阅读理解评测标准

# 评估原则

评测中，我们会同时计算ROUGE-L和BLEU-4两个指标，ROUGE-L将被作为主要参考指标用于排名。

本次评测中开放的数据除了可以支持基本的阅读理解任务之外，还提供了对问题类型的标注，对是非问题的答案所包含的观点的判断，以及对实体类问题对应的全部文章中所包含的命名实体集合的标注。我们鼓励参赛者在找到答案文本之后，能够对答案做进一步的理解。对于问题类型为‘YES\_NO’（是非类问题，指问是非观点的问题）的问题，希望参赛者能够对自己找到的答案做进一步的观点判断，如果判断正确，评估时将会得到一定的奖励；而对于问题类型为‘ENTITY’（实体类问题，指问“多少钱”，“什么地方”等期待回答是实体的问题）的问题，虽然参赛者不需要在答案的基础上做进一步的工作，但我们在评价时会针对答案中包含的命名实体进行一定的奖励。

# BLEU计算方法

N元文法的BLEU分数原理是在候选答案N元的准确率（右项，大括号）的基础上加上适当的长度惩罚（左项，BP：brevity penalty），因此相对侧重答案的准确率，其中每N元计算的准确率的公式为：

即候选答案命中参考答案的N元个数在候选答案N元中的占比（也就是N元准确率）。

其中黑色部分为计算BLEU时的标准方法，而红色部分是对是非类问题和实体类问题的额外奖励项。和分别是对是非奖励和实体奖励的权重。如果参赛者对答案的是非观点进行了正确的判断，那么他找到的答案将会和对应观点的参考答案进行匹配，匹配到的N元文法数量将会同时加到分子分母中，这样做了正确的是非观点判断的答案的得分将会增加；而对于实体类问题，如果用户找到的答案中包含了实体，那么被匹配的实体也会被作为被匹配的N元文法同时加入分子和分母，这样包含更多实体的答案的得分将会增加。

而长度惩罚则将参考答案的长度作为惩罚的参考值，将长于参考答案的候选答案进行一定程度惩罚，具体公式为：

其中多参考答案下的BP的计算方案有多种，一般可取最长参考答案的长度作为r值。

在本次测评中，我们采用，（即评价BLEU-4），，多参考答案的情况下，计算BP时取与预测答案长度最接近的参考答案的长度作为r值。

# ROUGE-L计算方法

ROUGE-L值计算的原理是考察候选答案和参考答案的最长公共子序列上的准确和召回情况，其中R为最长公共子序列在参考答案上的召回率，P为最长公共子序列在候选答案上的准确率，其中：

为最长公共子序列的长度与参考答案长度的比值，即召回率，而

为最长公共子序列的长度与候选答案长度的比值，即准确率。一般情况下公式中设置为较大的值，此时ROUGE-L值侧重考察答案的召回率，本次评测中取折衷考虑正确率召回率。

与前文类似，此处计算召回率和准确率的公式中红色部分为对是非问题和实体问题的额外奖励项。如果参赛者对是非类问题的答案做了进一步的观点判断并且判断正确的话，分子分母会同时加上这个答案和对应观点的参考答案之间的最长公共子串的长度，这样准确率和召回率都会提高；对于实体类问题，如果参赛者预测的答案中包含实体，那么这个实体的长度会也会同时被加到分子和分母中，同样会让准确率和召回率有一定提升。我们在本次测评中取。