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 Approach:

- ▷ **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

Evaluation:

We compare with:

- ▷ COMET
- ▷ COMET + augmentation

We measure:

- ▷ **Correlation with humans:** WMT22; MQM
- ▷ **Accuracy on challenge sets**: DEMETR, ACES

Extending COMET for Lexical Features Incorporation

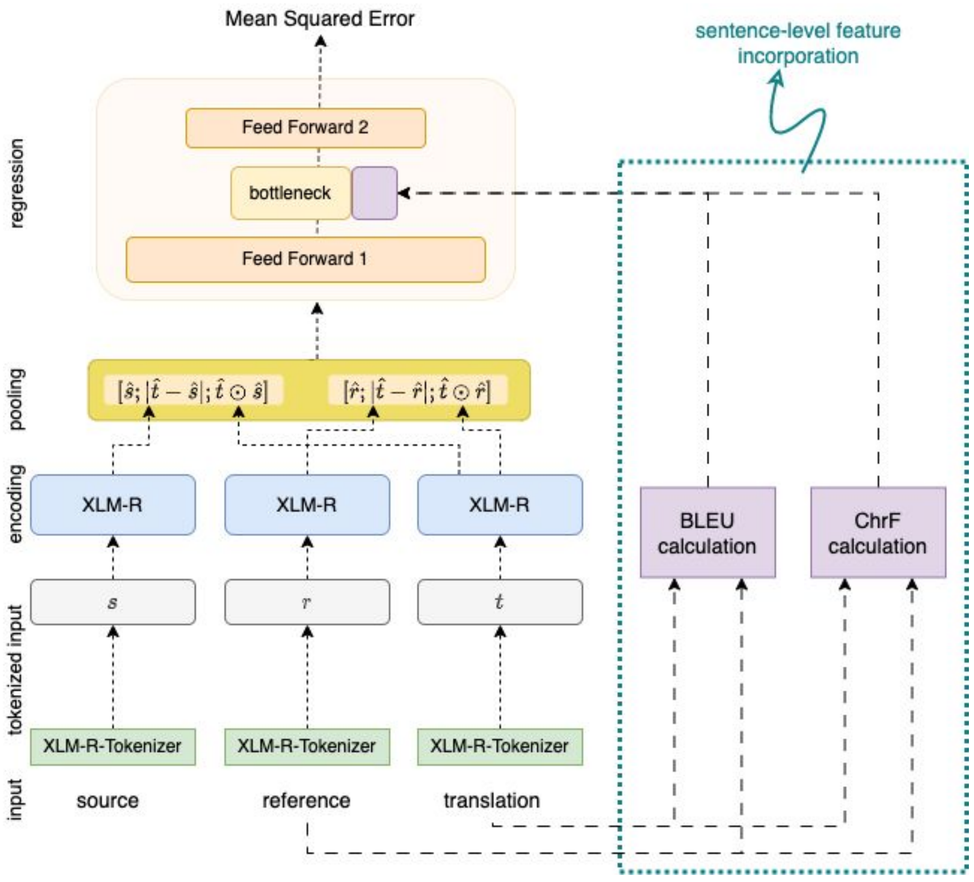

Figure 1: The architecture of the COMET model with incorporated sentence-level lexical features.

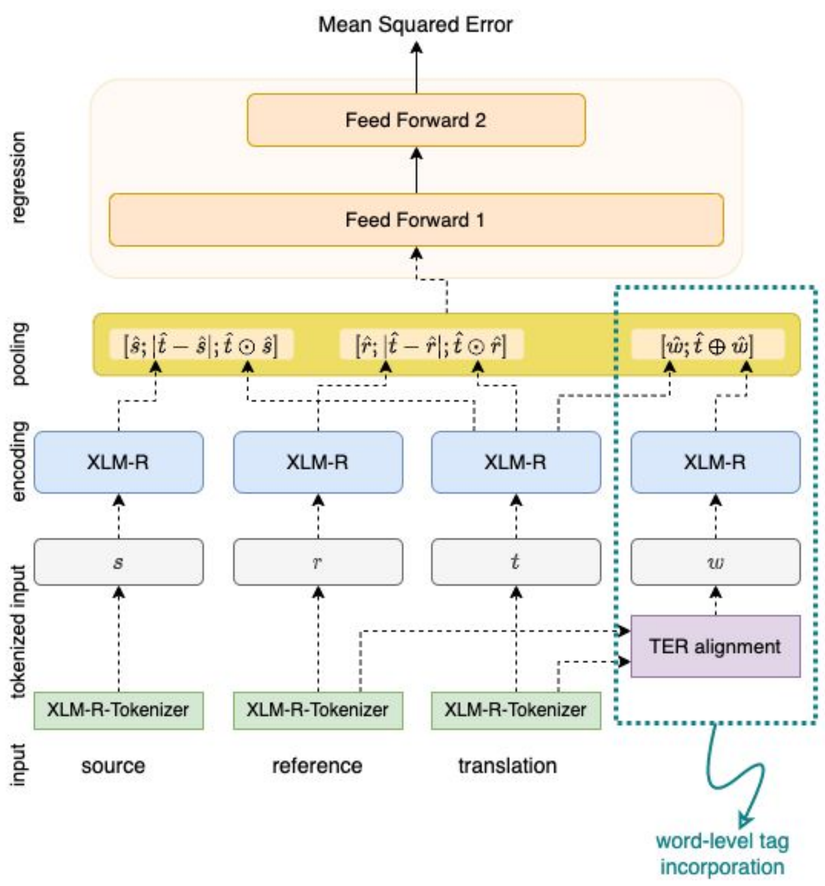

Figure 2: The architecture of the COMET model with incorporated word-level lexical features.

Segment-level Correlations

		BLEU	CHRF	COMET	ENSEMBLE	COMET+aug	COMET+SL-feat.	COMET+WL-tags
EN-DE	Conversation	0.201	0.257	0.308	0.309	0.296	0.310	**0.314**
	E-commerce	0.179	0.212	**0.326**	0.318	0.311	0.322	0.322
	News	0.167	0.202	0.361	0.356	0.330	0.355	**0.369**
	Social	0.130	0.168	**0.297**	0.292	0.277	0.294	0.293
EN-RU	Conversation	0.140	0.175	0.305	0.304	**0.328**	0.298	**0.328**
	E-commerce	0.202	0.221	0.372	0.371	0.382	0.369	**0.391**
	News	0.125	0.164	**0.373**	0.367	0.366	0.384	0.370
	Social	0.152	0.132	0.305	0.304	0.330	0.332	**0.349**
ZH-EN	Conversation	0.125	0.160	0.283	0.282	0.295	0.283	**0.298**
	E-commerce	0.174	0.187	0.326	0.325	0.342	0.335	**0.357**
	News	0.046	0.042	0.270	0.261	0.291	0.276	**0.292**
	Social	0.162	0.190	0.319	0.316	0.313	0.315	**0.330**
	AVG	0.150	0.176	0.321	0.317	0.322	0.323	**0.334**†

Table 1: Kendall's tau correlation on high resource language pairs using the MQM annotations for Conversation, E-commerce, News and Social domains collected for the WMT 2022 Metrics Task. **Bold** numbers indicate the best result for each domain in each language pair. † in the averaged scores indicates statistically significant difference to the other metrics [5].

Robustness to Different Types of Errors

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**

Table 3: Accuracy on DEMETR perturbations for both lexical-based and neural-based metrics, shown bucketed by error severity (base, critical, major, and minor errors), including a micro-average across all perturbations.

Acknowledgements:

This work was supported by the European Research Council (ERC StG DeepSPIN 758969), by EU's Horizon Europe Research and Innovation Actions (UTTER, contract 101070631), by P2020 project MAIA (LISBOA-01-0247- FEDER045909), by the Portuguese Recovery and Resilience Plan through project C645008882-00000055 (NextGenAI, Center for Responsible AI) and Fundação para a Ciência e Tecnologia through contract UIDB/50008/2020.