|  |  |
| --- | --- |
| 学号 | 2013302730014 |
| 密级 |  |

**武汉大学本科毕业论文**

|  |
| --- |
| **一种面向医学指南的知识图谱生成系统**  **设计与实现** |

|  |  |
| --- | --- |
| 院（系）名 称 ： | 计算机学院 |
| 专 业 名 称 ： | 物联网工程 |
| 学 生 姓 名 ： | 谭冉 |
| 指 导 教 师 ： | 何扬帆 副教授 |

二〇一七年五月

**郑 重 声 明**

本人呈交的学位论文，是在导师的指导下，独立进行研究工作所取得的成果，所有数据、图片资料真实可靠。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确的方式标明。本学位论文的知识产权归属于培养单位。

本人签名： 日期：

摘 要

医学指南是权威、详实的指导性医学文献，在医学实践活动中具有非常重要的指导意义，然而受制于其纸质文本的存在形式，在实际使用上有诸多不便，使用率很低，而知识图谱概念的应用有助于改变医学指南的窘迫现状。知识图谱是存储特定领域各种实体概念以及实体间关系的知识库，通常以本体文件形式存储，是一种结构化的概念模型。根据知识图谱的思想，将非结构化的医学指南文本转换为结构化的本体可以为医生使用医学指南提供很大的便利。

人工构建医学知识图谱需要构建者拥有很高的相关领域内的知识理解，而且费时费力，如果能够使用计算机进行辅助有望加快这一过程的实现。

本文设计并实现了一个面向医学指南的知识图谱生成系统，用于辅助医学专家构建医学指南的本体，具有结构性信息抽取、快速填充本体两大核心功能，可以在专家构建本体的过程中简化很多工作。

关键词：医学指南；知识图谱；自然语言处理；语义网；

**ABSTRACT**

Medical guidelines as authoritative and informative guide of medical literature, has a very important guiding significance in medical practice, but is subject to the existing form of paper text, there are a lot of inconvenience in the actual use, the use rate is very low, and the emergence of distress situation of knowledge map concept can change the medical guidelines. Knowledge map is a kind of knowledge base, which stores the concepts of various entities and the relationships among entities in a specific domain. According to the ontology of the knowledge map, it is very convenient for doctors to use the medical guide to convert the unstructured medical guide text into structured ontology.

The artificial construction of medical knowledge mapping has the need to build knowledge related fields in the very high, and time-consuming, if can use the computer should be able to improve the efficiency of the process a lot.

In this paper, the design and implementation of a guideline for medical knowledge map generation system for auxiliary medical experts to build medical guidelines with structural information extraction, ontology, ontology filled two core functions, can be in the process of Constructing Ontology in experts to simplify a lot of work.

**Key Words**: Medication guide; Knowledge graph; Natural language processing; Semantic Web;

目 录

[1 绪论 1](#_Toc481574809)

[1.1 研究背景和意义 1](#_Toc481574810)

[1.2 国内外发展趋势和研究现状 2](#_Toc481574811)

[1.2.1 医学指南 2](#_Toc481574812)

[1.2.2 知识图谱 3](#_Toc481574813)

[1.3 研究内容与意义 4](#_Toc481574814)

[1.4 本文的组织结构 5](#_Toc481574815)

[2 关键工具与平台 6](#_Toc481574816)

[2.1 本体构建工具 6](#_Toc481574817)

[2.1.1 Protégé简介 6](#_Toc481574818)

[2.1.2 Protégé 优点 7](#_Toc481574819)

[2.2 Jena 语义网框架 7](#_Toc481574820)

[2.2.1 Jena简介 7](#_Toc481574821)

[2.2.2 Jena框架 7](#_Toc481574822)

[2.2.3 Jena接口介绍 8](#_Toc481574823)

[2.3 THULAC分词工具包 9](#_Toc481574824)

[2.3.1 THULAC简介 9](#_Toc481574825)

[2.3.2 THULAC使用 9](#_Toc481574826)

[2.4 本体构建方法 10](#_Toc481574827)

[2.5 本章小结 14](#_Toc481574828)

[3 系统需求分析 15](#_Toc481574829)

[3.1 系统需求 15](#_Toc481574830)

[3.2 系统功能模块划分 15](#_Toc481574831)

[3.2.1 文本处理模块 15](#_Toc481574832)

[3.2.2 操作界面模块 16](#_Toc481574833)

[3.2.3 本体填充模块 17](#_Toc481574834)

[3.3 本章小结 17](#_Toc481574835)

[4 系统设计 18](#_Toc481574836)

[4.1 系统架构设计 18](#_Toc481574837)

[4.2 系统功能模块设计 19](#_Toc481574838)

[4.2.1 文本结构性信息抽取模块 19](#_Toc481574839)

[4.2.2 分词处理模块 22](#_Toc481574840)

[4.2.3 本体填充模块 24](#_Toc481574841)

[4.3 本章小结 24](#_Toc481574842)

[5 系统的部分实现 25](#_Toc481574843)

[5.1 目录/文件选择页面 25](#_Toc481574844)

[5.2 文本处理页面 27](#_Toc481574845)

[5.3 本体构建结果展示 28](#_Toc481574846)

[5.4 本章小结 30](#_Toc481574847)

[6 总结与展望 31](#_Toc481574848)

[6.1 工作总结 31](#_Toc481574849)

[6.2 研究展望 31](#_Toc481574850)

[6.2.1 分词处理 31](#_Toc481574851)

[6.2.2 结构性信息抽取 32](#_Toc481574852)

[参考文献 33](#_Toc481574853)

[致谢 37](#_Toc481574854)

# 绪论

## 研究背景和意义

如何提高医疗服务的质量和效率是全世界范围内共同关注的热门话题，医疗服务质量作为一项重要的民生指标，在发达国家和很多发展中国家中都备受重视。高质量的医疗服务不仅依赖于医生的诊疗水平，更是和配套的医疗条件、设施以及医疗信息化的程度密切相关。

医学指南在进行临床诊疗时有很高的参考价值，可以帮助医生对特定的临床问题做出正确的判断和处理[4]。医学指南内容丰富，对疾病的病因、症状、诊断、治疗等问题有具体的描述和系统的归纳[2]，贴合临床医生的工作实际。然而受限于文本的形式，在临床实践过程中使用医学指南的操作比较繁琐，所以医学指南在实际临床实践过程中的使用率很低。另外，由于医学指南的内容主要采用自然语言描述，缺乏计算机化的表现形式，在具体的诊疗过程中，难以与电子病历等结构化的医学信息资源进行整合。所以，为了推进医学指南的应用，需要研究如何将指南的结构化内容表现出来[5]，为医学专家做进一步的确认和补充奠定基础。

知识图谱是存储特定领域各种实体概念以及实体间关系的知识库[3]，通常以本体文件形式存储[9]，是一种结构化的概念模型。知识图谱于2012年5月17日被Google正式提出，用于增强搜索引擎的能力，提高搜索质量，改善用户体验[3]。随着语义技术[23]、智能信息处理技术等相关技术的发展，知识图谱的应用已经不再局限于智能搜索领域，而是延伸到了包括学科领域知识构建在内的各个领域。将学科领域知识构建成知识图谱不仅为将来“Web3.0”时代知识互联奠定了基础，也在当下为学科领域知识指导具体实践提供了方便。

常见的文献、资料等都是用自然语言描述的非结构化的信息，如何将非结构化的信息转换成结构化的知识框架并构建成本体是一个值得研究的问题[16]。如果采用人工构建的方法，首先需要在相关领域具有清晰、深刻理解的专家来进行知识的抽取；其次人工构建非常地费时费力，虽然正确率有很大的保障但是总体效率比较低。如果要用计算机全自动地进行构建则需要极高的自然语言处理技术的支持，在医学领域有很多的专业名词，要做到正确识别可能还需要对应领域的词库支持。综上，一种能够起到辅助专家的作用的半自动化的知识图谱生成系统无疑是很有价值的。

本文以将对医学指南结构化内容抽取方法进行研究，选择合适的方法从文本资料中抽取出关键性信息，为后期医学专家的进一步确认与完善提供支持。

## 国内外发展趋势和研究现状

### 医学指南

医学指南是一系列权威指导意见的整合，包含着医学理论知识和科学的临床诊疗建议[27]。医学指南从临床适用的角度出发，给临床医师提供明晰的诊疗指导，在理论知识与临床实践中架设了一座桥梁[1]。目前为止，医学指南一直是以语言对疾病症状、诊断意见、治疗方案选择等进行论述，以文本形式保存交流。虽然目前存在着很多中文或外文的医学指南，然而实际考察研究发现，这些临床指南并没有被很好地应用于临床实践，规整医疗工作者们的诊疗行为[29]。虽然医生普遍意识到临床指南的重要性和价值，但是限于文本的形式，在临床实践中从医学指南获取所需内容的效率比较低，所以只有很少数医生会选择在临床实践活动中使用医学指南，而且只在遇到复杂病情的时候才使用[18]。

研究表明，纸质医学指南不被使用的原因在于阅读使用起来不方便，涉及很多医学知识，难以记忆并应用。这些因素使得临床指南无法真正有效和实时地应用于临床实践[6]。

20世纪90年代之后，医疗信息化在提高医疗服务质量和效率、降低医疗卫生服务成本等方面具有显著成效。通过对纸质临床指南的数字化表达方法进行研究[31]，可以使得临床知识逐渐满足临床决策支持系统发展的需要和实际需求；使得临床研究向临床实践的转化；同时使得临床诊疗行为更加信息化、规范化和自动化；并促进临床指南的共享。与文本形式的临床指南相比，计算机可识别的数字化临床指南具有以下优势：（1）与患者电子病历数据相结合[7]，计算机根据临床指南可以自动地为医护人员提供诊疗建议，规范临床诊疗行为；（2）根据数字化临床指南为医护人员提供及时的决策支持，辅助临床实践，为医护人员减轻负担；（3）方便医护人员随时查找自己所需的临床指南，有利于医护人员自身的再学习过程[8]。

### 知识图谱

知识图谱（Knowledge Graph）作为一种新的知识表示方法，属于语义网范畴[10]，用于描述真实世界中存在的各种实体、概念，以及这些实体、概念间的关系[11]，抽取并表示领域概念之间的语义关系。知识图谱主要采用（实体1，关系，实体2）的三元组形式表示知识，实体（entity）对应现实世界中的各种事物或抽象概念，在知识图谱中以节点存在；关系（relation）表示实体之间的联系，在知识图谱中以节点间的连线存在。2012年5月，Google公司将知识图谱引入搜索引擎是大规模知识图谱的首次应用，Google的艾米特·辛格尔博士认为“The world is not made of strings , but is made of things.”推动大规模知识图谱的研究、应用进入高潮[24]。

知识图谱构建是知识图谱发展的前提，涉及了实体的抽取和实体间关系的建立，同时还需要很好地组织存储这些抽取到的实体及关系，便于快速访问、操作。在知识图谱的创建上有两种基本的方法[22]。一种是自下而上的方法，首先要收取大量的文本数据，然后通过自然语言的方法进行分词，实体发现和关系发现等处理，在这个基础上来创建特定应用论域的知识图谱[25]。有一些情况下，论域中没有具有权威性的概念体系，那么这种自下而上的方法比较适用。而有些领域已经发展了很多年，概念体系相对比较完善，那么根据权威文献建立概念体系，自上而下建立知识图谱则是更为合适的一种选择[26]。

目前国外比较成熟的大规模通用知识图谱有：WordNet[19]、FreeBase[20]、DBpedia[21]等，在知识图谱的应用领域有：Google的知识图谱、微软的Knowledge Vault[15]，利用多数据融合设计并被应用于其Satori搜索引擎；同时，知识图谱在生物医学领域也有一定建树。在国内，近几年知识图谱的研究也有了一定的进步并取得了不少成果，例如搜狗的知立方[22]、百度知心等。不过总的来说国内在相关领域的研究起步较晚，仍处在比较初级的阶段，和国外的水平还是有一定的差距，比如国外已经建立了许多本体库，可以从中下载、学习已经建立好的本体文件，国内在这个方面的积累要落后很多。

随着近几年大数据的不断发展以及国内本体研究整体理论水平的不断提高，国内本体的研究一定会更上一层楼。将本体的方法与医学指南结合，就是知识图谱在医学领域的又一探索性研究。

## 研究内容与意义

本文针对医学指南缺乏结构化表达的问题展开研究，分别从本体构建方法和自然语言处理两个角度对医学指南知识图谱构建的方法进行探究[32]。从本体构建的角度，根据本体构建的重用原则[17]，从本体库中找到相似本体，对已有本体进行修改后得到一个知识结构框架，在这个基础上，通过对医学指南的文本内容进行实体关系抽取并填充进这个知识图谱骨架中获得需要的医学指南知识图谱[12]。从自然语言处理[36]的角度[14]，根据医学指南行文规则对文本内容进行分段并识别抽取出主干内容[30]，以三元组的形式对这些内容进行表述，在人工确认无误后，把抽取出的这些实体关系三元组构建成为本体，然后用分词工具把分段后的内容按词性分词，分词后再从其中筛选出可能正确的实体、关系以供专家确认，经确认正确的内容会被填充入本体中。

本文的主要工作包括：

（1）介绍了医学指南的重要性以及医学指南在实际应用中存在的问题。

（2）简要介绍了知识图谱的定义、构建方法以及当前知识图谱的发展现状。

（3）对本文选取的工具：Protégé本体编辑器、Jena语义网框架[38]、THULAC分词工具包[38]进行介绍并分析其功能和特点。

（4）根据本体构建的重用原则，运用翻译法[40]将国外本体库中的一个英文心血管疾病本体转换成了中文本体。

（5）以《心血管疾病诊疗指南》为例，分析医学指南行文规律，制定了一种比较可靠的医学指南知识图谱构建方法，并将其与使用分词工具进行实体关系抽取的方法进行比较，最后综合两种方法设计了一个医学指南知识图谱生成工具。

（6）对本文的研究进行总结和展望。

本文的意义在于运用现有的一些工具，设计了一个能够从医学指南文本中提取出比较可靠的结构性内容和不完全可靠的非结构性内容的医学指南知识图谱生成工具，该工具可以简化专家创建医学指南知识图谱的部分工作，提高本体创建的效率。

## 本文的组织结构

第一章绪论，主要描述了本课题涉及的医学指南和知识图谱技术的发展现状、背景，并介绍了本文的研究内容与意义。

第二章关键技术、工具与平台，对本体的构建方法、本文用到的工具：Protégé本体编辑器、Jena 语义网框架、THULAC分词工具包进行介绍。

第三章系统需求分析，根据系统要完成的目标对系统需要哪些模块并且分别实现哪些功能进行分析。

第四章系统设计，详细描述系统的设计思路、各个模块的功能实现方法设计。

第五章系统部分实现，以《心血管疾病诊疗指南》中的某章节为例，介绍使用本文设计的系统如何工作以实现辅助用户构建医学知识图谱的目标。

第六章总结与展望，总结本文所做的研究，并对一些未能实现的想法做展望。

# 关键工具与平台

## 本体构建工具

随着对本体的研究不断深入，人们越发认识到本体在知识检索、人工智能等领域能够发挥的重要作用，因此对本体的需求也在不断提升，相应的用于支持本体建设、维护和应用的本体构建工具应运而生。

本体构建工具之于本体文件就相当于Word软件之于doc文档，故而本体构建工具也被称为本体编辑，它为用户屏蔽了底层具体的本体描述语言，方便了用户对本体的创建、编辑等一系列操作，而且具有在不同类型的本体语言格式之间进行转换的功能，帮助人们更方便快捷地进行本体的开发、应用[13]。

目前本体构建工具有很多，表2.1罗列了其中比较常见的几种：

|  |  |
| --- | --- |
| 表2.1 常见的本体构建工具 | |
| 名称 | 功能简介 |
| Protégé[41] | 斯坦福大学医学院生物信息研究中心开发的开源、免费本体构建工具，采用图形化界面，可通过插件进行功能扩展，操作简单友好，支持中文 |
| Ontolingua[42] | 斯坦福大学知识系统实验室（KSL）开发的本体开发环境，应用满足面向对象的框架视图表达和浏览知识，支持合作开发本体 |
| WebOnto[43] | 英国Open University开发的基于Web的本体编辑器，使用户可以合作地浏览、创建和编辑本体，并且提供多重继承和锁机制 |
| OntoEdit[44] | 德国卡尔斯鲁厄大学开发的本体构建工具，使用图形方法支持本体的开发和维护，将本体开发方法论与合作开发和推理的能力相结合，提供对于本体的并发操作 |
| LinKFactory[45] | 基于Java的商品化本体编辑器，具有多用户协作支持能力和高度的可伸缩性，全面支持OWL，处理时间短，支持多语言方法 |

本文选用的是Protégé本体构建工具，下面简要介绍Protégé工具。

### Protégé简介

Protégé是由斯坦福大学医学院生物信息研究中心基于Java语音开发的一款开源本体编辑工具，可以从protege.stanford.edu官网上免费下载[41]。Protégé历史悠久、延续性好，有很多本体编辑工具都是受到Protégé的设计思路的启发而开发出来。Protégé起源于1989年，由最初的一个简单的应用程序，不断地更新发展到如今的支持用户界面操作、跨平台的本体建模工具，其间发布了很多的版本，最新的版本是刚发布的Protégé 5.2.0。本文在进行本体编辑、查看时使用的是Protégé 5.1.0版本。

### Protégé 优点

（1）具有很强的扩展功能。Protégé提供了很多可下载的插件，通过这些插件可以使用Protégé的更多功能，比如可以通过多种可视化功能插件实现本体的各有侧重的图形化显示。

（2）开发界面友好。Protégé的界面风格简洁，而且还可以通过用户定制以适用不同的语言开发。Protégé中的本体以树形结构展示，结构清晰明了，方便用户查看、编辑，操作简单易于上手。

（3）可以对多种格式本体描述语言描述的本体进行操作，包括XML、RDF/RDFS、OWL等本体格式，而且在本体文件导出时可以便捷地切换格式。

（4）支持中文，使用GraphViz获取本体中概念的中文关系，为中文本体构建用户提供了便利。

由于Protégé是开源、免费的本体构建工具，而且使用起来功能完善、操作简单，并有配套的详细的开发手册帮助，因而Protégé深受国内外本体研究这的喜爱。综合以上所罗列的各种优点，本文选择Protégé作为编辑、查看医学指南本体文件的工具。

## Jena 语义网框架

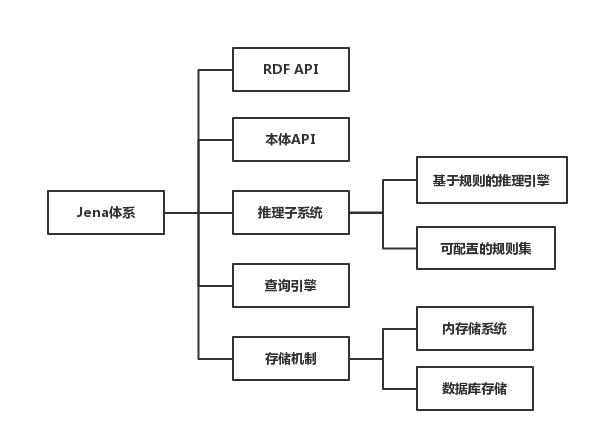
### Jena简介

Jena是用于构建语义Web和链接数据应用程序的免费开源Java框架。该框架由不同的API组成，以一起处理RDF数据，可以在jena.apache.org官网上下载最新或历史版本并查看对应的帮助文档。Jena提供了一个处理OWL、RDF等本体文件语言文件的开发平台，帮助用户对本体文件进行一些操作。Jena对于应用程序就好比Protégé对于我们：我们用Protégé进行本体的编辑，而应用程序则可以通过Jena提供的方法对本体进行操作[46]。

### Jena框架

Jena提供了一些帮助用户进行本体编辑的工具，包括本体解析器（用于不同本体语言格式文件进行解析、转换）、查询模型、存储、推理子系统等，用户可以使用Jena提供的类方便地操作本体。Jena的主要框架如图2.1所示。

图2.1 Jena体系结构图



（1）RDF应用程序接口（RDF API）：RDF API提供的接口帮助用户对RDF模型进行创建、读写、查询等操作。

（2）本体API能够处理基于RDF的本体数据，对于OWL、DAML+OIL和RDFS等格式都可以支持。通过与推理子系统结合能够将本体中的信息提取出来。

（3）推理子系统：Jena支持基于规则的简单推理，其推理机制支持将推理器导入Jena，创建模型时将推理器与模型关联以实现推理[50]。

（4）存储机制，Jena允许将数据存储到硬盘中，或OWL文件中，或关系数据库中[49]。

### Jena接口介绍

本文使用Jena用于操作OWL本体文件[48]，在这里介绍几个相关的Jena的类和接口。

（1）本体模型OntModel

本体模型由Jena RDF模型扩展而来，能够对本体数据进行操作。使用Jena的第一步就是创建本体模型，然后使用本体模型中提供的方法来对本体模型进行操作，例如：创建类、提取本体的信息、从磁盘文件中提取本体或将本体输出到磁盘文件中。

（2）接口OntClass

OntClass定义了一系列关于本体中类的方法，例如提取类的属性信息、添加或修改属性、增删子类等等。

（3）基本本体类型OntResource

本体API中表示本体的类均继承自OntResource，OntResource中可以放置所有类公用的功能。

## THULAC分词工具包

### THULAC简介

THULAC（THU Lexical Analyzer for Chinese）是清华大学自然语言处理与社会人文计算实验室研制推出的一套免费开源的中文词法分析工具包，具有中文分词和词性标注功能。THULAC分别有C++、Java、python三个版本可以在多种编程环境下运行。THULAC工具包有三种模型，分词模型1和分词及词性标注模型2都是由人民日报语料库训练得到，该语料库中约有一千二百万已被标注的字；更复杂、精确的分词及词性标注模型3由多语料联合训练得到，该语料包含约五千八百万已被标注的字[38]。THULAC的特点如下：

（1）分词能力强。利用规模庞大的语料库进行训练得到的模型具有很强大的分词标注能力。

（2）准确率高。THULAC工具包在标准数据集上分词的F1值高达97.3%。

（3）速度较快。

### THULAC使用

本文使用的是Java版，模型采用分词及词性标注模型2。将工具包和模型下载好之后，通过命令行调用运行，命令格式如下：

（1）java -jar THULAC\_lite\_java\_run.jar [-t2s] [-seg\_only] [-deli delimeter] [-user userword.txt] 从命令行输入输出

（2）java -jar THULAC\_lite\_java\_run.jar [-t2s] [-seg\_only] [-deli delimeter] [-user userword.txt] -input input\_file -output output\_file 从文本文件输入输出（注意均为UTF8文本）

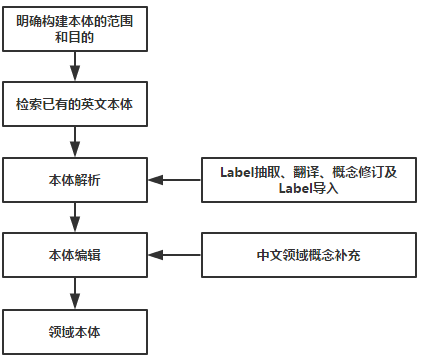
其中的参数含义如表2.2所示。

|  |  |
| --- | --- |
| 表2.2 调用THULAC命令行参数及含义 | |
| 参数 | 含义 |
| -t2s | 将句子从繁体转化为简体 |
| -seg\_only | 只进行分词，不进行词性标注 |
| -deli de | 将de作为词性与词的分隔符，默认为下划线\_ |
| -filter | 使用过滤器去除一些没有意义的词语，例如“可以”。[38] |
| -user userword.txt | 设置用户词典，用户词典中的词会被打上uw标签。词典中每一个词一行，UTF8编码(python版暂无)[38] |
| -model\_dir dir | 设置模型文件所在文件夹，默认为models/ |
| -input input\_file | 设置从文件读入，默认为命令行输入[38] |
| -output output\_file | 设置输出到文件中，默认为命令行输出[38] |

## 本体构建方法

常用的本体构建方法有骨架法、七步法等[47]，在这里不再赘述，因为医学指南中有很多的专业术语以及复杂的逻辑关系，需要很高的医学素养才能很好地构架出一个医学指南本体。考虑到自身实际情况、国外本体库的建设情况，结合本体的重用思想，即在构建本体时推荐在已有的类似本体上进行修改来创建新本体，本文决定采用翻译法：从国外丰富的本体库中下载英文的心血管疾病本体，进行翻译后得到我们需要的本体。用翻译法构建本体的流程如图2.2所示。

图2.2 翻译法构建本体流程图

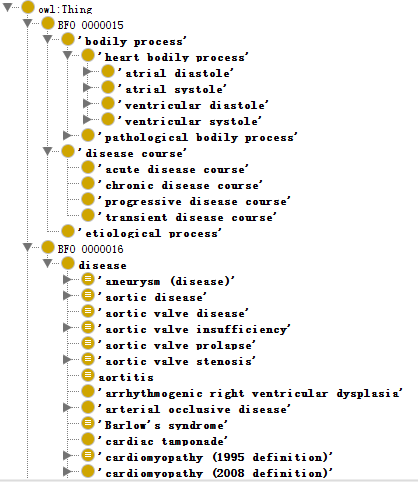


（1）明确构建本体的范围和目的

（2）在本体库中检索所需的英文本体

国外对本体的研究比国内起步早、水平高，而且已经建立了很多本体库，有大量公开的优质英文本体，比较有名的本体资源库有：Ontolingua ontology library、DAML ontology library、UNSPSC、DMOZ等。本文在本体库中成功地找到并下载了一个心血管疾病CVDO本体，其中包含了689个类，865条逻辑公理等。CVDO本体的部分类结构如图2.3所示。

图2.3 心血管疾病英文本体 部分类的结构



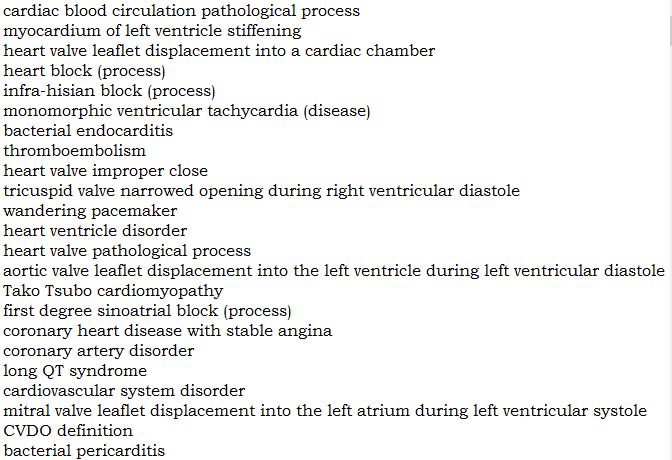
（3）本体解析

本体解析分为三个步骤：英文Label提取、Label翻译修订、中文Label导入。

①英文Label提取

使用Jena中提供的方法，首先用ModelFactory类创建一个本体模型，用以读取CVDO.OWL本体文件中的数据，然后以RDFS.label为参数使用listObjectOfProperty方法抽取出CVDO.OWL本体文件中所有类的名称标签label，通过BufferedWriter方法把提取出的label数据存入concept\_cvdo.txt文本文件中。提取出的部分信息如图2.4所示。

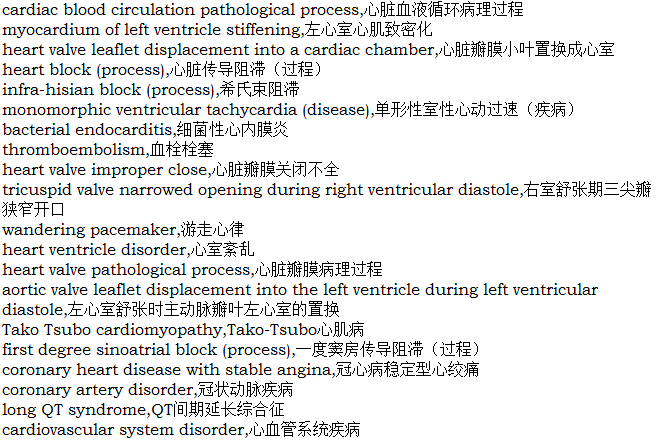
图2.4 部分英文label信息



②Label翻译修订

因为这些label信息大部分都是医学专业名词，所以直接用翻译软件进行翻译即可，对翻译后的结果人工矫正修订之后的结果如图2.5所示。

图2.5 部分中英文对照label信息

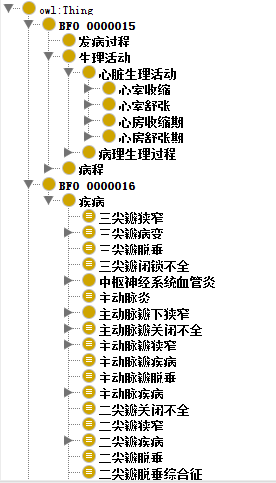


将英文label信息与中文label信息以逗号隔开，以csv格式进行存储。

③中文Label导入

与Label提取类似，首先用ModelFactory类创建一个本体模型，然后读取前面创建的中英文对照csv文件中的信息和CVDO.OWL本体文件中的英文label条目，分别从csv文件信息中为每一条英文label条目找到对应的中文翻译并进行替换，最后通过本体模型的write方法将改动后的label信息写入CVDO.OWL本体文件中。翻译后的本体文件部分类结构如图2.6所示。

图2.6 中文心血管疾病本体 部分类的结构



至此，一个中文心血管疾病本体已经大致建成了，接下来要完成心血管疾病诊疗指南知识图谱的构建还需要向其中添加很多内容，后续的工作是本文的重点，将在后续章节进行详细展开。

## 本章小结

本章介绍了知识图谱生成系统完成功能时要用到的Jena语义网框架和THULAC分词工具，以及知识图谱生成后用于查看、编辑本体的protégé本体构建工具。最后还详细介绍了本系统第一步用到的本体构建方法：翻译法。

# 系统需求分析

本文要实现的是面向医学指南的疾病知识图谱生成系统，重点在于从医学指南中提取出有效的三元组信息，也就是实体以及实体间的关系。开发软件首先要明晰软件需要具有哪些功能、需要完成哪些工作，对系统目的和功能进行分析，有利于后续软件的设计与功能的实现。

## 系统需求

前文已经提到，本文要设计的系统是一个半自动的知识图谱生成系统：从医学指南文本中提取信息以供专家确认，并把专家确认并修改后的信息填充到本体文件中。根据系统的定位可以得出以下结论：

（1）系统需要有较为可靠、高效的文本处理能力。系统的目的是简化专家构建本体的工作，如果经过文本处理得出的结果不能有足够的正确性，反而会适得其反，增加专家的工作量。

（2）系统需要有一个操作界面。因为是半自动的知识图谱生成系统，有人工参与所以要有一个操作界面。操作界面中需要有一个文本显示区域用来展示系统从医学指南中提取出的信息，还需要有一个文本输入区域以供用户输入想要填充入医学指南知识图谱中的内容。

（3）系统需要将用户输入内容填充入本体文件的能力。系统最终的目的是要帮助用户建立一个知识图谱也即本体，最后要把经过用户确认、修改、补充后的内容填充到指定本体文件中去。

## 系统功能模块划分

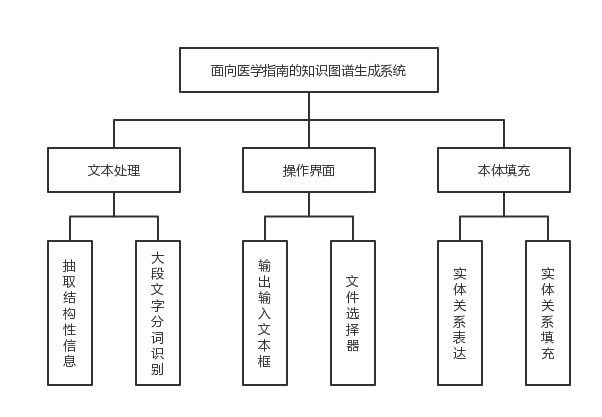
根据以上分析，可以将面向医学指南的疾病知识图谱生成系统划分为：文本处理、操作界面、本体填充三个功能模块。如图3.1所示。

### 文本处理模块

文本处理模块分为两个部分：抽取结构性信息模块和大段文字分词识别模块。

（1）抽取结构性信息模块：医学指南一般都有一定的行文规范，通过对其行文规则的研究可以找到一些规律，例如：章名和节名通常都是疾病的名称而且两者之间有包含关系。根据这些行文规律可以从医学指南中抽取出比较可靠的结构性信息。

图3.1 系统功能模块划分



（2）大段文字分词识别模块：在识别并抽取出了医学指南的结构性信息之后，还会剩下大段的没有处理的文字，而这些文字中可能还有很多有价值的实体及关系，所以需要对这些大段的文字进行处理。一般实体间比较强烈的关系都在一句话内进行体现，句间的逻辑关系由于过于复杂在这里不做更具体的研究。该模块在借助THULAC分词工具进行按词性分词后进行信息筛选。

### 操作界面模块

操作界面模块大致可分为输出输入文本框模块和文件选择器模块。

（1）输出输入文本框模块：输出文本框负责输出系统从医学指南中提取出的实体关系信息，输入文本框模块供用户将确认、修改、补充后的实体关系信息进行输入，然后交由实体关系填充模块输出到本体文件中。

（2）文件选择器模块：系统需要通过文件选择器定位THULAC分词工具（后期也可考虑打包后直接使用相对路径），另外要输入到系统中进行处理的医学指南文本文件和系统最终要输出到的本体文件路径都需要通过文件选择器模块进行确定。

### 本体填充模块

本体填充模块可分为两个部分：本体关系表达和本体关系填充，这两个部分其实是共同发挥作用的。

（1）实体关系表达模块：从医学指南文本中抽取出的实体关系需要以简单明了的方式呈现给用户，而且用户进行实体、关系输入时也需要让系统能够理解，因此就需要制定一种关系表达模式或者说规范，既清晰简洁又能够正确表达意思。

（2）实体关系填充模块：用户输入的实体关系信息经过实体关系表达部分“翻译”后需要通过实体关系填充模块添加到本体文件中。该模块需要运用Jena提供的方法将用户输入的信息与已知本体中的信息进行比对，把没有重复的部分填充到本体中。

## 本章小结

本章通过对系统要完成的目标进行分析，明确了系统应该有一个操作界面，后台要有文本处理的能力和将实体关系导入到本体文件中的能力。根据这些要求对系统的构成进行了模块划分，并且详细介绍了各个模块需要完成的功能。

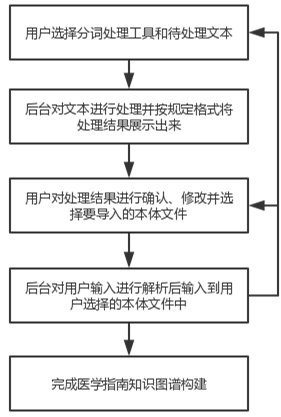
# 系统设计

经过上一章的需求分析，已经明确了系统需要有哪些模块、完成哪些功能，在此基础上要通过系统设计找到具体的方法实现需求的功能，这是整个系统开发的核心。本章将对系统总体架构设计、功能模块具体设计、实现展开详细描述。

## 系统架构设计

经过需求分析以及功能模块分析，面向医学指南的知识图谱生成系统大致可分为前台显示与后台处理两个方面，后台处理中又分为实体关系抽取和实体关系填充两部分。系统的使用、工作流程如图4.1所示。

图4.1 系统工作流程



为了完成医学指南知识图谱可能需要反复对本体进行填充，故而重复执行前四步操作。如果是要对多个文本进行处理则重复1到4步，如果是对单个文本的处理结果进行多次修改、补充则重复3到4步，直到最终实现医学指南知识图谱的构建。

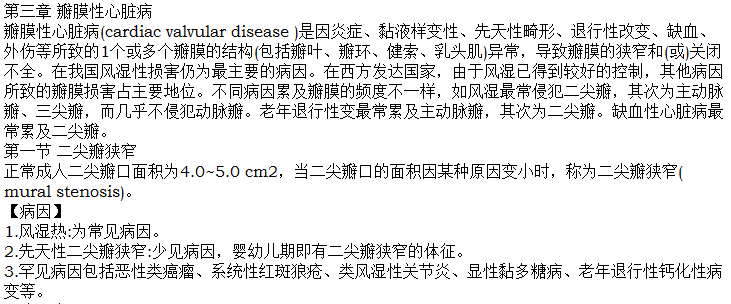
## 系统功能模块设计

上一章已经对系统功能模块的需求进行了介绍，本节将对各个功能模块的设计细节以及功能的实现方法进行具体的描述。

### 文本结构性信息抽取模块

要将文本中的结构性信息抽取出来，首先要对医学指南的行文规律进行研究，这里摘取文献[1]中第三章的部分内容如图4.2所示。

图4.2 心血管疾病诊疗指南第三章的部分内容



经过对文献[1]医学指南的结构进行分析，可以得出以下结论：

（1）章与节之间有包含关系，每一个小节描述的疾病都是该小节所在章节所描述的疾病的一个子类，如图4.2中所示的“二尖瓣狭窄”病就是“瓣膜性心脏病”中的一种，前者应是后者的子类。

（2）每一小节、每种疾病都从病因、症状、体征、特殊检查和治疗五个方面进行介绍，且病因和治疗由特殊符号“【】”进行强调，症状、体征和特殊检查则是形如“2.特殊检查”即“1位数字+1位符号+特定文字”。

（3）在上述五种小标题下通常都会有多个条目进行展开，其格式都是形如“1.风湿热：”即“1位数字+1位符号+若干文字+冒号”的组合。

通过第一条结论我们可以抽取出章名和节名作为疾病的名称，以及每一章描述的疾病和它下面各小节描述的疾病之间具有的包含关系。另外，通过对章名、节名的识别，将整篇医学指南按章节划分为了多个内容独立的段：每个段描述一种疾病，若干段组合成一个大段描述的是一类疾病，若干大段组合成整篇文章。

第二条结论中之所以不是单纯地识别“病因”、“症状”等文字而是要识别特定文字以及文字前后数字、符号的格式，是为了对这些小标题进行比较精准的定位，避免错误地定位到段落中的“病因”“症状”等文字上，造成错误的分段结果。在第一条结论的基础上，通过第二条结论又可以把按小节分的段再按病因、症状、体征等内容划分为五个独立区域，每个区域描述该疾病的一个方面。

第三条结论则是帮助识别疾病的某一方面的具体条目。

本文使用如表4.1所示的正则表达式对以上三条结论中提到的结构性信息进行表述。

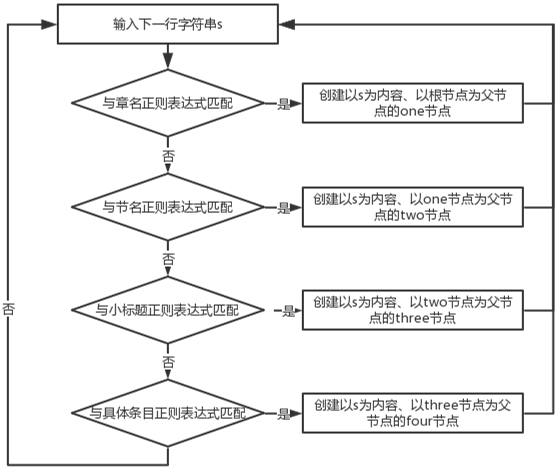
|  |  |
| --- | --- |
| 表4.1 结构性信息及对应正则表达式 | |
| 结构性信息 | 正则表达式 |
| 章名 | (\\s\*第)(.{1,9})[章](\\s\*)(.\*) |
| 节名 | (\\s\*第)(.{1,9})[节](\\s\*)(.\*) |
| 病因、治疗小标题 | 【(病因|治疗)】 |
| 症状、体征、特殊检查小标题 | [1-9]{1}.{1}(症状|体征|特殊检查) |
| 具体条目 | [0-9]{1}.{1}(.{0,10}): |

利用以上正则表达式即可正确匹配到医学指南中的章节名、小标题、小标题下的主要条目这些结构性的信息，然而只是单纯地提取出这些信息对应的实体意义并不大，还需要提取出他们之间的关系，本文采用树形数据结构对上述数据进行存储以保留它们之间的关系。树的数组结构示例如表4.2所示，根节点无父节点，用“-1”表示。

|  |  |  |
| --- | --- | --- |
| 表4.2 树的数组结构 | | |
| Index | Data | Parent |
| 0 | Root | -1 |
| 1 | B | 0 |
| 2 | C | 0 |
| 3 | D | 1 |
| 4 | E | 3 |

把提取到的信息存储到树中的流程如图4.3所示。

图4.3 将信息存储入树的流程



根据图4.3所示流程存储得到的树结构大致如图4.4所示。

图4.4 树的存储结构

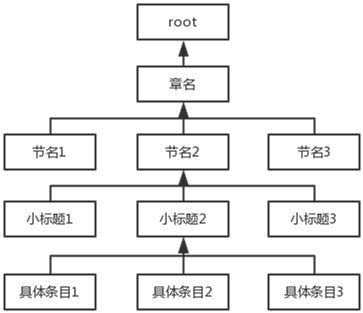


图4.4中以节名2、小标题2为例进行展示，箭头指向的是当前节点的父节点，树中隐含了以下两种关系：

（1）章名与节名之间的关系：节名1、2、3对应的疾病分别是章名对应疾病的一个子类。

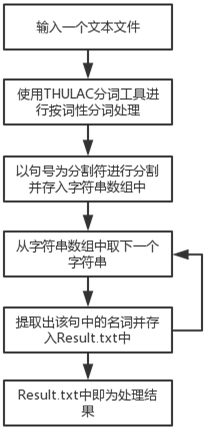
（2）节名、小标题和具体条目的关系：假设小标题2为“病因”则具体条目1、2、3分别为节名2对应疾病的一个病因，即节名2对应的疾病具有以小标题2为名的属性（病因），该属性含有具体条目1、具体条目2、具体条目3，这三个值。

### 分词处理模块

上一小节介绍的模块将医学指南中的诸如章节名、小标题等结构性信息抽取了出来，还有大段的正文没有进行处理，在正文中当然也隐含有很多有价值的信息，本模块就是要将这些信息尽量提取出来。

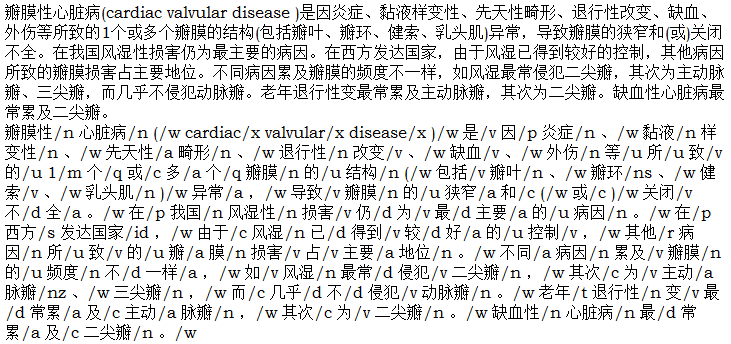
一般实体间比较强烈的关系都在一句话内进行体现，句间也可能会存在关系，但是这种关系一来比较隐晦，不容易发现，二来一旦识别出一个错误的结果会很有误导性，所以本模块在这里只提取句内的实体及关系。具体思路如图4.5所示。

图4.5 正文中实体抽取思路



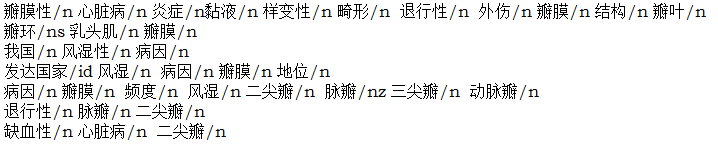
使用THULAC分词工具对文本进行按词性分词，该工具会进行分词后，在每一个词后进行词性标注，如：名词/n，动词/v，标点/w等，效果如图4.6所示。

图4.6 按词性分词示例



可以看出从分词后的文本中能够很容易地提取出名词来，对图4.6中的文本进行分句名词提取后的结果如图4.7所示。

图4.7 从分词结果中提取出名词



这里之所以只提取名词是因为句中的动词位置不固定而且常有副词、形容词等进行修饰，句内的逻辑关系通常比较复杂，限于自然语言处理能力的有限无法正确地识别句内复杂的逻辑关系，若是全部提取可能会生成一些南辕北辙的具有误导性的信息，这样反而增加了工作量。因此只对名词进行提取，将提取出的名词展示给用户以供用户筛选补充。

### 本体填充模块

本文暂时只涉及了两种实体-关系表达格式：

（1）实体A\实体B：表示实体B是实体A的子类，

（2）实体A\实体C\实体B：表示实体A具有以实体B为值的属性C且实体B为实体C的子类。

相比起来第一种格式的填充方法是对第二种格式的填充方法的简化，因此本文以第二种格式的填充方法为例进行介绍。

首先对本体进行填充要避免重复冗余信息的填充，在用ModelFactory的方法建立本体模型并把指定本体文件进行导入之后，使用以RDFS.Label为参数的listObjectOfProperty()方法得到所有类、属性的名称并与实体A、B、C的名称分别进行比较，如果本体中已经有对应的资源存在了就用getOntClass()方法返回该资源，否则用createClass()方法创建新的资源。

然后是对关系的填充，若实体A为疾病名称而且本体中还没有类A存在，则将类A添加为疾病类的子类，方法是addSubClass()；若本体中没有实体C对应的类C或属性C存在，则在新建类C的同时新建属性C，方法是createOntProperty()；若本体中没有实体B对应的类B存在，则将类B添加为类C的子类并且为类A添加以类B为值的属性C。

最后将以上改动输出写入到所选本体文件中。

本模块所使用的Jena的方法及其功能如表所示。

|  |  |
| --- | --- |
| 表4.3 结构性信息及对应正则表达式 | |
| 方法名 | 功能 |
| listObjectOfProperty() | 返回本体中所有指定属性的值 |
| getOntClass() | 返回本体中指定的类 |
| createClass() | 创建一个新类 |
| addSubClass() | 为当前类添加指定子类 |
| createOntProperty() | 为当前类添加指定属性 |

## 本章小结

本章介绍了系统中的三个核心功能模块：文本结构性信息抽取模块、分词处理模块和本体填充模块具体实现的方法，工作的流程。

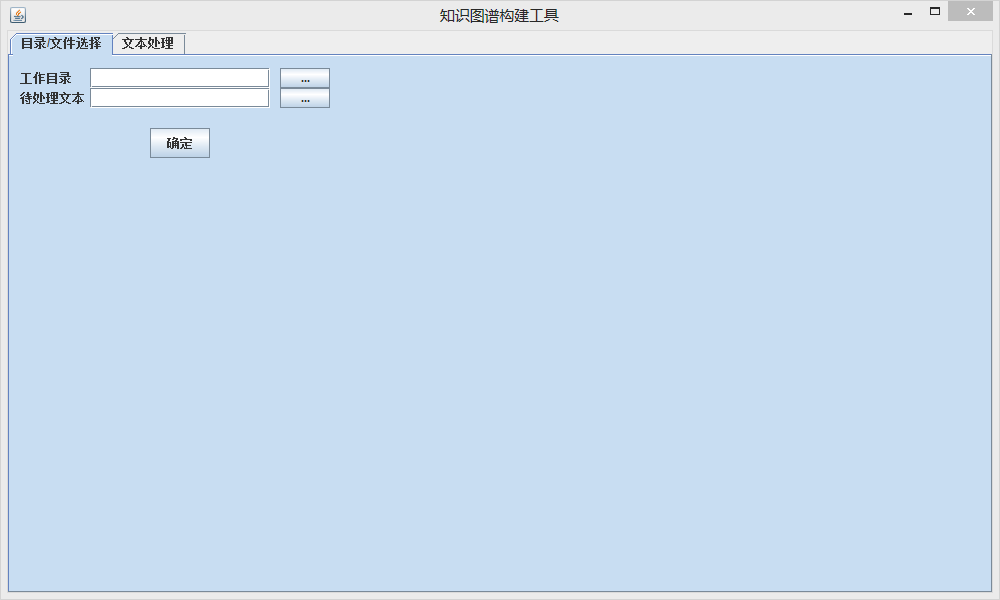
# 系统的部分实现

经过了系统需求分析和功能模块设计，并使用Eclipse进行编程开发之后，本文主要实现了结构性信息提取方面的功能。本系统采用Java语言开发，有后台的处理功能，也有前台的用户操作界面。本章将以对文献[1]的第三章第一节进行处理为例对本系统的操作界面和功能实现进行详细介绍。

## 目录/文件选择页面

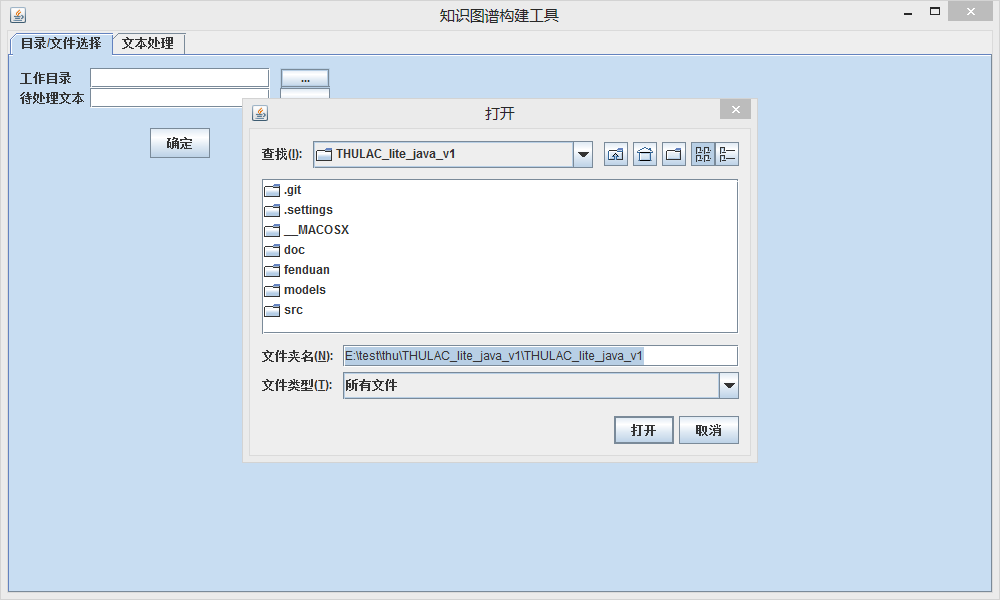
知识图谱构建工具有两个工作页面，首页如图5.1所示。

图5.1 知识图谱构建工具首页



首页中有一个工作目录选择器，通过点击“…”按钮打开文件夹选择器，如图5.2所示。这里工作目录的含义是分词工具可执行jar包所在的文件夹路径，文件夹选择器返回的路径会在“工作目录”标签后的文本域中显示，通过这个路径程序可以对分词工具进行调用，并且在该路径下新建一些用于存储中间结果的文件。工作目录选择器下方还有一个待处理文本选择器，该选择器与工作目录选择器的不同在于，前者只能选择到文件而后者只能选择到文件夹。待处理文本选择器的功能是选择需要处理的医学指南文本文件并返回其路径。

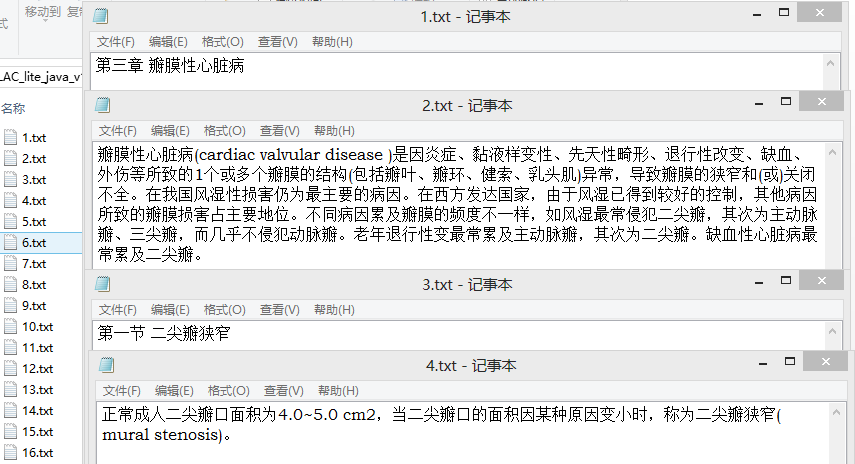
图5.2 工作目录文件夹选择器



选择好工作目录和待处理文本之后点击“确定”按钮则开始对选定的文本进行处理。处理分以下几个步骤。

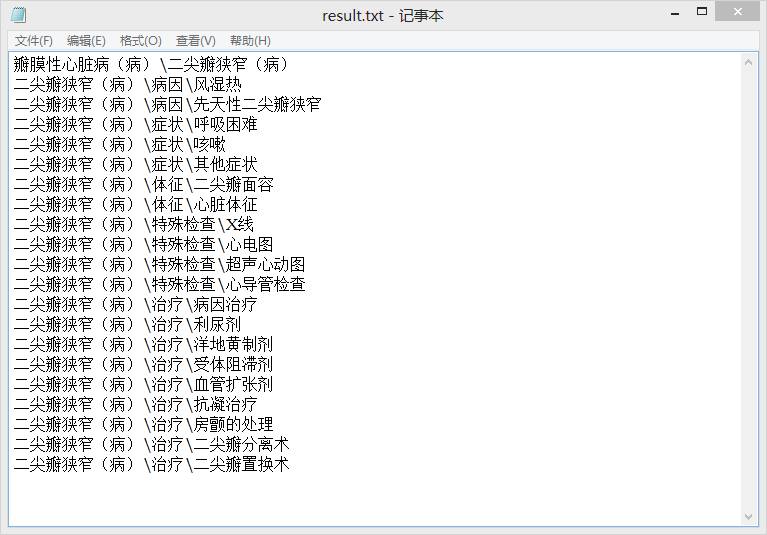
（1）分段：把医学指南文本通过输入流进行输入然后在工作目录下创建名为“textsplit”的文件夹，逐行（段）读取输入流中的内容并且在textsplit文件夹下新建以当前行数为名的TXT文本文件，最后把每一行字符串分别存到对应的TXT文本文件中。分段结果如图5.3所示。

图5.3 分段结果



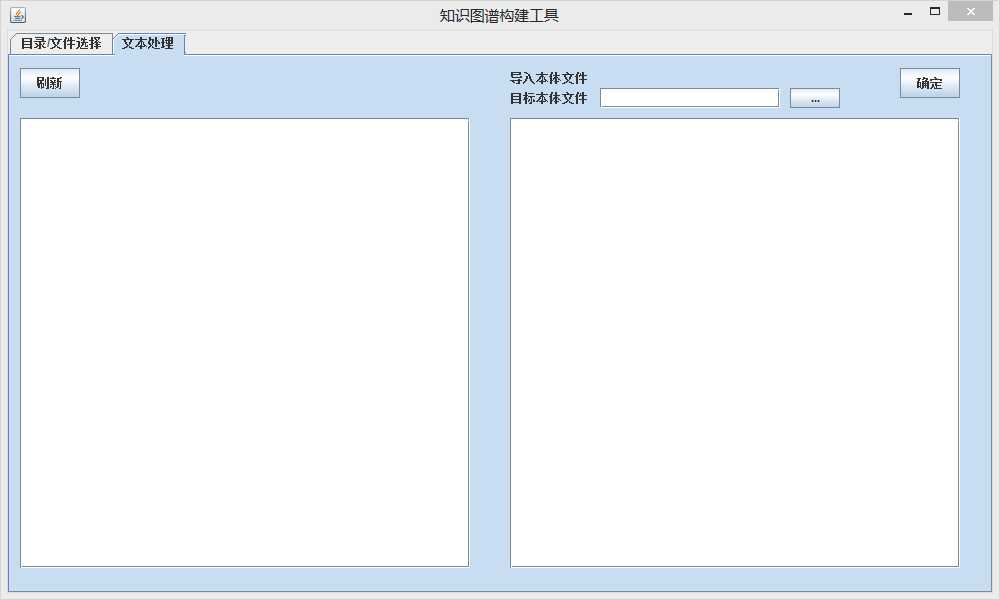
（2）信息提取：分别输入每一段的内容与第四章中设计好的正则表达式进行匹配，若匹配成功则将有效信息按设计好的规则存入树中，若匹配失败则记录其段落号留待分词处理模块处理。在所有段落都被处理过之后，把树中的信息提取出来并存入新建的“result.txt”文件中。提取结果如图5.4所示。

图5.4 信息提取结果



## 文本处理页面

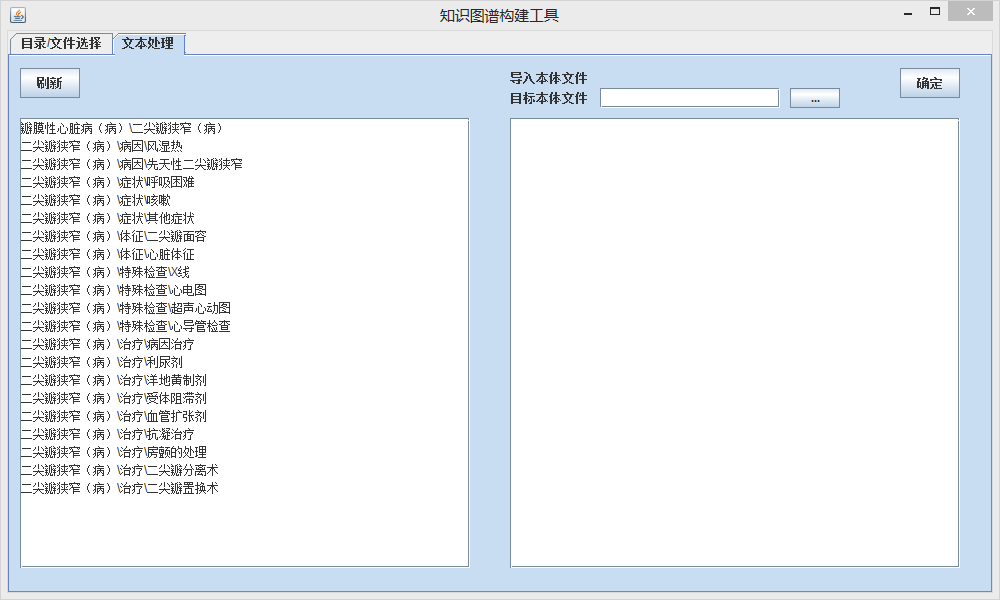
图5.5 文本处理界面



文本处理页面如图5.5所示。文本处理页面分左右两个部分，左边是处理结果显示区，右边是用户编辑输入区。在文本处理页面，用户的操作分为以下步骤：

（1）点击“刷新”按钮，程序就会把result.txt文本文件中的内容显示在下方的文本域中，效果如图5.6所示。

图5.6 刷新显示结果



（2）用户可以对左边文本域中显示的信息进行编辑修改、补充，把修改后经确认无误的内容直接粘贴到右边的文本域中。

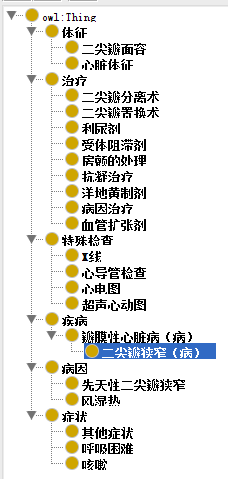
（3）通过目标本体文件选择器选择将要导入的本体文件例如heart disease.owl，点击“确定”按钮即可把确认后的内容填充到选定的本体文件中。

（4）使用protégé本体编辑器打开heart disease.owl本体文件查看构建结果。

## 本体构建结果展示

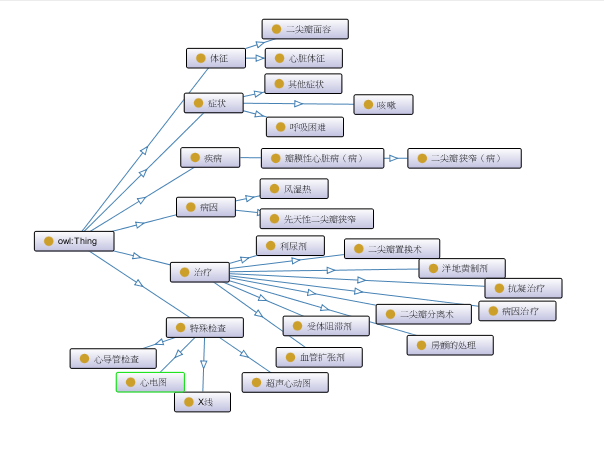
本体中所有的类如图5.7所示。

图5.7 本体中的类



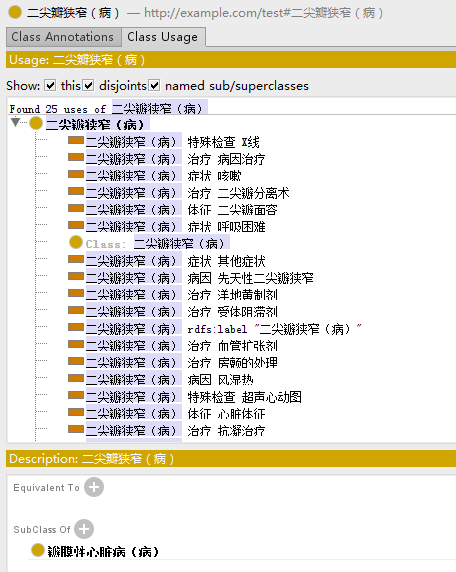
使用可视化工具Ontograf显示效果如图5.8所示。

图5.8 Ontograf显示



还可以直接查看某个类的所有属性，例如“二尖瓣狭窄（病）”的属性显示如图5.9所示。

图5.9 二尖瓣狭窄（病）类的属性



## 本章小结

本章从用户使用的角度，根据使用知识图谱构建工具的操作流程，对本系统的工作实现进行了展示，本系统能够有效地将医学指南中的结构性信息提取出来，并且可以帮助用户进行本体的构建，省去了用户手动添加本体中的类、属性的麻烦。

# 总结与展望

## 工作总结

医学指南作为权威、详实的指导性医学文献，本可以在临床诊疗实践中发挥很大的作用，然而受制于其纸质文本的存在形式，在实际生活中却鲜有医生会使用。随着知识图谱相关概念与技术的发展、成熟，根据知识图谱的思想，将非结构化的医学指南文本转换为结构化的本体可以为医生使用医学指南提供很大的便利。本文设计并实现了一个面向医学指南的知识图谱生成系统，用于辅助医学专家构建医学指南的本体，具有结构性信息抽取、快速填充本体两大核心功能，可以在专家构建本体的过程中简化很多工作。

本课题的主要研究工作有：

（1）分析总结了当前医学指南发展遇到的困境以及知识图谱相关概念及技术的发展现状，指出了本课题的研究背景和意义。

（2）简要介绍了本课题开发用到的工具：protégé本体构建工具、Jena语义网框架和THULAC分词工具，并且用翻译法构建了一个心血管疾病本体，作为心血管疾病知识图谱构建的基础本体。

（3）研究了如何有效地将医学指南中的结构性信息抽取出来、如何使用Jena提供的方法对本体进行编辑、填充，提供了使用THULAC分词工具对医学指南文本内容进行实体抽取的思路。

（4）通过编程对结构性信息提取、本体填充两大核心功能模块进行实现，最后整合出一个知识图谱构建工具，并以《心血管疾病诊疗指南》第三章为例使用开发出的构建工具构建了一个医学知识图谱。

## 研究展望

### 分词处理

本文在第四章的系统设计部分提供了使用THULAC分词工具对大段文字进行实体抽取的思路，然而受限于自然语言处理能力的不足，无法很好地将句中复杂的关系正确地提取出来，最终只能实现对名词的提取，实用价值不高，最终没有在本体构建工具中实现。对于这个思路的实现主要有以下难点：

（1）THULAC分词工具使用的分词模型是通过对人民日报语料库进行训练得到的，不能很好地对医学领域的专有名词进行分词识别，对医学指南进行按词性分词得到的结果正确率不高。若要提高正确率需要自己准备医学领域的语料进行训练得到专用分词模型。

（2）要有效识别医学术语可能需要有医学词库的辅助，具体如何实现还有待研究。

（3）中文博大精深，要正确识别中文语句中的逻辑关系是一门大学问，暂时还没能找到现成、有效的中文语义识别工具包。

### 结构性信息抽取

在结构性信息抽取模块中我们使用了正则表达式对结构性信息进行抽取，但是这种抽取在对其他可能格式不同的医学指南文本进行处理时无法正确地工作，下一阶段的研究可以有两个方向：

（1）设计更为通用、智能的识别方法。

（2）让用户人工地对行文格式进行设定。

参考文献

1. 汪勇, 刘海华, 李玲. 心血管疾病诊疗指南[J]. 2010.
2. 吴彬飞. 临床指南知识表达和应用方法研究 [D][D]. 浙江大学, 2010.
3. 徐增林, 盛泳潘, 贺丽荣, 等. 知识图谱技术综述[J]. 2016.
4. 陆玉婷. 基于句法分析的医学指南事件及事件关系提取[D]. 武汉科技大学, 2014.
5. 范玉玲, 顾进广, 黄智生. 中文医学指南的事件处理及其语义数据自动生成[J]. 中国数字医学, 2015, 10(9): 76-78.
6. 尹梓名, 吕旭东, 段会龙. 基于临床指南的决策支持系统[J]. 中国医疗器械信息, 2015, 21(3): 1-5.
7. 周群一,郭玮,吕旭东,段会龙. 数字化临床指南的可视化表达[J]. 生物医学工程学杂志,2009,(02):239-243.
8. 聂蕾. 高血压疾病智能分层系统开发与实现[D]. , 2015.
9. 王昊奋. 大规模知识图谱技术[J]. 中国计算机学会通讯, 2014, 10(3): 64-68.
10. 张晓孪,张蕾,王西锋. 基于知识图的汉语词语间语义相似度计算[J]. 计算机工程与应用,2007,(08):160-163.
11. 胡芳槐. 基于多种数据源的中文知识图谱构建方法研究[D]. 华东理工大学, 2015.
12. 张晓孪,张蕾,王西锋. 基于知识图的汉语词语间语义相似度计算[J]. 计算机工程与应用,2007,(08):160-163.
13. 张少敏. 基于FCA与Jena的领域本体半自动构建方法研究[D].昆明理工大学,2011.
14. Chowdhury G G. Natural language processing[J]. Annual review of information science and technology, 2003, 37(1): 51-89.
15. Dong X, Gabrilovich E, Heitz G, et al. Knowledge vault: A web-scale approach to probabilistic knowledge fusion[C]//Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2014: 601-610.
16. 宋绍成, 毕强, 杨达. 本体技术在学术研究领域中的应用[J]. 东北师大学报: 自然科学版, 2005, 37(1): 41-45.
17. 陈悦, 刘则渊. 悄然兴起的科学知识图谱[J]. 科学学研究, 2005, 23(2): 149-154.
18. 曾庆文. 基于语义的医疗 Web 服务发现研究[D]. 电子科技大学, 2013.
19. 姚天顺,张俐,高竹. WordNet综述[J]. 语言文字应用,2001,(01):27-32.
20. 孙文玲,张晓如. Freebase研究及应用[J]. 信息技术,2014,(09):201-204.
21. Lehmann J, Isele R, Jakob M, et al. DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia[J]. Semantic Web, 2015, 6(2): 167-195.
22. BaiduBake.IMDB[EB/OL].[2017-05-02].http://baike.baidu.com/link?url=SnhrJc0S-oEjRKLzfbUqM0q8cOGfLd\_j03--lUnLYTLnizL6WBoTkGW0TrTshC26xnSzCsNaYBlzSfpRmWcmlgq2g3qR-YvQd7X1ZHHIH3SWrhnX1AxBT5zYqFjK\_xOg0qgl8cjERNxmrgDOHsPQzK
23. Kawamura T, De Blasio J A, Hasegawa T, et al. Public deployment of semantic service matchmaker with UDDI business registry[C] //International Semantic Web Conference. Springer Berlin Heidelberg, 2004: 752-766.
24. Singhal A. Introducing the knowledge graph: things, not strings[J]. Official google blog, 2012.
25. Lin Y, Liu Z, Sun M, et al. Learning Entity and Relation Embeddings for Knowledge Graph Completion[C]//AAAI. 2015: 2181-2187.
26. Wang Z, Zhang J, Feng J, et al. Knowledge Graph Embedding by Translating on Hyperplanes[C]//AAAI. 2014: 1112-1119.
27. Wolf M S, Davis T C, Shrank W H, et al. A critical review of FDA-approved medication guides[J]. Patient education and counseling, 2006, 62(3): 316-322.
28. 杜捷夫, 李银平. 基于循证医学证据的《 急诊医学指南》[J]. 中国危重病急救医学, 2006, 18(1): 1-4.
29. 贾晓巍, 蒋朱明. 中国临床医学指南现状[J]. 中华临床营养杂志, 2011, 18(6): 327-329.
30. Wolf M S, King J, Wilson E A H, et al. Usability of FDA-approved medication guides[J]. Journal of general internal medicine, 2012, 27(12): 1714-1720.
31. 何雨生, 孙宏宇. 计算机化临床实践指南研究进展[J]. 中国数字医学, 2007 (1): 10-15.
32. 贾继东. 以循证医学的观点正确认识和对待临床诊疗指南[J]. 中华内科杂志, 2005, 44(9): 641-642.
33. 郑万松, 袁军, 黄志中, 等. 以病人为中心的自助医疗系统研究与应用[J]. 中国数字医学, 2013 (11): 40-42.
34. 刘光熠, 赵迎宾. 一种灵活的基于临床指南的临床决策支持系统[J]. 计算机应用与软件, 2011, 28(6): 189-191.
35. 叶青, 刘丹红, 杨喆, 等. 临床指南模型构建中医学知识的规范化表达方法[J]. 中国数字医学, 2011, 6(8): 14-17.
36. Liddy E D. Natural language processing[J]. 2001.
37. Zhongguo Li, Maosong Sun. Punctuation as Implicit Annotations for Chinese Word Segmentation. Computational Linguistics, vol. 35, no. 4, pp. 505-512, 2009.
38. 孙茂松, 陈新雄, 张开旭, 郭志芃, 刘知远. THULAC：一个高效的中文词法分析工具包. 2016.
39. 王磊, 周宽久, 仇鹏. 领域本体自动构建研究[J]. 情报学报, 2010 (1): 45-52.
40. 袁燕. 大肠癌病理本体的构建研究[D].浙江大学,2016.
41. Musen, M.A. The Protégé project: A look back and a look forward. AI Matters. Association of Computing Machinery Specific Interest Group in Artificial Intelligence, 1(4), June 2015. DOI: 10.1145/2557001.25757003.
42. 何平. 基于可拓模型的本体进化研究[D].广东工业大学,2011.
43. 李秀华. 本体构建系统的关键技术研究与实现[D].南京航空航天大学,2011.
44. 蒋祥杰. 基于用户情境本体的个性化知识服务研究[D].武汉理工大学,2010.
45. BaiduBake.IMDB[EB/OL].[2017-05-02].http://baike.baidu.com/link?url=eUNfAu1KZkjY5nQEt5C8B1mTq4Ca7Zkldw3EbEldmLp20fosezWJMaj2omawCQt4j1crOIaSTFztorM-hfuwkhzpkc1iQoacoewcMovhc-kPMjyl\_HLb6GTON09-NJ1\_M-5nEYIBRoMySxb4iFNrQa
46. Csdn.IMDB[EB/OL].http://blog.csdn.net/skiffloveblue/article/details/9355925
47. 王庆连. 基于本体的交通导航数据语义查询研究[D].大连海事大学,2009.
48. 李新霞. 基于本体的中医学脾胃病知识库的构建[D].南京理工大学,2008.
49. 刘雁昆. 基于本体的软件工程领域知识库构建方法研究[D].北方工业大学,2015.
50. 卢承山. 基于领域的主题信息采集技术研究[D].武汉理工大学,2011.

致谢

时光飞逝犹如白驹过隙，随着论文的完成，四年的本科生生涯也已经进入了尾声。回想这四年，从春风得意到迷茫沉沦再到重新确立目标，这一路上经历过很多挫折磨难，也收获过很多成功喜悦，不可谓不充实。在此向关心和帮助过我的老师、同学、家人以及朋友表示由衷的感谢。

首先，衷心感谢何扬帆老师，在刚开始选题的时候我对自己很没有自信，是何老师给了我很多鼓励让我有信心把这个课题做好。在课题开始之初老师帮我搜集了很多相关的论文、资料，我在学习过程中有疑问老师也会耐心地解答，每当研究陷入困境、死胡同老师总能另辟蹊径找到解决的办法，在整个毕业设计完成的过程中何老师给了我莫大的帮助。

感谢我的同学和室友们，他们是江凌峰、程环宇、孙柯、汪嘉祺。有了他们的陪伴，这大学四年才不孤单。也感谢他们在学习和生活中给我的很多帮助。

感谢我的父母，在我这一路成长的过程中，为我无私付出了很多，一直在鼓励我、关心我，是我心中永远的支柱和避风港。

向我人生路途中遇到的所有曾关心、爱护和帮助过我的人致以最衷心的感谢。

感谢参加论文评审和答辩的各位老师对本文的批评指正！