Alberto Andrés Valdés González.
Degree: Mathematical Engineer.
Work position: Data Scientist.
Mail: anvaldes@uc.cl/alberto.valdes.gonzalez.96@gmail.com
Location: Santiago, Chile.

BLEU Score

BLEU (Bilingual Evaluation Understudy) is an algorithm for evaluating the quality of text which has been machine-translated from one natural language to another. Quality is considered to be the correspondence between a machine's output and that of a human: "the closer a machine translation is to a professional human translation, the better it is" – this is the central idea behind BLEU. BLEU was one of the first metrics to claim a high correlation with human judgements of quality, and remains one of the most popular automated and inexpensive metrics.

Example:

French: Le chat est sur le tapis.

Reference 1: The cat is on the mat.

Reference 2: There is a cat on the mat.

MT output: The the the the the.

Now we going to compute the precision:

$$\text{Precision} = \frac{\text{Sum of indicator of each word if appears in some of the references}}{\text{Number of words}} = \frac{7}{7} = 1$$

Modified precision: Using bigrams.

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.

Sentences	Bigram 1	Bigram 2	Bigram 3	Bigram 4	Bigram 5	Bigram 6
Reference 1	The cat	cat is	is on	on the	the mat	-
Reference 2	There is	is a	a cat	cat on	on the	the mat
MT output	The cat	cat the	the cat	cat on	on the	the mat

Now we going to summary this table for MT output:

Bigram	Count	Appears (Dummy)
The cat	2	1
cat the	1	0
cat on	1	1
on the	1	1
the mat	1	1

$$\underline{\text{Modified precision}} = \frac{1+0+1+1+1}{2+1+1+1+1} = \frac{4}{6}$$

---

n-gram:

$$p_1 = \frac{\sum_{unigram \in \hat{y}} Countclip(unigram)}{\sum_{unigram \in \hat{y}} Count(unigram)}$$

$$\vdots$$

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