AliMe Chat: A Sequence to Sequence and Rerank based Chatbot Engine

Minghui Qiu, Feng-Lin Li, SiyuWang, Xing Gao, Yan Chen, Weipeng Zhao, Haiqing Chen, Jun Huang,Wei Chu
Alibaba Group, Hangzhou, China

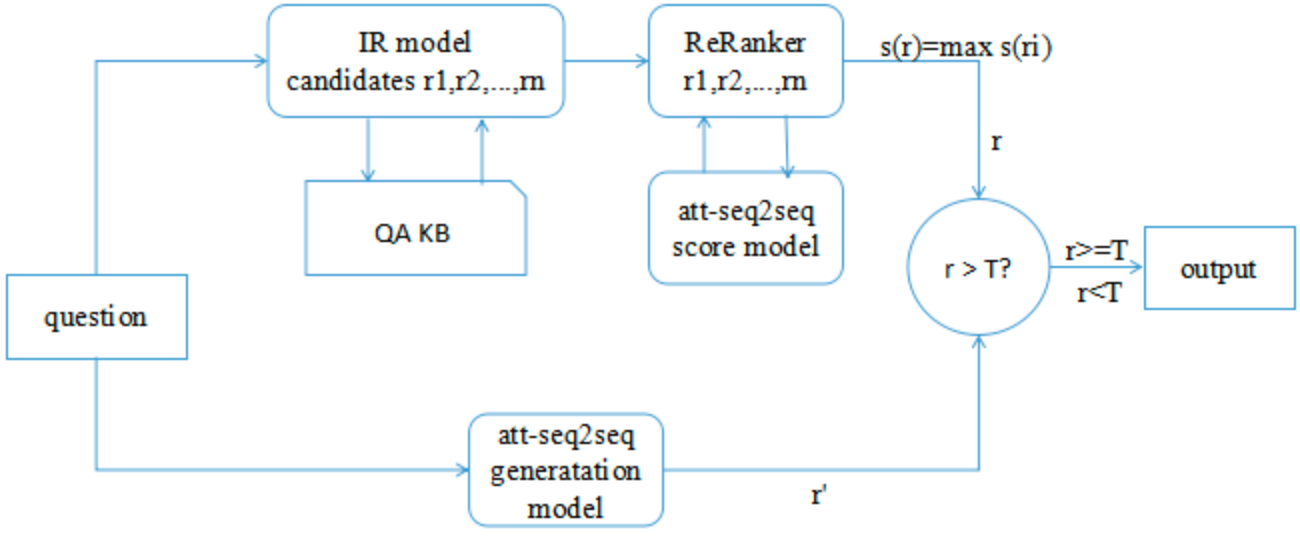
Motivation

目前对话生成主要基于两种方式：（1）信息检索（2）生成式模型。这两种方式都有其短处，对于长问题或者复杂问题，IR很难找到相关的匹配项，而生成式模型对于问题生成的质量是不可控的。因此，将两种方法结合是一种不错的折中解决方法。

Model

1.对于给定问题，通过检索模型，找到k个与之相关的候选<q,a>对。
2.然后利用一个预训练的attention-seq2seq模型对候选<q,a>对进行打分。
3.选取评分最高的<q,a>对的a作为预测答案。如果该得分超过已定阈值，那么就将该a作为输出。否则，将另一个att-seq2seq模型生成的答案作为输出。

Model


IR model
candidates r1,r2,...,rn
ReRanker
r1,r2,...,rn
s(r)=max s(ri)
r
QA KB
att-seq2seq
score model
r > T?
r>=T
r<T
output
question
att-seq2seq
generatation
model
r'