You Cannot Feed Two Birds With One Score

Samvit Mavinkurve
Mentor: Till Aczel

German Sentence: Meine Lehrerin ist nett.

Translation B: My teacher is nice. A

Translation B: My female teacher is nice. F

German: Meine Lehrerin ist nett.

Translation A: My teacher is nice.

Translation B: My female teacher is nice.

A

F

You Cannot Feed Two Birds with One Score: the Accuracy-Naturalness Tradeoff in Translation

Gergely Flamich* Imperial College London g.flamich@imperial.ac.uk

David Vilar **Jan-Thorsten Peter** **Markus Freitag** Google {vilar, jtp, freitag}@google.com

Motivation

- MT Evaluation traditionally cared about:
  - Accuracy
  - Fluency
- Modern MT often compresses into a single score.
- WMT24 observations:
  - best automatic metric systems ≠ best human-rated systems
- Is this:
  - Imperfect metrics?
  - Or a fundamental limitation?

Types of metrics

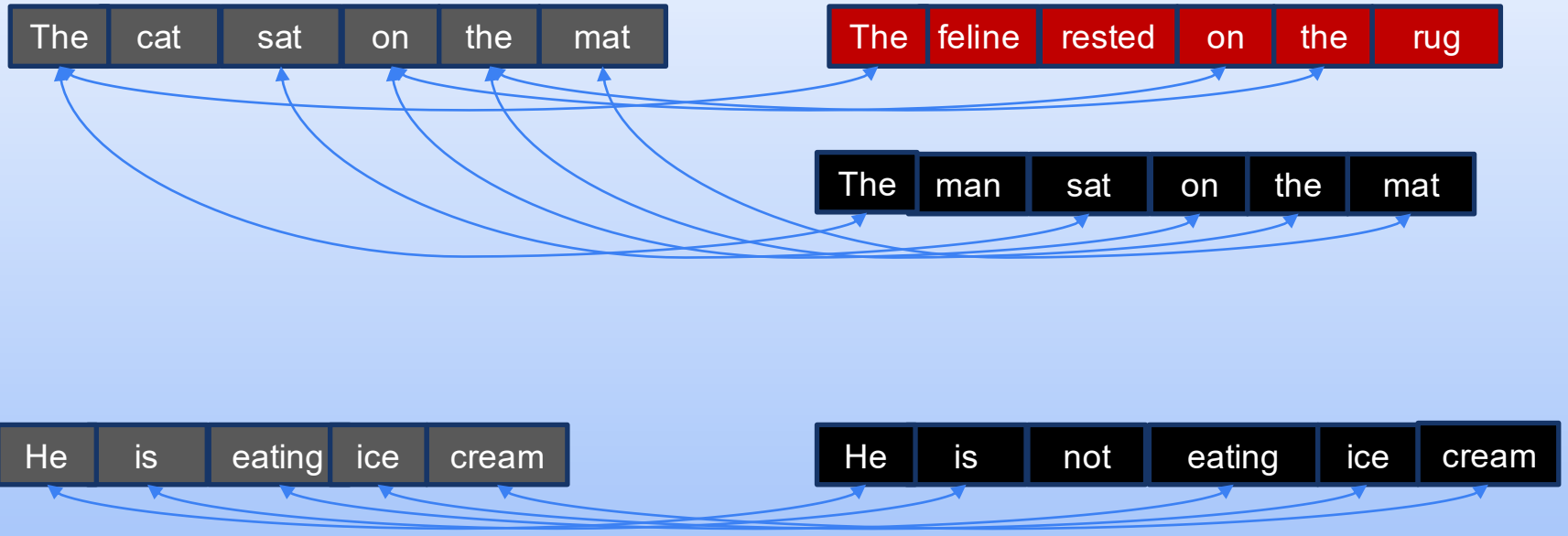
Two types of metrics:

- Accuracy metrics (Reference metrics)
- Fluency Metrics (No-Reference Metrics)

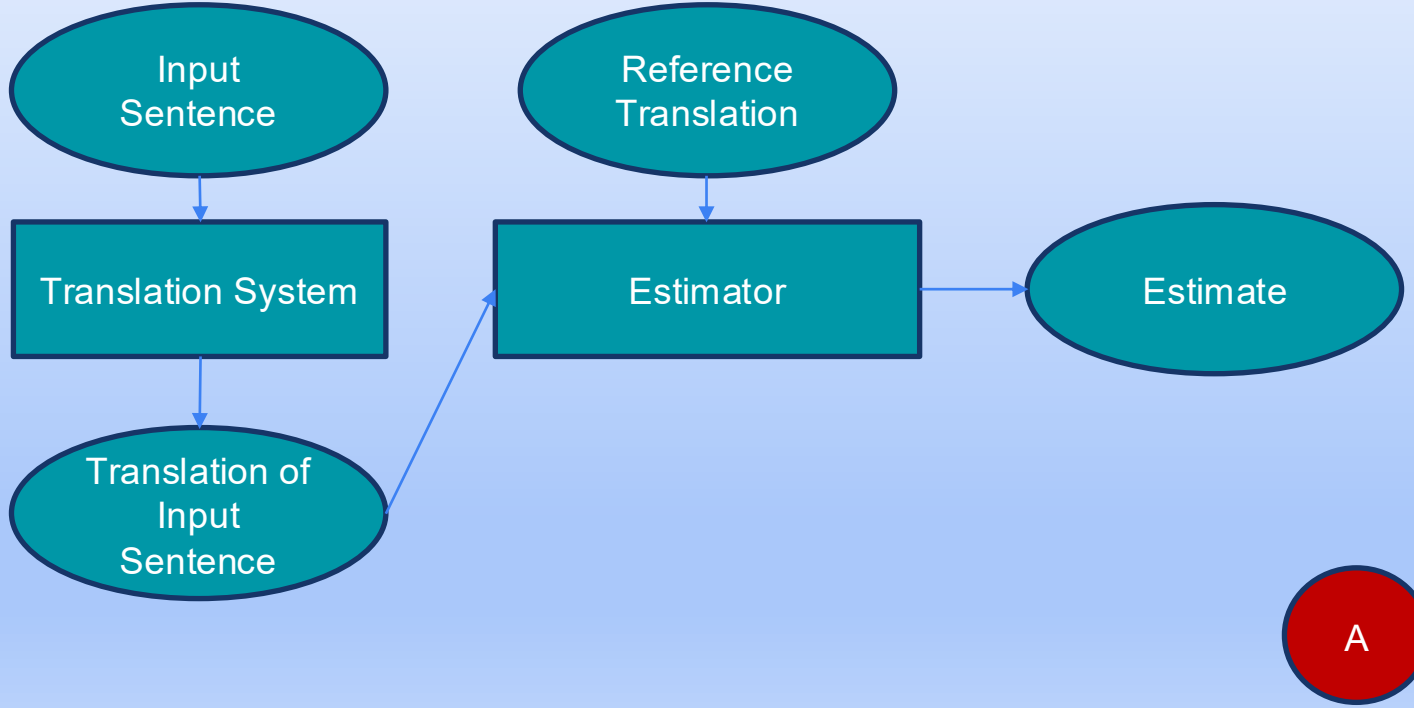
BLEU

- BLEU looks at geometric mean of n-grams and multiplies it by a penalty for brevity, ie, outputs that are too short are penalized.

BLEU Failures

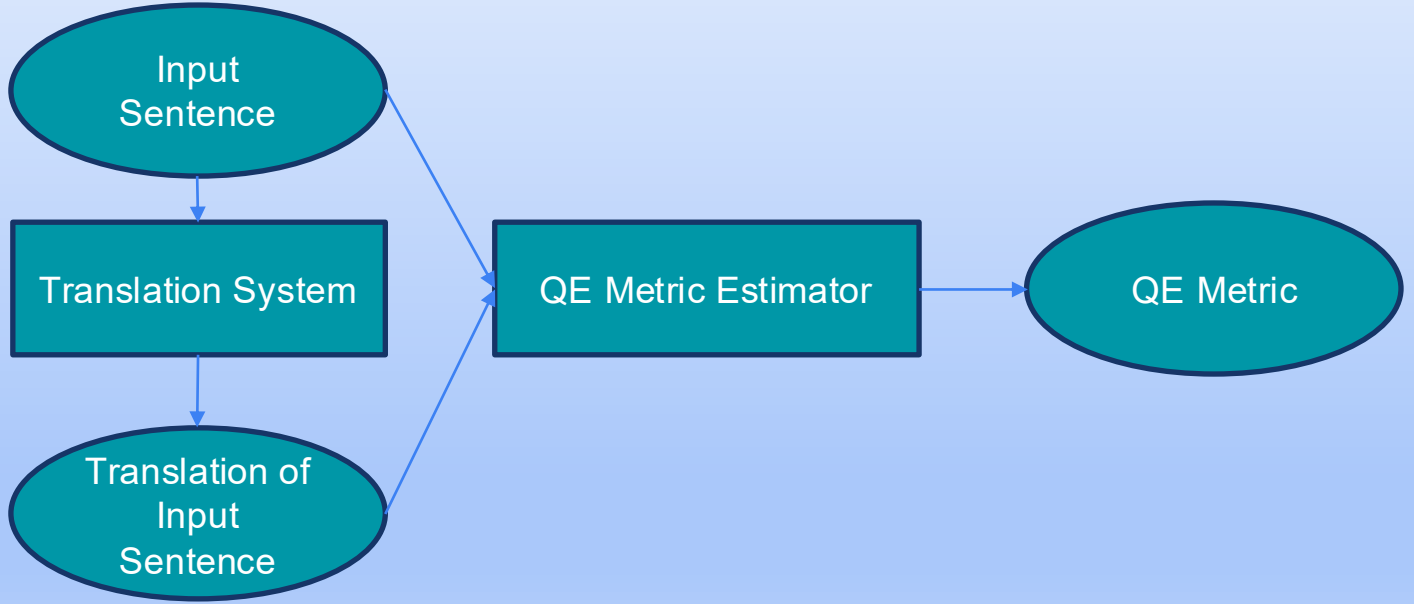


Neural Metrics



QE Metrics

- Tries to estimate quality without needing a translation



Naturalness metrics

- Evaluate how natural or fluent sentence is.
- Examples include perplexity and log perplexity.

Some Notation

- $\boldsymbol{\mathcal{X}}$ is the set of all possible source texts and $\boldsymbol{\mathcal{Y}}$ is the set of possible target texts
- $\mathbf{x} \in \boldsymbol{\mathcal{X}}$ is a randomly selected source text
- $\boldsymbol{y} \in \boldsymbol{\mathcal{Y}}$ is a randomly selected target text
- Source target pairs are drawn from a joint probability distribution $P_{\boldsymbol{X},\boldsymbol{Y}}$.
- $P_{\boldsymbol{Y}}$ is distributions of human translations into the target language
- P_X is the marginal over $P_{\boldsymbol{X},\boldsymbol{Y}}$ and represents the true monolingual distribution in the original language.

$$\mathrm{P_X} = \sum_{\mathrm{y}} \mathrm{P_{X.Y}(x, y)} = \mathbb{E}_{\mathrm{Y}}[\mathrm{P(X = x|Y)}]$$

Some Notation

- $P_{\mathrm{Y|X}}$ is the distribution of human translations into the target language given the source sentence
- $\mathrm{R_Y}$ is the “true” monolingual distribution in the target language
- A translation system is modelled as $Q_{\mathrm{Y}}(\cdot) = \mathbb{E}_{\boldsymbol{x}\sim P_{\mathrm{X}}}[Q_{\boldsymbol{y}|\boldsymbol{x}}(\cdot)]$and thus a translators marginal is defined as $\mathrm{Q_{Y|X}}$

A

Defining Accuracy Formally

- A distortion metric is a function $\Delta: \mathcal{X} \times \mathcal{Y} \times \mathcal{Y} \to \mathcal{R}$ such that:
  - It takes as an input the source sentence x, the candidate translation y^{c}, and the reference translation y^{r}.
  - Δ is bounded from below.
  - Δ is minimized when x or y^{c} is semantically equivalent to y^{r}.
- The accuracy then is:

$$\mathrm{A}(Q_{\boldsymbol{y}|\boldsymbol{x}}) = -\mathbb{E}_{\mathrm{x},\mathrm{y}^{\mathrm{r}} \sim \mathrm{P}_{\boldsymbol{x},\mathrm{y}}}[\mathbb{E}_{\mathrm{y}^{\mathrm{c}} \sim Q_{\boldsymbol{y}|\boldsymbol{x}}}[\boldsymbol{\Delta}(\mathrm{x}.\,\mathrm{y}^{\mathrm{r}}, \mathrm{y}^{\mathrm{c}})]]$$

Defining Naturalness formally

- A divergence function $\mathrm{D}(\mathrm{Q}, \mathrm{R})$ for any two probability distributions Q and R is such that
  - $\mathrm{D}(\mathrm{Q}, \mathrm{R}) \geq 0$
  - $\mathrm{D}(\mathrm{Q}, \mathrm{R}) = 0$ if and only if $\mathrm{Q} = \mathrm{R}$
- Examples of Divergence functions are KL-Divergence, Wasserstein distance, etc.
- A non-reference metric can be defined as:

$$\mathrm{N}(Q_{y|x}) = -\mathrm{D}(Q_y, \mathrm{R}_y)$$

Showing the Tradeoff

For perfectly accurate: $Q_Y^* = \underset{Q_{Y|X}}{\operatorname{argmax}} A(Q_{Y|X})$

For perfectly natural: $Q_Y^* = R_Y$

These two conditions cannot happen together as the natural distribution in a language cannot be identical to the distribution on the translation language.

Studying the Trade-off

- Consider the accuracy-naturalness(AN) function:

$$\mathrm{A(N)} = \max_{Q_{y|x}} \{\mathrm{A(Q_{y|x})}\} \text{ subject to } -\mathrm{D(Q_y, R_y)} \geq \mathrm{N}$$

- A(N) represents the best accuracy we can ever expect for a translation system, subject to achieving a naturalness of at least N.

Studying the Trade-off

- Let A(N) be the accuracy-naturalness function defined by distortion Δ, distributional distance D, parallel distribution $\mathrm{P}_{\mathrm{x,y}}$ and monolingual reference R_{y}. Then:
  1. A(N) is non-increasing in N.
  2. If D is convex in its first slot, then A(N) is concave in N.
- Higher Naturalness may require sacrificing some accuracy and vice-versa.
- As you demand more and more naturalness, each extra improvement in naturalness costs increasingly more accuracy.
- The trade-off defines a Pareto frontier.

Obtaining Naturalness from metrics

- Multiple critics: $C = \{f_1, f_2, \dots, f_K\}$
- Each critic assigns a score to generated outputs

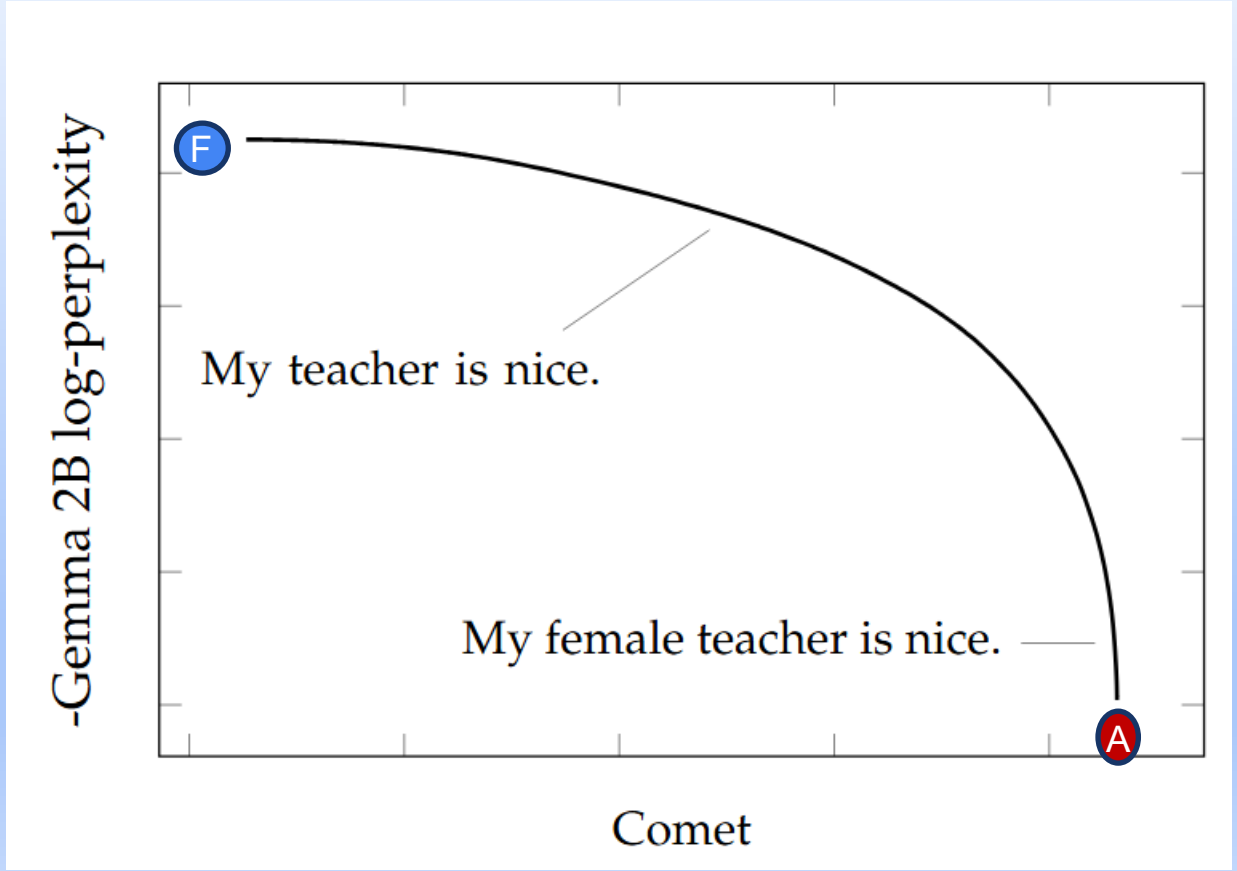
$$D_1(Q_y^m, R_y|\mathcal{P}) = \mathbb{E}_{f\sim\mathcal{P}}\left[\left|\mathbb{E}_{Y\sim Q}[f(Y)]\right| - \left|\mathbb{E}_{Y\sim R}[f(Y)]\right|\right]$$

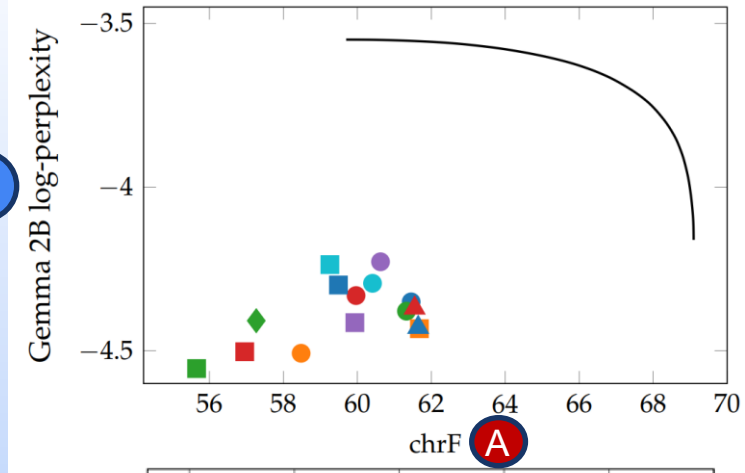
- It can also be shown that:

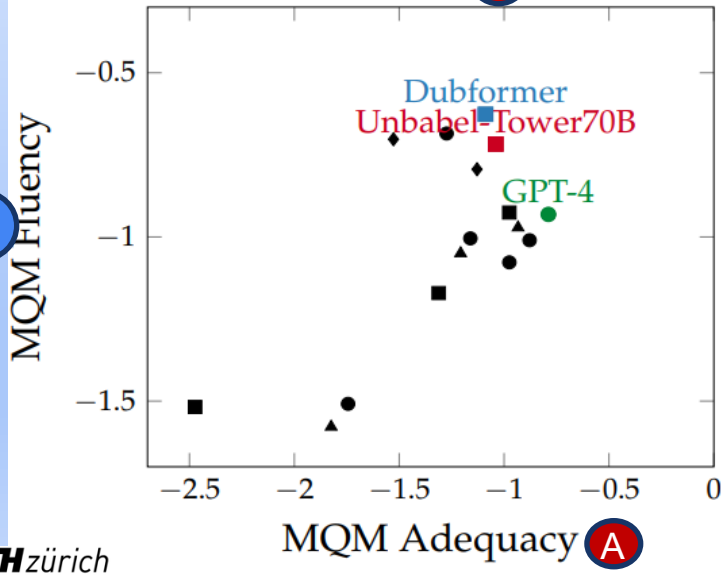
$$D_1(Q_y^m, R_y|\mathcal{P}) \approx (KN)^{-1}\sum_{k=1}^{K}\sum_{y\in Y^m} f_k(\boldsymbol{y})$$

F

The Accuracy Naturalness curve

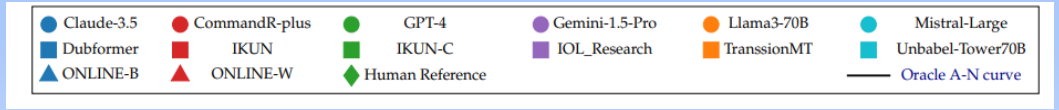






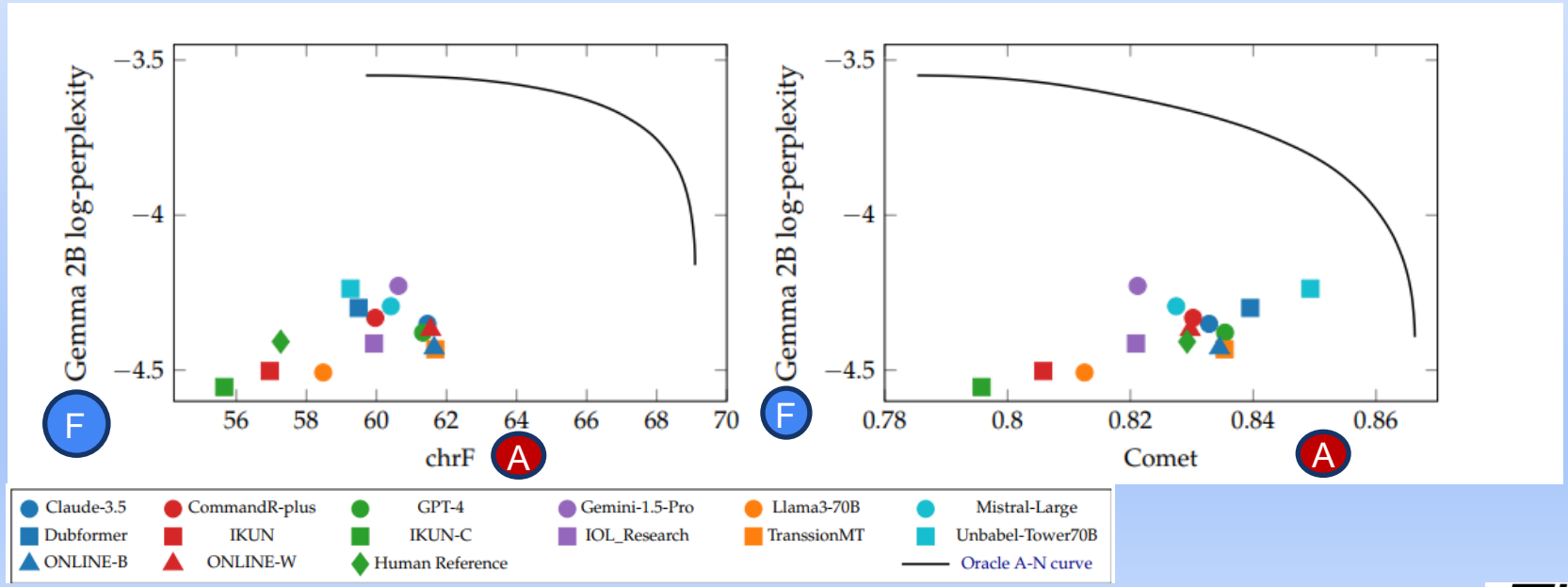
The Tradeoff

- All models are below the curve.
- The models which are best performing according to human evaluation lie closest to the curve.
- As one comes from the point furthest from the origin towards it, accuracy and fluency correlate until a point after which they anti correlate.



How Different metrics affect tradeoff?

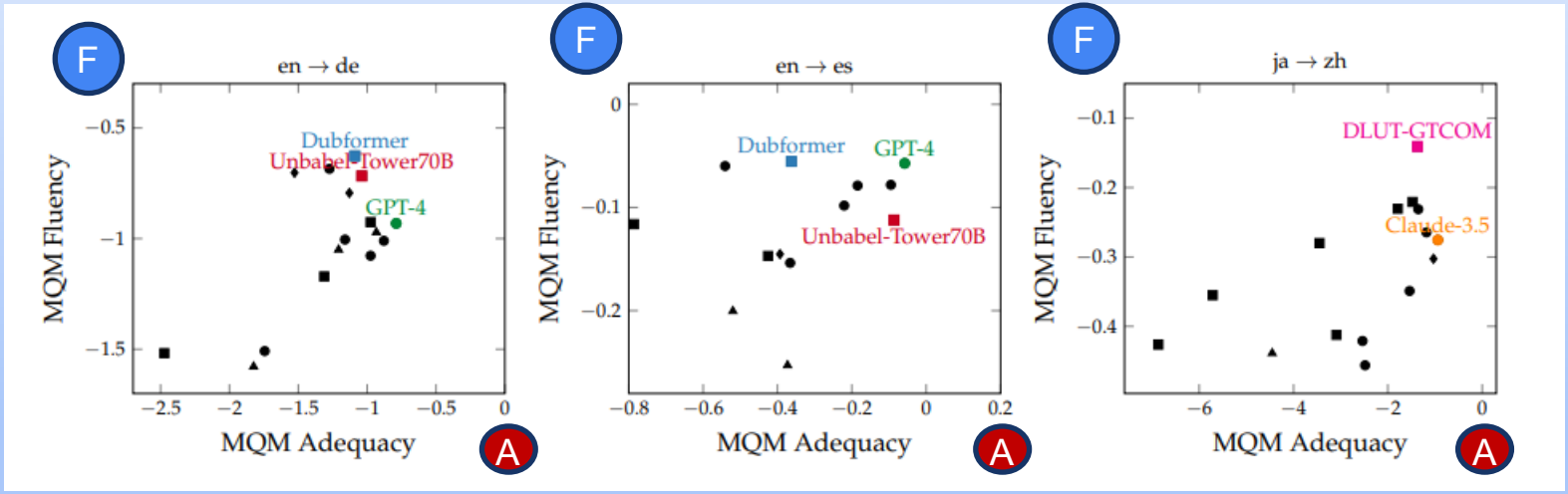
- COMET correlates for longer than chrF, and thus its tradeiff is much milder.



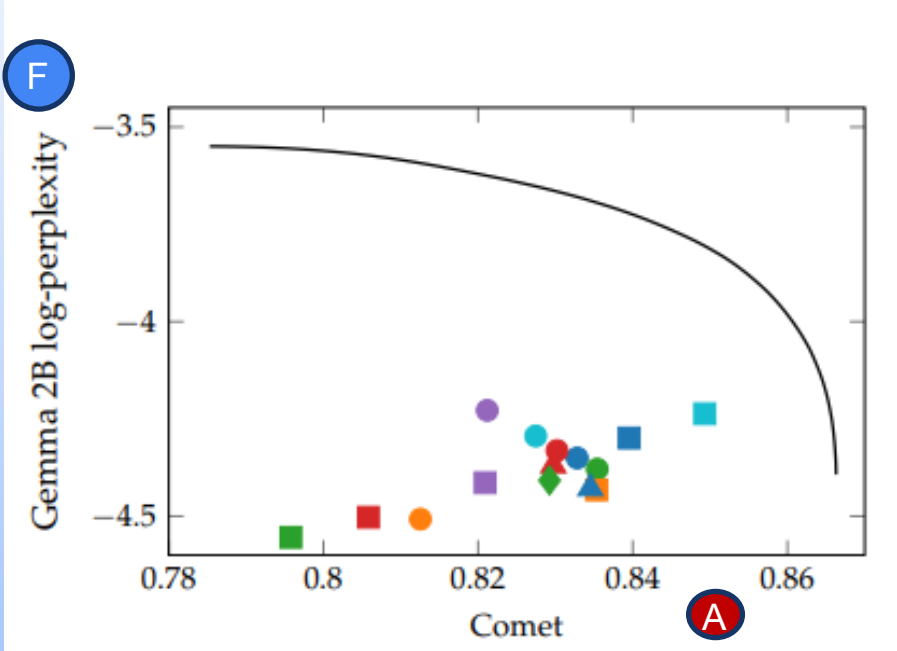
ETH zürich

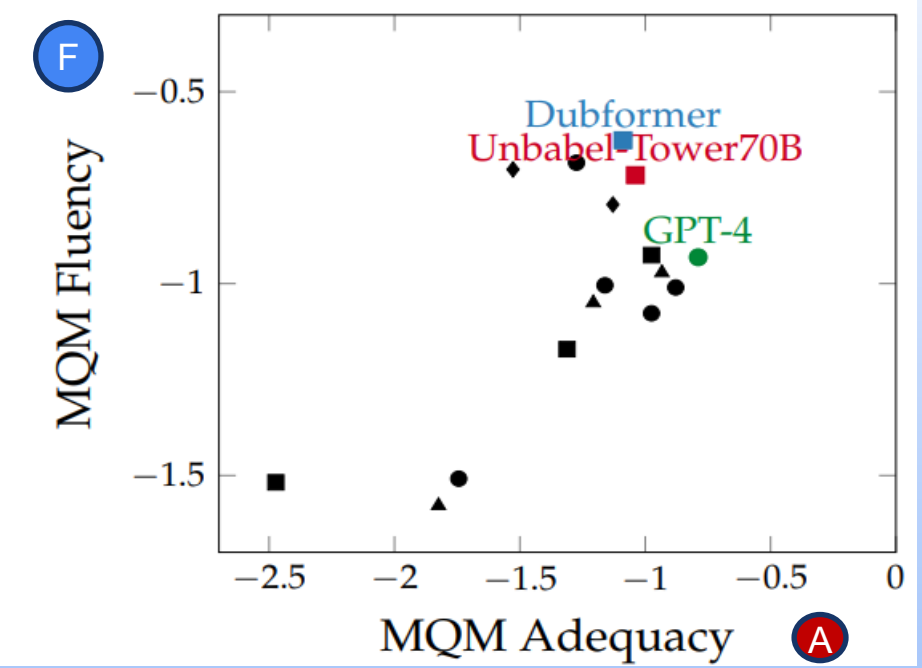
How does the tradeoff work across languages?

- English to german and japenese to Chinese, exhibits a more pronounced tradeoff compared to English to Spanish.



What can this tell us about individual metrics?





- Unbabel was trained on COMET.
- Frontier on the Gemma-2B-COMET curve.
- High MQM Adequecy but fluency suffers.

Implications and Future Work

- There is a fundamental tradeoff between accuracy and fluency.
- We may want prioritise or have a different preference or utility for the amount of inaccuracy or unnaturalness we may want to tolerate.
- Optimising the tradeoff directly.

Thank you!

Questions?