# BLEU Meets COMET: Combining Lexical and Neural Metrics Towards Robust Machine Translation Evaluation

**Taisiya Glushkova, Chrysoula Zerva, André F. T. Martins**

{taisiya.glushkova, chrysoula.zerva, andre.t.martins}@tecnico.ulisboa.pt

✔ **COMET outperforms lexical metrics** (BLEU, chrF) for MT evaluation

✘ … but it is **less sensitive to** specific **error patterns**

- ▸ e.g. changes in numbers, named entities, sentence polarity, …

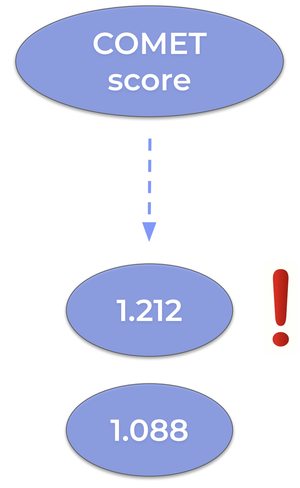
💡 What if we combine them and **enhance COMET** with some **lexical information**?

## Proposed Approaches:

- ▷ **Ensemble** sentence-level metrics
- ▷ Use **BLEU & chrF sentence-level scores** as extra features through a bottleneck layer
- ▷ Use **subword-level** quality features based on TER **alignements** between target and reference

## Problem

| | |
|---|---|
| Source | You only have **20** kilometres **of** range. |
| Reference | Só tem **20** quilómetros **de** autonomia.<br>(You only have **20** kilometres **of** range.) |
| Critical Error | Só tem **200** quilómetros de autonomia.<br>(You only have **200** kilometres of range.) |
| Non-Critical Error | Só tem 20 quilómetros **da** autonomia.<br>(You only have 20 kilometres **of** range.) |

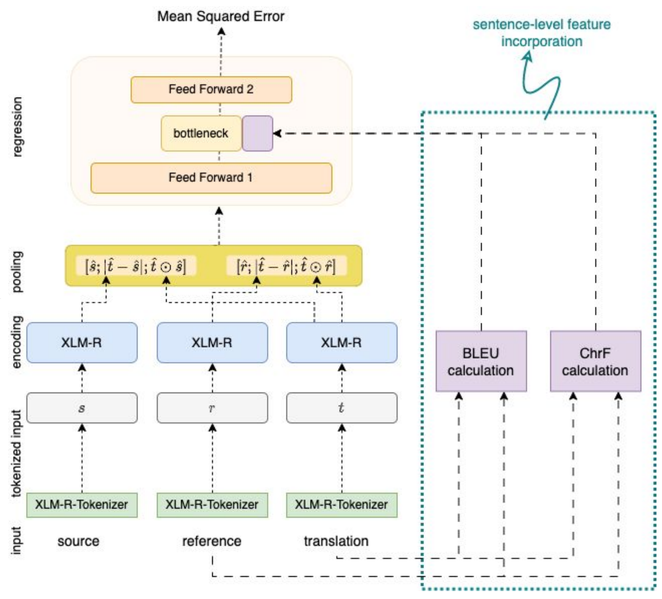


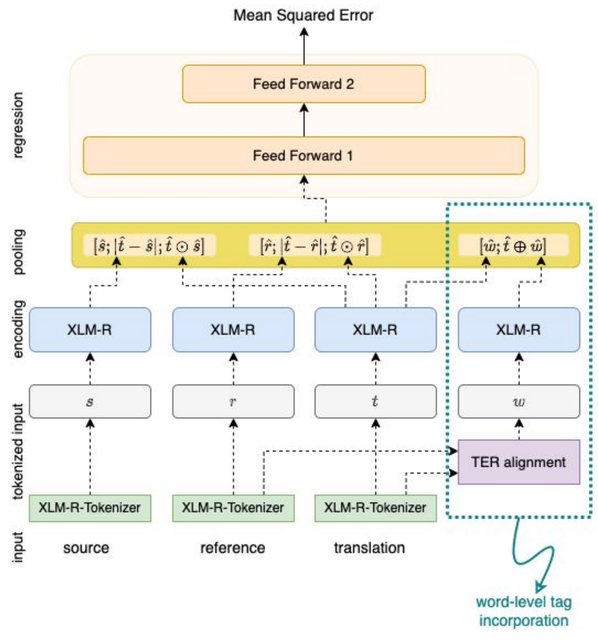

## Extending COMET for Lexical Features Incorporation

Sentence-level features: BLEU & chrF

COMET architecture

Word-level features: e.g. OK / BAD tags




## Segment-level Correlations

| | BLEU | CHRF | COMET | ENSEMBLE | COMET+aug | COMET+SL-feat. | COMET+WL-tags |
|---|---|---|---|---|---|---|---|
| AVG | 0.150 | 0.176 | 0.321 | 0.317 | 0.322 | 0.323 | **0.334**† |

Kendall's Tau correlation with human scores (MQM) on high resource language pairs (LPs). Average score across 4 different domains and 3 LPs (En-De, En-Ru, Zh-En).

## Robustness to Different Error *Severities*

| Metric | Base | Crit. | Maj. | Min. | All |
|---|---|---|---|---|---|
| *lexical-based metrics* | | | | | |
| BLEU | **100.0** | 79.33 | 83.76 | 72.6 | 78.52 |
| CHRF | **100.0** | 90.79 | 90.85 | 80.83 | 87.16 |
| *neural-based metrics* | | | | | |
| ENSEMBLE | 100.0 | 96.87 | 92.91 | 93.77 | 95.14 |
| COMET | 99.3 | 95.77 | 91.04 | 92.18 | 93.74 |
| + aug | 98.6 | 95.54 | 91.66 | 92.06 | 93.65 |
| + SL-feat. | 99.3 | **96.95** | 93.56 | 94.64 | 95.59 |
| + WL-tags | 99.2 | 96.48 | **93.9** | **96.36** | **96.2** |

Accuracy on errors (synthetically generated using perturbations) for the DEMETR dataset, bucketed by ***error severity.***

## Future Steps

Use additional ***context*** to inform word-level quality?

| | |
|---|---|
| Source | The nurse told the patient about his surgery. **He** then … |
| MT 1 | L'**infirmière** a parlé au patient de sa chirurgie. |
| MT 2 | L'**infirmier** a parlé au patient de sa chirurgie. |