(第四组)【综述研究】翻译质量评估模型（QA Model）已广泛应用于翻译行业内，当下比较著名的翻译质量量化评估模型包括本地化企业使用率高达20%的LISA QA Model，适用于汽车行业的SAE J2450，以及TMS classic等等.  
（1）查找文献或其他实用性资料，对**至少三种翻译质量量化评估模型**中的参数设置和运行环节进行综述。  
（2）对翻译质量质性评估模型进行学习，探讨**翻译质量质性模型**与**翻译质量量化模型**间转换的困难之处，当前使用的翻译质量量化模型还能在哪些方面进行改进。请自行构造文章，最好能反应这些模型特点的例子在里面，进行实例研究，对比各模型的差异  
（3）根据以上学习内容，尝试设计游戏翻译领域的翻译质量量化评估模型。

三种量化模型的参数设置和运行环节

质性与量化模型的转换困难

量化模型的改进

简单介绍质性模型与量量化模型

（1）概念

（2）评估对象：机器（A LPAC、BLEU、NIST）和人工翻译【分别讲两个模型？】

评估方法：调查翻译质量标准

参数设置（简单）和运行环节（？）

LISA QA Model 错误扣分模式

# <http://producthelp.sdl.com/SDL_TMS_2011/en/Creating_and_Maintaining_Organizations/Managing_QA_Models/LISA_QA_Model.htm>

质性与量化模型的转换困难

转换？为什么要转换？

不用对或错（扣分制？）对译文加以判断，而是依据它在具体情境中的表现对其加以评判。

质性模型质性/定性评价

量化模型改进：缺陷总结+对应改进

机器自动评估翻译质量（以BLEU为例）

BLEU算法

1. **算法详述**

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。最初是使用BLEU来对机器翻译进行评估，设计思想与评判机器翻译好坏的思想一致，即机器翻译的译文越接近人类专业译文，那么翻译的效果就越好。机器无法判断两个译文意思是否一致，于是就判断两个句子的相似程度。计算公式：



算法详解：

**（1）N-gram匹配规则（Pn计算规则）**

1. gram代表n元组精度。

原理：就是比较译文和参考译文之间n组词的相似的一个占比。

翻译结果叫做候选翻译集（candidate1, candidate2, ……）;

参考翻译叫做（reference1, reference2, ……）;

将精度记做Pn， n代表n-gram，又叫做n-gram precision scoring——多元精度得分，在这里一般n不大于4。

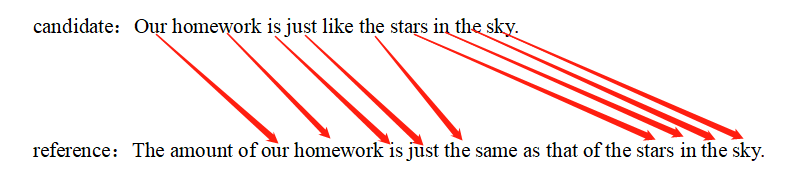
举例：

原文：我们的作业真像是天上的星星一样多。

candidate：Our homework is just like the stars in the sky.

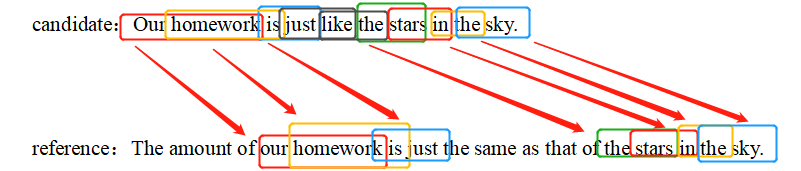
reference：The amount of our homework is just the same as that of the stars in the sky.

如果用1-gram匹配的话：



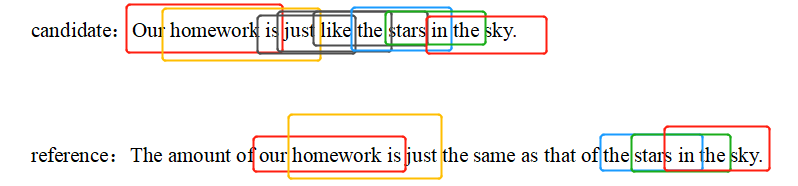
候选译文一共10个词，有9个词语命中了参考译文，所以1-gram的匹配度为9/10

2-gram：

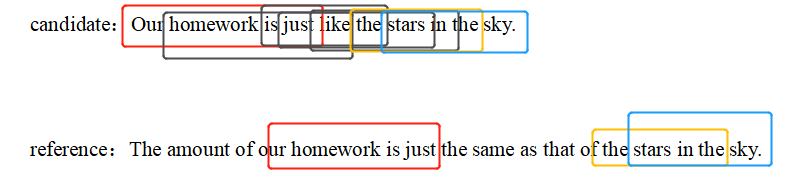


候选译文一共可以分为9个2-gram的词组，其中有7个可以命中参考译文，那么它2-gram的匹配度为 7/9

同理3-gram的匹配度为：5/8



1. gram的匹配度为：3/7



可以用程序来遍历计算N-gram的匹配度。n-gram精度得分可以视为衡量了翻译评估的充分性和流畅性两个指标：一般来说一元组（1-gram）属于字符级别，结果代表了文中有多少个词被单独翻译出来了，关注的是翻译的充分性，衡量逐字翻译能力；多元组上升到了词汇级别，结果反映的是译文的流畅度，词组越准确，文章的可读性就越好。

但光有N-gram匹配度对译文进行评判显然是不够的，上述方法会出现明显错误如下：

原文：我们的作业真像是天上的星星一样多。

candidate：homework homework homework homework homework

Reference1：The amount of our homework is just the same as that of the stars in the sky.

Reference2：We have lots of homework, the amount of the homework is like the stars in the sky.

当出现上述情况时，候选译文的一元组匹配度是5/5，这一结果显然不合理，主要是分子的统计不合理，因此需要对上面式子中的分子进行改进。规则（截断计数）如下（一个或多个参考译文都适用）：





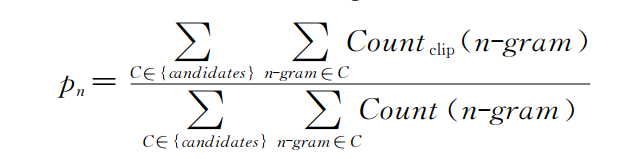
：单词Wi在候选译文里出现的次数

：单词Wi在第j个参考译文里出现的次数

：对于第j个参考译文，Wi的截断计数（因为可能有不止一个参考译文）

：Wi在所有参考译文里的综合截断计数

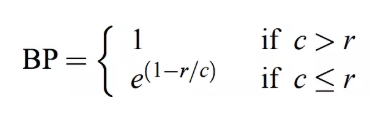
以上述例子计算，等于5，分别等于1，2，最终等于2，所以一元组匹配度最终为2/5，引入n元组之后，Pn的计算表达式如下：



简单来说，这个连加公式就是将候选段落的所有n-gram进行了截断统计作为分子，分母是候选段落的n-gram的个数。

**（2）BP计算规则**

上述算法已经足够可以有效的翻译评估了，然而N-gram的匹配度可能会随着句子长度的变短而变高，因此会存在这样一个问题：一个候选译文只翻译出了部分句子且翻译的比较准确，那么它的匹配度就会很高。为了避免这种评分的偏向性，BLEU在最后的评分结果中引入了长度惩罚因子(Brevity Penalty)。



其中c表示候选译文的长度，r表示参考译文的长度，当存在多个参考译文时，选取与候选译文最接近的句子长度。当翻译译文长度大于参考译文的长度时，惩罚系数为1，意味着不惩罚，只有候选译文长度小于参考译文惩罚因子才会生效。

**（3）最终公式**

随着n-gram的增大，精度得分总体上成指数下降，可以粗略的看成随着n指数级的下降。所以公式采取几何加权平均，最终得到的公式：（Wn=1/N权重系数，一般将N赋值为4）



**（4）实例计算**

用上述提到的例子计算一次：

原文：我们的作业真像是天上的星星一样多。

candidate：Our homework is just like the stars in the sky.

reference：The amount of our homework is just the same as that of the stars in the sky.

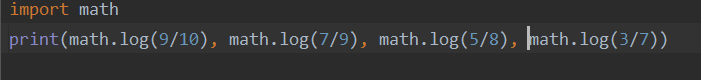
P1=9/10

P2=7/9

P3=5/8

P4=3/7

再计算logPn（这里用的是Python的自带函数）





求和再乘以Wn，也就是除以4取正值约等于0.418。

BP = e(1-17/10)约等于0.497

BLEU =0.497\* e0.418 ≈0.755

BLEU评分的范围是从0到1。很少有翻译得分为1，除非候选译文与参考译文完全相同。因此，即使是译员，也不一定会在一个大约500个句子（也就是40个普通新闻报道的长度）的测试语料上得1分，一个译员在四个参考翻译下的得分为0.3468，在两个参考翻译下的得分为0.2571。（信息来源：BLEU- a Method for Automatic Evaluation of Machine Translation）

|  |  |
| --- | --- |
| 优点 | 缺点 |
| ①方便 | ①BLEU 值仅仅只是在计算机翻译文与参考译文之间的相似性，而没有考虑语言表达（语法）上的准确性； |
| ②计算迅速 | ②常用词、译文长度、同义词等很多情况都会影响到 BLEU 值的评判，因此测评精度会受干扰：容易陷入常用词和短译句的陷阱中，短译句的测评精度有时会较高，没有考虑同义词或相似表达的情况，可能会导致合理翻译被否定； |
| ③结果有一定程度的参考价值 | ③计算 BLEU 值最重要的是提供参考译文，参考译文质量越高，对于同一句原文的参考译文越多，翻译模型的度量就越准确。但很多情况下高质量参考译文很难获得，或者成本非常高。 |

**二、优缺点总结：**