习法：使用统计知识和已经标好的数据集训练模型并预测的方法。本文的研究模型就是参考了该研究领域中的相关模型。

5.专家调查法：指围绕一个主题，通过专家的知识对相关内容进行判断。具体来说，需要由专家对研究的问题表达意见和看法。主要用于明确养老词库的范围、分词是否准确、词库是否可用等。

1.4 研究技术路线

本文有三个研究内容：一、设计并实现一个嵌套术语识别算法；二、基于嵌套术语识别算法识别出未登录词典的嵌套术语词汇，与基础词典进行整合，得到养老领域词库；三、将嵌套术语识别算法和养老领域词库应用到嵌套术语分词可视化网站建设中。本文的研究内容框架如图 1-3 所示。

第一阶段：国内外研究现状分析。笔者在中国人民大学线上图书馆数据库检索、阅读并分析了大量国内外相关的文献。首先，分析了文本挖掘领域的研究现状和方法，发现了分词和领域词库的重要性。其次，对现有的领域词库构建现状及方法进行了分析，发现了新词/未登录词识别是该领域的一个研究重点；接着，阅读并分析了新词/未登录词识别、术语识别等相关领域的文献及方法；

最后，深入研究了术语识别和领域词库构建方法。

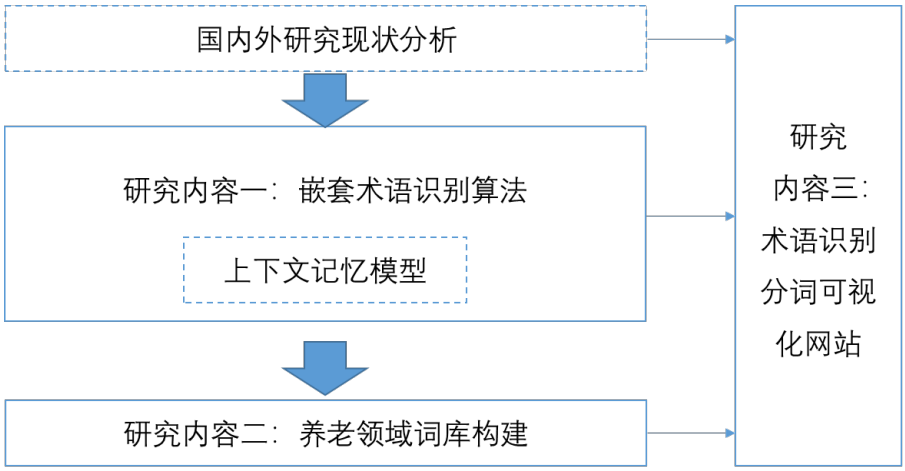

图 1-3 研究内容框架

第二阶段：嵌套术语识别算法设计。本文的算法借鉴和参考了深度学习中双向长短期记忆网络的模型以及统计学的点互信息 PMI,对二者进行融合和改造，并加入了一些文本语言特征，总结出了适用于本文的嵌套术语识别算法。最后，使用了准确率、召回率、F1 值分别评价了嵌套术语算法。

第三阶段：养老领域词库构建。首先，使用网络爬虫技术爬取与养老相关的网络文章作为基础语料。其次爬取老年健康领域词典、知乎上相关话题词、百度百科相关词条，将其整理成养老领域基础词典。再次，对基础语料使用嵌套术语识别算法识别出未登录的术语词汇。最后，将其与基础词典进行整合，得到最终的养老领域词库。

第四阶段：建设了基于嵌套术语识别算法和领域词典的可视化分词系统。将嵌套术语识别算法和养老领域词库应用到可视化分词系统中。

1.5 难点与创新点

1.5.1 研究中的难点

中文非结构化文本的分析工作是非常有难度的，往往面临着词语识别错误、语义歧义等问题。中文养老领域文本分析挖掘工作更是增加了领域专业性的难度，具体展开如下：

（1）由于汉语的博大精深，构词规则和语言规则都极其复杂，难以梳理出规则全集，所以中文的自然语言处理和文本分析的难度非常大，目前的分词工具对专业领域分词时可能会丢失重要的、未登录的术语或关键词，会影响文本分析结果。未登录词、术语识别是自然语言处理中需要长期解决的问题，还有较大的研究空间。

（2）信息化养老、智慧养老是近几年才开始进入人们视野的，还是一个新兴的、尚未成熟的领域。养老领域也是一个比较大的范畴，目前还没有与养老领域词库构建相关的研究，所以输出一个养老领域公开词库的工作量比较大且有一定的难度。

1.5.1 研究中的创新点

本文研究了嵌套术语识别算法、养老领域词库构建、术语识别分词系统这三部分内容，含有三个创新点，具体展开如下：

（1）本文提出的基于上下文记忆网络的嵌套术语识别算法，丰富了未登录词/新词识别、术语识别领域的研究方法，可为研究人员提供参考；

（2）本文创新性地输出了一个中文养老领域词库，该词库可以分为老年医学、老年社会学、老年生物学、老年心理学、养老政策、养老服务、养老模式、养老产品共计 8 个大类、10000 余条养老相关词汇，可为养老领域产业提供基础能力和参考；

（3）本文建设了一个可以识别嵌套术语的可视化分词系统，可提供给相关人员进行可视化分词和术语识别，降低了养老领域文本分析的难度。

1.6 论文结构

本文有六个部分，具体说明如下:

第一章：绪论。概述了养老领域产业、养老领域词库的背景以及研究现状、研究内容和拟采用的研究方法、理论意义和实践意义、技术路线、难点和创新点。

第二章：国内外研究现状分析。本部分介绍了文本挖掘研究的现状和方法、词库构建的现状、以及词库构建中用到的方法：分词、未登录词/新词、命名实体和术语识别的研究现状。

第三章：嵌套术语识别算法。本部分主要介绍了实现该算法会用到的相关技术，嵌套术语识别算法的设计以及模型图，并对算法进行了评价。

第四章：养老领域词库构建。本部分主要阐述了养老领域词库构建的相关工作：数据源获取、基础词典构建、嵌套术语识别、养老领域词库构建。

第五章：嵌套术语识别可视化分词系统。本部分主要介绍了该系统的功能以及系统页面，包括分词功能和术语识别功能。

第六章：总结与展望。本部分总结了成果和不足之处，并对后续可以研究的方向做了展望。

1.7 本章小结

本章从背景及现状、方法和技术路线、理论意义和实践意义、难点和创新点这几个方面分别进行了阐述。首先，介绍了养老产业和养老领域词库研究的背景现状；其次，介绍了本文重点研究的三个内容及其对应的方法；再次，阐述了本研究的理论意义和实践意义所在；接着，介绍了可行的技术路线；然后，阐述了研究中遇到的难点以及本文的创新点；最后，介绍了论文的整体结构，并对每一部分的内容做了简要概述。

第2章 国内外研究现状分析

本章对文本挖掘领域、词库构建领域以及词库构建中用到的方法，如分词、未登录词/新词识别、命名实体识别的研究现状进行了梳理和总结。

2.1 文本挖掘领域分析

文本挖掘(Text Mining)，是从文本形式的数据集合里，通过计算机技术挖掘并分析出对领域有启示的内容。目前，文本挖掘领域的方法可以总结为基于统计机器学习、句法分析、词典或规则进行匹配的方法。

（1）基于机器学习，如刘一宁等人（2019）使用基于终身机器学习机制的推荐模型提高了主题词分类的准确率。胡吉明等人（2014）提出了基于动态LDA 的主题挖掘模型,该方法可以有效地从网络文本中提取主题，并根据时间分析了主题的演化。还有学者使用主题模型与聚类方法的结合以挖掘主题短语（Li et al. 2019）。

（2）基于句法分析，如 He 等人（2017）提出了基于依存句法分析的观点挖掘方法，该方法提取特征词和情感词之间的特征并进行打分，实现了从评论中挖掘出有用的观点和意见。

（3）基于词典：如 Mandal（2017）设计了基于词典的观点挖掘和计算情感极性的算法。

在以上这几种文本挖掘的方法中，大部分方法都需要对文本语料进行一定的预处理，其中一步就是对文本进行分词操作，词典的质量会影响文本挖掘和分析的准确性。而且基于机器学习的方法通常需要提前准备已经标注好的数据，如果没有则需要人工标注，十分耗费人力，且扩展性较差。基于词典的匹配方法是目前该领域最有效的方法之一，因为该方法不需要对数据进行标注，使用成本比较低，且相对容易进行词典扩展。基于词典的匹配方法被广泛应用于多个领域（石玉鑫等，2019）。因此，本文要构建一个中文养老领域公开词库。

2.2 词库构建领域分析

在自然语言处理和文本分析中，分词是非常重要的步骤。但是在专业领域的文本挖掘中，其文本语料含有大量的专业术语，单纯依靠现有的分词算法难以准确地对术语进行切分，需要引入外部词库。词库构建领域目前已有很多研究，研究方法有基于知识库/语料库、机器学习以及基于语料库和机器学习集成的方法。

（1）基于知识库/语料库，如周咏梅等人（2014）设计了一个基于情感词典的词库构建方法，先从基础情感词典中筛选出种子词集，再从新闻评论中抽取出评论情感词集，最后使用 PageRank 算法和 PMI（点互信息）对种子词集进行扩展，得到了基于新闻评论领域的情感词典。杨鑫等人（2020）对民宿评论提取候选词集，使用 SO-PMI（情感点互信息）对基础情感词典进行了扩展，得到了基于民宿领域的情感词典，并分析了用户对民宿的评论的情感正负倾向性及原因。尹文科等人（2014）设计了一种基于图聚类的词典构建方法，该方法对 Wiki 的词条链接构建了有权无向的链接图，然后使用图聚类算法，得到了领域词典。Song 等人（2015）实现了基于 wiki 的命名实体半自动构建系统，使用了 Active Learning 和 BM25 信息检索模型，结构表明该系统有不错的表现。

（2）基于机器学习的方法：如 Ju 等人（2016）设计了基于 CRF 和临床语料库的算法，构建了一个中文的临床症状领域词库，该词库包含 22501 个临床症状词汇。Wu 等人（2017）标注了足球新闻语料，使用 Word2Vec 得到了词向量，然后计算 TF-IDF 值和余弦值进行筛选，最后生成基于足球领域的情感词典。杨秀璋等人（2019）提出了一个基于 LDA 模型的词典构建方法，该方法使用 LDA 主题模型抽取特征词词集，构建了一个特征词词典。

（3）基于语料库和机器学习集成的方法：如石玉鑫等（2019）设计了一个基于 LDA 和特征的方法，首先使用 LDA 提取主题词作为基础词典，然后使用点互信息和依存句法分析方法扩展基础词典，最终得到了一个洗衣液产品评论的领域词典。Tang 等人（2014）使用了 KNN 模型对扩展种子词得到词典；李伟卿等人（2018）基于商品评论数据，使用 Word2Vec 工具将评论数据转化为词向量，计算同义词词林和语料库的相似度，将相似度较高的候选词加入基础词典，得到了产品特征词典。Beigi 等人（2020）使用神经网络模型和多层感知器

自动构建了一个情感词典。

以上词库构建的方法中，基于机器学习往往需要很多已经标好的语料数据，十分耗费人力；基于知识库的网络数据构建的方法、遍历深度不好确定，深度过大时间空间复杂度大，深度过小词典不全，且词条为人工创建，未录入的词条无法识别。新词、未登录词是分词工作中，非常影响准确率的重要因素（刘浏，2018），所以**未登录词/新词、术语识别是该领域研究的重点，本文将从术语识别算法方面优化领域词库构建方法。**

2.3 分词领域分析

分词算法在自然语言处理中的应用十分广泛，如信息检索、语音识别、机器翻译、自动问答等（唐琳，2020）。目前，分词算法领域已有较多的研究：基于字符串匹配、机器学习、深度学习以及集成方法。

（1）基于匹配的方法；如常建秋等人（2016）使用了基于字符串匹配算法对中文文本进行分词研究。

（2）基于机器学习：如钱智勇等人（2014）研究了汉语自动分词技术，该方法使用统计学中的 HMM，取最大概率和加值平滑算法对《楚辞》进行自动分词和标注研究。张梅山等人（2012）使用条件随机场（CRF），并引入了词典信息特征，对中文文本进行分词并提高了领域自适应能力。Wu 等人（2010）使用 CSV- Markov 模型用于分词的研究, Zhang 等人（2010）使用 SVM 算法进行中文分词并在比赛中取得了优秀的成绩。

（3）基于深度学习的分词算法：如使用神经网络进行中文分词（Zheng et al.2013），使用 LSTM 分词（Chen et al.2015），该方法引入字与字之间的远距离依赖关系。张洪刚等人（2017）将 CNN、Bi-LSTM 等应用到中文分词中，并进行了改进。Peters 等人（2018）使用基于上下文的动态词向量训练模型进行分词研究。Wang 等人(2019)提出了一种双 LSTM-CRF 模型，该模型通过在训练过程中共享 LSTM 网络来有效地融合语言特征，从而提高分词的准确率。

（4）集成方法：如冯国明等人（2018）提出了一个 DBLC 分词模型，该模型结合了词典、统计和深度学习，基于专业语料实现了分词并提高了分词的

准确度。 张文静等人（2018）设计了多粒度中文分词算法,引入了多分词。夏松等人（2019）设计了一个名为 LBCP 的分词算法，该算法基于词位置等特征，提升了中文分词的效率和准确率。韩冬煦等人（2015）引入了边界熵和卡方作为特征进行分词。朱艳辉等人（2016）设计了基于 CRF 的中文分词方法，该方法使用条件随机场，引入领域词典，提升了中文分词的准确性。

以上几种分词方法，机器学习和深度学习都需要有大量已标注的数据（ 如 CRF 需要对词进行词位标注），而且往往不能充分利用汉语词典中的有用信息，且更为复杂，需要更多的计算资源。而且，新词/未登录词识别是一直需要并且值得研究的关键问题（唐琳，2020）。

2.4 术语识别领域分析

术语（Term）是使用文字表达特定专业概念的约定性符号（冯志伟，2011）。我们可以理解为特定领域内的、非所有领域通用的词汇为该领域的术语，在本文中即为养老领域的术语。术语在通用词典中比较少见，所以大量的术语未登录到通用词典中，属于新词/未登录词的范畴。本文认为，命名实体识别（Named Entity Recognition，NER）和术语识别（Term Recognition，TR）属于交叉领域，因此，也需要对 NER 领域进行分析。

2.4.1 新词/未登录词识别

该领域的研究方法有:基于规则、统计机器学习、还有将二者进行集成的方法。其中，基于规则的方法在特定领域识别效果较好，但是难以适用于其他领域；基于统计的新词识别方法，需要学习新词规律，那么就需要大量的标注语料库，会产生数据稀疏、正确率不高等问题；目前业界大多使用二者结合的方法。

Pecina 等人（2006）通过实验计算了 50 余种量化指标,最终发现了点互信息（Pointwise Mutual Information，PMI）是确定不同字符串相关性效果最好的指标之一。但只依靠 PMI 经常会识别出低频相邻但是没什么用的字符串。所以学者会对 PMI 做改进或与其他指标结合，如徐豪杰等人（2020）使用了改进的点互信息 PMI 和最小邻接熵进行了未登录词识别。夭荣朋等人（2016）利用 N-Gram、PMI 和邻接熵等特征实现了未登录词提取。杜丽萍等

人（2016）对 PMI 进行改造，将其应用在新词识别中，而后对中文分词系统也进行了优化，最后得出一个结论就是引入自定义词典可以提升分词系统的准确率。赵耀全等人（2021）使用 n-grams 切分中文文本从而实现了新词识别。

这些研究的共同特点是以字维度计算相邻字的 PMI 或者其他指标，识别出的基本都是 5 字以内的新词，对 5 字以上的长新词几乎没有研究，但事实上，专业领域的长新词/长术语识别是非常重要且必要的。

2.4.2 命名实体识别

命名实体识别（Named Entity Recognition，NER）是在文本语料中对命名实体进行识别，有非常广泛的应用，如信息抽取、观点挖掘、机器翻译、知识图谱、语义搜索等（Goyal 等，2018）。NER 常用的方法有：基于规则、机器学习、深度学习以及集成方法。

（1）基于规则的方法，如龚德山（2019）使用了 jieba 分词工具与关键词匹配的识别方法，基于 Bi-LSTM-CRF 的方法，在中药领域做了一个 NER 试验并进行了比较，实验表明，前者的效果更优。

（2）基于机器学习及其与其他方法结合的方法，如 Keretna 等人（2015）建立了一个扩展的分词表示（SR）技术，提高了 NER 在医学领域的实用性。Morwal 等人(2012)设计了基于 HMM 的 NER 方法。Shijia 等人（2017）设计了 CWME 的中文 NER 方法，该方法使用字词混合嵌入的方式，提高了中文 NER 的准确率。Peng N 等人（2015）设计了联合嵌入的 NER 方法，该方法使用的是微博数据集，通过对标记的 NE 和未标记的原始文本联合嵌入，实现了 NER 任务。

（3）基于深度学习的方法。LSTM 方法，如顾孙炎（2018）设计了基于深度神经网络的 NER 方法，该方法改进了 Bi-LSTM-CRF，提高了准确率；张瑞东（2018）设计了基于 Bi-LSTM-CRF 的 NER 方法,该方法在深度学习中引入了条件随机场,可以有效地进行 NER 任务。卷积神经网络方法，如 Kong 等人（2021）使用 CNN 和门结构模型实现了 NER。

（4）混合方法。Greenberg 等（2018）设计了基于 CRF 的 NER 方法，该方法使用了 CRF 条件随机场和异构标签集，在对领域语料的 NER 任务中有较好的表现。张春燕（2019）提出了一个基于神经网络、Semi-Markov-CRF 的命

名实体识别方法，该方法标记子序列并将其作为一个整体,解决了条件随机场中的局部依赖问题,实现了 NER 任务且识别性能有所提高。Wang 等人（2018）提出了一个基于序列变换模型的中文 NER 方法，该方法使用了神经网络模型和条件随机场（CRF），利用句子标记信息和注意机制来获取语义信息，从而实现了命名实体识别。Liu 等人(2019)提出了基于 Semi-Markov-CRF 的 NER 方法，该方法使用了混合半马尔可夫条件随机场并引入了地名词典。

可以看出，目前比较多的研究都是采用的混合方法，可以集合多种方法的优点。目前中文 NER 的研究热点主要是中文嵌套命名实体（Nested Named Entity，NNE）识别（陈曙东，2020）。而且，英文的嵌套命名实体研究较多，如 Xia 等人（2019）使用多粒度命名实体识别框架 MGNER 识别一句话中的多个实体或实体引用不重叠或完全嵌套的实体，Ju 等人（2018）使用动态堆叠平面 NER 层识别嵌套实体，该模型使用了 LSTM 和 CRF；Katiyar 等人（2018）利用递归神经网络提取术语特征，用超图表示来嵌套实体。目前与中文嵌套命名实体识别有关的文献还相对较少，主要是使用已标注数据、机器学习、深度学习结合的混合方法，如李雁群等人（2018）利用层次标记和层叠模型实现中文嵌套命名实体识别任务。许浩亮等人（2019）利用已经标好的中文嵌套命名实体语料库， 然后使用 SVM 和 RNN 实现了中文嵌套实体的识别；尹迪等人（2014）将 BIE 标注与联合模型结合进行中文嵌套实体识别。本文研究的未登录的嵌套术语词汇与之类似，其研究成果可做参考。

2.4.3 术语识别

术语识别是指从文本数据中识别出属于某领域的专业术语，是中文分词领域的一个重要步骤,也是文本分类、文本摘要等领域的关键研究问题（张雪等，2020）。

目前来看，该领域的方法包括：基于规则、规则和统计、外部知识/语料库、语义理解、机器学习以及深度学习的识别方法。

（1）基于规则的识别方法。如 Foo 等人（2010）提出了基于 Ripper 的术语识别方法，该方法采用机器学习算法来学习已经标注好的语料库的语言规则，主要有词性标记、语义信息、归一化词频等，从而构成了术语规则并进行识别。李思良等人（2018）提出了 DRTE 术语识别方法，该方法使用了术语的

定义及其关系、构词规则以及边界确认的方法实现了对教育语料库的术语抽取。

（2）混合方法。如 Yangyi 等人(2017)提出了一种混合的中文属于抽取方法，该方法使用了 TF-IDF 和信息熵作为特征，从而进行中文术语识别。陈梅婕等人（2020）使用双向聚合度特征提取方法识别专利领域的长新词。杨双龙等人（2016）提出专利领域的术语识别算法，通过学习专利文献的标题，生成对术语结构词性的筛选规则，使用该规则筛选出候选术语，然后针对候选术语使用新提出的 TermRank 算法计算分值并排序，实现了专利文献领域的术语识别。

（3）基于机器学习的方法。Wang 等人（2016）以迭代执行 CRF 的方式获取新的术语词集，并根据迭代出的规则识别新的术语。王密平等人（2016）提出了基于 CRF 的术语识别方法，该方法使用了条件随机场，实现了对冶金领域的中文专利文献的术语识别。Mishara 等人（2020）基于词嵌入的语义过滤器实现了术语识别。

（4）基于深度学习的方法。

学者们通常使用深度学习的方法和其他方法结合使用，如 Gao 等人（2019）提出了基于端到端的深度学习模型来学习候选术语的向量表示,然后输入到分类器,可以识别出嵌套术语。Zhao 等人（2018）基于 Bi-LSTM-CRF 进行术语识别。

以上术语识别的方法中，由于规则无法通用，所以仅仅基于规则的术语识别方法无法直接应用到其他领域；集成混合方法可以利用各种特征组合来抽取术语,提高了准确性和可移植性；基于机器学习的方法需要让计算机学习规律和特征，所以往往需要大量已经标注好的数据集，若没有则需人工进行标注；基于深度学习的方法同样需要已经标记完成的数据集和训练时间,相对来说模型在其他领域的泛化能力弱一些。而且，目前嵌套术语识别研究相对较少，截至行文前在中国人民大学图书馆仅搜到一篇 Gao 等人（2019）的应用于英文生物领域的嵌套术语识别研究，没有搜到“中文嵌套术语”的研究，有一篇基于电力客服文本的“复合术语”的研究，使用了该领域的规则，所以难以移植到其他领域（嵇友浪等，2021）。

本文将老年学领域和养老领域涉及到的老年生物学、老年社会学、养老政策、养老模式、老年人生活、老年医学、老年病学以及近些年兴起的养老服务等维度纳入考虑。首先，爬取网络数据整理出基础词典，并爬取大量养老领域文章作为本文的语料库。其次，参考以往关于术语识别、命名实体识别、新词识别的方法提出本文的嵌套术语识别算法。再次，使用该算法对养老文章语料识别出未登录词典的、新的养老领域嵌套术语。最后，利用嵌套术语对基础词典进行扩展和补充，得到最终的养老领域词库，以期指导未来智慧化养老领域的文本分析和系统开发实践。

2.5 本章小结

本章先从文本挖掘领域、词库构建领域进行了分析，发现了未登录词和术语识别的问题。其次，对分词领域、新词/未登录词识别、命名实体识别（NER）、术语识别等领域的研究现状分别进行了梳理和总结。经梳理发现，目前还未有养老领域词库的研究、且未登录的、嵌套术语的识别有一定的研究价值，因此，本研究将参考以上领域的方法提出本文的嵌套术语识别算法，并整理出养老领域词库，期望对养老研究和自然语言处理有一定的研究价值和参考。

第3章 嵌套术语识别算法

本章主要研究的是研究内容一：嵌套术语识别算法。首先介绍了主要的研究问题，其次介绍了相关的技术，然后设计了嵌套术语识别算法及模型结构，最后对本文提出的嵌套术语识别算法进行了评价。

3.1 研究问题

我们发现，现有中文分词工具对专业领域的中文文本进行分词操作时，存在将一些未登录的新词，特别是长术语和嵌套术语拆分成几个词的情况。如图3-1 所示，使用国内某在线分词工具对养老领域的专业术语“社区老年人日间照料中心”进行分词，结果是：“社区”、“老年人”、“日间”、“照料”、“中心”；将术语“互助幸福院”进行分词，结果是：“互助”、“幸福”、“院”。以本文研究的养老领域为例，专业领域文章的嵌套术语非常多，且大多数嵌套术语未登录词典或分词工具无法识别，对文本挖掘和分析会有影响。

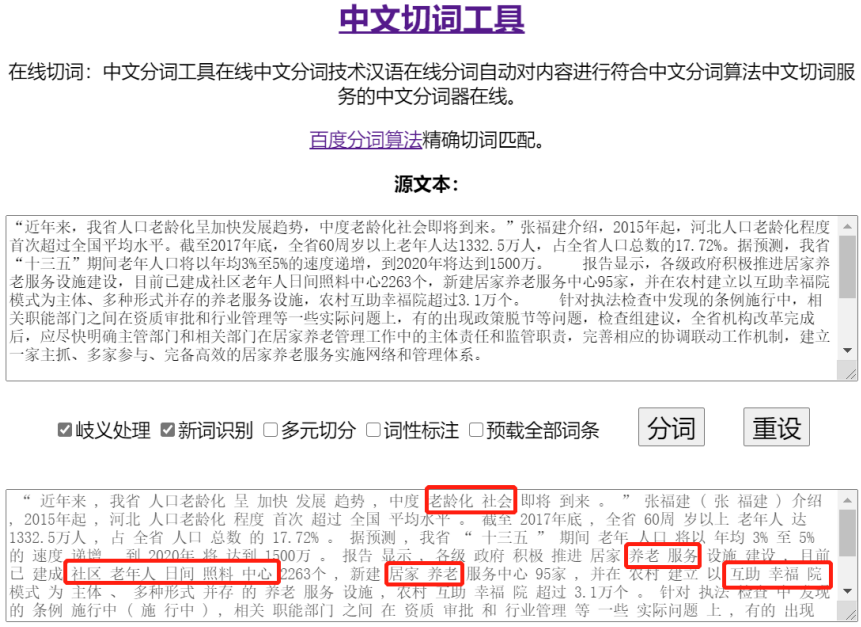

图 3-1 分词工具分词结果

资料来源：某在线中文分词工具

参考“嵌套命名实体”（陈曙东等，2020），**本文将含有多个（两个及以上）子术语的长术语词汇称为嵌套术语（Nested Term）**。如图 3-2 所示，嵌套术语“社区老年人日间照料中心”含有“照料中心”、“日间照料中心”、“老年人日间照料中心”三个子术语。若该嵌套术语未被记录在词典中，该嵌套术语则会被分成其多个子词，即如图 3-1 所示的结果。

图 3-2 嵌套术语举例

嵌套术语识别与术语识别流程相似，均为先获取候选词集，再根据特征值筛选过滤非术语，得到术语集合。嵌套术语识别的难点是边界确定和筛选过滤条件的设计，要充分考虑嵌套术语上下文的影响。

因此，嵌套术语识别非常值得进一步研究，亟需一种方法来识别嵌套术语。本文的嵌套术语识别算法关键步骤如下：

（1） 采用网络爬虫工具抓取网络文章作为语料。

（2） 使用 Python 第三方库 jieba 对语料进行分词以及词性标注工作。

（3） 将 N-GRAM、PMI、BILSTM 进行改造融合，提出本文的嵌套术语识别算法，该部分将在 3.3 节详细展开。

（4） 对嵌套术语识别算法进行评价。

以上关键步骤的详细内容在后续章节会进行展开描述。

3.2 相关技术介绍

3.2.1 网络爬虫及工具

网络爬虫（Web Crawler）是通过一段自动抓取网页的程序代码获取符合要求的内容的方法，是构建领域语料库的关键步骤。首先我们需要一个统一资源定位符 (Uniform Resource Locator，URL)集合，然后从 URL 集合中第一个网页开始爬虫，获取当前网页的内容，接着，获取新 URL 并将其放进 URL 队

尾，如此循环往复，直到满足结束条件（于娟，2015），最后将网页数据存储下来。

Python 的拓展性较强，在用 Python 爬取网页时可以在代码中导入专门用于网页爬虫的第三方库 Requests、Beautiful soup 等，功能强大，因此 python 爬虫是目前最常用的网页获取方式之一。Requests 的编程语言是 python，主要的功能是获取网页的 URL。Requests 常用的方法有 4 个，如表 3-1 所示。

表 3-1 Requests 的常用方法

方法	说明
requests.request()	创建一个请求，是基础方法，用于支撑其他方法。
requests.get()	获取 HTML 网页的主要方法，通过 URL 提交请求数据
requests.head()	获取 HTML 网页头部信息的方法
requests.post()	通过 html header 提交请求数据，相对安全

Beautiful soup 也是 Python 的第三方库，主要是为了抓取 HTML 网页代码和解析代码数据。使用基本元素识别和定位 HTML 标签，如表 3-2 所示。

表 3-2 Beautiful soup 常用元素说明

基本元素	说明
Tag	标签，是最基本的信息组织单元
Name	标签的名字
Attributes	标签的属性
NavigableString	标签内非属性字符串
Comment	标签内字符串的注释部分

常用的 Python 网络爬虫的流程：发送一个请求、获取响应的内容、解析网页代码、保存网页中的数据，具体说明如下：

（1） 发送请求

使用 requests 向目标网站发送请求

（2） 获取响应内容

服务器返回请求的内容：html、Json 字符串、图片、视频等。

（3） 解析内容

使用 Beautiful soup 识别和定位元素标签，抽取需要的数据。

（4） 保存数据

保存到本地，格式可以是 txt、csv 等，也可以存到数据库。

3.2.2 中文分词及工具

中文分词是通过某种方法，将连续的中文文本根据定义好的规则和规范切分成词（1993）。相邻英文单词间有空格，可以作为分界符，然而中文的句子和文章中，几乎所有词语都是紧挨着的，没有任何分界符。所以基于中文的分词工作要比基于英文的分词困难得多，而且中文分词工作是对中文本文进行挖掘和分析的重要基础，非常值得进一步研究。

该领域常用的方法有：基于字符串匹配、语义理解、统计学习、机器学习的方法。

（1）基于字符串进行匹配的方法，是按照预先写好的规则将文本依次与词典的每个词条进行匹配，如果在分词词典中能找到与之完全相同的词条，那么就可以识别出该词。

（2）基于语义理解的分词方法。可以让计算机在进行分词操作时能像人类一样理解文本的语义。为了避免因为歧义造成的错误分词，可以利用机器学习到的句法和语义信息处理有歧义的词语。但是，基于语义进行理解的方法要求计算机提前学习非常多的语言学知识，不过由于汉语非常复杂，所以目前仍无法将所有的语言知识都组织成计算机能够读取并理解的结构形式，因此，目前此种方法较少投入使用。

（3）基于统计的方法。主要统计给定的语料库中同时出现的字和字的组合次数，计算出所有字和字的组合的互现信息值从而识别出可能是词语的字串。该方法只需要足够的语料库即可，而不依赖分词词典，因此可以识别出未

登录词典的新词。然而，该方法有一个明显的缺点，因为它会识别出共现率高而非“词语”的“垃圾串”。所以在业界的实际应用中，该方法通常都会引入外部分词词典配合使用。

（4）基于机器学习的方法。是让计算机通过机器学习模型去学习大量已经分好词的数据的规律，在大量的学习之后，就能对其他文本分词，但是分词准确率难以保证，比较依赖标注数据和机器学习算法。该方法有一定的局限性，需要大量已经分好词的语料库，若没有预料库则需要人工对语料进行标注，需要耗费人力，机器学习的训练过程时间也比较久，换言之，该方法的成本较高。

目前业界常用的中文分词工具有 jieba、SnowNLP、北京大学 PKUseg、清华大学 THULAC、FoolNLTK、哈工大 LTP 等。这些工具的特点是可以导入文本分析的代码中，且可以自定义优化分词效果。还有一些在线分词网站，在线分词网站的优点是无需下载安装，可直接使用，缺点是无法自己优化分词结果。

本文使用的分词工具为 JIEBA。JIEBA 分词工具带有词性标注（Part-Of-Speech tagging, POS tagging）功能，POS tagging 被广泛应用在文本挖掘和 NLP 领域的各个任务中，是个非常重要的文本特征。

3.2.3 N-gram 模型

N-gram，是一种语言模型，它能在文本中截取出 N 个项目（item）序列。这里的项目（item）可以是字母、单字或单词等等。N-gram 有非常广泛的应用，如概率论、计算机语言 NLP 等领域。

N-gram 是指给定窗口 N 内的一组同时出现的项目（item，字或词），对于本文来说，项目（item）是分词后的 segment。以本文研究的养老领域为例，以 segment 维度对“我国将大力推动社区老年人日间照料中心的建设”这句话进行 Bi-gram（2-gram），得到的 Bi-gram 结果为“我国将”、“将大力”、“大力推动”、“推动社区”、“社区老年人”、“老年人日间”、“日间照料”、“照料中心”、“中心的”、“的建设”，滑动窗口为 2，所以每个 Bi-gram 由两个 segment 构成。当 N=3 时，滑动窗口为 3，可称 Tri-gram 或者 3-gram；以此类推，N 没有限制，通常最多为 5，常用的是 Bi-gram 和 Tri-gram。

3.2.4 点互信息（PMI）

互信息（Mutual Information，MI），来源于概率论和信息论，用于量化两个随机变量间相互依赖的程度，也就是相关性的指标。公式如下：

$$I(x;y) = \sum_{x,y} p(\mathrm{x,y}) \log \frac{\mathrm{p(x,y)}}{\mathrm{p}(x)p(y)} \tag{3-1}$$

点互信息(Pointwise Mutual Information，PMI)，是 MI 的演化，可以用于量化两个事物的相关性，在自然语言处理中，则是用来量化两个词或字之间的相关性。公式如下：

$$PMI（x;y） = \log \frac{\mathrm{p(x,y)}}{\mathrm{p}(x)p(y)} \tag{3-2}$$

公式（3-2）中，p(x)是指字或词 x 在文本中出现的概率，即 x 出现的次数除以该文本的总词数 N。p(x,y)是指 x 与 y 在一句话里同时出现的概率，即 x 与 y 在一句话里同时出现的次数除以 N 的平方。

3.2.5 文本逆文本频率（TF-IDF）

TF-IDF（term frequency–inverse document frequency）文本逆文本频率，可以用来量化一个字或词对于文档集合中某一份文档的重要程度。文档里词 w 的个数越多，其 TF-IDF 越大；w 在文档集合里出现的频率越大，w 的 TF-IDF 值越小。所以，TF-IDF 值可以筛掉在文档里高频出现但是没什么用的词，筛选出比较少见但是相对重要的词语。

TF-IDF 的公式如下：

$$TF - IDF = \mathrm{TF} * \mathrm{IDF} \tag{3-3}$$

其中，TF (Term Frequency)为词频，表示该词在文档中出现的频率，IDF (Inverse Document Frequency)为逆向文件频率。二者的公式如下：

$$TF = \frac{\mathrm{n(w)}}{\mathrm{N}} \tag{3-4}$$

$$IDF = \log \frac{\mathrm{F}}{f(w)+1} \tag{3-5}$$

公式（3-4）里，n(w)是指文本文件里词 w 的个数，N 为 w 所在文本文件的所有词的数量。公式（3-5）里，F 是指文本文件集合的文件总数，f(w)是指包含词 w 的文本文件的数量。含有词 w 的文本文件数量越小，该词的 IDF 越大，表示该词可以用来区分其所在文本文件与不含有该词的文本文件的类别。

3.2.6 Bi-LSTM 模型

（1）循环神经网络（Recurrent Neural Network, RNN）

Michael Jordan 首次在神经网络中引入了循环连接（1986）。而后，Jeffrey Elman 在 Jordan 的研究基础上，正式提出了循环神经网络（Recurrent Neural Network, RNN）模型（1990），当时叫 SRN (Simple Recurrent Network)。如图 3-3 所示，RNN 的核心思想我们可以简单理解为过去的信息被保留下来，并且会对将来产生影响。某一时间 t 的输入和前一个 cell 的输出相同，t 的输出同样也是后一个 cell 的输入。所以 RNN 实际上在 t 时刻得到的结果是基于前 t 个输入的，存在长期依赖的问题。因此，有学者提出了长短期记忆网络。

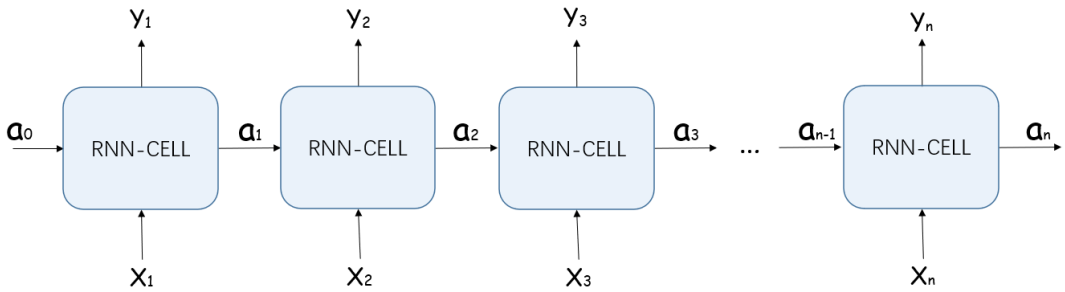

图 3-3 循环神经网络

资料来源：本文根据文献整理

（2）长短期记忆网络模型

长短期记忆网络模型（Long Short-Term Memory，LSTM）（Hochreiter S ,1997），如图 3-4 所示，也是一种循环神经网络。LSTM 中有三个门结构，

遗忘门：它可以决定上一个 cell 的信息是否保留；

输入门：它决定了哪些值需要被更新并输入；

输出门：它决定了哪些值可以被输出。

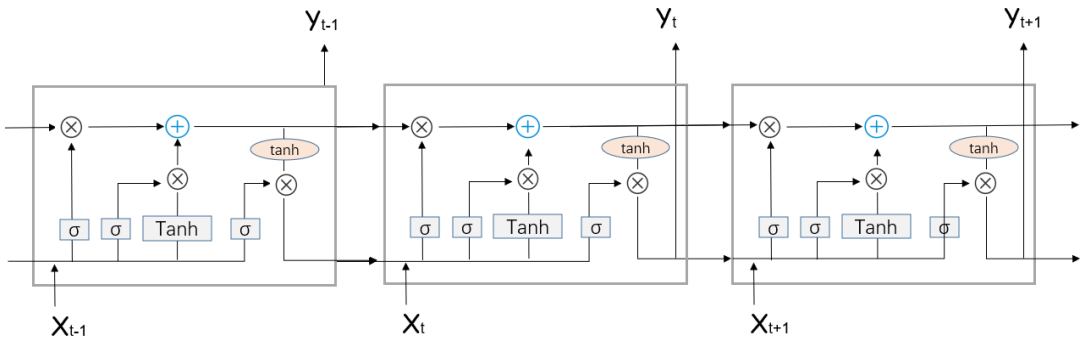

图 3-4 长短期记忆网络

资料来源：本文根据文献整理

（3）双向长短期记忆网络

双向长短期记忆网络（Bidirectional long-short term memory network，Bi-LSTM）是由正反两个长短期记忆网络构成，即当前的输出不仅会受到之前输入的影响，还会受到之后输入的影响，如图 3-5 所示。

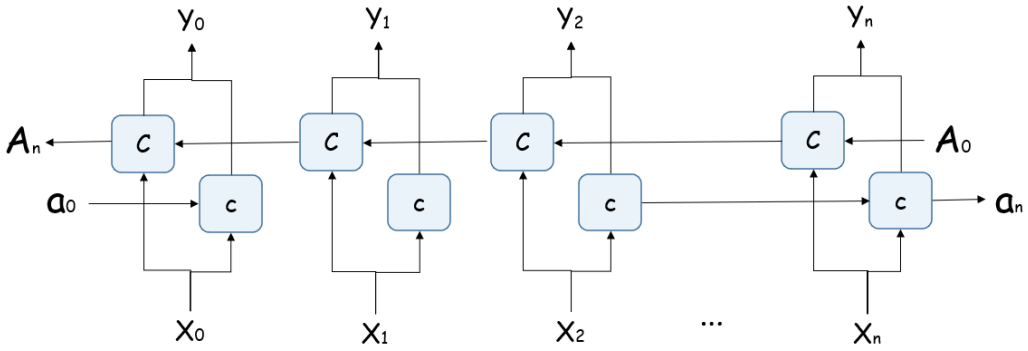

图 3-5 长短期记忆网络

资料来源：本文根据文献整理

3.3 嵌套术语识别算法设计

本文要识别的嵌套术语大部分都属于还未登录词典的新词，可以重点参考新词识别的一些方法。新词识别研究中主要有两个工作内容:候选新词的提取以及垃圾字串的过滤（张海军，2010），同理，本文的嵌套术语识别算法的主要任务就是嵌套术语的识别和垃圾串的过滤。

本文的嵌套术语识别算法流程如图 3-6 所示，首先，在网络上爬取养老领域相关的文章作为语料库，其次，通过特征工程提取语料的文本特征。接着，利用提取出来的特征设计出上下文记忆模型，利用该模型拼接分词结果 segment 并过滤垃圾串，得到候选嵌套术语。最后，人工筛选出与领域相关、有意义的术语词汇。

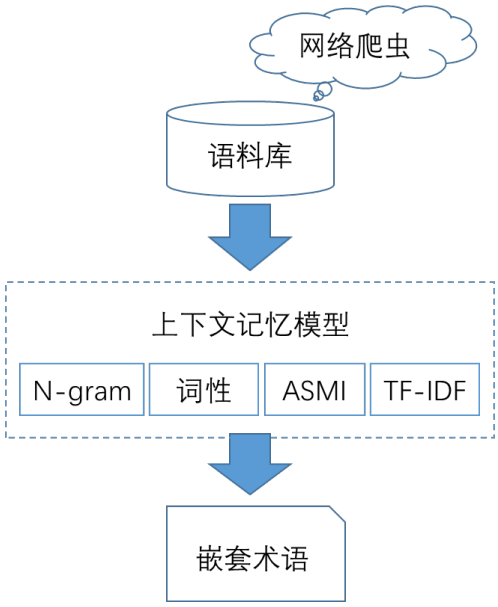

图 3-6 嵌套术语识别算法流程图

3.3.1 文本特征选取

观察本文中使用分词工具分词后的数据结果，具有将专有名词分得较短、较碎的特点。基于此，拟将分词后的段（segment）做拼接，得到完整的术语词汇。在这个过程中，文本特征是表征文本及其上下文信息的特征，文本的维度可以是字符、单词、句子、文档等。之前在第二章有提到，现有的新词识别

大多数是基于字维度及其上下文的特征。本文采用的文本维度是段（segment），segment 是指通过分词工具分词后结果，如对“我在社区老年人日间照料中心工作”这句话进行分词，分词结果为“我/在/社区/老年人/日间/照料/中心/工作”，其中“我”、“在”、“社区”、“老年人”、“日间”、“照料”、“中心”、“工作”分别是一个 segment。

1.segment 词性

JIEBA 分词工具的词性如表 3-3 所示，本文选择名词（n）、名语素（ng）、名形词（an）、地名（ns）、机构团体（nt）、其他专名（nz）、名动词（vn）、前接成分（h）、形容词（a）、时间词（t）、处所词（s）、后接成分（k）作为词性特征集合，词性特征集合外的 segment 作为天然的分隔符，不参与拼接。

表 3-3 词性表

Pos	**词性**	**Pos**	**词性**	**Pos**	**词性**
t	时间词	g	语素	d	副词
a	形容词	ag	形语素	tg	时语素
an	名形词	ad	副形词	r	代词
n	名词	ng	名语素	u	助词
nr	人名	c	连词	vg	动语素
ns	地名	e	叹词	v	动词
nt	机构团体	f	方位词	vd	副动词
nz	其他专名	p	介词	vn	名动词
s	处所词	dg	副语素	w	标点符号
h	前接成分	k	后接成分	x	非语素字
i	成语	l	习用语	y	语气词
m	数词	q	量词	z	状态词
o	拟声词	b	区别词	un	未知词

资料来源：本文根据 jieba 词性表整理

使用 jieba 对“我在社区老年人日间照料中心”这句话进行分词和词性标注，结果为：“我/r 在/p 社区/n 老年人/n 日间/t 照料/n 中心/n”，即“我/r 在/p”将不参与拼接，可视为天然的分隔符，其余部分符合词性规则，可参与 n-gram 拼接。

2.segment 数量

观察本文获取到的语料，我们发现其中有较多的嵌套术语，且长短不一，如比较短的嵌套术语“居家养老”的分词结果是“居家/养老”有 2 个 segment；比较长的嵌套术语“社区老年人日间照料中心”的分词结果“社区/老年人/日间/照料/中心”有 5 个 segment。

因此，本文在对语料进行 segment 拼接以获取较长的嵌套术语时，将 segment 数量限制为 2-5 个。

3.上下文信息

在新词识别的研究中，词的边界非常难以确定，通常要结合上下文信息共同分析和识别。本文基于点间互信息 PMI，提出相邻段互信息 ASMI（Adjacent Segment Mutual Information），公式如下：

$$ASMI（x;y）=\log\frac{\mathrm{P_{adj}(x,y)}}{\mathrm{p}(x)p(y)} \tag{3-6}$$

类似地，p(x)表示词/字 x 在语料里出现的频率，即语料里 x 的数量除以语料的总词数 N。$P_{adj}(x,y)$是指 x 和 y 相邻的概率，即 x 和 y 相邻出现的次数除以 N 的平方。

4.TF-IDF

术语具有领域专业性，往往只出现在少数的领域文章中，非专业文章非常少见，所以我们需要在领域语料中进行嵌套术语识别，并且需要筛掉一些非术语的常用的固定搭配。3.2.5 节有提到过，文本逆文本频率 TF-IDF 值可以筛掉在文档里高频出现但是没什么用的词，筛选出比较少见但是相对重要的词语。对于本文来说可以过滤掉非术语的常用固定搭配，保留频率相对较低的领域术语。

3.3.2 上下文记忆模型

计算上下文对当前单词的重要性可以确认词的边界（Song et al.2020）。因此本文参考双向长短期记忆网络结构，当前细胞的输出不仅会受到前几个细胞的输出影响，还会受到之后细胞输出的影响，提出本文的模型：上下文记忆模型（Context memory model，CMM），如图 3-7 所示。

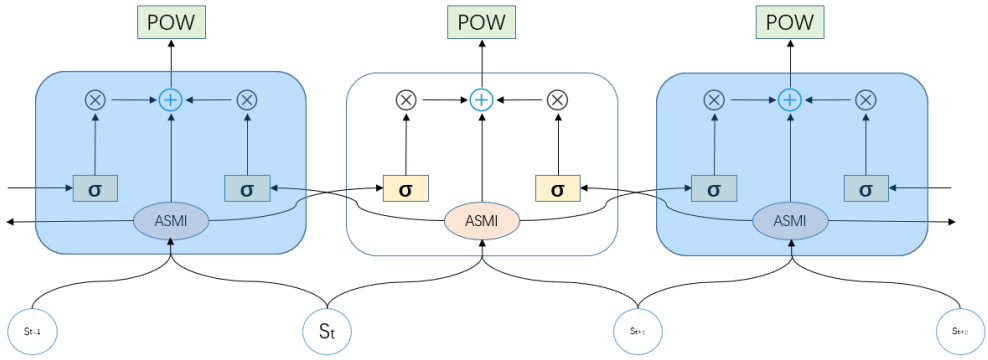

图 3-7 上下文记忆网络模型

模型中的 S 代表 segment，判断当前 segment S_t 与下文的 segment S_{t+1} 能否成词，除了需要计算 S_t 与 S_{t+1} 的 ASMI（S_t，S_{t+1}），还需要计算 S_{t-1} 与 S_t 的 ASMI（S_{t-1}，S_t）以及 S_{t+1} 与 S_{t+2} 的 ASMI（S_{t+1}，S_{t+2}），具体如下：。

（1）门结构及 sigmoid 函数

在自然语言处理中，当前词的上下文信息量是非常大的，对当前词会有一定的影响，我们需要确定这种影响的范围以及性质，需要使用激活函数 sigmoid 来判断是否保留该影响，公式如下：

$$sig = \frac{1}{1+e^{-x}}, x = \mathrm{ASMI}(S_t, S_{t-1}) \tag{3-7}$$

其中，$\mathrm{ASMI}(S_t, S_{t-1})$值越小，说明它对上下文词语的影响越小，sigmoid 越接近于 0；当$\mathrm{ASMI}(S_t, S_{t-1})$越大，说明影响越大，sigmoid 越接近于 1。影响可量化为输出值为$inf(S_t, S_{t-1})$，公式如下：

$$inf(S_t, S_{t-1}) = x * \mathrm{sig}, x = \mathrm{ASMI}(S_t, S_{t-1}) \tag{3-8}$$

（2）POW（the probability of words）成词概率

仅根据上下文 segment 的 ASMI 值无法确定能否成词，遂引入上下文特征，即左右相邻 segment 的 ASMI 值，使用如下公式计算得到成词概率POW。

$$POW(S_t; S_{t+1}) = \text{ASMI}(S_t，\ S_{t+1}) - \frac{inf(S_t，\ S_{t-1})}{N(S_t,S_{t+1})} - \frac{inf(S_{t+1}，\ S_{t+2})}{N(S_t,S_{t+1})} \quad (3\text{-}9)$$

其中，N 为 segment 的字符数量，如 N(老年人，社区)=5。 POW 越大，判断该字符串是一个词的概率就越高。

3.3.3 嵌套术语识别算法设计

基于上下文记忆模型和 TF-IDF 设计了一个嵌套术语识别算法（Nested term recognition algorithm，NTRA），算法步骤如下：

（1） 利用 jieba 对生语料分词和词性标注操作。

```
import jieba.posseg as pseg #导入 jieba 分词
import jieba.analyse

f = open('test/deeplearn.txt', 'r', encoding='utf-8')   # 待分词文本，utf-8 格式
outputs = open('dl.txt', 'w', encoding='utf-8')   # 加载待写入文本，utf-8 格式
inputs=''
for line in f.readlines():#读取待分词文本
    inputs=inputs+line
words =pseg.cut(inputs)#使用 jieba 分词和词性标注
```

（2）根据词性筛选 segment。

将表 3-4 分隔符词性表中的词性对应的 segment 作为分隔符。因为领域术语主要是以名词（n）、名形词（an）、地名（ns）、机构团体（nt）、其他专名（nz）、名动词（vn）、形容词（a）、时间词（t）、处所词（s）、名语素（ng）、前接成分（h）、后接成分（k）词性结合而来，所以在做 segment 拼接时，遇到表 3-4 中的词性对应的 segment 可跳出当前循环，继续执行下个 segment 的拼接循环。

表 3-4 分隔符词性表

Pos	**词性**	**Pos**	**词性**	**Pos**	**词性**
o	拟声词	g	语素	d	副词
b	区别词	ag	形语素	tg	时语素
q	量词	ad	副形词	r	代词
l	习用语	m	数词	u	助词
z	状态词	c	连词	vg	动语素
y	语气词	e	叹词	v	动词
un	未知词	f	方位词	vd	副动词
i	成语	p	介词	vn	名动词
dg	副语素	x	非语素字	w	标点符号

资料来源：根据 jieba 公开词性表整理

（2）使用 n-gram 模型对 segment 进行拼接。

常用的 n-gram 为 2-gram 或 3-gram，n 越大其模型的时间复杂度和空间复杂度就越高。但是本文为了尽可能多的识别出长的嵌套领域术语，n 从 2 开始递增，当 n=5 时，跳出当前循环，窗口移动到下个 segment，继续 n 元拼接操作，即 n=（2，3，4，5）。部分代码如下：

```
对语料分词后遍历所有 segment
for segment in lines:
    n=n+1
    if segment.flag in ('an','a','n','ng','nr','nrt','vn' ,'h','k','j') :#保留以上词性
        st0 = na[int(i+n-3)].word #St-1
        st = na[int(i+n-2)].word #St
        st1 = segment.word #St+1
        st2 = lines[int(n)].word    # St+2
        joint=joint+st1 #拼接 segment
          拼接相邻 segment
        ladj = str(st0) + str(st)    # St-1,St
        adj = str(st) + str(st1) #St,St+1
        radj=str(st1)+str(st2)    #St+1,St+2
```

（4）计算所有完成拼接的 segment 串的成词概率 POW 值，得到候选词集，部分代码如下：

```
#计算相邻 segment 的 ASMI
jcount = inputs.count(adj)   # 统计相邻 segment 出现次数
padj = jcount / num   # 统计相邻 segment 出现频率
px = inputs.count(st) / num   # 统计 px 频率
py = inputs.count(st1) / num   # 统计 py 频率
pxy = px * py
jf = padj / pxy
asmi = math.log(jf, 2)
#计算 sigmoid 函数
rsig = 1 / (1 + math.exp(-rasmi))# #sig（St+1,St+2）
rinf = rasmi * rsig #计算 inf
#计算 pow 值
pow = asmi - (linf + rinf)/len(joint) #计算 pow 值
```

（5）计算候选词集的 TF-IDF 值，部分代码如下：

```
ct=0
#使用本文的语料库作为 idf 计算的语料库
for yuliaopath in yuliaolist:
    with open(yuliao_path + '\\' + yuliaopath, 'r', encoding='utf-8') as fff:
        text = fff.read()
        tcj=text.count(adj)
        if   tcj> 0:
            ct = ct + 1
#计算 idf
idf = math.log((1000 / (ct + 1)), 2)
#计算 tf
tf=inputs.count(joint)/num
#计算 tfidf
tfidf=tf*idf
```

（6）筛掉小于阈值的候选词，输出所有符合上述规则的嵌套术语词汇。

（7）人工检查、筛选出领域内、有意义的嵌套术语。

3.4 算法评价

本文使用精确率、召回率、F1 值来评价本文提出的嵌套术语识别算法。

精确率：对本文来说，就是在给定的语料中，正确识别出领域术语的数量和总样本数的比值。公式如下：

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP+FN+FP}} \tag{3-10}$$

公式（3-10）中，TP 表示将术语正确识别的数量，FN 表示算法未能识别出来的术语的数量，FP 表示将非术语识别成术语的数量。

召回率：算法识别出来的术语数量与所有术语数量的比值，公式如下：

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \tag{3-11}$$

F1 值，公式如下：

$$F1 = \frac{2\mathrm{Precision}*\mathrm{Recall}}{\mathrm{Precision+Recall}} \tag{3-12}$$

本文使用吴俊等人（2020）使用的数据集“深度学习 500 问的前三章”内容（205KB，6.5w 字）作为测试数据。请两位专家标注语料，取一致结果作为标准数据。经过实验后，得到如表 3-5 所示结果。

表 3-5 术语识别算法比较

算法名称	**精确率（precision）**	**召回率（Recall）**	**F1 值**
左右熵+PMI+word2vec	0.6185	0.6564	0.6369
BiLSTM+CRF	0.8623	0.8916	0.8767
BERT-BiLSTM+CRF	0.9196	0.9343	0.9296
本文算法 NTRA	0.9144	**0.9480**	**0.9310**

如表 3-5 所示，本文提出的嵌套术语识别算法在同一测试集上精确率虽然不是最高的，但也达到了 91.44%；召回率和 F1 值是四个算法相比最高的，分别达到了 94.80%，93.10%。同时，本文提出的嵌套术语识别算法不需要进行语料标注工作，也不需要对模型进行训练，总体来说本文的算法表现最优。

3.5 本章小结

本章主要研究的是研究内容一：嵌套术语识别算法。首先介绍了主要的研究问题，其次介绍了相关的技术，然后设计了嵌套术语识别算法及模型结构，最后对嵌套术语识别算法进行了评价，实验表明，相比其他算法，本文的算法有较优秀的表现。

第4章 中文养老词库构建

本章主要阐述了养老词库的构建框架以及相关工作：数据源获取、基础词典构建、嵌套术语识别、最后输出一个养老领域词库。

4.1 研究问题

养老领域文本挖掘等文本分析的工作需要用到自然语言处理的分词技术，专业领域如果没有相应的领域词库，一些专业术语经常会被错误拆分，导致准分词准确率较低，对文本数据处理会有一定的影响。然而，目前业界还没有公开的养老领域词库，因此，本文的研究问题二就是构建一个中文的养老领域公开词库，期望对养老领域的研究和从业人员提供一定的参考。

4.2 词库构建框架

本文的养老领域词库构建框架如图 4-1 所示，主要包括四部分工作，具体如下：

1. 爬取相关语料和知识

使用 python 网络爬虫工具爬取构建基础词典需要用到的养老相关词条以及扩展基础词典需要的养老领域文章。

2. 整理基础词典

将从各种渠道获取到的养老领域词汇进行筛选、去重，整理成养老领域基础词典。

3. 从领域语料中识别出嵌套术语

使用本文在第三章中提出的嵌套术语识别算法对爬取的养老领域文章进行嵌套术语识别操作，得到嵌套术语。

4. 整理养老领域词库

将步骤三得到的嵌套术语和步骤二得到的基础词典整合去重，得到最终的养老领域词库。

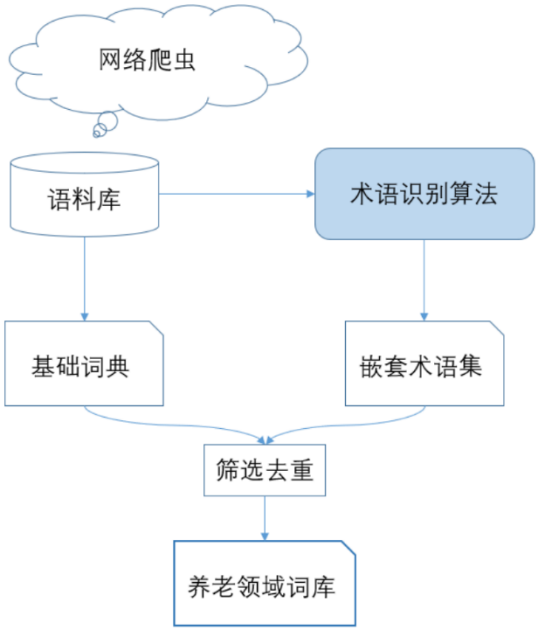

图 4-1 词库构建框架

4.3 实验

4.3.1 数据获取

一、基础词典数据

（1）贾岩于 1990 年出版的《简明老年学辞典》，内容主要包括 15 部分，如老年医学,老年人心理,老年营养学等。

李旭初等人于 2009 年出版的《新编老年学词典》，按照老年学的分支科目以及工作的实际需要分为 17 个类别,共收录辞条 1700 多条,如老年通论,老年医学、老年人保健,老年人心理学等。

筛选整理以上两本词典的词条，作为养老领域基础词典的一部分。

（2）在知乎平台搜索“养老”、“老人”、“老年”、“老龄”等关键词，获取其相关的话题词，以及知乎老年健康话题下问题中提取的相关词汇。

（3）在百度百科搜索养老相关的词语，爬取并收集所有相关的词语。

图 4-2 百度百科词条示例

资料来源：百度百科

本文在第二章的词库构建国内外研究现状分析中有提到，可以通过百科词条链接构建词库。所以本文在百度百科检索养老相关的词条，在具体的词条中会有相关词条，如图 4-2 所示，“养老”词条中有一些蓝色字体的词汇，可以链接到更多词条，但是百度百科是人工创建的，新词条的更新时效性较低，一些未创建的词条或嵌套术语使用该方法无法获取，如蓝框中的“城乡居民基本养老保险制度”这种还未来得及收录创建的嵌套术语。所以该方法获取的词汇只能作为基础词典，对于未创建的嵌套术语需要使用本文提出的嵌套术语识别方法。爬取百科词条的关键代码如下：

```
from bs4 import BeautifulSoup
import urllib.request
url = "https://baike.baidu.com/item/养老"
response = urllib.request.urlopen(url)  # 访问并打开 url
html = response.read()  # 创建 html 对象读取页面源代码
soup = BeautifulSoup(html, 'html.parser')  # 创建 soup 对象，获取 html 代码
ldiv = soup.find('div', class_="main-content")  # 定位到 div 标签，
labela = ldiv.find_all('a', target="_blank")  # 找到所有 a 标签，返回一个列表
wordlist = []
for i in labela:  # 将所有 dd 标签内容存入列表
    word = i.get_text()
    wordlist.append(word)
print(wordlist)
```

（4）获取张卓越（2020）在学位论文《养老服务本体构建及其应用研究 》中整理的养老服务本体的术语词汇，如图 4-3 所示。

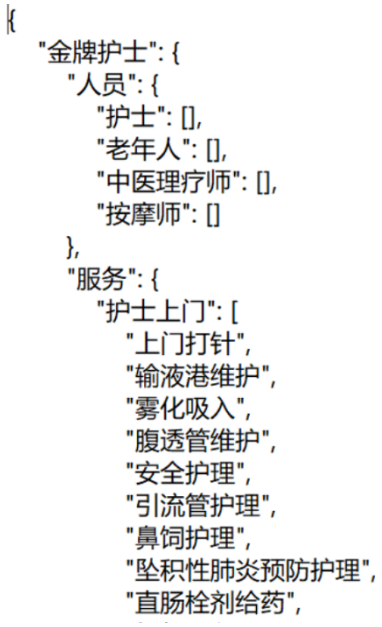

图 4-3 养老服务本体词汇部分截屏

资料来源：张卓越硕士学位论文

综合上述四种渠道获得的词汇，整理去重后得到养老领域基础词典，共计 2540 个词汇。

二、语料获取

全国中老年网，是一个由中国老龄协会老年人才信息中心主办的网站，如图 4-4 所示，网站主要包括 16 个模块，如老龄用品、政策法规、私房美食、服务机构、老年人才等。其中老龄新闻共 3000+篇，政策法规 3000+篇，才艺展示 3000+篇，老龄用品 3000+篇，老年人才 3000+篇，使用与前文提到的爬取百度百科词条相似的爬虫方法从全国中老年网爬取 1000 篇与养老领域相关的文章（3.93MB，约 130 万字）作为本文嵌套术语识别来源和逆文本频率 IDF 计算的语料库。

图 4-4 全国中老年网首页截屏

资料来源：全国中老年网

4.3.2 实验过程及结果

1. 爬取《简明老年学辞典》、《新编老年学词典》、知乎养老相关话题词、百度百科养老相关词条、养老服务本体词汇等，整理去重得到基础词典，共计 2540 个词汇，部分词汇如图 4-5 所示。

关节痛
霍乱
跟腱炎
跨国养老
骨密度
骨密度值
养老机构
肝脏虚弱
保健品
抽筋
高发人群
智能家电
老年乘车证

图 4-5 养老领域基础词典部分截屏

2.嵌套术语识别。首先，从全国中老年网爬取 1000 篇养老相关文章（3.93MB，约 130 万字）作为语料库；其次，使用本文在第三章中提出的嵌套术语识别算法 NTRA 对语料库进行嵌套术语识别操作，共识别出 10729 个养老领域的嵌套术语。

3.构建养老领域词库。将步骤 2 识别出的嵌套术语与步骤 1 整理的基础词典整合，并进行去重操作后，共获得 10293 个养老领域术语词汇。

本文构建的养老领域词库，共收纳养老领域词汇 1 万余条，共计 6 万余字。 该词库不仅包含了老年学领域，还涵盖了近些年兴起的**养老领域，共 2 个部分。其中，老年学领域可总结为 4 个大类：老年生物学、老年医学、老年心理学、老年社会学；养老领域可总结为 4 个大类：养老政策、养老服务、养老模式、养老产品**等养老相关内容，即该养老领域词库共计 8 个大类，其中，养老服务是个比较大的范畴，可再具体细分为 4 个小类：家政服务、康复护理、生活照护、老年饮食。如图 4-6 所示。

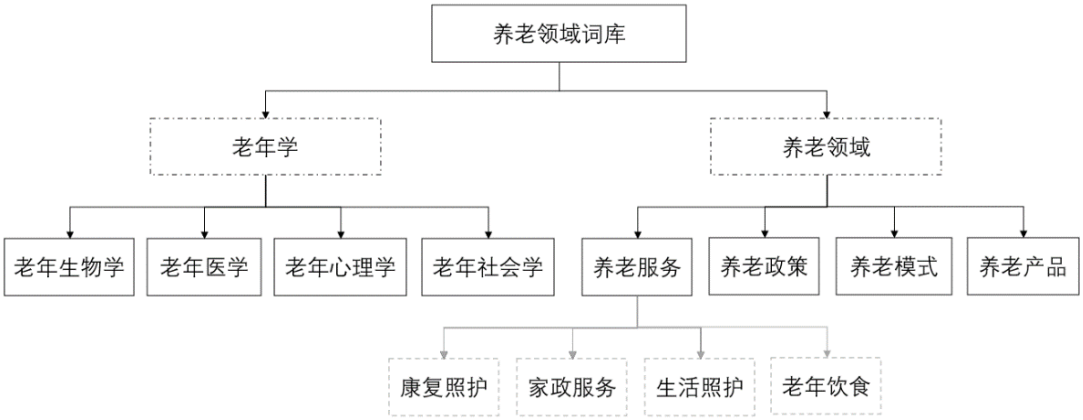

图 4-6 养老领域词库类别

根据以上类别将目前的养老领域词库进行分类统计：老年生物学含有 751 条词汇、老年医学含有 2960 条词汇、老年心理学含有 349 条词汇、老年社会学含有 4536 条词汇、养老政策含有 216 条词汇、养老服务含有 924 条词汇、养老模式含有 207 条词汇、养老产品含有 350 条词汇（数据将持续更新），各类别的部分词汇举例如表 4-1 所示。

表 4-1 词库分类举例

类别	老年社会学	老年生物学	老年医学	老年心理学	养老政策	养老服务	养老模式	养老产品
举例	人口老龄化 银发浪潮 银发经济 养老保险 基础养老金 老年人再婚 老年人再就业 老年教育 …	细胞衰老 大脑衰老 病理性衰老 抗衰老 肾脏衰老 器官衰老 消化系统 呼吸系统 生殖系统 …	老年痴呆症 老年肺炎 老年高血压 老年肺结核病 老年人贫血 …	黄昏心理 老年性抑郁症 老年性精神障碍 老年焦虑症 老年疑病症 神经衰弱 …	基本养老保险制度 城乡居民社会养老保险制度 北京市居家养老服务条例 老年人权益保障法 过渡性养老金 …	鼻饲护理 血糖监测 吸痰护理 压疮护理 灌肠护理 雾化治疗 体检陪诊 医护到家 家庭保洁 衣物清洗 冬令进补 流质饮食 …	社区居家养老 候鸟式养老 医养结合 机构养老 以房养老 旅居养老 社区养老 跨国养老 智慧养老 …	智能手环 智能水表 智能床垫 腕式血压计 助行器 康复训练器 急救腕表 智能马桶 无障碍扶手 …

4.4 词库评价

本文使用间接的评价方法对词库进行评价，即将养老领域词库应用于分词工作中，比较分词的效果，侧面对词库进行评价，根据分词效果判断养老领域词库的有用性。

4.4.1 评价数据获取

在中国新闻网、央视网等官方网站爬取养老领域文章，共计 11796 字。同样，请两位专家进行标注，取一致结果作为标准分词结果。评价数据共含有 268 个养老领域的嵌套术语，部分词汇如表 4-2 所示。

表 4-2 测试数据中部分嵌套术语

城乡居民基础养老金	银发经济	智能看护设备	护理型床位
异地就医结算	普惠型养老服务	适老化改造	农村互助幸福院
社区嵌入式养老	智慧养老	家庭养老照护床位	互助性养老
养老保险制度体系	信息无障碍建设	无围墙式健康养老	养老服务方式
基本养老保险基金	老年人权益保障	重大疾病医疗保险	日间照料中心
基本养老保险	高龄失能老年人	社区养老服务中心	半失能老人
城镇职工基本养老金	留守老年人	计划生育特殊困难家庭	基本医疗保险门诊共济保障机制

4.4.2 评价方法

使用 jieba、pkuseg、THULAC 这三种常用中文分词工具以及 jieba 与本文的养老领域词库结合的方法，如表 4-3 所示，分别对这三篇养老领域文章进行分词操作，对分词结果进行分析和比较。

表 4-3 分词工具

序号	分词工具	说明
1	jieba	是一款业界广泛使用的中文开源分词包，支持多种模式的中文分词、词性标注、关键词提取等多种功能。
2	pkuseg	是由北京大学研发的中文分词工具包。该分词工具简单易用，支持在不同的领域上进行中文分词。
3	THULAC	是由清华大学自研发的中文词法分析工具包，支持中文分词和词性标注。
4	jieba+养老领域词库	本文的分词方法

资料来源：本文根据文献资料整理

使用精确率、召回率以及 F1 值来评价分词结果。

精确率：在给定的语料中，正确识别出领域术语的数量和总样本数量的比值。公式如下：

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP+FN+FP}} \tag{4-1}$$

公式（4-1）中，TP 是指将术语正确识别的数量，FN 是指算法未能识别出来的术语的数量，FP 是指将非术语识别成术语的数量。

召回率：算法识别出来的术语数量与所有术语数量的比值，公式如下：

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \tag{4-2}$$

F1 值，公式如下：

$$F1 = \frac{2\mathrm{Precision}*\mathrm{Recall}}{\mathrm{Precision+Recall}} \tag{4-3}$$

4.4.3 评价结果

分词结果如表 4-4 所示。可以看出，引入本文构建的养老领域词库与 jieba 结合的分词工具的精确率为 86.39%，召回率为 94.51%，F1-score 为 90.16%，在各分词工具中表现最好。观察其他分词工具的分词结果可以发现，在未引入领域词库的情况下，一般的分词工具无法识别出养老领域文章中的嵌套术语。数据表明引入了养老领域词库之后明显提高了分词的准确率，可见本文构建的养老领域词库有一定的有用性，可为养老领域研究提供一定的参考。

表 4-4 分词工具比较

分词工具	**精确率 precision**	**召回率 recall**	**F1 值 F1-score**
Jieba	0.5731	0.7584	0.6526
THULAC	0.4973	0.7031	0.5824
pkuseg	0.5485	0.7430	0.6310
Jieba+养老领域词库	**0.8639**	**0.9451**	**0.9016**

然而，本文使用的分词方法并未能达到较高的准确率，一是因为 jieba 分词工具本身会存在将一些非领域词汇的常用词错误切分的情况，如“各方面”被切分为“各”、“方面”，“下一步”被切分为“下”、“一步”等；二是因为测试数据中含有少量养老领域词库未涵盖的新术语，如表 4-2 所示的“社区嵌入式养老”、“信息无障碍建设”等，说明了本文的养老领域词库还不够完善，应持续不断地扩充。

4.5 本章小结

本章主要研究的是研究内容二：养老领域词库的构建。首先介绍了主要的研究问题，其次介绍了词库构建的框架，然后阐述了实验过程，包括：数据源获取、基础词典构建、嵌套术语识别、养老领域词库整合，得到了 1 万余条养

老领域词汇，共计 6 万余字，并且数据还在持续更新。该词库包含老年学和养老领域两部分，可分为 8 个大类：老年生物学、老年社会学、老年医学、老年心理学、养老政策、养老服务、养老模式、养老产品，并分别展示了部分词汇。最后使用间接评价的方法对养老领域词库进行了评价，结果表明引入本文的词库可大大提升分词准确率，从而证明了词库的有用性。

第5章 嵌套术语分词系统的设计与实现

本文设计的分词工具使用的是基于统计分词和词典分词结合，将引入第三章提出的嵌套术语识别算法和第四章构建的中文养老领域词库。该方法既可以基于养老领域词库进行匹配分词，又可以基于嵌套术语识别算法识别新词，具有分词快，分词准确的特点。

5.1 系统搭建背景及框架

5.1.1 系统搭建背景介绍

目前市面上已有的分词工具大多数都是基于开发工具的安装包，无法直接使用并查看分词效果，可视化术语识别工具也较少。为了方便研究人员实现分词和术语识别的需要，本文搭建一个中文嵌套术语分词可视化系统，该系统可以对术语识别结果和分词结果进行可视化展示。

5.1.2 平台和开发环境介绍

本文的系统开发和运行环境如表 5-1 所示。

表 5-1 系统开发和运行环境

名称	说明
Windows10	操作系统
pycharm	开发工具
python3	编译器
Bootstrap	前端开发框架
Html5	前端语言
CSS	
JavaScript	

5.1.3 系统框架介绍

本文建设的嵌套术语分词可视化网站系统为常用的四层结构，系统框架图如图 5-1 所示。

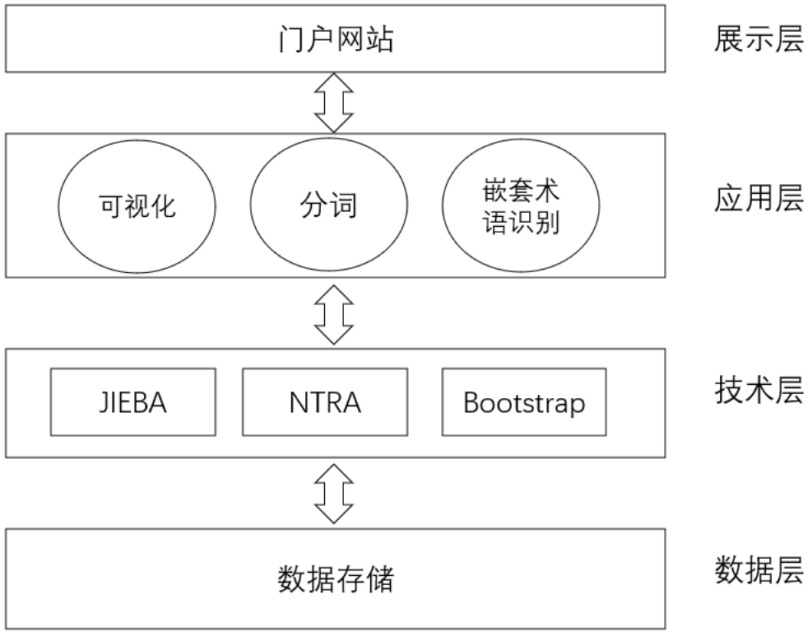

图 5-1 嵌套术语分词可视化系统框架

（1）展示层：展示层主要是通过 html5 网页的形式向用户直观地展示系统的功能，便于使用。
（2）应用层：应用层主要是为用户提供与系统的交互操作，可使用分词功能、术语识别功能。
（3）技术层：使用 Bootstrap 网站开发框架，使用 JIEBA 分词并引入本文提出的嵌套术语识别算法实现分词和术语识别的功能。
（4）数据层：数据层主要是为了给系统提供数据与规则。

5.2 系统实现

本文建设的嵌套术语分词可视化网站部分代码如下：

```
#前端显示
<form role="form" >
        <div class="form-group">
            <p align="center">请在下方输入待分词文本</p>
            <input  type="text"  class="form-control"  id="text"  style="width:1000px;
height:200px; align:center;"  />
        </div>
        <div class="checkbox">
```

```
                    <label><input type="checkbox" />识别术语</label>
                </div>
                    <button type="button" class="btn btn-default" id="btnsend" onclick="fenci()">分词</button>
                    <button type="button" class="btn btn-default" id="btnclear" onclick="clear()">清空</button>
            </form>
                <input type="text" id="result" style="width:1000px; height:200px; align:center;" readonly="readonly" >
#引入养老领域词库进行分词
jieba.load_userdict('yanglaociku.txt')
text=jieba.lcut(t)
```

本文建设的嵌套术语分词可视化网站截屏如图 5-2 所示，页面结构比较简单清晰。左上方是网站名称，右上方是中国人民大学智慧养老研究所 logo。正中央为文本输入框，可以在此输入待分词的文本。下方为识别术语功能的 checkbox，勾选该选项即可在分词结果中展示术语。最下方为分词结果展示框。

图 5-2 分词网站截屏

该网站具有分词和术语识别功能。在输入框中输入源文本，勾选“术语识别”即可使用术语识别功能，点击分词即可进行分词操作，分词结果将会在下方文本框展示，其中术语使用绿色加粗的特殊格式显示，便于研究人员查看。下面以一篇全国中老年网的养老领域文章《河北：推动居家养老仍需多方合力》为例，首先使用本文的分词网站进行分词和术语识别，分词结果如图 5-3 所示。

图 5-3 分词结果截屏

图 5-3 中，加粗标绿格式的为养老领域术语，如“老年人”、“中度老龄化社会”、“居家养老”、“河北省居家养老服务条例”等 20 个养老领域术语，在本文的网站中被正确拆分并标识，这将非常有利于后续的养老领域文本挖掘和分析工作。

图 5-4 为其他可视化分词工具的分词结果，可以看出，“中度老龄化社会”、“居家养老”、“居家服务设施”“社区老年人日间照料中心”、“居家养老服务中心”、“互助幸福院”等术语均被错误拆分，如“互助幸福院”被分为“互助、幸福、院”三个词，这会对养老领域文本分析有非常大的影响。从图 5-3 和图 5-4 的结果对比来看，在养老领域的文本挖掘和分析工作中，分词时使用本文的分词系统或在其他可自定义词典的分词工具中引入本文的养老领域词库是非常有必要的。

图 5-4 其他分词工具结果

资料来源：某在线分词网站

5.3 本章小结

本章主要阐述了中文嵌套术语分词可视化系统的构建方法，详细描述了系统的搭建环境、设计框架、系统功能及实现。该系统具有分词、术语识别的功能，能在网页上可视化展示分词结果、术语识别结果，并与其他可视化分词工具作了比较。

第6章 总结与展望

本章在国内外研究现状分析、嵌套术语识别算法设计、中文养老领域词库构建及实现中文嵌套术语分词可视化系统的基础上，对得出的研究结果进行了进一步分析，然后总结了研究成果具有的理论意义和实践意义、创新点和不足之处。最后，对可以后续研究的方向做了展望。

6.1 研究成果

本文从词库构建领域出发，结合了自然语言处理的新词/未登录词识别、术语识别、命名实体识别等领域的理论和方法，设计了嵌套术语识别算法NTRA。在嵌套术语识别算法的基础上，构建并输出了中文养老领域词库。最后结合二者实现了术语识别分词可视化系统。本文的主要研究成果如下：

1.本文通过回顾词库构建领域的现状后发现，目前国内外还没有关于中文养老领域的词库构建研究，这在养老文本挖掘中是研究空白。接着又回顾了新词/未登录词、术语、命名实体识别方法等，发现嵌套术语识别是非常值得研究的领域，而且养老领域有非常多的嵌套术语，基础词典目前无法覆盖嵌套术语，需识别出语料中的嵌套术语补充到养老领域词库中。**因此，本文基于 N-gram、PMI、Bi-LSTM 设计了上下文记忆模型 CMM 和判断嵌套术语是否成词的指标：成词概率 POW，并基于该模型提出了嵌套术语识别算法 NTRA，该算法可为新词识别、未登录词识别、术语识别、命名实体识别领域提供一些参考。**

2.本文在网络上爬取并整理了养老领域基础词典，包含 2540 个词汇。之后使用本文提出的嵌套术语识别算法识别出养老领域文章中的嵌套术语，扩展了养老领域基础词典，**构建了一个养老领域词库，该词库包括老年学和养老领域两个部分，可分为 8 个大类，分别为老年医学、老年社会学、老年生物学、老年心理学、养老政策、养老服务、养老模式、养老产品，共计 10000 余条养老相关词汇。本文提出的养老领域词库**可以为养老领域填补没有公开词库的空白，还可以为养老领域后续的文本挖掘、数据融合、知识图谱构建等研究提供词库的参考。

3.本文基于嵌套术语识别算法和养老领域词库设计并实现了中文养老领域术语分词可视化系统，该系统可以对养老领域非结构化数据，如网络新闻、文章等进行分词并且可以可视化地展示分词结果，特别是可以准确地对养老领域术语和嵌套术语进行识别，差异化地展示出该文本中包含的所有养老领域术语，可以为养老领域研究人员提供一个便捷的适用于养老领域中文文本的分词和术语识别工具。

6.2 研究不足

尽管本文取得了一些研究成果，但是受制于本人水平，本研究还有很多不足之处：

（1）由于本文爬取的语料和时间是有限的，所以识别出的嵌套术语和养老领域词库的范围以及数量有一定的局限性。

（2）本文的嵌套术语识别算法是在 jieba 分词结果基础上进行嵌套术语识别，所以嵌套术语识别结果比较依赖于 jieba 的准确性。

（3）本文搭建的术语识别可视化分词系统目前只有术语识别和分词功能，系统较简单且功能较少。

6.3 后续研究

本文将嵌套术语识别算法应用在了养老领域词库构建的工作中，将养老领域词库应用在了中文嵌套术语识别分词可视化系统的实现中。但目前来看，该系统的功能较少，系统架构也较为简单，因此有以下几个方面需要进一步研究：

（1）目前得到的嵌套术语和养老领域词库有一定的局限性，未来应使用更多的、新鲜的养老领域语料进行嵌套术语识别，不断地维护和更新养老领域词库。

（2）中文嵌套术语分词可视化系统可以增加更多的功能，如术语词频统计、词性标注、关键词提取等，也可以把该系统引入到其他自然语言处理的基础工作中。

（3）养老领域词库可以应用在文本挖掘领域，如热点识别；还可以应用在养老知识图谱、养老大数据融合等研究领域。

（4）未来应不局限于养老领域，可以用本文的词库构建方法构建其他领域的领域词库。

参考文献

[1] Beigi O M , Moattar M H . Automatic construction of domain-specific sentiment lexicon for unsupervised domain adaptation and sentiment classification[J]. Knowledge-Based Systems, 2020, 213(1–2):106423.

[2] Chen X , X Qiu, Zhu C , et al. Long Short-Term Memory Neural Networks for Chinese Word Segmentation[C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015.

[3] Dong Y , Li W , Hui Y . Intelligence Extraction Method of Domain Terms for Chinese Web Documents Based on Hierarchical Combination Strategy[C]// 2016 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS). IEEE Computer Society, 2016.

[4] Elman J L . Finding Structure in Time[J]. Cognitive Science, 1990, 14(2):179-211.

[5] Foo J,Merkel M.Using machine learning to perform automatic term recognition[C]//Proceedings of the LREC.European Language Resources Association,2010.49–54.

[6] Gao Y,Yuan Y.Feature-less end-to-end nested term extraction[C]//Proceedings of the CCF Int'l Conf.on Natural Language Processing and Chinese Computing.Cham:Springer-Verlag,2019.607–616.

[7] Goyal A , Gupta V , Kumar M . Recent Named Entity Recognition and Classification techniques: A systematic review[J]. Computer Science Review, 2018, 29(AUG.):21-43.

[8] Greenberg N , Bansal T , Verga P , et al. Marginal Likelihood Training of BiLSTM-CRF for Biomedical Named Entity Recognition from Disjoint Label Sets[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018.

[9] He T,Hao R,Qi H,et al.Mining Feature-Opinion from Reviews Based on Dependency Parsing[J].International Journal of Software Engineering & Knowledge Engineering,2017,26(9n10):1581-1591.

[10]Hochreiter S , Schmidhuber J . Long Short-Term Memory[J]. Neural Computation, 1997, 9(8):1735-1780.

[11]Jordan, M.I. Serial Order: A Parallel Distributed Processing Approach[R].Institute for Cognitive Science Report 8604, University of California, San Diego.1986.

[12] Ju M, Duan H, Li H. A CRF-based Method for Automatic Construction of Chinese Symptom Lexicon[C].International Conference on Information Technology in Medicine and Education.IEEE,2016:5-8.

[13]Ju M, Miwa M, Ananiadou S.A Neural Layered Model for Nested Named Entity Recognition[C] // Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies,2018(1):1446-1459.

[14]Katiyar A, Cardie C. Nested Named Entity Recognition Revisited[C] // Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies,2018(1):861-871.

[15]Keretna S, Lim C P, Creighton D, et al. Enhancing medical named entity recognition with an extended segment representation technique[J]. Computer methods and programs in biomedicine, 2015, 119(2): 88-100.

[16]Kong J , Zhang L , Jiang M , et al. Incorporating multi-level CNN and attention mechanism for Chinese clinical named entity recognition[J]. Journal of Biomedical Informatics, 2021, 116:103737.

[17]Li B , Yang X , Zhou R , et al. An Efficient Method for High Quality and Cohesive Topical Phrase Mining[J]. IEEE Transactions on Knowledge & Data Engineering, 2019, 31(1):120-137.

[18]Liu T , Yao J G , Lin C Y . Towards Improving Neural Named Entity Recognition with Gazetteers[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019.

[19]Mandal S,Gupta S.A novel dictionary-based classification algorithm for opinion mining[C].Second International Conference on Research in Computational Intelligence and Communication Networks.IEEE,2017:175-180.

[20]Mishra S, Sharma A. Automatic Word Embeddings-Based Glossary Term

Extraction from Large-Sized Software Requirements[C]//International Working Conference on Requirements Engineering: Foundation for Software Quality. Springer, Cham, 2020: 203-218.

[21]Morwal S，Jahan N，Chopra D.Named entity recognition using hidden markov model(HMM)[J].International Journal on Natural Language Computing,2012,1(4):15-23.

[22]Pecina P, Schlesinger P. Combining association measures for collocation extraction[C]//Proceedings of COLING/ACL on Main Conference Poster Sessions.Sydney,Australia.2006:651-658.

[23]Peng N , Dredze M . Named Entity Recognition for Chinese Social Media with Jointly Trained Embeddings[C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015.

[24]Peters M E， Neumann M， Iyyer M， et a1． Deep Contextualized Word Representations[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics，New Orleans，USA．New Y0rk，USA：ACL．2018：2227—2237．

[25]Shijia E，Xiang Y. Chinese named entity recognition with character-word mixed embedding[C]//Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, New York:ACM,2017:2055–2058.

[26]Song X， Feng A， Wang W , et al. Multidimensional Self-Attention for Aspect Term Extraction and Biomedical Named Entity Recognition[J]. Mathematical Problems in Engineering, 2020(2020):1-7.

[27]Song Y,Jeong S,Kim H.A Semi-automatic Construction method of a Named Entity Dictionary Based on Wikipedia[J]. Journal of KIISE,2015,42(11):1397-1403.

[28]Tang D Y，Wei F R，Qin B，et al.Building Largescale Twitter-specific Sentiment Lexicon:A Representation Learning Approach ［C］//Proceedings of the 25th International Conference on Computational Linguistics.New York，USA:ACM Press，2014:172-182.

[29]Wang J，Zhou J，Zhou J，et a1．Multiple Character Embeddings for Chinese Word Segmentation[C]//Proceedings of the 57th Annual Meeting of the Association

for Computational Linguistics，Florence，Italy．New York，USA：ACL，2019：210-216．

[30]Wang Q, Song Y, Liu H,et al. A Sequence Transformation Model for Chinese Named Entity Recognition[C]//Proceedings of the International Conference on Knowledge Science, Engineering and Management.Cham:Springer,2018:491-502.

[31] Wu J,Li Y.Research on construction of semantic dictionary in the football field[C].IEEE,International Conference on Software Engineering Research,Management and Applications. IEEE,2017:303-306.

[32]Wu Y C , Yang J C , Lee Y S . Chinese Word Segmentation with Conditional Support Vector In-spired Markov Models[C]// the CIPS-SIGHAN Joint Conference on Chinese Language Processing (CLP 2010). 2010.

[33]Xia C， Zhang C， Yang T , et al. Multi-Grained Named Entity Recognition[J]. the 57th Annual Meeting of the Association for Computational Linguistics (ACL), 2019.

[34]Zheng X， Chen H， Xu T . Deep Learning for Chinese Word Segmentation and POS Tagging[C]// Conference on Empirical Methods in Natural Language Processing. 2013.

[35]Zhang C , Chen Z , Hu G． A Chinese word segmentation system based on structured support vector machine utilization of unlabeled text corpus[C]//Proceedings of CIPS—SIGHAN Joint Conference on Chinese Language Processing 2010． Bejing，China，2010：221-227

[36]Zhao H,Wang F.A deep learning model and self-training algorithm for theoretical terms extraction[J].Journal of the China Society for Scientific and Technical Information,2018,37(9):923–938 .

[37]GB/T 13715-1992, 信息处理用现代汉语分词规范[S].北京：中国标准出版社，1993.

[38]常建秋,沈炜. 基于字符串匹配的中文分词算法的研究[J]. 工业控制计算机,2016.

[39]陈列蕾,方晖. 基于 Scopus 检索和 TFIDF 的论文关键词自动提取方法[J]. 南京大学学报(自然科学版), 2018(3):604-611.

[40]陈梅婕, 谢振平, 陈晓琪等. 专利新词发现的双向聚合度特征提取新方法[J]. 计算机应用, 2020, 40(3):631-637.
[41]陈曙东, 欧阳小叶. 命名实体识别技术综述[J]. 无线电通信技术, 2020, 第 46 卷(3):251-260.
[42]杜丽萍,李晓戈,于根,等.基于互信息改进算法的新词发现对中文分词系统改进. 北京大学学报(自然科学版), 2016, 52(1): 35-40.
[43]冯国明，张晓冬，刘素辉.基于自主学习的专业领域文本 DBLC 分词模型[J]．数据分析与知识发现，2018，2(5)：40．47．
[44]冯志伟. 现代术语学引论[M]. 商务印书馆, 2011.
[45]龚德山. 命名实体识别在中药名词和方剂名词识别中的比较研究[D]. 北京站中医药大学，2019.
[46]顾孙炎. 基于深度神经网络的中文命名实体识别研究[D].南京邮电大学，2018.
[47]韩冬煦,常宝宝.中文分词模型的领域适应性方法[J].计算机学报,2015,38(02):272-281.
[48]胡吉明,陈果.基于动态 LDA 主题模型的内容主题挖掘与演化[J].图书情报工作,2014,58(02):138-142.
[49]嵇友浪, 朱君, 顾晶晶等. 统计融合序列标注的电力客服术语识别[J]. 电子设计工程, 2021, 第 29 卷(2):29-33.
[50]贾岩. 简明老年学辞典[M]. 中国商业出版社, 1990.
[51]李思良, 许斌, 杨玉基. DRTE: 面向基础教育的术语抽取方法[J]. 中文信息学报, 2018, 第 32 卷(3):101-109.
[52]李伟卿,王伟军.基于大规模评论数据的产品特征词典构建方法研究[J].数据分析与知识发现,2018,2(1):41-50.
[53]李旭初，刘兴策. 新编老年学词典(精)[M]. 武汉大学出版社, 2009.
[54]李雁群, 何云琪, 钱龙华等. 中文嵌套命名实体识别语料库的构建[J]. 中文信息学报, 2018, (8):19-26.
[55]刘浏, 王东波. 命名实体识别研究综述[J]. 情报学报, 2018, 第 37 卷(3):329-340.
[56]刘一宁, 申彦明. 基于终身机器学习的主题挖掘与评分预测联合模型[J]. 计

算机工程, 2019, 第 45 卷(6):237-241，248.
[57]钱智勇,周建忠,童国平,等. 基于 HMM 的楚辞自动分词标注研究[J]. 图书情报工作,2014,58(4):105-110.
[58]石玉鑫,杨泽青,赵志滨,姚兰.一种面向商品评价对象挖掘的领域词典构建法[J].软件工程,2019,22(01):1-7.
[59]唐琳,郭崇慧,陈静锋.中文分词技术研究综述[J].数据分析与知识发现,2020,4(Z1):1-17.
[60]吴俊, 程垚, 郝瀚等. 基于 BERT 嵌入 BiLSTM-CRF 模型的中文专业术语抽取研究[J]. 情报学报, 2020, 第 39 卷(4):409-418.
[61]徐豪杰, 吴新丽, 杨文珍等. 基于改进 PMI 和最小邻接熵结合策略的未登录词识别[J]. 计算机系统应用, 2020, 第 29 卷(6):181-188.
[62]许浩亮, 李雁群, 何云琪等. 中文嵌套命名实体关系抽取研究[J]. 北京大学学报(自然科学版), 2019, 第 55 卷(1):8-14.
[63]徐建民,王金花,马伟瑜.利用本体关联度改进的 TF- IDF 特征词提取方法[J].情报科学,2011, 29(2): 279- 283.
[64]夏松, 林荣蓉, 刘勘. 网络谣言敏感词库的构建研究——以新浪微博谣言为例[J]. 知识管理论坛, 2019, v.4;No.23(05):5-13.
[65]杨双龙，吕学强，李卓等. 中文专利文献术语自动识别研究［J］. 中文信息学报，2016，30（3）:111-117，124.
[66]杨秀璋,夏换,于小民,武帅,赵紫如,窦悦琪.基于特征词典构建和 BIRCH 算法的中文百科文本聚类研究[J].计算机时代,2019(11):23-27+31.
[67]杨鑫,杨云帆,焦维,朱东霖,郑绍阳,袁中玉,杨秀璋,罗子江.基于领域词典的民宿评论情感分析[J].科学技术与工程,2020,20(07):2794-2800.
[68]夭荣朋, 许国艳, 宋健. 基于改进互信息和邻接熵的微博新词发现方法[J]. 计算机应用, 2016, 36(10):2772-2776.
[69]尹文科,朱明,陈天昊.基于 Wiki 链接结构图聚类的领域词典构建方法[J].小型微型计算机系统,2014,35(06):1286-1292.
[70]赵耀全, 车超, 张强. 基于新词发现和 Lattice-LSTM 的中文医疗命名实体识别[J]. 计算机应用与软件, 2021, 38(1):161-165，249.
[71]张春燕. 基于概率依赖关系的命名实体识别方法研究[D].北京交通大学，

2019.
[72]张海军,史树敏,朱朝勇,黄河燕.中文新词识别技术综述[J].计算机科学,2010,37(03):6-10+16.
[73]张洪刚, 李焕. 基于双向长短时记忆模型的中文分词方法[J]. 华南理工大学学报(自然科学版), 2017, 第 45 卷(3):61-67.
[74]张梅山,邓知龙,车万翔,等. 统计与词典相结合的领域自适应中文分词[J]. 中文信息学报,2012,26(2):8-13.
[75]张瑞东. 基于循环神经网络的中文命名实体识别研究[D].北京工业大学，2018.
[76]张文静，张惠蒙，杨麟儿等. 基于 Lattice—LSTM 的多粒度中文分词[J]. 中文信息学报，2019，33(1)：18—24.
[77]张雪, 孙宏宇, 辛东兴等. 自动术语抽取研究综述*[J]. 软件学报, 2020, 第 31 卷(7):2062-2094.
[78]张卓越. 养老服务本体构建及其应用研究[D].中国人民大学，2020
[79]周咏梅,阳爱民,杨佳能.一种新闻评论情感词典的构建方法[J].计算机科学,2014,41(08):67-69+80.
[80]朱艳辉, 刘璟, 徐叶强等.基于条件随机场的中文领域分词研究[J]. 计算机工程与应用, 2016, 第 52 卷(15):97-100.
[81]中国共产党中央委员会.中共中央关于制定国民经济和社会发展第十四个五年规划和二〇三五年远景目标的建议.[2020-11-03]http://www.gov.cn/zhengce/2020-11/03/content_5556991.htm
[82]中国互联网络信息中心. 第 47 次中国互联网络发展状况统计报告[R].2021

致谢

时光飞逝，三年的研究生学习生活就快要结束了。在这三年的学习和生活中，我从老师、师门的师兄师姐师弟师妹以及同学们的身上都学到了很多宝贵的知识和经验。

首先我要由衷地感谢我的学术导师左美云教授和企业导师王杰老师。感谢左老师为我创造了良好的学习环境和条件，以及实验室里小伙伴们浓厚的学习氛围。还要感谢左老师对我的学业和论文研究提供的帮助与指导，当我迷茫时左老师会给我指明方向，当我懈怠时左老师会激励我前行。。感谢王老师为我提供的实习工作机会和对论文研究的建议。左老师和王老师学识渊博、见多识广、兢兢业业、平易近人，是我们工作和生活中的好榜样。左老师经常教导我们：“不用扬鞭自奋蹄”，我们会永远铭记于心。左老师和王老师的教诲将激励我在以后的工作和学习中努力拼搏，正确做一个对社会和国家有用的人。

同时，也要感谢周季蕾老师和实验室的兄弟姐妹们，生活中我们一起玩耍，学习上我们互帮互助。非常感谢他们在组会上和私下里对我的论文研究提的宝贵建议。希望师兄师姐们可以多发论文，师弟师妹们学习成绩优异，祝大家都能顺利毕业，找到理想的工作。

其次，要感谢我身边的朋友们，感谢你们的爱护和照顾。希望我今天拥有的，永远都不会失去。

最后，要感谢我的父母和亲人，感谢父母对我的养育之恩、辛苦赚钱供我读书，但是我却不能常伴他们左右，希望将来有时间能够常回家看看，带他们游遍中国吧。

鉴于本人研究水平有限，对养老领域和术语识别领域的研究也只是基于文献阅读和实验室工作的总结和心得，还有很多不足之处，今后将在实际工作中结合理论知识继续思考和总结。

附录

第五章中术语识别可视化分词网站部分代码：

```
<!DOCTYPE html>
<html lang="zh-CN">
  <head>
    <meta charset="utf-8">
    <meta http-equiv="X-UA-Compatible" content="IE=edge">
    <meta name="viewport" content="width=device-width, initial-scale=1">
    <title>分词网站</title>
    <link href="bootstrap/css/bootstrap.min.css" rel="stylesheet">
  </head>
  <body>
<div class="container">
   <div class="row clearfix">
      <div class="col-md-6 column">
         <h3>
                  嵌套术语分词网站
         </h3>
      </div>
      <div class="col-md-6 column">
         <img src="logo.bmp"    align="right"/>
      </div>
   </div>
   <div class="row clearfix" align="center">
      <div class="col-md-12 column">
         <ul class="nav nav-tabs">
            <li class="active">
               <a href="#">首页</a>
            </li>
            <li>
               <a href="#">简介</a>
            </li>
            <li class="disabled">
               <a href="#">信息</a>
            </li>
            <li class="dropdown pull-right">
```

```
                    <a  href="#"  data-toggle="dropdown"  class="dropdown-toggle"> 下 拉 <strong
class="caret"></strong></a>
                    <ul class="dropdown-menu">
                        <li>
                            <a href="#">更多设置</a>
                        </li>
                    </ul>
                </li>
            </ul>
            <form role="form" >
                <div class="form-group">
                    <p align="center">请在下方输入待分词文本</p>
                            <br>
                            <input  type="text"  class="form-control"  id="text"    style="width:1000px;
height:200px; align:center;"   />
                </div>
                <div class="checkbox">
                    <label><input type="checkbox" />识别术语</label>
                </div>
                     <button type="button" class="btn btn-default" id="btnsend" onclick="fenci()">分
词</button>
                     <button type="button" class="btn btn-default" id="btnclear" onclick="clear()">清
空</button>
            </form>
            <h2>
                分词结果
            </h2>
                <input  type="text"  id="result"  style="width:1000px;  height:200px;  align:center;"
readonly="readonly" >

            <p align="center">
                @中国人民大学智慧养老研究所
            </p>
        </div>
    </div>
</div>
  </body>
```