项目介绍

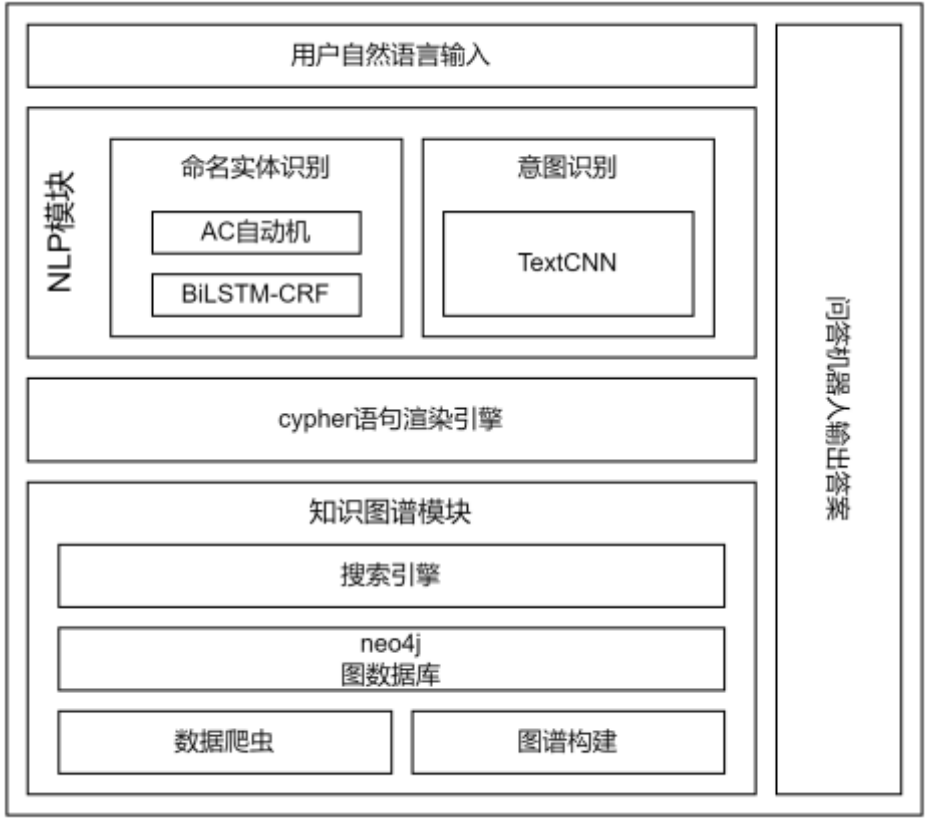
本项目是基于知识图谱的医疗问答机器人。使用者可以通过输入自然语言的问题来获取有关医疗健康方面的答案。

问答机器人的答案来源于知识图谱中已有的知识点和相关实体，并且采用了多种自然语言处理技术来提高识别和推理能力，最终能够给出准确、全面的答案。

项目介绍


欢迎使用医疗问答机器人
你好，我是您的医疗助手Tars，请问有什么可以帮到您?
你可以这样问我...
请输入您的问题，询问格式可参考左侧提示栏。
发送

整体架构

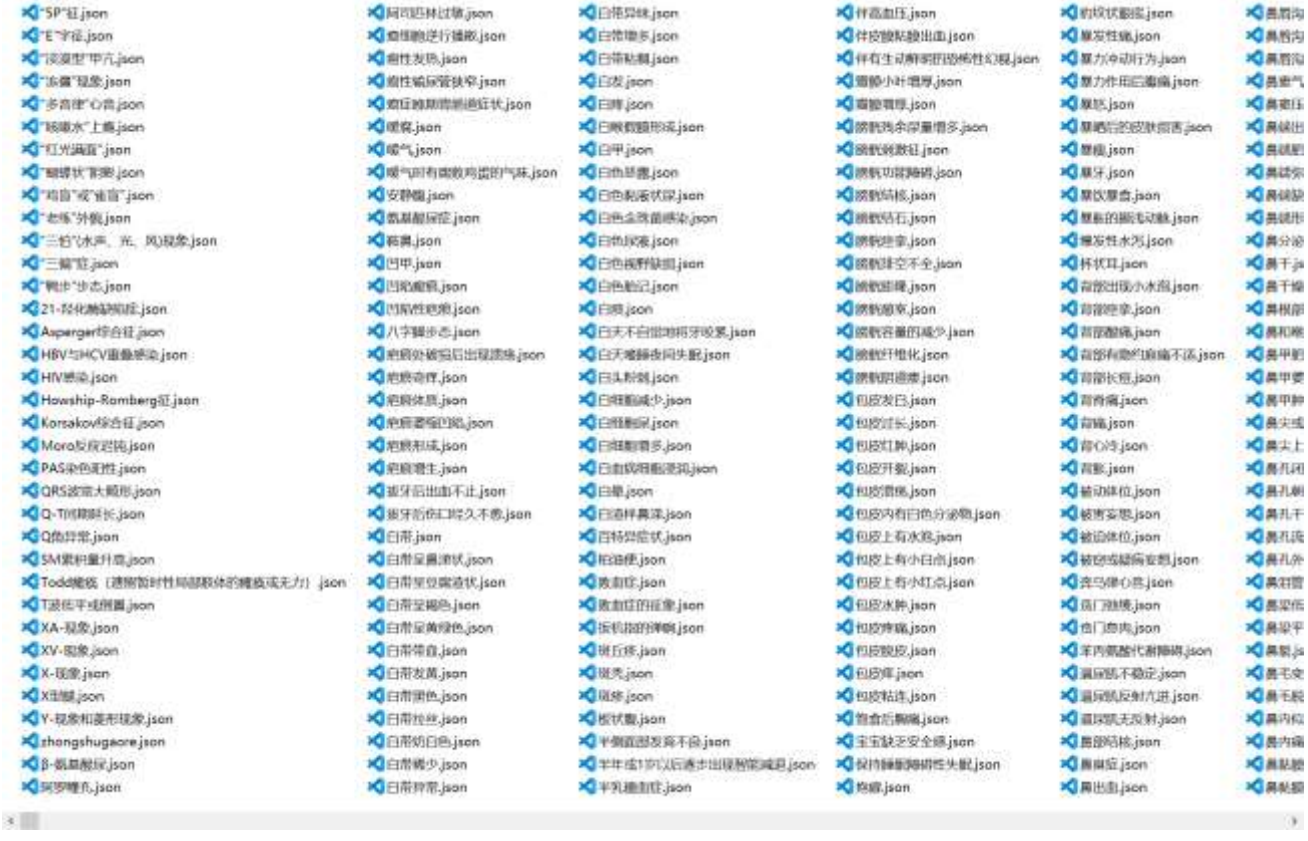


自然语言处理模块负责命名实体识别、意图识别等任务，cypher语句渲染完成后，由搜索引擎进行聚合与执行，知识图谱用于存储和管理医疗健康数据，整个流程中各模块协同工作，完成整个系统的问答过程。

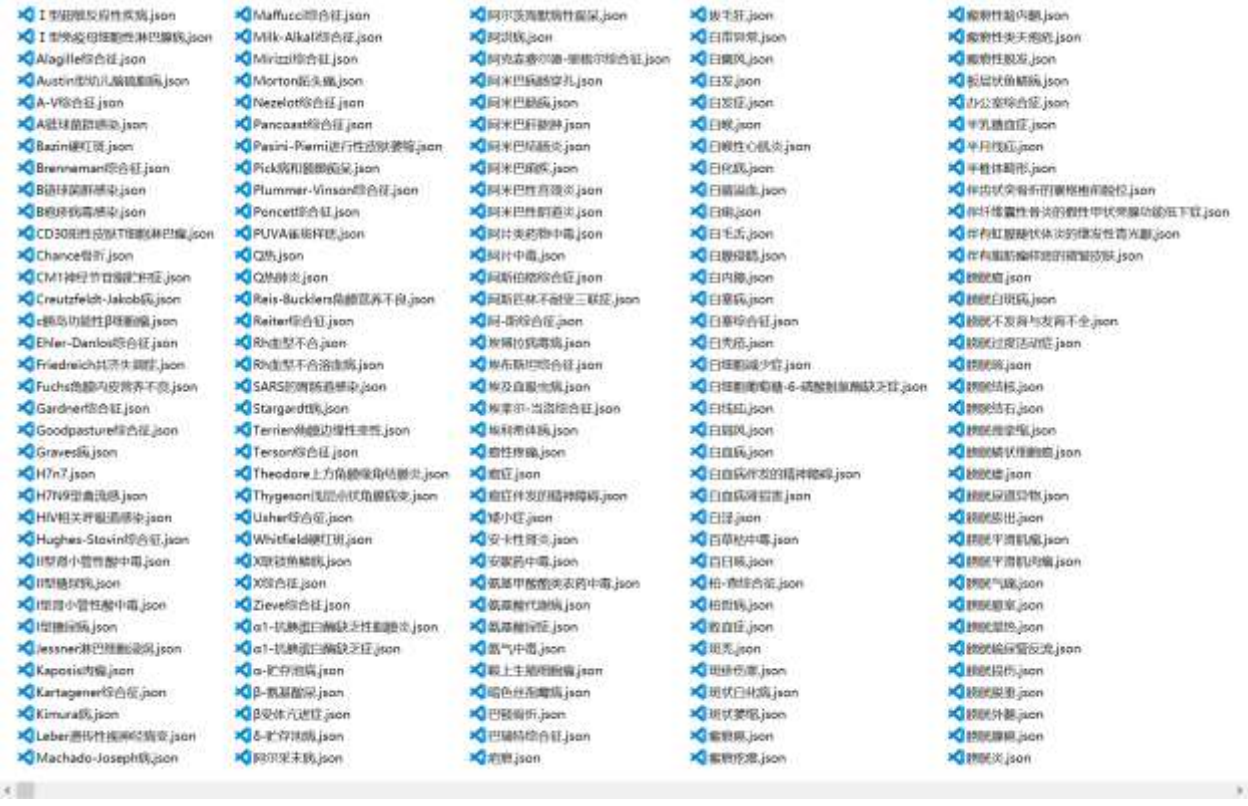
数据来源

构建知识图谱的数据来源于寻医问药网，通过爬虫进行抓取，经过数据清洗之后形成结构化数据，在本项目中主要抓取的是症状数据和疾病数据，其他实体（如医疗检查、科室部门和食物等）信息可从这两类数据中抽取。

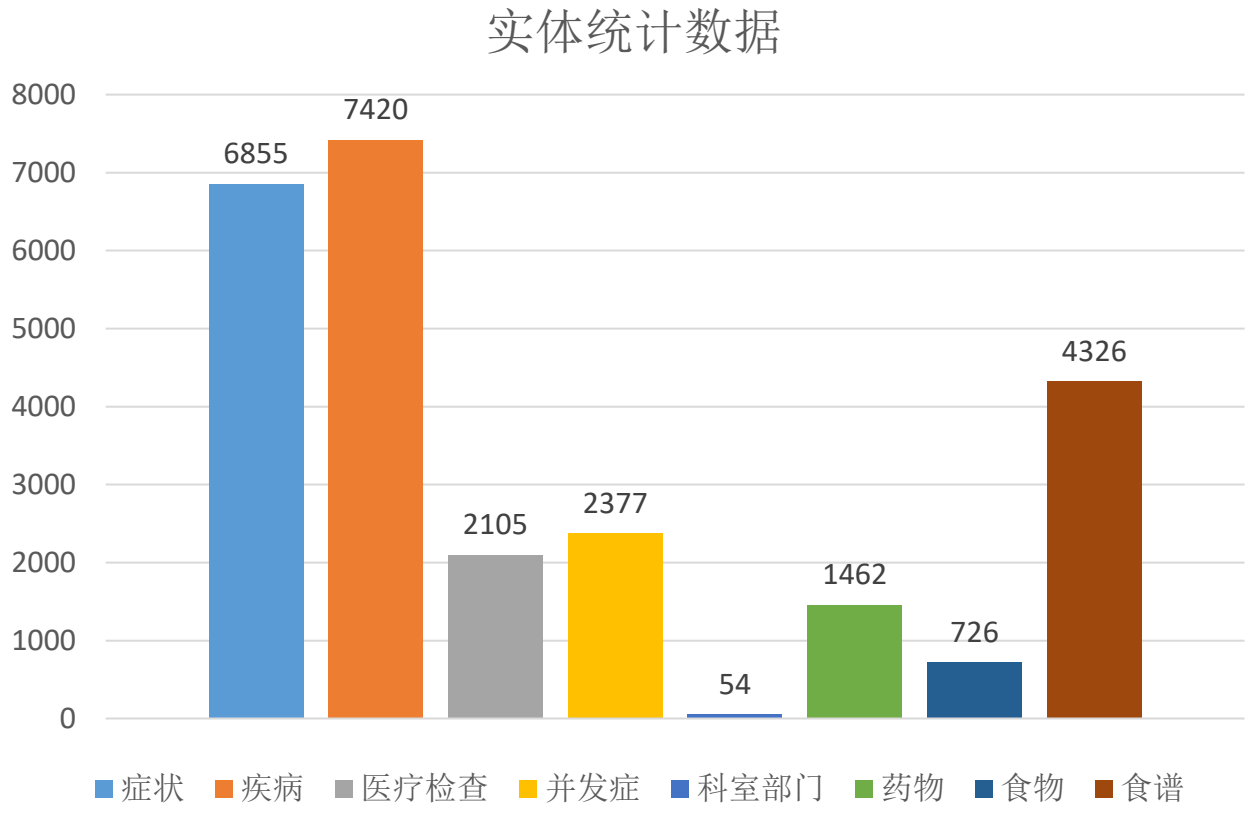
症状数据


"SP"征.json
"E"字征.json
"浸渍型"甲亢.json
"冰雹"现象.json
"步音律"心音.json
"铁锹水"上痿.json
"红光满面".json
"蝴蝶状"斑.json
"鸡盲"或"雀盲".json
"老练"外貌.json
"三怕"(水声、光、风)现象.json
"三偏"征.json
"鸭步"步态.json
21-羟化酶缺陷症.json
Asperger综合征.json
HBV与HCV重叠感染.json
HIV感染.json
Howship-Romberg征.json
Korsakov综合征.json
Moro反应迟钝.json
PAS染色呈阳性.json
QRS波增大畸形.json
Q-T间期缩长.json
Q波异常.json
SM累积量升高.json
Todd瘫痪（遗留暂时性局部肌体的瘫痪或无力）.json
T波低平或倒置.json
XA-现象.json
XV-现象.json
X-现象.json
X线.json
Y-现象和菱形现象.json
zhongshugaore.json
β-氨基酸尿.json
阿罗喹乐.json
阿司匹林过敏.json
癌细胞浸行播散.json
癌性发热.json
癌性输尿管狭窄.json
癌症晚期前驱症状.json
暗疾.json
暗气.json
暗气时有腐败鸡蛋的气味.json
安静型.json
氨基酸尿症.json
鞍鼻.json
凹甲.json
凹陷癌肿.json
凹陷性水肿.json
八字脚步态.json
把肩处被猫后出现溃疡.json
把肩奇痒.json
把肩休息.json
把肩萎缩凹陷.json
把肩和成.json
把肩增生.json
拔牙后出血不止.json
拔牙后伤口许久不愈.json
白带.json
白带呈黄泡状.json
白带呈豆腐渣状.json
白带呈褐色.json
白带呈黄绿色.json
白带带血.json
白带发黄.json
白带黑色.json
白带拉丝.json
白带奶白色.json
白带增少.json
白带异常.json
白带异味.json
白带增多.json
白带粘稠.json
白发.json
白痒.json
白癜风随形成.json
白甲.json
白色结痂.json
白色鳞屑状尿.json
白色念珠菌感染.json
白色绒毛.json
白色视野缺损.json
白色脓汁.json
白斑.json
白天不自觉地咬牙咬紧.json
白天嗜睡夜间失眠.json
白头粉刺.json
白细胞减少.json
白细胞尿.json
白细胞增多.json
白血病相酷使用.json
白癜.json
白血样真菌.json
百特异症状.json
柏油便.json
败血症.json
败血症的征象.json
扳机指的弹响.json
斑丘疹.json
斑秃.json
斑疹.json
板状腹.json
半侧面部发育不良.json
半年或1岁以后逐步出现智能减退.json
半乳糖血症.json
伴高血压.json
伴皮肤粘膜出血.json
伴有生动酣梦的睡眠性幻视.json
蝙蝠小叶增厚.json
蝙蝠翼厚.json
膀胱残余尿量增多.json
膀胱刺激征.json
膀胱功能障碍.json
膀胱结核.json
膀胱结石.json
膀胱挛缩.json
膀胱排空不全.json
膀胱癌肿.json
膀胱破章.json
膀胱容量的减少.json
膀胱纤维化.json
膀胱阴道瘘.json
包皮发巴.json
包皮过长.json
包皮红肿.json
包皮开裂.json
包皮溃疡.json
包皮内有白色分泌物.json
包皮上有水疱.json
包皮上有小白点.json
包皮上有小红点.json
包皮水肿.json
包皮痒疹.json
包皮脱皮.json
包皮垢.json
包皮粘连.json
暴食后酮嗅.json
宝宝缺乏安全感.json
保持睡眠障碍性失眠.json
抱膝.json
初级试剧症.json
暴发性痛.json
暴力冲动行为.json
暴力作用后遗痛.json
暴跌.json
暴晒后的皮肤损害.json
暴瘦.json
暴牙.json
暴饮暴食.json
暴躁的脑电动触.json
爆发性水泡.json
杯状耳.json
背部出现小水疱.json
背部驼挛.json
背部酸痛.json
背部利物引痛痛不适.json
背部长短.json
背脊痛.json
背镜.json
背心冷.json
背驼.json
被动体位.json
被害妄想.json
被迫体位.json
被恐惧缠绕妄想.json
贲门部心悸.json
贲门痉挛.json
贲门癌肉.json
苯丙氨酸代谢障碍.json
鼻腔肌不稳定.json
鼻腔肌反射亢进.json
鼻腔肌无反射.json
鼻部鸡核.json
鼻翼痒症.json
鼻出血.json
鼻唇沟
鼻唇沟
鼻唇沟
鼻窦气
鼻窦压
鼻缝出
鼻缝肥
鼻缝宽
鼻缝缺
鼻缝形
鼻分泌
鼻干ja
鼻干燥
鼻根部
鼻和颊
鼻甲肥
鼻甲萎
鼻甲肿
鼻尖或
鼻尖上
鼻孔闭
鼻孔扇
鼻孔干
鼻孔流
鼻孔外
鼻前庭
鼻梁低
鼻梁平
鼻软ja
鼻毛变
鼻毛脱
鼻内检
鼻内痛
鼻翼塌
鼻翼脂

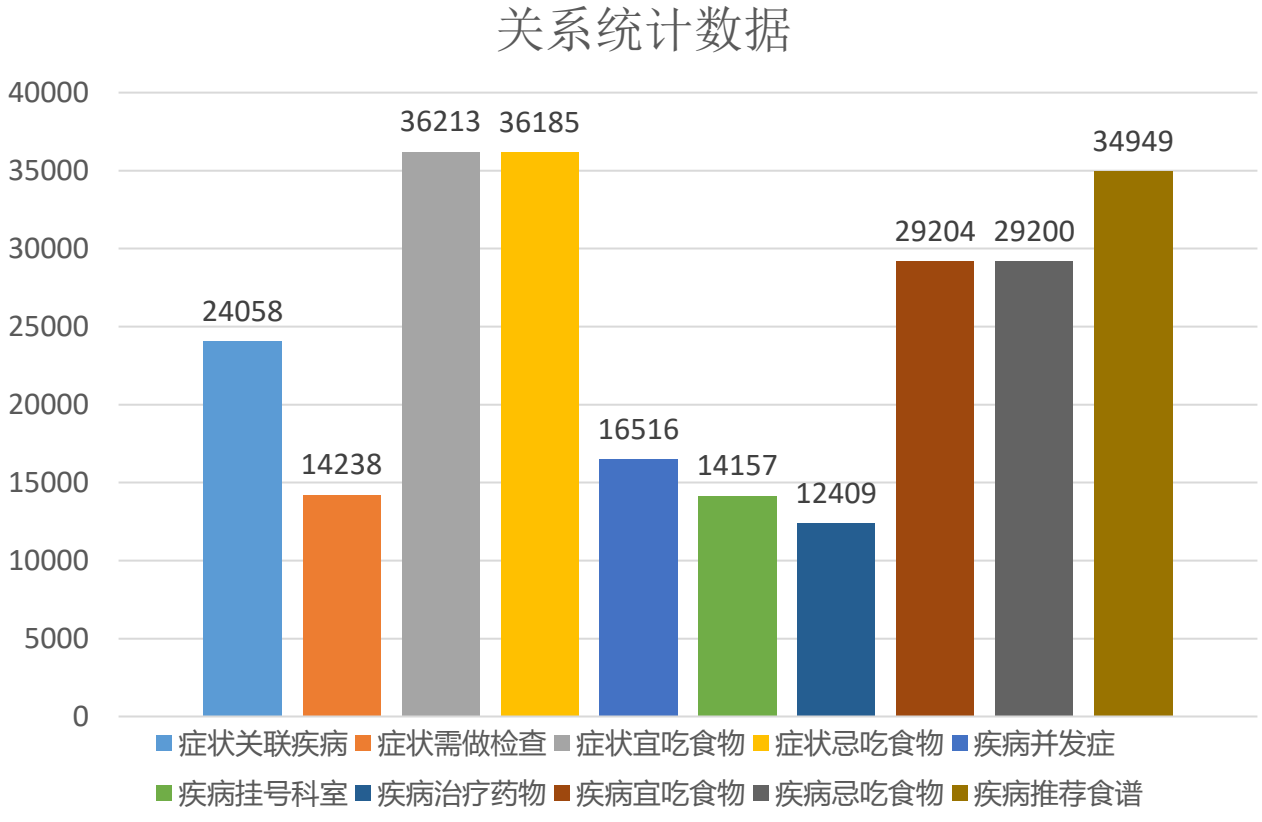
疾病数据


I型超敏反应性疾病.json
I型免疫母细胞性淋巴瘤病.json
Alagille综合征.json
Austin型幼儿脑病.json
A-V综合征.json
A组球菌咽感染.json
Bazin硬红斑.json
Brennemann综合征.json
B组球菌肺感染.json
B型球病毒感染.json
CD30阳性皮肤T细胞淋巴瘤.json
Chance骨折.json
CM1神经节苷脂贮积症.json
Creutzfeldt-Jakob病.json
c畸马功能性β球蛋白.json
Ehler-Danlos综合征.json
Friedreich共济失调症.json
Fuchs角膜内皮营养不良.json
Gardner综合征.json
Goodpasture综合征.json
Graves病.json
H7n7.json
H7N9型禽流感.json
HIV相关肝脏病感染.json
Hughes-Stavin综合征.json
II型高小管性酸中毒.json
II型糖尿病.json
I型高小管性酸中毒.json
I型糖尿病.json
Jessner淋巴细胞浸润病.json
Kaposi肉瘤.json
Kartagener综合征.json
Kimura病.json
Leber遗传性视神经病变.json
Machado-Joseph病.json
Maffucci综合征.json
Milk-Alkali综合征.json
Mirizzi综合征.json
Morton肥头病.json
Nezelof综合征.json
Pancoast综合征.json
Pasini-Pierni进行性皮肤萎缩.json
Pick病和脑额痴呆.json
Plummer-Vinson综合征.json
Poncett综合征.json
PUVA疗法相对症.json
Q热.json
Q热肺炎.json
Reis-Bucklers角膜营养不良.json
Reiter综合征.json
Rh虫型不合.json
Rh虫型不合溶血病.json
SARS型病毒感染.json
Stargardt病.json
Terrien角膜边缘性变性.json
Terson综合征.json
Theodore上方角膜缘角结膜炎.json
Thygeson浅层小伏角膜病变.json
Usher综合征.json
Whitfield硬红斑.json
X型结角膜病.json
X综合征.json
Zieve综合征.json
α1-抗胰蛋白酶缺乏性肝硬变.json
α1-抗胰蛋白酶缺乏症.json
α-贮存病.json
β-氨基酸尿.json
β受体亢进症.json
δ-贮存病.json
阿尔采末病.json
阿尔茨海默病性痴呆.json
阿洪病.json
阿克森尔迪-恩格尔综合征.json
阿米巴肺脓空孔.json
阿米巴脑病.json
阿米巴肝脓肿.json
阿米巴肝脓疮.json
阿米巴性肠炎.json
阿米巴性直肠炎.json
阿米巴性阴道炎.json
阿片类药物中毒.json
阿片中毒.json
阿斯伯格综合征.json
阿斯巨林不能型三联症.json
阿-维综合征.json
埃博拉病毒病.json
埃布斯坦综合征.json
埃及血吸虫病.json
埃莫尔-当洛综合征.json
埃利希体病.json
癌性疼痛.json
癌症.json
癌症伴发的精神障碍.json
矮小症.json
安卡性贫血.json
安眠药中毒.json
氨基甲酸酯类农药中毒.json
氨基酸代谢病.json
氨基酸病症.json
氨气中毒.json
鞍上生殖细胞瘤.json
暗色丝孢霉病.json
巴顿骨折.json
巴瑞特综合征.json
癌症.json
皮干症.json
白带异常.json
白癜风.json
白发.json
白发症.json
白喉.json
白喉性心肌炎.json
白化病.json
白细胞血.json
白眼.json
白毛虫.json
白塞病.json
白内障.json
白塞病.json
白塞综合征.json
白壳虫.json
白细胞减少症.json
白细胞葡萄糖-6-磷酸脱氢酶缺乏症.json
白线虫.json
白癣.json
白血病.json
白血病伴发的精神障碍.json
白血病神经症.json
白浊.json
百草枯中毒.json
白日梦.json
柏-查综合征.json
柏氏病.json
败血症.json
班关.json
斑块伤寒.json
斑状白化病.json
斑状萎缩.json
癌前病.json
癌症疼痛.json
癌性脑内酮.json
癌性关天疱疹.json
癌性脱发.json
板层状鱼鳞病.json
办公室综合征.json
半乳糖血症.json
半月板红.json
半椎体畸形.json
伴齿状突骨折并脊椎椎间盘脱位.json
伴纤维囊性骨炎的弥性甲状旁腺功能亢下症.json
伴有虹膜睫状体炎的慢性青光眼.json
伴有脑脊髓神经的脊髓炎.json
膀胱癌.json
膀胱白斑病.json
膀胱不发育与发育不全.json
膀胱过度活动症.json
膀胱病.json
膀胱结核.json
膀胱结石.json
膀胱痉挛症.json
膀胱癌状瘤.json
膀胱癌.json
膀胱尿道异物.json
膀胱外出.json
膀胱平滑肌瘤.json
膀胱平滑肌肉瘤.json
膀胱气肿.json
膀胱癌.json
膀胱结核.json
膀胱炎症反流.json
膀胱损伤.json
膀胱脱垂.json
膀胱外翻.json
膀胱阴癌.json
膀胱炎.json

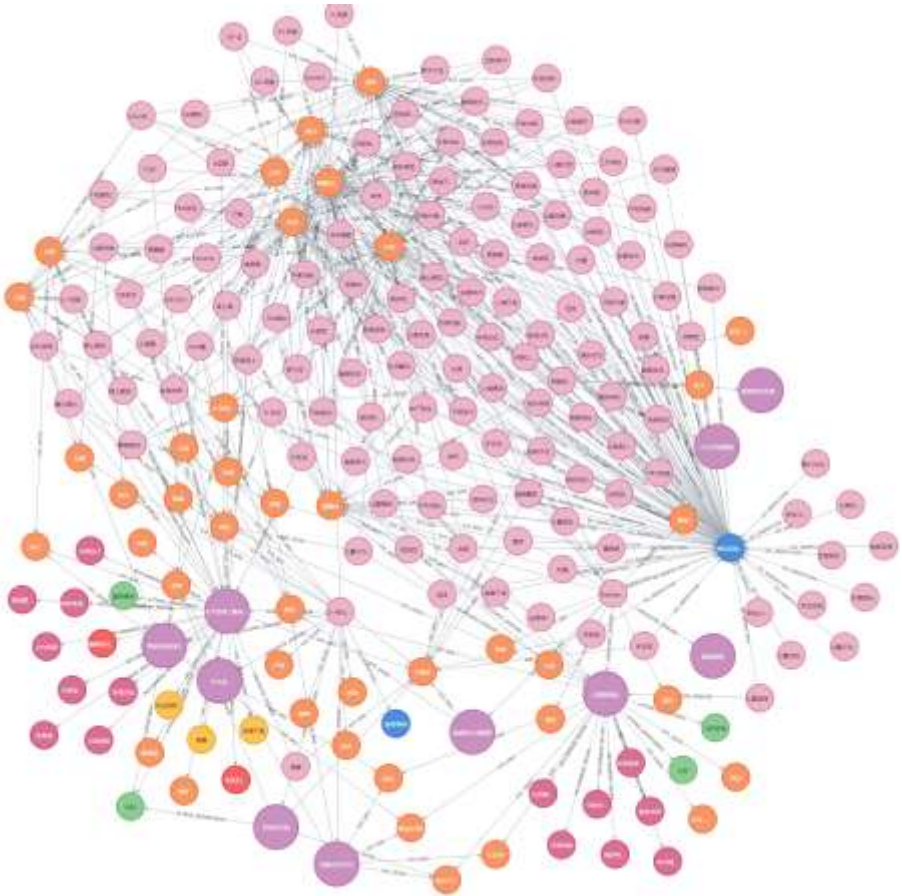
统计数据


实体统计数据
8000
7000
6000
5000
4000
3000
2000
1000
0
6855
7420
2105
2377
54
1462
726
4326
症状
疾病
医疗检查
并发症
科室部门
药物
食物
食谱

统计数据


关系统计数据
40000
35000
30000
25000
20000
15000
10000
5000
0
24058
14238
36213
36185
16516
14157
12409
29204
29200
34949
症状关联疾病
症状需做检查
症状宜吃食物
症状忌吃食物
疾病并发症
疾病挂号科室
疾病治疗药物
疾病宜吃食物
疾病忌吃食物
疾病推荐食谱

知识图谱



技术原理

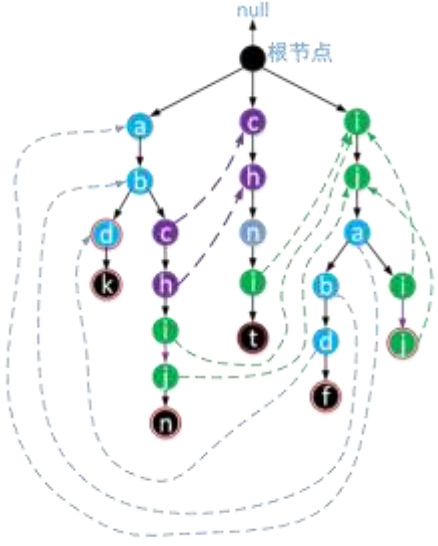
命名实体识别是自然语言处理的一项重要任务，目的是在文本中找到具有特定意义的实体，例如人名、地名、组织机构名等，本项目中对应的则是找出问句中的症状和疾病实体。具体来说，使用AC自动机和BiLSTM-CRF进行命名实体识别，有效提高了识别准确度和效率。

AC自动机

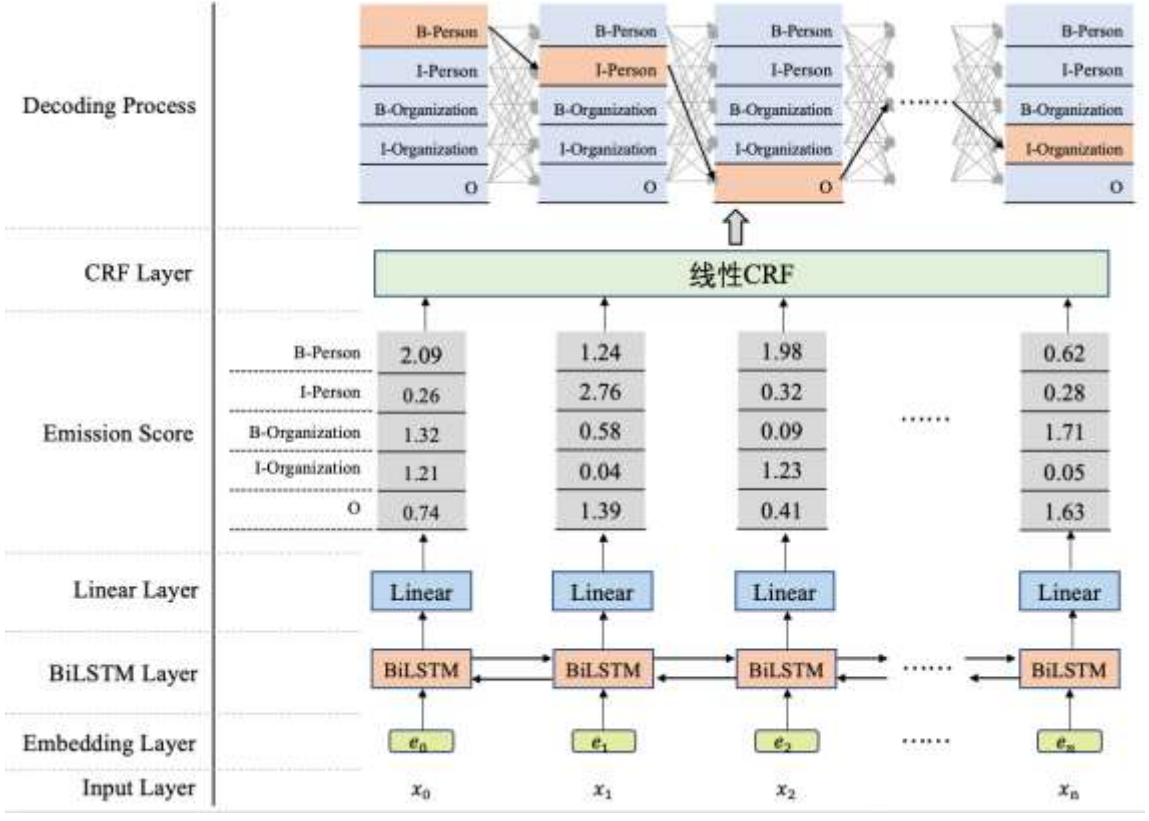
AC自动机可以用于模式匹配和字符串查找等问题，其核心思想是构建一个字典树，然后将字典树上的每个节点和某一模式串的前缀对应，构成一个状态机，即AC自动机。

从AC自动机的根节点开始搜索，在搜索的过程中，如果当前节点存在失配指针，则跳转到失配指针对应的节点；如果当前节点是一个匹配成功节点，则将其所对应的模式串加入到匹配结果中。

AC自动机的时间复杂度是O(n)，其中n为目标字符串的长度。



BiLSTM-CRF


Decoding Process
B-Person
I-Person
B-Organization
I-Organization
O
CRF Layer
线性CRF
Emission Score
B-Person
I-Person
B-Organization
I-Organization
O
2.09
0.26
1.32
1.21
0.74
1.24
2.76
0.58
0.04
1.39
1.98
0.32
0.09
1.23
0.41
0.62
0.28
1.71
0.05
1.63
Linear Layer
Linear
BiLSTM Layer
BiLSTM
Embedding Layer
Input Layer

BiLSTM-CRF

BiLSTM-CRF是一种序列标注模型，其结构包括BiLSTM和CRF两部分。BiLSTM是一种双向长短时记忆网络，可以对输入的序列进行有效的编码和提取特征。CRF是一种条件随机场，可以对标注序列进行联合概率建模，从而提高序列标注的准确性。

在BiLSTM-CRF中，首先使用BiLSTM对输入序列进行编码，得到每个时刻的隐状态表示。然后将隐状态表示传递给CRF层，进行联合概率建模并得到最优的标注序列。最终输出的标注序列即为模型预测的结果。

BiLSTM-CRF

相比于传统的序列标注模型，BiLSTM-CRF具有以下优点：

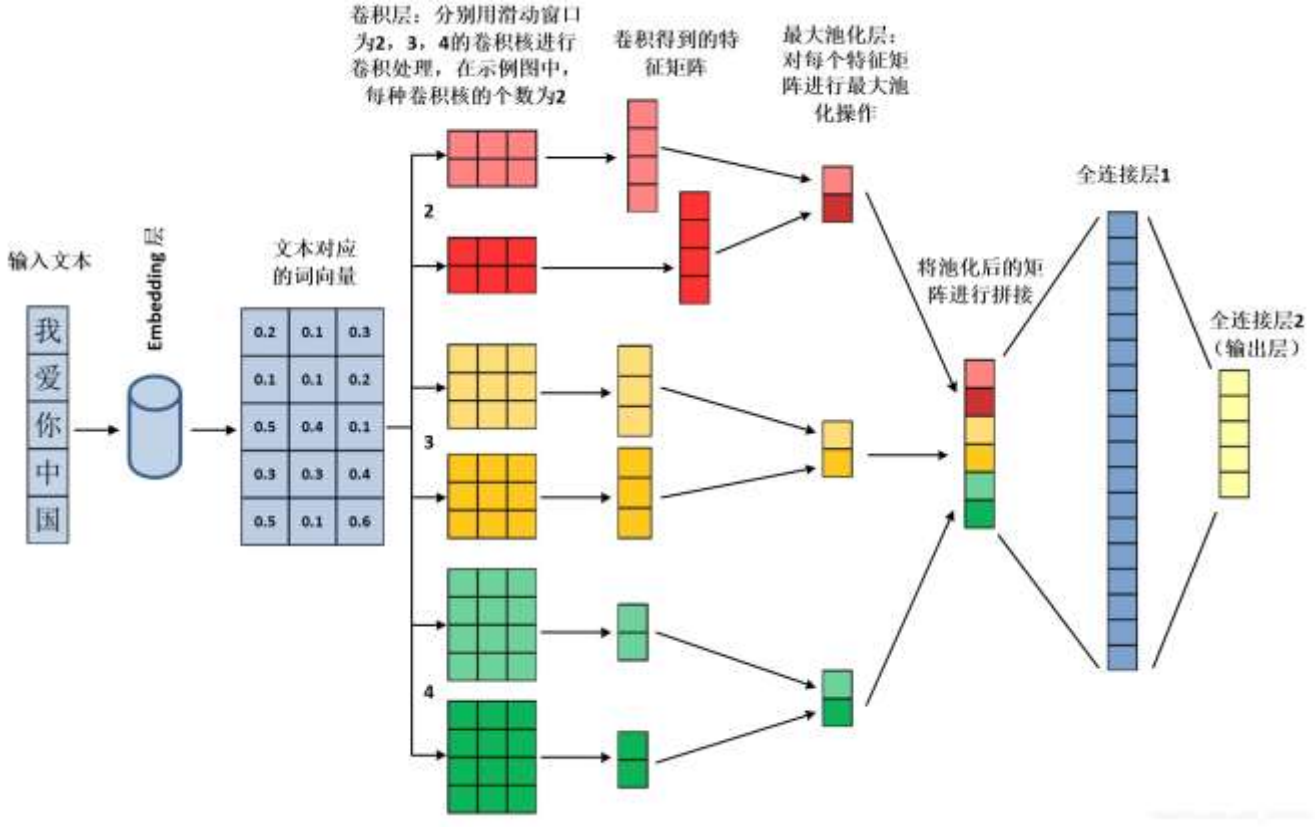
1. BiLSTM能够有效地捕捉输入序列中的上下文关系，从而提高特征表示的能力。
2. CRF层能够考虑相邻标签之间的依赖关系，并且利用全局约束来解决标签依赖问题，使得标注结果更加合理和一致。
3. BiLSTM-CRF可以进行端到端的训练和预测，避免了传统模型中需要进行特征工程和后处理的麻烦。

BiLSTM-CRF已经在自然语言处理领域的诸多任务中取得了良好的表现，例如命名实体识别、句法分析、情感分析等。

技术原理

意图识别是指要确定对问题的回答属于特定类别，例如查询疾病、药品、治疗方法等。本项目采用了TextCNN模型作为意图识别的核心算法，结合NER层人工嵌入特征后的输入数据，实现快速、准确的意图分类。

TextCNN


卷积层：分别用滑动窗口为2，3，4的卷积核进行卷积处理，在示例图中，每种卷积核的个数为2
卷积得到的特征矩阵
最大池化层：对每个特征矩阵进行最大池化操作
全连接层1
全连接层2（输出层）
输入文本
Embedding层
文本对应的词向量
我
爱
你
中
国
0.2 0.1 0.3
0.1 0.1 0.2
0.5 0.4 0.1
0.3 0.3 0.4
0.5 0.1 0.6
2
3
4
将池化后的矩阵进行拼接

TextCNN

TextCNN（Text Convolutional Neural Network）是一种卷积神经网络（CNN）的变种，主要用于自然语言处理（NLP）任务，例如文本分类、情感分析和命名实体识别。

TextCNN的核心思想是将文本转化为一个矩阵，并使用卷积层进行特征提取，然后将提取出的特征传递给全连接层进行分类。

TextCNN

TextCNN的优点包括：

1. 高效的文本分类模型，可以在短时间内处理大量的文本数据。
2. 利用卷积神经网络的局部关系特性，能够更好地提取文本中的特征，无需人工提取特征。
3. 采用多个不同大小的卷积核对数据进行卷积和池化操作，可以对不同大小的语义信息进行提取，从而更加全面的理解文本内容。

TextCNN的缺点包括：

1. 在处理较长文本时，可能会出现信息丢失或混淆的问题。
2. 最大的问题全局max pooling丢失了结构信息，很难发现文本中的转折关系等复杂模式。
3. 只知道关键词是否在文本中出现了，以及相似度强度分布，不可能知道关键词出现了几次，以及出现这些关键词的顺序。

数据集

数据冷启动指的是当一个系统或模型开始运行时，缺乏足够的已有数据以支撑其学习和运作，从而导致性能不佳、准确率低下的情况。

在机器学习领域中，数据冷启动是指在训练模型之前，无足够的已有数据集来进行模型训练的情况，例如：

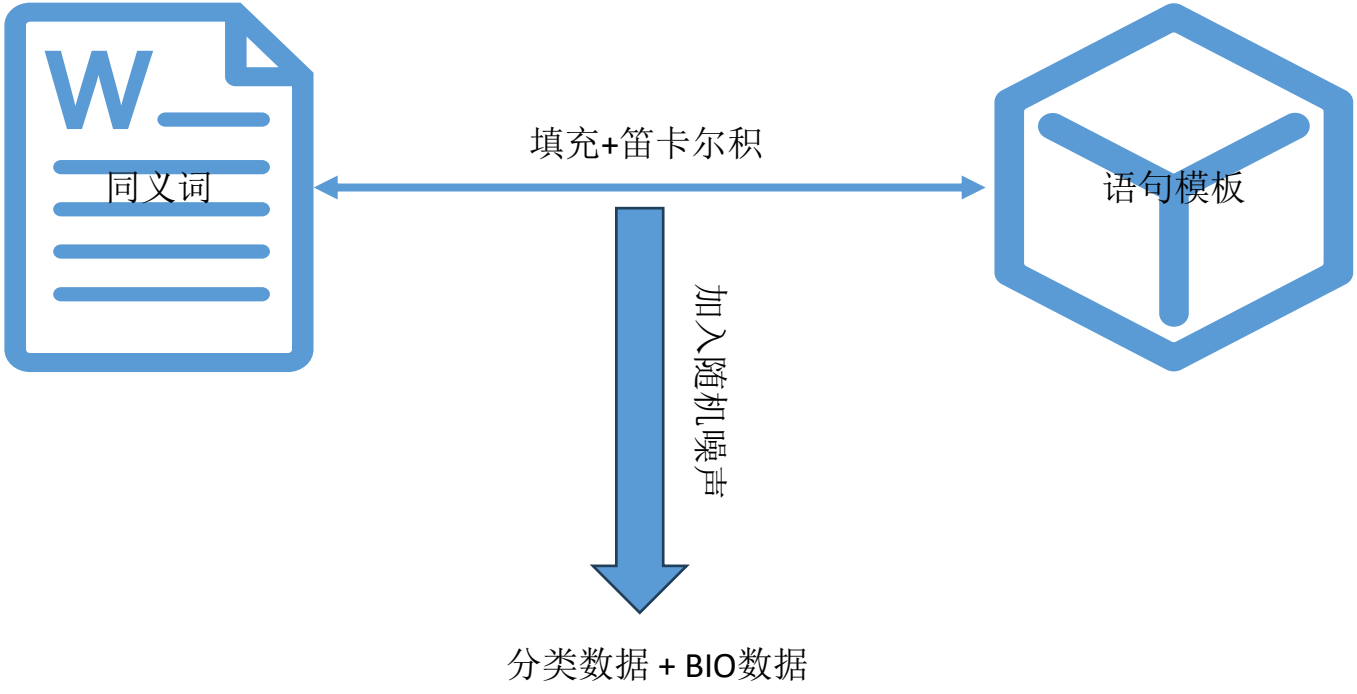
数据集不足：当某一特定领域的数据集非常珍贵，如医疗、金融等领域，可能由于数据保密、获取难度大等原因，导致缺乏足够的数据进行模型训练。

新领域探索：当研究领域刚开始探索一种新的现象时，缺乏足够的数据集来支持训练模型。

数据集

对于数据冷启动问题，常用的解决方法包括利用模拟生成数据、利用迁移学习、利用半监督学习等方法来缓解数据缺失的问题，以提高模型训练和性能。

数据集



数据集


```
# 同义词, 用于构造语句数据
WORDS_SEQ = {
    'how': ['怎么样', '如何', '何以', '怎么', '怎么', '怎么', '怎样', '怎么', '怎样', '咋样', '咋样', '咋样', '咋地', '咋地', '咋地'],
    'why': ['为什么', '为何', '何故', '何以', '为啥', '何以故'],
    'what': ['什么', '何', '甚', '哪', '哪些', '那', '有何', '何事', '何物', '什么事', '什么物'],
    'do': ['办', '办', '办', '办', '办', '做', '做', '做', '做', '做', '处理', '解决', '操作'],
    'is': ['是', '在', '属于', '在于', '代表', '表示', '指的是', '意味着', '包括', '归属于'],
    'may': ['可能', '或许', '可能会', '恐相', '大概', '也许', '有可能', '很可能'],
    'should': ['应该', '必须', '需要', '要', '希望', '愿意', '可以', '理应', '当然', '有必要'],
    'feel': ['感受', '觉得', '意识到', '认为', '知觉', '发觉', '感应', '体验', '体识', '感觉', '察觉', '出现', '显现'],
    'cure': ['治疗', '医治', '治愈', '康复', '痊愈', '治疗', '止疼', '治好', '缓解', '减轻'],
    'illness': ['疾病', '病症', '病患', '疾患', '病病', '疾疾', '流行病', '病毒性疾病', '传染病'],
    'suffer': ['得', '患', '感染', '患染', '发病', '罹患', '患有'],
    'explain': ['介绍', '了解', '告知', '理解', '解释', '详情', '信息', '定义'],
    'prevent': ['预防', '防范', '抵制', '抵御', '防止', '避免', '免得', '避开', '免于'],
    'check': ['检查', '检测', '体检', '化验', '验证', '查验', '查询', '查', '测', '检', '监测'],
    'take': ['吃', '食', '服', '摄入', '饮', '喝', '食用', '服用', '饮食', '饮用', '饮食', '膳食', '忌口', '补品',
             '食谱', '菜谱', '食物', '补品'],
    'good': ['好', '有益', '改善', '提升', '增加', '增进', '增强', '益处', '好处', '适合', '适宜', '宜', '宜于'],
    'bad': ['不好', '不益', '差', '降低', '减少', '不良', '糟糕', '弊处', '恶劣', '有害', '不宜', '忌'],
    'suggest': ['建议', '提示', '忠告', '注意', '推荐', '帮助', '指南', '提醒']
}
```

数据集


```
# 问句模板，用于构造训练数据
TEMPLATES_SEQ = {
    '症状问办法': [
        ('entity#symptom', 'how', 'do'),
        ('how', 'do', 'entity#symptom'),
        ('entity#symptom', 'how', 'cure'),
        ('how', 'cure', 'entity#symptom'),
        ('how', 'cure', 'entity#symptom'),
        ('entity#symptom', 'what', 'do'),
        ('what', 'do', 'entity#symptom'),
        ('entity#symptom', 'what', 'cure'),
        ('what', 'cure', 'entity#symptom'),
        ('what', 'cure', 'entity#symptom'),
    ],
    '症状问疾病': [
        ('entity#symptom', 'suffer', 'what', 'illness'),
        ('suffer', 'what', 'illness', 'entity#symptom'),
        ('why', 'feel', 'entity#symptom'),
        ('feel', 'entity#symptom', 'why'),
    ],
    '症状问预防': [
        ('how', 'prevent', 'entity#symptom'),
        ('entity#symptom', 'how', 'prevent'),
        ('what', 'prevent', 'entity#symptom'),
        ('prevent', 'entity#symptom', 'what'),
        ('do', 'what', 'prevent', 'entity#symptom'),
        ('prevent', 'entity#symptom', 'do', 'what'),
    ],
```

数据集

意图分类数据

训练集：12831条

测试集：1431条

```
text,label
呀腱反射消失symptom什么事办,症状问办法
潮汗symptom那止痊,症状问办法
怎么医治先兆流产symptom4,症状问办法
免疫性肾炎symptom怎么办,症状问办法
何p治瘪饭后困倦symptom,症状问办法
晶状体青光眼斑symptom咋地办,症状问办法
肢体运动不协调symptom5咦咋地做,症状问办法
咀嚼肌痉挛symptom哪些止痊,症状问办法
什么事治好腹泻symptom,症状问办法
皮肤色素沉着呈青铜色symptom如何！做,症状问办法
1咋样办运动耐力降低symptomj,症状问办法
肝淀粉样变性symptom什么事缓解,症状问办法
哈g那缓解前臂及手部肌群的缺血性挛缩symptom,症状问办法
哈甚1治疗脾胃不和symptom,症状问办法
```

数据集

命名实体识别数据

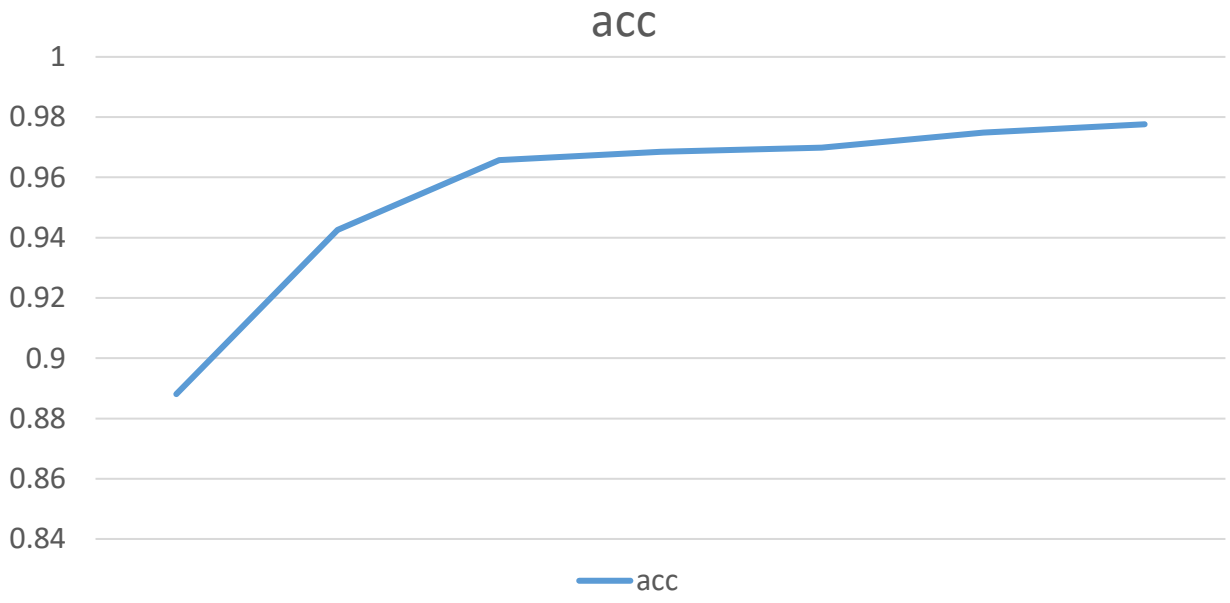
训练集：12848条

测试集：1429条

```
p哪验证生化妊娠,O O O O B-SYMPTOM I-SYMPTOM I-SYMPTOM I-SYMPTOM
何以康复黄色结节,O O O O B-SYMPTOM I-SYMPTOM I-SYMPTOM I-SYMPTOM
粘膜疹咋地医治,B-SYMPTOM I-SYMPTOM I-SYMPTOM O O O O
提示咋地操作癫痫性精神障碍,O O O O O O B-DISEASE I-DISEASE I-DISEASE I-DISEASE I-DISE
啊管型尿咋地检查,O B-SYMPTOM I-SYMPTOM I-SYMPTOM O O O O
甚不好婴儿青铜综合征,O O O B-DISEASE I-DISEASE I-DISEASE I-DISEASE I-DISEASE I-DISEA
下腹剧痛并渐向腹中线扩散咋样医治,B-SYMPTOM I-SYMPTOM I-SYMPTOM I-SYMPTOM I-SYMPTOM I-S
哪处理盘状红斑狼疮v、,O O O B-DISEASE I-DISEASE I-DISEASE I-DISEASE I-DISEASE I-DISE
骶骼筋膜脂肪疝注意何,B-DISEASE I-DISEASE I-DISEASE I-DISEASE I-DISEASE I-DISEASE I-D
那表示膀胱损伤,O O O B-DISEASE I-DISEASE I-DISEASE I-DISEASE
咋地治愈脊椎病1,O O O O O B-DISEASE I-DISEASE I-DISEASE O
w葡萄球菌感染有害啊哪,O B-DISEASE I-DISEASE I-DISEASE I-DISEASE I-DISEASE I-DISEASE
哪降低皮样表皮样囊肿,O O O B-DISEASE I-DISEASE I-DISEASE I-DISEASE I-DISEASE I-DISEA
有何恶劣小儿乳积,O O O O B-DISEASE I-DISEASE I-DISEASE I-DISEASE
梅尼埃病mr信息,B-DISEASE I-DISEASE I-DISEASE I-DISEASE O O O O
```

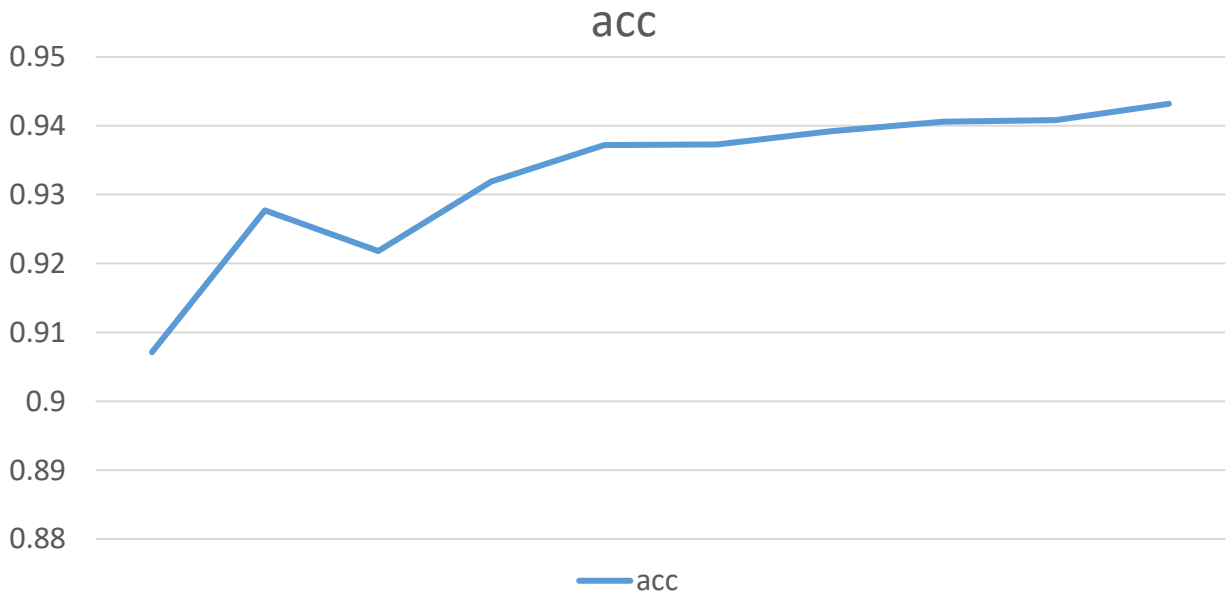
模型评估

意图识别模型：7个epoch后acc达到97.76%



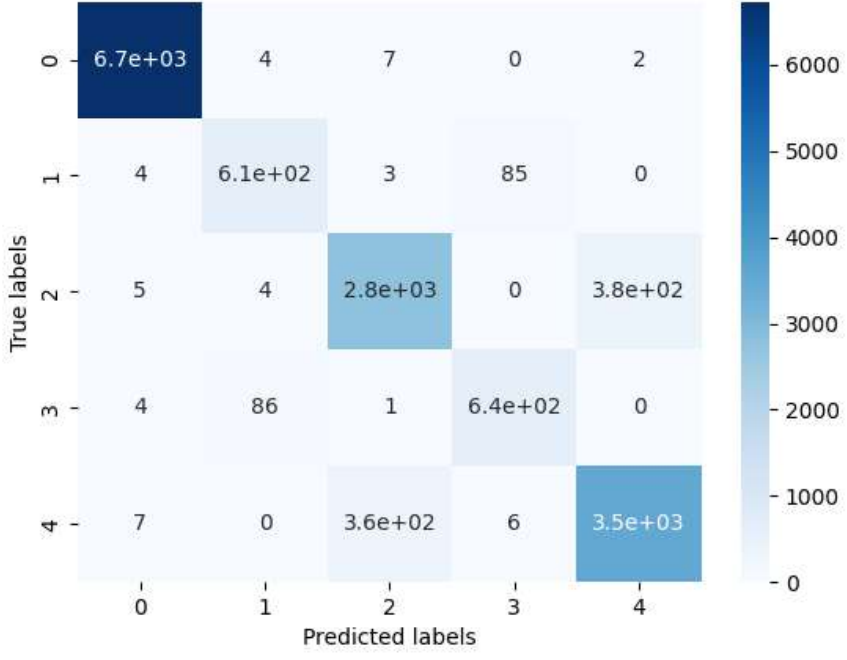
模型评估

命名实体识别模型：10个epoch后，标签整体acc达到97.76%



模型评估

命名实体识别模型混淆矩阵



优化与总结

TextCNN只知道关键词是否在文本中出现了，以及相似度强度分布，不可能知道关键词出现了几次，以及出现这些关键词的顺序。

针对这个问题，可以尝试k-max pooling优化，针对每个卷积核都不只保留最大值，而是保留前k个最大值，并且保留这些值出现的顺序，即按文本的顺序排列k个最大值。

优化与总结

可以结合Bert优化BiLSTM-CRF模型的表现。

Bert是一种预训练模型，在自然语言处理领域已经产生了广泛的应用，包括命名实体识别、句法分析、情感分析等。Bert的优点在于其在大规模语料库上训练得到了丰富的语言表示，这些表示可以应用在不同的自然语言处理任务上。

使用Bert对输入词汇进行编码，用Bert提取的语义特征替换Embedding层的词向量，以提高模型的语义表示能力。

优化与总结

目前通过NER输出的实体信息，是通过计算Jaccard相似度，映射到知识图谱中的实体，这种方法没有考虑两个词之间的语义相似度，单纯的依赖字符交并集的做法无法保证较好的映射效果。

可用通过预训练的Word2Vec结合本地知识图谱的数据，进行调优，使用调优后的模型能更好的计算两个词之间的语义相似度。