Beyond Single Scores:

Transparent Evaluation through Fine-Grained Error Detection

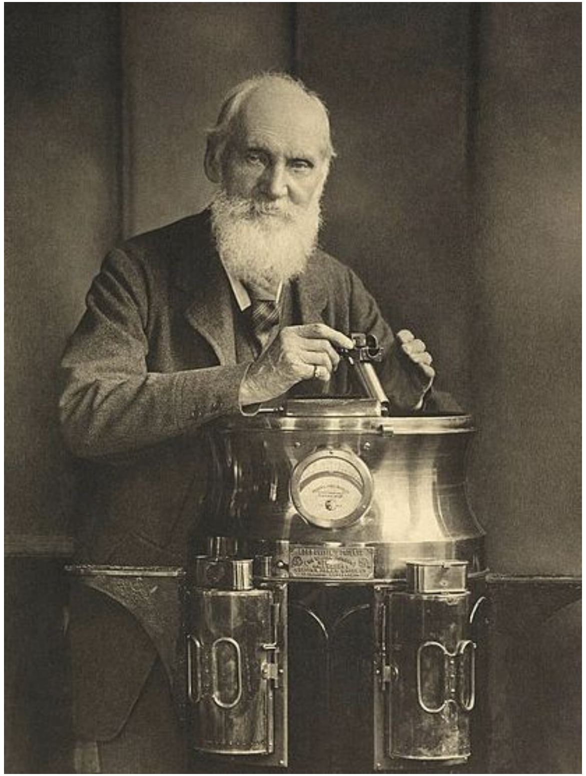
André Martins

NAACL SemEval 2024, June 21, 2024

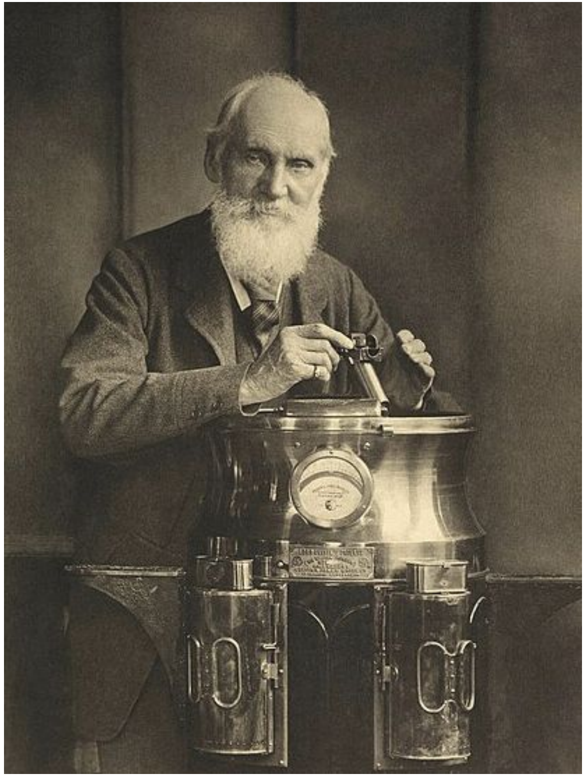
Two Amazing Teams: Unbabel + SARDINE Lab

SARDINE: Structure **A**wa**R**e mo**D**el**I**ng for **N**atural languag**E**

No science without measuring



No science without measuring



“When you can measure what you are speaking about and express it in numbers you know something about it; but ***when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind****: it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of science.”*

— Lord Kelvin, 1883

Evaluation shapes and guides research

We use it to:

- compare experiments,
- understand if one method / model is better than another,
- identify weaknesses and determine what to work on,
- decide which model we want to deploy / use.

Evaluation shapes and guides research

We use it to:

- compare experiments,
- understand if one method / model is better than another,
- identify weaknesses and determine what to work on,
- decide which model we want to deploy / use.

But is a "number" (a single score) enough to make progress? 😕

This talk: Evaluation in Machine Translation (MT)

- MT is a good example where evaluation research is quite advanced
  - WMT shared tasks
  - Lots of human annotated data, publically available
  - Very active meta-evaluation research.

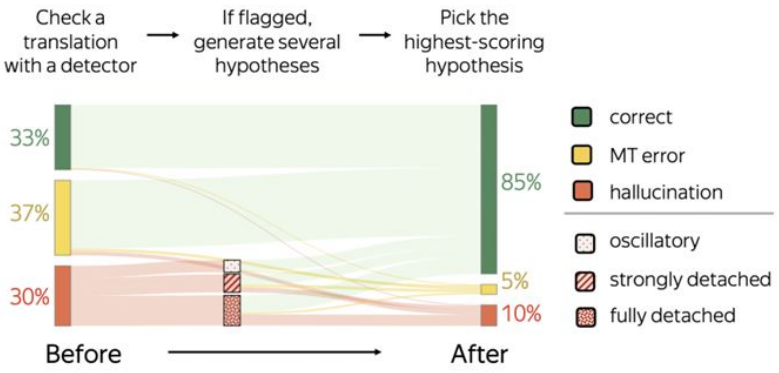
- Everything in this talk can equally be applied to other NLP tasks.

This talk

- Two recent open-source projects led by our team:
  - **xCOMET**: Fine-Grained Automatic MT Evaluation
  - **Tower**: A Multilingual LLM for Translation-Related Tasks

MT Hallucinations

Category	Source Sentence	Reference Translation	Hallucination
Oscillatory	Ist ein Kompromiss aufgrund des zugrundeliegenden Regelsystems unmöglich, so spricht man von Aporie.	The case where, based on the pertinent system of regulations a compromise is not possible, is referred to as Aporia.	Aporia is the name of aporia , which is the name of aporia.
Strongly Detached	Tickets für Busse und die U-Bahn ist zu teuer, vor allem in Stockholm.	Tickets for buses and the subway is too expensive, especially in Stockholm.	The hotel is located in the centre of Stockholm, close to the train station.
Fully Detached	Die Zimmer beziehen, die Fenster mit Aussicht öffnen, tief durchatmen, staunen.	Head up to the rooms, open up the windows and savour the view, breathe deeply, marvel.	The staff were very friendly and helpful.



"[Looking for a Needle in a Haystack: A Comprehensive Study of Hallucinations in NMT]". N. Guerreiro, E. Voita, A. Martins. EACL 2013.

"[Optimal Transport for Unsupervised Hallucination Detection in NMT]". N. Guerreiro, P. Colombo, P. Piantanida, A. Martins. ACL 2013.

"[Hallucinations in Large Multilingual Translation Models]". N. Guerreiro, D. Alves, J. Waldendorf, B. Haddow, A. Birch, P. Colombo, A. Martins. TACL 2013.



Evaluation in Machine Translation

Two Choices

Automatic (e.g. BLEU)

Fast, scalable, often unreliable

VS.

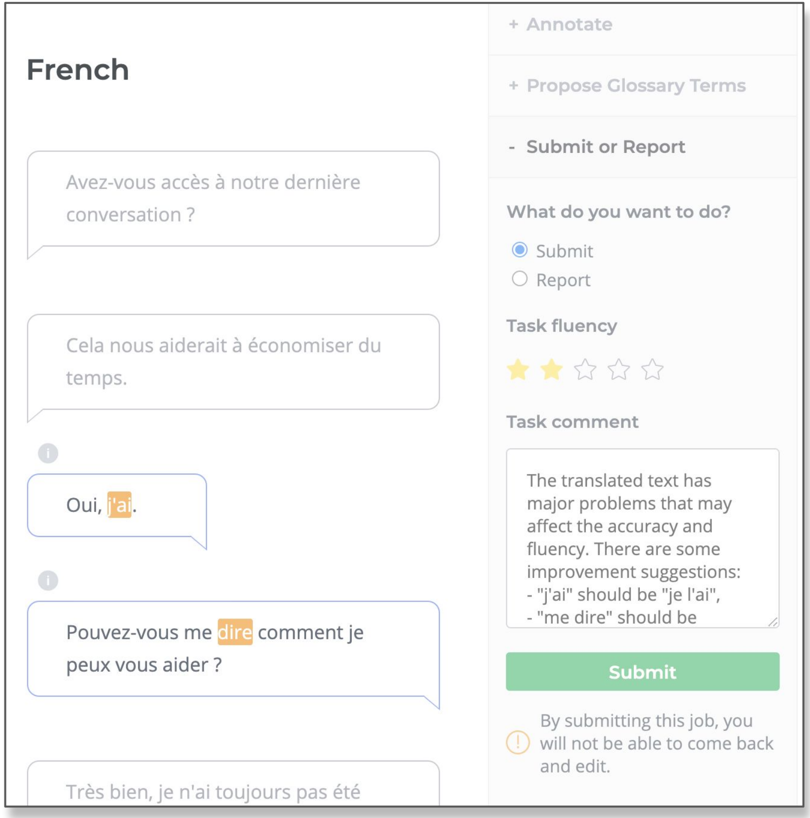
Human (e.g. MQM)

Slow, expensive, more reliable

Human Evaluation

Some examples:

- Ranking – compare translations *relative* to each other
- Direct assessments – assign an *absolute* score
- **Multidimensional quality metrics (MQM)**



Multidimensional Quality Metrics (MQM)

- Ask annotators to highlight errors according to an internal error typology (for things like 'style', 'fluency' and 'accuracy') and rank the error as either **minor, major** or **critical.**

- Calculate a document-level score as a function of the **number** and **severity** of errors in the translation.

$$\text{MQM score} = 100 - \frac{I_{\text{Minor}} + 5 \times I_{\text{Major}} + 10 \times I_{\text{Crit.}}}{\text{Sentence Length} \times 100}$$

(http://www.qt21.eu/mqm-definition/definition-2015-12-30.html)

CUA: Customer Utility Analysis

Excellent

The translation is practically fluent! There are almost no mistakes, and the occasional flaw does not affect the meaning and communication.

Good

Almost there! There are a few grammatical issues or inaccuracies in meaning, but the translation is generally understandable.

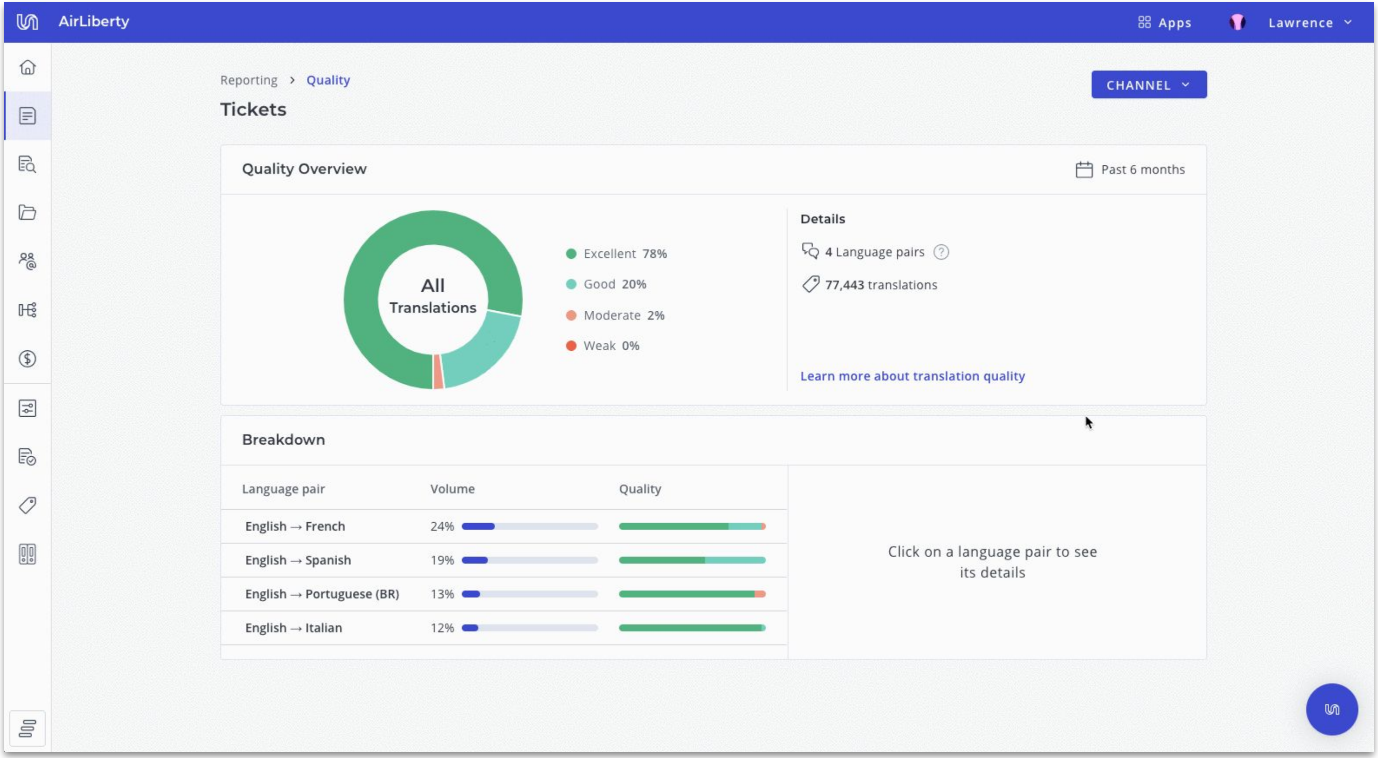
Moderate

The translation has quite a few errors. The message and communication may only be partially understandable.

Weak

The translation has errors that critically impact the overall communication and meaning. The message may not be understandable at all.

CUA: Customer Utility Analysis



Requirements for Automatic Metrics

1. Strong correlation with human judgments,
2. Applicable to a wide range of languages, domains, and scenarios,
3. Interpretable, and
4. Fast and lightweight.

Does BLEU Satisfy Our Requirements?

Re-evaluating the Role of BLEU in Machine Translation Research

Chris Callison-Burch   Miles Osborne   Philipp Koehn
School on Informatics
University of Edinburgh
2 Buccleuch Place
Edinburgh, EH8 9LW
callison-burch@ed.ac.uk

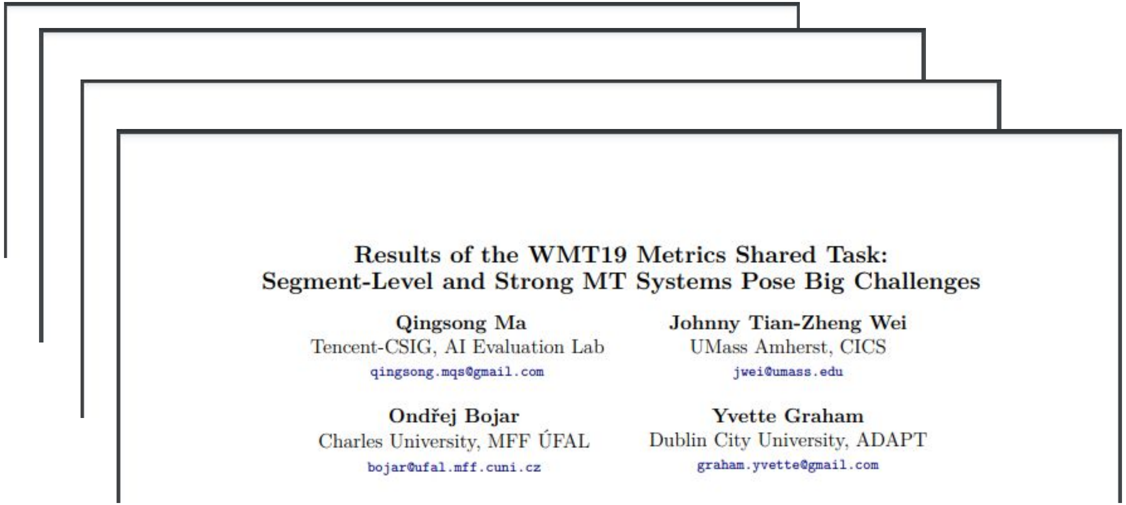
Does BLEU Satisfy Our Requirements?

Comparing Automatic and Human Evaluation of NLG Systems

Anja Belz
Natural Language Technology Group
CMIS, University of Brighton
UK
A.S.Belz@brighton.ac.uk

Ehud Reiter
Dept of Computing Science
University of Aberdeen
UK
ereiter@csd.abdn.ac.uk

Does BLEU Satisfy Our Requirements?



Results of the WMT19 Metrics Shared Task:
Segment-Level and Strong MT Systems Pose Big Challenges

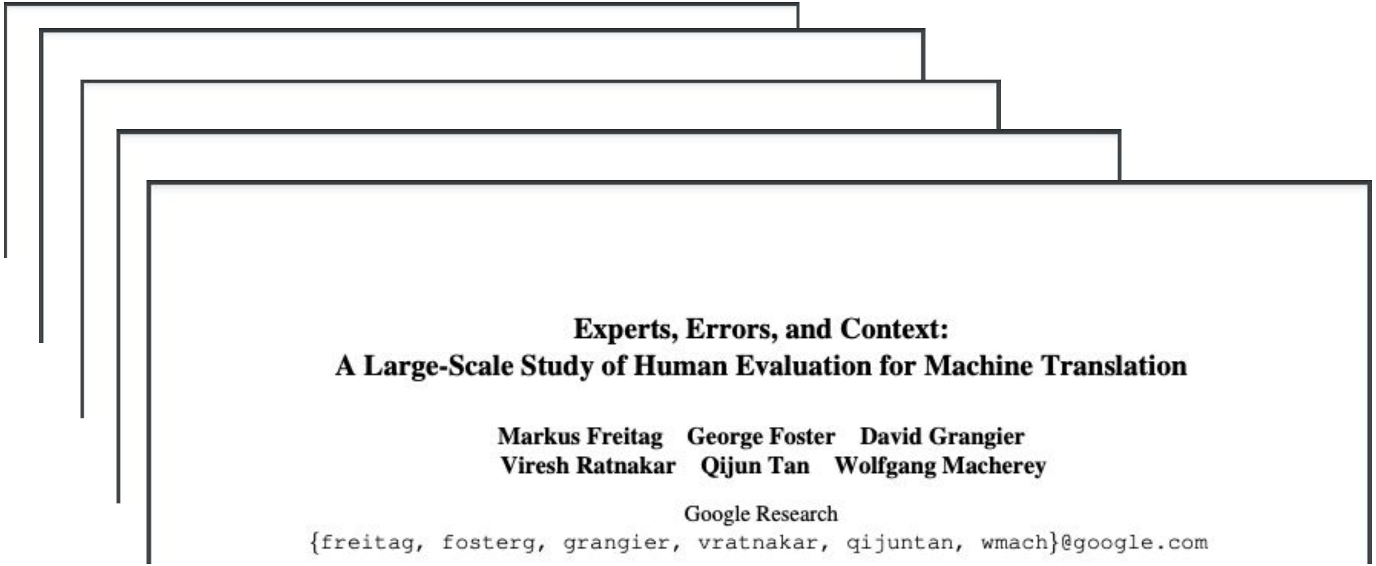
Qingsong Ma
Tencent-CSIG, AI Evaluation Lab
qingsong.mqs@gmail.com

Johnny Tian-Zheng Wei
UMass Amherst, CICS
jwei@umass.edu

Ondřej Bojar
Charles University, MFF ÚFAL
bojar@ufal.mff.cuni.cz

Yvette Graham
Dublin City University, ADAPT
graham.yvette@gmail.com

Does BLEU Satisfy Our Requirements?



Experts, Errors, and Context:
A Large-Scale Study of Human Evaluation for Machine Translation

Markus Freitag George Foster David Grangier
Viresh Ratnakar Qijun Tan Wolfgang Macherey

Google Research
{freitag, fosterg, grangier, vratnakar, qijuntan, wmach}@google.com

.... and many more works show many flaws of BLEU!

12 Critical Flaws of BLEU

Does BLEU Satisfy Our Requirements?

	BLEU
Strong correlation with human judgments	❌
Applicable to a wide range of languages and domains	❓
Interpretable	❓
Fast and lightweight	✅

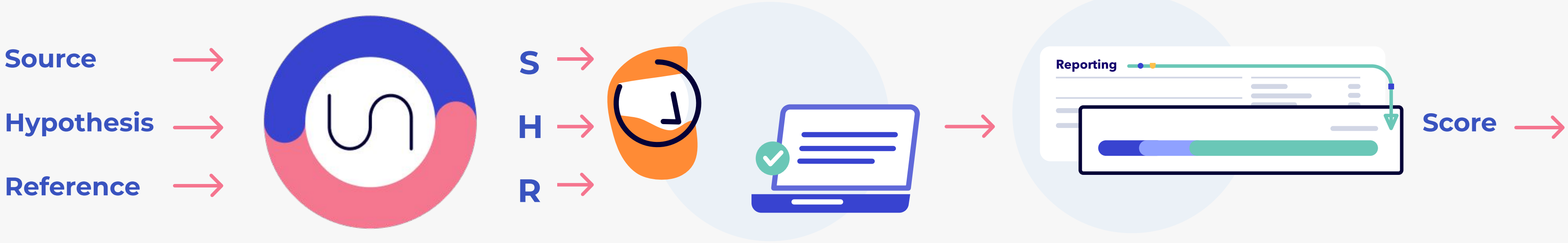
Does BLEU Satisfy Our Requirements?

	BLEU
Strong correlation with human judgments	❌
Applicable to a wide range of languages and domains	❓
Interpretable	❓
Fast and lightweight	✅

Not really :(We need better automatic evaluation!

Can we *learn* an automatic metric to predict a quality score?

COMET (Cross-lingual Optimized Metric for Evaluation of Translation)



Large, pre-trained Language Model

Combination of embeddings

Neural Network regresses on score

"COMET: A Neural Framework for MT Evaluation".
Ricardo Rei, Craig Stewart, Ana C Farinha, Alon Lavie. EMNLP 2020.

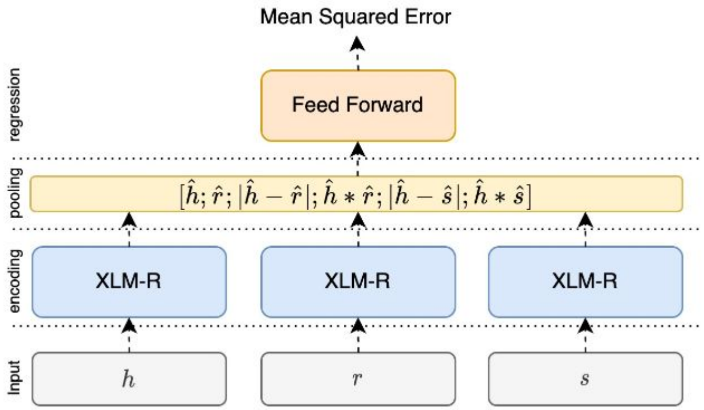
COMET (Cross-lingual Optimized Metric for Evaluation of Translation)

Idea:

Train a neural network to perform evaluation!

How? Taking advantage of human evaluation:

1) Human-mediated Translation Edit Rate (HTER)
2) Multidimensional Quality Metrics (MQM)
3) Direct Assessments (DA)

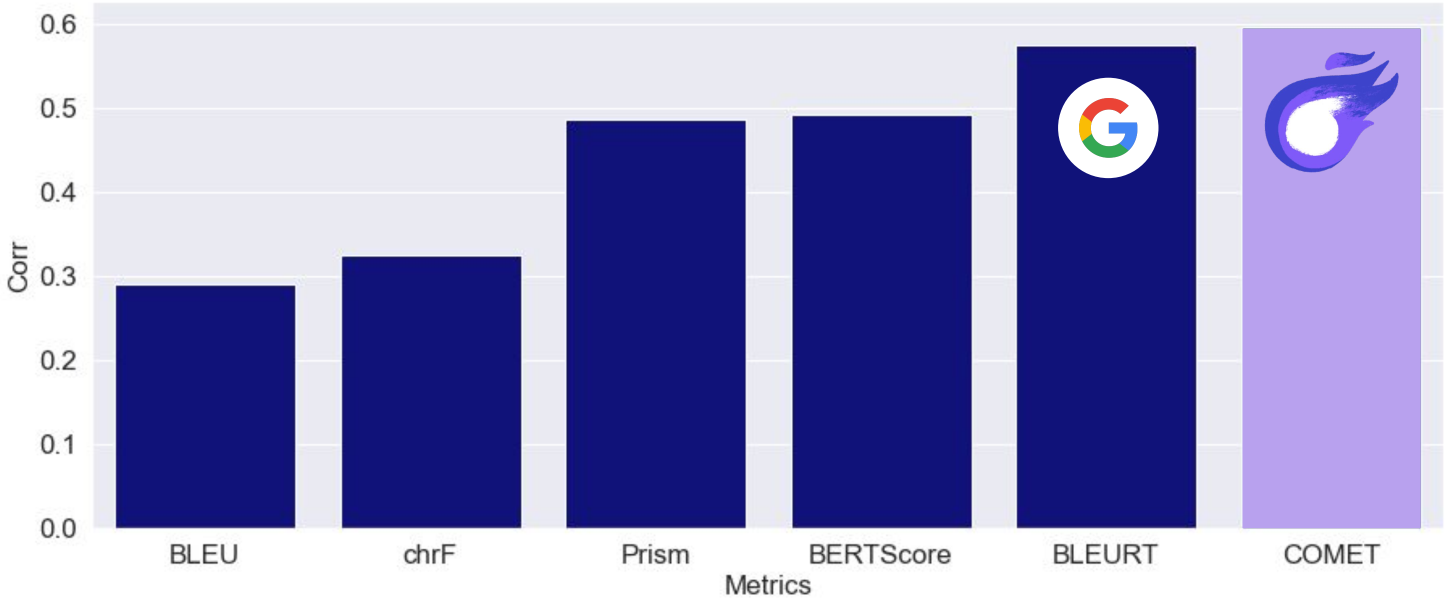


Since **human evaluation is primarily source-based**, there is value in including the source!

"[COMET: A Neural Framework for MT Evaluation]".
Ricardo Rei, Craig Stewart, Ana C Farinha, Alon Lavie. EMNLP 2020.

COMET: Performance

Spearman on segment level with MQM annotations for WMT21 (development data)



Can we estimate MT quality *without* references?

Motivation:

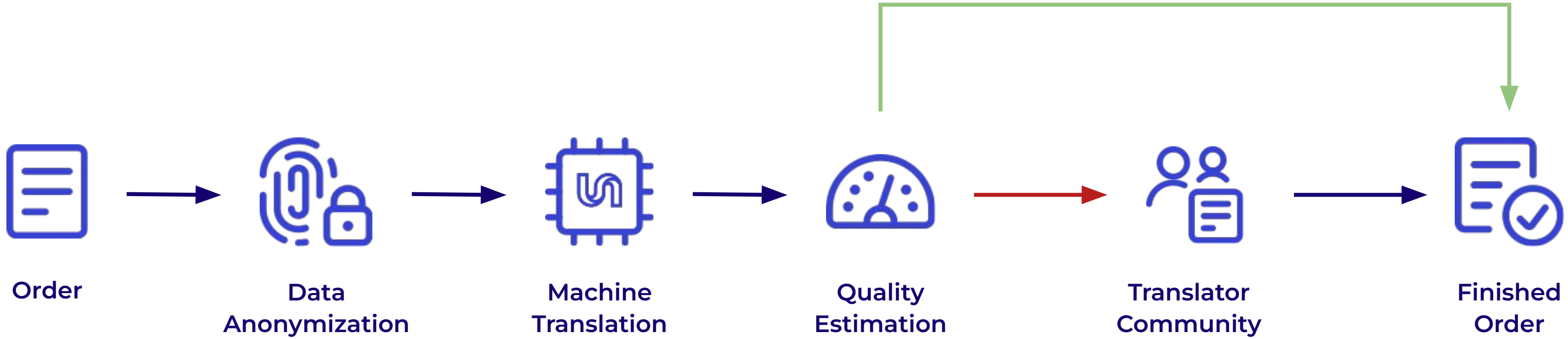
What can we do if we knew the **quality of a translation on-the-fly?**

1) If it is good we can trust it and use it.

2) If it is not good we need to improve it (e.g. asking a human to post edit)

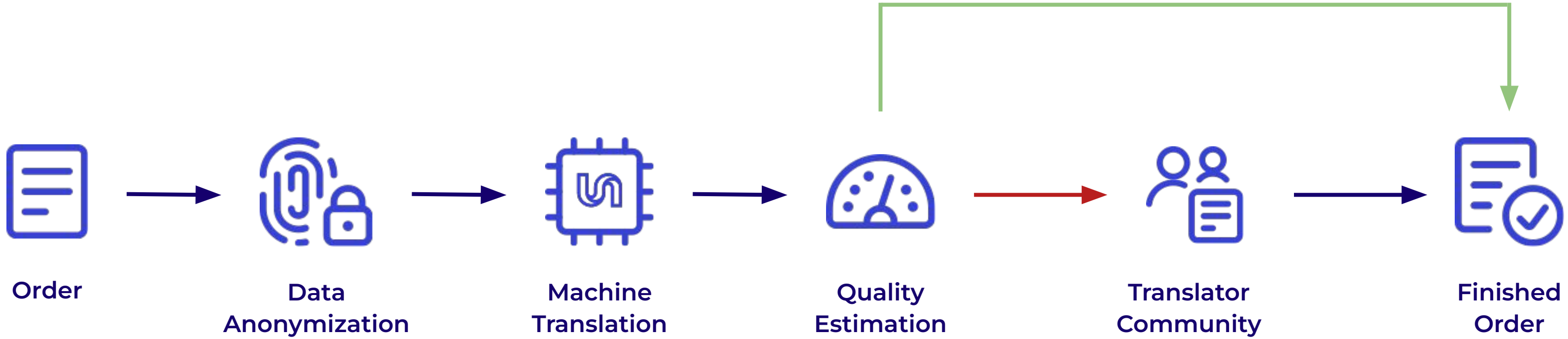
Motivation:

What can we do if we knew the **quality of a translation on-the-fly?**



Motivation:

What can we do if we knew the **quality of a translation on-the-fly?**



Quality estimation ensures that the delivered quality is higher (better MQM) and reduces post-edit costs!

Quality Estimation vs Automatic Metrics

OpenKiwi
By Unbabel

- Estimates translation quality (without seeing a reference)
- Is this translation OK to send out? (QE skips)
- Learns from what annotators highlight (MQM annotations)
- Does not provide a direct estimation of MQM but rather tries to identify major/critical translation problems

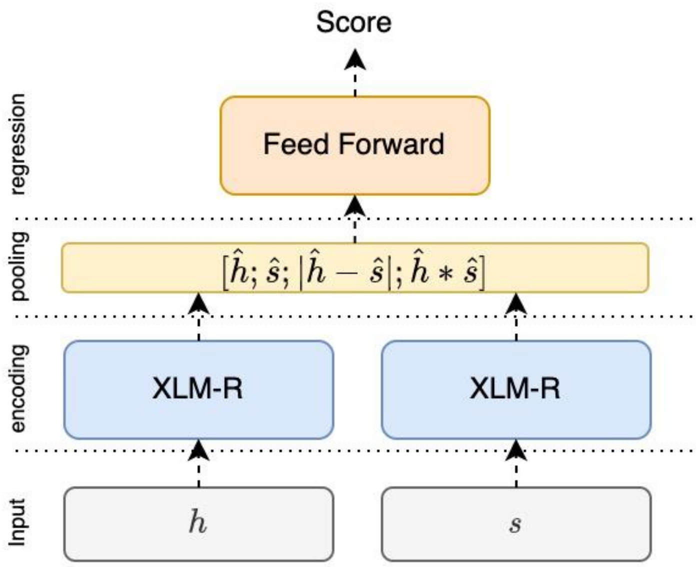
COMET

- Measures MT Model quality (with the aid of a reference)
- Is this MT model OK to deploy? (MT retrainings)
- Learns from what annotators highlight (MQM annotations)
- Provides a direct estimation of MQM but the data requires more precious human effort

COMET-QE Dual Encoder

COMET was first developed for **reference-based MT evaluation** but it has been extended for **QE** as well!

- Sentence embeddings are created through average pooling
- Along with source and target embeddings we extract the element-wise difference and product between embeddings
- A feed forward is used to predict a quality assessment (MQM or DA).



QE is competitive with reference-based metrics!

Results of the WMT20 Metrics Shared Task

Nitika Mathur
The University of Melbourne
nmathur@student.unimelb.edu.au

Johnny Tian-Zheng Wei
University of Southern California,
jwei@umass.edu

Markus Freitag
Google Research
freitag@google.com

Qingsong Ma
Tencent-CSIG,
AI Evaluation Lab
qingsong.mqs@gmail.com

Ondřej Bojar
Charles University,
MFF ÚFAL
bojar@ufal.mff.cuni.cz

To summarize, we see that the current MT metrics generally struggle to score human translations against machine translations reliably. Rare exceptions include primarily trained neural metrics and reference-less COMET-QE. While the metrics are not really prepared to score human translations, we find this type of test relevant as more and more language pairs are getting closer to the human translation benchmark. A general-enough metric should be thus able to score human translation comparably and not rely on some idiosyncratic properties of MT outputs. We hope that human translations will be included in WMT DA scoring in the upcoming years, too.

To Ship or Not to Ship: An Extensive Evaluation of Automatic Metrics for Machine Translation

Tom Kocmi **Christian Federmann** **Roman Grundkiewicz** **Marcin Junczys-Dowmunt** **Hitokazu Matsushita** **Arul Menezes**
Microsoft
1 Microsoft Way
Redmond, WA 98052, USA
{tomkocmi,chrife,rogrundk,marcinjd,himatsus,arulm}@microsoft.com

	All	0.05	0.01	0.001	Within
n	3344	1717	1420	1176	541
COMET	**83.4**	**96.5**	**98.7**	**99.2**	**90.6**
COMET-src	83.2	95.3	97.4	98.1	89.1
Prism	80.6	94.5	97.0	98.3	86.3
BLEURT	80.0	93.8	95.6	98.2	84.1
ESIM	78.7	92.9	95.6	97.5	82.8
BERTScore	78.3	92.2	95.2	97.4	81.0
ChrF	75.6	89.5	93.5	96.2	75.0
TER	75.6	89.2	93.0	96.2	73.9
CharacTER	74.9	88.6	91.9	95.2	74.1
BLEU	74.6	88.2	91.7	94.6	74.3
Prism-src	73.4	85.3	87.6	88.9	77.4
EED	68.8	79.4	82.4	84.6	68.2

WMT21 Metric task Results

Metric	Total "wins"	Language Pair			Granularity		Data condition		
		en→de	en→ru	zh→en	sys	seg	news w/o HT	news w/ HT	TED
C-SPECpn	11	4	3	4	6	5	3	5	3
bleurt-20	10	4	5	1	4	6	4	3	3
COMET-MQM_2021	10	3	3	4	3	7	3	2	5
tgt-regEMT	4	1	1	2	3	1	2	1	1
COMET-QE-MQM_2021	3	1	1	1	3			3	
OpenKiwi-MQM	3	2		1	3		1	2	
RoBLEURT*	3			3	1	2	1		2
cushLEPOR(LM)	2	1		1	2		1		1
BERTScore	2	1	1		2		1		1
Prism	2		2		2		1		1
YiSi-1	2		2		2		1		1
MEE2	2	2			2		1		1
BLEU	1	1			1		1		
hLEPOR	1		1		1				1
MTEQA*	1			1	1				1
TER	1			1	1				1
chrF	1			1	1				1

Results of the WMT21 Metrics Shared Task: Evaluating Metrics with Expert-based Human Evaluations on TED and News Domain (Freitag et al., WMT 2021)

Does COMET Satisfy Our Requirements?

	BLEU
Strong correlation with human judgments	❌
Applicable to a wide range of languages and domains	❓
Interpretable	❓
Fast and lightweight	✅

Does COMET Satisfy Our Requirements?

	BLEU	COMET
Strong correlation with human judgments	❌	✅
Applicable to a wide range of languages and domains	❓	✅
Interpretable	❓	❌
Fast and lightweight	✅	❌

Does COMET Satisfy Our Requirements?

	BLEU	COMET
Strong correlation with human judgments	❌	✅
Applicable to a wide range of languages and domains	❓	✅
Interpretable	❓	❌
Fast and lightweight	✅	❌

How can we make COMET more interpretable?

How can we make COMET more interpretable?

We need to go beyond a single score!

Examples (next):

- MT Telescope
- Explainable QE
- COMET with uncertainty quantification
- AutoMQM
- xCOMET

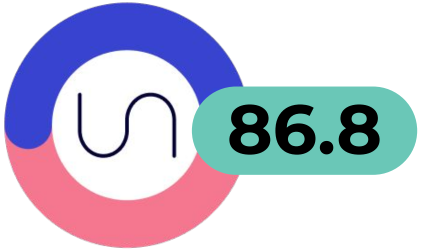
MT Telescope

An open-source tool which enables **fine-grained comparative analysis of MT system performance**.

Translation quality is extremely difficult to pin down. Standard practice uses tools to assign a quality score to translations. This score usually determines which translation systems we use:

'Scores' don't tell us the full story:

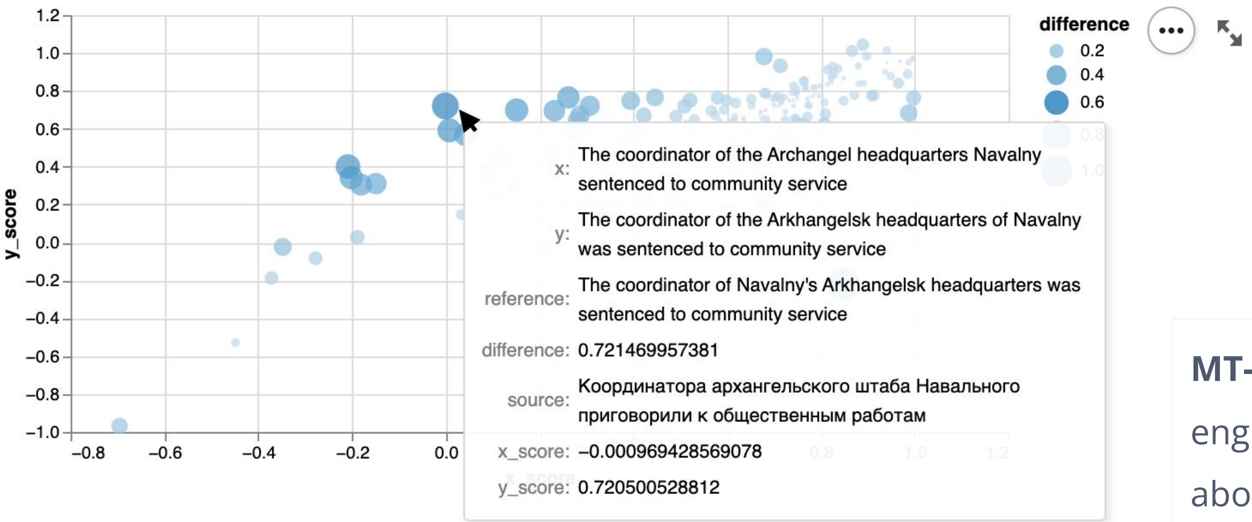
A system with a higher score is 'better' **but what is it better at?** Translating customer names? Greetings?...

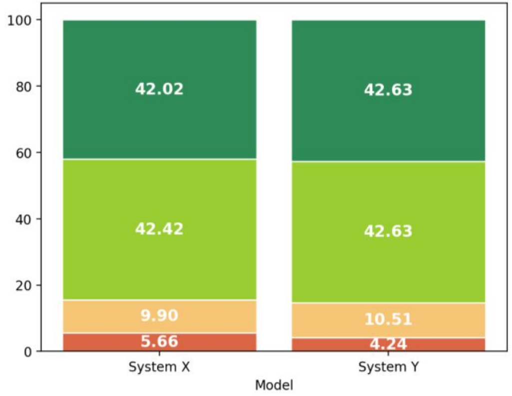


MT Telescope

MT-Telescope allows MT engineers to fully understand the capabilities of a translation system.

It is an **easy to use, web-based, interactive interface** that exposes how different models translate.





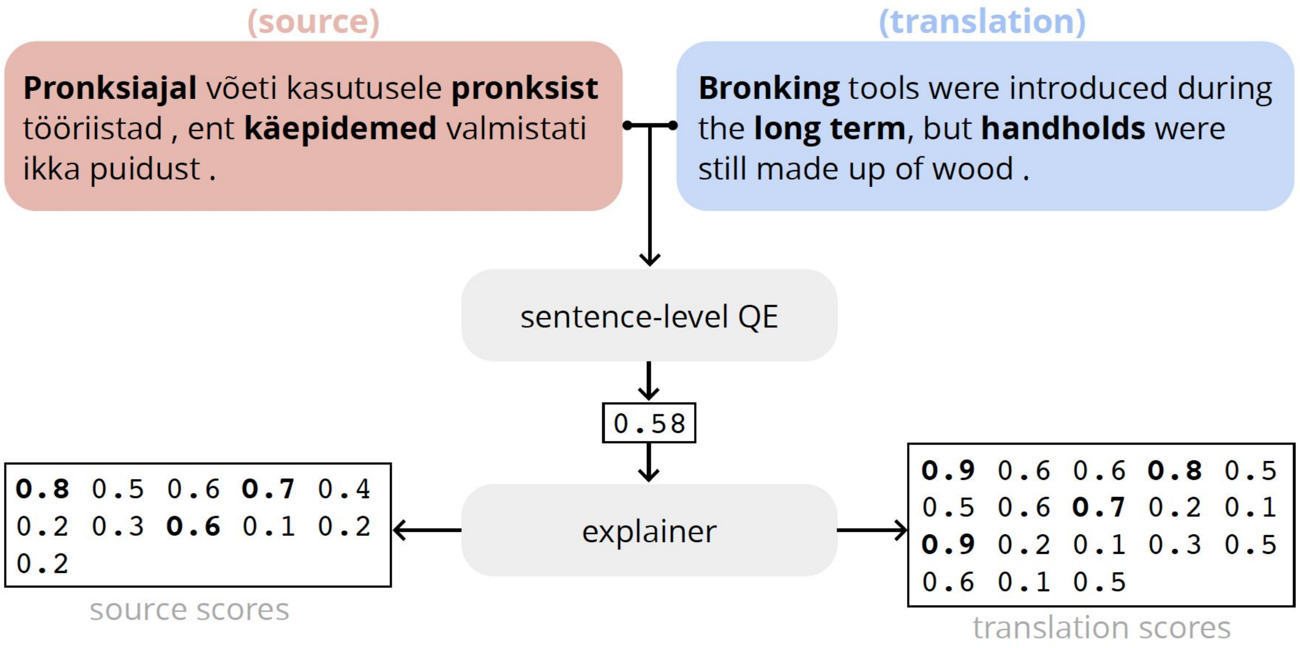
MT-Telescope tools empowers engineers to make better decisions about translation quality.

Can we “explain” low scores with attribution methods?

WMT 2022 QE Task: Unbabel-IST Submission

Explainable QE shared task objective:

Identify translation errors via explainability methods (without any word-level supervision)



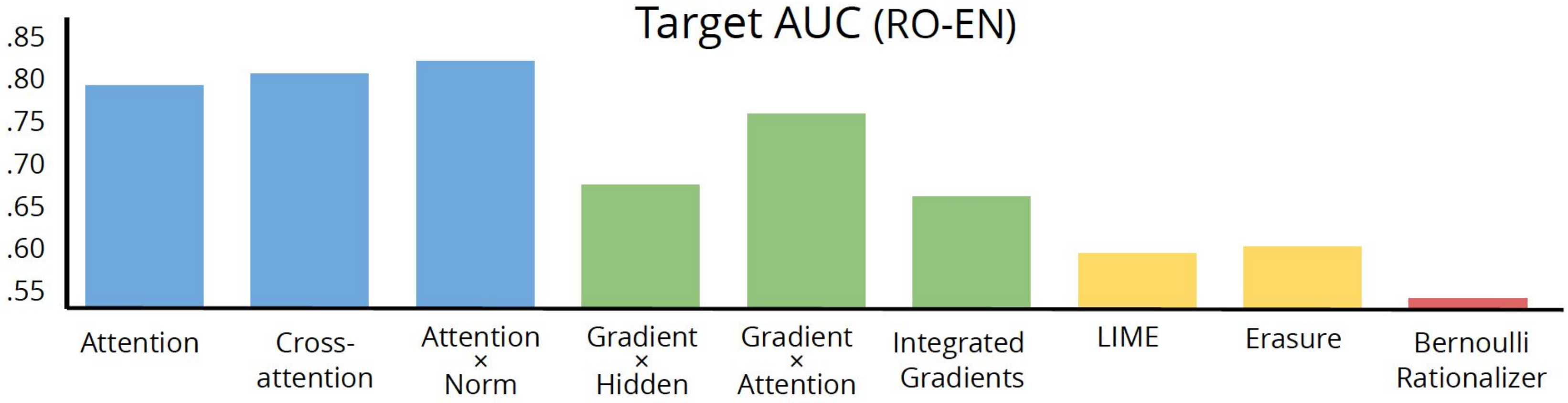
WMT 2022 QE Task: Unbabel-IST Submission

• **Attention-based**	attention weights cross-attention weights attention weights × L2 norm of value vectors [1]
• **Gradient-based**	gradient × hidden state vector gradient × attention output integrated gradients [2]
• **Perturbation-based**	LIME [3] erasure
• **Rationalizers**	Relaxed-Bernoulli (reparam. trick)

[1] Kobayashi, Goro, et al. "Attention is not only a weight: Analyzing transformers with vector norms." EMNLP (2020)
[2] Sundararajan, Mukund, Ankur Taly, and Qiqi Yan. "Axiomatic attribution for deep networks." ICML (2017)
[3] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier." SIGKDD (2016).

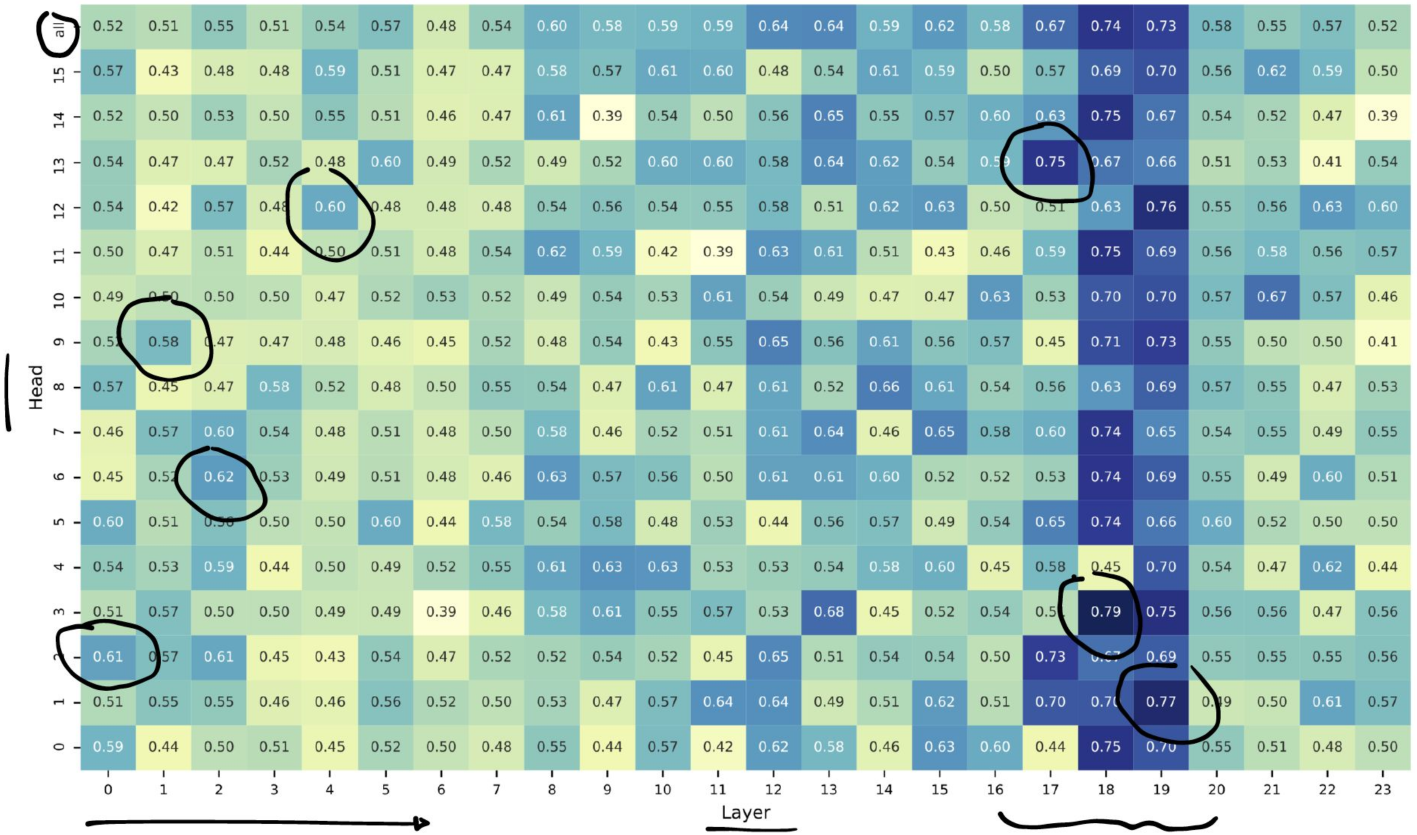
WMT 2022 QE Task: Unbabel-IST Submission

Attention heads provide good explanations!



* Results from IST-Unbabel 2021 Submission for the Explainable Quality Estimation Shared Task (Treviso et al., Eval4NLP 2021)

WMT 2022 QE Task: Unbabel-IST Submission



* Results from [IST-Unbabel 2021 Submission for the Explainable Quality Estimation Shared Task](#) (Treviso et al., Eval4NLP 2021)

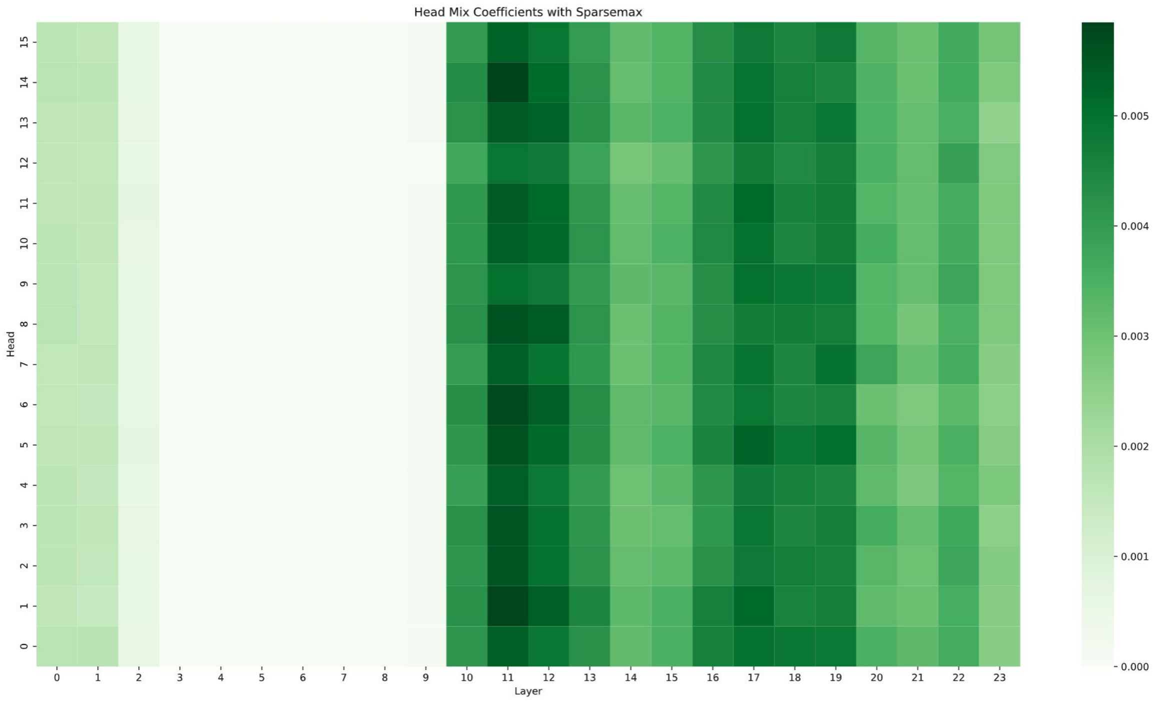
WMT 2022 QE Task: Unbabel-IST Submission

We take advantage of the results from last year and we build a **final layer that produces an output vector by attending on a subset of attention heads using sparsemax**

This means that the model will learn to ignore several heads.. This has two effects:

1) Forces the model to focus on relevant heads

2) Reduces the search space for heads that correlate with MT errors.



WMT 2022 QE Final Results

Official results: https://www.statmt.org/wmt22/quality-estimation-task_results.html

	DA								MQM		
Team	**en-cs**	**en-ja**	**en-mr**	**en-yo**	**km-en**	**ps-en**	**all**	**all/yo**	**en-ru**	**en-de**	**zh-en**
Sentence-level QE											
Baseline	0.560	0.272	0.436	0.002	0.579	0.641	0.415	0.497	0.333	0.455	0.164
Alibaba	-	-	-	-	-	-	-	-	0.505	0.550	0.347
NJUQE	-	-	0.585	-	-	-	-	-	0.474	**0.635**	0.296
Welocalize	0.563	0.276	0.444	-	0.623	-	0.448	0.506	-	-	-
hui	0.562	0.318	0.568	0.064	0.610	0.656	0.463	0.542	0.334	0.501	0.240
joanne.wjy	0.635	0.348	0.597	-	0.657	0.697	-	0.587	-	-	-
HW-TSC	0.626	0.341	0.567	-	0.509	0.661	-	-	0.433	0.494	**0.369**
Papago	0.636	0.327	**0.604**	0.121	0.653	0.671	0.502	0.571	0.496	0.582	0.325
IST-Unbabel	**0.655**	**0.385**	0.592	**0.409**	**0.669**	**0.722**	**0.572**	**0.605**	**0.519**	0.561	0.348
Word-level QE											
Baseline	0.325	0.175	0.306	0.000	0.402	0.359	0.235	0.257	0.203	0.182	0.104
NJUQE	-	-	0.412	-	0.421	-	-	-	0.390	**0.352**	0.308
HW-TSC	0.424	**0.258**	0.351	-	0.353	0.358	-	0.218	0.343	0.274	0.246
Papago	0.396	0.257	**0.418**	0.028	**0.429**	0.374	0.317	0.343	0.421	0.319	0.351
IST-Unbabel	**0.436**	0.238	0.392	**0.131**	0.425	**0.424**	**0.341**	**0.361**	**0.427**	0.303	**0.360**
Explainable QE											
Baseline	0.417	0.367	0.194	0.111	0.580	0.615	0.381	0.435	0.148	0.074	0.048
f.azadi	-	-	-	-	0.622	0.668	-	-	-	-	-
HW-TSC	0.536	0.462	0.280	-	**0.686**	**0.715**	-	0.535	0.313	0.252	0.220
IST-Unbabel	**0.561**	**0.466**	**0.317**	**0.234**	0.665	0.672	**0.486**	**0.536**	**0.390**	**0.365**	**0.379**

Table 6: Official results for sentence-level QE (top) in terms of Spearman's correlation, word-level QE (middle) in terms of MCC, and explainable QE (bottom) in terms of R@K.

WMT 2022 QE Final Results

Official results: https://www.statmt.org/wmt22/quality-estimation-task_results.html

	DA								MQM		
Team	**en-cs**	**en-ja**	**en-mr**	**en-yo**	**km-en**	**ps-en**	**all**	**all/yo**	**en-ru**	**en-de**	**zh-en**
Sentence-level QE											
Baseline	0.560	0.272	0.436	0.002	0.579	0.641	0.415	0.497	0.333	0.455	0.164
Alibaba	-	-	-	-	-	-	-	-	0.505	0.550	0.347
NJUQE	-	-	0.585	-	-	-	-	-	0.474	**0.635**	0.296
Welocalize	0.563	0.276	0.444	-	0.623	-	0.448	0.506	-	-	-
hui	0.562	0.318	0.568	0.064	0.610	0.656	0.463	0.542	0.334	0.501	0.240
joanne.wjy	0.635	0.348	0.597	-	0.657	0.697	-	0.587	-	-	-
HW-TSC	0.626	0.341	0.567	-	0.509	0.661	-	-	0.433	0.494	**0.369**
Papago	0.636	0.327	**0.604**	0.121	0.653	0.671	0.502	0.571	0.496	0.582	0.325
IST-Unbabel	**0.655**	**0.385**	0.592	**0.409**	**0.669**	**0.722**	**0.572**	**0.605**	**0.519**	0.561	0.348
Word-level QE											
Baseline	0.325	0.175	0.306	0.000	0.402	0.359	0.235	0.257	0.203	0.182	0.104
NJUQE	-	-	0.412	-	0.421	-	-	-	0.390	**0.352**	0.308
HW-TSC	0.424	**0.258**	0.351	-	0.353	0.358	-	0.218	0.343	0.274	0.246
Papago	0.396	0.257	**0.418**	0.028	**0.429**	0.374	0.317	0.343	0.421	0.319	0.351
IST-Unbabel	**0.436**	0.238	0.392	**0.131**	0.425	**0.424**	**0.341**	**0.361**	**0.427**	0.303	**0.360**
Explainable QE											
Baseline	0.417	0.367	0.194	0.111	0.580	0.615	0.381	0.435	0.148	0.074	0.048
f.azadi	-	-	-	-	0.622	0.668	-	-	-	-	-
HW-TSC	0.536	0.462	0.280	-	**0.686**	**0.715**	-	0.535	0.313	0.252	0.220
IST-Unbabel	**0.561**	**0.466**	**0.317**	**0.234**	0.665	0.672	**0.486**	**0.536**	**0.390**	**0.365**	**0.379**

Table 6: Official results for sentence-level QE (top) in terms of Spearman's correlation, word-level QE (middle) in terms of MCC, and explainable QE (bottom) in terms of R@K.

WMT 2022 QE Final Results

Official results: https://www.statmt.org/wmt22/quality-estimation-task_results.html

	DA								MQM		
Team	**en-cs**	**en-ja**	**en-mr**	**en-yo**	**km-en**	**ps-en**	**all**	**all/yo**	**en-ru**	**en-de**	**zh-en**
Sentence-level QE											
Baseline	0.560	0.272	0.436	0.002	0.579	0.641	0.415	0.497	0.333	0.455	0.164
Alibaba	-	-	-	-	-	-	-	-	0.505	0.550	0.347
NJUQE	-	-	0.585	-	-	-	-	-	0.474	**0.635**	0.296
Welocalize	0.563	0.276	0.444	-	0.623	-	0.448	0.506	-	-	-
hui	0.562	0.318	0.568	0.064	0.610	0.656	0.463	0.542	0.334	0.501	0.240
joanne.wjy	0.635	0.348	0.597	-	0.657	0.697	-	0.587	-	-	-
HW-TSC	0.626	0.341	0.567	-	0.509	0.661	-	-	0.433	0.494	**0.369**
Papago	0.636	0.327	**0.604**	0.121	0.653	0.671	0.502	0.571	0.496	0.582	0.325
IST-Unbabel	**0.655**	**0.385**	0.592	**0.409**	**0.669**	**0.722**	**0.572**	**0.605**	**0.519**	0.561	0.348
Word-level QE											
Baseline	0.325	0.175	0.306	0.000	0.402	0.359	0.235	0.257	0.203	0.182	0.104
NJUQE	-	-	0.412	-	0.421	-	-	-	0.390	**0.352**	0.308
HW-TSC	0.424	**0.258**	0.351	-	0.353	0.358	-	0.218	0.343	0.274	0.246
Papago	0.396	0.257	**0.418**	0.028	**0.429**	0.374	0.317	0.343	0.421	0.319	0.351
IST-Unbabel	**0.436**	0.238	0.392	**0.131**	0.425	**0.424**	**0.341**	**0.361**	**0.427**	0.303	**0.360**
Explainable QE											
Baseline	0.417	0.367	0.194	0.111	0.580	0.615	0.381	0.435	0.148	0.074	0.048
f.azadi	-	-	-	-	0.622	0.668	-	-	-	-	-
HW-TSC	0.536	0.462	0.280	-	**0.686**	**0.715**	-	0.535	0.313	0.252	0.220
IST-Unbabel	**0.561**	**0.466**	**0.317**	**0.234**	0.665	0.672	**0.486**	**0.536**	**0.390**	**0.365**	**0.379**

Table 6: Official results for sentence-level QE (top) in terms of Spearman's correlation, word-level QE (middle) in terms of MCC, and explainable QE (bottom) in terms of R@K.

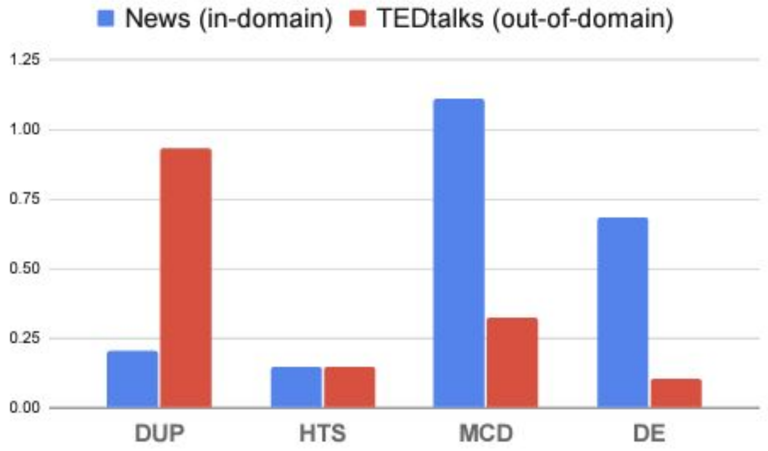
Can we handle *uncertainty* in quality scores?

Uncertainty-Aware MT Quality Evaluation

- Instead of predicting a quality score, predict a **confidence interval.**

- Some methods can capture both
  - **epistemic** (model) uncertainty (e.g. out-of-domain data, complex sentences)
  - **aleatoric** (data) uncertainty (e.g. noisy references, annotator disagreement)

MT	DA	COMET	UA-COMET
Она сказала, ’Это не собирается работать.	-0.815	*0.586*	0.149 [-0.92, 1.22]
Gloss: “*She said, ‘that’s not willing to work*”			
Она сказала: «Это не сработает.	0.768	1.047	1.023 [0.673, 1.374]
Gloss: “*She said, «That will not work*”			

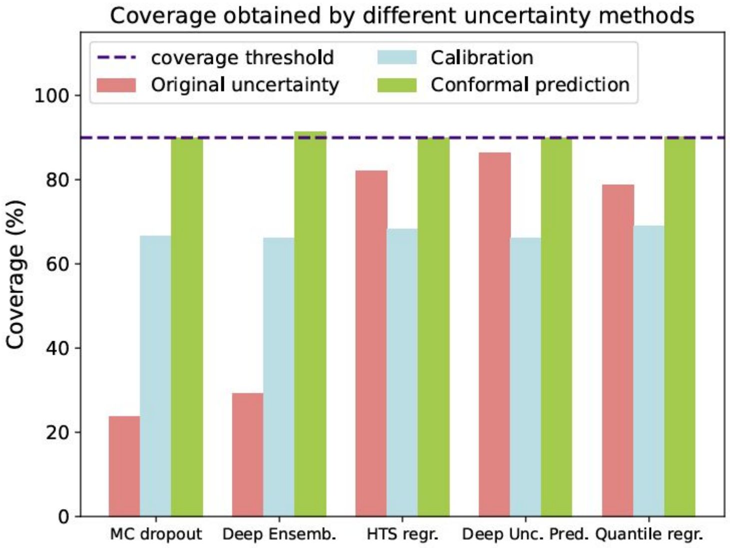


“Uncertainty-Aware Machine Translation Evaluation”. T. Glushkova, C. Zerva, R. Rei, A. Martins. Findings of EMNLP 2021.

“Disentangling Uncertainty in Machine Translation Evaluation”. C. Zerva, T. Glushkova, R. Rei, A. Martins. EMNLP 2022.

Conformalizing MT Quality Evaluation

- Returns a confidence interval with guaranteed **coverage** (contains the true score with 90% probability)
- Can also do **equalized coverage – e.g. coverage spread equally across languages.**



	QNT	MCD	DE	HTS	DUP
En-Cs	0.982	0.959	0.939	0.875	0.931
En-De	0.973	0.971	0.925	0.863	0.927
En-Ja	0.990	0.978	0.987	0.886	0.972
En-Pl	0.977	0.948	0.914	0.882	0.914
En-Ru	0.974	0.958	0.936	0.862	0.926
En-Ta	0.970	0.952	0.949	0.892	0.858
En-Zh	0.934	0.983	0.991	0.919	0.945
Cs-En	0.890	0.871	0.884	0.898	0.875
De-En	0.880	0.888	0.867	0.896	0.902
Ja-En	0.883	0.856	0.921	0.910	0.887
Kn-En	0.881	0.875	0.948	0.943	0.840
Pl-En	0.862	0.833	0.825	0.873	0.849
Ps-En	0.851	0.854	0.932	0.922	0.786
Ru-En	0.851	0.828	0.831	0.879	0.888
Ta-En	0.793	0.809	0.878	0.898	0.883
Zh-En	0.861	0.833	0.868	0.886	0.827

Non-equalized

→

	QNT	MCD	DE	HTS	DUP
En-Cs	0.893	0.917	0.888	0.892	0.902
En-De	0.902	0.902	0.902	0.896	0.893
En-Ja	0.909	0.891	0.900	0.891	0.904
En-Pl	0.882	0.905	0.895	0.900	0.898
En-Ru	0.900	0.898	0.908	0.906	0.903
En-Ta	0.903	0.895	0.883	0.886	0.903
En-Zh	0.880	0.890	0.884	0.896	0.896
Cs-En	0.890	0.917	0.909	0.904	0.894
De-En	0.897	0.901	0.901	0.897	0.903
Ja-En	0.900	0.912	0.899	0.894	0.902
Kn-En	0.896	0.903	0.902	0.904	0.894
Pl-En	0.900	0.905	0.893	0.894	0.877
Ps-En	0.905	0.899	0.900	0.884	0.907
Ru-En	0.910	0.896	0.907	0.900	0.900
Ta-En	0.884	0.901	0.886	0.901	0.908
Zh-En	0.900	0.910	0.908	0.900	0.905

Equalized

"Conformalizing Machine Translation Evaluation". C. Zerva and A. Martins. 2023.

Can we learn to predict *error spans* from human annotations?

Looking back at MQM:

English to Spanish

Source	Translation (Spanish Informal)	Quality
I am giving a talk in Mexico.	Estoy dando una charla en México.	Best
I am giving a talk in Mexico.	Estoy dando un charla en México.	Weak
I am giving a talk in Mexico.	Estoy visitando las pirámides en México.	Weak

https://qi.unbabel.com/

Looking back at MQM:

English to Spanish

Source	Translation (Spanish Informal)	Quality
I am giving a talk in Mexico.	Estoy dando una charla en México.	Best
I am giving a talk in Mexico.	Estoy dando un charla en México.	Weak
I am giving a talk in Mexico.	Estoy visitando las pirámides en México.	Weak

https://qi.unbabel.com/

How can we predict errors and their severities?

xCOMET:
Fine-Grained Automatic MT Evaluation

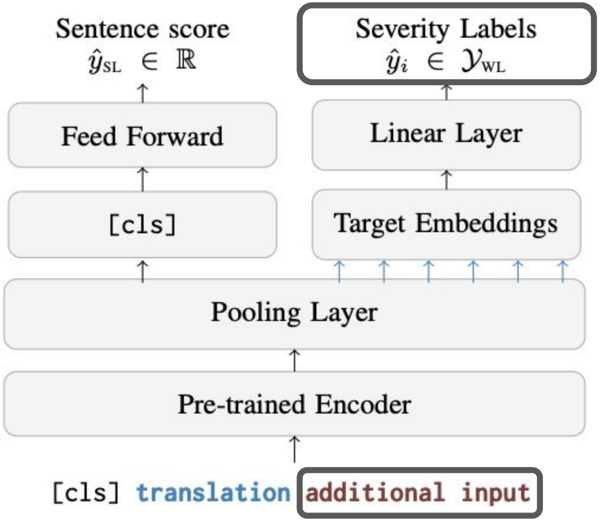
xCOMET: Transparent Machine Translation Evaluation through Fine-grained Error Detection

Nuno M. Guerreiro*[1,3,4,5], Ricardo Rei*[1,2,5], Daan van Stigt[1], Luisa Coheur[2,5], Pierre Colombo[4], André F. T. Martins[1,3,5]
[1]Unbabel, Lisbon, Portugal, [2]INESC-ID, Lisbon, Portugal
[3]Instituto de Telecomunicações, Lisbon, Portugal
[4]MICS, CentraleSupélec, Université Paris-Saclay, France
[5]Instituto Superior Técnico, University of Lisbon, Portugal

New: xCOMET

Single model that:

- **can be used as a metric or as a QE system:**
    - Reference-based (ref-only and src+ref)
    - Quality estimation (src-only)
- **can be used to score translations at the sentence level but also predict error spans (as MQM annotations)**



“[xCOMET: Transparent Machine Translation Evaluation through Fine-grained Error Detection]”.
N. Guerreiro, R. Rei, D. Stigt, L. Coheur, P. Colombo, A. Martins.
TACL 2024.

Curriculum learning

xCOMET models undergo a 3-phase curriculum training.

- **Phase 1:** the model is trained exclusively on DA data, with sole focus on sentence-level regression

Warm-up

- **Phase 2:** we introduce word-level supervision; we continue training the model on MQM data (most emphasis on word-level task)

Shift the focus to word-level without compromising sentence-level capabilities

- **Phase 3:** we unify both tasks; we give more emphasis on sentence-level and use very high-quality MQM data

Mitigate potential decline of sentence-level capabilities from Phase 2

Correlation with human judgments

Sentence-level (WMT 22 News)

Metric	zh-en		en-de		en-ru		Avg.	
	ρ	τ	ρ	τ	ρ	τ	ρ	τ
BLEURT-20	0.462	0.336	0.568	0.380	0.498	0.379	0.509	0.365
COMET-22	0.423	0.335	0.581	0.369	0.516	0.391	0.507	0.361
METRICX	**0.573**	**0.415**	**0.640**	0.405	0.581	0.444	0.598	0.421
GEMBA-GPT4-DA*	0.318	0.292	0.508	0.387	0.454	0.383	0.427	0.354
xCOMET-XL	0.556	0.399	**0.653**	0.414	0.611	0.448	0.607	0.420
xCOMET-XXL	0.554	0.390	**0.644**	**0.435**	**0.628**	**0.470**	**0.609**	**0.432**
Predicted MQM scores from the error spans ($\hat{y} = \hat{y}_{\text{MQM}}$)								
xCOMET-XL (MQM)	0.447	0.374	0.561	0.389	0.534	0.445	0.514	0.402
xCOMET-XXL (MQM)	0.446	0.332	0.597	0.415	0.533	0.439	0.525	0.395

State-of-the-art metric, outperforming both MetricX and GPT-4 based sentence-level evaluation.

Correlation with human judgments

Sentence-level (WMT 22 News)

	zh-en		en-de		en-ru		Avg.	
METRIC	ρ	τ	ρ	τ	ρ	τ	ρ	τ
BLEURT-20	0.462	0.336	0.568	0.380	0.498	0.379	0.509	0.365
COMET-22	0.423	0.335	0.581	0.369	0.516	0.391	0.507	0.361
METRICX	**0.573**	**0.415**	**0.640**	0.405	0.581	0.444	0.598	0.421
GEMBA-GPT4-DA*	0.318	0.292	0.508	0.387	0.454	0.383	0.427	0.354
xCOMET-XL	0.556	0.399	**0.653**	0.414	0.611	0.448	0.607	0.420
xCOMET-XXL	0.554	0.390	**0.644**	**0.435**	**0.628**	**0.470**	**0.609**	**0.432**
Predicted MQM scores from the error spans ($\hat{y} = \hat{y}_{\text{MQM}}$)								
xCOMET-XL (MQM)	0.447	0.374	0.561	0.389	0.534	0.445	0.514	0.402
xCOMET-XXL (MQM)	0.446	0.332	0.597	0.415	0.533	0.439	0.525	0.395

The inferred MQM scores via xCOMET's error span predictions are very competitive with widely used metrics.

Correlation with human judgments

System-level evaluation

Metric	zh-en	en-de	en-ru	Avg.
BLEURT-20	0.762	0.771	0.743	0.759
COMET-22	0.705	0.800	0.733	0.746
MetricX	0.762	0.781	0.724	0.756
GEMBA-GPT4-DA	0.752	**0.848**	**0.876**	**0.825**
xCOMET-XL	**0.800**	0.743	0.790	0.778
xCOMET-XXL	**0.800**	**0.829**	**0.829**	**0.819**
MQM scores from the error spans ($\hat{y} = \hat{y}_{\text{MQM}}$)				
xCOMET-XL (MQM)	0.781	0.762	0.762	0.768
xCOMET-XXL (MQM)	0.781	**0.838**	0.810	0.810

WMT 22 News

Metric		avg corr
XCOMET-Ensemble	**1**	**0.825**
XCOMET-QE-Ensemble*	2	0.808
MetricX-23	2	0.808
GEMBA-MQM*	2	0.802
MetricX-23-QE*	2	0.800
mbr-metricx-qe*	3	0.788
MaTESe	3	0.782
CometKiwi*	3	0.782
COMET	3	0.779
BLEURT-20	3	0.776
KG-BERTScore*	3	0.774
sescoreX	3	0.772
cometoid22-wmt22*	4	0.772
docWMT22CometDA	4	0.768
docWMT22CometKiwiDA*	4	0.767

WMT 23 Metrics Shared Task

Correlation with human judgments

System-level evaluation

Metric	zh-en	en-de	en-ru	Avg.
BLEURT-20	0.762	0.771	0.743	0.759
COMET-22	0.705	0.800	0.733	0.746
METRICX	0.762	0.781	0.724	0.756
GEMBA-GPT4-DA	0.752	**0.848**	**0.876**	**0.825**
xCOMET-XL	**0.800**	0.743	0.790	0.778
xCOMET-XXL	**0.800**	**0.829**	**0.829**	**0.819**
MQM scores from the error spans ($\hat{y} = \hat{y}_{\mathrm{MQM}}$)				
xCOMET-XL (MQM)	0.781	0.762	0.762	0.768
xCOMET-XXL (MQM)	0.781	**0.838**	0.810	0.810

MQM inferred scores doing really well again!

WMT 22 News

Metric		avg corr
XCOMET-Ensemble	**1**	**0.825**
XCOMET-QE-Ensemble*	2	0.808
MetricX-23	2	0.808
GEMBA-MQM*	2	0.802
MetricX-23-QE*	2	0.800
mbr-metricx-qe*	3	0.788
MaTESe	3	0.782
CometKiwi*	3	0.782
COMET	3	0.779
BLEURT-20	3	0.776
KG-BERTScore*	3	0.774
sescoreX	3	0.772
cometoid22-wmt22*	4	0.772
docWMT22CometDA	4	0.768
docWMT22CometKiwiDA*	4	0.767

WMT 23 Metrics Shared Task

Correlation with human judgments

Error span prediction

METRIC	zh-en	en-de	en-ru	Avg.
AutoMQM (GPT3.5)	0.143	0.160	0.166	0.156
AutoMQM (GPT4)	0.248	0.257	**0.281**	0.262
XCOMET-XL	0.237	0.290	**0.281**	0.269
XCOMET-XXL	**0.257**	**0.320**	0.262	**0.280**
Error spans detected with source-only				
XCOMET-XL (SRC)	0.208	0.264	0.252	0.242
XCOMET-XXL (SRC)	0.229	0.298	0.238	0.255

LLM-based evaluation

QE-style span detection outperforms AutoMQM (ref-based) w/ GPT3.5

State-of-the-art metric in error span prediction, outperforming AutoMQM approaches w/ generative LLMs.

Does xCOMET Satisfy Our Requirements?

	BLEU	COMET
Strong correlation with human judgments	❌	✅
Applicable to a wide range of languages and domains	❓	✅
Interpretable	❓	❌
Fast and lightweight	✅	❌

Does xCOMET Satisfy Our Requirements?

	BLEU	COMET	xCOMET
Strong correlation with human judgments	❌	✅	✅
Applicable to a wide range of languages and domains	❓	✅	✅
Interpretable	❓	❌	✅
Fast and lightweight	✅	❌	❓

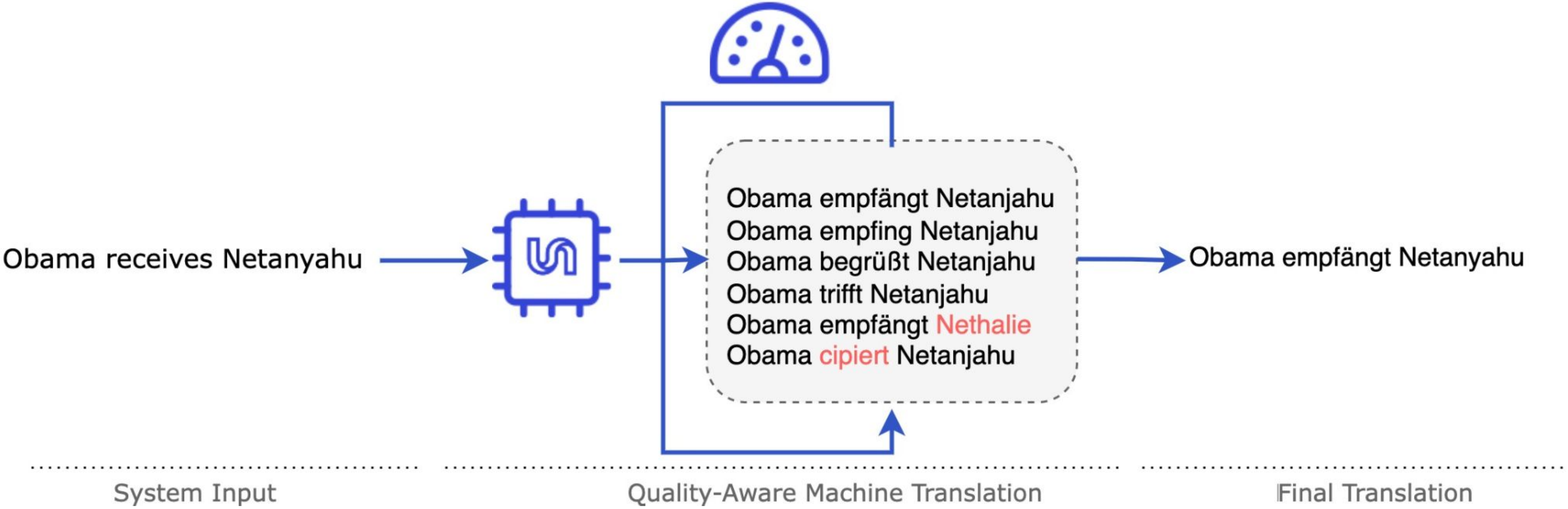
Does xCOMET Satisfy Our Requirements?

	BLEU	COMET	xCOMET
Strong correlation with human judgments	❌	✅	✅
Applicable to a wide range of languages and domains	❓	✅	✅
Interpretable	❓	❌	✅
Fast and lightweight	✅	❌	❓

COMETinho is a step in this direction!
(Rei et al., EAMT 2022)

Can we use QE to make MT better?

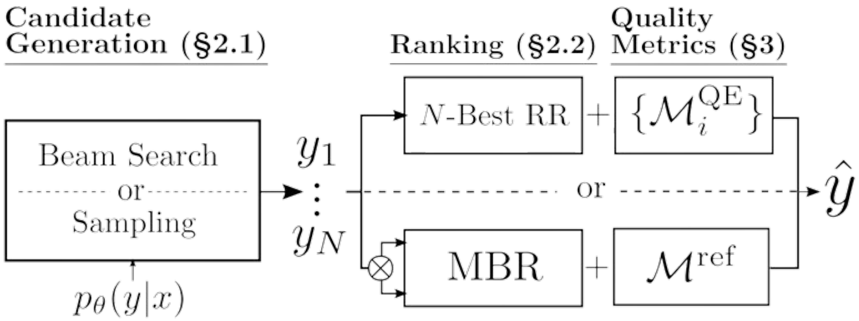
Quality Aware Decoding*:



* Quality-Aware Decoding for Neural Machine Translation (Fernandes et al., NAACL 2022)

Quality Aware Decoding

1) Translation **candidates are generated** according to the model;
2) Using reference-free and/or reference based MT metrics, these **candidates are ranked**;
3) The **highest ranked one is picked** as the final translation.



* Quality-Aware Decoding for Neural Machine Translation (Fernandes et al., NAACL 2022)

Quality Aware Decoding:
Impact on MQM

	EN-DE (WMT20)				EN-RU (WMT20)			
	Minor	Major	Critical	MQM	Minor	Major	Critical	MQM
Reference	24	67	0	97.04	5	11	0	99.30
Baseline	8	139	0	95.66	17	239	49	79.78
F-RR w/ COMET-QE	15	204	0	93.47	13	254	80	76.25
T-RR w/ COMET	12	109	0	**96.20**	9	141	45	85.97[†]
MBR w/ COMET	11	161	0	94.38	8	182	40	83.65
T-RR + MBR w/ COMET	10	138	0	95.44	11	134	45	**86.78**[†]

Also Works With LLM-based MT (even with few samples)

"An Empirical Study of Translation Hypothesis Ensembling with LLMs".
A. Farinhas, J. Souza, A. Martins. EMNLP 2023.

Coming next: LLM-based QE

GEMBA

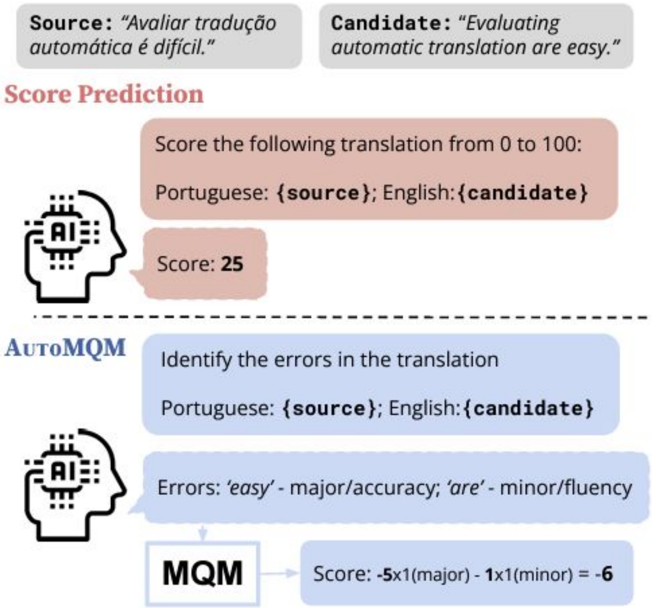
```
Score the following translation from {source_lang} to {target_lang} with respect
to the human reference on a continuous scale from 0 to 100, where score of zero means
"no meaning preserved" and score of one hundred means "perfect meaning and grammar".

{source_lang} source: "{source_seg}"
{target_lang} human reference: {reference_seg}
{target_lang} translation: "{target_seg}"
Score:
```

Metric	Acc	en-de	en-ru	zh-en
GEMBA-GPT4-DA	89.8%	0.36	0.36	0.38
GEMBA-Dav3-DA	88.0%	0.31	0.33	0.37
GEMBA-GPT4-DA[noref]	87.6%	0.31	0.40	0.41
GEMBA-Dav3-DA[noref]	86.1%	0.18	0.26	0.29
MetricX XXL	85.0%	0.36	**0.42**	**0.43**
BLEURT-20	84.7%	0.34	0.36	0.36
COMET-22	83.9%	**0.37**	0.40	**0.43**
UniTE	82.8%	**0.37**	0.38	0.36
COMETKiwi[noref]	78.8%	0.29	0.36	0.36
COMET-QE[noref]	78.1%	0.28	0.34	0.36
chrF	73.4%	0.21	0.17	0.15
BLEU	70.8%	0.17	0.14	0.14

Table 4: Kendall's Tau (τ) segment-level evaluation.

"Large Language Models Are State-of-the-Art Evaluators of Translation Quality". Tom Kocmi, Christian Federmann. EAMT 2023.

AutoMQM




```
Based on the given source and reference, identify the major and minor errors in this
translation. Note that Major errors refer to actual translation or grammatical errors,
and Minor errors refer to smaller imperfections, and purely subjective opinions about
the translation.

{src_lang} source: "{source}"
{tgt_lang} human reference: "{reference}"
{tgt_lang} translation: "{candidate}"
Errors: {error1:span} - {error1:severity}/{error1:category}; {error2:span} - ...
```

"[The devil is in the errors: Leveraging LLMs for fine-grained machine translation evaluation]".
P. Fernandes, D. Deutsch, M. Finkelstein, P. Riley, A. Martins, G. Neubig, A. Garg, J. Clark, M. Freitag, O. Firat.
WMT 2023.

This is becoming a very active area of research

See also:

- Gemba-MQM (Kocmi & Federmann, WMT 2023)
- InstructScore (Xu et al., EMNLP 2023)
- LLM-Refine (Xu et al., NAACL 2024)
- etc.

Tower:

An LLM for Translation-Related Tasks

TOWER: An Open Multilingual Large Language Model for Translation-Related Tasks

Duarte M. Alves[†2,4] José Pombal[†1] Nuno M. Guerreiro[†1,2,4,5]
Pedro H. Martins[1] João Alves[1] Amin Farajian[1] Ben Peters[2,4]
Ricardo Rei[1,3] Patrick Fernandes[2,4,7] Sweta Agrawal[⋆2]
Pierre Colombo[5,6] José G.C. de Souza[1] André F.T. Martins[1,2,4]

[1]Unbabel, [2]Instituto de Telecomunicações, [3]INESC-ID, [4]Instituto Superior Técnico & Universidade de Lisboa (Lisbon ELLIS Unit), [5]MICS, CentraleSupélec, Université Paris-Saclay, [6]Equall, [7]Carnegie Mellon University

A big team’s effort


André Martins


José Souza


Pierre Colombo


Graham Neubig


Nuno Guerreiro


João Alves


José Pombal


Pedro Martins


Ricardo Rei


Sweta Agrawal


Amin Farajian


Duarte Alves


Manuel Faysse

Ben Peters


Patrick
Fernandes

A big team’s effort


André Martins


José Souza


Pierre Colombo


Graham Neubig


Nuno Guerreiro


Duarte Alves

Instruction Tuning


João Alves


Manuel Faysse

Data


José Pombal

Ben Peters

Multilingua-lization


Pedro Martins


Patrick Fernandes

Pretraining


Ricardo Rei


Sweta Agrawal

Alignment


Amin Farajian


Vera Cabarrão


Marianna Buchicchio

Evaluation

Why the name Tower?

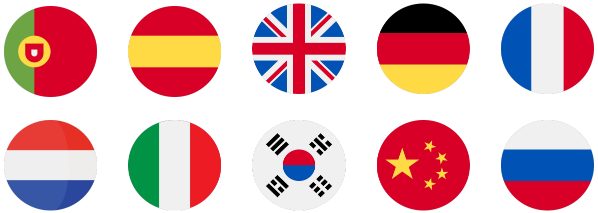
Unbabel

TÉCNICO LISBOA

université PARIS-SACLAY

The vision for Tower

Goal: create the best open multilingual LLM.



Focus (for now): **~10 languages** (mostly European).

In the future: more languages.

Optimized for **translation-related tasks**:

- Machine translation (MT)
- Quality Estimation (QE)
- Error span (MQM) prediction / explanations

- MT evaluation
- Source correction
- Automatic post-editing

The first suite of Tower models

Just released: Tower models that run at 7 and 13B params.

TowerBase

Base model with **improved multilingual performance.**

TowerInstruct

Optimized model (built on top of TowerBase) for **translation-related tasks.**

TowerBase

From LLaMA-2 to TowerBase.

Llama 2

- ✓ Suite of models of different size
- ✓ A lot of open research on top of the models
- ✗ Not great for multilingual tasks

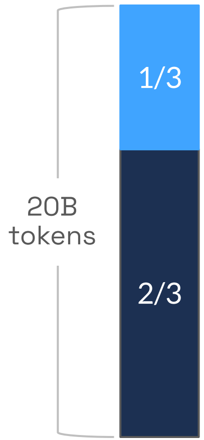
>

Extended multilingualization

How can we improve Llama 2 for multiple languages without compromising its general capabilities?

- **A** Just instruction-tuning for the tasks of interest
- **B** Continue pre-training on a large multilingual corpus (billions of tokens)
  - **B1** Use only monolingual data
  - **B2** Mix monolingual and parallel data

We built a corpus of 20B tokens with monolingual and parallel data



Parallel data

> We used **OPUS** data for each of the 20 language pairs with English.
> Filtering with **Bicleaner** and **CometKiwi-22**.
> **Uniform** weight across all language pairs.

Monolingual data

> We used data from **mC4** for each of the 10 languages.
> Filtering with **deduplication**, **language identification**, **perplexity**.
> **Uniform** weight across languages.

Details on training TowerBase

Addition of parallel data

We append the parallel data as different documents of the format:

```
{SRC_LANG}: {SRC}\n{TGT_LANG}:
{TGT}<EOS>
```

Training Conditions

Single node of 8 x A100 GPUs

We used Megatron-LM to train TowerBase

Training Time

10 days for TowerBase 7B

18-20 days for TowerBase 13B

TowerInstruct

From TowerBase to TowerInstruct.

TowerBase

- Multilingual capabilities
- Good few-shot performance
- No capability to follow instructions
- Suboptimal 0-shot performance

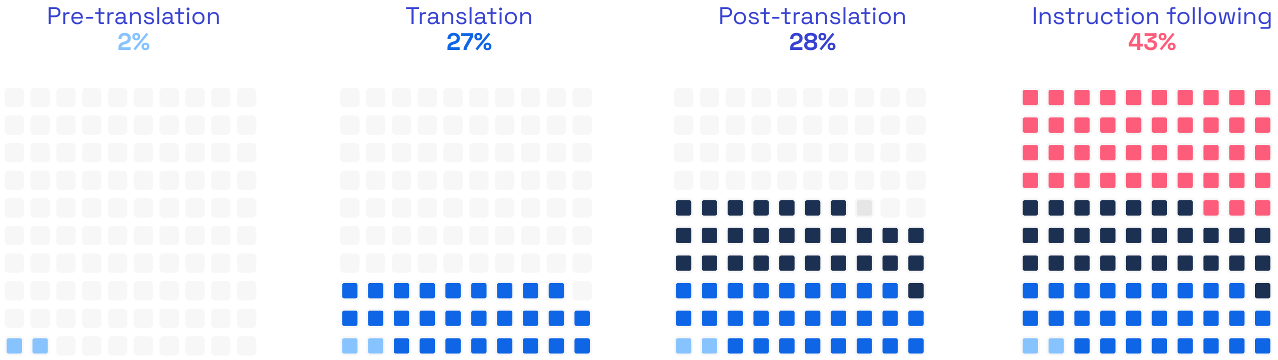
>

Instruction Tuning

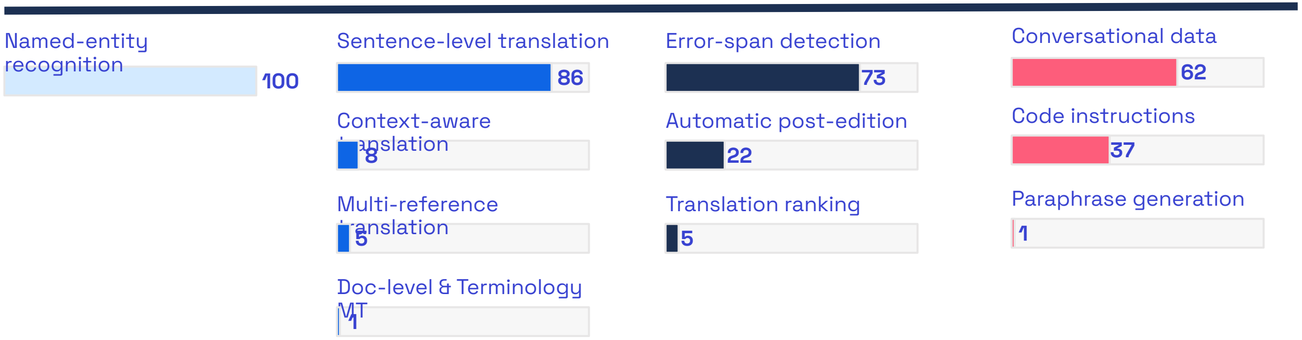
How can we improve Tower's capabilities for tasks of interest? How can we make it a conversational model?

- **A** Collect lots of supervised data and just train on that data
- **B** Collect fewer samples but guarantee they are high-quality
  - **B1** Use only supervised data
  - **B2** Leverage conversational data and synthetic data from SOTA LLMs (e.g., GPT-4)

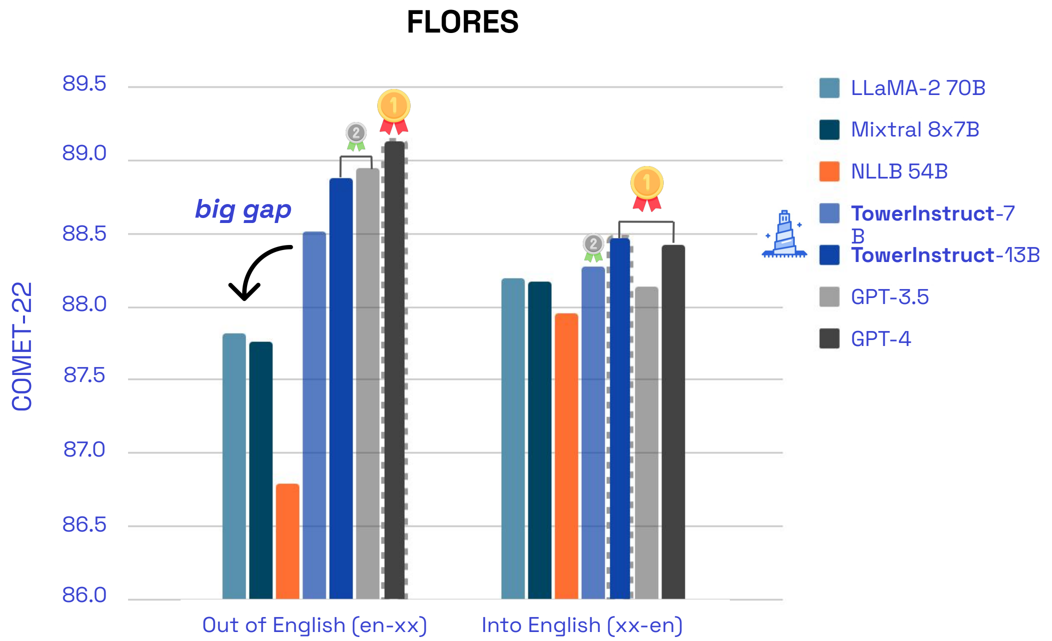
TowerBlocks balances translation-related data with instruction following data



Share of each task in its corresponding branch of **TowerBlocks**, %



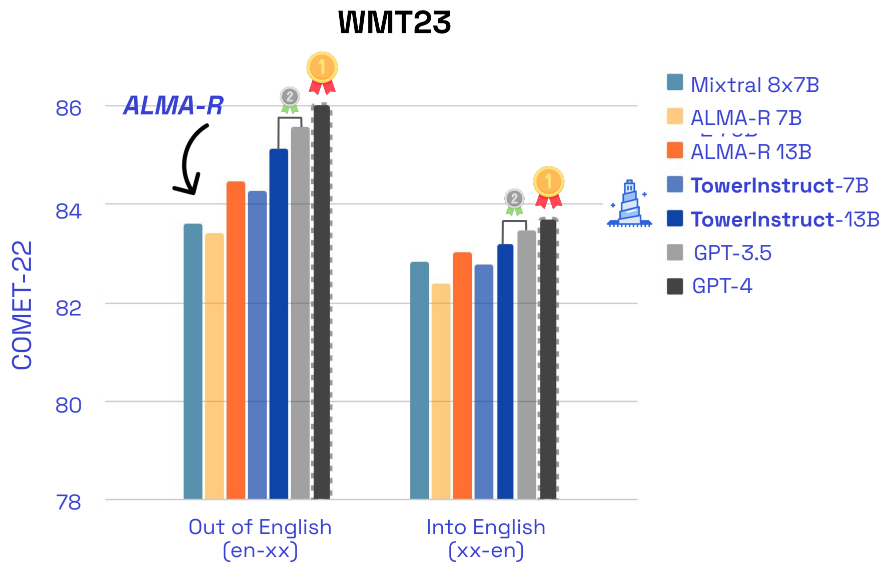
TowerInstruct outperforms all open-weight alternatives in sentence-level translation

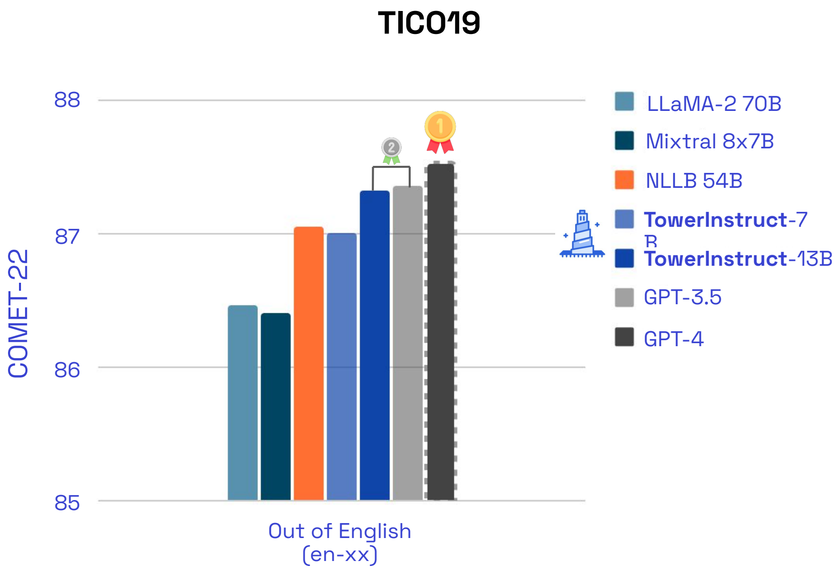


>

- **TowerInstruct (even the 7B)** models outperform other open-weight alternatives and dedicated models (even of much larger scales)
- **TowerInstruct** can be competitive with GPT-3.5 and GPT-4
- Performance in out-of-English could possibly be improved with further continued pre-training

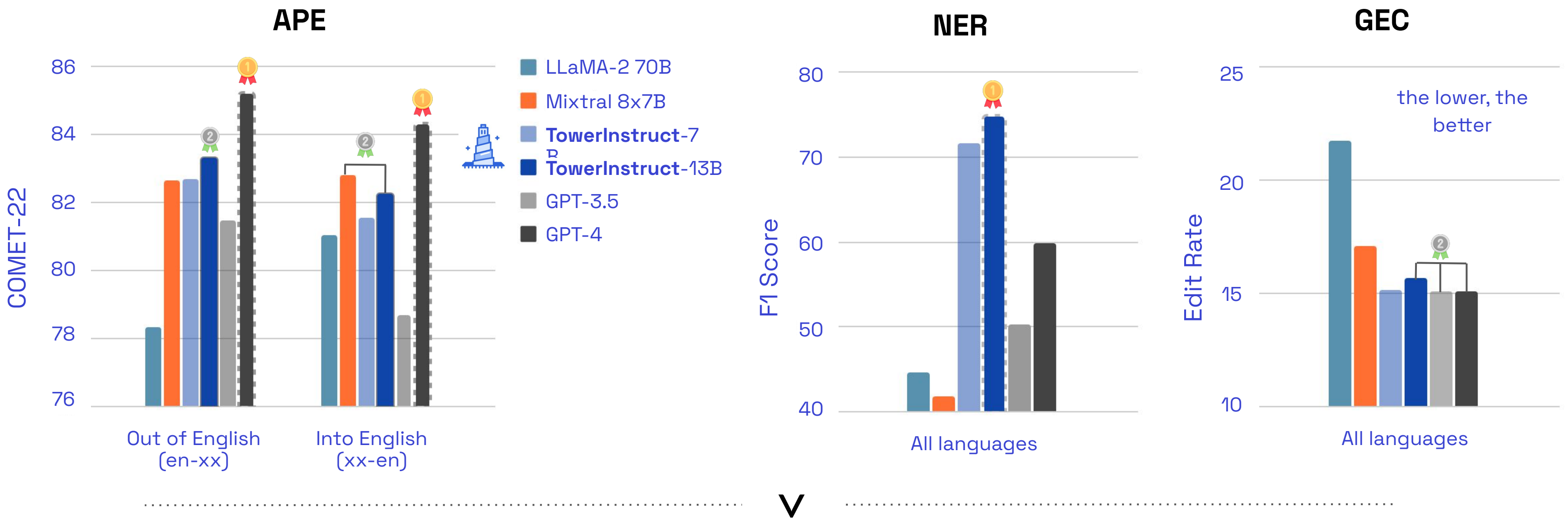
TowerInstruct is competitive with GPT-3.5 and outperforms ALMA-R, a dedicated LLM-based MT model.





- **TowerInstruct** outperforms ALMA-R (continued pre-trained LLaMA-2 + MT alignment) models across the board.
- **TowerInstruct** is competitive with GPT-3.5; still lags behind GPT-4.

TowerInstruct also showcases great performance in translation-related tasks



- **TowerInstruct** is an effective post-editor, second only to GPT-4.
- **TowerInstruct** outperforms all other models in NER.
- There is room for improvement in GEC, possibly because it is a held-out task.

Next steps: on the (modeling) road to EuroLLM…

We have been testing our codebase and experimental setup extensively on various pre-training runs at smaller scales.

Tower-1B model trained from scratch

- A **1.6B model trained from scratch on 100B tokens** on 12 different languages:
    - developed several scaling laws to predict the performance of the 1B model;
    - prevent problems in future runs (e.g., tokenization issues, etc.);
    - tested the pre-training codebase built on top of Megatron-Deepspeed.

Croissant LLM – a French & English model trained from scratch

- A **1.3B model trained from scratch on 3T tokens** for French and English:
    - tested the codebase for multi-node runs;
    - issue proofing modeling and tokenization;
    - study the impact of incorporating parallel data during pre-training.

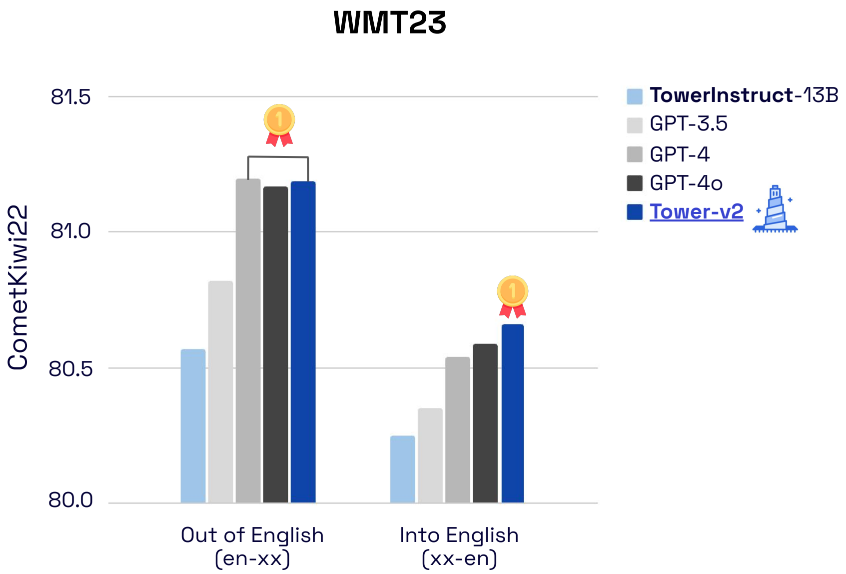
- **CroissantLLM** is a great bilingual model with exceptional performance in translation.

Funded by the European Union

A first look at... Tower v2

Tower-v2 models

- **At 7B parameters,** it outperforms across the board the previous TowerInstruct-13B model
- **Tower-v2** supports system prompts for better steerability and flexibility
- **Tower-v2 is now a fine-grained machine translation evaluator** with similar correlations as COMET-22
- **Improved translation capabilities** across all language pairs



Funded by the European Union

and EuroLLM

A suite of models for European languages to be trained on EuroHPC – MareNostrum 5

1

Dense models of 7B and 30B parameters.

We will train from scratch 7B and 30B parameter models.

These sizes will fit most needs for LLMs and go according to recent releases by big players.

2

The models will be trained on 4T tokens.

The best models out there are trained way beyond Chinchilla optimal.

These 4T tokens will include data for all official EU languages.

3

We will use scaling laws to predict our training.

We are currently running several scaling laws on data mixes in order to predict the quality of our models, including training a 1B model.

This gives us a principled way to guide all our design choices from architecture to data mix.

Conclusions

- Trained automatic metrics (e.g. COMET) can get **high correlations** with human judgments – however, **a single score is not enough**
- These metrics can be modified to provide fine-grained information such as **error spans** → **xCOMET**
- We can obtain strong **multilingual LLMs** by continued pretraining and careful instruction tuning of English-centric LLMs → **Tower**
- **Tower** is a state-of-the-art model for MT and other MT-related tasks
- **Tower v2** (to come soon) can also perform fine-grained MT evaluation.

TÉCNICO LISBOA Unbabel Center for Responsible AI

Questions?

andre.t.martins@tecnico.ulisboa.pt