# Survey on NMT Metric

## KU NMT

**이승준**
고려대학교 컴퓨터학과
2023/03/31

# 목차

# Why need NMT Metric?

| Sentence 1 | Sentence 2 | Similarity | BLEU | BLEURT |
|---|---|---|---|---|
| An arrest warrant claimed Bryant assaulted the woman 30 June at a hotel. | According to an arrest warrant, Bryant, 24, attacked a woman on 30 June. | 85.0 | 8.72 | 71.28 |
| Two white dogs are swimming in the water. | The birds are swimming in the water. | 16.0 | 66.42 | 43.38 |

## Evaluation Challenges

- Language variability
- Subjectivity of human evaluation
- Determining "good enough" quality
- Lack of universally accepted approach

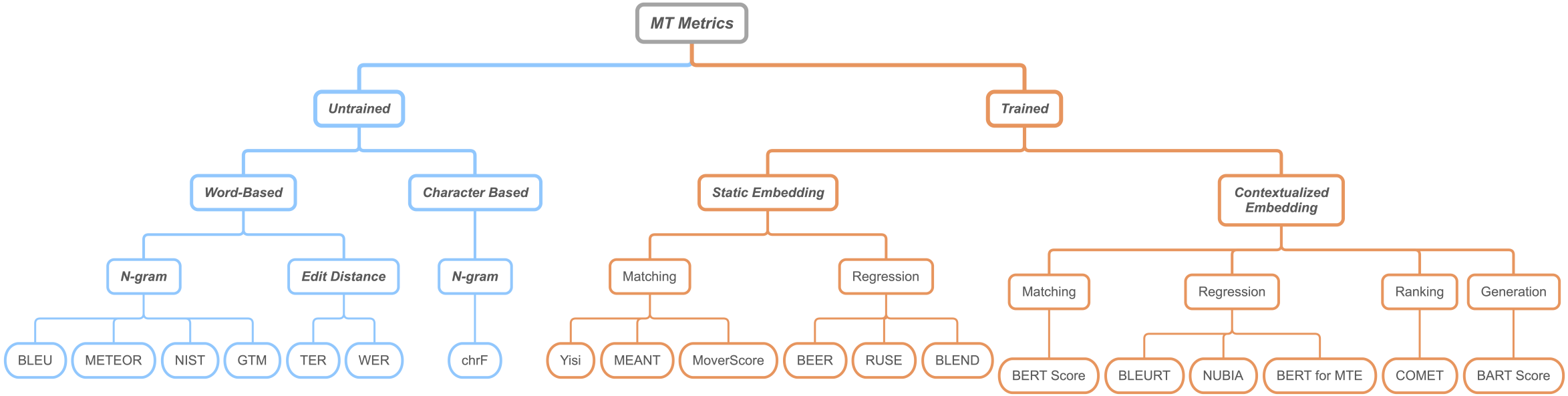
## Human Evaluation

- Types of human evaluation (adequacy, fluency, ranking, post-editing)
- Challenges and limitations of human evaluation
- The role of inter-annotator agreement (IAA)
- Challenges:
  - Maintaining consistency
  - Evaluating large translation units
  - High costs and substantial human labor

## Automatic Evaluation

- Benefits of automatic evaluation metrics
  - Cost-effective and minimal human labor
  - Comparing performance of multiple translation systems
  - Limitations in quality
- Traditional metrics and their limitations (e.g., BLEU)
- Deep learning-based metrics and their improvements

# Taxonomy of NMT Metric



# Taxonomy of NMT Metric

## Traditional Automated Evaluation Metrics

- Word-based Metrics (BLEU, NIST, TER, METEOR)
- Limitations in capturing semantic, grammatical diversity, and sentence structure

## Deep Learning-Based Evaluation Metrics

- Using embeddings from deep learning or Transformer-based language models (BERT, BART)
- Improved semantic similarity and higher correlation with human evaluation

# Taxonomy of NMT Metric

1. Matching
   - token or character level similarity
2. Regression
   - translation score annotated by human for the prediction
3. Ranking
   - learns to assign higher scores to better machine translation output than poor quality output.
4. Generation
   - high-quality hypothesis will be easily generated based on source or reference

## Definition

$$\textbf{Precision:} \frac{\#\ of\ matching\ n\text{-}grams}{\#\ of\ total\ n\text{-}grams\ in\ hypothesis} \quad \textbf{Recall:} \frac{\#\ of\ matching\ n\text{-}grams}{\#\ of\ total\ n\text{-}grams\ in\ reference}$$

- Source: Original Sentence
- Hypothesis: Machine Translation Output
- Reference: Correct Translation

```
Source: 배누르면 털 나와요
Hypothesis: It sheds when you brush it.
Reference: If you squeeze my stomach, the pubic hair will come out.
```

- Generation Metric (**Reference**-based):
  Hypothesis와 **Reference** 사이의 유사도
- Quality Estimation (**Source**-based):
  Hypothesis와 **Source** 사이의 유사도

## NMT Metric

$$\text{NMT Metric} \in \text{Generation Metric} \cup \text{Quality Estimation}$$

## BLEU: Bilingual Evaluation Understudy

- n-gram을 통한 순서쌍들이 얼마나 겹치는지 측정(precision)
- 문장길이에 대한 과적합 보정 (Brevity Penalty)
- 같은 단어가 연속적으로 나올때 과적합 되는 것을 보정(Clipping)

$$BLEU = \min\left(1, \frac{\text{hypothesis length ( 예측 문장 )}}{\text{reference length ( 실제 문장 )}}\right)\left(\prod_{i=1}^{4} \text{precision}_i\right)^{\frac{1}{4}}$$

## BLEU 예제

### 1. n-gram(1~4)을 통한 순서쌍들이 얼마나 겹치는지 측정(precision)

**Hypothesis**: 빛이 쐬는 노인은 완벽한 어두운곳에서 잠든 사람과 비교할 때 강박증이 심해질 기회가 훨씬 높았다
**Reference**: 빛이 쐬는 사람은 완벽한 어둠에서 잠든 사람과 비교할 때 우울증이 심해질 가능성이 훨씬 높았다

```
1-gram: 10/14,  2-gram: 5/13,  3-gram: 2/12,  4-gram: 1/11
```

$$\left(\prod_{i=1}^{4} \text{precision}_i\right)^{\frac{1}{4}} = \left(\frac{10}{14} \times \frac{5}{13} \times \frac{2}{12} \times \frac{1}{11}\right)^{\frac{1}{4}}$$

https://donghwa-kim.github.io/BLEU.html

## BLEU 예제

### 2. 같은 단어가 연속적으로 나올때 과적합 되는 것을 보정(Clipping)

**Hypothesis**: 배 누르면 털 나와요 털 배 아저씨 X 배 즙 아저씨
**Reference**: 털 배 사랑해요

- 1-gram Precision:

$$\frac{\text{일치하는 } 1\text{-gram 의 수(hypothesis)}}{\text{모든1-gram쌍 (hypothesis)}} = \frac{5}{9}$$

- (clipping) 1-gram precision: (hyp: 배: 3, 털: 2 vs. ref: 배: 1, 털: 1 )

$$\frac{\min\left(\#\text{ n-gram of hypothesis}, \#\text{ n-gram of reference}\right)}{\text{모든1-gram쌍 (hypothesis)}} = \frac{2}{5}$$

## 3. 문장길이에 대한 과적합 보정 (Brevity Penalty)

$$\min\left(1, \frac{\text{예측된 sentence의 길이(단어의 갯수 )}}{\text{true sentence의 길이(단어의 갯수 )}}\right)$$

Hypothesis의 길이가 Reference의 길이보다 길면 1, 작으면 0에 가까운 값이 나온다.
기계 번역 모델이 짧은 문장을 생성할 때 높은 BLEU 점수를 얻을 가능성 때문

## BLEU 예제

### 최종 BLEU Score

`Hypothesis`: 빛이 쐬는 노인은 완벽한 어두운곳에서 잠든 사람과 비교할 때 강박증이 심해질 기회가 훨씬 높았다
`Reference`: 빛이 쐬는 사람은 완벽한 어둠에서 잠든 사람과 비교할 때 우울증이 심해질 가능성이 훨씬 높았다

$$BLEU = \min\left(1, \frac{\text{output length ( 예 측 문장 )}}{\text{reference length ( 실제 문장 )}}\right)\left(\prod_{i=1}^{4} \text{precision}_i\right)^{\frac{1}{4}}$$

$$= \min\left(1, \frac{14}{14}\right) \times \left(\frac{10}{14} \times \frac{5}{13} \times \frac{2}{12} \times \frac{1}{11}\right)^{\frac{1}{4}}$$

## BLEU의 한계

- BLEU는 Recall를 고려하지 않는다. Only Precision
- 동의어(어간) 고려 X -> 다양한 형태론적 표현이 있는 언어 평가 어려움

## Precision vs. Recall

- 번역된 문장이 얼마나 정확하게 참조 문장을 포착하는지 평가
- **Precision** : Hypothesis 관점에서 올바르게 번역된 토큰(단어 또는 구)의 비율
- **Recall**: Reference 관점에서 모든 토큰 중 번역된 문장에서 올바르게 번역된 토큰의 비율
- BLEU를 보완한 **METEOR** -> **동의어** 그리고 **Recall** 고려

## METEOR: Metric for evaluation of translation with explicit ordering

- 기존 BLEU의 한계를 보완하기 위해 제안된 Metric
  - 어간과 동의어 고려
  - 재현율 (Recall): Precision과 Recall의 조화평균

$$P = \frac{\text{matched unigrams}}{\text{unigram in hypothesis}}, \; R = \frac{\text{matched unigrams}}{\text{unigram in reference}}$$

$$F\text{score} = \frac{10PR}{R + 9P}$$

- R과P의 조화평균: 번역의 정확성과 완전성을 동시에 고려

## TER: Translation Edit Rate

- Hypothesis와 Reference 사이의 편집거리를 측정 (편집률)
- reference의 평균 길이로 정규화된 최소 편집 작업
  - Multi reference 일 경우, 평균 길이를 가진 참조를 사용
  - 최소 수정 횟수를 계산할 때 가능한 적은 토큰을 반영하기 위함
- 번역 결과의 Edit은 이동, 대체, 삭제, 삽입

$$\text{TER} = \frac{\#\text{ of edits}}{\text{average }\#\text{ reference words}}$$

## chrF: Character n-gram F-score

- 단어 단위의 n-gram이 아닌, 문자 단위의 n-gram을 사용
- Recall과 Precision을 동시에 고려
- Tokenization에 종속적이지 않음
  - CJK (Chinese, Japanese, Korean) 언어에 적합
- stem과 morpheme errors가 발생하기 쉬운 언어에 대해 높은 성능을 보임

# Embedding-based Metric

- Word Embedding
  - MEANT
- Contextual Embedding
  - YiSi
  - BERT Score
  - Bart Score

## BERT Score

- MLM을 통해 얻은 Contextualized 임베딩을 이용하여 문장의 유사도를 측정
- BERT의 Token Embedding
- Hypothesis: $h = \{h_1, h_2, ..., h_n\}$ (tokenized)
- Reference: $r = \{r_1, r_2, ..., r_m\}$ (tokenized)
- Greedy Matching
- F1 score, Precision, Recall

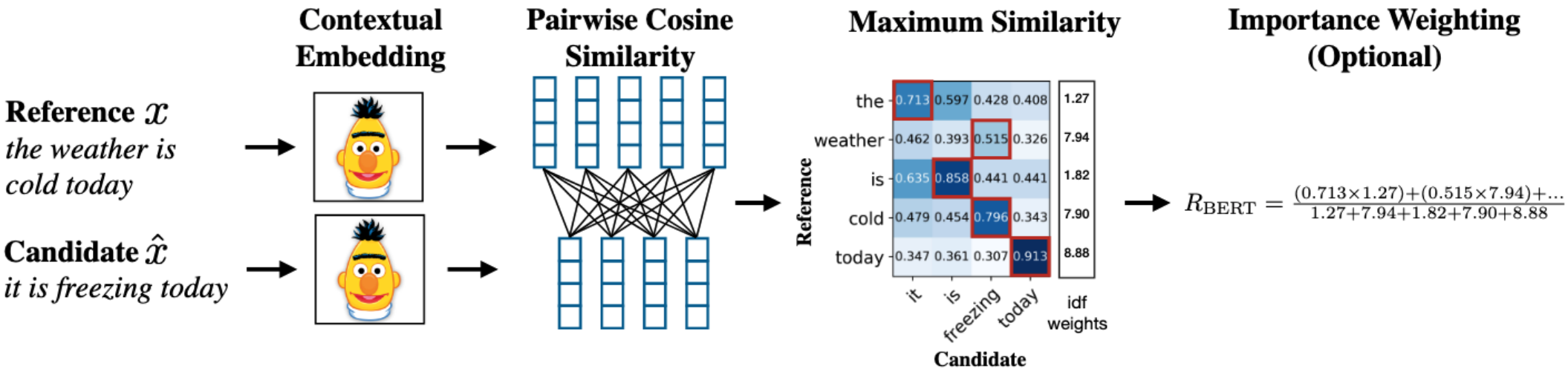
$$R_{\text{BERT}} = \frac{1}{|r|} \sum_{r_i \in r} \max_{h_j \in h} \mathbf{r}_i^\top \mathbf{h}_j, \; P_{\text{BERT}} = \frac{1}{|h|} \sum_{h_j \in h} \max_{r_i \in r} \mathbf{r}_i^\top \mathbf{h}_j$$

$$\text{BERT score} = F_{\text{BERT}} = 2 \frac{P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}}$$

# BERT Score

## Greedy Matching

- Hypothesis와 Reference의 각 토큰에 대해 가장 유사한 토큰을 찾음



## BERT Score

### Discussion

- BLEU에서 synonym을 고려하지 못하는 것과 대조적으로 BERT Score는 synonym(유사성)을 고려
- Context를 고려하기 때문에, BERT Score는 BLEU보다 더 정확한 평가를 할 수 있음
- 단점으로는 Hypothesis와 Reference의 토큰이 일치하지 않는 경우, 가장 유사한 토큰을 찾음

## Definition

- trained by machine learning or deep learning using labeled data.
- labeled data is WMT Direct Assessment (DA) dataset
  - human judgment for machine translation output
- It shows a higher correlation with human evaluation than other metrics

## Better Evaluation as Ranking

- training translation quality scores using labeled data to increase the resemblance to human ranking
- Features: Unigram statistics
  - word pair
  - function word
  - content word

$$\text{BEER score}(h, r) = \sum_i W_i \times \phi_i(h, r)$$

## BLEND

- combining multiple untrained metrics.
  - various perspective of hypothesis and reference
- Feature : 57 metric scores and DA evaluated by a human annotator
  - Lexical
  - Syntactic
  - Semantic
- Model: trained through an SVM regressor

## BERT for MTE

- BERT Score는 Embedding을 matching 방식으로 사용, BERT for MTE는 Regressor로 사용
- concatenating the hypothesis and reference
  - input it into BERT to obtain sentence-pair encoding
- final hidden state of [CLS] is used for the MLP regressor

$$\vec{v} = \text{BERT pair-encoder}\left([\text{CLS}]; \text{h}; [\text{SEP}]; \text{r}; [\text{SEP}]\right)$$

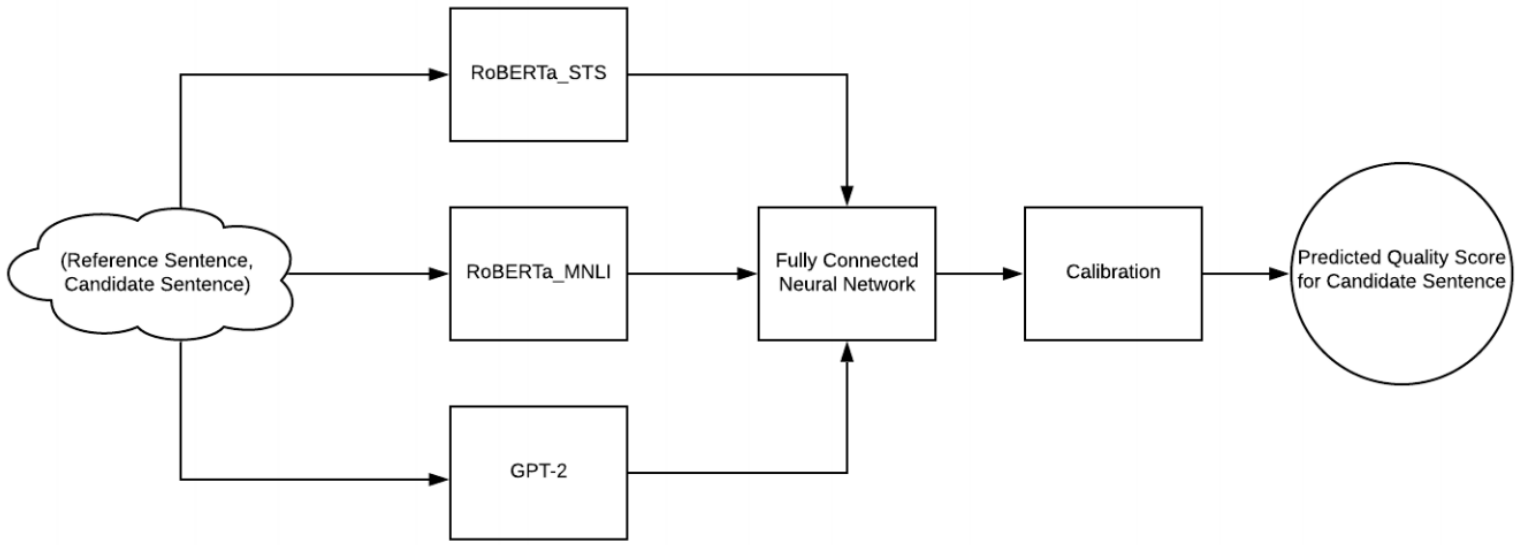
$$\text{BERT for MTE} = \text{MLP-Regressor}\left(\vec{v}_{[\text{CLS}]}\right)$$

## BLEURT

- multi domain에서 좋은 성능을 목표로 함
- Data Augmentation for scarcity of human ratings
  - mask-filling
  - back-translation
  - dropping words
- train regression models to predict human ratings
- 현재까지 가장 좋은 성능을 보이는 대표적인 생성 metric

# NUBIA

- combination of three modules for translation evaluation
  - neural feature extractor
  - aggregator
  - calibrator



## NUBIA

- neural feature extractor
  - semantic similarity (STS-B)
  - logical entailment (MNLI)
  - sentence intelligibility (ppl of GPT-2)
- aggregator: regression model to predict human evaluation
- calibrator: normalize the scores to the range of 0 to 1

# COMET: Cross-lingual Optimized Metric for Evaluation of Translation

- multilingual machine translation using ranking and regression
- estimator and translation ranking model based on human determination
- training objectives
  - estimator: regression
  - translation ranking model: minimize the distance between the ranking of the human and the ranking of the machine translation

# Evaluation of MT Metrics

- 좋은 성능의 메트릭이란?: 두 변수 간의 선형 상관관계를 측정

## Pearson correlation coefficient

- the combination of the two variables is a normal distribution
- the two variables have a linear relationship

$$\rho_{xy} = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\sqrt{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}}$$

## Spearman correlation coefficient

- between two variables is the same as the Pearson correlation calculated by rank.

$$r_{xy} = \frac{\sum_{i=1}^{n} \left(r_{x_i} - \bar{r}_x\right)\left(r_{y_i} - \bar{r}_y\right)}{\sqrt{\sum_{i=1}^{n} \left(r_{x_i} - \bar{r}_x\right)^2} \sqrt{\sum_{i=1}^{n} \left(r_{y_i} - \bar{r}_y\right)^2}}$$

## Kendall's $\tau$ coefficient

- computes the number of concordant and discordant of the ordered pair
- the variable does not follow a normal distribution

# Tips for NMT Evaluation

## BLEU Usage

- 가장 많이 사용되는 metric임으로 여러 variant 존재
  - nltk, sacrebleu, google, moses -> **SacreBLEU** (=huggingface's bleu)
- tokenizing 방법에 따라 성능이 달라짐
  - word_tokenize, moses -> 영어: 13a, 한국어: ko-mecab

```
pip install "sacrebleu[ko]"
```

```
import sacrebleu
sacrebleu.corpus_bleu(hypotheses=hypo, references=ref, tokenize='ko-mecab')
```

## Sentence-level BLEU vs. Corpus-level BLEU

- Sentence-level BLEU
  - 각 문장에 대해 BLEU를 계산 (문장 단위)
  - 문장 단위 성능을 비교 하고자 할 때 사용
- Corpus-level BLEU
  - 모든 문장(=전체 코퍼스, 문서 단위)에 대해 BLEU를 계산
  - 모든 문장의 n-그램 일치 횟수를 누적하여 계산
  - 기계 번역 모델을 비교하거나 모델의 전체 성능을 평가할 때 사용

```python
import sacrebleu

sacrebleu.sentence_bleu(translated_sentence, [reference_sentence])
# 번역된 문장과 참조 문장
translated_sentence = "이것은 예제 문장입니다."
reference_sentence = "이것은 샘플 문장입니다."

# 번역된 문장 목록과 참조 문장 목록
translated_sentences = ["이것은 예제 문장입니다.", "안녕하세요, 반갑습니다."]
reference_sentences = [["이것은 샘플 문장입니다."], ["안녕하세요, 만나서 반가워요."]]

# Corpus-level BLEU 계산
sacrebleu.corpus_bleu(translated_sentences, reference_sentences)
```

# Tips for NMT Evaluation

## 사소한 팁들 (1)

- BLEURT는 한국어도 사용이 가능하다.
  Currently, BLEURT-20 was tested on 13 languages: Chinese, Czech, English, French, German, Japanese, Korean, ...(these are languages for which we have held-out ratings data)

## 사소한 팁들 (2)

- BERT Score에서 한국어를 사용하고자 할 경우, `lang=others`로 설정
- Ko-BERTScore도 존재 한다.

---

BLEURT git: https://github.com/google-research/bleurt
Ko-BERTScore git: https://github.com/lovit/KoBERTScore

## 사소한 팁들 (3)

- 완벽한 Metric이란 존재 하지 않는다.
  - 각각의 장점 및 capture할 수 있는 특징이 다름
  - BLEU가 Semantic Similarity를 잘 캡쳐 하지 못한다고 해서 안 좋은 Metric이 아니다.
  - 각각의 Metric을 여러 다방면으로 활용하여 NMT 성능을 평가해야 한다.
- 일반적으로 BLEU, METEOR, chrF, TER, BLEURT, BERTScore 등을 사용
  - 단, 연구의 흐름은 계속 해서 바뀌니, 본인의 연구에 적합한 메트릭을 선택하는 것이 중요
- 최근 좋은 성능을 보이는 metric은 단연, COMET
  - Reference-free, QE, DA with Regression & ranking

# Thank you