Unbabel

inesc id lisboa
DEFINING TECHNOLOGY

Pushing the Limits of Machine Translation Quality Evaluation


Ricardo Rei, Research Engineer, PhD student


Catarina Farinha, Research Engineer, PhD

Deep Learning Sessions, 26th January 2022

Agenda

MAIA project and Unbabel's pipeline

COMET:

- Why we developed COMET?
- It's architecture
- Data used
- Results

More than just a score:

- MT-Telescope

Ongoing work:

- Reference-free Evaluation: COMET-QE
- A faster and lighter COMET: COMETinho
- Uncertainty-Aware MT Evaluation

Multilingual AI Augmented Agents

Unbabel

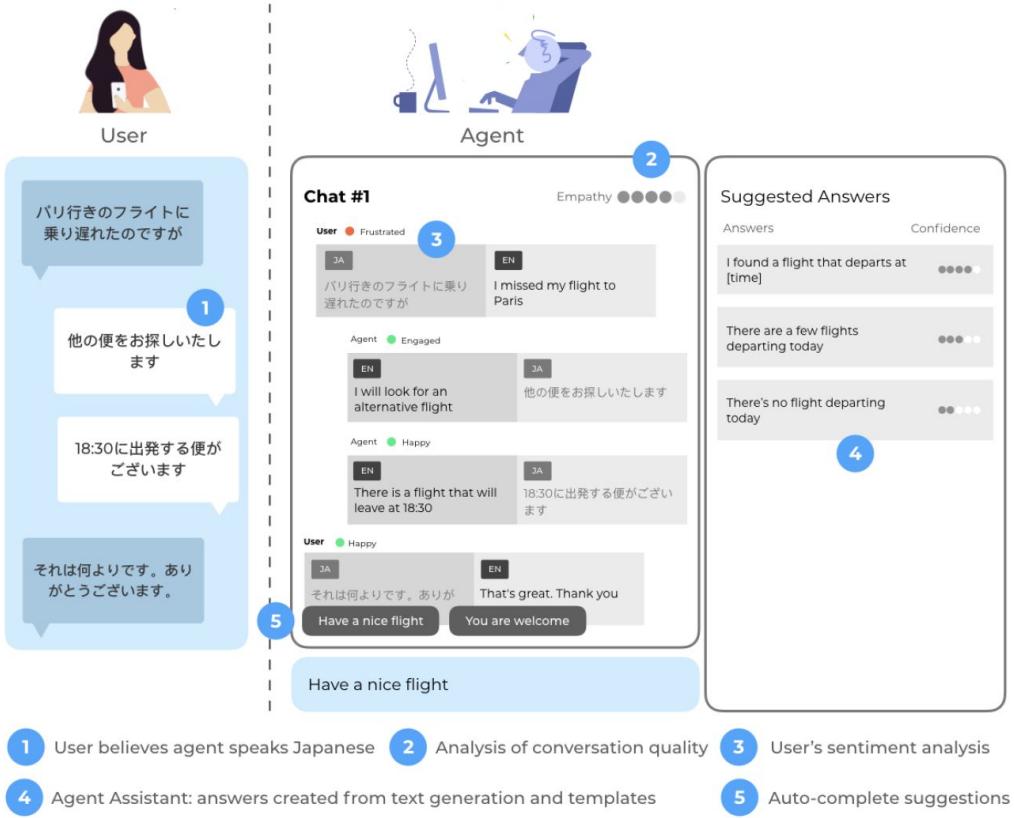
Carnegie Mellon University

inesc id lisboa

instituto de telecomunicações

Augmenting agents with multilingual superpowers.

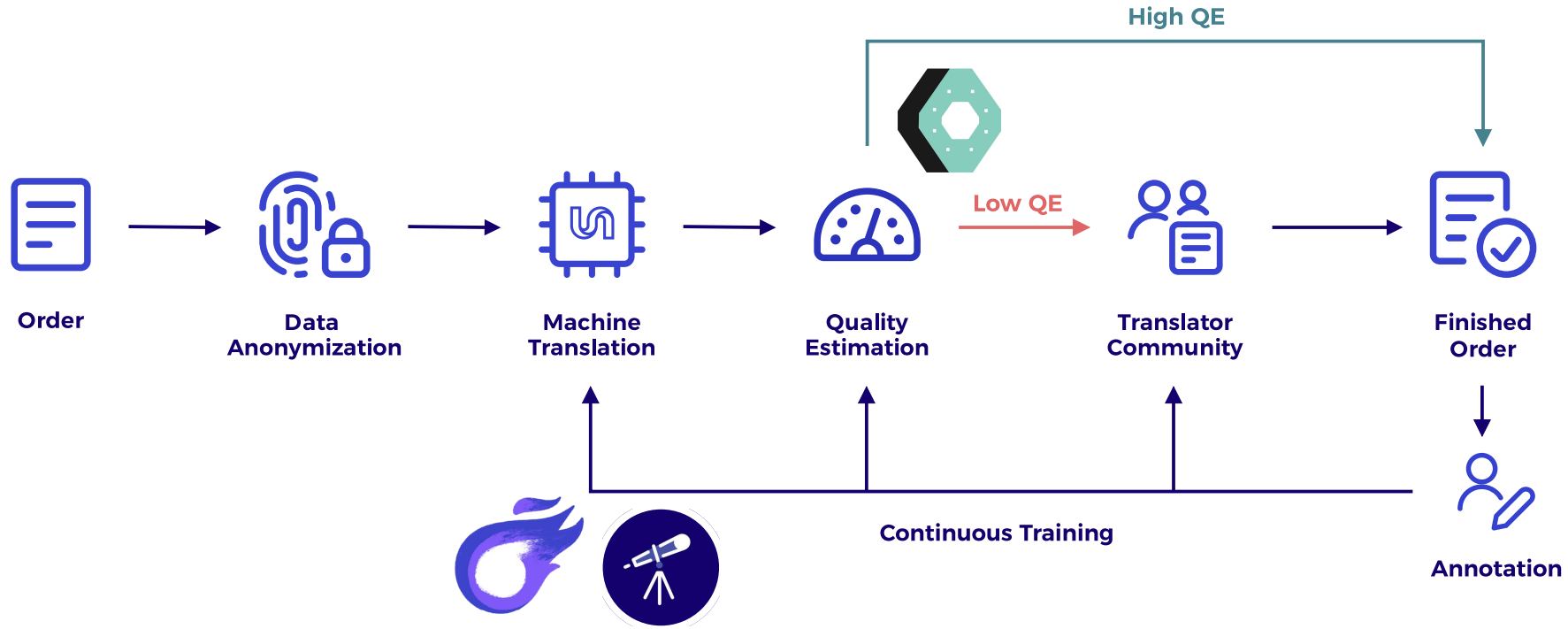
Multilingual AI Augmented Agents



COMING SOON

MAIA corpus
Genuine bilingual conversational corpus soon to be released!

Unbabel's pipeline



01

Introduction COMET

Motivation

Evaluation metrics shape the direction of research:

- We use them to compare experiments
- To decide what gets published
- To identify weakness and determine what to work on
- To decide which model we want to deploy
- etc...

But MT evaluation is hard:

- The reference isn't the only valid hypothesis/response
- Human evaluations are the gold standard but are costly and time consuming

Motivation

- Popular n-gram matching metrics: e.g. BLEU, METEOR, and chrF fail to recognize and capture semantic similarity beyond the lexical level

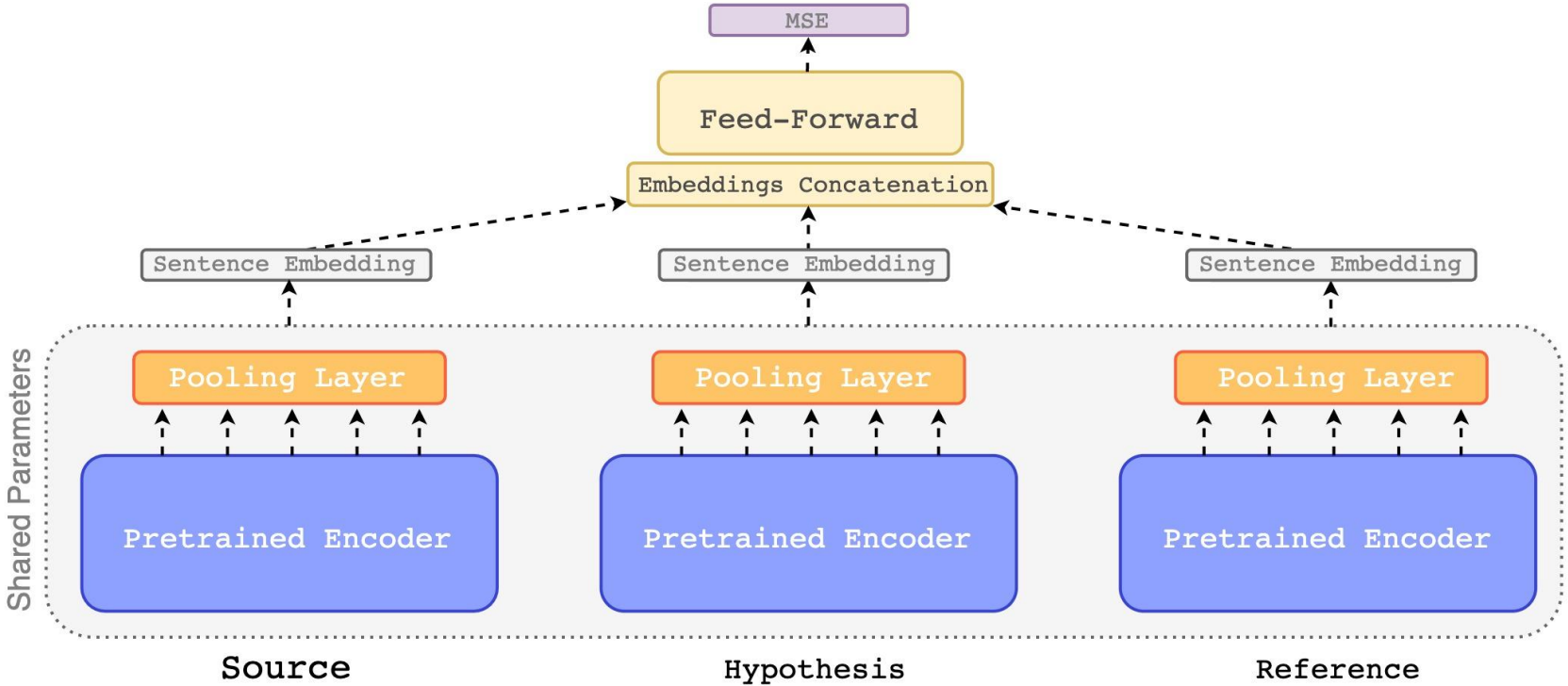
- Embedding-based metrics: e.g. BLEU2VEC, YiSi-1, MoverScore, and BERTscore try to capture semantic similarity but results are still far from perfect

- Learnable metrics: e.g. RUSE, BLEURT and PRISM attempt to directly optimize the correlation with human judgments

02

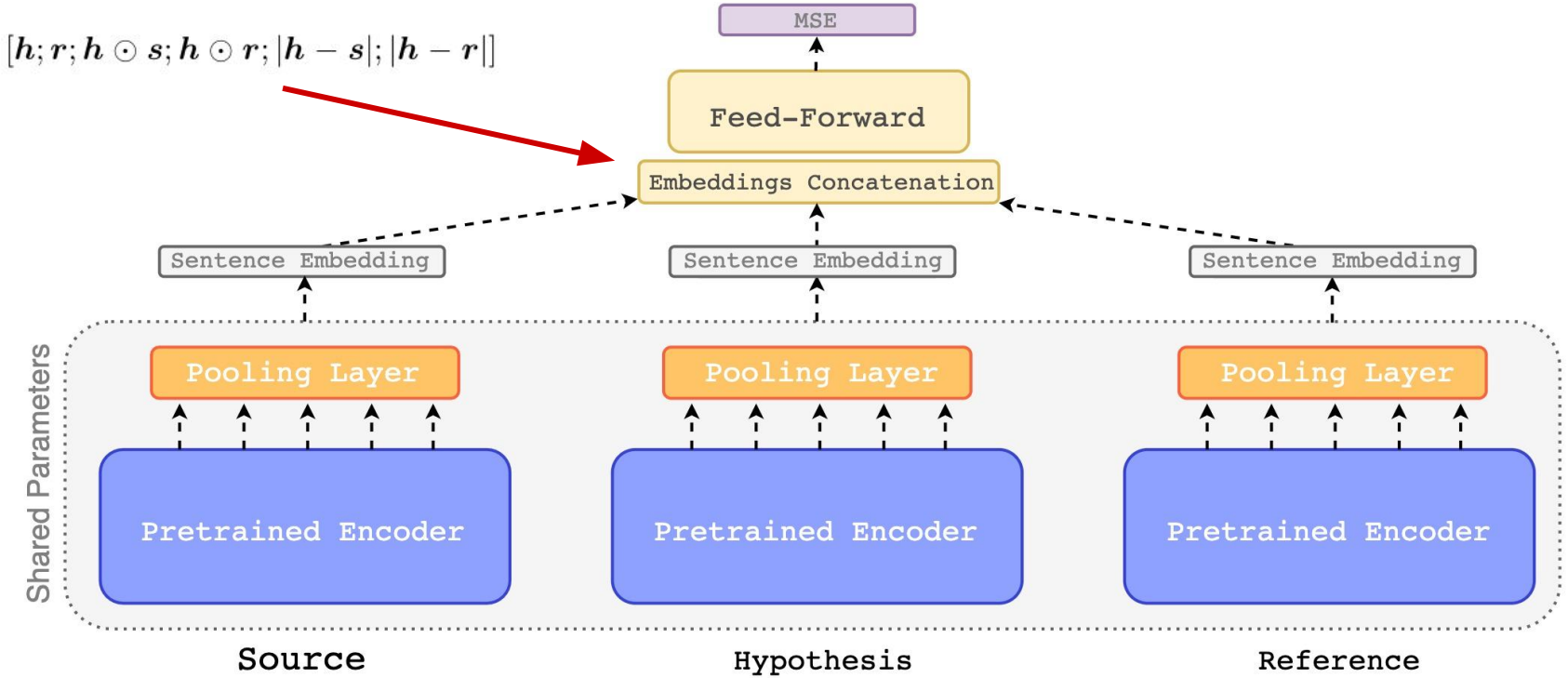
COMET Architecture

The Estimator Model



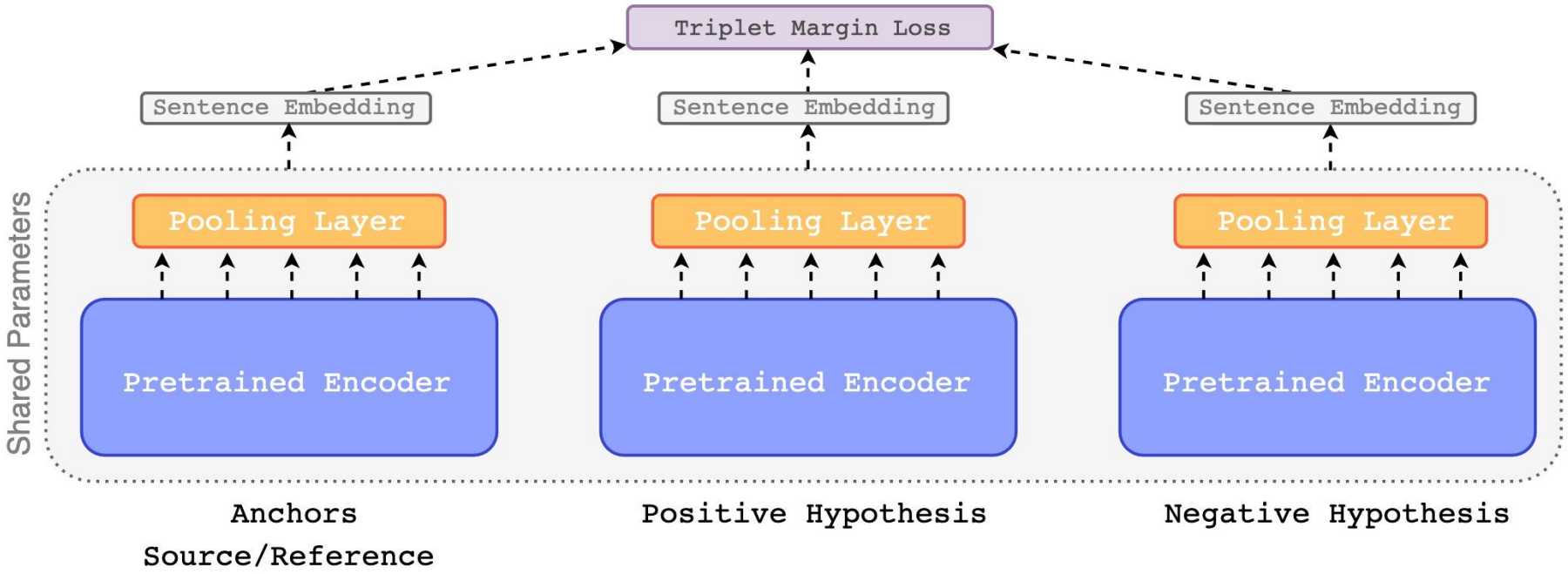
(Rei et al, 2020) - https://aclanthology.org/2020.emnlp-main.213/

The Estimator Model



(Rei et al, 2020) - https://aclanthology.org/2020.emnlp-main.213/

Translation Ranking Model



(Rei et al, 2020) - https://aclanthology.org/2020.emnlp-main.213/

03

Corpora

Types of human assessments:

1) Human-mediated Translation Edit Rate, HTER (for regression)
2) Multidimensional Quality Metrics, MQM (for regression)
3) Direct Assessments, DA (for regression and ranking)

QT21 Corpus (Specia, et al. 2017)

173K tuples with source sentence, respective human-generated reference, MT hypothesis (either from a phrase-based statistical MT or from a neural MT), and post-edited MT (PE). The language pairs represented in this corpus are: English to German (en-de), Latvian (en-lt) and Czech (en-cs), and German to English (de-en).

Example from en-de:

Source: The line in the preview window defines the light direction and angle, and the handles define the edges of the ellipse.

Hypothesis: Die Linie im Vorschaufenster definiert die Lichtrichtung und den Winkel und die Griffe der Kanten der Ellipse zu definieren.

MT Post-Edit: Die Linie im Vorschaufenster definiert die Richtung und den Winkel des Lichts, und die Griffe definieren den Rand der Ellipse.

Reference: Die Linie im Vorschaufenster definiert Lichtrichtung und -winkel, die Griffe definieren die Kanten der Ellipse.

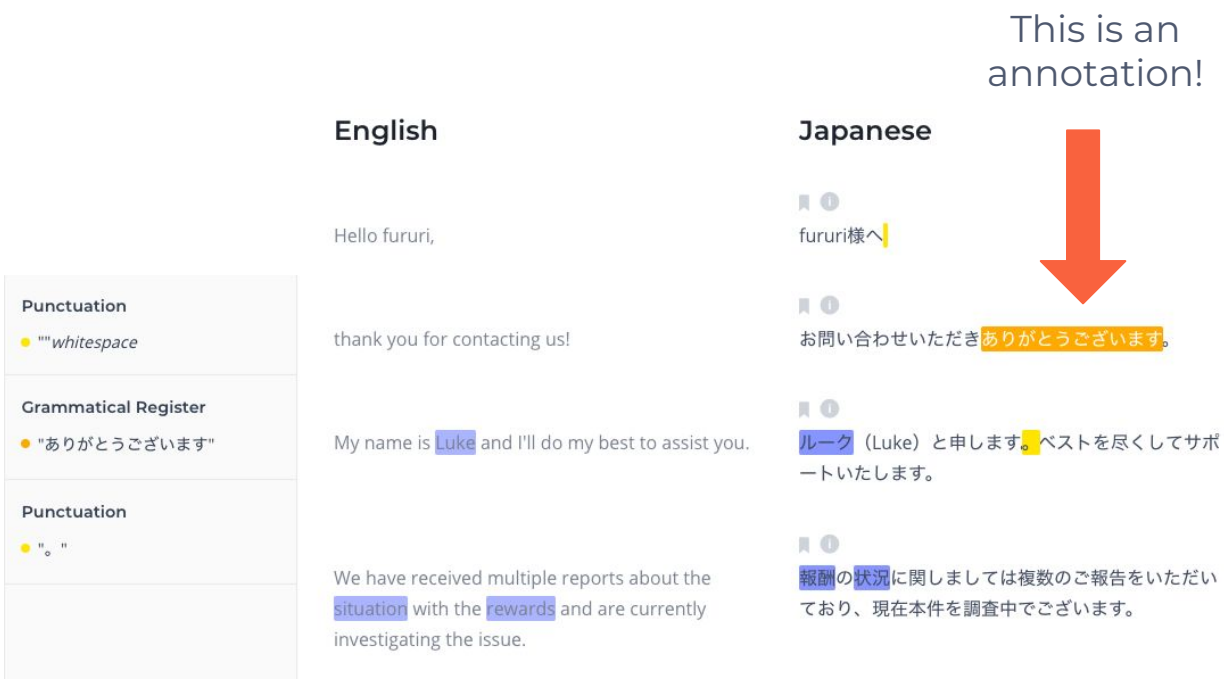
0.4090
HTER

APE-QUEST Corpus (Ive, et al. 2020)

31K tuples with source sentence, respective human-generated reference, MT hypothesis from a neural NMT system, and post-edited MT (PE) from English into three European languages: Dutch, French and Portuguese.

By **concatenating both QT21 with APE-QUEST** we end up with 211k samples covering 7 different European languages. For the WMT2020 Metrics shared task we participated with an HTER metric trained on this concatenated data.

MQM Corpus



Unbabel's:

MQM = 100 -100*SUM(
1*MINORS
+5*MAJORS
+10*CRITICALS
)/Words

Google's:

Severity	Category	Weight
Major	Non-translation	25
	all others	5
Minor	Fluency/Punctuation	0.1
	all others	1
Neutral	all	0

DA Corpus

Every year, since 2008, the WMT News Translation shared task organizers collect human judgements in the form of DAs. Since 2017, due to a lack of annotators, these scores are mapped to relative rankings (DARR). We take advantage of this data in two ways:

1) we use the scores directly in order to train an estimator model,

2) we use the DA relative-ranks to train a translation ranking system.

DA Corpus

Example from fi-en:

Source: Estlander kertoo kyseessä olleen noin 50-vuotias mies.

Reference: Estlander says that the man was close to 50 years of age.

		Human Scores
JUCBNMT:	Estlander people say about 50 years of age.	0
talp-upc:	Estlander says that it was a 50-year-old man.	90
...	...	
online-B:	Estlander tells the man about 50 years old.	50

DA Relative-Ranks

Example from fi-en:

Source:	Estlander kertoo kyseessä olleen noin 50-vuotias mies.
Reference:	Estlander says that the man was close to 50 years of age.
Better Hypothesis:	Estlander says that it was a 50-year-old man.
Worse Hypothesis:	Estlander people say about 50 years of age.

04

Experiments and Results

Evaluation Setup

To evaluate our metrics performance we employ the standard setup from WMT17, 18 and 19. This means that the DA scores are converted into relative-ranks:

Example:

source: 基于上述的定位，可以大胆的设想出未来国产护卫舰的大概模样。

Reference: Based on the above positioning, we can boldly imagine the rough appearance of the new frigates in the future.

Better Hypothesis: Based on the above positioning, you can boldly imagine the future of the domestic frigate general appearance.

Worse Hypothesis: Based on the above localization, can the bold tentative plan the general appearance of future domestic-made escort ship.

How many times a metric agrees in which hypothesis is the “better”?

$$\tau = \frac{Concordant - Discordant}{Concordant + Discordant}$$

WMT19 Metrics DA$_{RR}$ corpus Into-English Results

For the WMT shared task we improved our HTER/MQM models using XLM-R large and more data.

We also trained estimator models to regress directly on the standardized DA scores (instead of using only the relative-ranks). Overall we were able to improve our results.

	de-en	fi-en	gu-en	kk-en	lt-en	ru-en	zh-en	
Nº Tuples	85365	32179	20110	9728	21862	39852	31070	avg.
BLEU	0.054	0.236	0.194	0.276	0.249	0.115	0.321	0.206
CHRF	0.123	0.292	0.240	0.323	0.304	0.177	0.371	0.261
BERTSCORE (F1)	0.191	0.354	0.292	0.351	0.381	0.221	0.433	0.318
BLEURT (large-512)	0.174	**0.374**	0.313	0.372	0.388	0.220	0.436	0.325
PRISM	0.189	0.366	**0.320**	0.362	0.382	0.220	0.434	0.325
COMET-MQM (large)	0.191	0.360	0.289	0.346	0.373	0.213	0.426	0.314
COMET-HTER (large)	0.193	0.351	0.286	0.340	0.375	0.209	0.429	0.312
COMET-DA (large)	**0.220**	0.368	0.316	**0.378**	**0.405**	**0.231**	**0.462**	**0.340**
COMET-RANK (base)	0.202	**0.399**	**0.341**	0.358	**0.407**	0.180	0.445	0.333

(Rei et al, 2020) - https://aclanthology.org/2020.wmt-1.101/

WMT19 Metrics DA$_{RR}$ corpus
Out-of-English Results

	en-cs	en-de	en-fi	en-gu	en-kk	en-lt	en-ru	en-zh	
Nº Tuples	27178	99840	31820	11355	18172	17401	24334	18658	avg.
BLEU	0.364	0.248	0.395	0.463	0.363	0.333	0.4691	0.235	0.410
CHRF	0.444	0.321	0.518	0.548	0.510	0.438	0.548	0.241	0.510
BERTSCORE (F1)	0.486	0.350	0.526	0.559	0.534	0.464	0.581	0.350	0.550
PRISM	0.580	0.416	0.590	-	0.529	0.555	0.581	0.373	0.518
COMET-MQM (large)	0.595	0.405	0.594	0.580	0.546	0.607	0.693	0.400	0.553
COMET-HTER (large)	0.610	0.427	0.610	0.587	0.569	0.615	0.707	0.405	0.566
COMET-DA (large)	**0.618**	**0.435**	0.620	**0.617**	0.585	0.619	**0.711**	0.427	0.579
COMET-RANK (base)	0.603	0.427	**0.664**	0.611	**0.693**	**0.665**	0.580	**0.449**	**0.587**

(Rei et al, 2020) - https://aclanthology.org/2020.wmt-1.101/

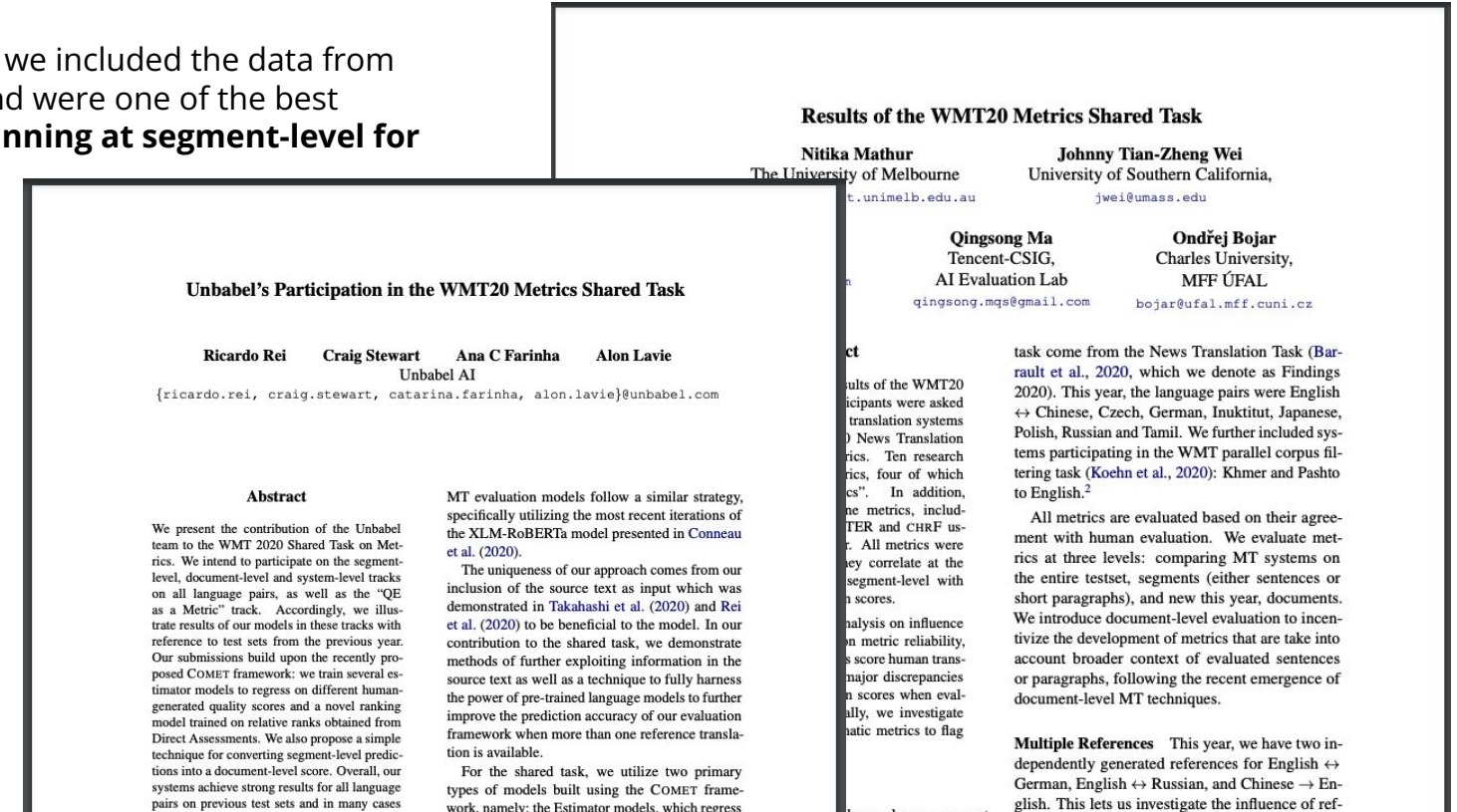
WMT19 Metrics DA$_{RR}$ corpus
Results for LPs not involving English

	de-cs	de-fr	fr-de	
Nº Tuples	23194	4862	1369	avg.
BLEU	0.222	0.226	0.173	0.207
CHRF	0.341	0.287	0.274	0.301
BERTSCORE (F1)	0.356	0.330	0.277	0.321
PRISM	0.452	0.443	**0.421**	0.439
COMET-MQM (large)	0.413	0.422	0.327	0.387
COMET-HTER (large)	0.425	0.449	0.381	0.418
COMET-DA (large)	**0.471**	**0.469**	**0.420**	**0.453**
COMET-RANK (base)	0.389	0.444	0.331	0.388

(Rei et al, 2020) - https://aclanthology.org/2020.wmt-1.101/

Unbabel's Participation in the WMT20 Metrics shared Task

In our final submission we included the data from WMT19 into training and were one of the best performing metrics! **Winning at segment-level for most language pairs!**



Unbabel's Participation in the WMT20 Metrics Shared Task

Ricardo Rei Craig Stewart Ana C Farinha Alon Lavie
Unbabel AI
{ricardo.rei, craig.stewart, catarina.farinha, alon.lavie}@unbabel.com

Abstract

We present the contribution of the Unbabel team to the WMT 2020 Shared Task on Metrics. We intend to participate on the segment-level, document-level and system-level tracks on all language pairs, as well as the "QE as a Metric" track. Accordingly, we illustrate results of our models in these tracks with reference to test sets from the previous year. Our submissions build upon the recently proposed COMET framework: we train several estimator models to regress on different human-generated quality scores and a novel ranking model trained on relative ranks obtained from Direct Assessments. We also propose a simple technique for converting segment-level predictions into a document-level score. Overall, our systems achieve strong results for all language pairs on previous test sets and in many cases

MT evaluation models follow a similar strategy, specifically utilizing the most recent iterations of the XLM-RoBERTa model presented in Conneau et al. (2020).

The uniqueness of our approach comes from our inclusion of the source text as input which was demonstrated in Takahashi et al. (2020) and Rei et al. (2020) to be beneficial to the model. In our contribution to the shared task, we demonstrate methods of further exploiting information in the source text as well as a technique to fully harness the power of pre-trained language models to further improve the prediction accuracy of our evaluation framework when more than one reference translation is available.

For the shared task, we utilize two primary types of models built using the COMET framework, namely: the Estimator models, which regress

Results of the WMT20 Metrics Shared Task

Nitika Mathur
The University of Melbourne
t.unimelb.edu.au

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

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

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

ct

ults of the WMT20
icipants were asked
translation systems
) News Translation
rics. Ten research
rics, four of which
cs". In addition,
ne metrics, includ-
TER and CHRF us-
r. All metrics were
ey correlate at the
segment-level with
n scores.

nalysis on influence
n metric reliability,
s score human trans-
major discrepancies
n scores when eval-
ally, we investigate
natic metrics to flag

task come from the News Translation Task (Barrault et al., 2020, which we denote as Findings 2020). This year, the language pairs were English ↔ Chinese, Czech, German, Inuktitut, Japanese, Polish, Russian and Tamil. We further included systems participating in the WMT parallel corpus filtering task (Koehn et al., 2020): Khmer and Pashto to English.[2]

All metrics are evaluated based on their agreement with human evaluation. We evaluate metrics at three levels: comparing MT systems on the entire testset, segments (either sentences or short paragraphs), and new this year, documents. We introduce document-level evaluation to incentivize the development of metrics that are take into account broader context of evaluated sentences or paragraphs, following the recent emergence of document-level MT techniques.

Multiple References This year, we have two independently generated references for English ↔ German, English ↔ Russian, and Chinese → English. This lets us investigate the influence of ref-

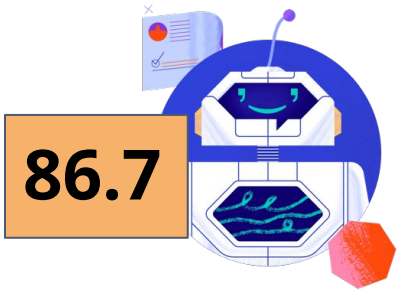
MT-Telescope

More than just a score

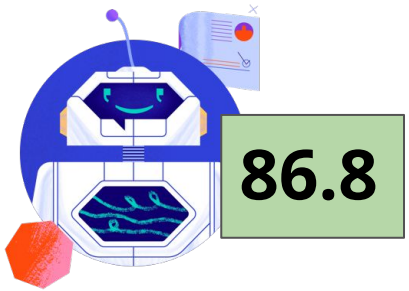
MT Telescope

MT-Telescope is a new, 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?...

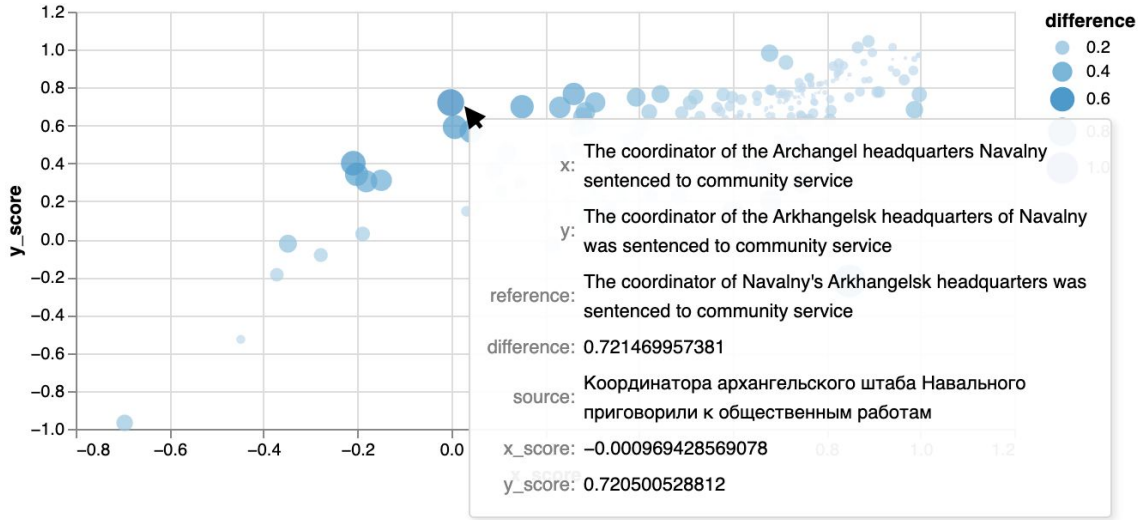


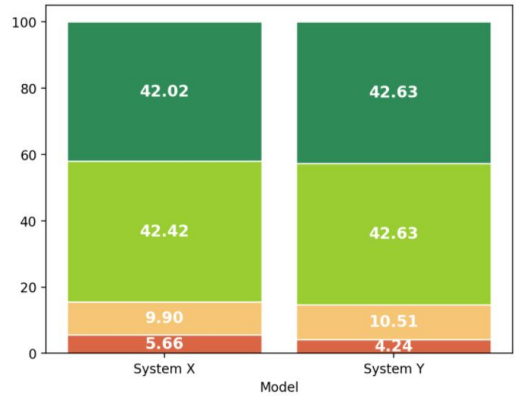
(Rei et al, 2021) - https://aclanthology.org/2021.acl-demo.9/

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.

MT Telescope

Main Features for system-level comparison:

- **SOTA MT evaluation metrics:** COMET (Rei et al. 2020), Prism (Thompson et al, 2020), BLEURT (Sellam et al. 2020), BERTScore (Zhang et al. 2020) along with traditional lexical metric such as BLEU, chrF, TER, etc...
- **Statistical significance** between two systems using **Bootstrap Resampling** (Koehn et al. 2004).
- **Dynamic Corpus Filtering** for named entities, terminology and length.

Ongoing work

Reference-free eval: COMET-QE

COMETinho

Are References Really Needed?

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

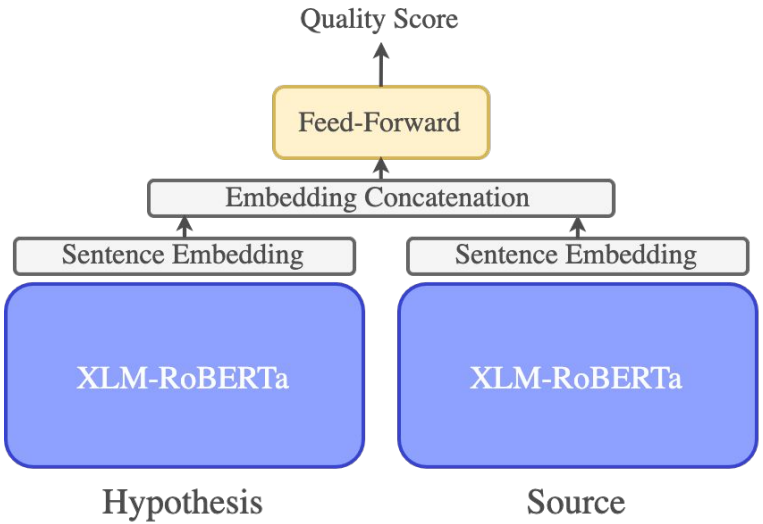
To summarize, we see that the current MT metrics generally struggle to score human translations against machine translations reliably. Rare exceptions include primarily trained neural metrics and reference-less COMET-QE. While the metrics are not really prepared to score human translations, we find this 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

COMET-QE



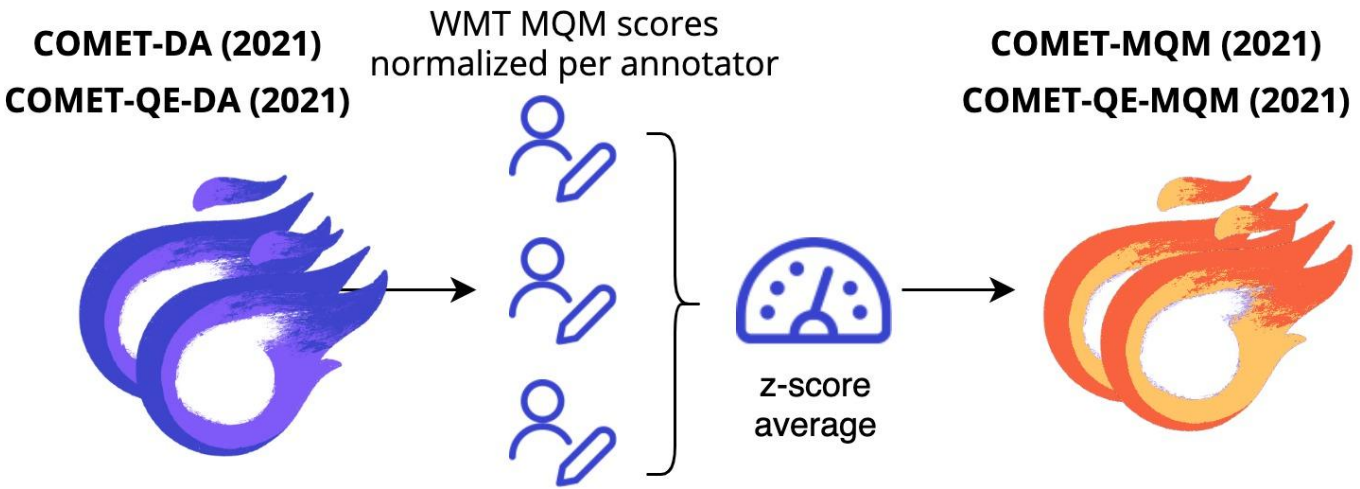
COMETinho

(faster and lighter)

- XLM-R large → **MiniLMv2** (Wang et al., 2020)
- 19x **faster** on CPU (estimate)
- 14.3x **faster** on GPU (estimate)
- Disk footprint: 5x **smaller**

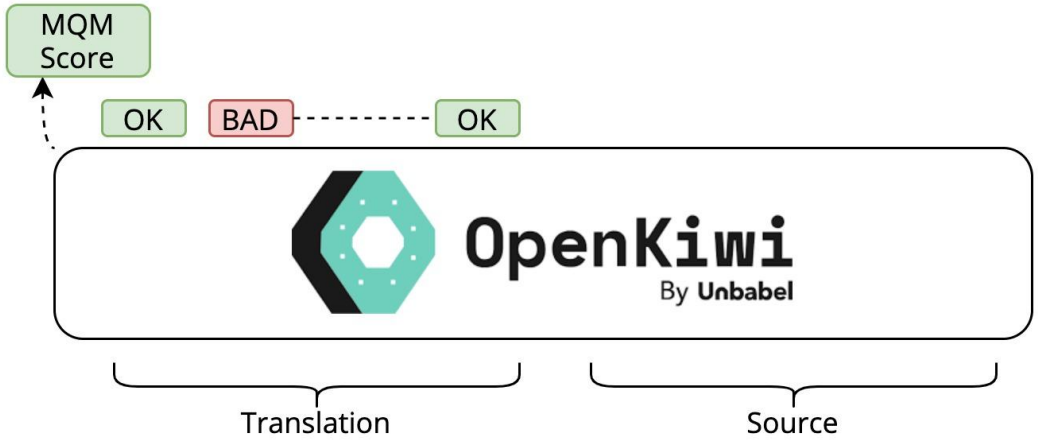
(Rei et al, 2021) - http://statmt.org/wmt21/pdf/2021.wmt-1.111.pdf

COMET-MQM (2021)



(Rei et al, 2021) - http://statmt.org/wmt21/pdf/2021.wmt-1.111.pdf

OpenKiwi-MQM: a reference-free tagging model



Tags	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	OK	BAD	BAD
MT	the	main	purpose	of	this	project	is	to	design	a	car	for	blind	driving.

Source: 这个项目的主要目的 是设计一辆盲人驾驶的车。
Reference: the main goal of this project is to develop a car for the blind.

(Rei et al, 2021) - http://statmt.org/wmt21/pdf/2021.wmt-1.111.pdf

WMT21 MQM devset

	Nº Segments	zh-en 4400		en-de 2950			
		Pearson	Kendall	Pearson	Kendall	Pearson Avg.	Kendall Avg.
Baselines	BLEURT	0.492	0.405	0.107	0.060	0.299	0.232
	PRISM	0.399	0.337	0.072	0.020	0.235	0.178
	BERTSCORE	0.441	0.344	0.116	0.060	0.279	0.202
	BLEU	0.196	0.275	0.062	0.004	0.129	0.140
	CHRF	0.267	0.219	0.119	0.059	0.193	0.139
	COMET-DA (2020)	0.538	0.435	0.425	0.282	0.481	0.359
Ref. based	COMET-DA (2021)	0.559	0.454	0.464	0.309	0.511	0.382
	COMET-MQM (2021)	**0.717**	**0.546**	**0.488**	**0.361**	**0.602**	**0.454**
	COMETINHO-DA	0.484	0.386	0.299	0.204	0.392	0.295
	COMETINHO-MQM	0.670	0.496	0.311	0.237	0.490	0.367
Ref. Free	COMET-QE-DA (2021)	0.567	0.436	**0.497**	0.308	0.532	0.372
	COMET-QE-MQM (2021)	**0.720**	**0.531**	0.470	**0.359**	**0.595**	**0.445**
	OPENKIWI	0.522	0.385	0.448	0.287	0.485	0.336

Table 2: Segment-level correlations on the *en-de* and *zh-en* testset.

(Rei et al, 2021) - http://statmt.org/wmt21/pdf/2021.wmt-1.111.pdf

WMT21 Official Results Summary

Metric	Total "wins"	Language Pair en→de	en→ru	zh→en	Granularity sys	seg	Data condition 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

(Freitag et al, 2021) - http://statmt.org/wmt21/pdf/2021.wmt-1.73.pdf

Ongoing work

Uncertainty-Aware MT Evaluation

Uncertainty-Aware MT Evaluation

We made a good progress in automatic MT Evaluation metrics recently

Yet, metrics such as COMET only provide a single point estimate

In this paper we propose simple ways to get *distribution of scores* -- **confidence interval estimates**

(Glushkovato et al, 2021) - https://aclanthology.org/2021.findings-emnlp.330/

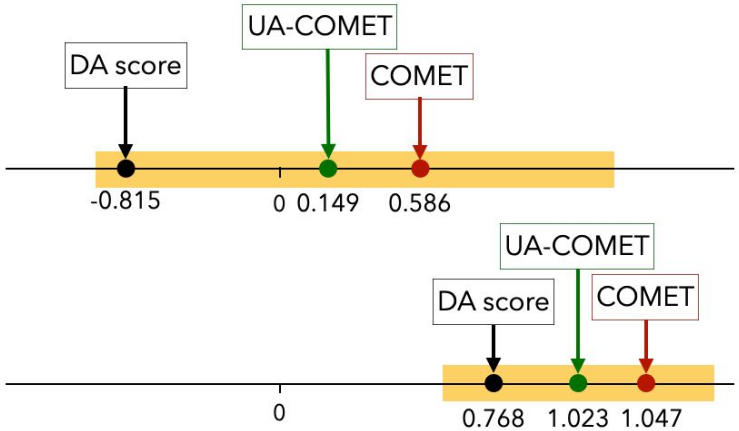
Uncertainty-Aware MT Evaluation

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»



(Glushkovato et al, 2021) - https://aclanthology.org/2021.findings-emnlp.330/

Uncertainty-Aware MT Evaluation

Methods:

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

Experiments:

- Uncertainty-Aware reference-free evaluation
- Impact of reference quality
- Detection of critical errors

(Glushkovato et al, 2021) - https://aclanthology.org/2021.findings-emnlp.330/

Uncertainty-Aware MT Evaluation

- Monte Carlo Dropout and Deep Ensembles show consistent improvements over baseline in all correlation metrics for all LPs
- DE provide more accurate predictions and narrower confidence intervals
- MCD is cheaper and competitive to DE performance

(Glushkovato et al, 2021) - https://aclanthology.org/2021.findings-emnlp.330/

Questions?

Checkout our github:

https://github.com/Unbabel/COMET

https://github.com/Unbabel/MT-Telescope


Unbabel