Bleu

BLEU，全称为Bilingual Evaluation Understudy（双语评估替换），是一个比较候选文本翻译与其他一个或多个参考翻译的评价分数。

Bleu据我的理解,他主要是应用于生成对话领域，我们的检索式chatbot使用这个得分可能不是很合理。

定义

这种评测方法通过对候选翻译与参考文本中的相匹配的n元组进行计数，其中一元组（称为1-gram或uni-gram）比较的是每一个单词，而二元组（bigram）比较的将是每个单词对。这种比较是不管单词顺序的。
n元组匹配的计数结果会被修改，以确保将参考文本中的单词都考虑在内，而不会对产生大量合理词汇的候选翻译进行加分。在BLEU论文中这被称之为修正的n元组精度。
BLEU评分是用来比较语句的，但是又提出了一个能更好地对语句块进行评分的修订版本，这个修订版根据n元组出现的次数来使n元组评分正常化。

Bleu分为Bleu-1,2,3,4。
bleu-1就是1-gram的权重为1其他权重为0
Bleu-2就是1和2gram的权重为0.5，0.5其他为0
以此类推

简单的来说Bleu就是比较两个句子，取n=1的时候就是比较相同单词的个数，n=2的时候就是比较单词对相同的个数...

方法

如何评估一个模型的Bleu得分：（最原始的Bleu）

1. 对于一个模型，我们会为测试集句子（问题）产生一些候选回应（candidate1, candidate2, ……）
2. 这些句子在测试集中有参考回应（reference1, reference2, ……）
3. $score = \frac{|candidate \cap reference|}{|candidate|}$ (以词为单位的交集比较)

总结

Bleu在评价生成式的模型的时候能给出合理的解释，因为生成式是由machine给我们生成的句子，比较一个未知的句子跟参考句子的相似度比较合理。如果把他用在检索模型上面，可以想象一个场景，我们模型在检索的时候很多都是检索对了，那么bleu就为1，因为检索的句子肯定跟参考句子一某一样，这就会导致很高的bleu。同理，在我们检索错了的时候，但是我们检索的回应中很多词跟参考回应的句子中的词是一样的，那么他也会得到一个高的bleu得分，这是非常不合理的。
在之前我做过的调查中，我在李纪为的论文《ResponseGenerationbyContext-awarePrototypeEditing》使用的人工评价指标可以尝试下（在很多chatbot论文中也采用了人工评估的方法，可能具体的方法不一样），他在论文中使用的4个人工指标loss/tie/win/k，论文中还采用了机器评估方法：fluency, relevance, diversity and originality4个评估指标
loss：代表自己模型提供的回应相对与baseline或者是参考（很多语聊库参考回应也不合理）差的比例
win：与loss相反
Tie：表示人工认为候选回应和参考回应那个更加合理
k：表示人工认为模型给出的回应的合理程度

References

[1] 《BLEU: a Method for Automatic Evaluation of Machine Translation》
[2] https://cloud.tencent.com/developer/article/1042161
[3] 《ResponseGenerationbyContext-awarePrototypeEditing》