# Disentagling Uncertainty for Machine Translation Evaluation

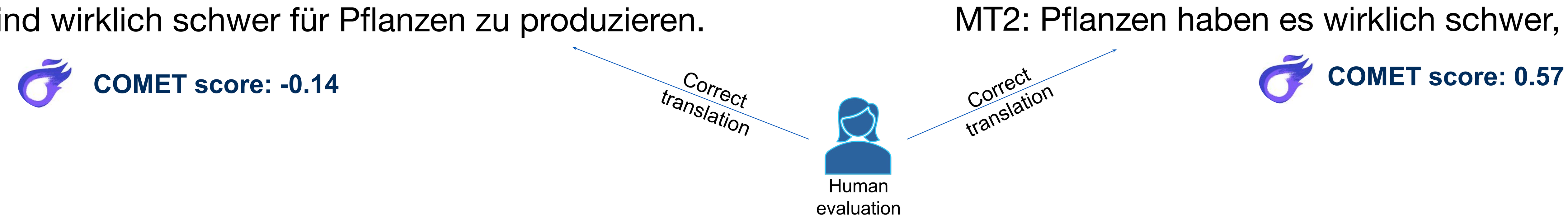
Chrysoula Zerva[1,4], Taisiya Glushkova[1,4], Ricardo Rei[2,3], Andre F. T. Martins[1,2,4]

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

They're really difficult for plants to produce.
*Pflanzen haben grosse Mühe sie zu produzieren.*

MT1: Sie sind wirklich schwer für Pflanzen zu produzieren.

MT2: Pflanzen haben es wirklich schwer, sie zu produzieren.




## Can we determine how **confident** our metric is and **why**?

## Motivation

MT evaluation metrics share a list of limitations:

- Limited reliability
- Lack of robustness
- Lack of interpretability for the predicted scores

We aim to fill this gap by investigating **uncertainty quantification** methods for MT evaluation that target specific **sources of uncertainty**

## Sources of uncertainty

- Out-of-domain data
- Insufficient training
- Complex sentences
- Low quality references
- Annotator disagreements

Epistemic uncertainty

Aleatoric uncertainty

Total uncertainty

## Methods

### Baselines

Variance-based methods which do not target specific uncertainty sources

- $\sigma^2$-fixed: minimise $\frac{1}{|\mathcal{D}|}\sum_{\langle s,t,\mathcal{R},q^*\rangle\in\mathcal{D}}(q^*-\hat{q})^2$
- MC dropout (MCD): calculate STD over multiple (100) inference runs
- Deep Ensembles (DE): calculate STD over 5 checkpoints

### Aleatoric

Can we learn from annotator disagreement?

- KL-divergence minimisation:

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

estimate uncertainty from annotator disagreement (STD), when multiple annotations are available for each example

If we do not have access to annotator disagreement?

- Heteroscedastic uncertainty:

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

learn to predict heteroscedastic noise variance from the training data

*Combine with **MCD** for total uncertainty prediction

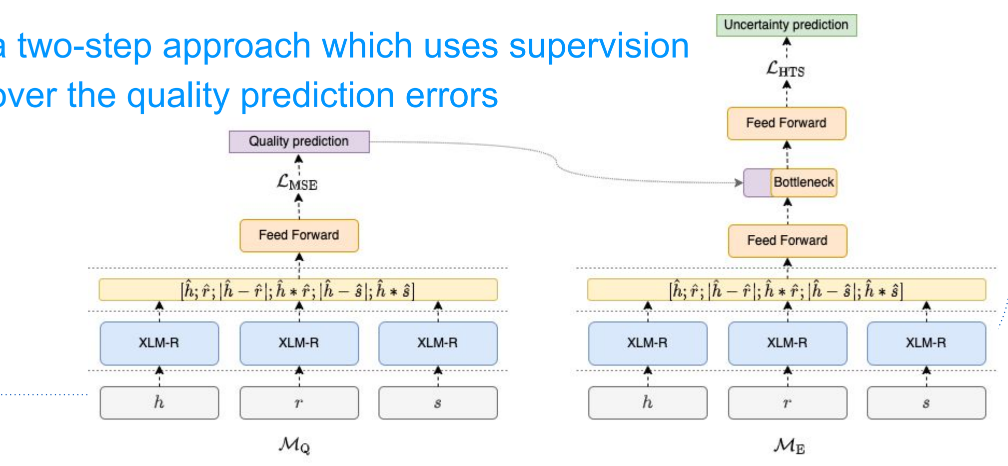
### Epistemic/Total uncertainty

Can we learn **directly** from the metric **error ($\epsilon$)**?

- Direct Uncertainty Prediction (DUP)

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

a two-step approach which uses supervision over the quality prediction errors




## Evaluation

What indicates a **good** uncertainty prediction method?

Accurate & representative uncertainty intervals:

- Uncertainty Pearson Score (UPS) $r(|q^*-\hat{q}|,\hat{\sigma})$
- Estimated Calibration Error (ECE) $\frac{1}{M}\sum_{b=1}^{M}|\mathrm{acc}(\gamma_b)-\gamma_b|$
- Sharpness (sha) $\frac{1}{|\mathcal{D}|}\sum_{\langle s,t,\mathcal{R}\rangle\in\mathcal{D}}\hat{\sigma}^2$

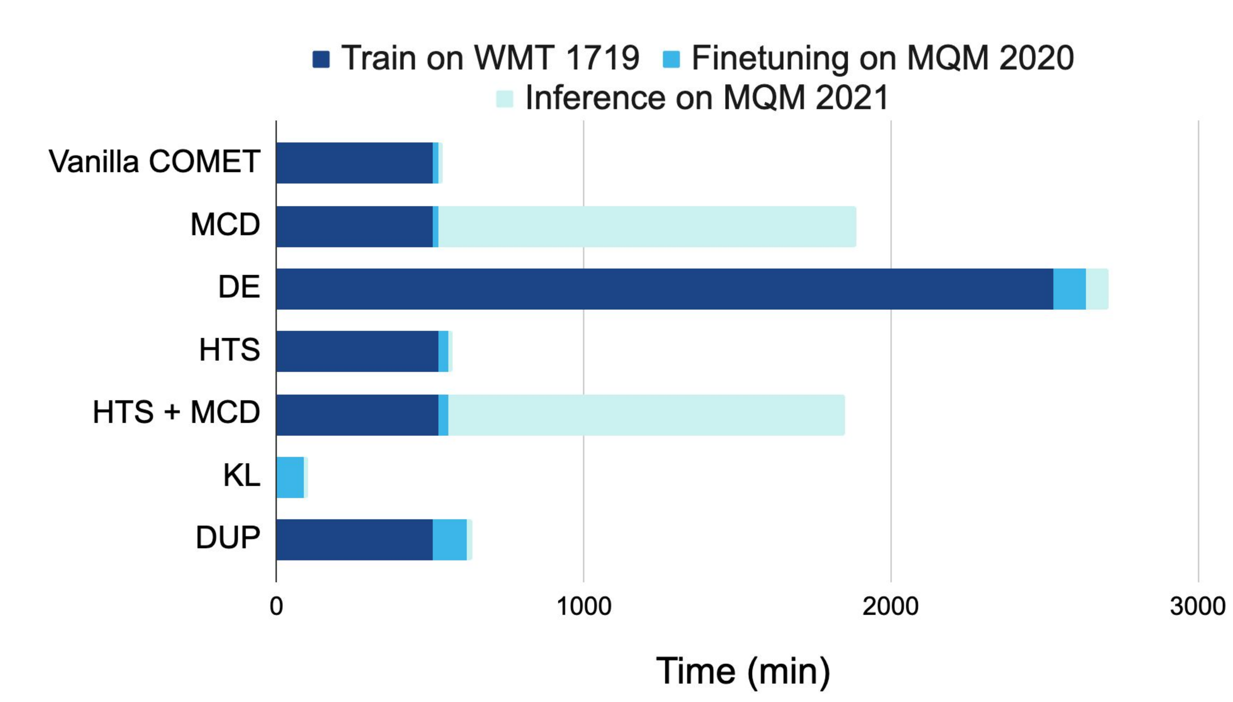
… without compromising the quality prediction accuracy:

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

| | | UPS ↑ | ECE ↓ | Sha. ↓ | PPS ↑ |
|---|---|---|---|---|---|
| WMT20 DA | $\sigma^2$-fixed | – | 0.019 | 0.415 | 0.444 |
| | MCD | 0.106 | 0.016 | 0.377 | 0.443 |
| | DE | 0.134 | 0.019 | 0.366 | 0.460 |
| | HTS | 0.177 | 0.015 | 0.450 | 0.444 |
| | HTS+MCD | 0.254 | 0.013 | 0.528 | 0.429 |
| | DUP | 0.182 | 0.014 | 0.437 | 0.444 |
| WMT21 MQM | $\sigma^2$-fixed | – | 0.055 | 0.371 | 0.377 |
| | MCD | 0.179 | 0.024 | 0.334 | 0.460 |
| | DE | 0.128 | 0.051 | 0.236 | 0.479 |
| | HTS | 0.307 | 0.041 | 0.284 | 0.445 |
| | HTS+MCD | 0.311 | 0.037 | 0.388 | 0.445 |
| | KL | 0.296 | 0.046 | 0.273 | 0.443 |
| | DUP | 0.285 | 0.039 | 0.634 | 0.377 |

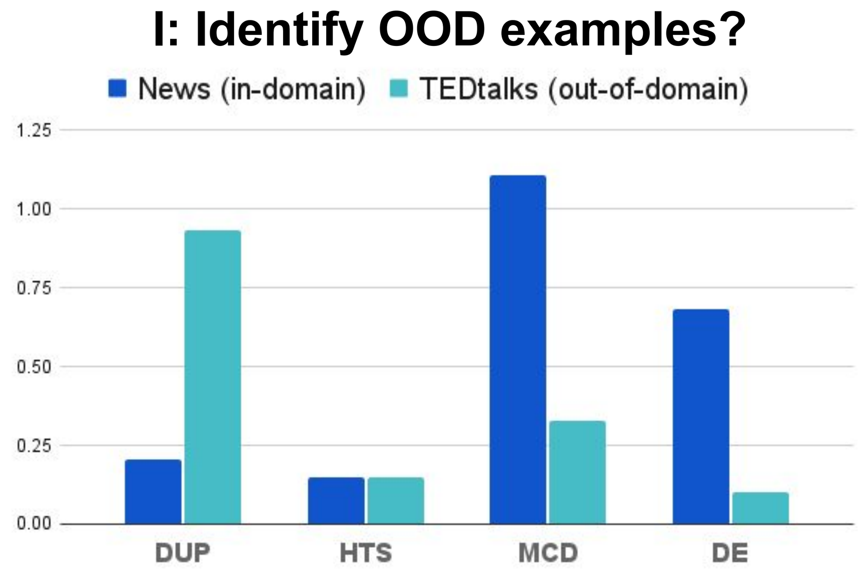
Performance on WMT 2020 (DA) and 2021(MQM) metrics data; averaged over all language pairs

## Computational Cost



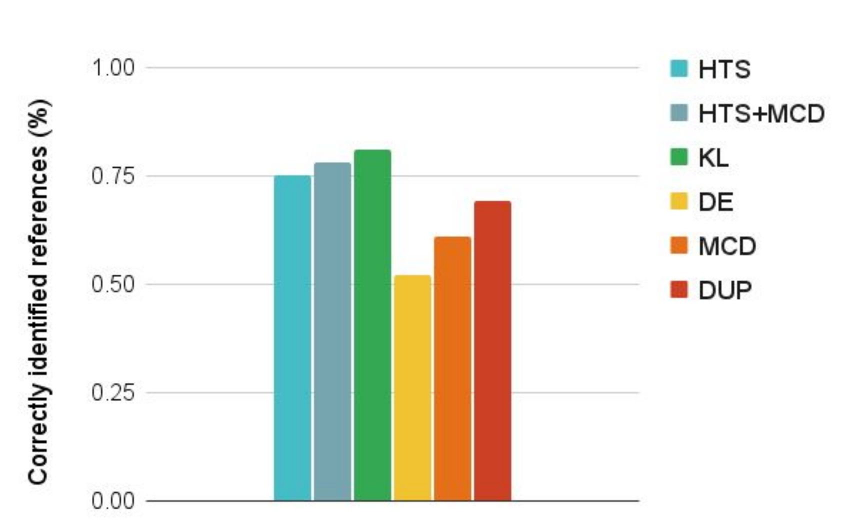

## Case-studies: Can we use predicted uncertainty to

### I: Identify OOD examples?




Sharpness (average uncertainty) on two En-Ru test sets from the WMT21 metrics task

### II: Identify high quality references?




Correctly recognized references with higher quality ($r_+$ vs $r_-$) by different uncertainty predictors on the En-De news 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

Github