Disentangling Uncertainty in Machine Translation Evaluation

Chrysoula Zerva [1,4]
Taisiya Glushkova [1,4]
Ricardo Rei [2,3,4]
André F. T. Martins [1,2,4]

[1]Instituto de Telecomunicações, [2]Unbabel, [3]Inesc ID,
[4]Instituto Superior Técnico & LUMLIS (Lisbon ELLIS Unit)

CMU PORTUGAL SUMMIT 2022
NEW FRONTIERS IN TECH

Introduction

Lexical and Neural-based metrics share a **list of limitations**:

- lack of robustness
- lack of reliability
- lack of interpretability in scores
- ...

Our goal:

to fill this gap and improve the existing metrics

Methods

- **Direct uncertainty prediction (DUP)**
  a two-step approach which uses supervision over the quality prediction errors
- **Heteroscedastic regression (HTS)**
  estimates input-dependent aleatoric uncertainty and can be combined with MC dropout
- **KL-divergence minimization (KL)**
  estimates uncertainty from annotator disagreements, when multiple annotations are available for the same example

$$\epsilon^* = |\hat{q} - q^*|$$

$$\mathcal{L}^{\mathrm{E}}_{\mathrm{HTS}}(\hat{\epsilon};\epsilon^*) = \frac{(\epsilon^*)^2}{2\hat{\epsilon}^2} + \tfrac{1}{2}\log(\hat{\epsilon})^2$$

$$\mathcal{L}_{\mathrm{HTS}}(\hat{\mu},\hat{\sigma}^2;q^*) = \frac{(q^*-\hat{\mu})^2}{2\hat{\sigma}^2} + \tfrac{1}{2}\log\hat{\sigma}^2.$$

$$\mathcal{L}_{\mathrm{KL}}(\hat{\mu},\hat{\sigma}^2;\mu^*,\sigma^{*2}) = \frac{(\mu^*-\hat{\mu})^2+\sigma^{*2}}{2\hat{\sigma}^2} + \frac{1}{2}\log\frac{\hat{\sigma}^2}{\sigma^{*2}} - \frac{1}{2}$$

Experiments

Results for segment-level DA and MQM

		UPS ↑	ECE ↓	Sha. ↓	NLL ↓	PPS ↑
WMT20 DA	σ^2-fixed	–	0.019	0.415	1.236	0.444
	MCD	0.106	0.016	0.377	1.199	0.443
	DE	0.134	0.019	0.366	1.156	0.460
	HTS	0.177	0.015	0.450	1.201	0.444
	HTS+MCD	0.254	0.013	0.528	1.167	0.429
	DUP	0.182	0.014	0.437	1.190	0.444
WMT21 MQM	σ^2-fixed	–	0.055	0.371	2.090	0.377
	MCD	0.179	0.024	0.334	1.686	0.460
	DE	0.128	0.051	0.236	2.631	0.479
	HTS	0.307	0.041	0.284	2.264	0.445
	HTS+MCD	0.311	0.037	0.388	1.614	0.445
	KL	0.296	0.046	0.273	2.595	0.443
	DUP	0.285	0.039	0.634	1.778	0.377

Evaluation Metrics

Quality prediction accuracy:

Predictive Pearson Score (PPS) .... $r(q^*,\hat{\mu})$

Uncertainty-related accuracy:

Uncertainty Pearson Score (**UPS**) .... $r(|q^*-\hat{\mu}|,\hat{\sigma})$

Sharpness (sha) .... $\mathrm{sha}(\hat{p}_Q) = \frac{1}{|\mathcal{D}|}\sum_{\langle s,t,\mathcal{R}\rangle\in\mathcal{D}}\hat{\sigma}^2.$

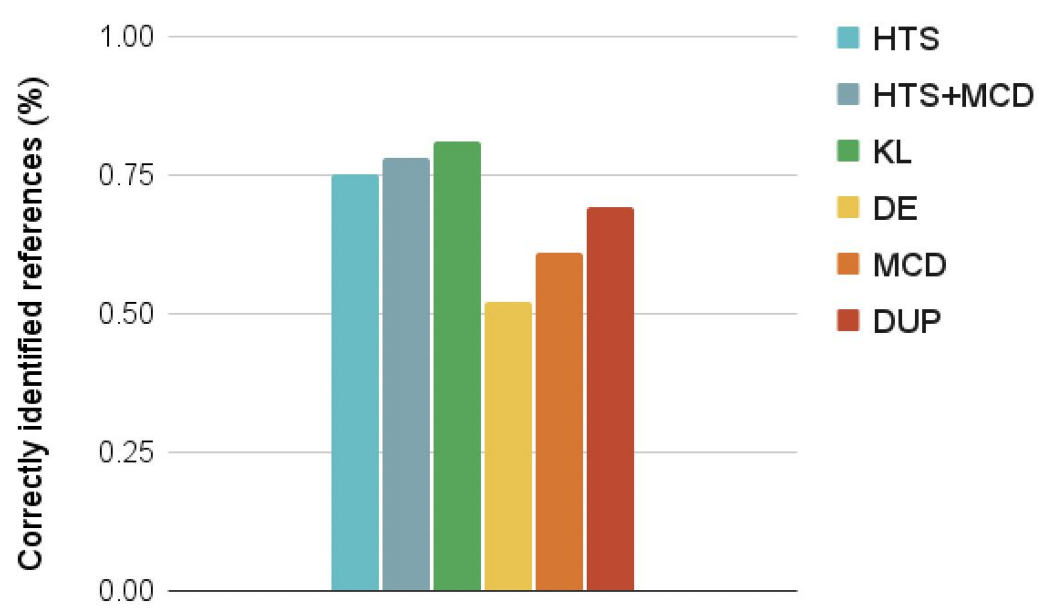
Expected Calibration Error (ECE) .... $\mathrm{ECE} = \frac{1}{M}\sum_{b=1}^{M}|\mathrm{acc}(\gamma_b)-\gamma_b|,$

Combination:

$\mathrm{acc}(\gamma_b) = \frac{1}{|\mathcal{D}|}\sum_{\langle s,t,\mathcal{R},q^*\rangle\in\mathcal{D}}\mathbb{1}(q^*\in I(\gamma_b)).$

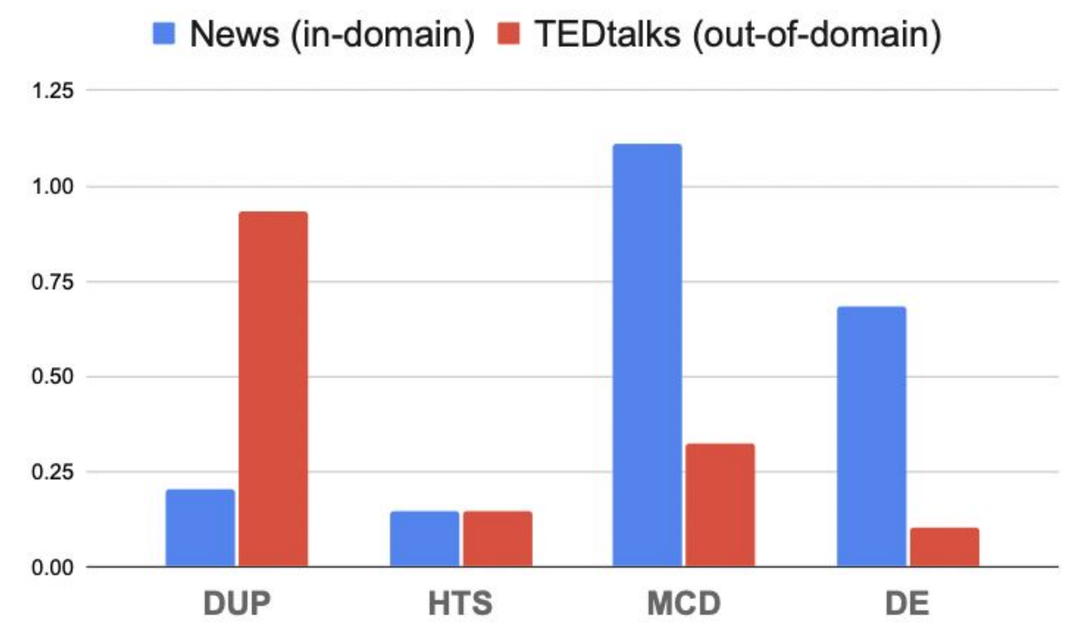
Negative Log-Likelihood (NLL) .... $\mathrm{NLL} = -\frac{1}{|\mathcal{D}|}\sum_{\langle s,t,\mathcal{R},q^*\rangle\in\mathcal{D}}\log\hat{p}(q^*\mid\langle s,t,\mathcal{R}\rangle).$

Identification of noisy references



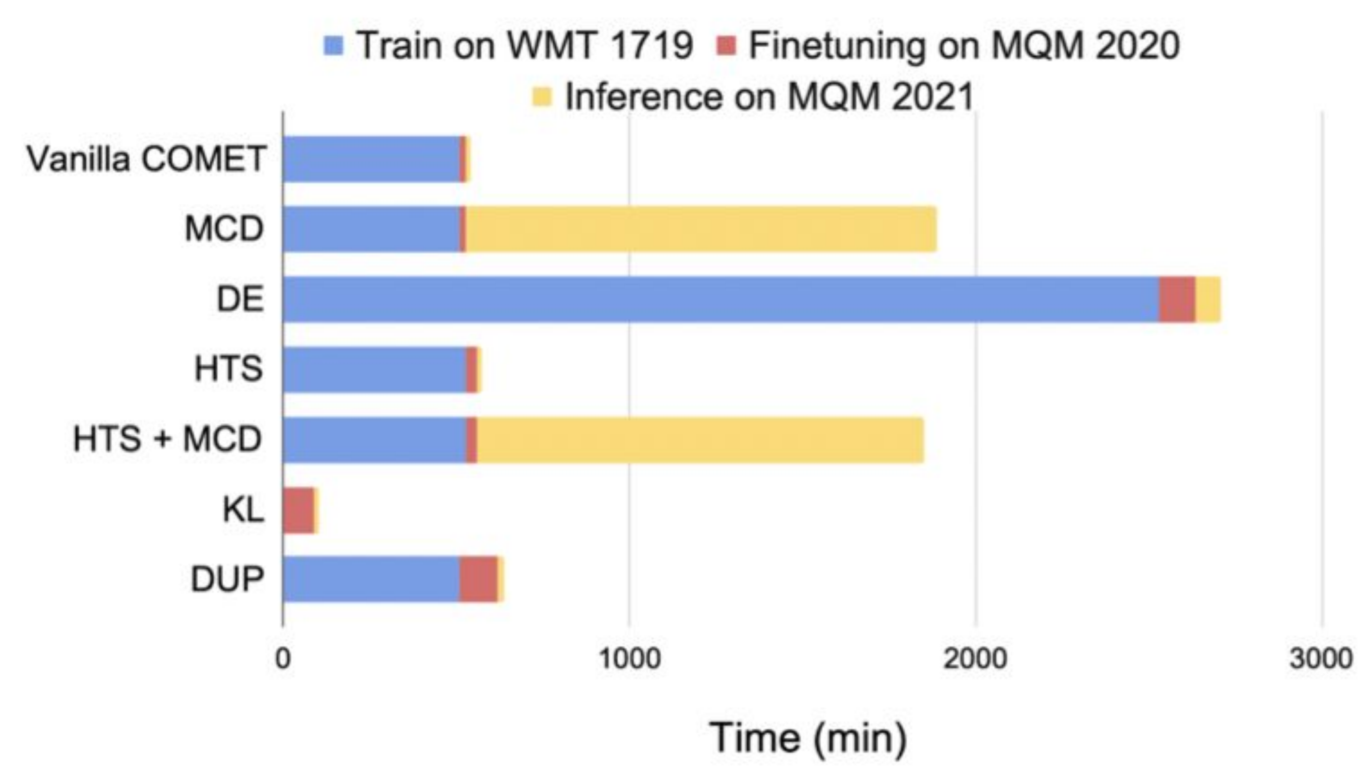
Results for segment-level DA and MQM

Sharpness: Epistemic uncertainty caused by out-of-domain data



Main Takeaways

- **improved results on uncertainty prediction** for the WMT metrics task datasets
- a substantial **reduction in computational costs** (compared to MCD and DE)
- **the ability** of new uncertainty predictors **to target different aleatoric and epistemic uncertainty sources** in MT evaluation, such as:
  - *low quality references*
  - *out-of-domain data*



This work was supported by P2020 project MAIA (LISBOA-01-0247- FEDER045909).

Carnegie Mellon Portugal

fct Fundação para a Ciência e a Tecnologia

instituto de telecomunicações

TÉCNICO LISBOA

Unbabel

inesc id lisboa