2018年3月28日 星期三

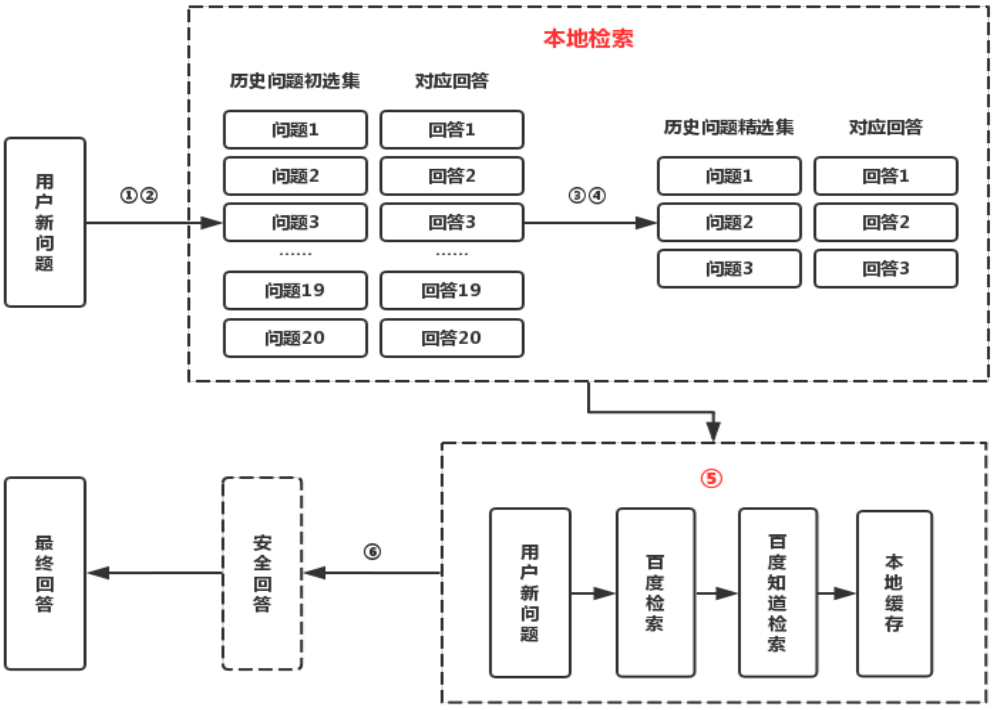
基于社区问答的单轮检索式通用知识问答工程实现

设计思路

目前，自动问答技术的实现主要有阅读式、检索式和生成式三大类。其中，阅读式是指机器阅读[1]：通过给定文章和针对文章提出的具体问答对训练机器的阅读理解能力，并对针对文章的新问题作出回答。阅读式的问题是知识边界狭窄，训练语料构筑困难，较适合于专业领域的机器人。与之相反的是生成式：生成式理论上可实现不受知识局限的通用机器人。但目前生成式最大的问题是答非所问、胡言乱语，无法表现出人类在解决问题时的逻辑自洽、思维发散和自然表达等能力。

而检索式则是一种取长补短的中间技术。检索式要求搜集众多在线问答网站海量的历史问答对语料作为知识库，通过检索与新问题最接近的历史相似问题和最匹配这一新问题的历史最佳问答，变相完成自动问答任务。检索技术既避免了知识的局限，又利用了人类的先验经验积累，是当前处理通用问答任务较合适的方法之一[2]。

结合实际，我的设计为：通过爬虫和用户历史数据建立QA式本地知识库，当给定用户新问题时，系统先通过ES[3]进行本地检索得到参考问答初选集，其次使用语义深度匹配和问答质量评估从初选集中精选出参考问答精选集，按得分降序返回最终参考问答；若没有任何问答通过精选，则启动在线搜索获取标识有“最佳问答”的参考问答返回给用户，并将其中的优质问答缓存到本地ES索引中。



[1] SQuAD和Dureader；方式有passage-based, assertion-based, sentence-based, answer-span-based等不同粒度

[2] 用学术的话来说就是：融合了本地知识与开放知识的一种实现思路

[3] ElasticSearch(ES)：ElasticSearch是一个基于Lucene的搜索服务器。它能够提供分布式、多用户的企业级全文搜索能力。原理：http://blog.csdn.net/cyony/article/details/65437708?locationNum=9&fps=1

2018年3月28日 星期三

从技术实现角度，本系统包括本地知识库、本地ES初选模块、问答质量评价模块、基于语义匹配的精选模块、在线搜索及缓存模块、异常处理机制等部分。

①本地知识库：百万条历史问答对数据经格式清洗、涉黄涉政检查、重复问答合并等预处理，构建完成“一问一答+点赞+点踩”格式的结构化本地知识库。

②本地ES初选模块：基于本地知识库搭建了ES索引(词义级初选)。当给定新问题时，可实现对本地百万级数据的毫秒级响应，返回结果为初选相似问题及对应答案。

③问答质量评价模块：对于给定新问题，给出候选问题对应答案与该新问题组合形成新问答对时该问答对的质量得分。本模块详述见下文。

④基于语义匹配的精选模块：训练基于N-gram和CNN的语义匹配模型、调用问答质量评价模块实现基于语义的精选。最后组合获得不超过3条的精选问答对及其得分。本模块详述见下文。

⑤在线搜索及缓存模块：此模块负责在线搜索用户新问题的回答，并将从百度知道源获取的优质问答对增量缓存到本地知识库中。本模块详述见下文。

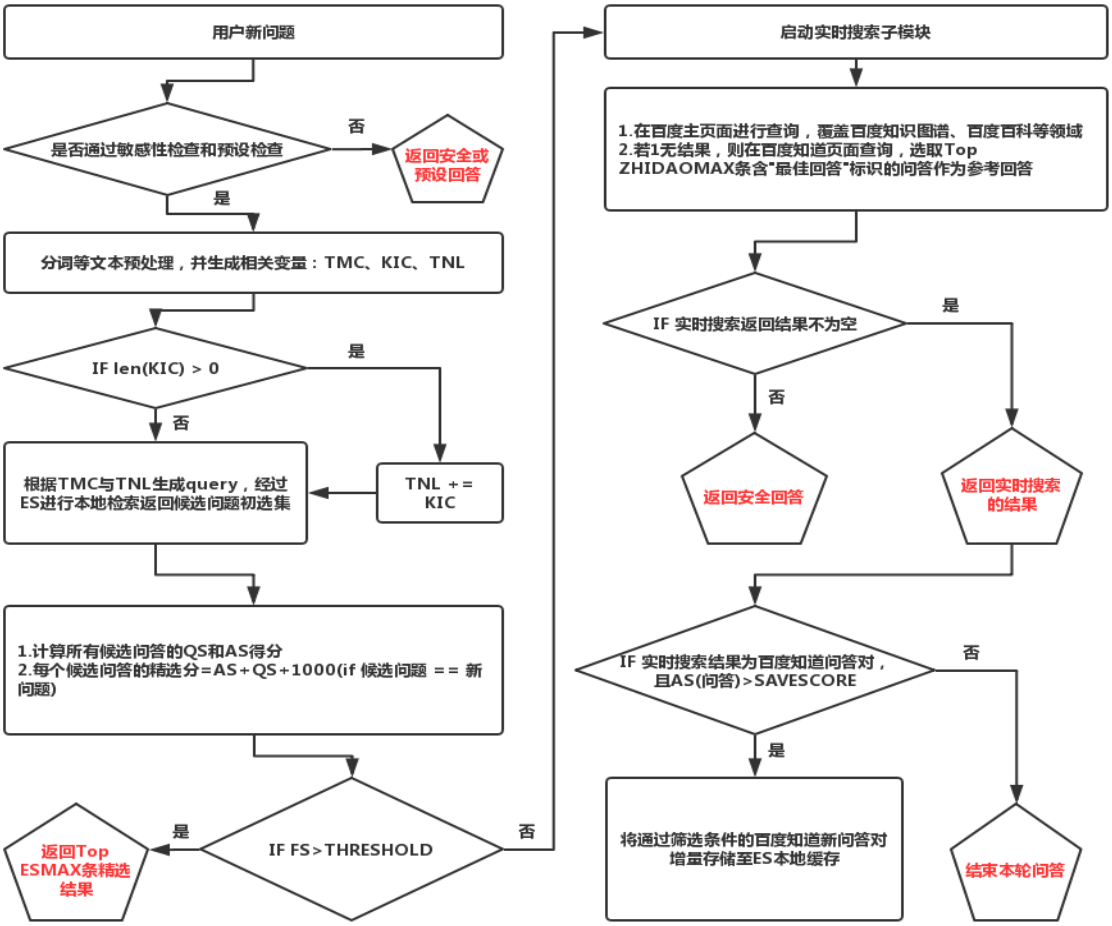
⑥异常处理机制：当所有检索途径都不能返回适合用户新问题的回答时，返回安全回答，或调用基于生成式问答技术的闲聊模块产生回答。

主程序设计

1.变量、参数命名及说明

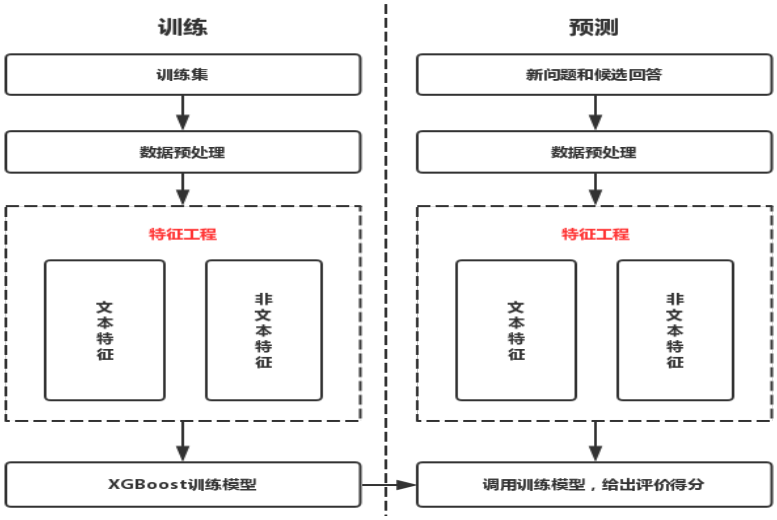
变量/参数名	说明
StopWords	停用词表
KeyWords	核心主题词表
Text_Main_Content(TMC)	用户新问题经分词、去除停用词后的分词结果
Keywords_in_Content(KIC)	用户新问题命中的核心主题词
Text_NER_List(TNL)	用户新问题中的分词后的NER
Question_Main_Content(QMC)	ES本地检索返回的候选问题经分词、去除停用词后的分词结果
QuestionScore(QS)	语义匹配模块返回的候选问题与新问题的相似度得分
AnswerScore(AS)	通过答案排序模块计算出的，候选问题对应的候选答案与用户新问题组成“好问答对”的得分
FinalScore(FS)	候选问答在精选规则下的最终得分
THRESHOLD	控制在精选集中记录ES初选结果的阈值。
ZHIDAOMAX	控制在线搜索时从百度知道返回的问答数量。
SAVESCORE	控制增量存储模块中的百度知道问答对的阈值
ESMAX	控制精选集中的精选问答数量

2.每轮问答程序流图



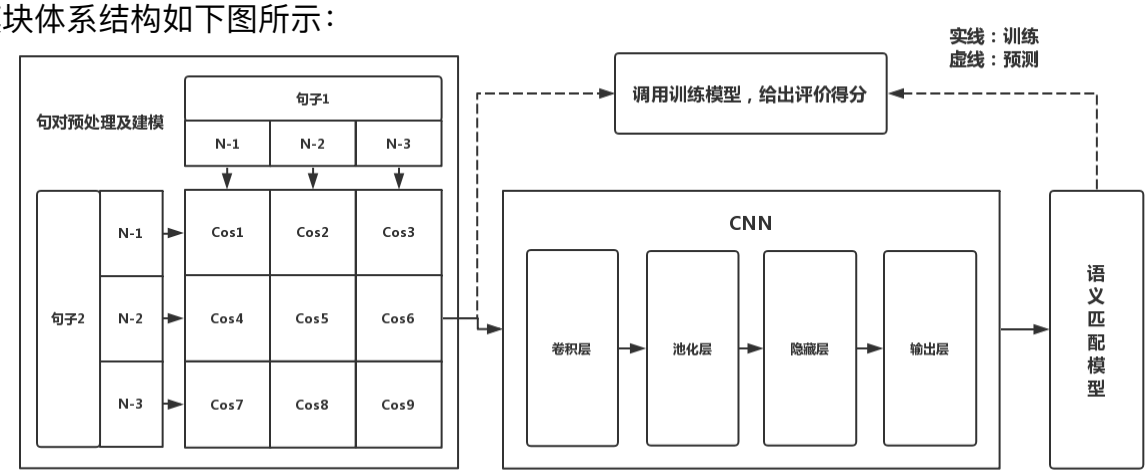
子模块解读

1.问答质量评价子模块：包括数据预处理、问答特征工程[4]、模型训练与调用等多个组成部分，模块体系结构如下图所示：



[4] Toba H, Ming Z Y, Adriani M, et al. Discovering high quality answers in community question answering archives using a hierarchy of classifiers[J]. Information Sciences, 2014, 261(5):101-115.

2.语义匹配子模块：包括句对预处理及建模[5]、模型训练与调用等多个组成部分，模块体系结构如下图所示：



句对预处理及建模：将输入的句子1与句子2进行分词，然后将两组切分后的结果两两计算基于Word Embedding的余弦相似度，并填充成上图所示的2D Feature Map形式(图示采用了3-gram的切分)。当句对的分词list长度不同时，使用0填充缺失值。

模型训练与调用：训练时，将训练集句对的Feature Map输入到CNN模型中，获取语义匹配模型并保存；预测时，调用预训练好的语义匹配模型，输入新句对的Feature Map，将返回这组句对的相似概率得分。

这个方法存在一定的**问题**[6]：负样本很难标注。

3.在线搜索及缓存子模块：包括百度主页搜索、百度知道搜索和本地缓存优质问答等多个组成部分，模块体系结构如下图所示：

百度主页搜索: 向百度主页请求输入用户新问题后的检索结果，若检索结果的第一条中包含以上结果链接，则获取内容并向用户返回。

百度知道搜索：当百度主页搜索没有结果时，在百度知道页面请求输入用户新问题后的检索结果，若前ZHIDAOMAX条结果中包含带有“最佳回答”标识的回答，则获取对应问答内容并返回。

---

[5] 会计家园社区问答系统：http://blog.csdn.net/malefactor/article/details/50374237

[6] 数据量越大，我们需要的负样本越接近词义上相似但语义不相似的句对

2018年3月28日 星期三

本地缓存优质问答：当百度知道搜索有结果返回时，调用问答质量评价子模块对该问答对进行质量判断，将高于设定阈值(SAVESCORE)的问答对缓存到本地知识库。缓存机制有助于后续提升系统整体的响应速度及性能。

未来思路

1.为CNN语义匹配积累语料，为问答子模块积累新特征和数据。

2.使用开发信息抽取技术OIE抽取断言，实现知识结构化，提升系统的推理和知识理解能力。