**E2E数据集：**

1. 参考文本表现出更多的词汇和句法丰富性，包括各种话语现象。
2. 该数据集上的文本生成需要进行内容选择。

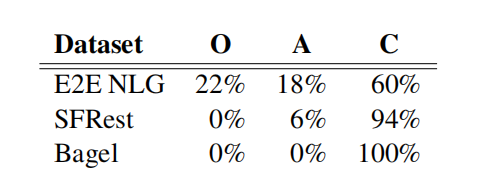
与其他数据集不同的是，我们人工对MR中的所有有用信息进行口头表达，并被允许跳过被认为不重要的属性值。这一特性使得从我们的数据集生成文本更具挑战性，因为NLG系统还需要学习实现哪些内容。

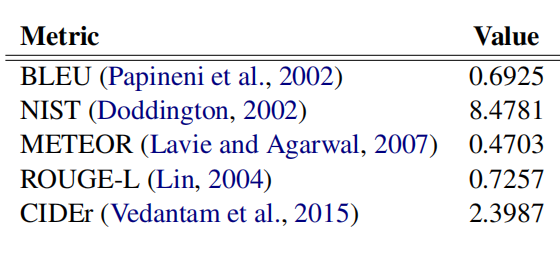
为了说明这种现象，我们随机抽取了50对MR-REF对

fully covered ：NL出现与MR相同的信息 （C)

additional ：NL 中包含MR中未提到的 (A)

omitted ：MR包含NL中未提及的信息 (O)



Baseline-Tgen:

**评估方法：**

1. **BLEU**

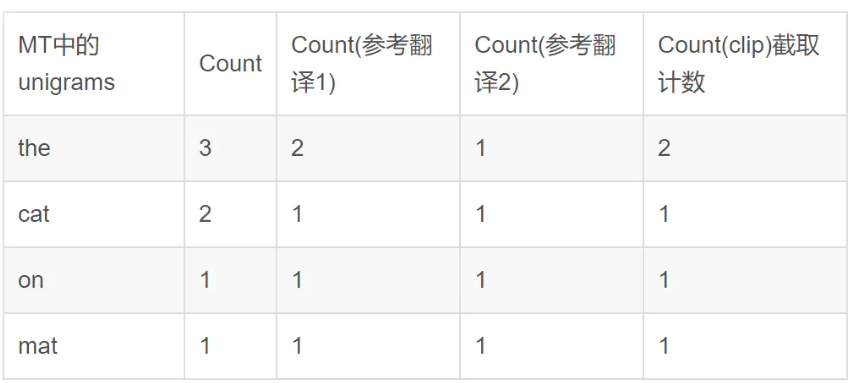
BLEU名为双语评估替换，是衡量一个有多个正确输出结果的模型的精确度的评估指标。

Ref1：The cat is on the mat

Ref2：There is a cat on the mat

Candidate: the cat the cat on the mat

**BLEU 1-gram**



(2+1+1+1) / (3+2+1+1) = 5/7

**BLEU 2-gram**



(1+0+1+1+1) / (2+1+1+1+1) = 2/3

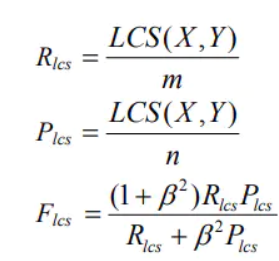
1. **gram、4-gram。。。**

最后再乘上一个“简短惩罚” BP（brevity penalty），即最后的bleu得分为：Bleu(total)=BP \* bleu(avg)。



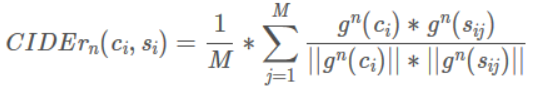
1. **ROUGE-L**

LCS（最长公共子序列）：给定两个序列X和Y，使得公共子序列长度最大的序列是X和Y的最长公共子序列。



1. **CIDEr**

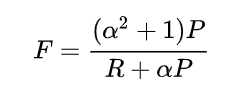
CIDEr采用了一种TF-IDF+余弦相似度的组合。



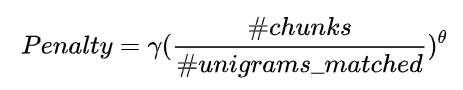


**4.METEOR**

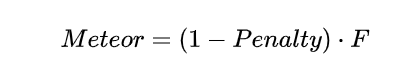
METEOR是一种基于单精度的加权调和平均数和单字召回率的度量方法，目的是解决BLEU标准中的一些固有缺陷。



Meteor的特别之处在于，它不希望生成很“碎”的译文：比如参考译文是“A B C D”，模型给出的译文是“B A D C”，虽然每个unigram都对应上了，但是会受到很严重的惩罚。惩罚因子的计算方式为：



最终评分：



**评分：**

