COMET-22:
Unbabel-IST 2022 Submission for the Metrics Shared Task

Ricardo Rei, José G. C. de Souza, Duarte M. Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, André F. T. Martins

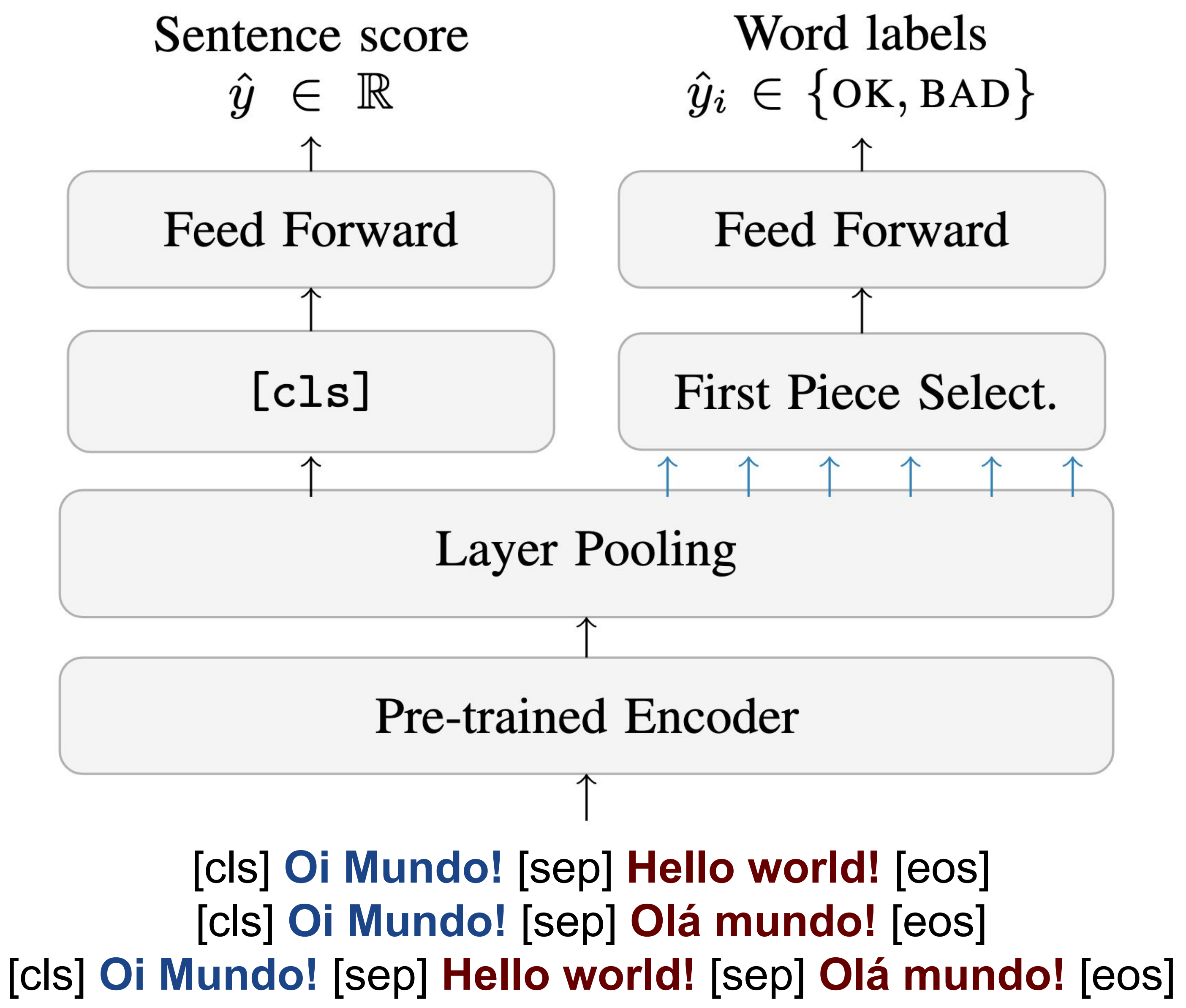
Our submission is an **ensemble between two evaluation models:**

- COMET estimator trained on DA's (similar to the model from Rei et al. 2020)
- New Multitask Model trained on MQM data.

Our **new architecture was specifically designed to learn from MQM data** by taking advantage of MQM error spans along with the final sentence score.

Together these models show **improved correlations** compared to state-of-the-art metrics from last year as well as **increased robustness** to critical errors.

Extending COMET for Sequence Tagging



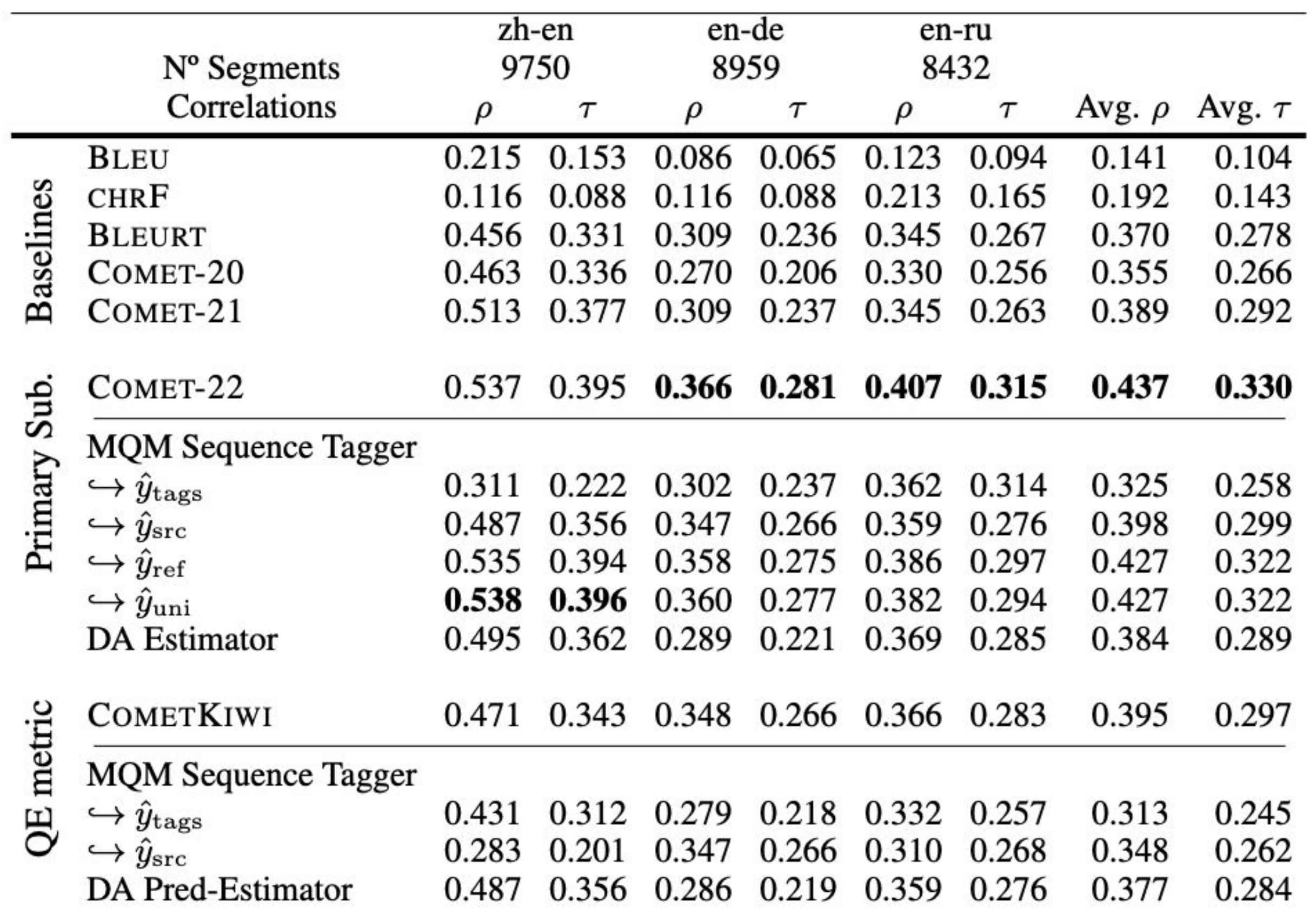
[cls] **Oi Mundo!** [sep] **Hello world!** [eos]
[cls] **Oi Mundo!** [sep] **Olá mundo!** [eos]
[cls] **Oi Mundo!** [sep] **Hello world!** [sep] **Olá mundo!** [eos]

Our new multitask model is inspired by OpenKiwi (Kepler et al. 2019) and UniTE (Wan et al. 2022) and performs both regression and sequence tagging.

From this model we produce 4 scores:
1) Reference score
2) Source score
3) Unified score
4) Tagging score

Since the model is trained with and without references we can also use it for QE-as-a-metric by restricting the access to the reference translation during inference.

Segment-level Correlations



	Nº Segments	zh-en 9750		en-de 8959		en-ru 8432			
	Correlations	ρ	τ	ρ	τ	ρ	τ	Avg. ρ	Avg. τ
Baselines	BLEU	0.215	0.153	0.086	0.065	0.123	0.094	0.141	0.104
	CHRF	0.116	0.088	0.116	0.088	0.213	0.165	0.192	0.143
	BLEURT	0.456	0.331	0.309	0.236	0.345	0.267	0.370	0.278
	COMET-20	0.463	0.336	0.270	0.206	0.330	0.256	0.355	0.266
	COMET-21	0.513	0.377	0.309	0.237	0.345	0.263	0.389	0.292
Primary Sub.	COMET-22	0.537	0.395	**0.366**	**0.281**	**0.407**	**0.315**	**0.437**	**0.330**
	MQM Sequence Tagger								
	$\hookrightarrow \hat{y}_{\text{tags}}$	0.311	0.222	0.302	0.237	0.362	0.314	0.325	0.258
	$\hookrightarrow \hat{y}_{\text{src}}$	0.487	0.356	0.347	0.266	0.359	0.276	0.398	0.299
	$\hookrightarrow \hat{y}_{\text{ref}}$	0.535	0.394	0.358	0.275	0.386	0.297	0.427	0.322
	$\hookrightarrow \hat{y}_{\text{uni}}$	**0.538**	**0.396**	0.360	0.277	0.382	0.294	0.427	0.322
	DA Estimator	0.495	0.362	0.289	0.221	0.369	0.285	0.384	0.289
QE metric	COMETKIWI	0.471	0.343	0.348	0.266	0.366	0.283	0.395	0.297
	MQM Sequence Tagger								
	$\hookrightarrow \hat{y}_{\text{tags}}$	0.431	0.312	0.279	0.218	0.332	0.257	0.313	0.245
	$\hookrightarrow \hat{y}_{\text{src}}$	0.283	0.201	0.347	0.266	0.310	0.268	0.348	0.262
	DA Pred-Estimator	0.487	0.356	0.286	0.219	0.359	0.276	0.377	0.284

Table 1: Segment-level Spearman R (ρ) and Kendall-Tau (τ) correlations for zh-en, en-de and en-ru 2021 MQM annotations for the News Domain.

Robustness to Critical Errors

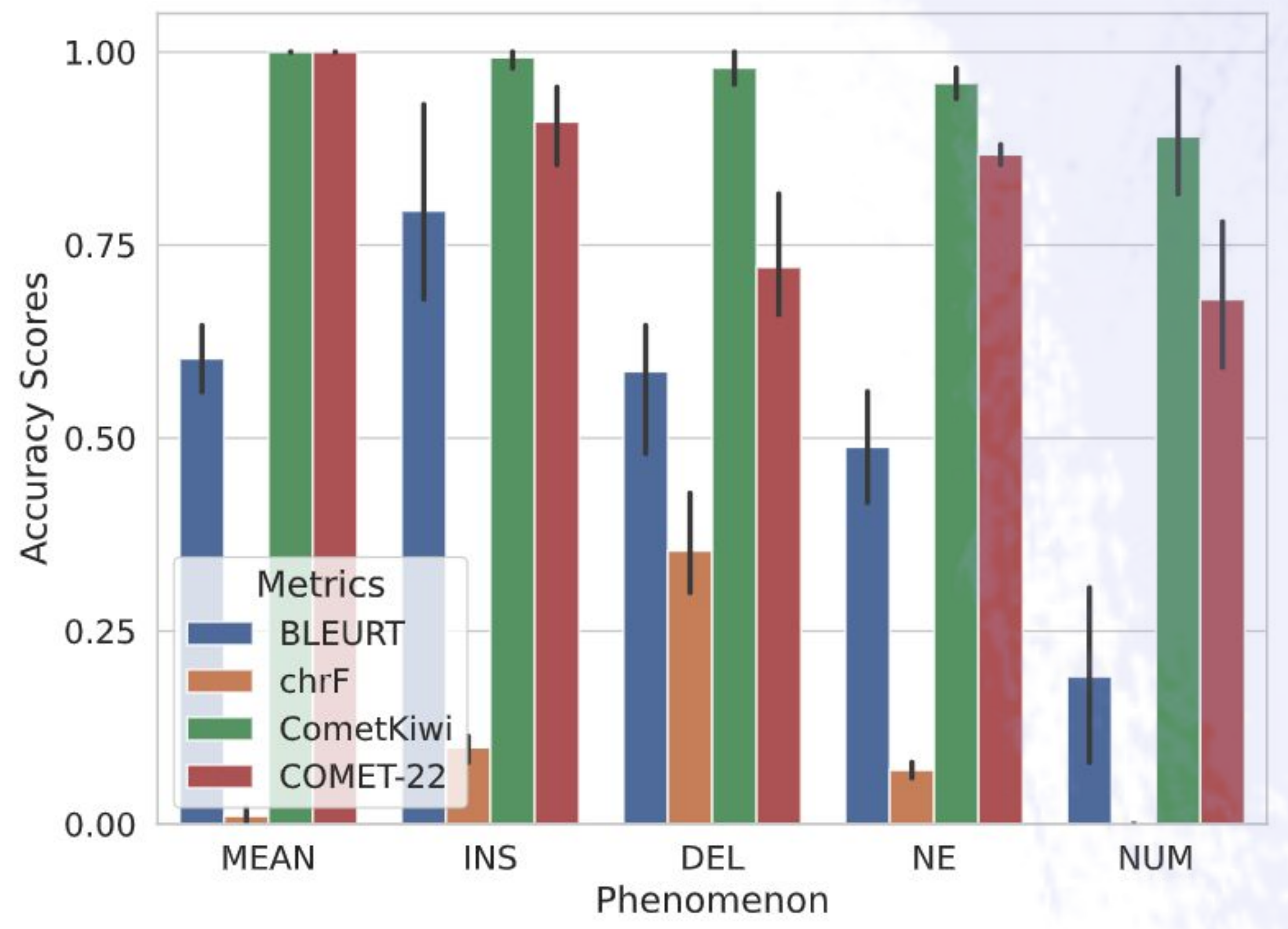


Figure 1: Accuracy Scores on the SMAUG Challenge Set for the baseline and submitted metrics.

Acknowledgements:
Work was supported by the P2020 programs (MAIA, contract 045909), by EU's Horizon Europe Research and Innovation Actions (UTTER, contract 101070631), by the European Research Council (ERC StG DeepSPIN 758969), and by the Funda\c{c}\~{a}o para a Ci\^{e}ncia e Tecnologia through contract UIDB/50008/2020.

References:
Unbabel's Participation in the WMT20 Metrics Shared Task (Rei et al., WMT 2020)
OpenKiwi: An Open Source Framework for Quality Estimation (Kepler et al., ACL 2019)
UniTE: Unified Translation Evaluation (Wan et al., ACL 2022)

UNIÃO EUROPEIA
Fundo Europeu de Desenvolvimento Regional
AGÊNCIA NACIONAL DE INOVAÇÃO