Unbabel

6th October 2020

AMTA 2020

COMET - Deploying a New State-of-the-art MT Evaluation Metric in Production

Craig Stewart

Research Scientist

Unbabel AI Metrics

Craig Stewart

Research Scientist

craig,stewart@unbabel.com

Ricardo Rei

Research Engineer

ricardo.rei@unbabel.com

Catarina Farinha

Research Engineer

catarina.farinha@unbabel.com

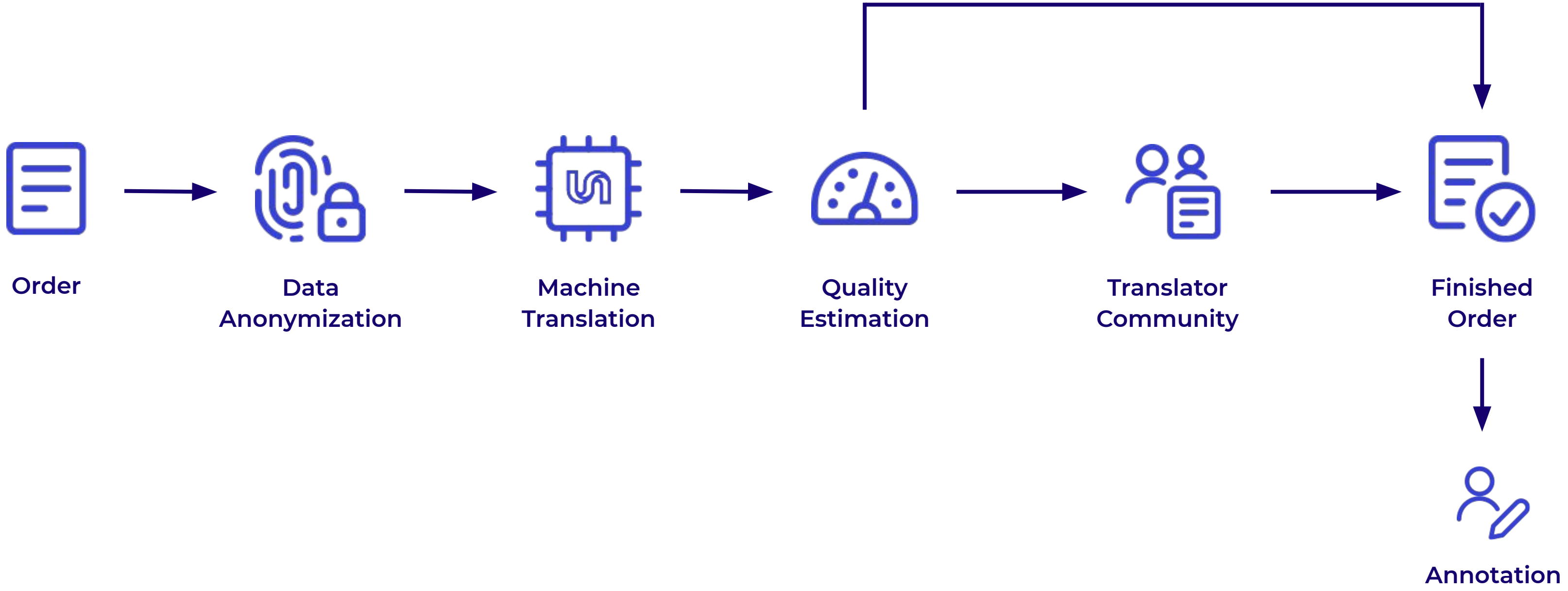
Alon Lavie

VP of Language Technologies

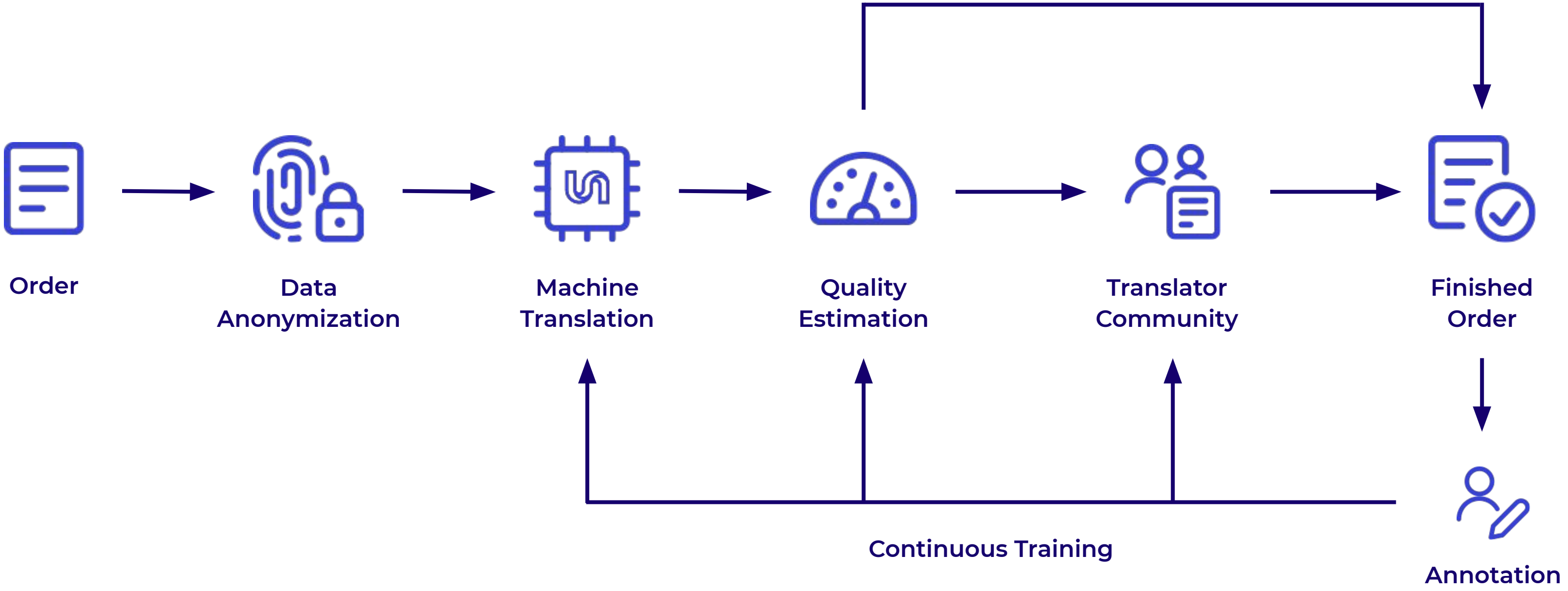
alon.lavie@unbabel.com

Why is Automatic Evaluation important at Unbabel?

Unbabel’s Translation Pipeline



Unbabel's Translation Pipeline



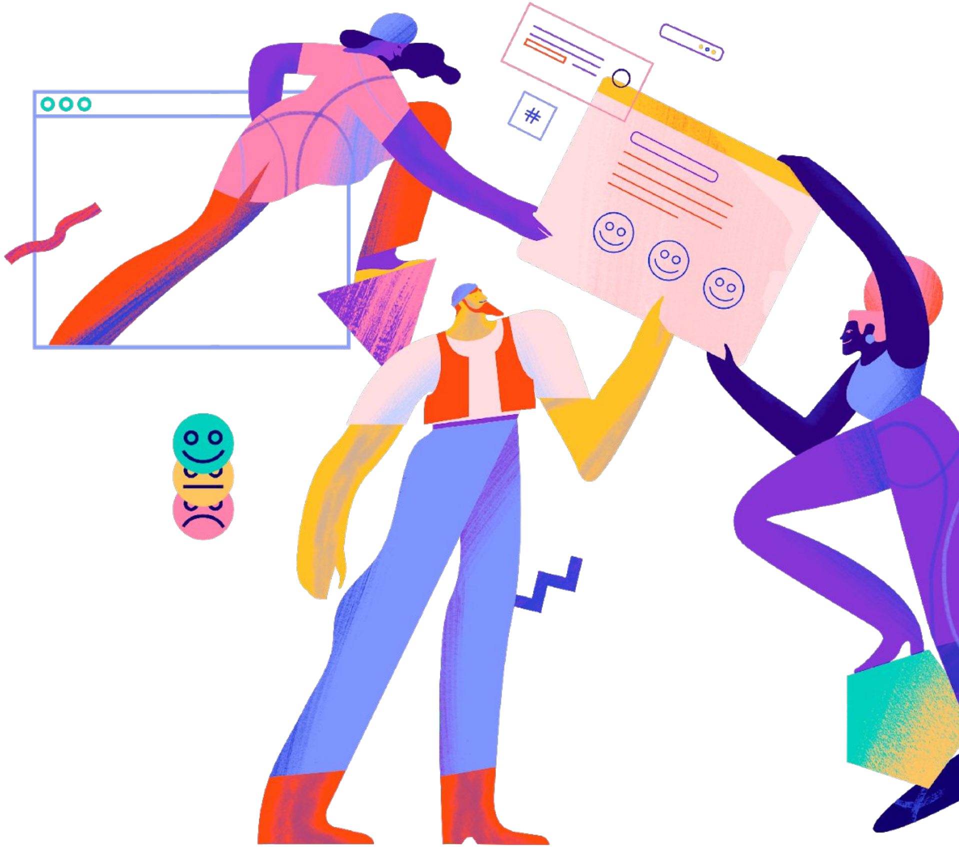
Evaluation at Unbabel

We process high volumes of translations using highly specialized models for customer service solutions in a wide range of domains.

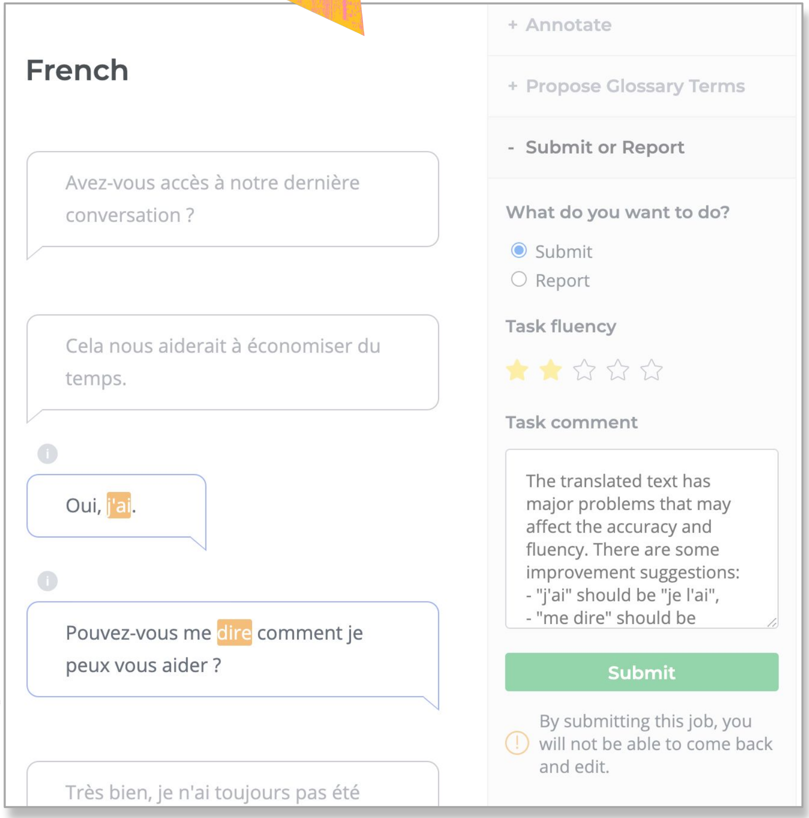
Our MT engines are continually retrained to ensure that we maintain the highest quality of translation and robustness to new content.

How do we know that MT Engine A is better than MT Engine B?

- Our engineers and scientists rely on existing metrics such as BLEU and METEOR to make initial modelling decisions
- We leverage our community of linguists to provide human evaluation using MQM



Multidimensional Quality Metrics (MQM)



French

Avez-vous accès à notre dernière conversation ?

Cela nous aiderait à économiser du temps.

Oui, j'ai.

Pouvez-vous me dire comment je peux vous aider ?

Très bien, je n'ai toujours pas été

What do you want to do?

● Submit
○ Report

Task fluency

★★☆☆☆

Task comment

The translated text has major problems that may affect the accuracy and fluency. There are some improvement suggestions:
- "j'ai" should be "je l'ai",
- "me dire" should be

By submitting this job, you will not be able to come back and edit.

Our primary method of evaluating MT quality involves sending batches of translations to our community for annotation.

We ask annotators to highlight errors according to an internal error typology (for things like ‘style’, ‘content and ‘accuracy’) and rank the error as either **minor, major** or **critical.**

We then calculate a segment-level score as a function of the **number** and **severity** of errors in the translation. Post-edition by our community of editors provides us with a ‘gold-standard’.

What's wrong with using existing metrics like BLEU?

Automatic VS Human evaluation of MT

VS.

Automatic (BLEU)

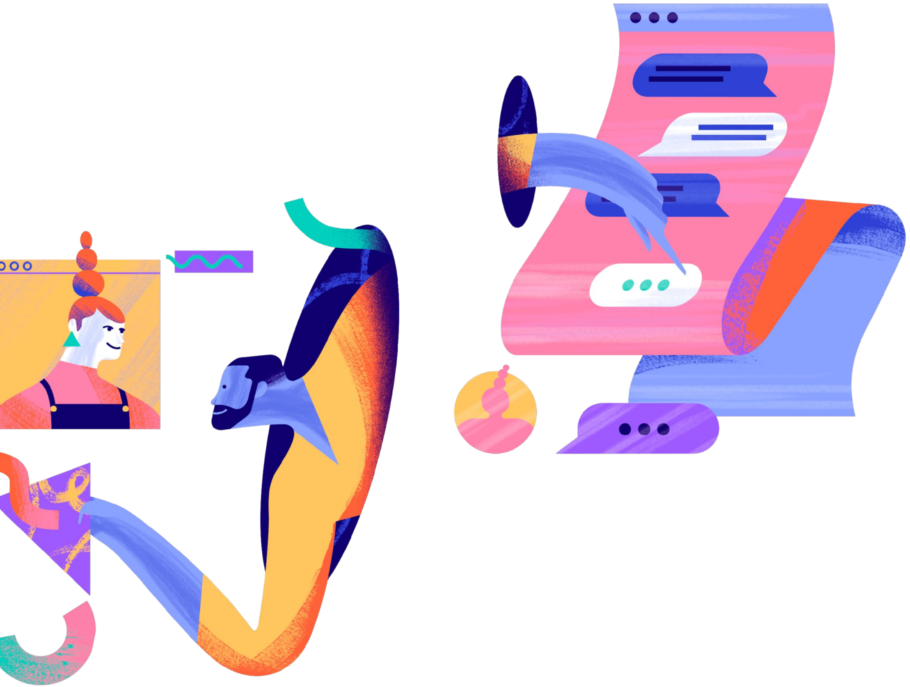
PRO: Allows our scientists and engineers to iterate quickly over MT models

CON: Less reliable and not sensitive to granular error

Human (MQM)

PRO: More reliable and sensitive to nuanced error

CON: Slow and expensive



Inability to differentiate high-performing systems

Much of the time in developing or retraining MT engines we are comparing two systems or versions of the same system that already perform very well. The gap in performance of the two iterations might be very small.

One of the key findings of the WMT 2019 Metrics Shared Task was that **even modern metrics struggle to successfully rank high-performing systems**.

Correlation with Human Judgement

In general, metrics such as BLEU and METEOR (based on n-gram matches with a reference translation) correlate poorly with human judgement.

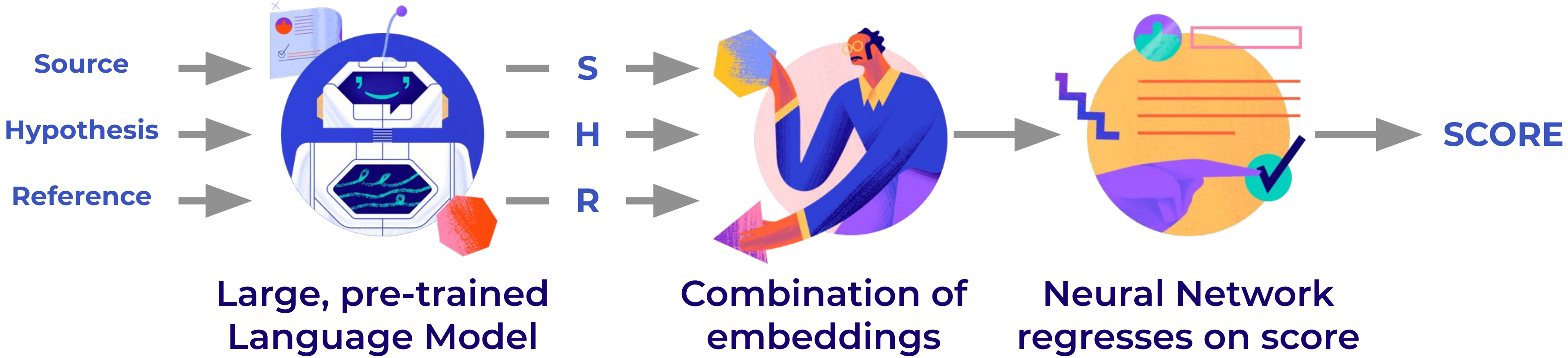
What does this mean for us and our customers?

- Modelling decisions are poorly informed and often don't align with human opinion
- Cost of verifying and rectifying modelling decisions is huge
- Degradation of performance downstream results in unhappy customers

COMET: A neural framework for MT evaluation

6 October 2020

COMET: Basic Modelling Approach



Source

Hypothesis

Reference

S

H

R

SCORE

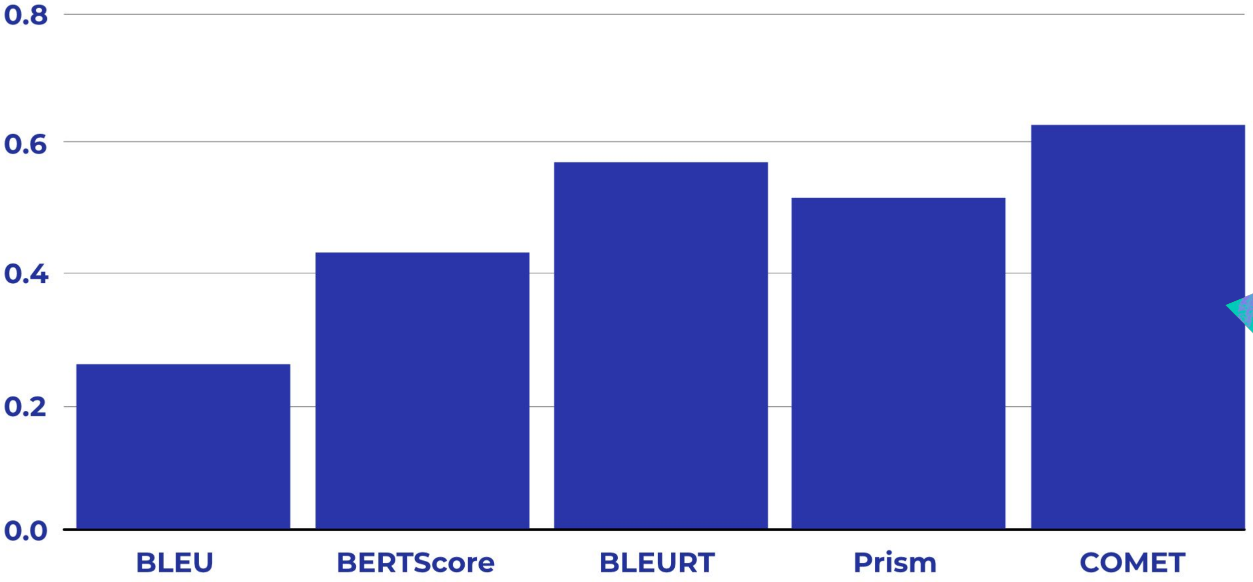
Large, pre-trained Language Model

Combination of embeddings

Neural Network regresses on score

COMET: Performance

Kendall's Tau on segment level WMT 19 Metrics Shared Task



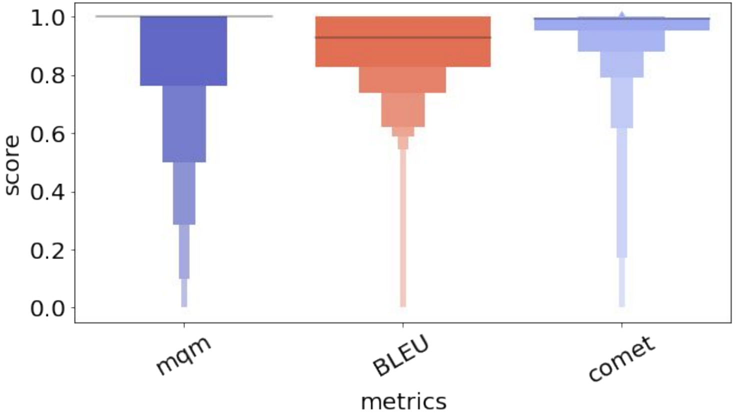
COMET: Strengths and weaknesses

EN-PT_BR

SRC: “Is there anything else I can help with?”
REF: “Existe mais alguma coisa com a qual eu possa ajudar?”
MT: “Posso ajudar com mais alguma coisa?”

MQM	100
BLEU	0.5696
COMET	0.9689

COMET can capture semantic similarities even where there is lexical disparity.



COMET has a tendency to overestimate which presents a challenge for interpretation

COMET: The Importance of Good References

EN-DE

Reference*	Adequacy	r (1-ref)	r (2-refl)
WMT	85.3	**0.523**	-
AR	86.7	0.539	**0.555**
WMTp	81.8	0.470	**0.529**
ARp	80.8	0.476	**0.537**

DE-EN

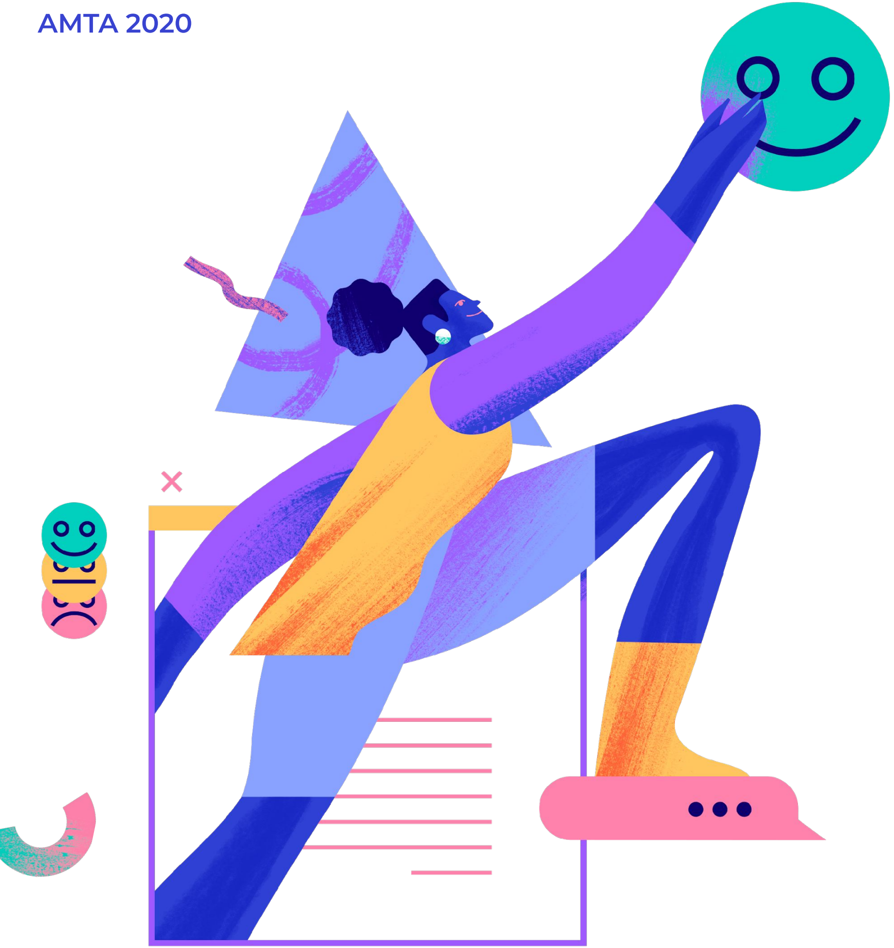
Reference	r (1-ref)	r (2-refl)
WMT	**0.42**	-
ALT	0.34	**0.40**

More references doesn’t, necessarily, mean a higher correlation.

Using more references can even hurt the correlation!

* *Data from Freitag et al (2020)* - https://arxiv.org/pdf/2004.06063.pdf

Evaluating COMET metrics for deployment



How do we know that COMET is good enough?

We started by assessing the different use cases for COMET internally and realized that these fall into two fairly distinct categories:

- **Single model evaluation** - we just want a score to tell us how well our model is doing
- **Dual model comparison** - particularly in retrainings, we have two systems (usually very close in performance) and we want to know which is better

What do we want out of COMET?

High Quality Assurance

If our ultimate goal is high quality translation, we want to ensure that our engineers have the best tools to make well-informed modelling decisions. Fundamentally we want a metric that performs better than BLEU.

Low Risk Cost Reduction

Having humans verify our engine deployment with MQM annotation is not cost effective or scaleable. We want a metric that aligns well enough with human judgement that we can make deployment decisions based on COMET alone.

Tiered Evaluation

In light of the above we defined a tiered system of evaluation whereby we calculate a Pearson's *r* correlation score on internal test sets to assess how closely the metric aligns with MQM. We start by figuring out what we think is an **acceptable risk margin** which we set at **+/-0.1 Pearson**

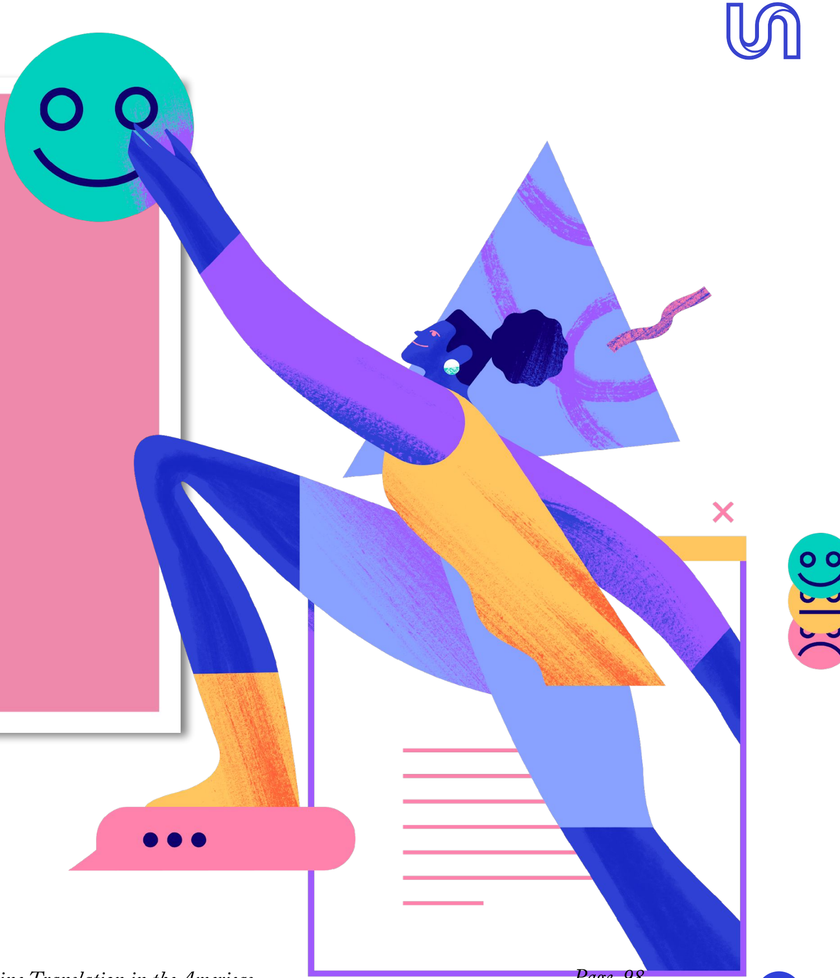
TIER 1 (near enough to human parity)

- Internal analysis revealed that human annotators correlate with each other at around 0.6-0.7 Pearson
- **Does COMET achieve a Pearson of >0.5** (i.e. is it within our risk margin of human agreement)?

TIER 2 (better than BLEU)

- **Does COMET perform better than BLEU** at a level exceeding our risk margin?

en-zh	en-ja	en-id	en-bg
en-fi	en-ko		en-hu
en-ro	en-cs		
en-es	en-es		
en-nl	en-vi		
en-ru	en-da		
en-pl	en-fr		
en-no	en-tr		
en-de	en-pt		
en-sv	en-it		
en-th			



Tiered LPs for out of English Ticket Products

6 October 2020

No Language left behind

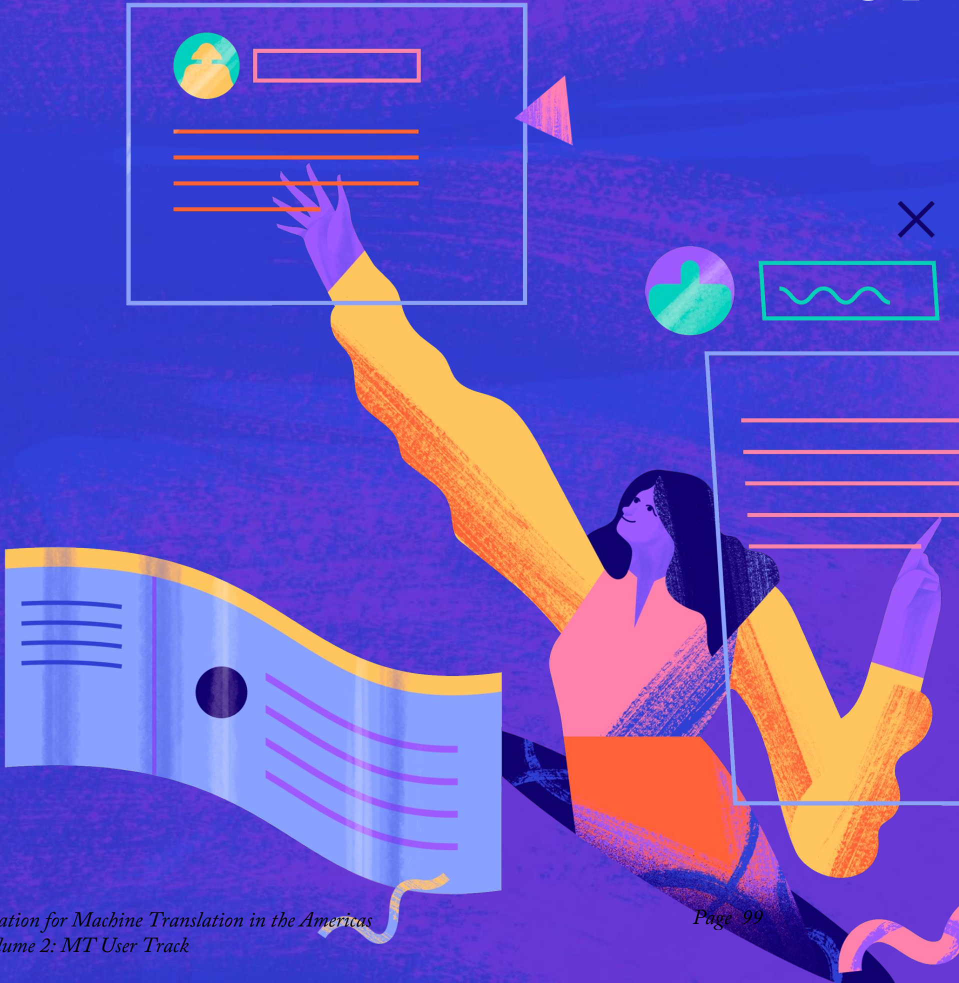
The ideal scenario for COMET is that it puts us in a postition where all of our products can rely on COMET scores without the need for human annotation (i.e. that all LPs land in Tier 1).

For LPs in Tier 2:

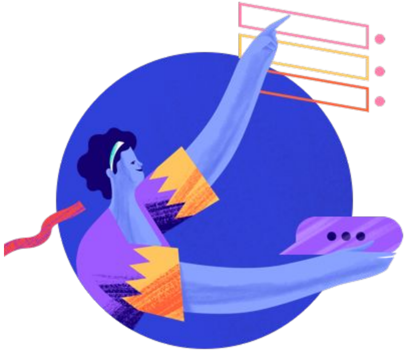
- We are actively seeking opportunities to improve COMET performance on these LPs. This involves both general model improvement and augmentation of our datasets.

For other LPs:

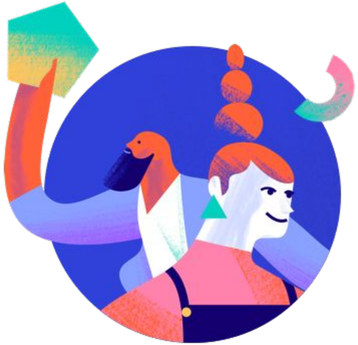
- Where we don't have data for existing LPs we rely on our editors to generate more data for testing and training.



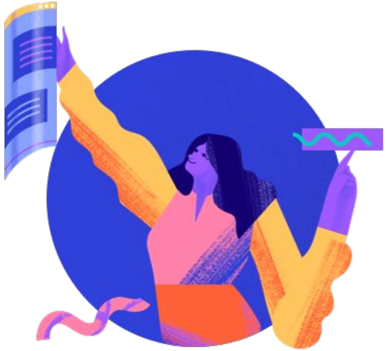
Evaluation Process



Identify products and language pairs across the business and collect data sufficient to give a reasonably reliable Pearson’s r score



Evaluate iterations of COMET across settings and compare results with human assessments



Based on our tiered evaluation scheme, assess reliability of COMET in each use case and iterate until we are satisfied of the impact of the model

Deploy the best model and provide clear information to our engineers about how and where to use COMET

COMET in deployment

To provide an extra layer of certainty and trust in COMET for our engineers, we are implementing **statistical significance testing** in our retrainings evaluation.

In deciding whether to deploy a retrained system we apply a bootstrapped t-test for significance to determine, with a 95% confidence interval, that the new system is better than the old.

We also complement our COMET evaluation with a range of other metrics to ensure that our engineers have a full toolkit when making modelling decisions.

What to do when metrics disagree?

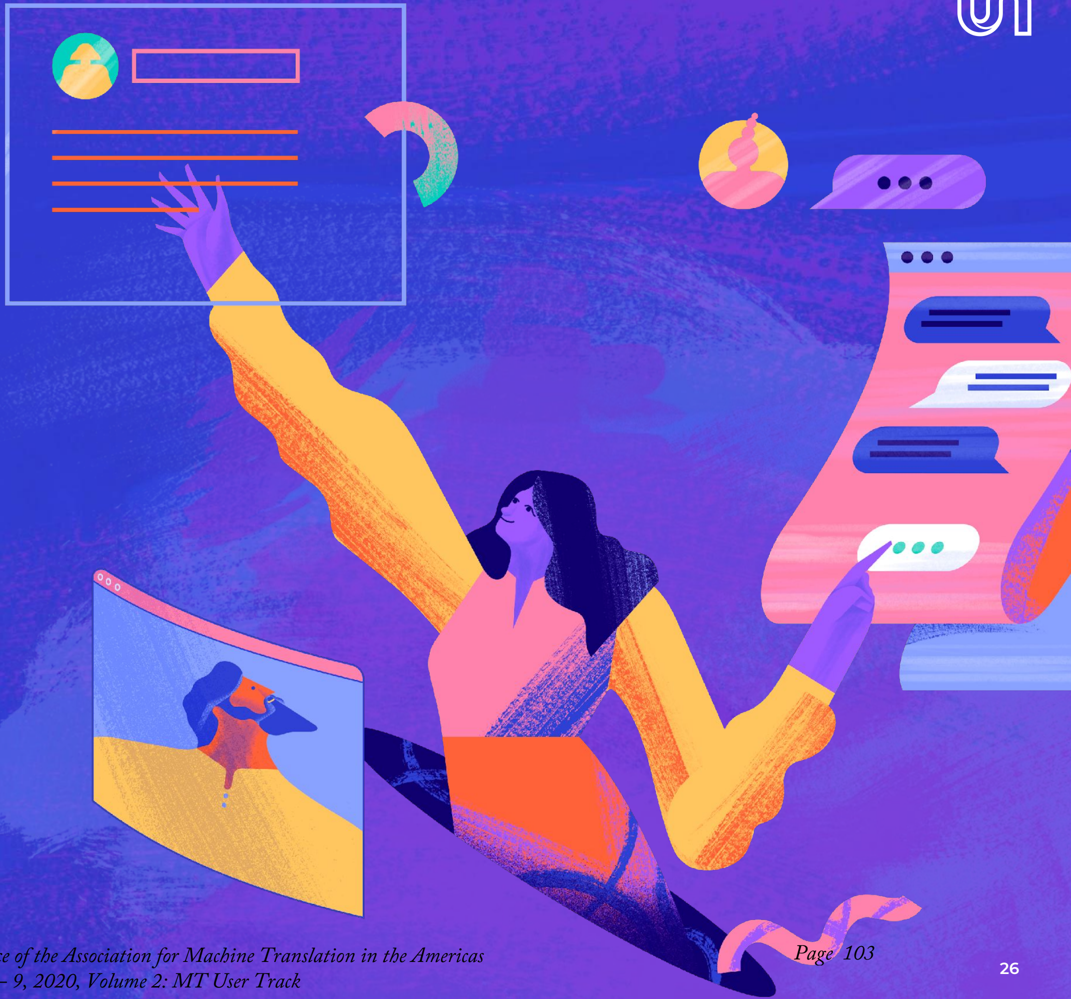
It is important to note that even where metrics like BLEU don't correlate well with human judgement, their input is still valuable, if only because metrics based on lexical similarity tell us something unique from metrics such as COMET which are more grounded in semantics.

As such we encourage our engineers to look at a variety of metrics including BLEU, METEOR, TER, BERTScore and COMET to get a fuller picture of what our models are doing.

Where all metrics agree the decision to deploy is black and white. Where it isn't:

- **COMET and other semantic metrics (e.g. BERTScore) agreeing? Good chance that MT is semantically accurate**
- **COMET disagrees with everyone? Check the magnitude of the difference before discarding and consider the statistical significance of the improvement**

Keeping tabs on COMET over time



How do we continue to adapt COMET?

As the range of products and languages at Unbabel grows, we need to ensure that COMET is keeping up.

With COMET in production, we are developing a procedure to re-evaluate COMET on a rolling basis by sampling retrainings for annotation with MQM.

We are also coordinating with product managers to anticipate future product and language demand and perform evaluations and adaptation on new data.

Outside of Unbabel

We we plan to release an open source version of the COMET framework to benefit the wider MT community, and we are hopeful that development will continue over the next year.

The code will be available at:

https://github.com/Unbabel/COMET

Key takeaways

Metrics in a commercial setting:

- Automatic metrics like BLEU are of limited use
- Adaptive evaluation frameworks trained to correlate well provide an attractive solution
- Our COMET framework is publicly available

Evaluating Metrics:

- Metrics can have different use cases and applications
- A tiered evaluation method can help to align expectations
- Considering the statistical significance of modelling decisions can be insightful

Questions?

Craig Stewart

Research Scientist

Unbabel

Thank you

Craig Stewart

Research Scientist, Unbabel

craig.stewart@unbabel.com

Unbabel

Unbabel