Evaluation Metrics in NLP

- Accuracy
- Precision
- Recall
- F1-score
- → BLEU score (ex. Machine Translation)

NLP

Accuracy ←

- How many samples did your model predict w/ the correct label out of all samples?

Precision and Recall

		Prediction	
		Cat	Dog
Actual	Cat	15	35
	Dog	40	10

falsely predicted cat as dog

falsely predicted dog as cat

TN

TP	FN
FP	TN

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{15}{15+40} = 27\%$$

Inversely proportional

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{15}{15+35} = 30\%$$

F1 - score

$$2 * \frac{\text{precision} * \text{Recall}}{\text{precision} + \text{Recall}}$$

BLEU (bilingual evaluation understudy)

ROUGE

French : Le chat est sur le tapis

Human
Reference 1 : The cat is on the mat. ←
Reference 2 : There is a cat on the mat. ←

Model Output : the the the the the the the ←

Modified Precision (BLEU) = $\frac{2}{7}$

count_clip

count

$= \frac{4}{6} = \frac{2}{3}$

sum ↓ / sum ↓

BLEU w/ bigrams

Model Output : The cat the cat on the mat

bigrams	Count	count_clip
The cat	2	1
cat the	1	0
cat on	1	1
on the	1	1
the mat	1	1