MT Evaluation
Automatic Evaluation

Multilingual NLP — lecture n°5

Guillaume Wisniewski

guillaume.wisniewski@u-paris.fr

UPCité Master in Computational Linguistics

November 4th, 2025

Our Goal

- evaluate translation quality...
- ...automatically

An Optimistic Note

- ambitious to try to automate a task that humans themselves cannot reliably perform

Loss Function for Translation Evaluation

Loss Function for Translation Evaluation

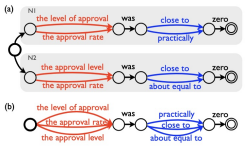
Reminder

- loss function: $\ell : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}^+$
- ↪ comparison between a translation hypothesis and a gold translation
- ↪ $\mathcal{Y}$ = sentence / paragraph / document / set of translations
- cornerstone of ML

For Translation

- very limited evaluation: only measures the "overall" quality of a translation (does not identify or explain errors)
- does a single reference even exist?

Counting Reference Translations



- evaluation metric HyTER introduced by Dreyer and Marcu 2012
- ↪ never used in practice (except in Apidianaki et al. 2018 😀)
- interesting side product: automatic generation of all possible references for a given sentence
- ↪ from a set of human translations of a given sentence
- empirically, naturally occurring sentences have billions of valid translations:
- ↪ median number of reference paths per sentence produced by a single annotator ranges from 7.7×10^5 up to 5.9×10^8

Part I

Historical Metrics

The Prehistory of Evaluation

- simplest idea: same word in the reference and in the translation hypothesis
  ↪ word matching
- 2 key criteria:
  1. fluency
  2. adequacy

  ⇒ same as for human evaluation
- intuition: match words (adequacy) + word sequences (fluency)

The Bleu Score

- introduced by Papineni et al. 2002
- main metric for many years
- "cooking recipe" with no theoretical justification

  ↪ no explanation of what was tested
  ↪ reported to correlate with human judgements on a dataset that nobody has access to
- metric created with a simple goal: measure the impact of modifying a system

  ↪ not designed to compare systems
  ↪ not designed to measure translation quality

Key Takeaways

good workers know their tools

↪ 97.3 % (I counted) of the discussions about BLEU could have been avoided if people had read the original paper and understood the limitations it explicitly states

The BLEU score: intuition

References

1. It is a guide to action that ensures that the military will forever heed Party commands.
2. It is the guiding principle which guarantees the military forces always being under the command of the Party.
3. It is the practical guide for the army always to heed the directions of the party

- **hyp n°1** It is to insure the troops forever hearing the activity guidebook that party direct.
- **hyp n°2** It is a guide to action which ensures that the military always obeys the command of the party.

The BLEU score: intuition

References

1. It is a guide to action that ensures that the military will forever heed Party commands.
2. It is the guiding principle which guarantees the military forces always being under the command of the Party.
3. It is the practical guide for the army always to heed the directions of the party

- **hyp n°1** It is to insure the troops forever hearing the activity guidebook that party direct.
- **hyp n°2** It is a guide to action which ensures that the military always obeys the command of the party.

The BLEU score: intuition

References

1. It is a guide to action that ensures that the military will forever heed Party commands.
2. It is the guiding principle which guarantees the military forces always being under the command of the Party.
3. It is the practical guide for the army always to heed the directions of the party

- **hyp n°1** It is to insure the troops forever hearing the activity guidebook that party direct.
- **hyp n°2** It is a guide to action which ensures that the military always obeys the command of the party.

The BLEU score: intuition

At the end:

- **hyp n°1** It is to insure the troops forever hearing the activity guidebook that party direct.
- **hyp n°2** It is a guide to action which ensures that the military always obeys the command of the party.

↪ more common, longer n-grams in the second hypothesis ⇒ better translation

The BLEU Score

- BLEU evaluates the quality of a machine translation by comparing it to one or more reference translations.
- It combines:
  - modified n-gram precision
  - a brevity penalty (to discourage overly short translations)
- ⚠⚠ BLEU is not the average of sentence-level scores — it is computed from aggregated n-gram counts over the whole corpus. ⚠⚠
  ⇒ corpus-level metric
- such an adhoc metric 🤨

Formalization

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{1-\frac{r}{c}} & \text{if } c \leq r \end{cases}$$

- p_n: modified n-gram precision
- w_n: weights (usually $w_n = \frac{1}{N}$)
- c: length of candidate translation
- r: length of reference translation(s)

Example

References
R1: `the cat is on the mat`
R2: `there is a cat on the mat`
Candidates
C1: `the cat sat on the mat`
C2: `there is cat on mat`

Example

References
R1: `the cat is on the mat`
R2: `there is a cat on the mat`
Candidates
C1: `the cat sat on the mat`
C2: `there is cat on mat`

Step 1: Modified n-gram counts (aggregated over the corpus)

- Total candidate tokens: $c = 6 + 5 = 11$
- Total reference tokens: $r = 6 + 7 = 13$

Example

References
R1: `the cat is on the mat`
R2: `there is a cat on the mat`
Candidates
C1: `the cat sat on the mat`
C2: `there is cat on mat`

Step 1: Modified n-gram counts (aggregated over the corpus)

- Total candidate tokens: $c = 6 + 5 = 11$
- Total reference tokens: $r = 6 + 7 = 13$

$$p_1 = \frac{9}{11} \approx 0.8182$$
$$p_2 = \frac{5}{10} = 0.5$$
$$p_3 = \frac{2}{9} \approx 0.2222$$
$$p_4 = \frac{1}{8} = 0.125$$

Example

References
R1: `the cat is on the mat`
R2: `there is a cat on the mat`
Candidates
C1: `the cat sat on the mat`
C2: `there is cat on mat`

Step 2: Brevity Penalty

$$\text{BP} = e^{1-r/c} = e^{1-\frac{13}{11}} \approx 0.8337$$

Example

References
R1: `the cat is on the mat`
R2: `there is a cat on the mat`
Candidates
C1: `the cat sat on the mat`
C2: `there is cat on mat`

Step 3: Corpus BLEU-4

$$\text{BLEU} = \text{BP} \cdot \exp\left(\frac{1}{4}\sum_{n=1}^{4} \log p_n\right) \approx 0.8337 \times 0.3754 \approx 0.3130$$

An Important Detail

- impact of tokenizer, especially for morphologically-rich languages
- see Benjamin Marie's Post ("Science Left Behind")

The Limitations of BLEU

- a very naïve approach (at least, when used outside its intended scope 🤔)
- so many criticisms of BLEU that I stopped counting
- common joke in the MT community: more papers on new metrics than on translation itselfn 🤣
- synthesis of criticisms (+ new strong arguments) in Callison-Burch, Osborne, and Koehn 2006
  1. an improvement in BLEU does not necessarily correspond to an improvement in human judgements: *"We show that an improved Bleu score is neither necessary nor sufficient for achieving an actual improvement in translation quality."*
  2. there are thousands of transformations of a sentence that do not change the BLEU score

  ↪ all permutations that keep the matching n-grams unchanged

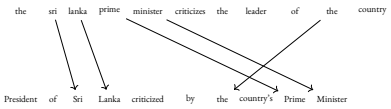
Consequence

- a natural consequence: numerous attempts to propose alternatives
  ↪ *Naturam expellas furca, tamen usque recurret*, Horace, Ep. 2, 10, 24
- however, criticising is easy — designing a better metric is much harder
  ↪ *La critique est aisée, et l'art est difficile.*, Philippe Néricault Destouches, *Le Glorieux*, II, 5
- to spare you the boredom: I will only cover the two that are still used (IMHO)

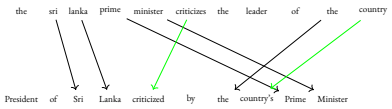
Meteor

- Metric for Evaluation of Translation with Explicit Ordering Banerjee and Lavie 2005
- rely on unigram recall and precision
  ↪ align/match words of the hypothesis and of the reference
- matching takes into account translation variability via word inflection variations, synonymy and paraphrasing matches
- direct penalty for word order: how fragmented is the matching?
- weighted metrics: the weights of the difference components can be optimized to improve correlation with human judgments

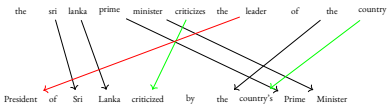
Alignment



Alignment



Alignment



↪ different kind of matches, with different weights

↪ optimal search is NP-complete (but clever search with pruning is very fast and has near optimal results)

The Full Meteor Metric

Fragmentation

- to take fluency into account
- frag = $\frac{\#\text{'group' of word that are in matches}-1}{\#\text{words in matches}-1}$

Final Score

- discounting factor: DF = $\gamma \times \text{frag}^{\beta}$
- F_α score: $F_\alpha = \frac{P \times R}{\alpha \cdot P + (1-\alpha) \cdot R}$
- original parameter settings: $\alpha = 0.9, \beta = 3.0, \gamma = 0.5$
  - ↪ now parameters are estimated on training data (to maximize correlation with human judgments)
- final score: $F_\alpha \cdot (1 - \text{DF})$

Meteor example

Example

- **Reference:** "the Iraqi weapons are to be handed over to the army within two weeks"
- **Hypothesis:** "in two weeks Iraq's weapons will give army"

Meteor example

Example

- **Reference:** "the Iraqi weapons are to be handed over to the army within two weeks"
- **Hypothesis:** "in two weeks Iraq's weapons will give army"

↪ Precision =

Meteor example

Example

- **Reference:** "the Iraqi weapons are to be handed over to the army within two weeks"
- **Hypothesis:** "in two weeks Iraq's weapons will give army"

↪ Precision = $\frac{5}{8}$ = 0.625

Meteor example

Example

- **Reference:** "the Iraqi weapons are to be handed over to the army within two weeks"
- **Hypothesis:** "in two weeks Iraq's weapons will give army"

↪ Precision = $\frac{5}{8}$ = 0.625
↪ Recall =

Meteor example

Example

- **Reference:** "the Iraqi weapons are to be handed over to the army within two weeks"
- **Hypothesis:** "in two weeks Iraq's weapons will give army"

↪ Precision = $\frac{5}{8}$ = 0.625
↪ Recall = $\frac{5}{14}$ = 0.357

Meteor example

Example

- **Reference:** "the Iraqi weapons are to be handed over to the army within two weeks"
- **Hypothesis:** "in two weeks Iraq's weapons will give army"

↪ Precision = $\frac{5}{8}$ = 0.625
↪ Recall = $\frac{5}{14}$ = 0.357
↪ Fragmentation =

Meteor example

Example

- **Reference:** "the Iraqi weapons are to be handed over to the army within two weeks"
- **Hypothesis:** "in two weeks Iraq's weapons will give army"

↪ Precision = $\frac{5}{8}$ = 0.625
↪ Recall = $\frac{5}{14}$ = 0.357
↪ Fragmentation = $\frac{3-1}{5-1}$ = 0.5

Meteor example

Example

- **Reference:** "the Iraqi weapons are to be handed over to the army within two weeks"
- **Hypothesis:** "in two weeks Iraq's weapons will give army"

↪ Precision = $\frac{5}{8}$ = 0.625
↪ Recall = $\frac{5}{14}$ = 0.357
↪ Fragmentation = $\frac{3-1}{5-1}$ = 0.5

Weighted combination: 0.3498

The problem with Meteor

- easier to interpret?
- rely on external resources (e.g. paraphrase table) that are not always available
- computational cost
- fuzzy matches have very low impact

Why not training MT metrics?

You can!

and it has been done many times

But...

- which objective function → human judgment are not consistent (more on this latter)
- you need training data!
  - ↪ only available in a few languages
  - ↪ we want to be able to evaluate translations between any languages
- Comets → sBERT ⊕ fine-tuned for predicting human scores

↪ best correlation with human judgments

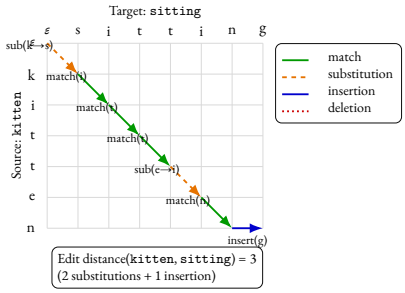
TER

- Translation Edit (Error) Rate (Snover et al. 2009)
- edit-based measure: minimal number of edits to transform hypothesis into reference
- add the notion of 'block movements' as single edit operation
- exact matches only — extension TERp: near-matches ⊕ weights
- rough estimate of post-editing effort: minimal number of transformation to "correct" the translation hypothesis

TER

- Translation Edit (Error) Rate (Snover et al. 2009)
- edit-based measure: minimal number of edits to transform hypothesis into reference
- add the notion of 'block movements' as single edit operation but NP-complet
- exact matches only — extension TERp: near-matches ⊕ weights
- rough estimate of post-editing effort: minimal number of transformation to "correct" the translation hypothesis

Edit Distance (Levenshtein) — Visual Intuition



⇒ dynamic programming: explore an exponential number of combinations in a polynomial time

The chrF (character n-gram F-score) Metric

- introduced by Popović 2015
- char-based and not word-based

↪ overlap of character n-grams between translation hypothesis and gold translation

- ↪ word matching is too rigid (especially for morphologically rich languages)
- ↪ character n-grams capture morphology and inflection more effectively
- Combines character precision and recall using a standard F_β-score (typically $\beta = 2$ to give more weight to recall).

Why chrF?

- Major advantages over BLEU:
  - better correlation with human judgements, particularly for morphologically rich languages
  - less sensitive to tokenisation and exact wording
  - captures partial matches (e.g. morphological variants, spelling variations)
- Works well for:
  - languages with rich morphology
  - low-resource settings where lexical variation is high
  - evaluating systems that produce paraphrastic translations
- As usual: chrF++: combines character n-grams with word n-grams ⇒ improved robustness.

chrF: Corpus-Level Example

References
R1: `The cats are playing outside`
R2: `A dog is sleeping on the sofa`

Candidates
C1: `The cat is playing outdoors`
C2: `A dog sleeps on the couch`

Step 1: Extract character n-grams (e.g. n = 1–6) for the whole corpus → counts are **summed across all sentences** before computing precision and recall.

Example with character 3-grams only (illustrative):

Total matches = 31, Total candidate 3-grams = 46, Total reference 3-grams = 49

$$P = \frac{31}{46} \approx 0.674, \qquad R = \frac{31}{49} \approx 0.633$$

And:

$$F_2 = \frac{(1+\beta^2)\cdot P \cdot R}{(\beta^2 \cdot P) + R} = \frac{5 \cdot 0.674 \cdot 0.633}{4 \cdot 0.674 + 0.633} \approx 0.642$$

Bilan de toutes ces recherches

Principle

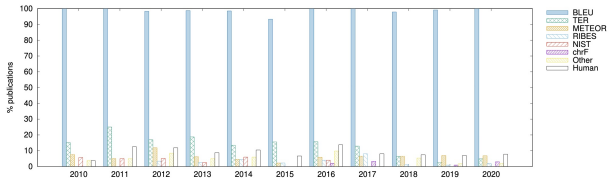
- (Marie, Fujita, and Rubino 2021): Scientific Credibility of Machine Translation Research: A Meta-Evaluation of 769 Papers
- meta-evaluation of 769 peer-reviewed MT papers (2010–2020)

↪ manual annotation of each paper for evaluation practices use of automatic metrics, human evaluation, statistical testing, references, ...

Key Findings

- Limited human evaluation
- Statistical testing lacking
- 98.8% of the annotated papers reported BLEU scores
- 74.3% of papers use only BLEU
- At least 108 new metrics were introduced between 2010 and 2020, claiming to be better than BLEU but 89% of them have never been used in any Anthology paper

The picture that says it all...



and then...

for anyone who needs to (re)acquaint themselves with French culture

and then...

word embeddings/Transformer/BERT/(L)LM arrived

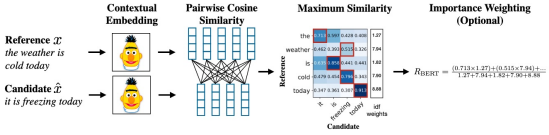
Key Concept

- word embeddings: capture semantic ⊕ syntactic information about a sentence
- ↪ all the information we need to compare a reference and a translation hypothesis
- but we need not combine/compare them!

BERTScore (Zhang et al. 2020)

- measure semantic similarity to references using contextual embeddings
- simple approach:
  - Encode tokens in candidate (C) and reference (R) with a pre-trained model (e.g., BERT/RoBERTa).
  - Compute pairwise cosine similarities; greedily match each token to its most similar counterpart.
  - Alternative: use Earth Moving Distance → (Zhao et al. 2019)
  - Aggregate into Precision, Recall, and F_1 (optionally IDF-weighted and baseline-rescaled).
- **Pros:** Captures paraphrases; correlates better with human judgements than n-gram metrics; multilingual variants available.
- **Cons:** Sensitive to model/layer choice; can be misled by antonyms in similar contexts; computationally heavier than BLEU.

With a Picture



A Personal Note

- simple idea
- so much unanswered question in the paper:
- ↪ impact of tokenization (is matching subword making sense?)
- ↪ 1:1 or n:1 or 1:n matching?
- ↪ representations anistropy

BLEURT: Core Idea

BLEURT (Sellam, Das, and Parikh 2020): evaluation metric trained to predict human judgements of text quality.

1. Start from a pre-trained language model (BERT or RoBERTa).
2. ↪ use `[CLS]` to represent sentences
3. Add synthetic noisy examples and fine-tune on them (noisy pre-training).
4. Fine-tune on human-annotated evaluation datasets (e.g., WMT DA scores).
5. ↪ single scalar score predicting text quality.

COMET: Core Principle

COMET ((<empty citation>) → COMET-22/23 versions) is a family of learned, reference-based and reference-free evaluation metrics

Principle

1. Encode source, candidate (and optionally reference) with a multilingual Transformer encoder (e.g., XLM-R).
2. Extract contextual embeddings for each sequence.
3. Apply a pooling layer (often mean-pooling + attention pooling) to derive sentence-level representations.
4. Combine embeddings (e.g., concatenation of source–candidate–reference vectors).
5. Feed into a regression network trained to predict human quality scores.

COMET: Variants

- **Reference-based** (e.g., `comet-wmt20`, `comet-22`):
  - (Rei et al. 2022)
  - Input: source + candidate + reference
  - Best correlation with human judgements
- **Reference-free** (e.g., `comet-qe-da`):
  - (Wieting et al. 2019)
  - Input: source + candidate
  - Quality Estimation (QE) setting
- **Task-specific fine-tuned models**:
  - Fine-tune on domain/human ratings for domain adaptation
  - E.g., medical, legal, conversational MT

COMET: Performance Across Languages

General Trends from WMT Shared Tasks (2020–2023)

- COMET consistently achieves the highest correlation with human judgements across most language pairs.
- Best performance on high-resource languages (e.g., En–De, En–Fr, En–Es) due to abundant training data.
- Robust for distant or low-resource languages (e.g., En–Zh, En–Hi), but correlation slightly lower due to:
  - data sparsity
  - encoder pre-training biases
  - cultural/linguistic differences not captured by the model

Reference-Free COMET (QE):

- Strong performance for high-resource pairs
- Larger drop for low-resource languages than reference-based COMET

COMET: Correlation with Human Judgements

Pearson correlation with human evaluation (WMT20–WMT23)

Language Pair	BLEU	BERTScore	COMET
En → De	0.55	0.72	**0.90**
En → Zh	0.35	0.60	**0.82**
Zh → En	0.40	0.67	**0.88**
En → Ru	0.50	0.69	**0.87**

COMET-MQM (MQM annotation correlation)

Language Pair	COMET-DA	COMET-MQM
En → De	0.88	**0.94**
En → Zh	0.78	**0.90**
En → Ja	0.75	**0.88**
En → Hi	0.70	**0.84**

Takeaway: COMET typically reaches **0.85–0.95** correlation with humans, and **COMET-MQM** is even more aligned with fine-grained human error ratings.

Part III

From Embeddings to Prompting

GEMBA

- **GEMBA** (General Evaluation of Machine Translation Benchmarking with Agents)

↪ introduced in Kocmi, Avramidis, et al. 2023

- Evaluation relies on structured prompts guiding the LLM to act as a consistent reviewer.
- Produces fine-grained scoring (e.g., adequacy, fluency, style) or holistic ratings.
- Designed to mimic expert human judgement while scaling to large evaluation sets.

GEMBA Prompt (FR → EN)

Prompt "Principle"

> You are an expert MT evaluator. Rate the English translation of the French source.
> **Source (FR):**
> "Le projet a été terminé en avance grâce à une excellente collaboration."
> **MT Output (EN):**
> "The project was finished early thanks to an excellent collaboration."
> Evaluate from 1–5:
>
> - Adequacy (meaning preserved?)
> - Fluency (natural English?)
>
> Give scores + a one-sentence justification.

⇒ better output: justification (but can't be analysed readily)

Auto-MQM: Automated MQM Scoring

- Auto-MQM uses LLMs to generate MQM-style error annotations and severity scoring.
  ↪ introduced in Fernandes et al. 2023
- Labels errors by category (accuracy, fluency, terminology, style, etc.).
- Reduces human effort in detailed MQM annotation campaigns.
- Evaluation with and without reference
- Same idea: GEMBA-MQM (Kocmi and Federmann 2023)

Auto-MQM Prompt (FR → EN)

Prompt "Idea"

You are an MQM evaluator. Identify translation errors in the MT output.
Source (FR):
"Le projet a été terminé en avance grâce à une excellente collaboration."
MT Output (EN):
"The project was finished sooner thanks to an excellent collaboration."
Instructions:

- Highlight text spans containing errors.
- For each error: provide category + severity (Minor/Major/Critical).
- Use MQM categories only.

Respond in a table with columns: Span | Error Type | Severity.

More Precise Results

	MT System	
	DeepL	ChatGPT
# texts	35	25
# gold errors	399	193
long prompt		
# pred. errors	384	224
precision	0.792 ± 0.0396	0.47 ± 0.0989
recall	0.653 ± 0.0488	0.57 ± 0.107
F_1	0.707 ± 0.0393	0.496 ± 0.0933
% correctly labeled	64.1 %	45.3 %
short prompt		
# pred. errors	417	—
precision	0.745 ± 0.0575	—
recall	0.671 ± 0.0505	—
F_1	0.702 ± 0.0531	—
% correctly labeled	46.9 %	—

- results detailled in Minder, Wisniewski, and Kübler 2025
- Prompt to identify errors in translation of NLP articles
- Compare errors identified by an expert and errors identified by an LLM
  ↪ at least one character in common

Scientific Questions: LLM-as-a-Judge

- **Reliability**: How consistent are LLM judgements across prompts, languages and model versions?
- **Bias**: Do LLM judges favour outputs resembling their own generation patterns?
- **Alignment**: Should LLM evaluation replicate human MQM standards or define its own?
- **Meta-Evaluation**: How do we evaluate the evaluator? What is the ground truth of judgement?
- **Ethics & Governance**: Are opaque LLM-based rankings acceptable for high-stakes MT decisions?

Part IV

References

- Apidianaki, Marianna et al. (June 2018). "Automated Paraphrase Lattice Creation for HyTER Machine Translation Evaluation". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Ed. by Marilyn Walker, Heng Ji, and Amanda Stent. New Orleans, Louisiana: Association for Computational Linguistics, pp. 480–485.
- Banerjee, Satanjeev and Alon Lavie (June 2005). "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments". In: *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ed. by Jade Goldstein et al. Ann Arbor, Michigan: Association for Computational Linguistics, pp. 65–72.
- Callison-Burch, Chris, Miles Osborne, and Philipp Koehn (Apr. 2006). "Re-evaluating the Role of Bleu in Machine Translation Research". In: *11th Conference of the European Chapter of the Association for Computational Linguistics*. Ed. by Diana McCarthy and Shuly Wintner. Trento, Italy: Association for Computational Linguistics, pp. 249–256.

- Dreyer, Markus and Daniel Marcu (June 2012). "HyTER: Meaning-Equivalent Semantics for Translation Evaluation". In: *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Eric Fosler-Lussier, Ellen Riloff, and Srinivas Bangalore. Montréal, Canada: Association for Computational Linguistics, pp. 162–171.
- Fernandes, Patrick et al. (Dec. 2023). "The Devil Is in the Errors: Leveraging Large Language Models for Fine-grained Machine Translation Evaluation". In: ed. by Philipp Koehn et al. Singapore, pp. 1066–1083.
- Kocmi, Tom, Eleftherios Avramidis, et al. (Dec. 2023). "Findings of the 2023 Conference on Machine Translation (WMT23): LLMs Are Here but Not Quite There Yet". In: ed. by Philipp Koehn et al. Singapore, pp. 1–42.
- Kocmi, Tom and Christian Federmann (June 2023). "Large Language Models Are State-of-the-Art Evaluators of Translation Quality". In: ed. by Mary Nurminen et al. Tampere, Finland: European Association for Machine Translation, pp. 193–203.

- Marie, Benjamin, Atsushi Fujita, and Raphael Rubino (Aug. 2021). "Scientific Credibility of Machine Translation Research: A Meta-Evaluation of 769 Papers". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Ed. by Chengqing Zong et al. Online: Association for Computational Linguistics, pp. 7297–7306.
- Minder, Joachim, Guillaume Wisniewski, and Natalie Kübler (June 2025). "Testing LLMs' Capabilities in Annotating Translations Based on an Error Typology Designed for LSP Translation: First Experiments with ChatGPT". In: *Proceedings of Machine Translation Summit XX: Volume 1*. Ed. by Pierrette Bouillon et al. Geneva, Switzerland: European Association for Machine Translation, pp. 190–203.
- Papineni, Kishore et al. (July 2002). "Bleu: a Method for Automatic Evaluation of Machine Translation". In: *Proceedings of the 40th Annual Meeting of the Association for*

Computational Linguistics. Ed. by Pierre Isabelle, Eugene Charniak, and Dekang Lin. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, pp. 311–318.
- Popović, Maja (Sept. 2015). "chrF: character n-gram F-score for automatic MT evaluation". In: *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Ed. by Ondřej Bojar et al. Lisbon, Portugal: Association for Computational Linguistics, pp. 392–395.
- Rei, Ricardo et al. (Dec. 2022). "COMET-22: Unbabel-IST 2022 Submission for the Metrics Shared Task". In: *Proceedings of the Seventh Conference on Machine Translation (WMT)*. Ed. by Philipp Koehn et al. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, pp. 578–585.
- Sellam, Thibault, Dipanjan Das, and Ankur Parikh (July 2020). "BLEURT: Learning Robust Metrics for Text Generation". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, pp. 7881–7892.

- Snover, Matthew et al. (Mar. 2009). "Fluency, Adequacy, or HTER? Exploring Different Human Judgments with a Tunable MT Metric". In: *Proceedings of the Fourth Workshop on Statistical Machine Translation*. Ed. by Chris Callison-Burch et al. Athens, Greece: Association for Computational Linguistics, pp. 259–268.
- Wieting, John et al. (July 2019). "Beyond BLEU: Training Neural Machine Translation with Semantic Similarity". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Anna Korhonen, David Traum, and Lluís Màrquez. Florence, Italy: Association for Computational Linguistics, pp. 4344–4355.
- Zhang, Tianyi et al. (2020). *BERTScore: Evaluating Text Generation with BERT*.
- Zhao, Wei et al. (Nov. 2019). "MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by Kentaro Inui et al. Hong Kong, China: Association for Computational Linguistics, pp. 563–578.