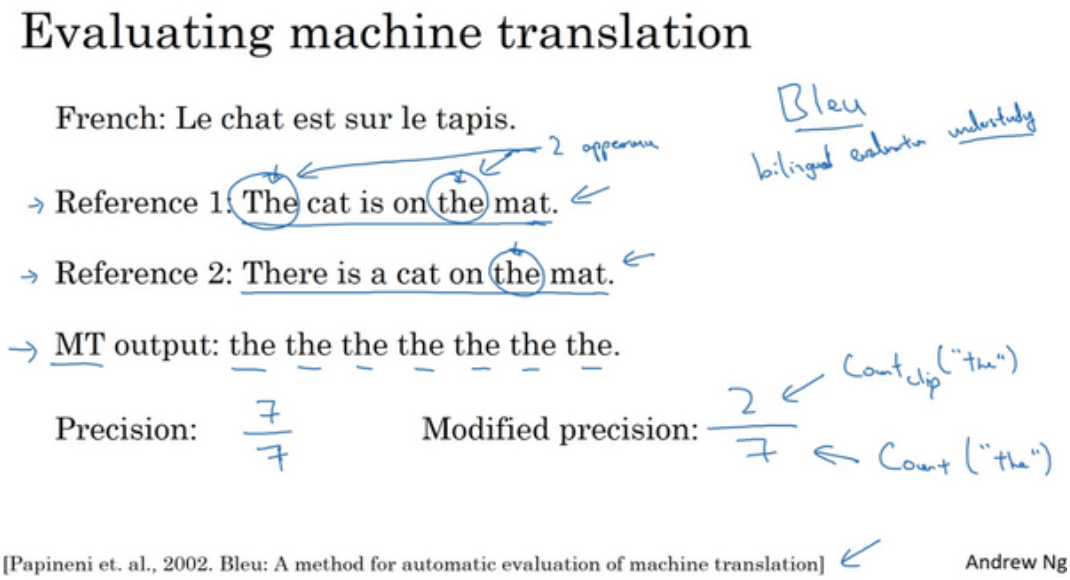
3.6 Bleu 得分（选修）（Bleu Score (optional)）

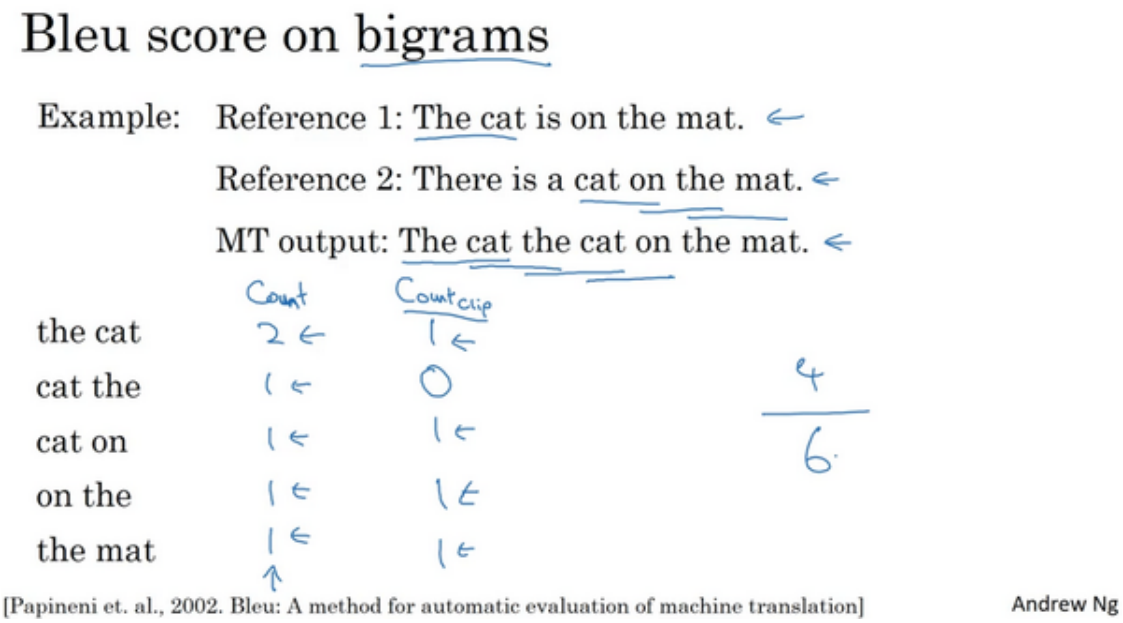
机器翻译（**machine translation**）的一大难题是一个法语句子可以有多种英文翻译而且都同样好，所以当有多个同样好的答案时，怎样评估一个机器翻译系统呢？不像图像识别（**image recognition**），只有一个正确答案，就只要测量准确性就可以了。如果有多个不错的答案，要怎样衡量准确性呢？常见的解决办法是，通过一个叫做 **BLEU 得分（the BLEU score）**的东西来解决。所以，在这个选修视频中，我想与你分享，我想让你了解 **BLEU** 得分是怎样工作的。

假如给你一个法语句子：**Le chat est sur le tapis**，然后给你一个这个句子的人工翻译作参考：**The cat is on the mat**。不过有多种相当不错的翻译。所以一个不同的人，也许会将其翻译为：**There is a cat on the mat**，同时，实际上这两个都是很好的，都准确地翻译了这个法语句子。**BLEU 得分做的就是，给定一个机器生成的翻译，它能够自动地计算一个分数来衡量机器翻译的好坏。**直觉告诉我们，只要这个机器生成的翻译与任何一个人工翻译的结果足够接近，那么它就会得到一个高的 **BLEU** 分数。顺便提一下 **BLEU** 代表 **bilingual evaluation understudy (双语评估替补)**。在戏剧界，侯补演员(**understudy**)学习资深的演员的角色，这样在必要的时候，他们就能够接替这些资深演员。而 **BLEU** 的初衷是相对于请评估员（**ask human evaluators**），人工评估机器翻译系统（**the machine translation system**），**BLEU** 得分就相当于一个侯补者，它可以代替人类来评估机器翻译的每一个输出结果。**BLEU** 得分是由 **Kishore Papineni, Salim Roukos**，**Todd Ward** 和 **Wei-Jing Zhu** 发表的这篇论文十分有影响力并且实际上也是一篇很好读的文章。所以如果有时间的话，我推荐你读一下。**BLEU** 得分背后的理念是观察机器生成的翻译，然后看生成的词是否出现在少一个人工翻译参考之中。因此这些人工翻译的参考会包含在开发集或是测试集中。



（**参考论文：Papineni, Kishore & Roukos, Salim & Ward, Todd & Zhu, Wei-jing. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. 10.3115/1073083.1073135.**）

现在，我们来看一个极端的例子。我们假设机器翻译系统缩写为 **MT**。机器翻译 **(MT)**的输出是：**the the the the the the the**。这显然是一个十分糟糕的翻译。衡量机器翻译输出质量的方法之一是观察输出结果的每一个词看其是否出现在参考中，这被称做是机器翻译的精确度（**a precision of the machine translation output**）。这个情况下，机器翻译输出了七个单词并且这七个词中的每一个都出现在了参考 1 或是参考 2。单词 **the** 在两个参考中都出现了，所以看上去每个词都是很合理的。因此这个输出的精确度就是 **7/7**，看起来是一个极好的精确度。这就是为什么把出现在参考中的词在 **MT** 输出的所有词中所占的比例作为精确度评估标准并不是很有用的原因。因为它似乎意味着，例子中 **MT** 输出的翻译有很高的精确度，因此取而代之的是我们要用的这个改良后的精确度评估方法，我们把每一个单词的记分上限定为它在参考句子中出现的最多次数。在参考 1 中，单词 **the** 出现了两次，在参考 2 中，单词 **the** 只出现了一次。而 2 比 1 大，所以我们会说，单词 **the** 的得分上限为 2。有了这个改良后的精确度，我们就说，这个输出句子的得分为 **2/7**，因为在 7 个词中，我们最多只能给它 2 分。所以这里分母就是 7 个词中单词 **the** 总共出现的次数，而分子就是单词 **the** 出现的计数。我们在达到上限时截断计数，这就是改良后的精确度评估（**the modified precision measure**）。



到目前为止，我们都只是关注单独的单词，在 **BLEU** 得分中，你不想仅仅考虑单个的单词，你也许也想考虑成对的单词，我们定义一下**二元词组（bigrams）的 BLEU 得分**。**bigram** 的意思就是相邻的两个单词。现在我们来看看怎样用二元词组来定义 **BLEU** 得分，并且这仅仅只是最终的 BLEU 得分的一部分。我们会考虑一元词组（**unigrams**）也就是单个单词以及二元词组（**bigrams**），即成对的词，同时也许会有更长的单词序列，比如说三元词组（**trigrams**）。意思是三个挨在一起的词。我们继续刚才的例子，还是前面出现过的参考 1 和 2，不过现在我们假定机器翻译输出了稍微好一点的翻译:**The cat the cat on the mat**，仍然不是一个好的翻译，不过也许比上一个好一些。这里，可能的二元词组有 **the cat** ，忽略大小写，接着是 **cat the**， 这是另一个二元词组，然后又是 **the cat**。不过我已经有了，所以我们跳过它，然后下一个是 **cat on**，然后是 **on the**，再然后是 **the mat**。所以这些就是机器翻译中的二元词组。好，我们来数一数每个二元词组出现了多少次。**the cat** 出现了两次 ，**cat the** 出现了一次，剩下的都只出现了一次。最后 ，我们来定义一下**截取计数（the clipped count）**。也就是 **Count_clip**。为了定义它，我们以这列的值为基础，但是给算法设置得分上限，上限值为二元词组出现在参考 1 或 2 中的最大次数。**the cat** 在两个参考中最多出现一次，所以我将截取它的计数为 1。**cat the** 它并没有出现在参考 1 和参考 2 中，所以我将它截取为 0。**cat on** ，好，它出现了一次，我们就记 1 分。**on the** 出现一次就记 1 分，**the mat** 出现了一次，所以这些就是截取完的计数（**the clipped counts**）。我们把所有的这些计数都截取了一遍，实际上就是将它们降低使之不大于二元词组出现在参考中的次数。最后，修改后的二元词组的精确度就是 **count_clip** 之和。因此那就是 4 除以二元词组的总个数，也就是 6。因此是 4/6 也就是 2/3 为二元词组改良后的精确度。

Bleu score on unigrams

Example: Reference 1: The cat is on the mat.

Reference 2: There is a cat on the mat.

→ MT output: The cat the cat on the mat. ($\hat{y}$)

$p_1, p_2, \ldots = 1.0$

$$p_1 = \frac{\sum_{\text{unigram} \in \hat{y}} \text{Count}_{\text{clip}}(\text{unigram})}{\sum_{\text{unigram} \in \hat{y}} \text{Count}(\text{unigram})}$$

unigram

$$p_n = \frac{\sum_{\text{n-gram} \in \hat{y}} \text{Count}_{\text{clip}}(\text{n-gram})}{\sum_{\text{n-gram} \in \hat{y}} \text{Count}(\text{n-gram})}$$

n-gram

现在我们将它公式化。基于我们在一元词组中学到的内容，我们将改良后的一元词组精确度定义为P_1，P代表的是精确度。这里的下标 1 的意思是一元词组。不过它定义为一元词组之和，也就是对机器翻译结果中所有单词求和，**MT** 输出就是$\hat{y}$，**Countclip(unigram)**。除以机器翻译输出中的一元词组出现次数之和。因此这个就是最终结果应该是两页幻灯片前得到的 2/7。这里的 1 指代的是一元词组，意思是我们在考虑单独的词，你也可以定义P_n为n元词组精确度，用 **n-gram** 替代掉一元词组。所以这就是机器翻译输出中的n元词组的 **countclip** 之和除以n元词组的出现次数之和。因此这些精确度或说是这些改良后的精确度得分评估的是一元词组或是二元词组。就是我们前页幻灯片中做的，或者是三元词组，也就是由三个词组成的，甚至是n取更大数值的n元词组。这个方法都能够让你衡量机器翻译输出中与参考相似重复的程度。另外，你能够确信如果机器翻译输出与参考 1 或是参考 2 完全一致的话，那么所有的这些P_1、P_2等等的值，都会等于 1.0。为了得到改良后的 1.0 的精确度，只要你的输出与参考之一完全相同就能满足，不过有时即使输出结果并不完全与参考相同，这也是有可能实现的。你可以将它们以另一种方式组合，但愿仍能得到不错的翻译结果。

Bleu details

p_n = Bleu score on n-grams only

p_1, p_2, p_3, p_4

Combined Bleu score: $\text{BP} \exp\left(\frac{1}{4}\sum_{n=1}^{4} p_n\right)$

BP: brevity penalty

$$\text{BP} = \begin{cases} 1 & \text{if MT\_output\_length} > \text{reference\_output\_length} \\ \exp(1 - \text{MT\_output\_length}/\text{reference\_output\_length}) & \text{otherwise} \end{cases}$$

最后，我们将这些组合一下来构成最终的 **BLEU** 得分。P_n就是n元词组这一项的 **BLEU** 得分，也是计算出的n元词组改良后的精确度，按照惯例，为了用一个值来表示你需要计算P_1，P_2， P_3，P_4。然后将它们用这个公式组合在一起，就是取平均值。按照惯例 **BLEU** 得分被定义为，$exp(\frac{1}{4}\sum_{n=1}^{4} P_n)$，对这个线性运算进行乘方运算，乘方是严格单调递增的运算，我们实际上会用额外的一个叫做 **BP** 的惩罚因子（**the BP penalty**）来调整这项。**BP** 的意思是“简短惩罚”（ **brevity penalty**）。这些细节也许并不是十分重要，但是你可以大致了解一下。 事实表明，如果你输出了一个非常短的翻译，那么它会更容易得到一个高精确度。因为输出的大部分词可能都出现在参考之中，不过我们并不想要特别短的翻译结果。因此简短惩罚**(BP)**就是一个调整因子，它能够惩罚输出了太短翻译结果的翻译系统。**BP** 的公式如上图所示。如果你的机器翻译系统实际上输出了比人工翻译结果更长的翻译，那么它就等于 1，其他情况下就是像这样的公式，惩罚所有更短的翻译，细节部分你能够在这篇论文中找到。

再说一句，在之前的视频中，你了解了拥有单一实数评估指标（**a single real number evaluation metric**）的重要性，因为它能够让你尝试两种想法，然后看一下哪个得分更高，尽量选择得分更高的那个，**BLEU** 得分对于机器翻译来说，具有革命性的原因是因为它有一个相当不错的虽然不是完美的但是非常好的单一实数评估指标，因此它加快了整个机器翻译领域的进程，我希望这节视频能够让你了解 **BLEU** 得分是如何操作的。实践中，很少人会从零实现一个 **BLEU** 得分（**implement a BLEU score from scratch**），有很多开源的实现结果，你可以下载下来然后直接用来评估你的系统。不过今天，**BLEU** 得分被用来评估许多生成文本的系统（**systems that generate text**），比如说机器翻译系统（**machine translation systems**），也有我之前简单提到的图像描述系统（**image captioning systems**）。也就是说你会用神经网络来生成图像描述，然后使用 **BLEU** 得分来看一下，结果在多大程度上与参考描述或是多个人工完成的参考描述内容相符。**BLEU** 得分是一个有用的单一实数评估指标，用于评估生成文本的算法，判断输出的结果是否与人工写出的参考文本的含义相似。不过它并没有用于语音识别（**speech recognition**）。因为在语音识别当中，通常只有一个答案，你可以用其他的评估方法，来看一下你的语音识别结果，是否十分相近或是字字正确（**pretty much, exactly word for word correct**）。不过在图像描述应用中，对于同一图片的不同描述，可能是同样好的。或者对于机器翻译来说，有多个一样好的翻译结果，**BLEU** 得分就给了你一个能够自动评估的方法，帮助加快算法开发进程。说了这么多，希望你明白了 **BLEU** 得分是怎么运行的。