Deploying MT Quality Estimation on a large scale: Lessons learned and open questions

Aleš Tamchyna
ales.tamchyna@memsource.com

MEMSOURCE

OUTLINE

- MTQE in Memsource
- Defining “quality” in QE
- Academic tasks vs. applications
- Reference-free metrics for MTQE
- What factors affect post-editing effort?
- Conclusion

ABOUT MEMSOURCE

- Cloud-based translation management system (TMS).
- Includes translation editors (CAT tool).
- Diverse customer base:
  - Freelance translators.
  - Language service providers (LSPs).
  - Enterprises (Uber, Supercell, ...).

MTQE IN MEMSOURCE

Predict MT quality category for each input segment:

- **100** (perfect), **99** (near perfect), **75** (high quality) or **0** (low quality)

Source: Machine translation is hard.
MT: Strojový překlad je těžký.

Internally, MTQE is a classifier based on a deep neural network, trained on large-scale datasets of MT outputs and their post-edits.

MTQE IN MEMSOURCE

Use cases:

- Predict overall savings thanks to MT before manual translation.
- Help translators choose when to start from scratch and when to post-edit the MT output.
- Routing: high-quality translations may even skip manual post-editing.
- Calculate translator compensation.

Interesting facts:

- First version deployed in 2018.
- We process around 10 million segments monthly.
- We support over 130 language pairs, models are updated every month with new data.

WHAT IS QUALITY?

- HTER represents the **amount of post-editing** required.
- Direct assessment (DA) represents overall **quality** as perceived by human annotators.

HTER and DA may not correlate very much and DA may be somewhat easier to predict, see Fomicheva et al. (2020) and Specia et al. (2020).

Which metric to choose probably depends on the use case.

At Memsource, we use a customized version of **(H)chrf3**.

- Essentially post-editing effort but more robust w.r.t. tokenization, morphology,...

ACADEMIC TASKS VS. REAL WORLD APPLICATIONS

WMT QE Shared tasks are a standard benchmark.

- How well do they capture realistic settings and challenges?
- Some doubts outlined by Sun et al. (2020) -- imbalanced score distributions, statistical artifacts, able to perform well when looking only at source or MT.

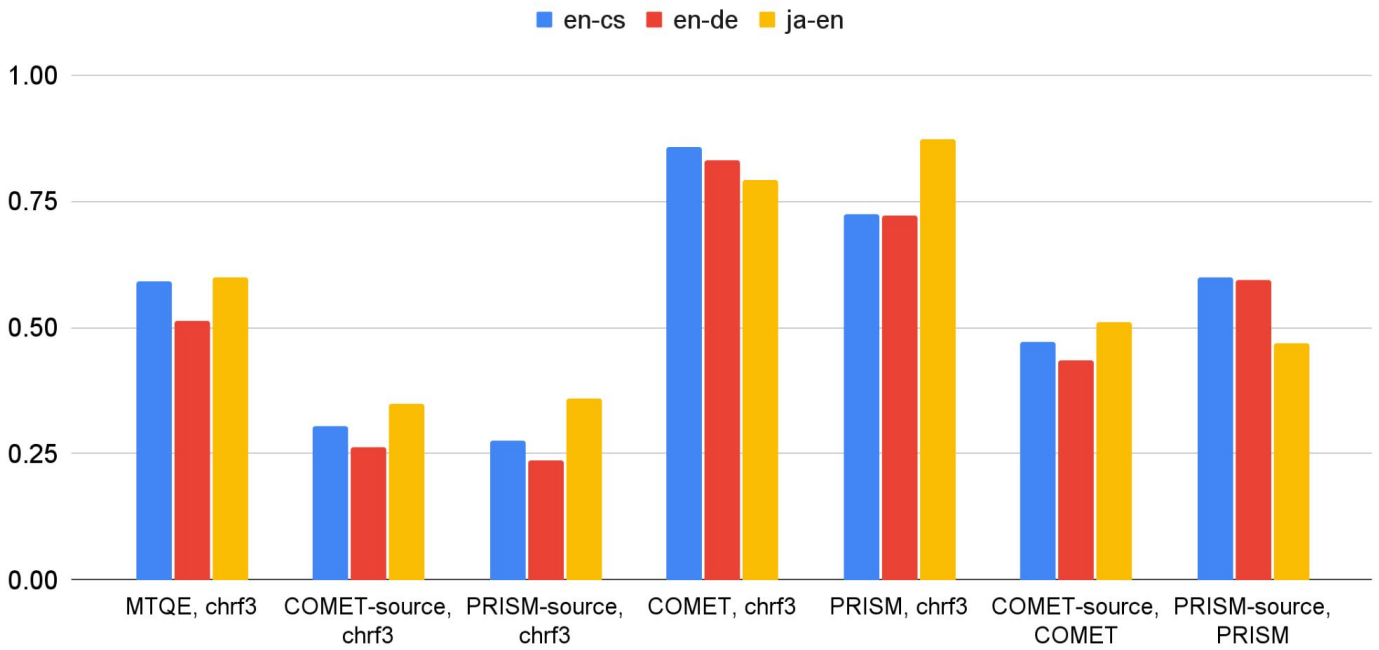
	WMT	**In practice**
Training data	~10^3 sentences	~10^6 sentences
Domains	Few	Many
Quality target	Fixed	Varied

Good systems for WMT may not necessarily perform well in practical settings.

- On our datasets, fine-tuned multilingual pre-trained models work comparably or better than QE-specific approaches.

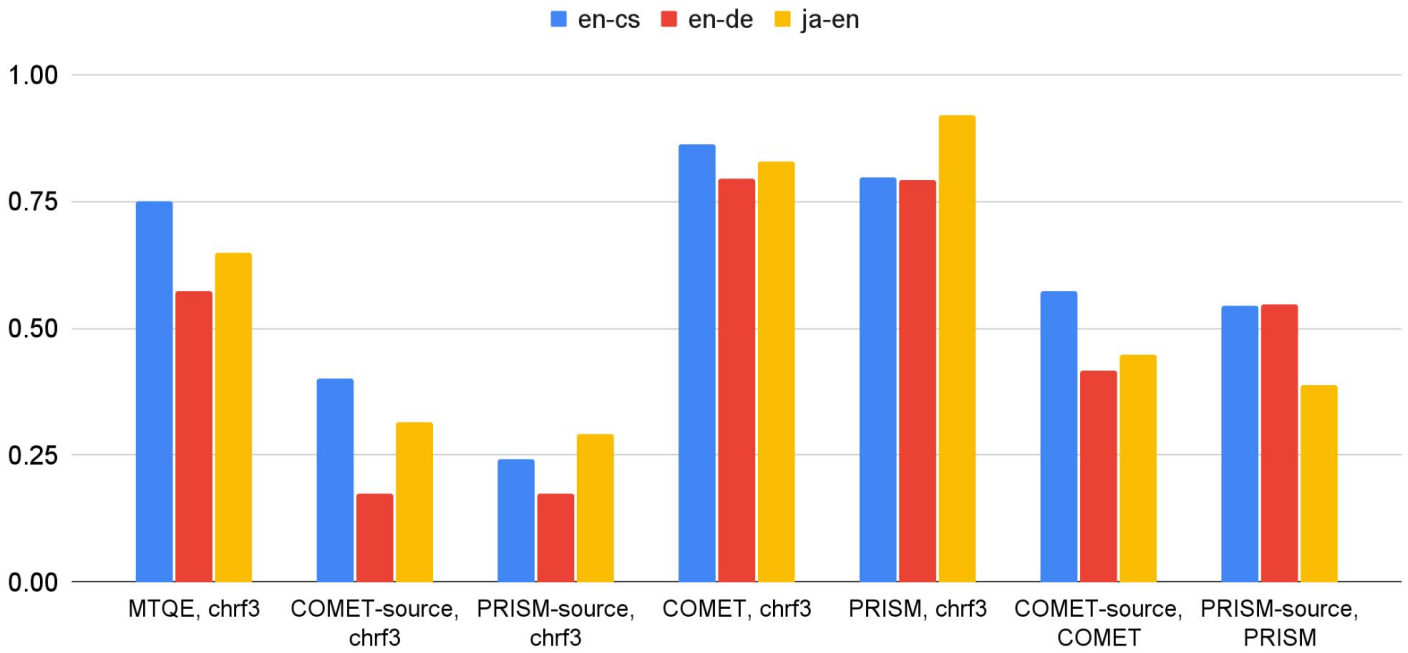
REFERENCE-FREE METRICS FOR MTQE

Segment-level evaluation, Spearman correlation



REFERENCE-FREE METRICS FOR MTQE

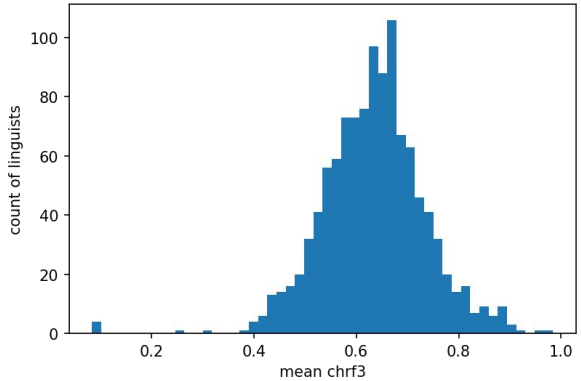
Document-level evaluation, Spearman correlation



MT QUALITY IS NOT ONLY ABOUT MT

Various factors play a role:

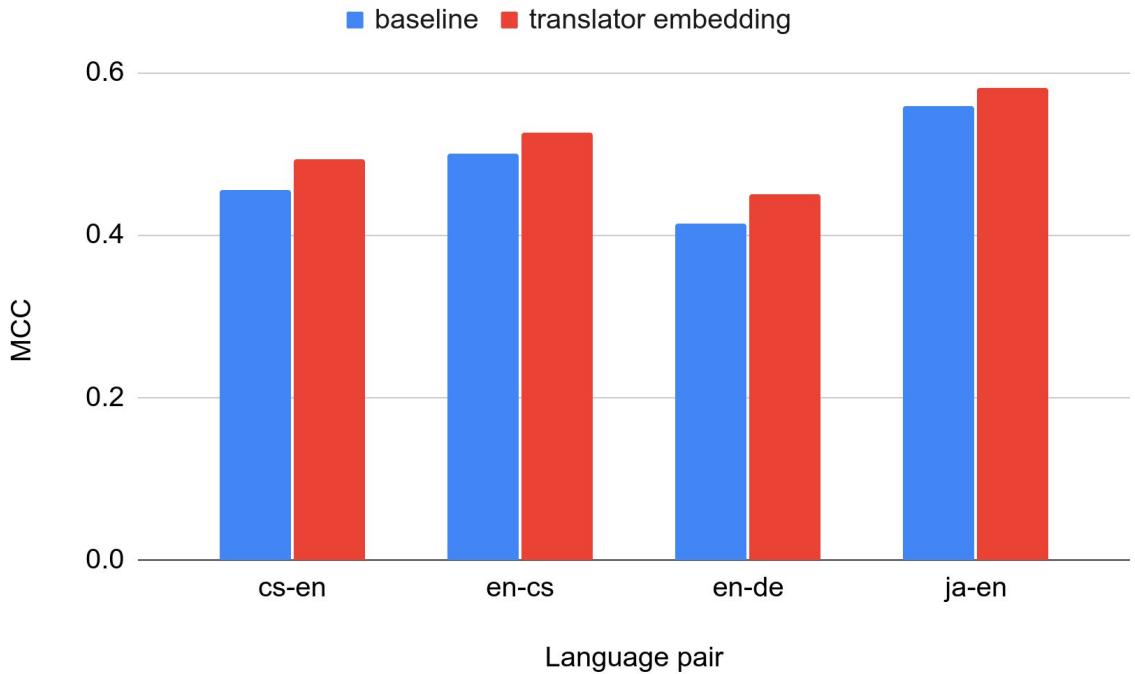
- Customer and domain.
- Customer budget.
  - Is light post-editing okay?
  - Will there be (multiple rounds of) manual revisions?
- Translator attitude towards MT.
  - Some translators like to overedit, others like to underedit the MT outputs.



In a way, when we get a post-edited translation, we're really getting just a **random sample from some distribution of possible post-edits**. This distribution may have quite a large variance.

- Corollary: completely accurate MTQE is impossible.

EFFECT OF POST-EDITORS



CONCLUSION

- MT quality has various definitions.
- Results on academic tasks do not always translate to real-world performance.
- Post-editing effort is influenced by various factors.
  - Translator attitude plays an important role.
- As MT systems approach human quality, we may need to revisit the definition of MTQE entirely.

REFERENCES

- Agrawal, S. et al. (2021). Assessing Reference-Free Peer Evaluation for Machine Translation. https://arxiv.org/abs/2104.05146
- Graham, Y. et al. (2016): Is all that Glitters in Machine Translation Quality Estimation really Gold? https://aclanthology.org/C16-1294/
- Fomicheva, M. et al. (2020). MLQE-PE: A multilingual quality estimation and post-editing dataset. https://arxiv.org/pdf/2010.04480.pdf
- Specia, L. et al. (2020). Findings of the WMT 2020 Shared Task on Quality Estimation. https://aclanthology.org/2020.wmt-1.79/
- Sun, S. et al. (2020). Are we Estimating or Guesstimating Translation Quality? https://aclanthology.org/2020.acl-main.558/

Q&A

MEMSOURCE

THANK YOU