**北京航空航天大学软件学院**

工程实践中期报告

**课题名称：** **基于医药知识图谱的智能问答系统设计与实现**

**成员： 刘亚楠 ZF1921220**

**陈一**

**秦景建 ZF1921249**

**南晴 ZF1921248**

**赵伟超 ZF1921256**

**专业方向：人工智能**

**所属院系：北京航空航天大学软件学院**

**指导教师：康一梅**

**实践起止时间：2020年2月至2020年6月**

**目录**

[1. 需求分析 2](#_Toc41733994)

[1.1课题背景 2](#_Toc41733995)

[1.2选题意义 3](#_Toc41733996)

[2. 方案设计 4](#_Toc41733997)

[2.1. 系统架构设计 4](#_Toc41733998)

[2.2. 模块详细设计 4](#_Toc41733999)

[4.2.1用户UI页面模块功能设计 5](#_Toc41734000)

[4.2.2 WEB服务器模块功能设计 5](#_Toc41734001)

[4.2.3 自然语言处理模块功能设计 5](#_Toc41734002)

[4.2.4 知识图谱模块功能设计 5](#_Toc41734003)

[3. 中期进展情况 7](#_Toc41734004)

[3.1. 完成情况说明 7](#_Toc41734005)

[3.2. 取得的成绩 7](#_Toc41734006)

[3.3. 遇到的问题 7](#_Toc41734007)

[3.4. 后续实施计划及风险 8](#_Toc41734008)

# 需求分析

## 1.1课题背景

随着我国社会经济的飞速发展，人们的生活水平质量得到了很大提升，人们的健康意识也不断地提高，医疗与医药服务质量亟待提升。另一方面，随着互联网技术的发展，使医疗服务的信息化、智能化程序不断提高。现阶段已经有很多基于互联网提供的在线医疗服务产品。例如药房网能够提供根据医药信息搜索药品的服务和购药服务、好医生在线能够提供在线医生咨询服务等。但这些产品很多时候不能够满足人们对医疗服务的需求。大部分人对于医药信息知之甚少，那么就无法根据医药信息进行搜索和购药，而在线医生咨询服务的等待周期又无法确定，因此基于互联网的医疗与医药服务质量仍有很大的提升空间。医疗服务的交互性和时效性比较强，当用户需要医疗服务时对于答案的寻找不仅要准确，还要快。而现有的医疗领域的互联网产品不能够很好的满足用户需求，因此用户经常会选择使用搜索引擎搜索想要知道的内容。通过搜索引擎在一定程序上的确能够帮助人们快速获取信息，但也有一定的局限性[1],主要表现在：搜索引擎搜索关键词只停留在语言表层没有涉及语义；搜索引擎只能搜索与关键词相近的网页，而不能够根据用户搜索内容进行回答；搜索引擎搜索结果数据量较大，对于用户较难筛选有用信息。

在这种情况下，人们需要一款科技产品来满足人们对于高质量医疗服务的诉求。而面向医药领域的智能问答系统能够很好的解决这些问题。智能问答系统通过对用户的提问内容进行语义分析，并按照一定的规则和算法生成答案，最后将答案返回给用户。面向医药领域的智能问答系统不仅能够给用户提供准确的医疗服务，而且能够及时与用户交互，很大程序上满足了用户对医疗信息准确快速获取的需求。

智能问答系统区别于简单的信息检索系统，它允许用户使用自然语言进行搜索或提问，并能够用准确的、简洁的自然语言对用户的提问进行回答[2]。问答系统有多种分类方法，根据其问答问题的范畴可分为：面向开放领域[3]和面向特定领域[4]。面向开放领域的智能问答系统可回答的问题不限定于某个特定领域，回答问题的方向较多，在不同的语境下，同一单词可能在说不同的事物。而面向特定领域的问答系统所处理的问题只限定于某个领域或者某个内容范围，比如只限定医学、化学、生物等。

相比于传统的信息检索信息和开放领域的问答系统，本文设计的面向医药领域的智能问答系统有以下优点：

（1）能够针对用户的输入进行回答，而不是搜索与用户输入近似内容的列表；

（2）面向医药领域，对于用户的提问内容分析限定了具体的范围，因此更容易分析用户的提问需求，从而用户能够得到更想要的答案；

（3）使用NLP（Natural Language Processing, 自然语言处理）技术正确识别医药领域词汇，并对用户输入进行语言分析、使用多种算法生成答案。

## 1.2选题意义

人们的健康意识随着社会的发展也不断提高。医疗服务又是我们生活中必不可缺的一部分，但并不是每个人都能够诊断疾病，也不是每个人对医药信息都能够了如指掌，而且现有的产品不能够准确、及时地为我们提供有效的医疗服务。基于这种现实情况，为了帮助人们能够准确的获取医药信息，自我诊断疾病和指导用药，设计并实现一个面向医药领域的智能问答系统势在必行。

智能问答系统作为人工智能最具代表性的应用之一，得益于其智能化、人性化地为人们提供最便捷有效的知识服务，一直备受产业界关注。有很多公司已经开发面向开放领域的智能问答系统，而垂直领域的智能问答系统则较为罕见，尤其是面向医药类领域。面向医药领域的智能问答系统，为患者提供疾病答疑解惑、疾病自诊、用药推荐的内容与知识等智能医药服务，成为下一代人工智能在医药领域中最有可能的井喷口之一。另外，由于面向特定领域的智能问答系统统易迁移的特点，使得这项技术同时可以应用到其他的专业领域，具有非常广阔的应用前景。但是涉及专业特定领域的智能问答系统鲜有实际实用成功的案例，主要由于其专业知识要求较高、技术门槛高、研发成本高，极大制约着医药智能问答系统产业化的发展。现有的大部分应用智能化、精准度不够高，用户在使用过程中很难体会到智能与便捷。

本文旨在设计并研发一款智能问答系统，该系统能够识别人类自然语言，并且提供基于医药知识的咨询服务。该系统对于提升用户用药体验和帮助用户解决医药问题具有重要作用。

# 相关研究分析

## 2.1实体识别与实体关系抽取任务

首先，命名实体识别是问句语义解析任务中的重要技术手段，是本文知识图谱问答系统设计与实现中采用的关键技术。其次，实体识别与关系抽取是以自动抽取方式实现知识抽取任务的核心技术方法，是本文医疗知识图谱构建流程中知识抽取模块所采用的关键技术。因此，本节将分别对命名实体识别和实体关系抽取进行概述。

２．１．１命名实体识别

命名实体识别（ＮａｍｅｄEｎtｉｔｙＲｅｃｏｇｎｉｔｉｏｎ，ＮＥＲ），又叫做“专名识别”，指的是从非结构化的输入文本中识别出具有特定类型或者指代性强的实体，是ＮＬＰ领域中的一项基础任务，同时该任务也可以用来为关系抽取等任务做铺垫。在特定领域中，会相应地定义领域内的各种实体类型，例如临床医疗领域内的疾病、症状、药物、检查、手术、科室等。当然，同样可以按照业务需求定义并识别出更多类型的实体，比如药物的剂型、药物的价格等，只要是业务需要的具有某种特定特征的文本片段都可以称为实体。

实体识别任务通常会采用精确匹配的原则去评判一个命名实体是否被正确的识别，其中包括实体边界识别和实体类型确定两部分，即实体的类型是否标记正确，实体的边界是否被正确识别。ＮEＲ的错误类型主要有三种，一种是文本的边界预测错误但其包含的主要实体词和实体类型标记是正确的，一种是文本边界预测正确但实体类型标记错误，还有一种是实体的边界和实体的类型全部标记错误。

英文命名实体识别因为存在空格分隔符或首字母大写等使得实体边界识别任务变得相对容易。但是，因为中文语言的复杂性，导致中文命名实体识别任务变得复杂和困难，尤其是实体边界识别任务，具体地如下：

（１）由于中文语句没有明显的词语边界分隔符，所以基于分词的中文实体识别方法可能会很大程度上受到中文分词、浅层语法等因素的影响，一旦分词错误就可能会导致实体边界的识别错误。

（２）某些类型的实体词没有统一的构词规范，包括长度上没有限制，中英文交替使用、嵌套、别名、缩略词、数字／汉字数值不同表述形态等情况，由于没有严格统一的规律可以遵循，使得命名实体识别任务的召回率相对偏低。例如，在医疗领域中经常会出现中英文交替使用或者使用各种医学术语缩略词的情况，这增加了实体识别任务的难度。

（３）相同的命名实体可能具有不同的特征（如词性、语义特征等），不同的命名实体也可能具有不同的特征，想要正确标记这些命名实体的类型，往往需要涉及到上下文语义层面的分析，不能用统一的规则来刻画所有的实体特征，这都给命名实体的识别任务带来困难。

命名实体识别任务通常有以下几种方法：

（１）基于规则的方法：利用人工编写规则，通过文本与规则的匹配来识别出实体词。但是，这种手工构建规则的方法通常需要大量的中文语言学知识，由于可能出现中英文交替使用的情况导致无法构建统一不同语言的识别规则，而且规则之间的冲突问题同样需要谨慎地处理；此外，构建规则的过程花费大量的人力、而且不具备良好的移植性。

（２）基于机器学习结合特征模板的方法：ＮＥＲ任务可以通过机器学习的方法转化成序列标注问题。常用的模型包括隐马尔科夫模型（ＨＭＭ）、条件随机场（ＣＲＦ）等，通过定义特征模板结合ＣＲＦ模型进行序列标注的训练，通过维特比解码来得到最优的标记序列。

（３）基于神经网络的方法：同样将ＮＬＰ作为一个序列标注任务处理，通过分布式表示，将出入映射到低维稠密的向量，输入到某个神经网络中（如ＬＳＴＭ、ＲＮＮ等），用神经网络学习自动提取特征，最后通过接入ｓｏｆｔｍａｘ分类层或者ＣＲＦ模型层来完成序列标注。

对于ＮＥＲ任务来说，通常是采用Ｐｒｅｃｉｓｉｏｎ（精确率）、Ｒｅｃａｌｌ（召回率）和ＦＩＭｅａｓｕｒｅ三种评价指标衡量信息抽取系统的性能。Ｐｒｅｃｉｓｉｏｎ代表的是查准率，等于系统正确识别的结果占所有识别结果的比例；Ｒｅｃａｌｌ代表的是查全率，等于

系统正确识别的结果占所有可能正确结果的比例；Ｆ１值是精确率和召回率的加权调和平均值，ＦＩＭｅａｓｕｒｅ能够更加综合地评价系统的性能。其中，单条记录标签的预测结果使用精确匹配方式进行衡量，即当实体的边界和实体与关系的标签全部正确时才认为预测成功。因此，命名实体识别有三种结果：正确（ｃｏｒｒｅｃｔ）：系统识别结果和标准结果相同；虚假（ｓｐｕｒｉｏｕｓ）：系统识别但标准结果中没有；丢失（ｍｉｓｓｉｎｇ）：系统未识别而标准结果中有。

精确率 ＝ 

召回率＝ 

F1值＝ 

２．１．２实体关系抽取

除了命名实体识别之外，实体关系抽取（ＲｅｌａｔｉｏｎＥｘｔｒａｃｔｉｏｎ，缩写ＲＥ）和属性抽取也是知识抽取任务中非常重要的部分。其中，属性抽取实现比较简单，可以通过网络爬虫方式去百度百科、维基百科等网站进行规则抽取。实体关系抽取也是自然语言处理（ＮａｔｕｒａｌＬａｎｇｕａｇｅＰｒｏｃｅｓｓｉｎｇ，缩写ＮＬＰ）中最常遇到的问题，也一直是学术界和业界研究的热点方向之一。

实体关系抽取任务指的是给定用户输入的一段自然语言文本，在正确识别实体的基础上，抽取它们之间的关系，该任务被广泛应用于数据简化和构建知识图谱任务中。目前解决实体关系抽取问题可以分为流水线（Ｐｉｐｅｌｉｎｅｄ）和联合抽取两类方法。

流水线方法指的是首先通过实体识别提取出实体，然后再识别这些实体之间的关系，这种方式是一个分离的框架，使得实体识别与关系分类两个子任务易于处理且更加灵活。但是，实体识别和关系抽取都是一个独立的模型，忽略了这两个子任务之间的相关性，而且这种方法在实体识别任务的结果是否准确将会影响到关系抽取任务的结果，前后容易导致错误传播。

针对这一问题，基于机器学习的联合学习模型（Ｊｏｉｎｔｍｏｄｅｌ）被提出并逐步应用于对实体识别和实体关系抽取任务中进行联合学习。与流水线方法不同，联合学习模型是利用单个神经网络的端到端模型同时实现实体识别和关系抽取任务，将实体与关系一起抽取出来。该方法能够有效地将实体和关系的信息整合起来，在实体识别与关系抽取任务中取得了较好的效果。然而，现有的联合学习方法大多是基于特征的结构化系统，它们需要复杂的特性工程，并且严重依赖于其他ＮＬＰ工具包，也可能导致错误信息的传播。为了减少特征提取的人工工作量，最近Ｚｈｅｎ[1]等提出了一种混合神经网络模型来提取实体及其关系，而不需要任何手工制作的特征，该混合神经网络分别采用基于双向ＬＳＴＭ的编码器－解码器实现了实体提取模块，采用ＣＮＮ模型实现了关系分类模块。通过底层的模型参数共享，在训练时两个任务都会通过后向传播算法来更新共享参数来实现两个子任务之间的依赖，实体抽取模块中提取到的实体及其上下文信息会进一步传递给关系分类模块，用来提高关系分类的性能。虽然联合模型可以在单个模型中同时表示具有共享参数的实体和关系，但是它们也可以单独提取实体和关系，并产生冗余信息。Ｌｉ［2］等将联合学习的方法用于生物医学实体识别和关系抽取当中，在关系分类时，输入的语句首先进行依存分析构建起依存句法树，然后将这种树状结构输入到Ｂｉ－ＬＳＴＭ＋ＲＮＮ的网络中进行关系分类，两个任务的网络通过共享参数的方式联合学习，训练先进行ＮＥＲ，再根据ＮＥＲ的结果进行关系分类。

ＺｈｅｎｇｔＭ等人提出了一种新的标注策略来进行关系抽取，采用一种联合的方法把命名实体识别和关系抽取两步结合到一起：通过一种新的标注策略（ｔａｇｇｉｎｇｓｃｈｅｍｅ）把抽取任务转换为序列标注任务，然后利用深度学习的方法通过一个端到端的模型来抽取出最终的结果。该模型同样是使用ＢｉＬＳＴＭ来进行编码，然后使用参数共享中的ＬＳＴＭ来进行解码。本文也将借鉴该模型的思想应用到医疗知识图谱构建的知识抽取任务中。

## ２．２实体对齐

知识图谱形式的知识库己被广泛应用于问答系统，同一实体可能以不同的形式存在于不同的知识图谱中。通常，这些知识图谱在完整性方面是互补的，我们可以将这些知识图谱整合成一个更大的图谱来进行知识推论及应用。想要集成知识图谱（知识融合），一个基本的问题就是标识不同知识图谱中的实体，这些实体表示相同的实际对象，这通常称为实体对齐问题。

实体对齐目前包括有成对实体对齐与集体实体对齐、图型网络嵌入计算等方法，但仍然缺乏能够兼顾准确率、召回率、运行效率问题的算法。下文将对实体对齐的方法进行概述，在本文中也将研究两个知识图谱之间的实体对齐问题。其中，成对实体对齐方法是一种只考虑实体实例及其实体的属性的相似度计算的方法，集体实体对齐方法在计算实体相似度时是在成对实体对齐的基础之上考虑加入了实体之间的相互关系，具体地可分为全局集体实体对齐和局部集体实体对齐。全局集体实体对齐通过不同匹配决策之间的相互影响来调整实体间的相似度；局部集体实体对齐算法主要是使用余弦相似度和向量空间模型计算实体的相似性，具有比较不错的召回率和运行速度，但是准确率并不高。

最近，基于嵌入式的模型被提出用于实体对齐任务。这种模型建立在知识图谱嵌入模型的基础上，例如ＴｍｎｓＥ学习实体的嵌入，这种嵌入基于知识图谱中的关系三元组捕捉知识图谱中实体之间的语义相似性。为了适应两个知识图谱之间实体对齐的图谱嵌入，基于嵌入的模型要求两个图谱的嵌入落在相同的向量空间中。为了解决这个问题，可以分别学习每个知识图谱的嵌入空间，并提出使用转换矩阵将每个知识图谱的嵌入空间映射到另一个。这种模型依赖于大量的种子排列（从两个知识图谱中对齐的三元组的种子集）来计算转换矩阵。然而，两个知识图谱之间的种子排列很少，因此很难获得，因为需要付出昂贵的人力。

但是，实体对齐同样存在一些问题和挑战。实体在不同的知识库中存在严重的实体共指消解的问题，这不仅包含在对齐的实体对之间，同时包含在实体周边关系或属性实体的对齐场景。因此，在解决了共指消解问题之后，需要构建实体归一化处理，统一规定出标准的医学实体名称，以及同义医学词库。例如：“流行性感冒”在寻医问药网中包含症状表现、并发疾病、常用药品等，而在３９健康网中实体名为“流感”，而且在疾病的症状、并发症关系中同样存在共指消解问题，例如：前者的症状中有“胃寒”，而在后者中是“怕冷”，因此若采用实体对齐算法在考虑实体周边实体相似性时同样存在着共指消解的问题。而且，当对齐了同义的实体之后，仍然需要考虑两个知识源的可靠性，为两个网站设定置信度，用来确定哪个知识来源更可信。因此，实体对齐是本文实现医学知识融合过程中非常重要的任务。

## ２．３基于语义解析的知识库问答

目前世界上的海量信息存储在结构化和半结构化知识库中，查询这些知识库有效而简单的方法是必不可少的，而且不应仅限于那些具有正式查询语言专业知识的人。跨越自然语言处理、信息检索、智能问答和人机交互的一个重要研究领域是自然语言理解，而语义分析是这一目标的核心，它是一种从自然语言表达到语义表达的映射。

语义解析（ｓｅｍａｎｔｉｃｐａｒｓｉｎｇ）是将自然语言的语句转换成可以在知识库上轻松执行的逻辑形式。语义分析领域的发展是从早期使用手工制作的基于规则的系统到今天人们尝试使用神经网络将自然语言转换成可执行的正式查询语言。这种方式生成的语义表示超越了对句子中角色和对象的浅显识别，从而支持自动推理。此外，有必要认识到语义解析与其他序列预测任务本质上是不同的，因为它涉及对固有结构化对象的预测，此外，它还必须遵守某些约束以便在引入特定领域的知识库上执行。

基于语义解析方法的知识库问答（ＫＢＱＡ）是指先利用语义解析的方法理解自然语言问题的语义，将问题转化为具备相同语义的逻辑形式，再通过查询引擎对生成的逻辑形式进行查询处理，得到最终结果。上述过程分别对应语义解析和查询执行两个主要阶段。这类方法的优点在于如果解析成功，则能完整获得提问者的意图，从而精确地返回查询结果，这在实际的系统应用上有很高的应用价值。与此同时，将生成的逻辑形式展示给用户可以让用户检验系统是否理解正确，哪一步解析发生了错误，从而利用用户反馈进一步改善系统的精度。因此这种方法是一种“可解释”的面向知识图谱的自然语言问答方法。语义解析式的知识库问答系统所生成的逻辑形式一般与对应数据库的查询语言相同，例如ＲＤＦ知识库对应的ＳＰＡＲＱＬ语言；Ｎｅｏ４ｊ图数据库对应的Ｃｙｐｈｅｒ语言等。

作为基于语义解析的知识库问答系统中最重要的阶段，语义解析（问题理解）一般可以拆解为三个任务，分别是实体识别、关系／属性映射和语义组合。

（１）实体识别

实体识别的目标是将问题中蕴含有效信息的核心实体词抽取出来。蕴含有效信息指的是能够对应到知识库中的特定元素（实体、关系、属性）。该任务的难点就是上文提到的在实体边界识别和实体类型确定上的难点。同时，在知识库问答场景中会存在问句中实体词与知识库中多个实体的冲突问题，例如“副流行性感冒的症状是什么”，而知识库中既有“副流行性感冒”也有“流行性感冒”，如何选择哪个实体短语作为识别结果存在着挑战。

（２）关系／属性映射

给定自然语言问题中的一个或多个主体实体短语，需要实现关系／属性映射的任务，其目标是将自然语言问句中的关系描述映射到知识库中的关系或属性。关系解决方案主要包括依赖于预定义的模板进行关系匹配、根据构建好的短语关系同义／释义词典进行识别、通过神经网络模型进行相似度计算等。关系映射任务的难点在于隐式关系和复杂关系。

当然，这个过程中有涉及到实体的链接问题，实体链接的难点在于如何消除歧义。知识库中的实体在自然语言中可能有多种表述形式，自然语言短语也可能指代不同的实体。如何选择合适的实体需要考虑到上下文环境、字符及语义层次的相似度，以及实体本身的流行度等相关因素。

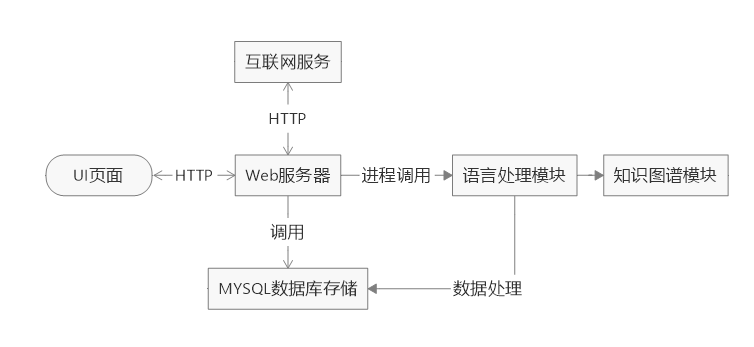
（３）语义组合

将自然语言问题中的主体实体词抽取出来，并将它们与知识库中的实体、关系、属性等元素对应起来。语义解析的最后一步就是将这些元素组合成对应的逻辑形式。对于简单的问题，一般只包含单一实体和关系，只需要将其相连即可。对于复杂的问题，可能包含多个实体和关系则需要考虑通过某种规则生成各实体和关系之间的配对组合。当问题被转化为结构化的逻辑形式时，只需要使用相应的数据查询引擎执行查询就可以了。

# 方案设计

## 系统架构设计

本次工程实践目的在于实现一个完整的医药领域的智能问答系统，系统能够接收用户输入，通过系统的语义识别转换与知识图谱分析匹配，最终以迅速返回给用户正确合理的语音回复为系统输出。为实现该系统，具体将系统分为以下几个功能模块：



**整体功能模块示意图**

1. 用户UI页面模块

前端展示为PC端web页面，主要功能为登录/注册页面、问答页面，还会有一些操作指引以及关于我们等。

1. WEB 服务器模块

承接UI模块输入的文字信息，保存用户的登录信息，用户基本信息，存储查询数据库信息，为UI模块提供必要的接口，对用户输入做处理，传递用户输入给语言处理模块。

1. 自然语言处理模块

从web服务器接口获取用户输入信息，对输入的文本信息进行预处理，提取实体和关系，理解用户意图，生成查询关键词和句子级embedding，方便后续通过关键字或者文本相似度查询数据库。

1. 知识图谱模块

知识图谱模块主要功能为从爬取的语料库中通过规则和深度学习的方法抽取实体以及关系，构造实体（entity）-关系（relation）-实体（entity）三元组，并通过多渠道知识融合，将相同实体的多种关系合并起来，组成丰富的知识图谱，通过Neo4j数据库来检索和存储。

## 模块详细设计

### 4.2.1用户UI页面模块功能设计

1. 前端框架采用Vue.js ，配合Element UI ，axios 等工具，实现web 端页面。
2. 用户登录/注册页面，实现基本的登录/注册以及用户信息存储功能。
3. 问答页面，部分数据前端固定，其他数据由后端返回。UI交互上面，尽量参考市场上主流的问答系统或网页聊天样式进行设计。
4. 前后端接口格式，采用RESTful规范。
5. 前端代码部署到Linux服务器上，用nginx 做转发请求。

### 4.2.2 WEB服务器模块功能设计

(1) Web服务器模块使用java语言开发，与UI模块交互协议为HTTP协议，主要功能为接收UI模块用户请求数据，响应用户的语音请求,为UI页面提供必须的接口。

(2)具体的功能可分为：用户注册功能，用户登录服务功能，语音信息接收，语音信息转换文字信息，数据库服务调用，数据存储与查询，自然语音处理模块进程调用，返回处理结果给用户。

(3) 调用互联网现有的一些接口进行语音处理，对一些简单的回答可以调用数据库或者也互联网已有接口进行快速回复。

(4) 对用户的操作历史记录进行处理，并且给登录用户推送信息。

(5) 整体架构采用springboot框架,maven 项目构建管理工具,以及tomcat web服务器。

### 4.2.3 自然语言处理模块功能设计

自然语言处理模块使用python语言开发，该模块主要功能为对用户输入的文本信息进行预处理。

1. 设置停用词例如语气助词、无关的标点符号等，对句子级文本分词，去除干扰信息，为后续语义理解做预处理。
2. 使用传统基于规则的方法，抽取英文单词（药物名称或专用名词）。
3. 使用word2vec(或one-hot)对中文分词进行embedding，构建embedding矩阵。

### 4.2.4 知识图谱模块功能设计

知识图谱模块包括知识图谱的构建以及检索。

1. 构建知识图谱

知识图谱的语料来源有两种类型，结构化数据和非结构化数据。对于结构化数据来说，使用传统基于规则的方法提取实体和抽取关系效率更高。对于非结构化数据需要使用基于深度学习的方法。

抽取实体即将语料库中的地名、人名、药物名称、医疗专业术语等提取出来，实体是后续检索的一个重要特征。

关系提取需要将不同实体之间的属性、关系等依赖信息提取出来，将不同的实体链接和实体组成三元组（triple），方便存储至数据库存储和检索。

1. 检索知识图谱

根据用户输入的文本信息进行分词和embedding，对中文句子的语义理解，提取句子特征，根据传统规则或者实体识别提取关键词。

根据提取的关键词检索知识图谱数据库，将检索的实体关系三元组生成文本，通过计算语句的文本相似度来选取最为相近的答案输出。

算法模块设计

算法部分包括NER（命名实体识别）和自然语句生成两部分模型。

NER部分使用BiLSTM-CRF模型来进行命名实体识别，命名实体识别属于自然语言处理领域的序列标注问题，在解决序列标注问题时，待标注序列的前后关系是研究的重点，而标注序列的前后关系，可以通过BiLSTM(Bidirectional LSTM)来获得。BiLSTM借助存储单元的结构来保存较长的依赖关系，并且通过输入门、输出门和遗忘门来调整之前状态对当前存储单元状态的影响。然而BiLSTM缺乏在整句层面的特征分析，所以需要借助条件随机场(Conditional Random Fields, CRF)的帮助。CRF将序列标注的重点放在句子级别上，根据特征模板来进行标注，通过Viterbi解码来获得最优解。然而CRF有提取特征困难，适用性不够广的问题，因此可以将CRF和LSTM结合起来，这样在保证能够提取足够整句特征的同时，使用有效的序列标注方法进行标注。

服务器端模块设计

# 中期进展情况

## 完成情况说明

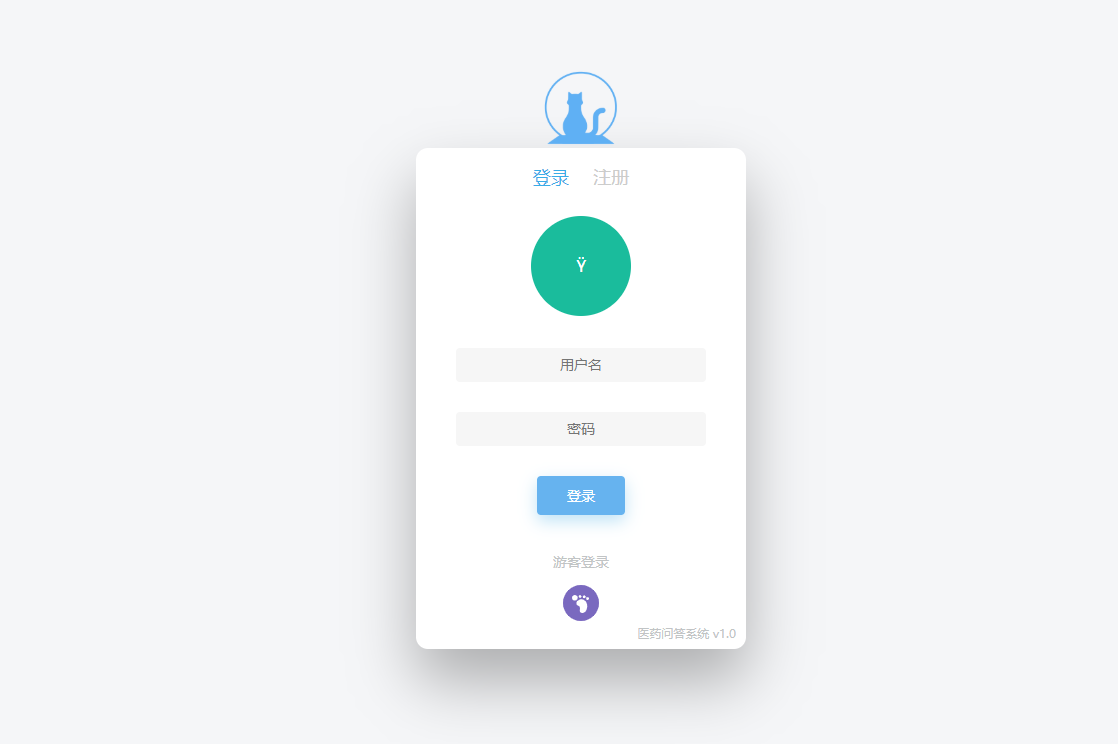
|  |  |  |  |
| --- | --- | --- | --- |
| 任务 | 分工 | 完成情况 | 说明 |
| 用户登录/注册页面 | 陈一 | 100% | 用户登录/注册功能，并存储用户信息。 |
| 问答页面 | 陈一 | 70% | 医药问答页面交互设计。 |
| 后端Rest接口封装 | 秦景建 | 100% | 和前端定义接口交互，实现数据传输。 |
| 后端业务逻辑处理 | 赵伟超 | 80% | 定义用户登录及交互的业务规则。 |
| 交互记录的存储 | 赵伟超 | 80% | 实现交互信息的落地，并基于此记录进行相应的规则处理。 |
| 提供后端与模型的接口联通 | 南晴 | 30% | 利用django封装模型，提供对应接口服务。 |
| 实体识别模型搭建及训练 | 刘亚楠 | 90% | 使用医药专业的语料进行模型训练、根据实验结果调整网络参数。 |
| 搭建知识图谱 | 南晴 | 90% | 针对语料进行实体和关系识别，并处理成知识图谱的结构 |

表1中期完成情况说明

## 取得的成绩

* + - 1. 前端界面

主要功能的页面已经大致完成，但是数据还处于前端模拟阶段，后续将同后端进行联调。



登录/注册页面



问答页面

* + - 1. WEB服务模块

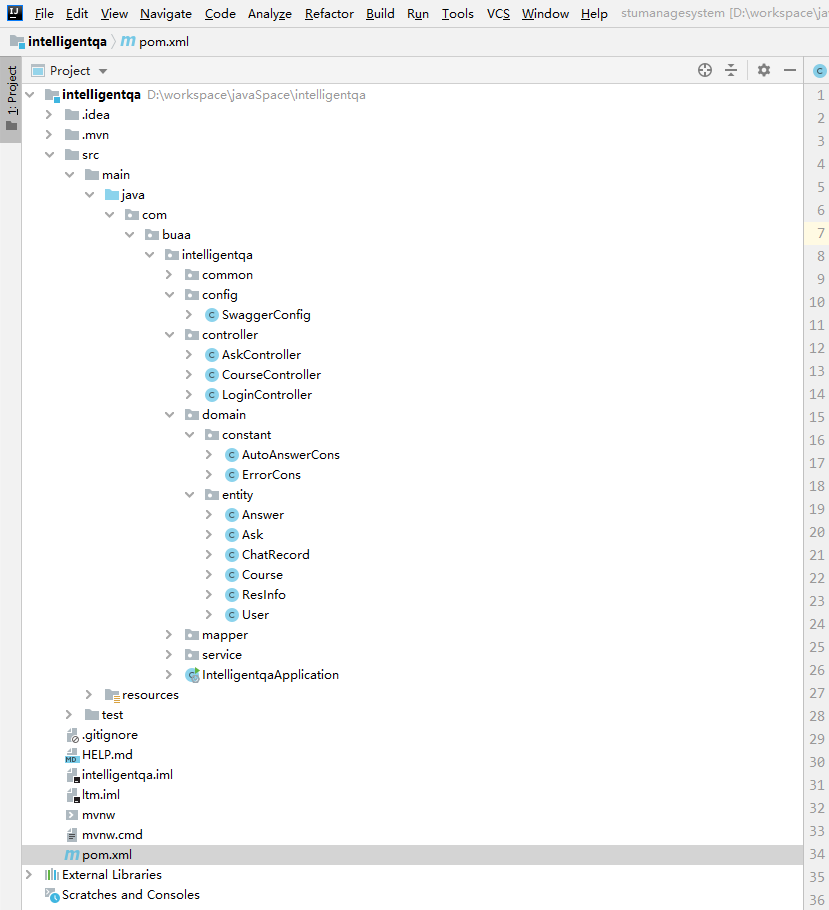
Web服务端整体框架搭建与设计已经完成，服务端开发语言主要为JAVA，JDK版本1.8，工程模块整体使用Springboot Springbook版本为2.2.2,服务端构建打包使用Aache Maven,数据库使用的是Mybatis 版本为8.0，Web容器使用的Apache Tomcat 版本为8.5

目前服务端已经完成整体框架设计，整体框架搭建，服务端参数配置,包括Springboot配置，ApacheMaven配置，Mybatis配置，Swagger配置，HTTP接口访问测试

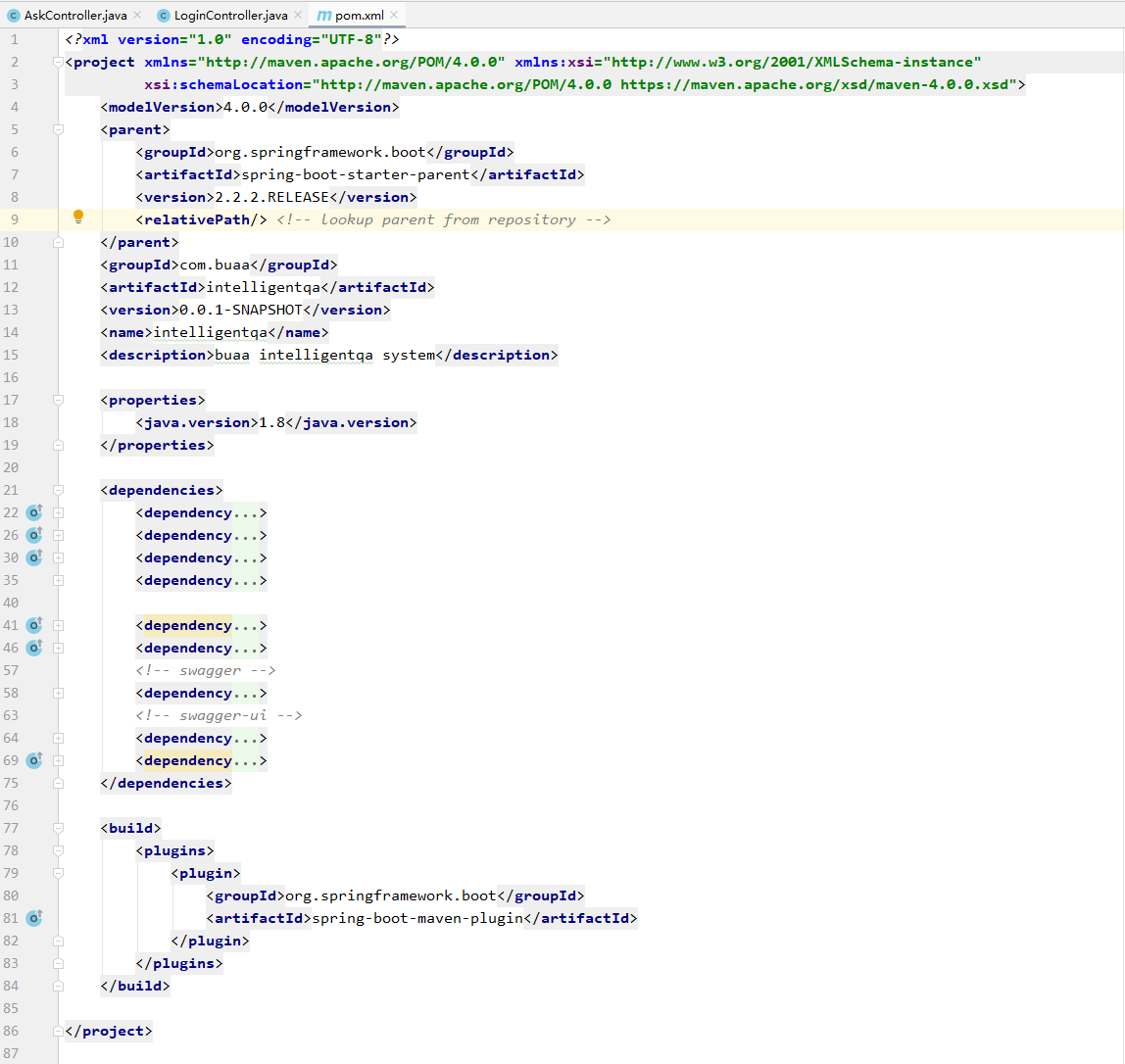
数据库部分已经完成数据库数据模型定义，数据表结构设计，数据库存储SQL设计

WEB接口部分已经完成业务对象设计与创建，用户登录，用户注册，用户信息查询，问题接口，信息保存，信息查询等等，部分接口都已经测试通过。

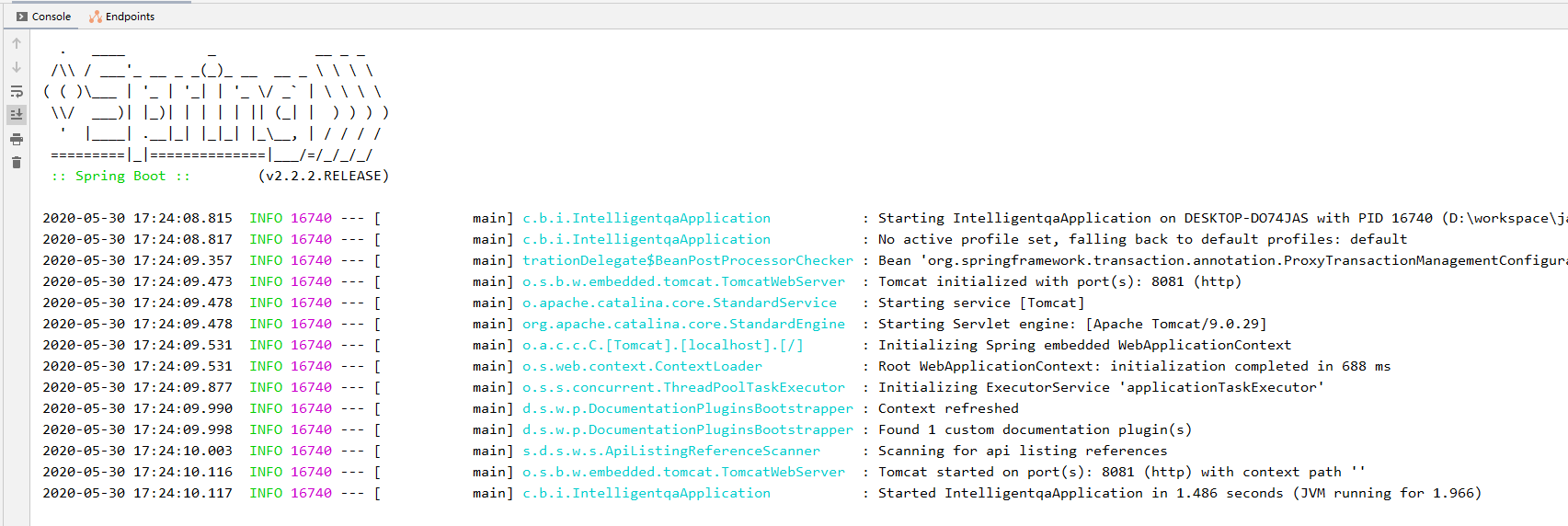
工程组织结构图如下所示



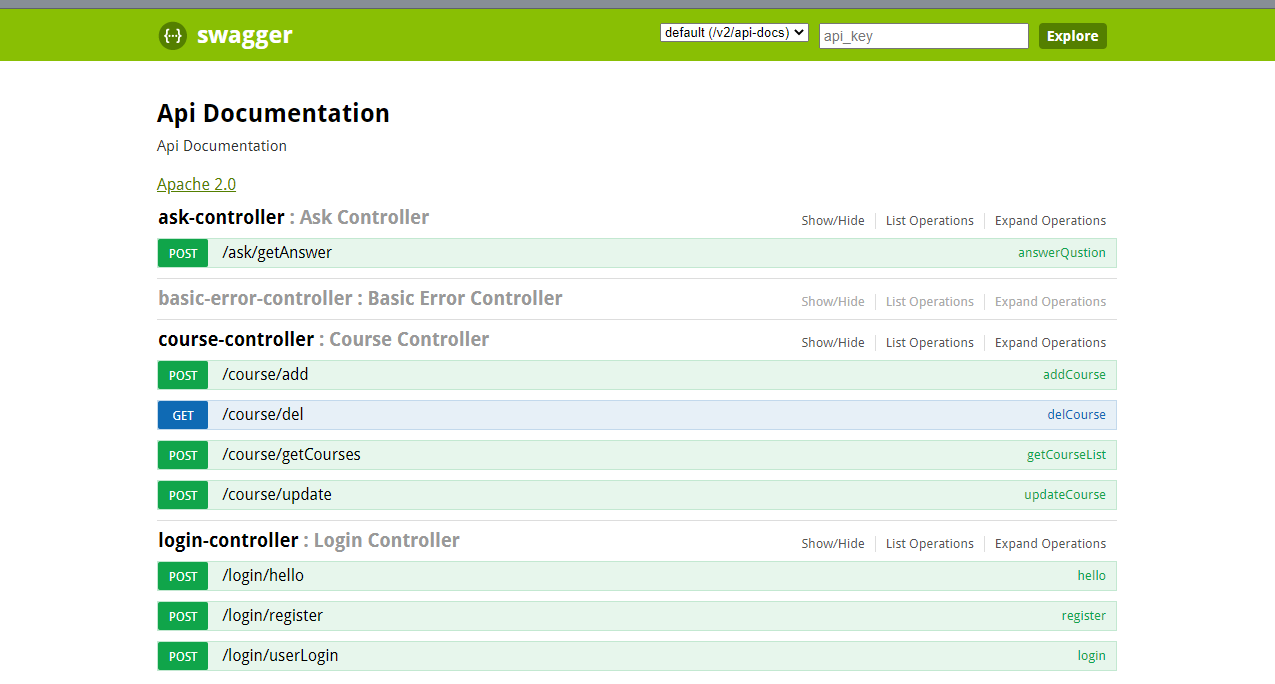
Apache Maven 配置文件pom.xml配置截图如下所示：



服务端监听8081端口，服务器启动截图：



WEB服务对外提供HTTP服务接口（部分）示意图如图所示：



* + - 1. 算法模型

根据语料信息使用BiLSTM-CRF进行模型训练，已初步实现一个模型

* + - 1. 知识图谱搭建

根据语料信息已经搭建了完整的知识图谱数据，但结构较简单，无法实现复杂的语义理解和推理过程

## 遇到的问题

|  |  |
| --- | --- |
| 问题 | 说明 |
| 模型训练速度较慢 | 算法使用的语料较大，计算机算力有限。 |
| 模块间调试困难 | 现在模块化开发，各自模块进度可观，但因人员分散，整体流程尚未调试通过，存在集成风险。 |

遇到的问题

## 后续实施计划及风险

|  |  |  |
| --- | --- | --- |
| 计划 | 说明 | 风险 |
| 基于BiLSTM-CRF进一步优化模型参数 | 对语料数据做具体分析，进行针对性的调优 | 模型算法理解深度不到位，出现进行参数调整后，模型改进不明显 |
| 组织周会进行进度跟进 | 组内分模块开发，整体进度不能及时同步 | 组内人员分散，且工作时间不稳定。不确定能否组织固定时间全员进度汇报。 |
| 进行全流程功能调试 | 组内任务分模块开发， |  |

后续计划及风险