基于传染病医疗知识图谱的知识问答项目介绍

> 本项目由@陈屹峰20373860完成，代码量约为700行，工作量约两周。有问题请联系20373860@buaa.edu.cn

项目简介

本项目实现了基于`fasttext`和基于`jieba`的的智能问答系统，通过对于用户问题进行分类确定问题的询问对象，对问题进行分词抽取问题中包含的关键词，进而抽取出实体，再将两者结合进入neo4j搭建的知识图谱中进行查询，返回用户需要的答案。

项目原理

项目主要的技术点在fastext分类和jieba分词。这里简单介绍一下。

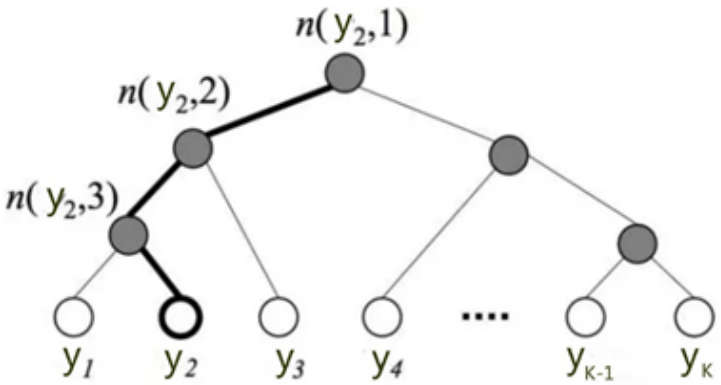
fasttext原理

fastText的核心思想就是：将整篇文档的词及n-gram向量叠加平均得到文档向量，然后使用文档向量做softmax多分类。这中间涉及到两个技巧：字符级n-gram特征的引入以及分层Softmax分类。

n-gram文本特征提取的主要思想是按照一定大小的滑窗按顺序对文本的内容进行分割，形成一系列的字节片段序列，作为文本特征的候选集。例如：

我来到达观数据参观 - 相应的bigram特征为：我来 来到 到达 达观 观数 数据 据参 参观

分层softmax的主要思想是利用书的层级结构代替扁平化的标准softmax是的在计算某个概率时只需要访问从树的根节点到叶子结点的路径即可。



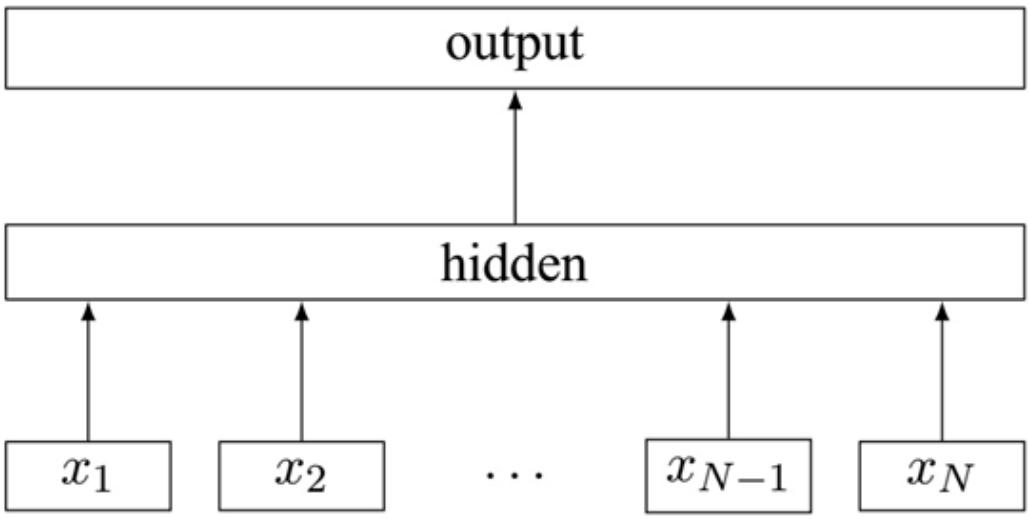


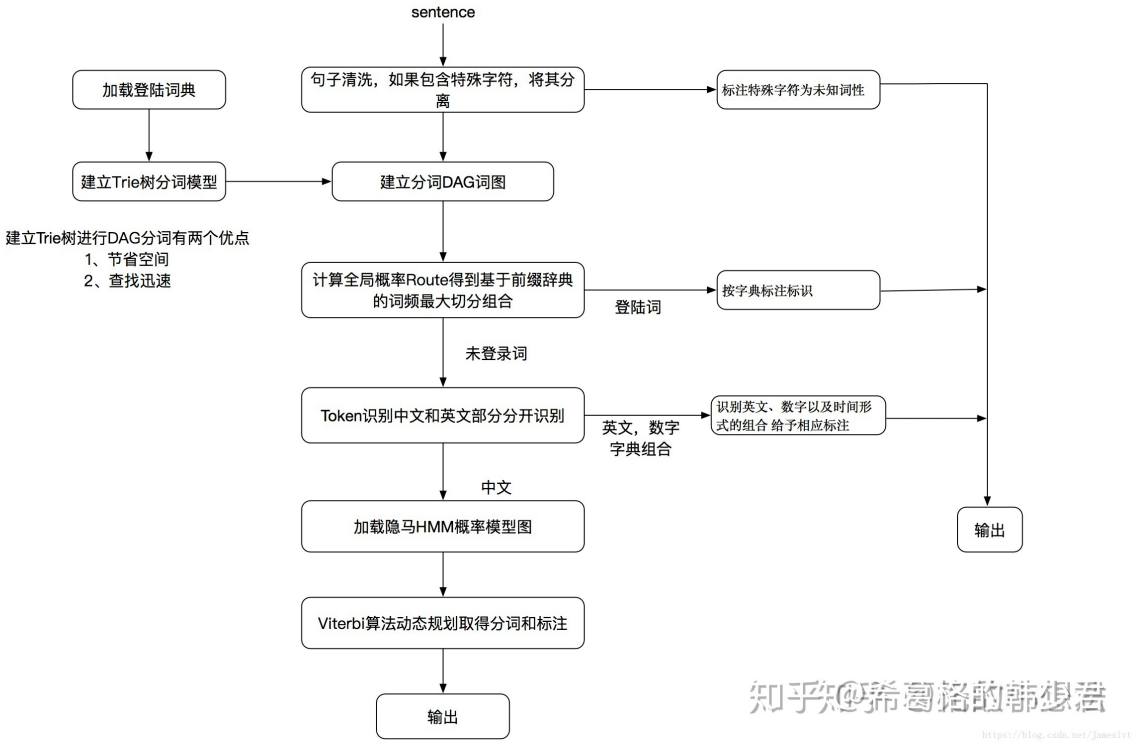
Figure 1: Model architecture of `fastText` for a sentence with N ngram features $x_1, \ldots, x_N$. The features are embedded and averaged to form the hidden variable.

fastText模型只有三层：输入层、隐含层、输出层。输入是多个单词及其n-gram特征，这些特征用于表示整个文档；隐含层是对多个词向量的叠加平均，而输出层则采用了分层softmax，对隐含层的输出做一个多分类。

jieba原理

分词的步骤

1. 根据jieba自带dict.txt词典生成trie树，便于前缀查询
2. 给定待分词的句子，使用政策获取连续的中文字符，切分成短语列表，对每个短语列表，基于前缀词典构建句子中汉字所有可能成词的情况所构成的有向无环图，再使用动态规划得到基于词频的最大概率路径（也就是最大切分组合）
3. 对于在有向无环图中那些没有在字典中查询到的字，组合成一个新的片段短语，使用隐马尔科夫模型进行分词，识别新词



具体实现

1、基于前缀词典实现高效的词图扫描

- 结巴分词**自带了一个叫做dict.txt的词典**，里面有349046条词，其每行包含了词条、词条出现的次数（这个次数是结巴作者自己基于人民日报语料等资源训练得出来的）和词性。
- 把这34万多条词语，放到一个trie树的数据结构中，也就是说一个词语的前面几个字一样，就表示他们具有相同的前缀，就可以使用trie树来存储，具有查找速度快的优势。
- 查询的过程对于查询词来说，从前往后一个字符一个字符的匹配。**对于trie树来说，是从根节点往下匹配的过程。**

2、找出基于词频的最大切分组合

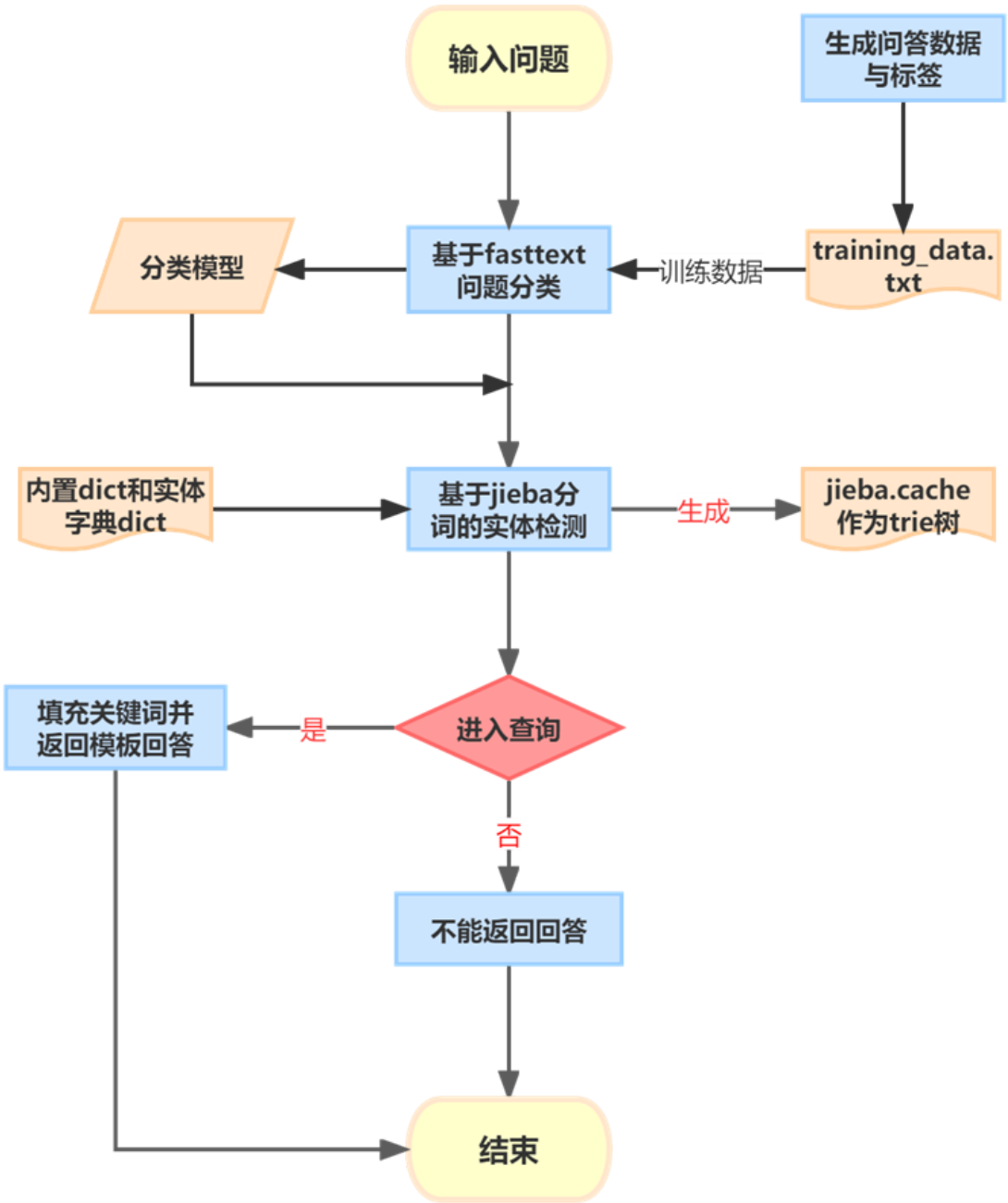
- 生成句子中汉字所有可能成词情况所构成的**有向无环图**，给定一个待分词的句子，将它的所有词匹配出来，构成词图，即是一个有向无环图DAG。
- 作者的代码中将字典在生成trie树的同时，也把每个词的出现次数转换为了频率。频率其实也是一个0~1之间的小数，是事件出现的次数/实验中的总次数。
- 动态规划中，先查找待分词句子中已经切分好的词语，对该词语查找该词语出现的频率(次数/总数)，**如果没有该词**(既然是基于词典查找，应该是有可能没有该词)，就**把词典中出现频率最小的那个词语的频率作为该词的频率。**

3、基于汉字成词能力的 HMM 模型

- 一些人可能会想到**把dict.txt中所有的词汇全部删掉**，然后再试试结巴能不能分词。结果会发现，结巴依然能够分词，不过分出来的词，大部分的长度为2。
- 作者采用了HMM模型，中文词汇按照BEMS四个状态来标记，B是开始begin位置，E是是结束end位置，M是中间middle位置，S是single，单独成词的位置。也就是说，他采用了状态为(B,E,M,S)这四种状态来标记中文词语。
- **经过作者对大量语料的训练**，得到了finalseg目录下的三个文件
  位置转换概率，即B(开头)，M(中间)，E(结尾)，S(独立成词) 四种状态的转移概率，该表存放于prob_trans.py中
  位置到单字的发射概率
  词语以某种状态开头的概率，其实只有两种，要么是B，要么是S

- **将一个给定的待分词的句子视为一个观察序列**，对HMM(BEMS)四种状态的模型来说，就是为了找到一个最佳的BEMS隐状态序列，这个就需要使用Viterbi算法来得到这个最佳的隐藏状态序列。通过提前训练好的HMM转移概率、发射概率，使用基于动态规划的viterbi算法的方法，就可以找到一个使概率最大的BEMS序列。

项目流程



文件列表

- 主程序：chatbot_graph_test.py
- 文本分类训练数据生成程序：data_producer_2.py
- 文本分类程序：question_classifier_test.py
- 实体检测并生成查询语句程序：entity_detection_test.py
- 查询程序：answer_search.py
- 实体字典（从爬取的数据中选区的实体名称表）：dict_for_trans_diseases.txt
- 文本分类训练数据（由data_producer生成）：training_data_for_trans_diseases.txt

- 文本分类模型（由question_classifier生成）：data_dim200_lr00.5_iter1000.model

数据流

在主程序中

```
def chat_main(self, sent):
    res_classify = self.classifier.main(sent)
    res_sql = self.detector.entity_dectect(res_classify, sent)
    final_answers = self.searcher.search_main(res_sql)
```

第一行调用分类程序，对于用户输入的问题进行分类，第二行调用实体检测程序，对输入问题进行分词并查找实体词汇，根据问题类别和实体生成查询语句，第三行调用查询程序，在知识图谱中查询相关实体，根据问题类别调用相应的回复模板，并返回最终答案。

在文本分类训练数据生成程序中，加载实体特征词（从实体字典中）和问句指示词（自定义），将他们组合在一起生成类似于人类提问的训练数据，并将每类问题打上标签，生成训练文件training_data_for_trans_diseases.txt。问句指示词如下：

```
 self.symptom_qwds = ['什么症状', '什么症候', '有什么表现', '是怎样的', '是什么样的',
'哪里不舒服', '哪里难受', '反应如何', '什么不适', '会怎样']
```

```
__label__symptom_disease , 前列腺肥大 可能 是 什么 病
```

在文本分类程序中，调用fasttext包，在初始化阶段，使用前述训练文件，训练得到分类模型classifier。后续数据到来的时候直接使用模型进行分类。

```
  self.classifier = self.train_model(ipt=train_file,
                                     opt=model,
                                     model=model,
                                     dim=dim, epoch=epoch, lr=0.5
                                     )
```

在实体检测程序中，遍历jieba分词后的句子，对照前面生成的各个类别的实体字典，查找相应的实体，并根据问题类型和实体生成查询语句：

```
if question_type == 'disease_cause':
    sql = ["MATCH (m:Disease) where m.name = '{0}' return m.name,
m.cause".format(i) for i in entities]
```

最后在查询程序中，进行查询，并返回相应的回复模板：

```
# 查询
question_type = sql_['question_type']
queries = sql_['sql']
ress = self.g.run(query).data()
# 回复模板
if question_type == 'disease_symptom':
    desc = [i['n.name'] for i in answers]
    subject = answers[0]['m.name']
    final_answer = '{0}的症状包括：{1}'.format(subject, '；'.join(list(set(desc))
[:self.num_limit]))
```

项目运行

如果是首次运行：

运行 `data_producer_2.py`，生成训练样本文件 `training_data_for_trans_diseases.txt`，再运行 `chatbot_graph_test.py`，自动训练问答系统，开始问答。

否则：

直接运行 `chatbot_graph_test.py`，使用已经训练好的问答系统，开始问答。

问答系统运行实例：

```
Building prefix dict from the default dictionary ...
Loading model from cache C:\Users\10140\AppData\Local\Temp\jieba.cache
模型已训练完成!
Loading model cost 0.555 seconds.
Prefix dict has been built successfully.
医疗字典已录入!
用户:得了登革出血热有什么症状
disease_symptom
小勇： 登革出血热的症状包括：肝大；昏迷；昏睡；发烧；休克；咯血；淋巴结肿大；医师；出血倾向
用户:患上痢疾应该吃什么药
disease_drug
小勇： 痢疾通常的使用的药品包括：盐酸左氧氟沙星片；盐酸左氧氟沙星胶囊；穿心莲内酯片；诺氟沙星片；乳酸左氧氟沙星片
用户:应该去哪个科室检查痢疾
disease_check
小勇： 痢疾通常可以通过以下方式检查出来：纤维结肠镜检查；胸部B超；粪便脓液；钼靶X线检查；小肠镜检查；痢疾杆菌检测；粪便显微镜检查
用户:患了血吸虫病推荐什么食谱
disease_do_food
小勇： 血吸虫病宜食的食物包括有：南瓜子仁;腰果;芝麻;松子仁
推荐食谱包括有：豆浆南瓜汤;扁豆糕;白扁豆参米粥;薏米扁豆老黄瓜汤;苋菜豆腐汤;豆腐苋菜羹;绿豆杂面条;玉米粉燕麦粥
```