

上海对外经贸大学

# 课程论文

论文名称:	基于中医知识图谱的任务型问答系统设计与实现
课程名称:	自然语言处理
学 院:	统计与信息学院
专 业:	数据科学与大数据技术
学 号:	18076014      18076016
学生姓名:	陈郁欣      徐颖
指导教师:	刘亮亮

2021 年 6 月

# 目 录

一、绪论.....	3
1.1 研究背景与意义.....	3
1.2 研究现状综述.....	3
1.3 研究内容.....	4
二、中医知识图谱构建.....	5
2.1 知识图谱构建流程.....	5
2.2 知识来源.....	5
2.3 知识抽取.....	5
2.4 知识存储.....	8
三、基于中医知识图谱的问答系统.....	10
3.1 问答系统构建流程.....	10
3.2 问句解析.....	10
3.3 实体链接.....	15
3.4 知识获取与答案返回.....	16
四、实验与结果分析.....	18
4.1 中医知识图谱构建.....	18
4.2 基于中医知识图谱的问答系统.....	19
五、改进方向.....	22
5.1 知识图谱.....	22
5.2 问答系统.....	22
六、结论.....	23
七、团队分工.....	24
参考文献.....	25

## 摘 要

问答系统采用自然语言作为交互方式,针对用户提出的问题给出简短、有效的答案。利用问答系统对信息进行检索符合现代社会对于高效且精准地获取信息的需求趋势,同时知识图谱技术的发展为问答系统提供了充裕的数据支撑,更加有效地促进了问答提供的发展。知识问答系统以存储结构化知识知识图谱为数据基础,对用户提出的自然语言形式的问题进行处理与理解,从而对答案进行搜索并返回给用户。中医作为中国的传统医学,积累了深刻且有效的医疗实践理论。基于中医知识图谱的任务型问答系统的研究主要包含以下内容:

首先,本文以《本草纲目》作为全文数据基础,以草药为中心,利用爬虫技术从中医网站抽取有效的知识信息并进行结构化,从而构建出包括 6 种实体类型,多种关系类型的中医知识图谱。

其次,对用户的自然语言问句进行中医实体的抽取和类型的划分。在实体识别中,通过嵌入层将训练问句转换为特征向量,作为 BiLSTM 的输入,最后通过 CRF 层约束结果标签,输出命名实体识别结果,模型识别准确率为 98.73%。对抽取出实体的问句的其他部分进行朴素贝叶斯分类从而得出问句的类型,分类准确率为 93.98%。

最后,获取到问题的中医关键实体和对应的问句类型,并对实体进行切分和相似度处理后,将问题转化为中医知识图谱上的查询语言,在中医知识图谱中进行信息检索,按照模板生成答案并返回给用户。

实验结果表明,基于中医知识图谱的任务型问答系统针对中医领域的相关知识,完整地实现了从数据采集到问答检索应用的过程,能够实现对用户的自然语言问句进行较为准确的回答。

**关键词:** 知识图谱, 中医, 问答系统

# 基于中医知识图谱的任务型问答系统设计与实现

## 一、绪论

### 1.1 研究背景与意义

中医药知识传承的前提条件，是对中医基础理论和知识体系的整理和分析（张德政等，2017）。如何借助信息科学技术对中医理论和知识体系进行组织和分析，挖掘古籍或文献中隐藏的学术思想、临床经验和辨证方法，是值得探索的重要问题。

随着计算机科学技术不断发展，现代信息化技术为大量的中医药信息数字化提供了技术支持。近年来，知识图谱已成为知识管理领域中的一项新兴技术，并得到了广泛的应用。把知识图谱应用到重要领域，有助于实现中医药信息的知识化结构重组，为智能搜索、自动问答、辅助诊疗、知识推荐、决策支持等智能医疗领域提供了技术手段。

### 1.2 研究现状综述

#### 1.2.1 知识图谱

知识图谱在中医药领域的应用实例归纳如表 1 所示。

表 1 知识图谱在中医药领域的应用

应用领域	应用实例	应用阶段	应用展望
中医基础	构建了基于《脾胃论》的方剂本体	智能检索	实现本体在智能检索、诊断，图书情报和生物信息学等领域的应用
	构建了中医脾脏象理论知识图谱	知识表示	推广应用于中医领域
	构建了以五脏为中心的中医知识本体	知识表示	1. 自定义规则； 2. 实现本体语义关系的量化 3. 利用本体自学习机制，实现本体概念表达体系自动构建
	构建了舌象、脉象本体	知识表示	1. 完善舌象、脉象的属性体系。 2. 完善软件功能，将规范舌象、脉象部分的功能做成输入法的形式
中医临床	构建了中医药知识图谱	决策支持	利用中医药知识图谱，对电子病历进行自动化解析和标注，形成基于知识图谱技术的临床病例库

	构建了哮喘疾病本体	知 识 表示	1. 构建整个哮喘疾病的语义网络 2. 将哮喘本体应用于临床医学领域
	构建了糖尿病医案本体库	智 能 检索	辅助用户查找, 学习中医医案信息, 从而发扬中医思想
	构建了医学诊断知识图谱	决 策 支持	将致力于辅助检查等方面的研究
	构建了中医脾胃病本体	决 策 支持	进一步构建可扩展、可重用、更广泛的中医辅助诊断系统
	构建了中医师辨证论治知识围语	知 识 表示	将名老中医经验传承与创新进行可视化
中医养 生保健	构建了中医养生知识图谱	知 识 推荐	对中医养生知识进行深度挖掘并促进其共享、传播与利用
	构建了中医健康知识图谱	知 识 表示	1. 对知识图谱内的知识进行查询; 2. 对实体重要程度进行排序。
	构建了面向中医养生的冠心病知识本体	知 识 表示	将进行实证性研究, 以证明本体的有效性。
其他	构建了中药功效的语义网络	知 识 表示	1. 改善方剂总体功效算法 2. 方剂智能检索
	构建了针灸传统知识本体	智 能 检索	进一步实现中医古籍针灸知识的表示和利用

资料来源: 孙华君, 李海燕, 聂莹, 甄思圆. 知识图谱及其在中医药领域应用研究进展[J]. 世界科学技术-中医药现代化, 2020, 22 (06) :1969-1974.

### 1.2.2 问答系统

在中医领域, 智能问答系统的研究也取得了一定的成果。

2012 年, 中国工程科技知识中心启动了中草药知识服务系统课题建设, 其中包括对智能问答系统的建设。钱宏泽等人(钱宏泽, 2016)设计并实现了中草药语义网的构建, 综合规则和统计的方法对用户提出的问题进行理解与回答。陈程等人(陈程, 翟洁等, 2018)将中医药知识图谱以及智能问答系统相结合, 借用知识图谱技术对中医药知识进行可视化。

## 1.3 研究内容

经过对知识内容和结构的分析, 本文以草药为中心, 在中医领域中, 证型、方剂和重要之间表达形式多样化、关系复杂, 贯穿整个中医治疗过程。

为了更加有效地分析中医药知识之间的联系, 优化知识的检索, 共享中医领域知识, 促进中医知识一体化, 本文旨在实现建立中医药领域知识图谱, 利用规则的方法对本体进行知识推理, 完善中医知识, 并基于中医知识图谱的构建建立

问答系统。

## 二、中医知识图谱构建

### 2.1 知识图谱构建流程

知识抽取的过程如图 1 所示。

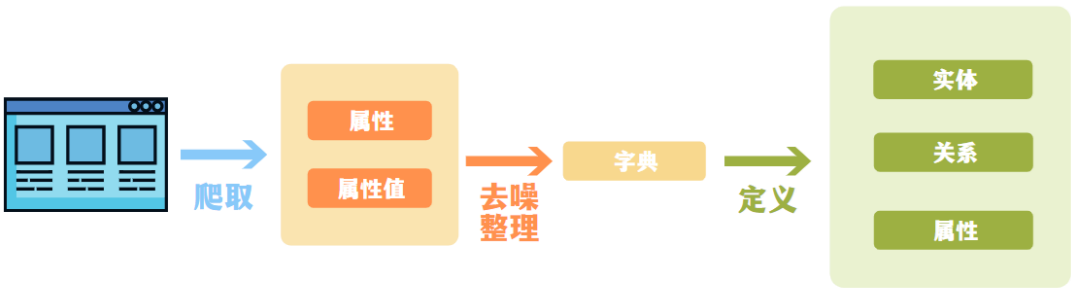


图 1 知识抽取过程

### 2.2 知识来源

本文的数据主要来源于中医世家网站，获取方法为从中医世家网页爬取 (<http://www.zysj.com.cn/lilunshuji/bencaogangmu/index.html>)。

### 2.3 知识抽取

爬取的文本不可避免地存在数据不干净、表达不规范等噪声问题。因此，在使用这些中医数据之前需要进行去噪处理，剔除无效信息。

具体去噪方式如下：

表 2 去噪方式

去噪方法	去噪前	去噪后
去除数据中的空白字符	戴  疹	戴疹
去除数据中的非法字符	出圣□方	出圣方

爬取的书籍是以“药”为中心整理的。爬取的药可以分为两种结构，其一为单层结构，如“石硫赤”，见表 3；其二为多层结构，如“鸽”，该药可以分为“白鸽肉”与“屎（左盘龙）”，见表 4。

表 3 石硫赤结构

石硫赤			
草药部门	释名	气味	主治

金石部	亦名石亭脂、石硫丹、石硫芝。为硫磺之呈现红色者，功同硫磺。	苦、温、无毒。	<p>1、赤鼻作痛。用呈紫色的石亭脂（红色者次之，黄色者勿用），研细，冷水调搽患处，半月可愈。</p> <p>2、风温脚气。用生石亭脂一两、生川乌头一两、无名异二两，共研为末，葱白捣汁和药做成丸子，如梧子大。每服一钱，空心服，淡茶加生葱送下。</p>
-----	-------------------------------	---------	---

表 4 鸽结构

鸽				
草药部门	释名	气味	主治	
禽部	鹑鸽、飞奴	白鸽肉：咸、平、无毒。鸽尿：辛、温、微毒。	白鸽肉	尿（左盘龙）
			解药毒，治恶疮、疥癣、白癜风等	<p>1、带下排脓。用野鸽尿一两，炒至微焦。白术、麝香各一分，赤芍药、青木香各半两，延胡索（炒赤）一两，柴胡三分，共研为末。每服一钱，空心服，温酒调下，脓排尽后，可服其他药物补养身体。</p> <p>2、蛔虫寄生。用白鸽尿烧过，研细，水送服适量。</p> <p>3、项上瘰癧。用鸽尿炒过，研为末。加饭做成丸子，如梧子大。每服三、五十丸，米汤送下。</p> <p>4、头痒生疮。用白鸽尿五合，加醋煮开三次，捣烂敷涂。一天三次。</p>

为了减少存储冗余以及挖掘更多关系，需要对爬取的数据进行整理以及实体对齐，建立新的数据结构。需要对文字进行切割，以便于实体的建立，在经过整理与划分之后，单层结构的草药结构可以表示为如表 5 所示，而多层架构的草药可以切分为多个结构，如表 6-表 8 所示。

表 5 切割后的石硫赤结构

石硫赤
-----

草药部门	释名	气味	主治	
			症状	方法
金石部	石亭脂	苦、温、无毒。	赤鼻作痛	用呈紫色的石亭脂(红色者次之,黄色者勿用),研细,冷水调搽患处,半月可愈。
	石硫丹		风湿脚气	用生石亭脂一两、生川乌头一两、无名异二两,共研为末,葱白捣汁和药做成丸子,如梧子大。每服一钱,空心服,淡茶加生葱送下。
	石硫芝			

表 6 切割后的鸽结构

鸽				
草药部门	释名	气味	主治	
			症状	方法
禽部	鹑鸽	\	\	\
	飞奴			

表 7 切割后的白鸽肉结构

白鸽肉				
释名	气味	归属	主治	
			症状	方法
\	咸、平、无毒	鸽	解药毒	\
			治恶疮	\
			疥癣	\
			白癜风	\

表 8 切割后的鸽屎结构

鸽屎				
释名	气味	归属	主治	
			症状	方法
左盘龙	辛、温、微毒。	鸽	项上瘰癧	用鸽屎炒过,研为末。加饭做成丸子,如梧子大。每服三、五十丸,米汤送下。
			头痒生疮	用白鸽屎五合,加醋煮开三次,捣烂敷涂。一天三次。



将所有文本都整理成如表 5-表 8 所示的结构，则可以对此进一步进行知识图谱的建立。基于上述结构，本文以“草药”实体为中心，设定了“释名”、“气味”、“方法”、“症状”、“草药部门”五种实体，实体间的关系如表 9 所示。

表 9 实体关系

实体	关系	实体
释名	又名	草药
草药	气味	气味
草药	属于	部门
草药	归属	草药
方法	使用	草药
症状	被治疗	方法

## 2.4 知识存储

为了方便问答系统对知识图谱进行查询、分析，本系统采用 Neo4j 图数据库对中医知识图谱进行建立。Neo4j 将知识存储在网络上，利用节点和关系两种基本元素对知识体系进行组织。其中，节点指向现实世界中实际存在的实体对象，关系则对实体对象之间存在的关系进行描述，通过关系类型，将节点相互连接起来，形成了一种关系型的网络结构。

在文本的知识图谱构建中，实体与结构的关系如图 2 所示，其中多个中药实体可以指向同一个中药部门实体，多个中药实体可以归属于同一个中药实体，多个释名实体可以指向同一个中药实体，多个方法实体可以指向同一个草药实体，多个症状实体可以指向同一个方法实体。

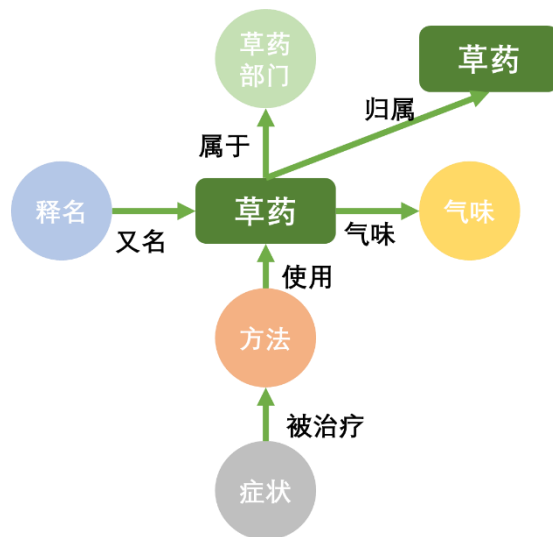


图2 实体与结构关系

在实例中，上图可以表示为：

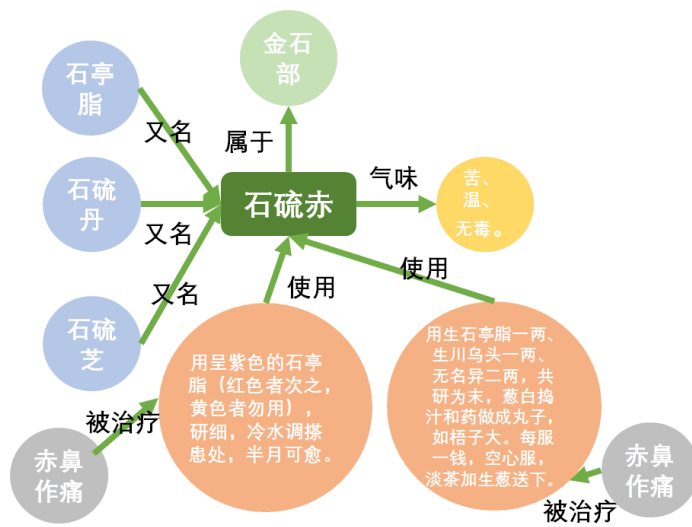


图3 石硫赤关系图

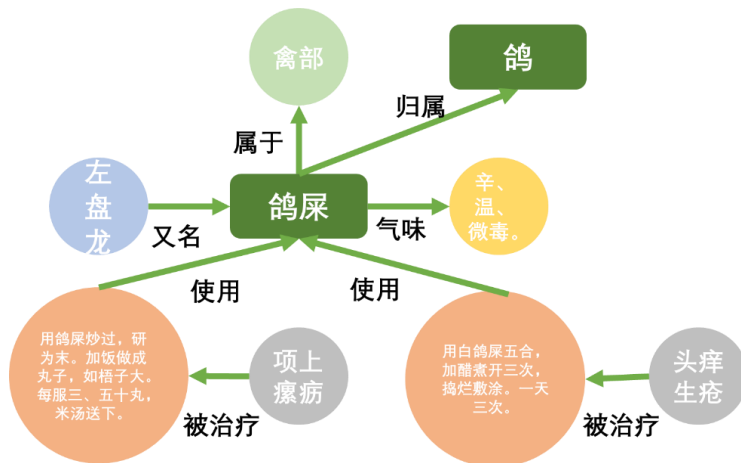


图4 鸽屎关系图

### 三、基于中医知识图谱的问答系统

#### 3.1 问答系统构建流程

该任务型问答系统采用了规则匹配方式完成。该问答系统支持的问答类型为：

表 10 系统支持问答类型

问句类型	回答实体	举例
L1-回答方法	方法实体	我感冒了怎么办
L2-回答症状	症状实体	甘草对治疗什么比较有用
L3-回答释名	释名实体/草药实体	紫檀还能叫啥
L4-回答气味	气味实体	甘草有什么气味吗
L5-回答子部	草药实体	鸽下面还有哪些部
L6-回答部门	部门实体	草部包含了哪些草药



图 5 问答系统构建流程

#### 3.2 问句解析

##### 3.2.1 基于 BiLSTM-CRF 的命名实体识别

###### ● BiLSTM

基于 BiLSTM-CRF 的命名实体识别模型主要包含嵌入层、BiLSTM 层、CRF 层和输出层，如图 2 所示。命名实体识别时，首先通过嵌入层将句子转换为特征向量，然后送入 BiLSTM 层，随后通过 CRF 层约束结果标签，最后在输出层输出分类结果。BiLSTM-CRF 模型的基本架构如图 6 所示。

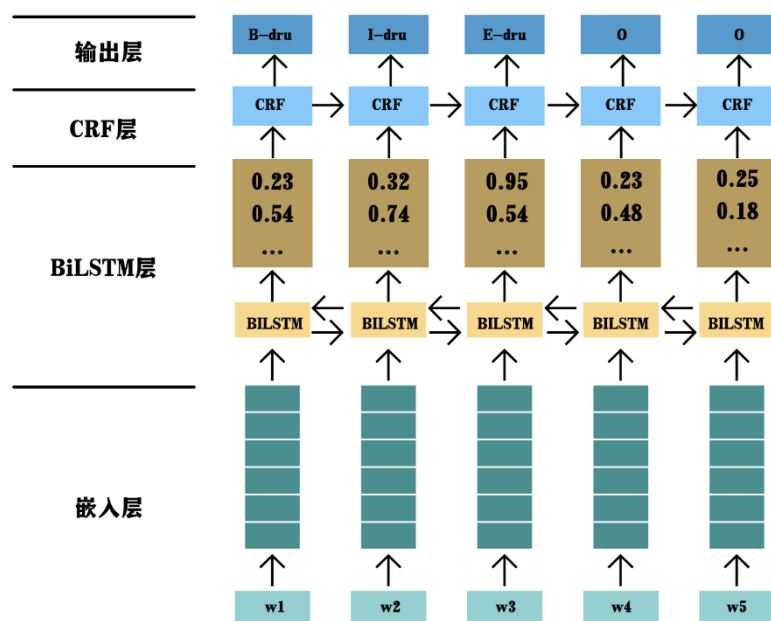


图 6 BiLSTM-CRF 模型架构

采用“BIO”标注规范对数据进行标注。其中，命名实体首字符用“B”表示，结尾命名实体用“I”表示，非命名实体用“O”表示。例如：对于句子“我最近头疼，流鼻涕，还咳嗽，请问有什么方法吗？”可以标注成“O O O B-SYM E-SYM O B-SYM I-SYM E-SYM O O B-SYM E-SYM O O O O O O O O O”

在该任务中，命名实体主要包括三大类，分别为“部门”、“草药”、“症状”，分别使用“DEP”“DRU”“SYM”来表示各个命名实体标签，下面是以“月经不调吃哪些药物”为例，采用 BIO 标注规范标注该句的结果如表 11 所示。

表 11 命名实体标注范例

序号	实体	标签
1	月	B-SYM
2	经	I-SYM
3	不	I-SYM
4	调	E-SYM
5	吃	O
6	哪	O
7	些	O
8	药	O
9	物	O

BiLSTM -CRF 模型的第二层为 BiLSTM 层，其结构如图 7 所示。

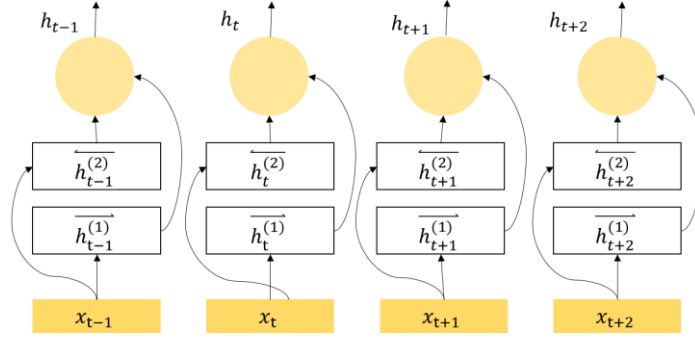


图 7 BiLSTM 层模型架构

BiLSTM 层的具体步骤如下：

- (1) BiLSTM 的输入层接收嵌入层生成的特征向量，分别将特征向量的正序和逆序作为前向 LSTM 和后向 LSTM 的输入；
- (2) 前向 LSTM 和后向 LSTM 分别计算得到前向和后向隐状态向量；
- (3) 隐藏层将前向和后向隐状态向量拼接，形成完整的隐状态向量，并将隐状态向量映射到  $k$  维空间( $k$  为标注集合的总标签数，在此总标签数为 17)；
- (4) 隐藏层的输出作为 CRF 的输入。

图 6 中 BiLSTM 层的输出是每个标签的分数。例如，对于  $w_1$ ，BiLSTM 节点的输出为 0.23 (B-sym)、0.54 (I-sym)、...，这些分数将作为 CRF 层的输入。

#### ● CRF

条件随机场是类似于 HMM 的序列建模框架。该模型的目标是学习映射函数  $x_s \rightarrow y_s$ ，使得正确的输出标签最大化。但是，每个输出  $y_s$  并不是独立的。CRF 模型能够通过计算给定观察到的特征向量  $x = (x_1, x_2, x_3 \dots x_t)$  的随机变量条件概率来预测输出向量  $y = (y_1, y_2, y_3 \dots y_t)$ 。

假设 CRF 的输入和输出是线性链，并且  $P(Y|X)$  服从马尔可夫性质。需要找到参数来建立模型，所以通过给定输入的单词序列  $X = (X_1, X_2, X_3 \dots X_n)$ 。  $P(Y|X)$  对应的方程式为：

$$P(Y|X) = \frac{1}{Z(X)} \exp \left( \sum_{i,k} \lambda_k t_k(Y_{i-1}, Y_i, X, i) + \sum_{i,l} \mu_l s_l(Y_i, X, i) \right)$$

$$Z(X) = \sum_Y \exp \left( \sum_{i,k} \lambda_k t_k(Y_{i-1}, Y_i, X, i) + \sum_{i,l} \mu_l s_l(Y_i, X, i) \right)$$

其中， $t_k(Y_{i-1}, Y_i, X, i)$  表示在给定序列  $X$  的情况下，序列  $Y$  在位置  $i-1$  处的值

转换为 $i$ 的概率。 $s_l(Y_i, X, i)$ 表示在给定序列 $X$ 的情况下，序列 $Y$ 的相应值在位置 $i$ 处的权重。 $\lambda_k$ 和 $\mu_l$ 是两个函数的权重。

$t_k(Y_{i-1}, Y_i, X, i)$ 和 $s_l(Y_i, X, i)$ 是特征函数，如果设 $s_l(Y_i, X, i) = s_l(Y_{i-1}, Y_i, X, i)$ ，则可以通过下面的公式来设置它们：

$$\lambda_k = \begin{cases} 1, & \text{关于 } Y_{i-1}, Y_i \text{ 的条件} \\ 0, & \text{其他} \end{cases}$$

因此，对于给定的句子(词) $X$ 可以给标签序列 $Y$ 打分：

$$\text{score}(Y|X) = \sum_{j=1}^m \sum_{i=1}^n \lambda_j F_j(Y'_{i-1}, Y'_i, X, i)$$

然后可以计算得分概率：

$$P(Y|X) = \frac{\exp \text{score}(Y|X)}{\sum_Y \exp \text{score}(Y|X)}$$

使用对数似然算法作为估计函数来估计 CRF 的权值：

$$L(\lambda) = \sum_{j=1}^m \sum_{i=1}^n \lambda_j F_j(Y'_{i-1}, Y'_i, X, i) - \sum_{j=1}^m \log z_\lambda(X_n)$$

使用改进的迭代缩放算法通过以下公式求出向量增量 $\delta$ ：

$$E_{P'}(t_k) = \sum_{x,y} P'(x, y) \sum_{i=1}^{n+1} t_k(Y_{i-1}, Y_i, X, i)$$

$$E_{P'}(s_l) = \sum_{x,y} P'(x, y) \sum_{i=1}^{n+1} s_l(Y_i, X, i)$$

然后使用 $\delta$ 更新当前参数 $\lambda = \lambda + \delta$ 。如果并非所有 $\lambda$ 都收敛，重复该步骤。

对于给定的特征向量、权重向量和观测序列，可使用维特比算法来预测单词的分类，步骤如下：

1) 通过以下公式进行初始化

$$\delta_1(j) = wF_1(y_0 = \text{start}, y_1 = j, x), j = 1, 2, \dots, m$$

2) 找出每个标签  $l$  在位置 $i$  的最大概率：

$$\delta_t(l) = \max_{1 \leq j \leq m} \delta_{t-1}(j) + wF_1(y_{t-1} = j, y_t = l, x), l = 1, 2, \dots, m$$

3) 记录路径：

$$\psi_t(l) = \arg \max_{1 \leq j \leq m} \delta_{t-1}(j) + wF_1(y_{t-1} = j, y_t = l, x), l = 1, 2, \dots, m$$

4) 停在 $i = n$ ，在这个位置，最大概率是：

$$\max_y (wF(y, x)) = \max_{1 \leq j \leq m} \delta_n(j)$$

5) 最佳路径的终点：

$$y_n = \arg \max_{1 \leq j \leq m} \delta_n(j)$$

6) 从最佳路径返回：

$$y_t = \psi_{t+1}(y_{t+1}), i = n - 1, n - 2 \dots, 1$$

最后得到最优路径 $y = (y_1, y_2 \dots y_n)^T$ 。输出路径 $y$ 即预测命名实体序列。

### 3.3.2 贝叶斯分类

由于缺少问句训练语料，将上述命名实体识别后的问句中标记为“SYM”、“DRU”、“DEP”的部分挖空后，根据问句模板，并结合人工对问句模板内部分词语进行近义词替换，对其进行问句生成。例如：句子“月经不调吃哪些药物”将会被命名实体识别为“B-SYM I-SYM I-SYM E-SYM O O O O O”，对其进行挖空后为“吃那些药物”。

将所有生成的问句通过贝叶斯分类进行问句分类。其中，问句可以分为如表12。

表 12 问句分类

问句类别	举例
L1-回答方法	可以吃什么缓解
L2-回答症状	能缓解什么
L3-回答释名	还有什么其他名字吗
L4-回答气味	闻起来啥样
L5-回答子部	下面还包含了什么部吗
L6-回答部门	包含了些啥草药

本文采用先验为多项式分布的朴素贝叶斯方法对问句进行六个类别的划分，朴素贝叶斯在贝叶斯对求解条件联合分布 $P(\mathbf{X}|\mathbf{C} = c_k)$ 的时候，假设在给定的类别 $\mathbf{C} = c_k$ 下，不同维度特征的取值之间是相互独立的，即 $n$ 维特征中某个维度的概率跟另一个维度无关

$$P(X_1 = x_1 | \mathbf{C} = c_k) = P(X_1 = x_1 | X_2 = x_2, \mathbf{C} = c_k)$$

其中 $X_1$ 和 $X_2$ 为 $n$ 维特征向量 $\mathbf{X}$ 里的两个维度。

在朴素贝叶斯条件独立性的假设下，给定类别的条件概率求解如下

$$P(C = c_k | X_1, X_2, \dots, X_n) = \frac{P(C = c_k) \prod_{i=1}^n P(X_i | C = c_k)}{P(X_1, X_2, \dots, X_n)}$$

使用标注好的训练集数据，估算出所有的 $P(C = c_k)$ 以及 $P(X_i = x_i | C = c_k)$ 即可。对用户提出的新待分类问句，使用问句的特征向量的取值对每个类别求出 $P(C = c_k) \prod_{i=1}^n P(X_i | C = c_k)$ ，选择最大值即可确定此问句的分类标签

$$C \leftarrow \arg \max_{c_k} P(C = c_k) \prod_{i=1}^n P(X_i | C = c_k)$$

为了避免新样本问句中有词没有出现在训练数据中导致概率为 0 的情况，本文使用拉普拉斯平滑技术对条件概率进行计算。

### 3.3 实体链接

对于基于知识图谱的问答系统来说，实体链接时用户输入的自然语言形式的问句中询问的实体和知识图谱中该实体对应节点的相互映射过程。因而，本步骤的主要目的是从自然语言形式的问句中提取出所提问的中医实体对象。在命名实体识别中，识别出问句中所提问的中医实体对象有“部门”、“草药” / “释名” / “子部位” / “父部位”、“症状”，故在此对识别出的三种实体进行链接。

#### 3.3.1 “部门”实体的链接

实体链接的难点在于两个方面，即多词一义和一词多义，对于“部门”实体的识别，由于不存在一词多义或者多词一义，所以直接以原词查找。

#### 3.3.2 “草药” / “释名” / “子部位”实体的链接

对于“草药”实体的链接，考虑三个方面：

寻找“草药”节点

寻找“草药”节点的释名

寻找“草药”节点的子部位或者父部位

#### 3.3.3 “症状”实体的链接

由于用户输入的问句中，通常包含多个症状，所以要对组合症状进行分词，以此拆分症状。

对最大概率法分词是在最大匹配分词算法上的改进。在某些语句切分时，按最大长度切分词语可能并不是最优切分。而不按最优长度切分词语，则同一语句会出现多种切分结果。计算每种切分结果的概率，选取概率最高的切分作为最优分词切分。



### 1) 计算切分概率

以 $S$ 表示多症状组合句， $W$ 表示所有可能的切分组合，那么对于输入句 $s$ ，其最佳切分可以视为：

$$\arg \max_w P(W|S = s)$$

在给定组合句 $s$ 的情况下，切分概率为：

$$P(w|s) = \frac{P(w)P(s|w)}{P(s)}$$

其中， $P(s)$ 为定值， $P(s|w)$ 恒等于 1，故仅需计算 $P(w)$ 。

在词表中需记录每个词语的词频，计算每个独立词语出现的联合概率。使用一元语法（不考虑词语上下文关系）则：

$$P(w_1^n) = P(w_1, w_2 \dots w_n) = P(w_1)P(w_2) \dots P(w_n)$$

在实践中一般用对数求和替代求积，原因是词语概率一般会很小，多次连乘后可能造成溢出。

$$P(w_1^n) = \sum_{i=1}^n \log P(w_i)$$

### 2) 平滑

同上朴素贝叶斯算法的平滑方法，使用拉普拉斯平滑计算词表。

### 3) 切分词语

切分词语时从语句中每个字作为词语前缀字符查询词表，查询结果构造为 DAG 图，如此表中无此词汇，则单字成词。有两种方法计算，其一是对 DAG 图做 DFS 搜索，得到所有切分组合，分别计算各组合概率，取最大概率组合输出，其二，对 DAG 图做动态规划 DP 算法，从深度搜索这里可以看到，其实很多局部计算过程是重复的。因此这里从图后面向前面反向计算，每次累计最大概率结果。计算每个字为前缀的不同切分词语概率和后续最大切分概率的乘积，取最大结果为最佳切分，依次向语句前方递推。最后得到语句第一个字开始的最佳切分词语位置，并由第一个字开始的最佳切分词语位置向后依次查询切分词语，最终得到最佳切分。由于第一种方案计算复杂度高，所以在本次实验种采用第二种方案。

## 3.4 知识获取与答案返回

### 3.4.1 知识获取

在前文的基础上，获得了问句所询问的中医实体对象和目的意图。本步骤主

要是利用这两类信息在中医知识图谱上进行知识的查询、匹配，从而获得问题的答案。

该知识图谱的实体类型有“草药”、“释名”、“部门”、“气味”、“方法”、“症状”六类，结合前文提到的问句分类，具体可以表示为以下几种类型，如表 13 所示：

表 13 问句实体匹配

问句类别	实体	支持提问的 实体数目	意图	获取方式
C1	草药/释名/子 部位	1 个	部门	①首先，尝试从草药实体寻找关系。②若返回为空，尝试从释名实体链接到相应草药，并寻找关系。③若返回再次为空，尝试从子部位实体链接到相应草药，并寻找关系。
C2	草药/释名/子 部位	1 个	气味	
C3	草药/释名/子 部位	1 个	方法	
C4	草药/释名/子 部位	1 个	症状	
C5	部门	1 个	草药	尝试从部门实体寻找关系。
C6	症状	4 个以内	方法	尝试与提供的多个症状出/入度最多的方法实体。

### 3.4.2 答案返回

当按照表 13 所列规则将问句转化成存储中医知识图谱的 Neo4j 数据库上的 Cypher 查询语句后，系统返回的结果是中医知识图谱中的结点信息。

为了使系统返回的答案对用户更友好，更符合人类的认知习惯，本文创建不同类型的答案模板，将查询到的结点信息以占位符的形式进行赋值，从而以完整的形式将答案返回给用户。答案模板集合如表 14 所示。

表 14 实体对应答案模板

实体	意图	答案模板
草药/释名/子部位	部门	① <u>草药/释名/子部位</u> 所属的部门是 <u>部门</u> 。

		② <u>草药/释名/子部位</u> 是属于 <u>部门</u> 的。
草药/释名/子部位	气味	① <u>草药/释名/子部位</u> 闻起来 <u>气味</u> 。 ② <u>草药/释名/子部位</u> 的气味是 <u>气味</u> 。
草药/释名/子部位	方法	① <u>草药/释名/子部位</u> 的使用方法是 <u>方法</u> 。 ② <u>草药/释名/子部位</u> 可以这样用： <u>方法</u> 。
草药/释名/子部位	症状	① <u>草药/释名/子部位</u> 可以用来治疗 <u>症状</u> 。 ② 可以治如果你有以下症状： <u>症状</u> ，可以使用 <u>草药/释名/子部位</u> 。
部门	草药	① <u>部门</u> 下有这些草药： <u>草药</u> 。 ② 属于 <u>部门</u> 的草药有： <u>草药</u> 。
症状	方法	① <u>方法</u> 可以治疗的 <u>症状</u> 有 <u>症状</u> 。 ② <u>症状</u> 的治疗方法为 <u>方法</u> 。

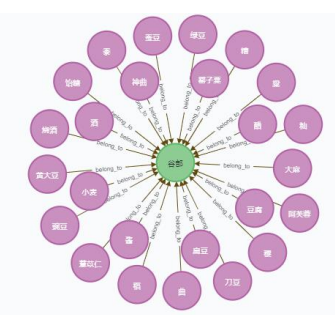
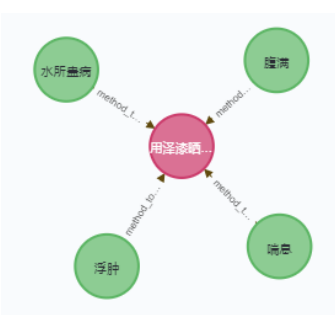
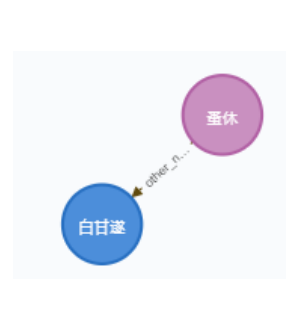
注：下划线表示具体实体

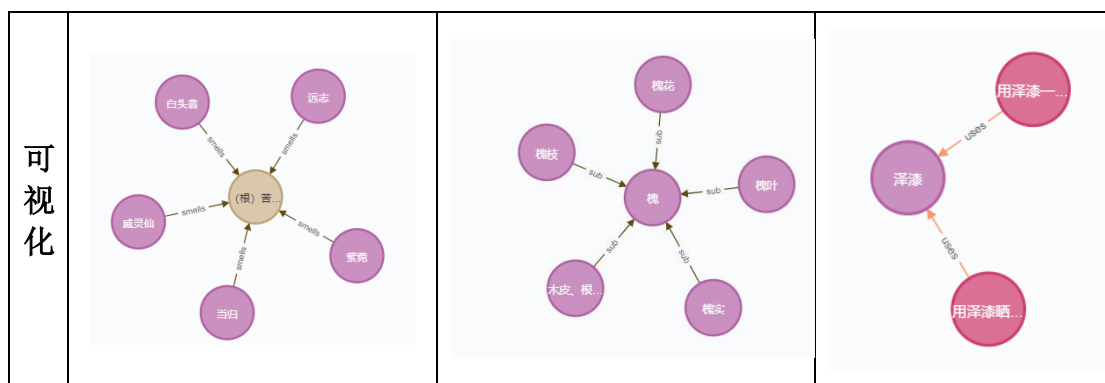
## 四、实验与结果分析

### 4.1 中医知识图谱构建

在本文所构建的知识图谱中，具有 14022 个实体，实体间的关系有 15055 条。其中，实体与实体间的关系可视化展示如表 15：

表 15 实体间关系可视化

实体	草药	症状	草药
关系	属于	被治疗	又名
实体	部门	方法	释名
可视化			
实体	草药	草药	方法
关系	气味	归属	使用
实体	气味	草药	草药



## 4.2 基于中医知识图谱的问答系统

### 4.2.1 数据构建

在本文对问句进行处理的数据构造过程中，对《本草纲目》全书进行数据清洗与结构化处理，共提取出草药名 2931 条作为草药名语料库，症状短语 5123 条作为症状语料库，以及 15 条草药总部名作为部门名语料库。使用 (<https://github.com/baiyang2464>) 医药问答系统中的现代医疗问句作为基础数据，抽取实体以外的部分作为问句模板，并人工制作针对本文特有数据的问句模板，共生成问句模板 329 条。针对问句模板，随机提取语料库中不同数量的实体对模板进行不同位置的填充，模拟用户可能的问法，最终生成标注好实体位置的完整问句共 128025 条。

### 4.2.2 命名实体识别

使用问句库中的 89617 条数据作为命名实体识别的训练数据，并取其中的 10% 作为训练过程中的验证集，其余 38408 条数据作为测试数据。采用双向 LSTM-CRF 的方式，在多次使用小批量数据进行试验后，选择词向量维度为 128，隐层状态含有 128 个神经元，在序列最长为 100 的情况下，迭代 16 次，训练中一次性批量选取 64 个数据，模型训练损失变动如下。

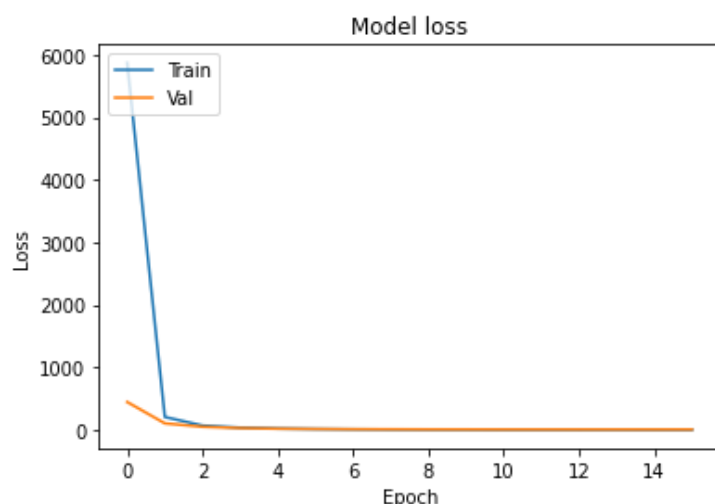


图 7 BiLSTM 层模型架构

从损失图可以看出，模型训练速度较快且训练集和验证集的损失在训练后期都趋近于平稳，损失趋近于 0，模型训练过拟合或欠拟合的情况较小。使用训练完毕的模型对测试集的识别结果如下，测试准确率为 98.73%，准确率较高。

表 16 命名实体识别测试结果

	precision	recall
B-DEP	1	1
B-DRU	1	1
B-SYM	0.99	0.99
E-DEP	1	1
E-DRU	0.95	1
E-SYM	0.99	0.99
I-DEP	1	1
I-SYM	1	1
I-DRU	0.99	0.99
O	1	1
S	0.99	1

#### 4.2.3 贝叶斯分类

使用 329 条问句模板，随机选择 75% 的数据作为训练集，在对训练数据进行切分并且按 0.5 为阈值忽略含停用词的数据，计算每条数据的 TF-IDF 值，使用 TF-IDF 值对朴素贝叶斯模型进行训练，取平滑值为 0.001，训练标签为 6 个。训练完毕的模型对测试集的分类准确率为 93.98%，测试数据如下，模型分类效果较好。

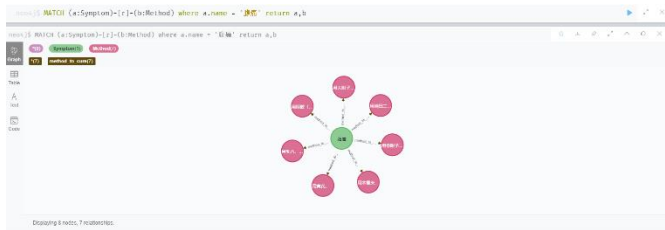
表 17 问句分类效果

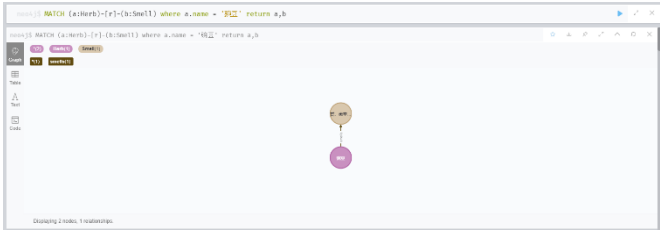
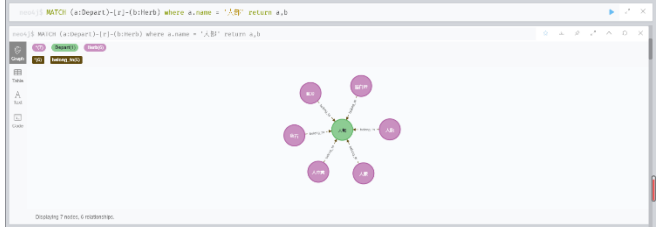
	precision	recall
L1	0.91	0.87
L2	0.88	0.93
L3	1	0.92
L4	1	0.88
L5	1	1
L6	0.9	1

#### 4.2.4 答案返回

使用部分问句对系统进行测试，返回数据如下。

表 18 部分实验结果

问题	腹痛怎么改善？
回答	<p>腹痛的治疗方法为：</p> <p>用木香末一钱，放入木瓜汁一杯中，加热酒调服。</p> <p>用香附子二两、艾叶半两，蒸日上在醋汤中蒸煮熟，去艾叶，炒香附子末，加醋、糊做成丸子，如梧子大。每服五十丸，开水送下。此方名“艾附丸”。</p> <p>用当归二两、芍一两，碎为粗末。每服三钱，以水一碗煎至将干，加酒一碗再煎开水后温服。过半小时，又服一次。不过三、五服，即可见效。此方名“神妙佛手散”。子尚活，可保胎；子已死，即产下。</p> <p>用大附子一个（去皮，切成片）、甘草（炙）二两、干姜三两，加不沽升，煮成一升，分两次温服，脉出现即愈。面赤，加葱九根；腹痛，加芍药二两；干呕，加生姜二两；咽痛，加桔梗一两；利止，而脉不出，加人参二两。此方名“脉四逆汤”。</p> <p>用阿胶（炒过，水化成膏）一两、黄连三两、茯苓二两，共捣匀做成丸子，如梧子大。每服五十丸，粟米汤送下。一天服三次，此方名“黄连阿胶丸”。</p> <p>用乳香、没药、木香等分，水煎服。</p> <p>用黄芪、芍各一两，糯米一合，水一升，一起煮到半到。分次服下。</p>
Cypher 语句	MATCH (a:Symptom)-[r]-(b:Method) where a.name = '腹痛' return a,b
可视化	
问题	豌豆闻起来怎么样？
回答	豌豆的气味是：甘、微辛、平、无毒。

Cypher 语句	MATCH (a:Herb)-[r]-(b:Smell) where a.name = '豌豆' return a,b
可视化	
问题	人部下有哪些草药？
回答	属于人部的草药有： 人尿 人胞 溺白沂 乱发 秋石 人中黄
Cypher 语句	MATCH (a:Depart)-[r]-(b:Herb) where a.name = '人部' return a,b
可视化	

## 五、改进方向

### 5.1 知识图谱

- 对现有知识进一步结构化，挖掘更多的关系。。如“方法”实体中，可以对中药、剂量再结构化；对症状、病名加以区分。
- 扩充知识库。若知识库中存在多个来源的数据，需要对其进行知识融合。
- 提高回答准确度。连接 API 对回答进行检查与补充。
- 在线补充。基于在线搜索得到的知识进行自动处理，对知识库自动更新。

### 5.2 问答系统

- 进一步对用户输入的预处理：消歧、纠错等。
- 训练过程中分类样本不平衡。
- 多轮回答。

- 一句多问。

## 六、结论

任务型问答系统的最终目的在于为用户提供直接并高效地获取知识的方式，相对于较传统的基于搜索引擎的方式，本文设计的问答系统能够更精准地限定于某些特定领域并且排除了广告等因素的干扰，给人们获取知识提供了新的便利。本文从知识图谱入手，对中医领域的问答系统进行研究。

本文研究的基于中医知识图谱的任务型问答系统主要完成了以下工作：

- 1) 面对网上中医知识图谱较少且中医相关知识较复杂的情况，直接选择对《本草纲目》原书进行处理，通过利用爬虫技术有针对性地对网页进行选择，并且对数据进行清洗以及结构化等操作，以草药名作为中心，构建特定领域的知识图谱，从而能实现整个系统研究的数据支持功能。
- 2) 自然语言问句的解析，在对用户问句进行解析时，首先使用 BiLSTM-CRF 模型抽取问句中的中医实体，实体类别为草药、症状和部门名，在导入训练好的模型后，每个问句的抽取动作能快速完成。其次，对问句实体以外的部分进行分类，分析问句的提问目的，使用朴素贝叶斯算法进行标签的分类，从而确定问句提问意图。在判断抽取实体是否合法及判断实体类别和问句是否可以配对后，进行下一步操作。
- 3) 对问句进行回答，利用解析中得到的实体以及提问目的，与知识图谱中的实体和关系属性进行链接并匹配，返回相应答案，其中对于症状实体，切分并按照相似度提取出具体知识图谱中含有的数据。
- 4) 在本文的相关理论中，主要包括知识图谱理论、词语向量化表示、词语相似度判别，以及基于词典的词语切分技术、基于深度学习的实体识别技术和基于统计的问句分类技术，对各个模块任务的准确率提供了保证。

通过对部分问句的测试，实验结果表明，系统能够对包含单实体且具有有效提问目的的自然语言问句给出较为理想的回答结果，但是本文研究还较为基础，与达到实际应用水平还有部分差距，例如本文只能根据单个实体问题进行回答、无法对用户问句实现进行消歧且图谱中实体数据量较少等。因此，在接下来的研究中，完善并优化以上问题是本文的研究方向。



## 七、团队分工

表 19 团队分工

成员	工作
徐颖	文本结构化 知识图谱构建 实体链接 知识获取 论文撰写 PPT 制作
陈郁欣	文本结构化 语料生成 问句分类 答案返回 论文撰写 PPT 制作

## 参考文献

- [1] 钱宏泽. (2016). 基于中草药语义网的自动问答系统的研究与实现. 浙江大学.
- [2] 陈程, 翟洁, 秦锦玉, 江嘉, 武海霞, & 蔡婷婷. (2018). 基于中医药知识图谱的智能问答技术研究. 中国新通信(2).
- [3] 张德政, 谢永红, 李曼, & 石川. (2017). 基于本体的中医知识图谱构建. 情报工程(1).
- [4] baiyang2464/chatbot-base-on-Knowledge-Graph. (2019). Retrieved 20 June 2021, from <https://github.com/baiyang2464/chatbot-base-on-Knowledge-Graph/#E5%8C%BB%E7%96%97%E5%91%BD%E5%90%8D%E5%AE%9E%E4%BD%93%E8%AF%86%E5%88%AB>
- [5] 基于 sklearn 和 keras 的数据切分与交叉验证 - 焦距 - 博客园. (2018). Retrieved 20 June 2021, from <https://www.cnblogs.com/bymo/p/9026198.html>
- [6] tensorflow2 实现 BiLSTM+CRF 中文命名实体识别 - 打工小黄人 - 博客园. (2021). Retrieved 20 June 2021, from <https://www.cnblogs.com/huanghaocs/p/14673020.html>
- [7] 毛宇. (2017). 中医药症状的中文分词与句子相似度研究. Retrieved 20 June 2021, from <https://www.doc88.com/p-7718626778042.html>
- [8] 刘焕勇. (2018). liuhuanyong/MedicalNamedEntityRecognition. Retrieved 20 June 2021, from <https://github.com/liuhuanyong/MedicalNamedEntityRecognition>
- [9] 王,岳. (2020). 基于知识库的问答 KBQA: seq2seq 模型实践. Retrieved 20 June 2021, from <https://zhuanlan.zhihu.com/p/34585912>
- [10] 刘,焕勇. (2020). liuhuanyong/QASystemOnMedicalKG. Retrieved 20 June 2021, from <https://github.com/liuhuanyong/QASystemOnMedicalKG>