Uncertainty-Aware Machine Translation Evaluation

Taisiya Glushkova, Chrysoula Zerva, Ricardo Rei, André F. T. Martins

Automatic MT Evaluation Metrics

Recently there has been a very good progress in automatic MT Evaluation metrics [1,2]

METEOR, BLEU, BERTScore, COMET, BLEURT, PRISM, …

but they all share the same limitation…

a single point estimate output

This paper: a **simple** way of getting a *distribution* of scores -- **confidence interval estimates**.

What are we trying to achieve?

Example of uncertainty-aware MT evaluation for a sentence in the WMT20 dataset (Mathur et al., 2020).

Source: ``She said, `That's not going to work.''

Reference: ``Она сказала: ``Не получится.''

Translation #1:

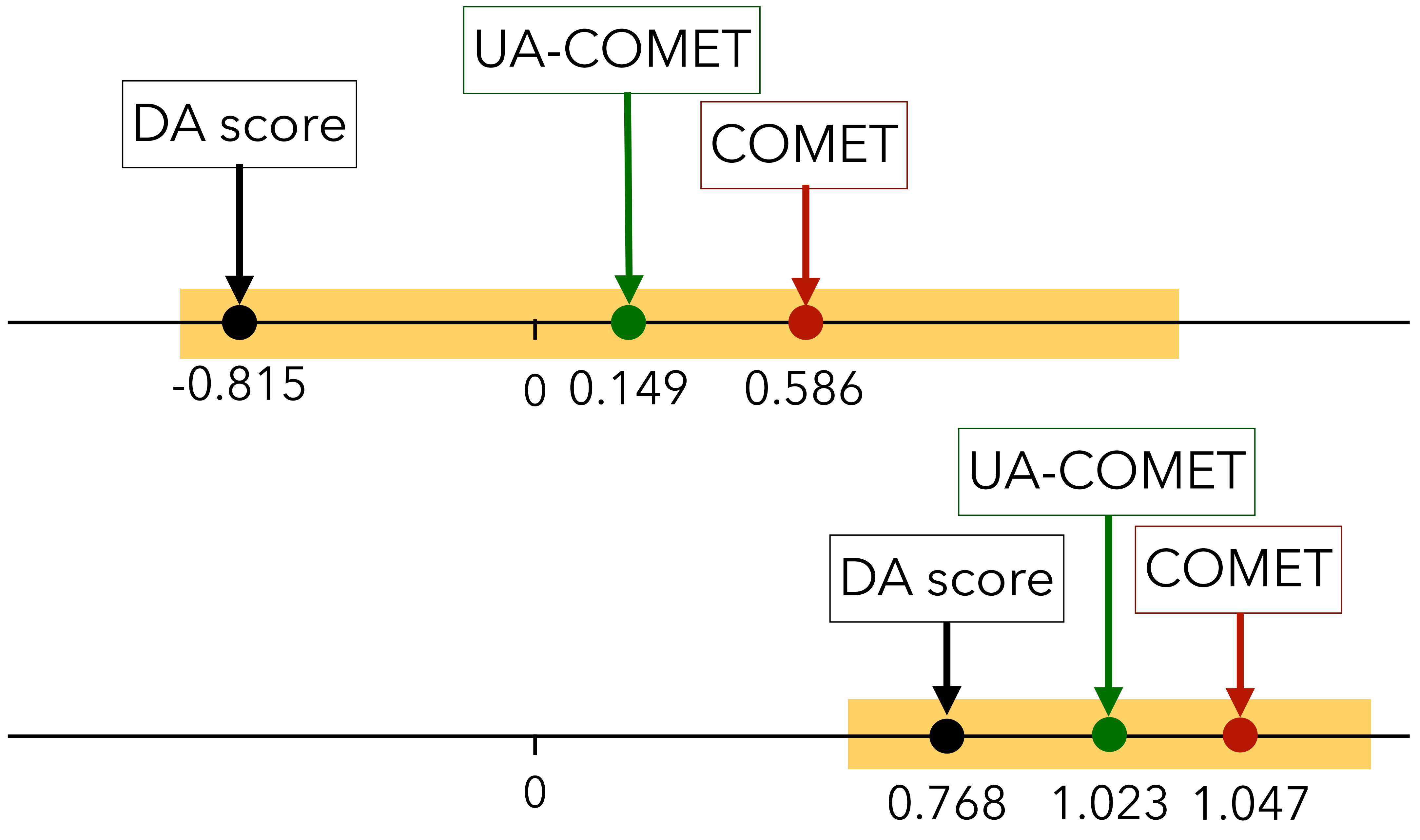
Она сказала, 'Это **не собирается** работать.

Gloss: «She said, that is not willing to work»

Translation #2:

Она сказала: «Это не сработает.

Gloss: «She said, «That will not work»



Sources of uncertainty in MT evaluation

- **Noisy DA/MQM scores**

  Low inter-annotator agreement

- **Noisy or insufficient references**

aleatoric (data) uncertainty

- **Complex non-literal translations**

- **Out-of-domain text**

  Domains of train and test data are different

epistemic (model) uncertainty

Uncertainty-aware MT Evaluation

Methods:

- **Monte Carlo Dropout** (MCD) (Gal et al, 2016)
- **Deep Ensembles** (DE) (Lakshminarayanan et al., 2017)

Framework of choice: COMET

Experiments:

- Uncertainty-aware MT evaluation on segment-level
- Impact of reference quantity
- Detection of critical translation mistakes

Well established for many ML tasks

- Including MT (Fomicheva et al., 2020)

While previous work on MT evaluation that uses gaussian processes is not easy to integrate into NN (Beck et al., 2016), MCD and DE are **easily applicable** to different NN

Notation – Problem Definition

Typical MT evaluation

input: $\langle s, t, \mathcal{R} \rangle$, where $\mathcal{R} = \{r_1, \ldots, r_{|\mathcal{R}|}\}$

ground truth score: q^* (DA, MQM or HTER)

output: $\hat{q} \in \mathbb{R}$

Uncertainty-Aware MT evaluation

input: $\langle s, t, \mathcal{R} \rangle$, where $\mathcal{R} = \{r_1, \ldots, r_{|\mathcal{R}|}\}$

ground truth score: q^* (DA, MQM or HTER)

output: $\hat{p}_Q(q)$ – a **distribution**, as apposed to a point estimate $\hat{q}$

assumption: Gaussian distribution

$$\hat{p}_Q(q) = \mathcal{N}(q; \hat{\mu}, \hat{\sigma}^2)$$

so that we can estimate: $\hat{\mu}$, $\hat{\sigma}^2$

Using the predicted variance, $\hat{\sigma}^2$ we can estimate the desired confidence intervals!

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)

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

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

Combination:

Negative Log-Likelihood (NLL)

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

Experiment 1 – Segment-level

Baseline

Original COMET score with Fixed (optimised) variance

$$\sigma^2_{\text{fixed}} = \frac{1}{|\mathcal{D}|} \sum_{\langle s,t,\mathcal{R},q^* \rangle \in \mathcal{D}} (q^* - \hat{\mu})^2$$

MC dropout (MCD)

Dropout probability: 0.1

Number of runs: N = 100

Deep Ensembles (DE)

N=5 models with random initialisation

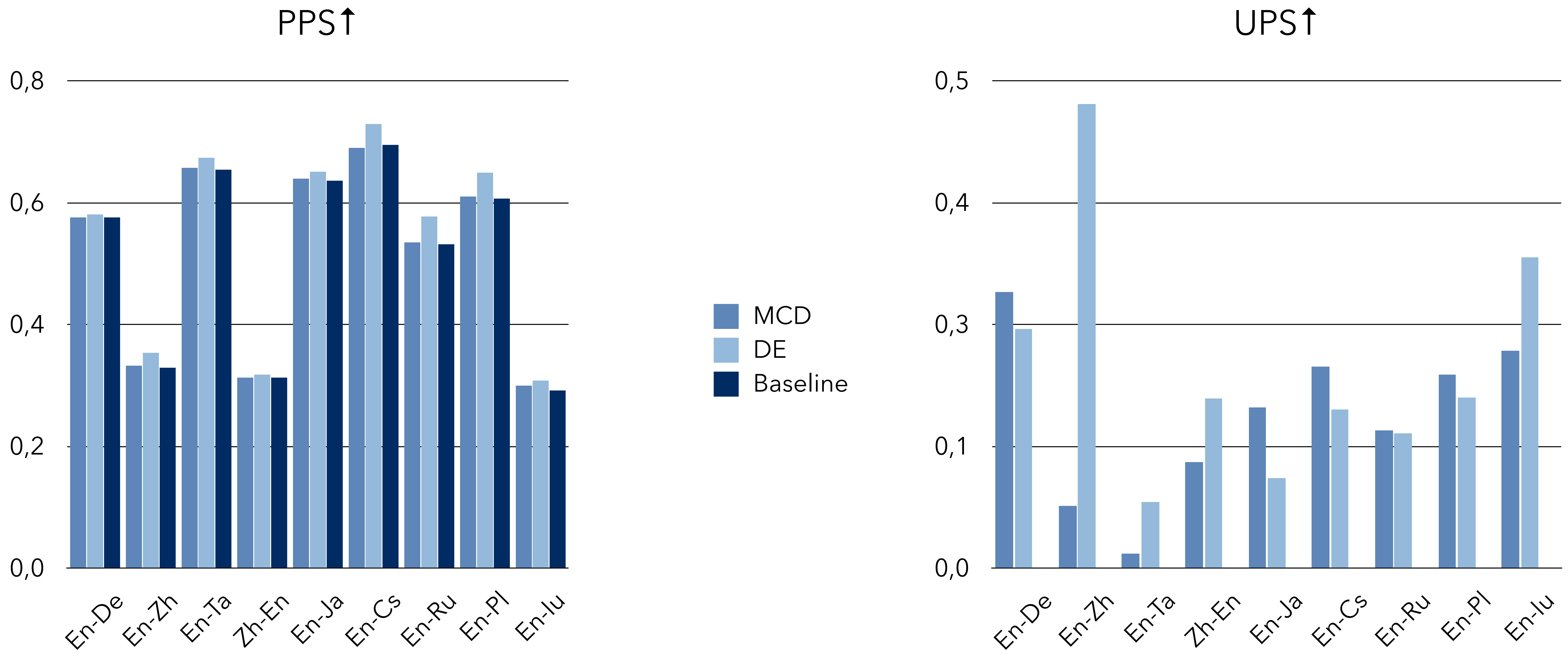
Train data

- WMT17-19 with DA scores
- 30 language pairs (LPs)

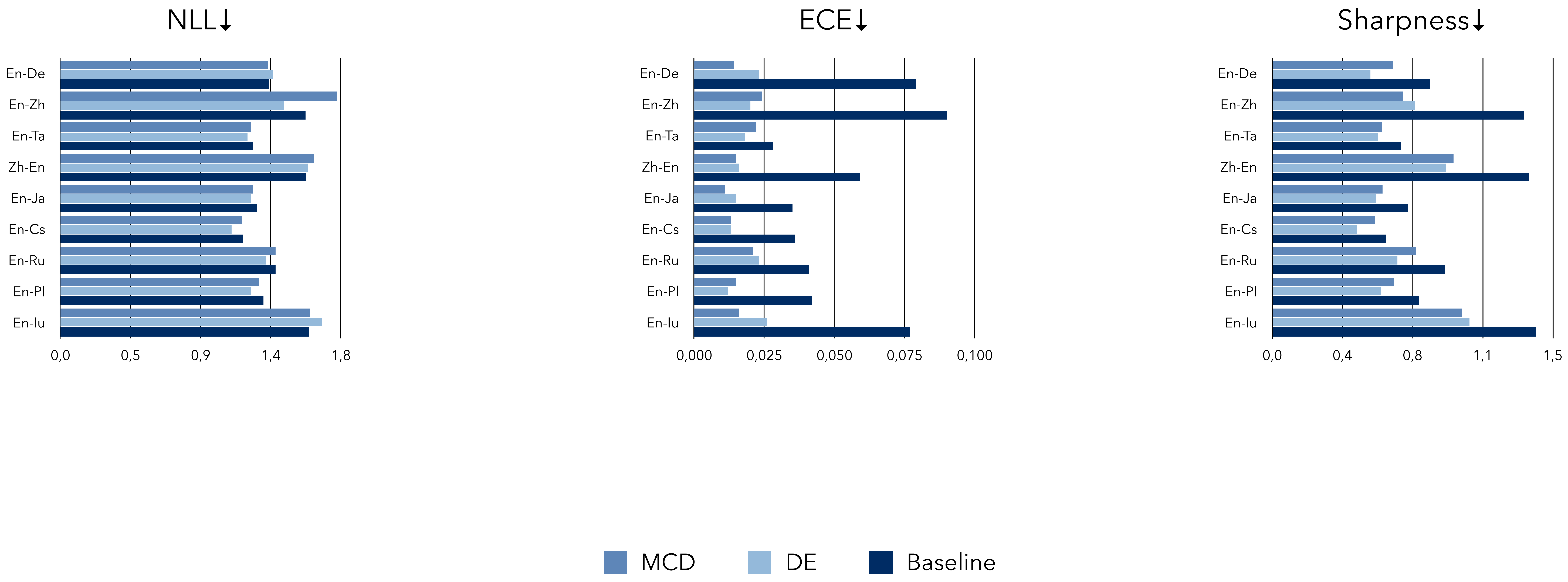
Test data

- WMT20 with DA scores, 9 LPs
- WMT20 with MQM, 2 LPs
- QT21 with HTER scores, 4 LPs

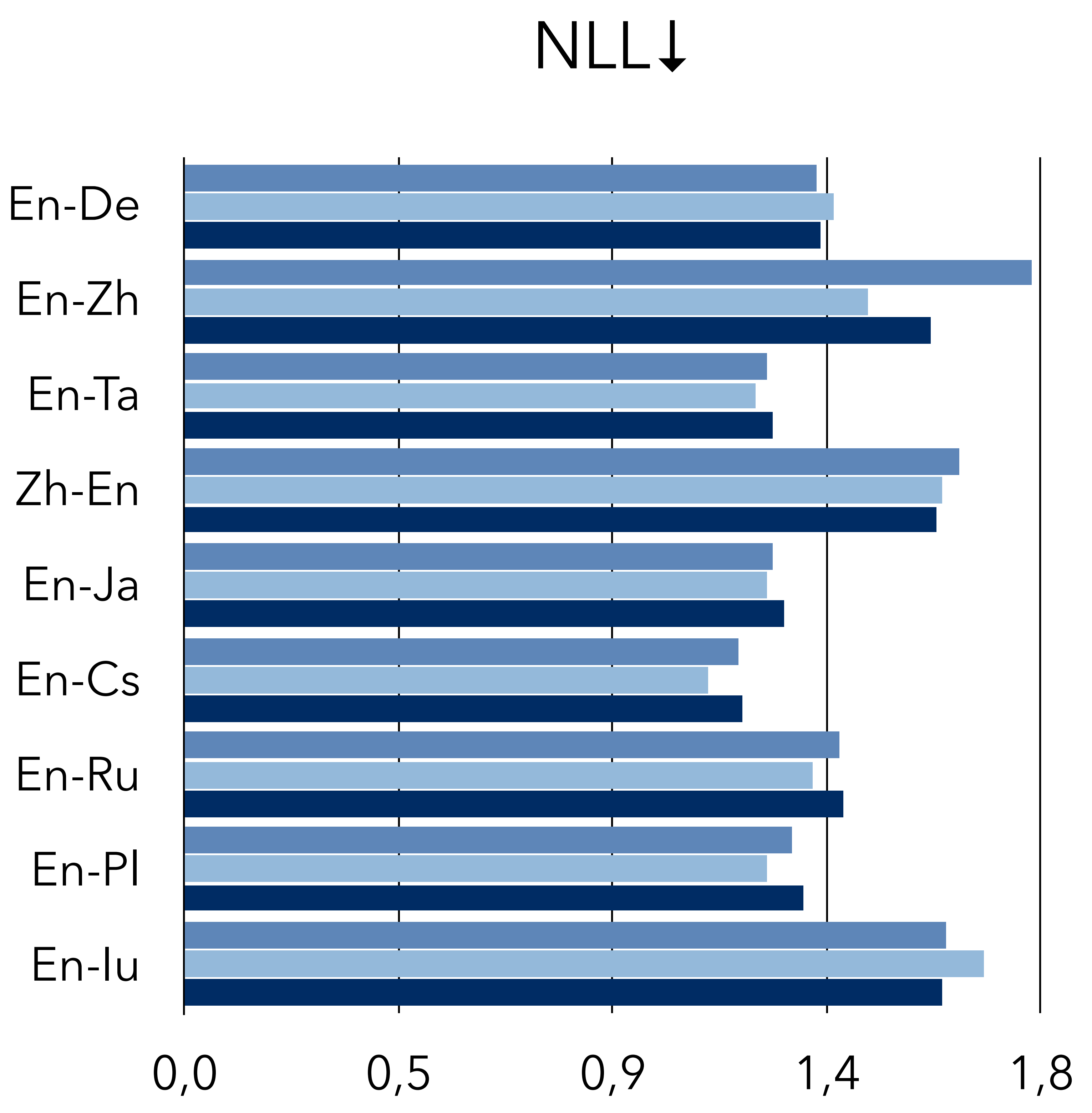
Experiment 1 – Results for segment-level DA predictions

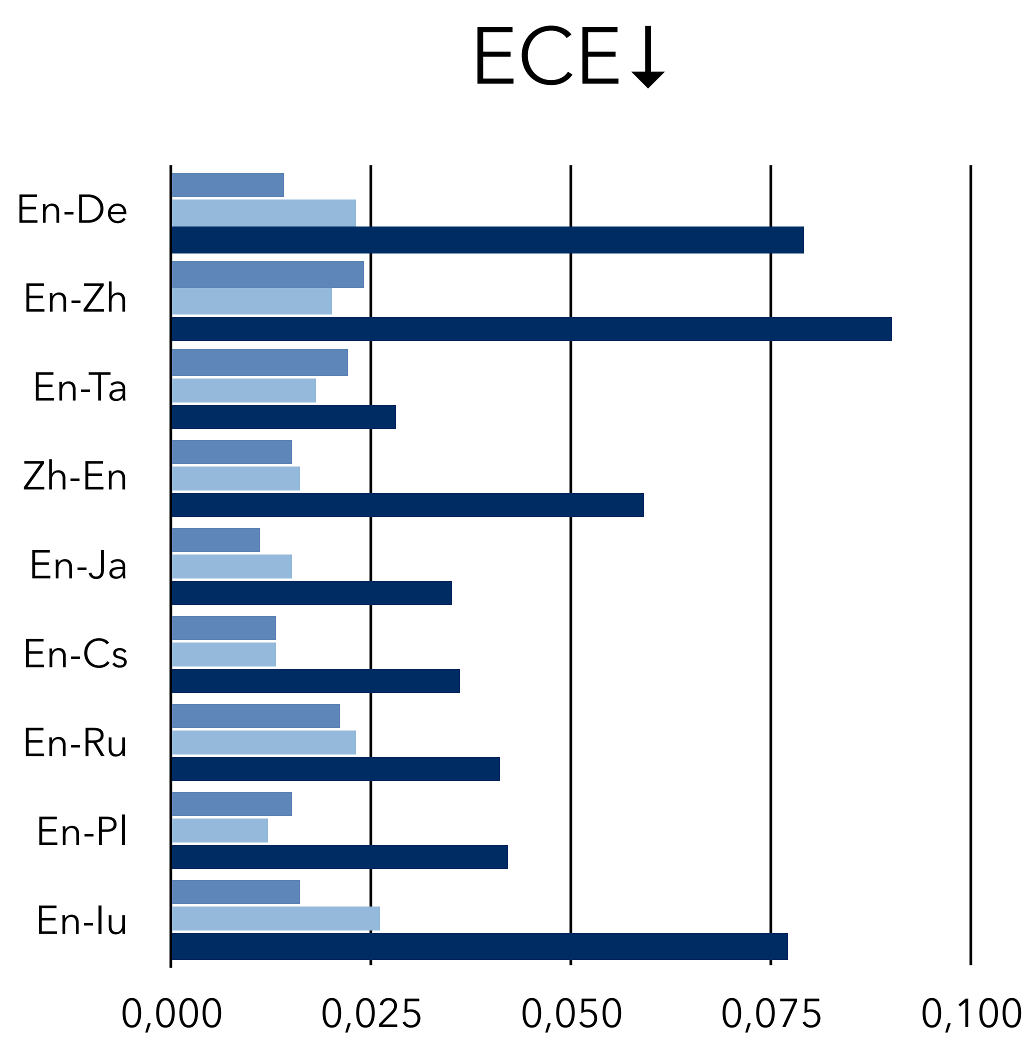


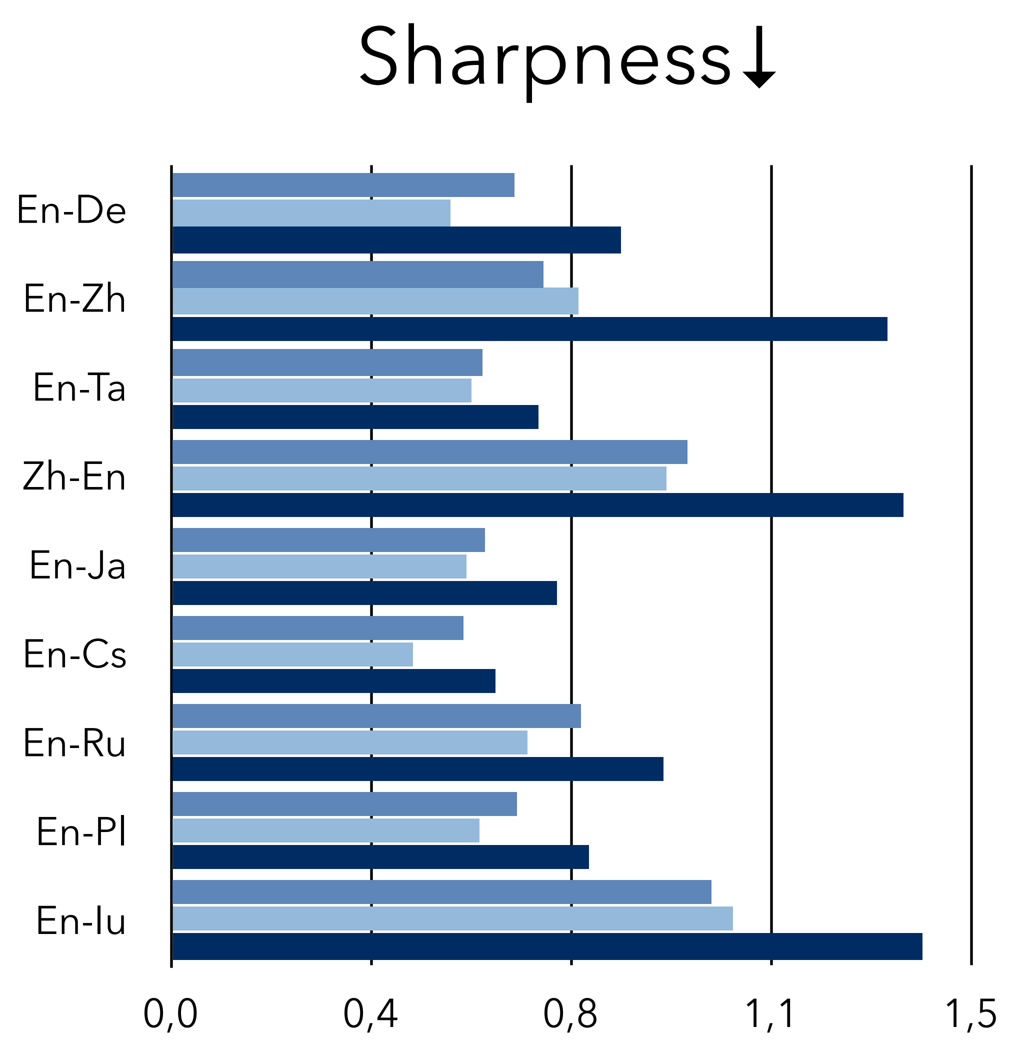
Experiment 1 – Results for segment-level DA predictions



Experiment 1 – Results for segment-level DA predictions

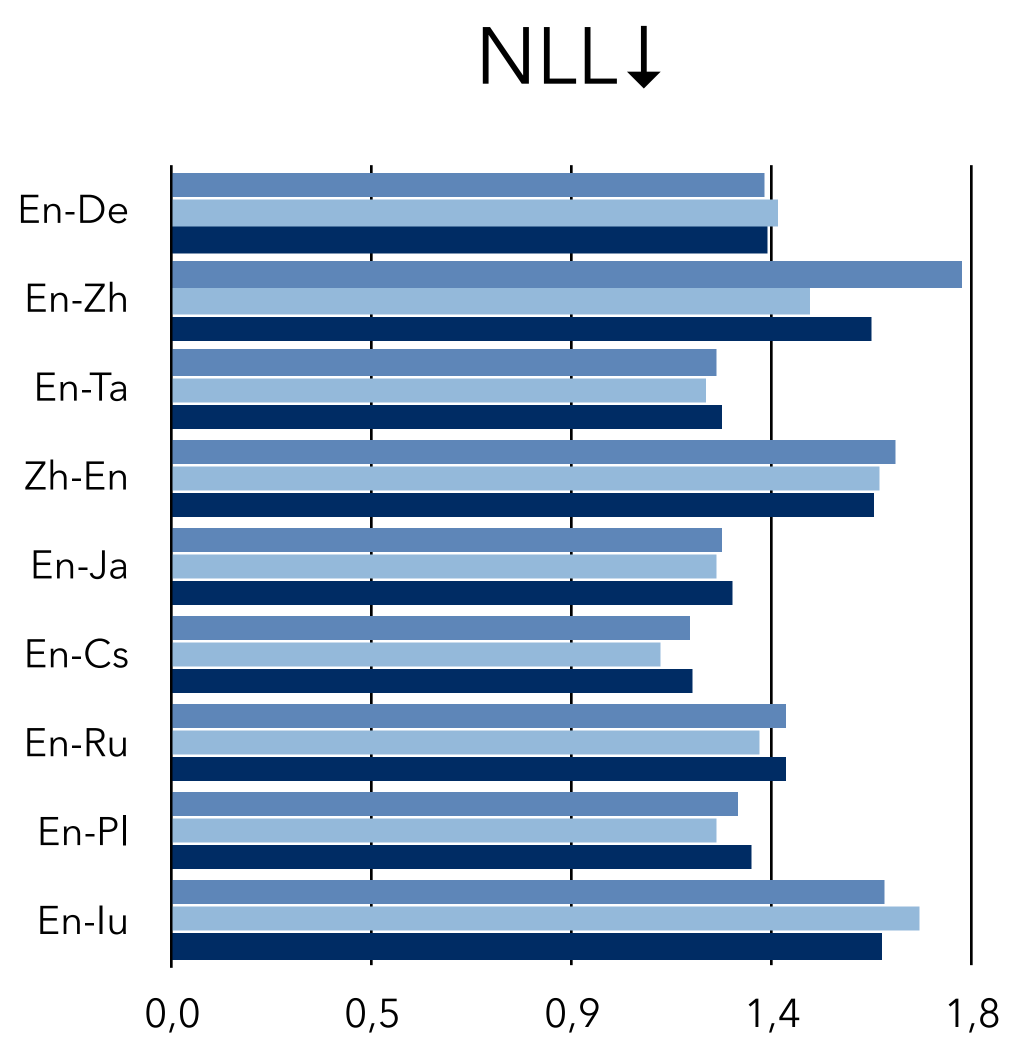


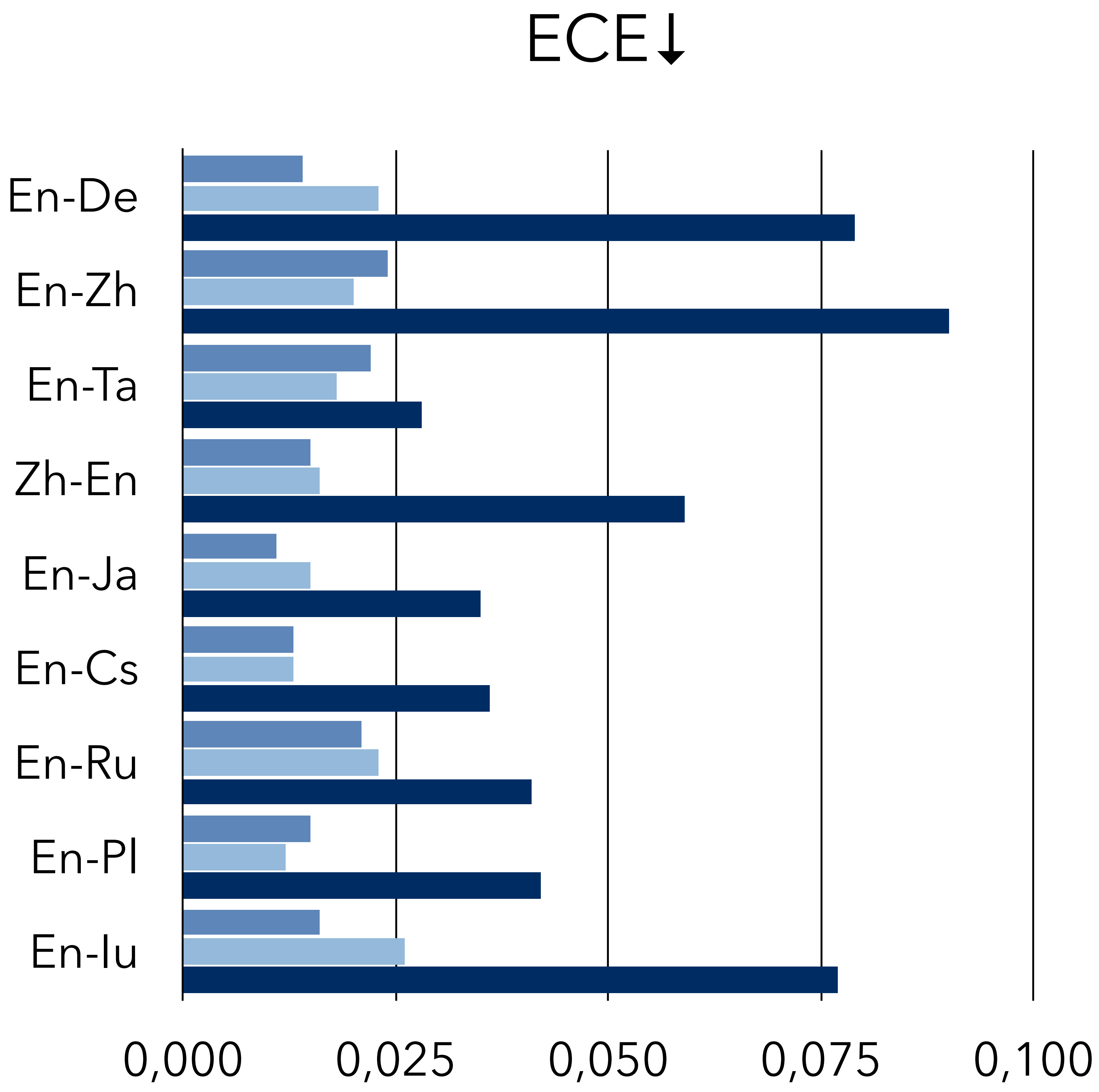


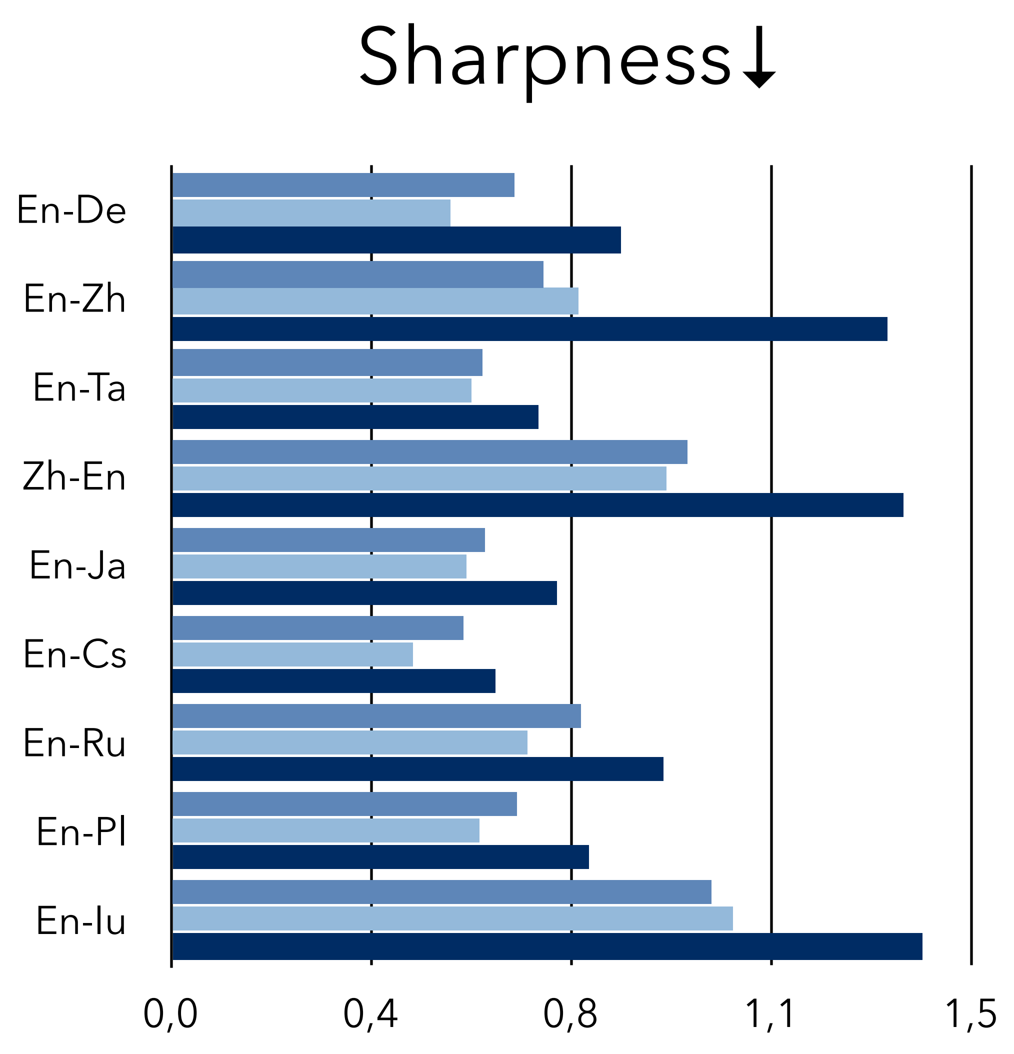


MCD DE Baseline

Experiment 1 – Results for segment-level DA predictions

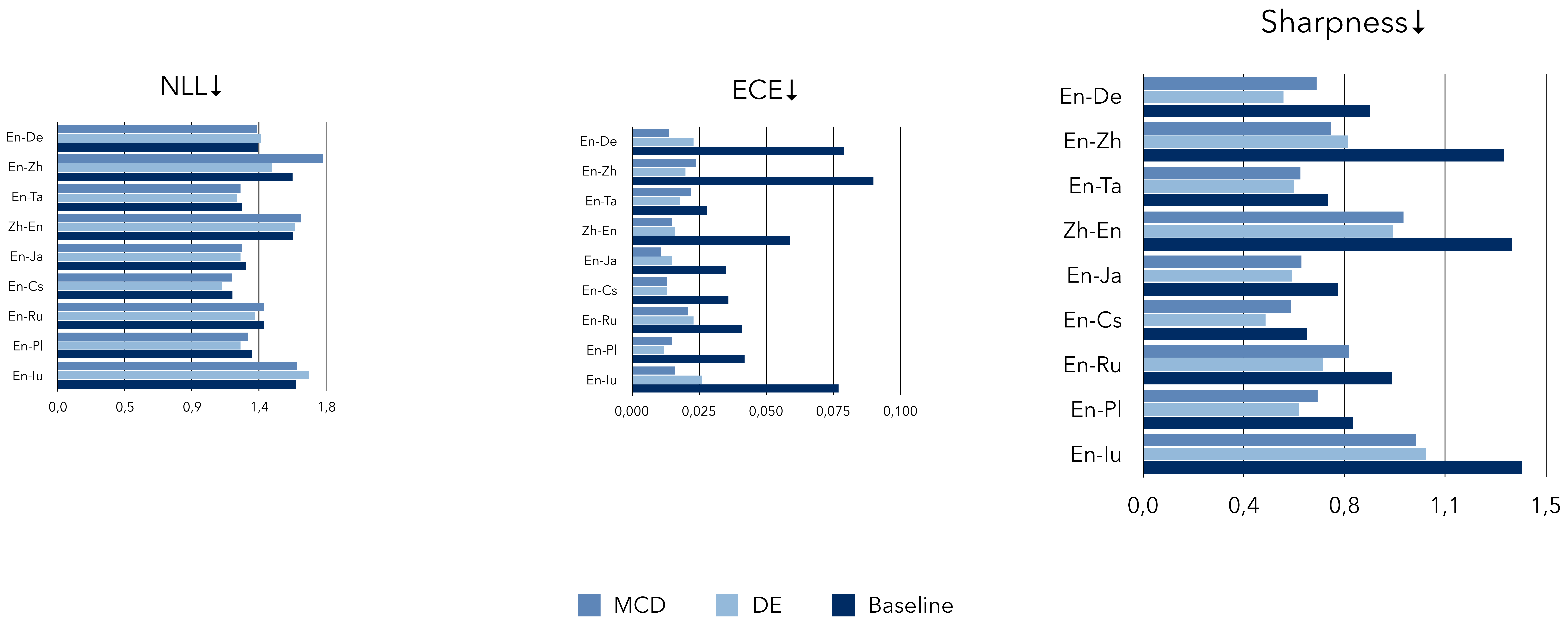






MCD DE Baseline

Experiment 1 – Results for segment-level DA predictions



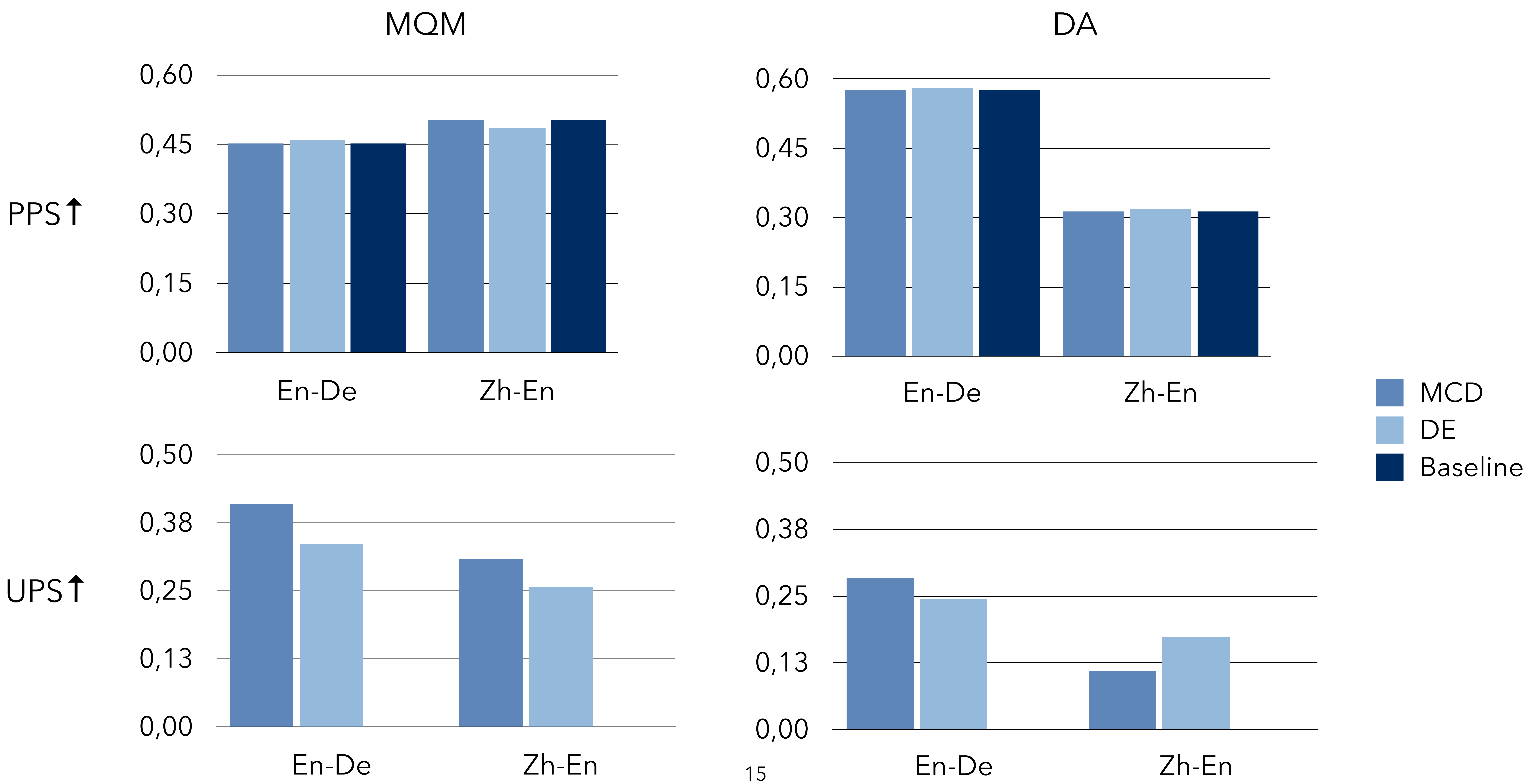
Experiment 1 – Results for segment-level DA predictions

MCD and DE show **consistent improvement** over the baseline in all metrics and LPs

DE provide more accurate predictions and narrower confidence intervals

MCD is cheaper and competitive to DE performance

Experiment 1 – Results for segment-level MQM predictions



Experiment 2 – Multi-reference

Impact of reference quantity

Goal:

Simulate access to multiple references of varying quality

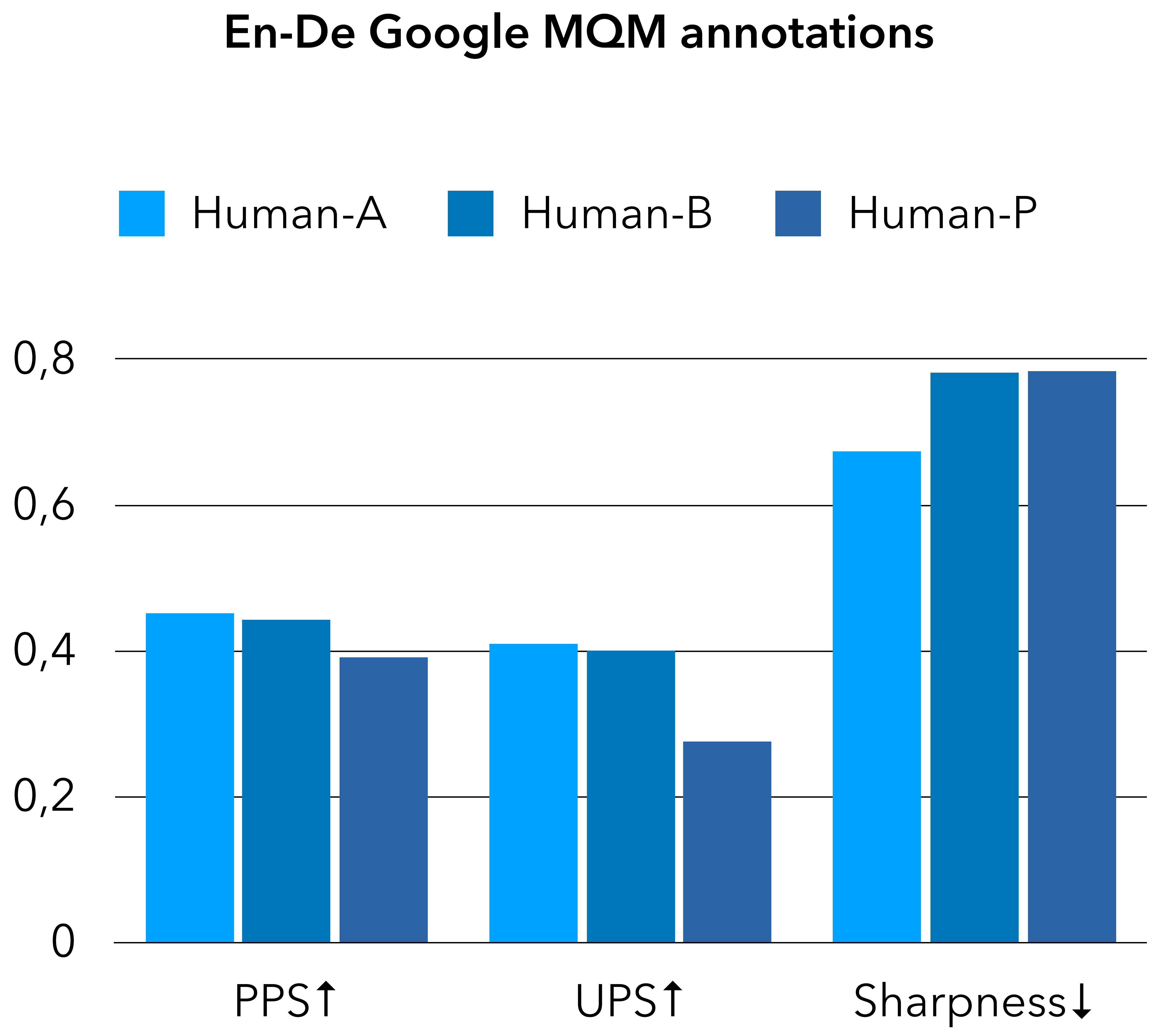
Hypothesis:

More references == Less uncertainty

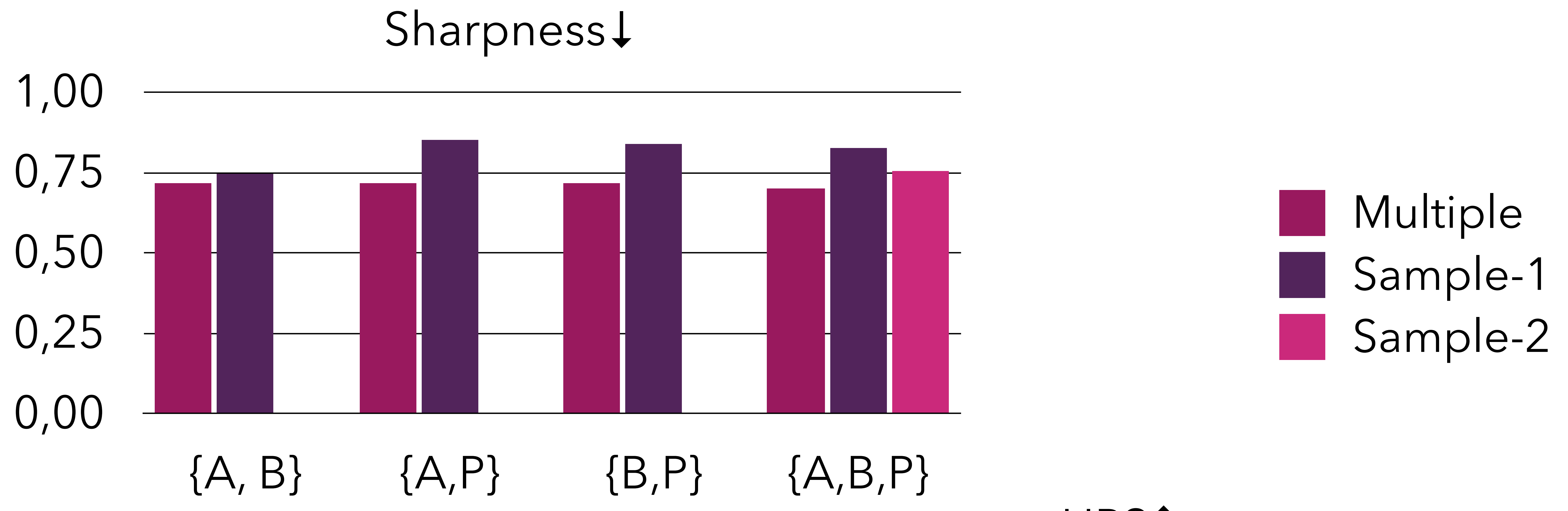
Experimental setup:

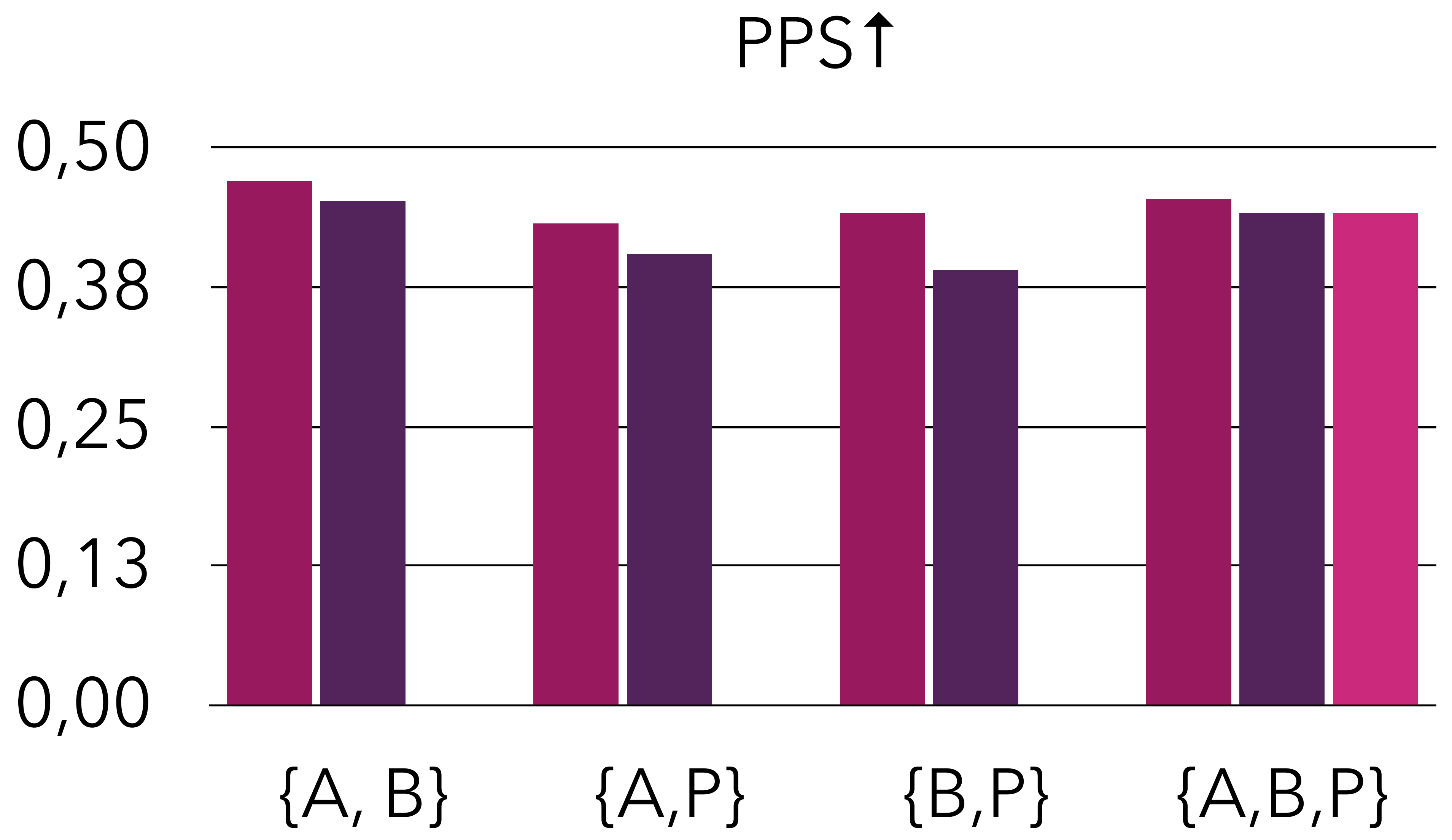
Compare

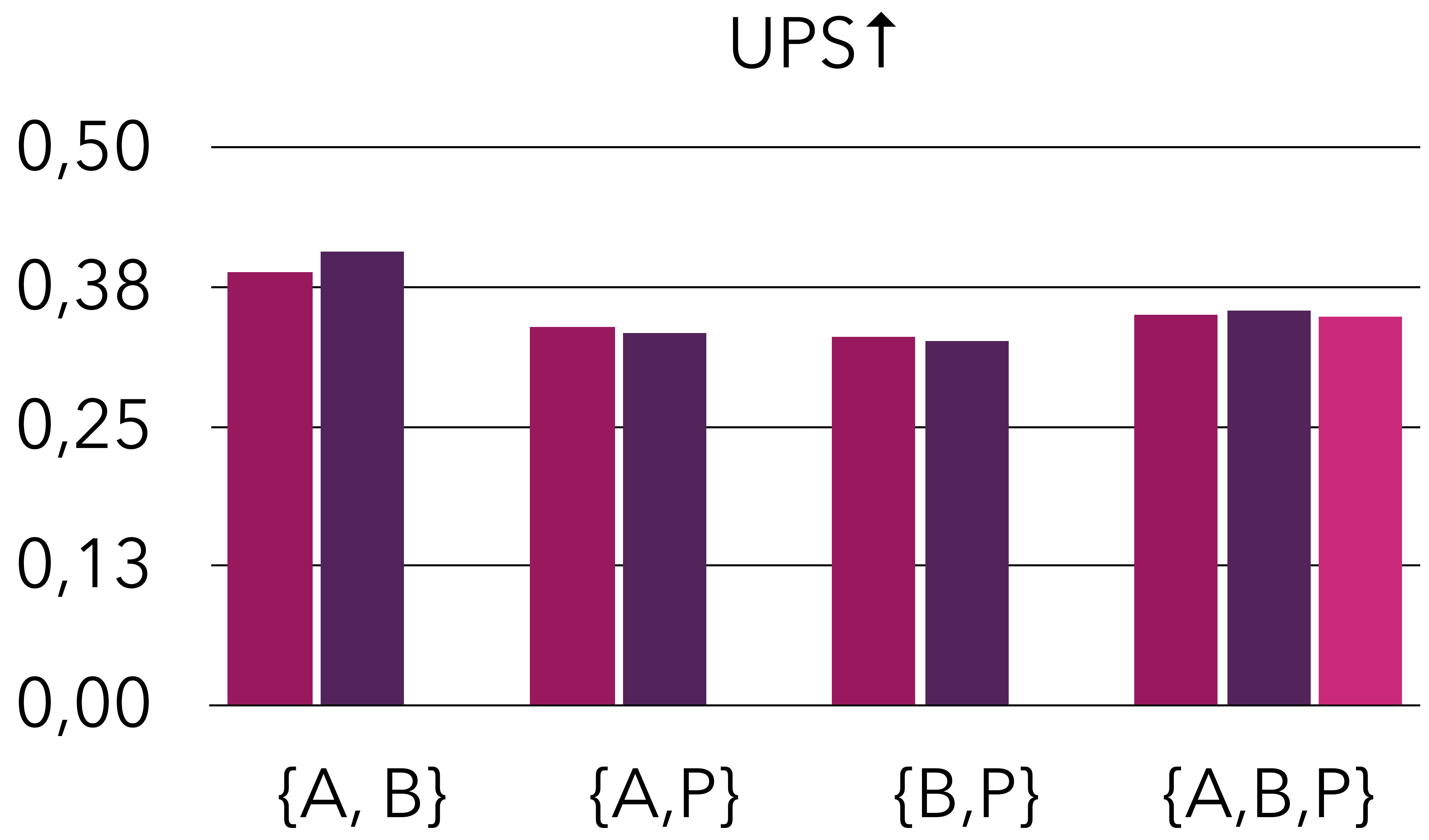
- **S-1** – Sampling single references
- **S-2** – Sampling pairs of references
- **MUL** – Combining all available references in *R* (averaging)



Experiment 2 – Multi-reference







Experiment 3 – Critical translation mistakes

Goal:

Improve retrieval of critical translation errors

Dataset:

WMT20, DA and MQM

Experimental setup:

- Rank segments by normalised MQM scores
- Normalize for MT length
- Target the lowest N%
- Assume no references --> Pseudo-references (PRISM)

Hypothesis:

We can use the cumulative distribution function over Q for each $\langle s, t, \mathcal{R} \rangle$ to predict $P(Q \leq q_{\text{err}})$

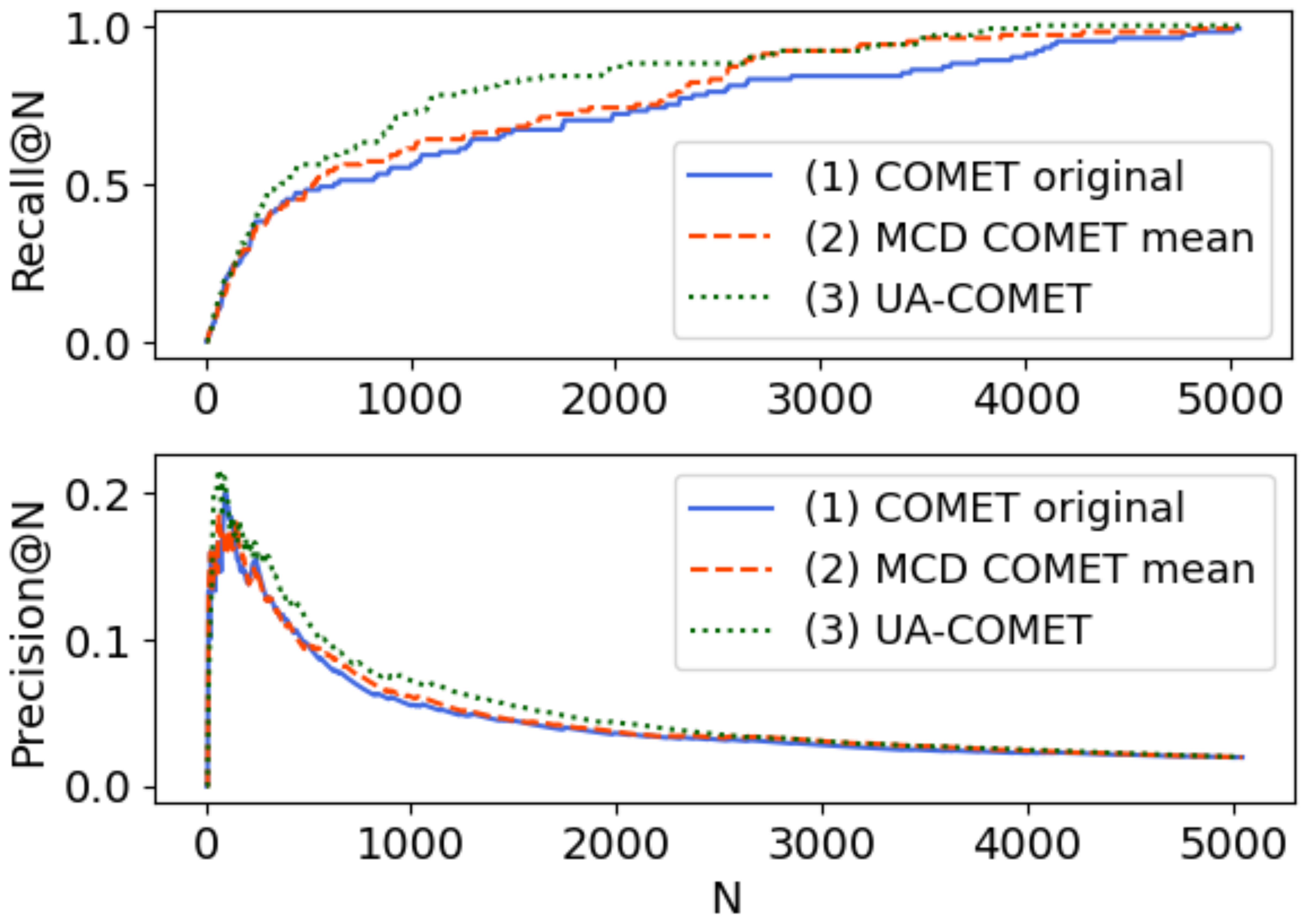
q_{err} – tuned threshold for Recall@N

Experiment 3 – Critical translation mistakes

UA-COMET:
- Better Recall as N increases
- Better Precision for small N

Overall:
- Recall & Precision are low for small N
- Room for improvement for all 3 methods

2% lowest quality MTs for En-De MQM



Conclusions

- A simple strategy for making MT evaluation metrics **uncertainty-aware**:
  - MC Dropout
  - Deep Ensembles
- UA-COMET matches COMET's prediction accuracy
  - but is **informative towards the reliability of the predicted quality scores**
- When number of (reliable) references **increases**, confidence intervals **shrink**
  - but bad references may be harmful!
- Confidence intervals show potential in detecting critical MT mistakes
- **Future work:** more sophisticated techniques for uncertainty quantification

Thank you!

taisiya.glushkova@tecnico.ulisboa.pt
@glushkovato

chrysoula.zerva@tecnico.ulisboa.pt
@chryssaZrv

ricardo.rei@unbabel.com
@RicardoRei7

andre.t.martins@tecnico.ulisboa.pt
@andre_t_martins