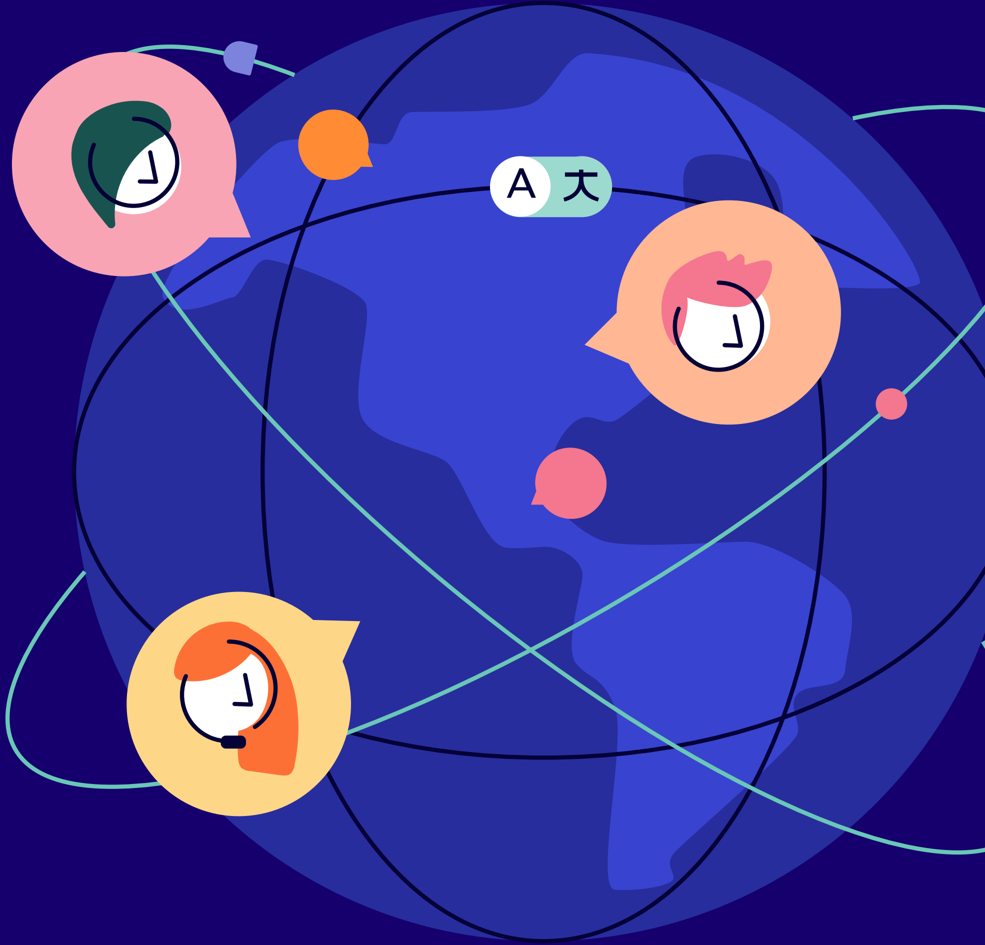
Quality Estimation for Machine Translation

Ricardo Rei

Unbabel AI

September 2022



Unbabel

Amazing Research Team

André Martins

Catarina Farinha

José Souza

João Alves

Alon Lavie

And many other research scientists/engineers split across product teams!

- **We actively collaborate with Instituto Superior Técnico and CMU**

Agenda

01 Definition

02 Models

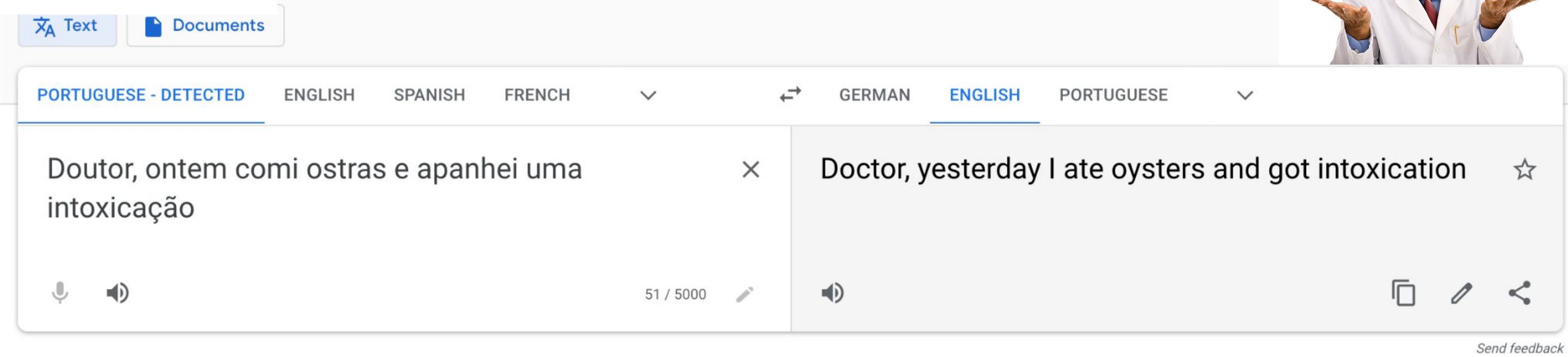
03 WMT Evaluation Shared tasks

04 QE for Decoding

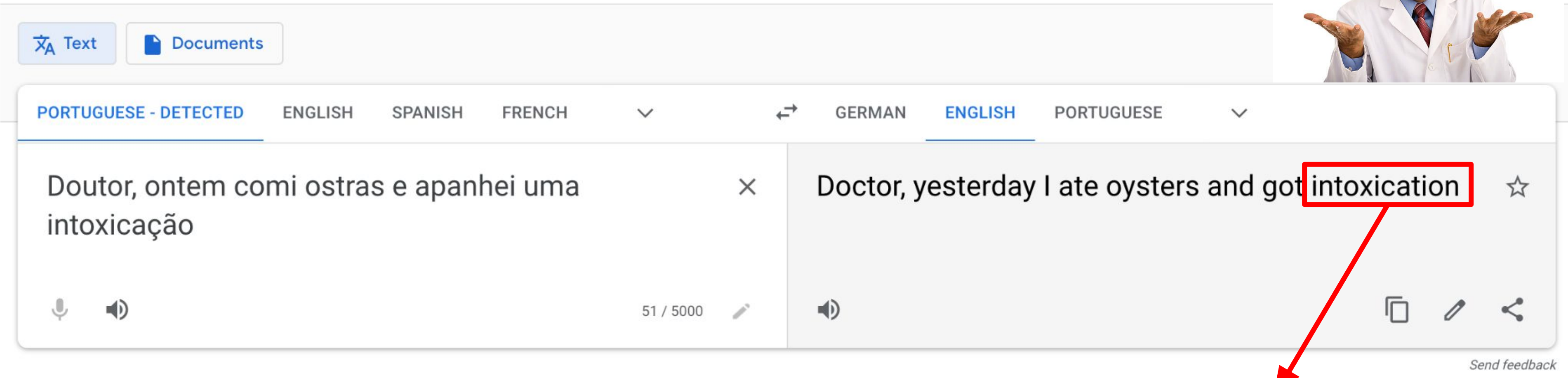
05 Take home messages

Why Quality Estimation?

Is Machine Translation solved?

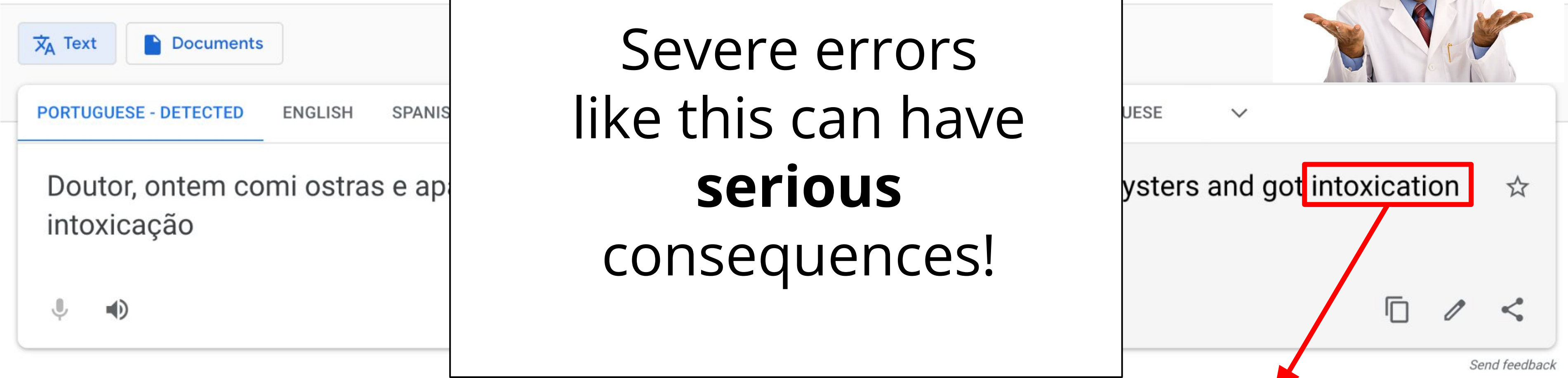


Is Machine Translation solved?



Should be **food poisoning**!

Is Machine Translation solved?



Severe errors like this can have **serious** consequences!

Should be **food poisoning**!

Motivation:

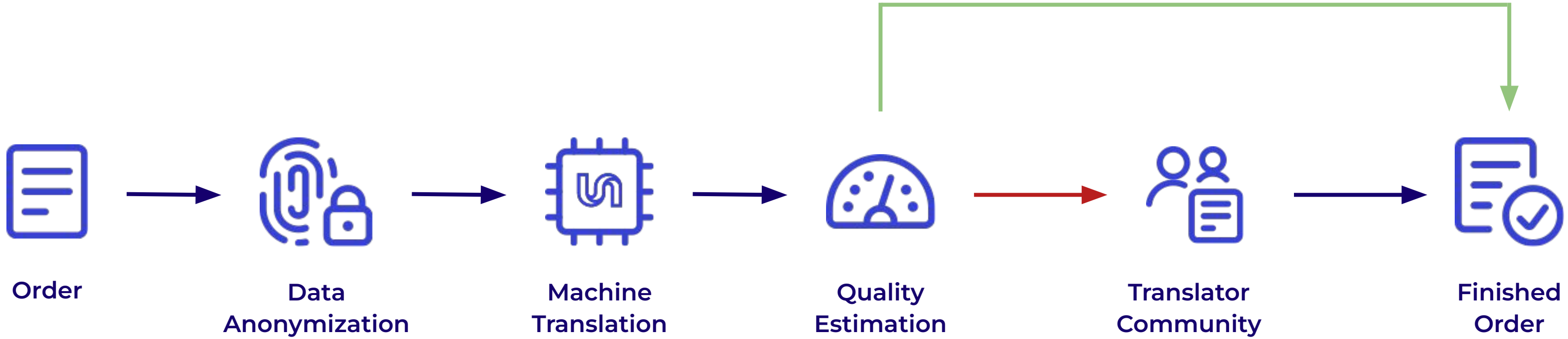
What can we do if we knew the **quality of a translation?**

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)

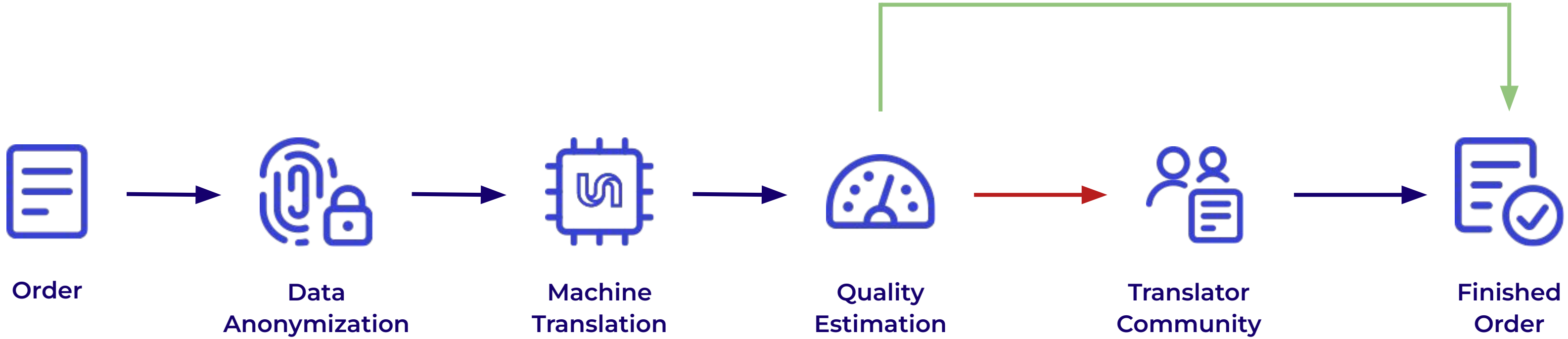
Motivation:

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



Motivation:

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



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

Definition

MT Quality Estimation (QE):

- Use a separate system to estimate **how good a translation is**
  - Typically coming from a **black box MT system**.

- **No access to a reference translation**

- With **different levels of granularity**
  - Word
  - Sentence
  - Document ?

Datasets:

- QE data requires:
  - **SOURCE:** text in the original language
  - **MT:** translation in the target language
  - **Quality assessment** (HTER, MQM or DA)
    - Word level tags (optionally)

- **Source and MT are inputs**

Datasets: Post edit data

"Classical" QE data comes from post-edits:

Sentence-level score

$$\text{HTER} = \frac{\text{edit distance}}{\text{PE words}} = \frac{3}{5} = 0.6$$

Word-level tags OK BAD BAD OK OK BAD

MT: I really like Machine Translation

delete replace insert

PE: I love Machine Translation !

Source: Eu adoro Tradução Automática!

Datasets: Multidimensional Quality Metrics*

Portuguese

Tarde :) Como posso ajudá-lo?

Comprei um monitor cardíaco mas não consegui colocar em funcionamento.

Já atualizei o sistema e tetei colocar a recarregar, mas parece que não carrega.

English

Afternoon :) How may I help you?

I bought a heart monitor but I couldn't get it up and running

Already updated the system and tetetei to recharge, but it does not charge.

Missing Punctuation Untranslated "tetetei" Omitted Pronoun

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

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

Datasets: Multidimensional Quality Metrics*

													MAJOR	
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.

We ask annotators to highlight errors according to an internal error typology (for aspects such as ‘lexical’, ‘fluency’ and ‘register’) and rank the error severity as 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’.

Datasets: Multidimensional Quality Metrics*

													MAJOR	
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.

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.

Datasets: Direct Assessments

Direct Assessments are only used for **sentence level evaluation**.

Example:

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

Models

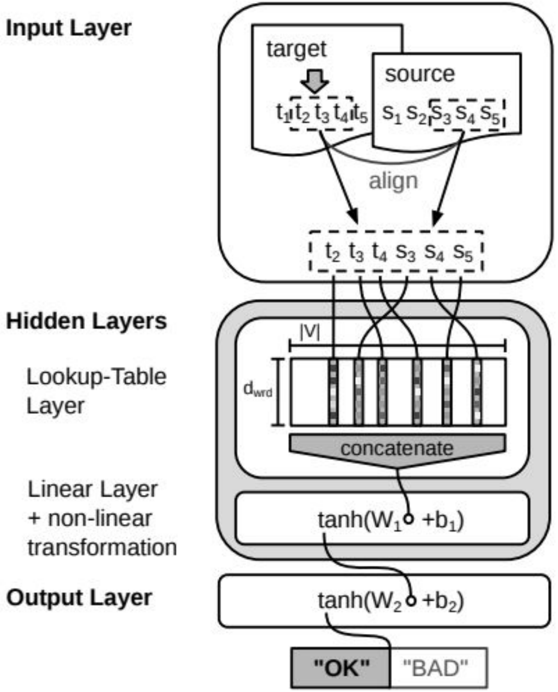
QUETCH: QUality Estimation from scraTCH*

First neural model for QE

Very simple architecture

Source embeddings are aligned and concatenated to MT embeddings

Only works for word-level.



* [QUality Estimation from ScraTCH (QUETCH): Deep Learning for Word-level Translation Quality Estimation] (Kreutzer et al., 2015)

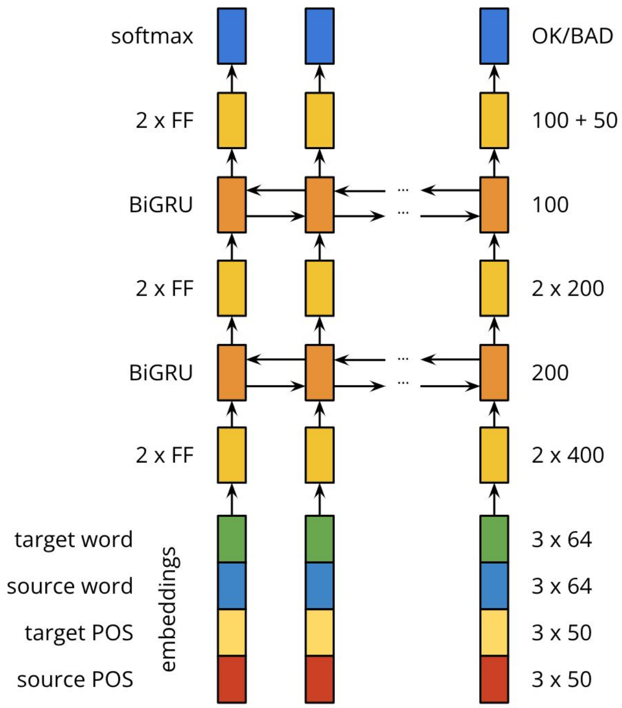
NuQE: Neural Quality Estimation*

Deeper version of QUETCH using recurrent layers

Source embeddings are aligned and concatenated to MT embeddings

Uses POS tags as input

First used in Unbabel's winning participation in WMT16



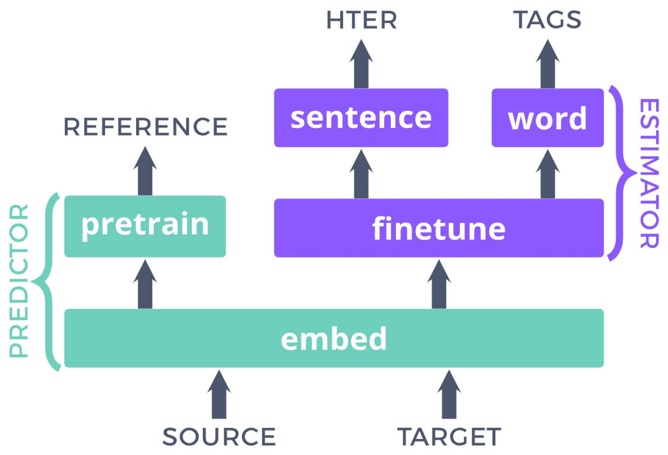
* Unbabel's Participation in the WMT16 Word-Level Translation Quality Estimation Shared Task (Martins et al., 2016)

Predictor-Estimator

Uses a two-stage neural model that is pretrained with large parallel data

- Deep contextualized language model pretraining

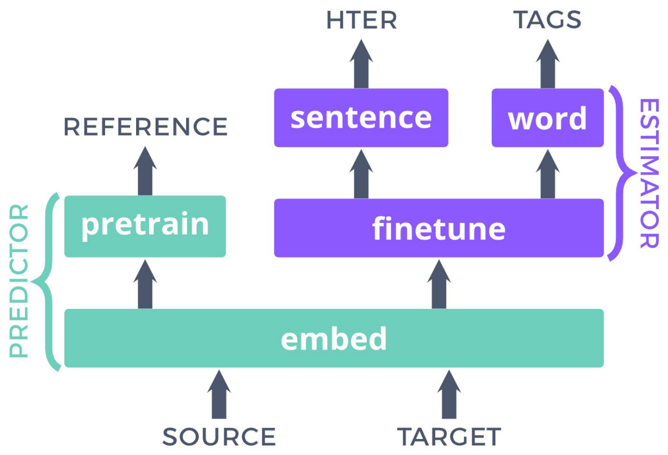
- 1 year ahead of muppet models!



* Predictor-Estimator using Multilevel Task Learning with Stack Propagation for Neural Quality Estimation (Kim et al., 2017)

Predictor-Estimator

The **predictor** is trained to predict every token of the **TARGET side given its left and right context** produced by two uni-directional LSTM's

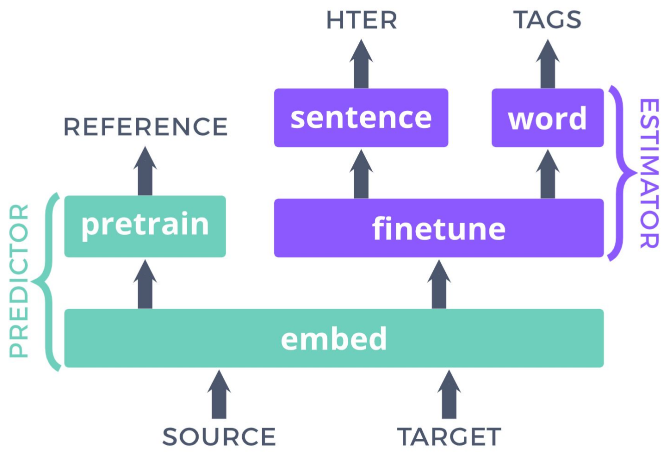


* Predictor-Estimator using Multilevel Task Learning with Stack Propagation for Neural Quality Estimation (Kim et al., 2017)

Predictor-Estimator

The **predictor** is trained to predict every token of the **TARGET side given its left and right context** produced by two uni-directional LSTM's

The **estimator** is fine-tuned to predict sentence scores and word-level tags.



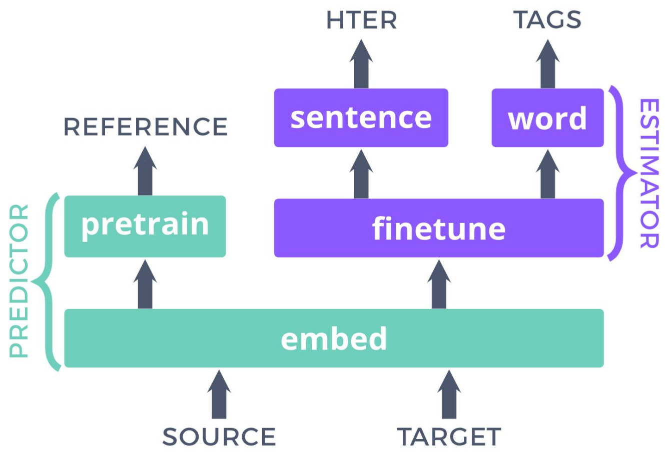
* Predictor-Estimator using Multilevel Task Learning with Stack Propagation for Neural Quality Estimation (Kim et al., 2017)

Transformer Predictor-Estimator

The **predictor** is trained to predict every token of the TARGET side given its **Bidirectional context** produced by a pretrained transformer (e.g. BERT)

The **estimator** is fine-tuned to predict sentence scores and word-level tags.

Unbabel's winning participation in WMT19

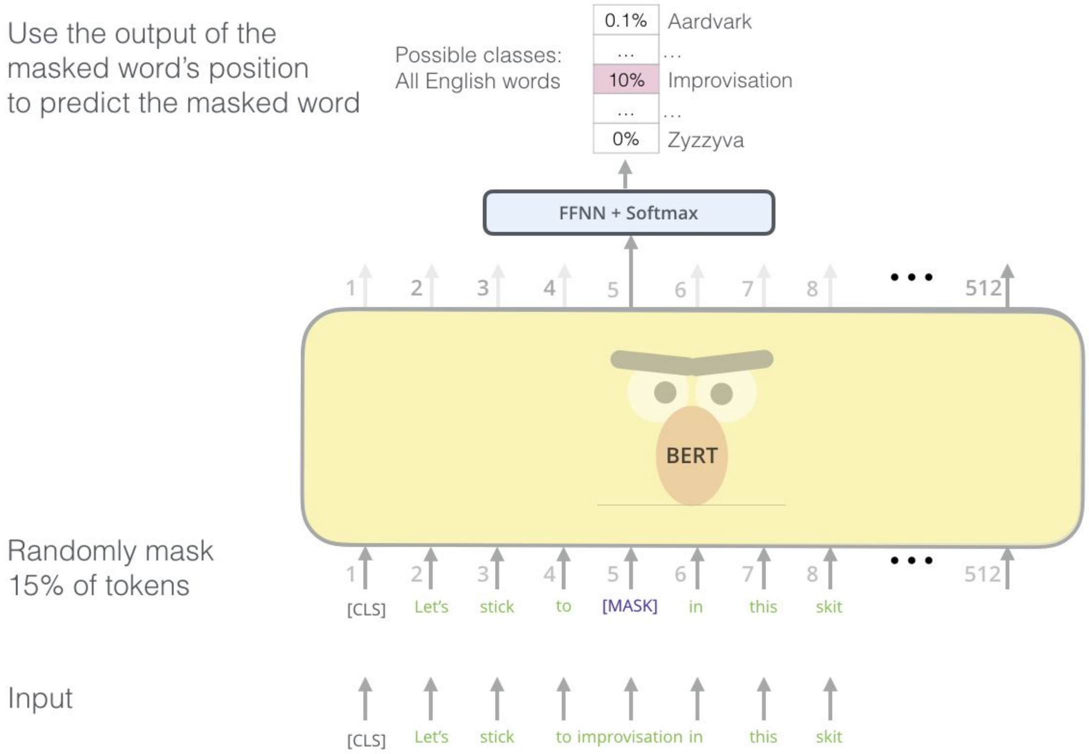


* OpenKiwi: An Open Source Framework for Quality Estimation (Kepler et al., ACL 2019)

* TransQuest: Translation Quality Estimation with Cross-lingual Transformers (Ranasinghe et al., COLING 2020)

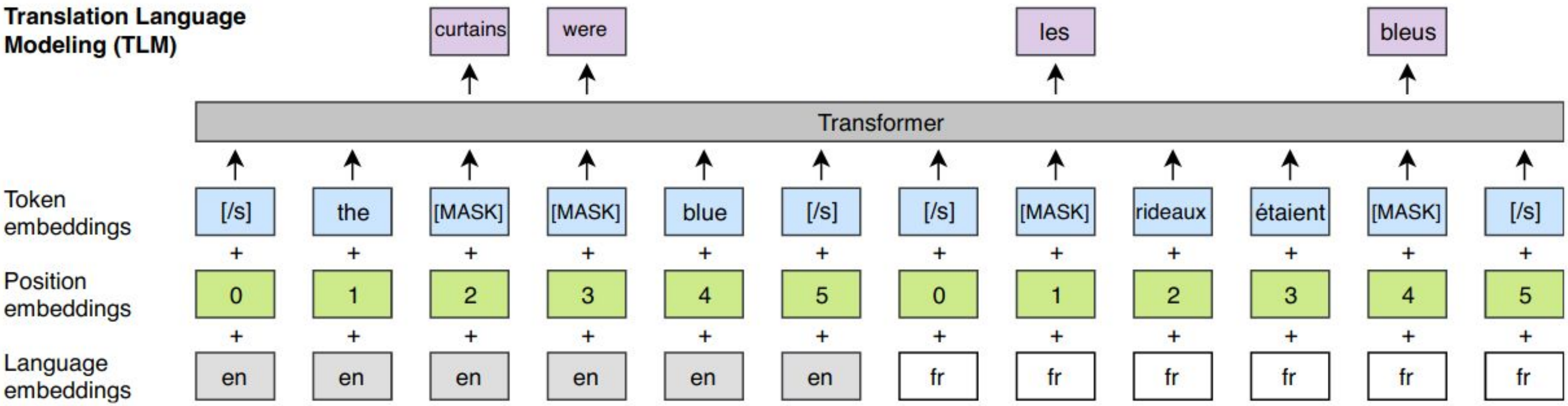
We will release this architecture also in COMET

Predictor: BERT & XLM-R

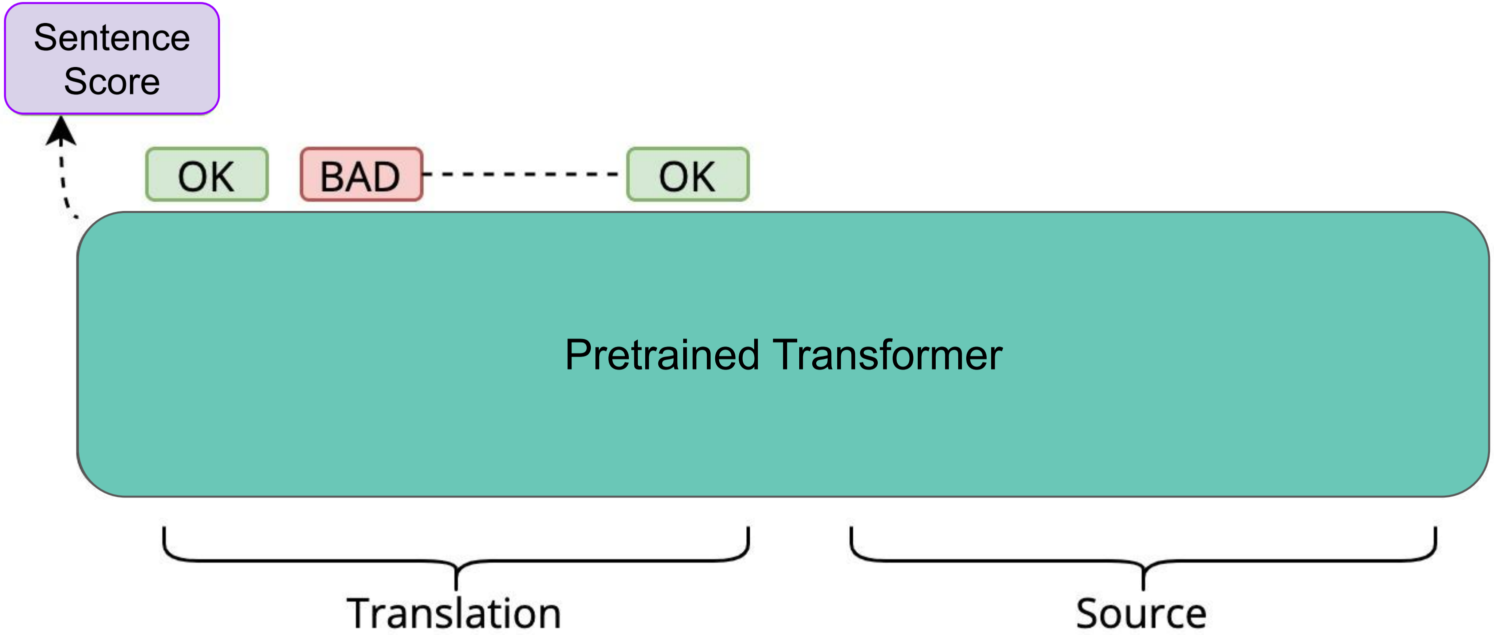


Source: The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning), Jay Alammar, 2019.

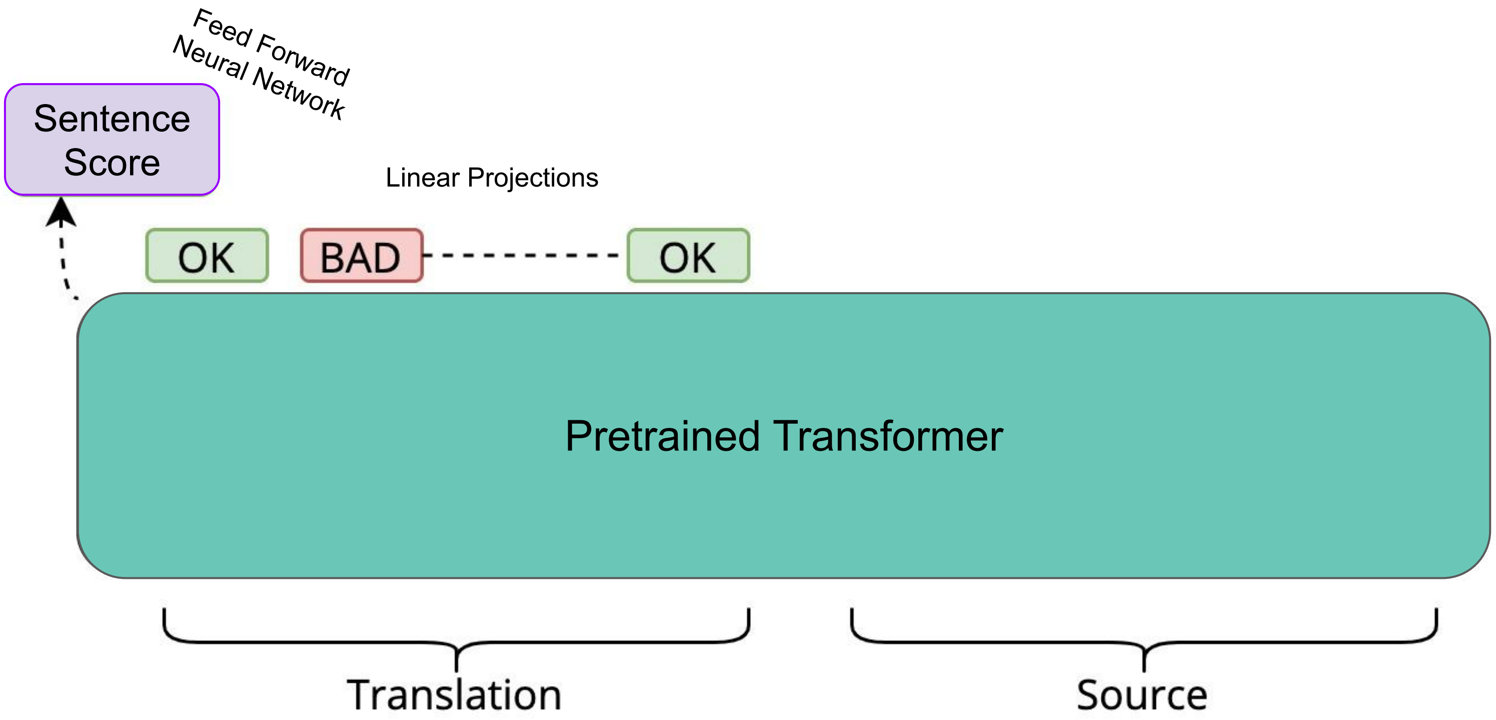
Predictor: XLM & InfoXLM



Estimator:

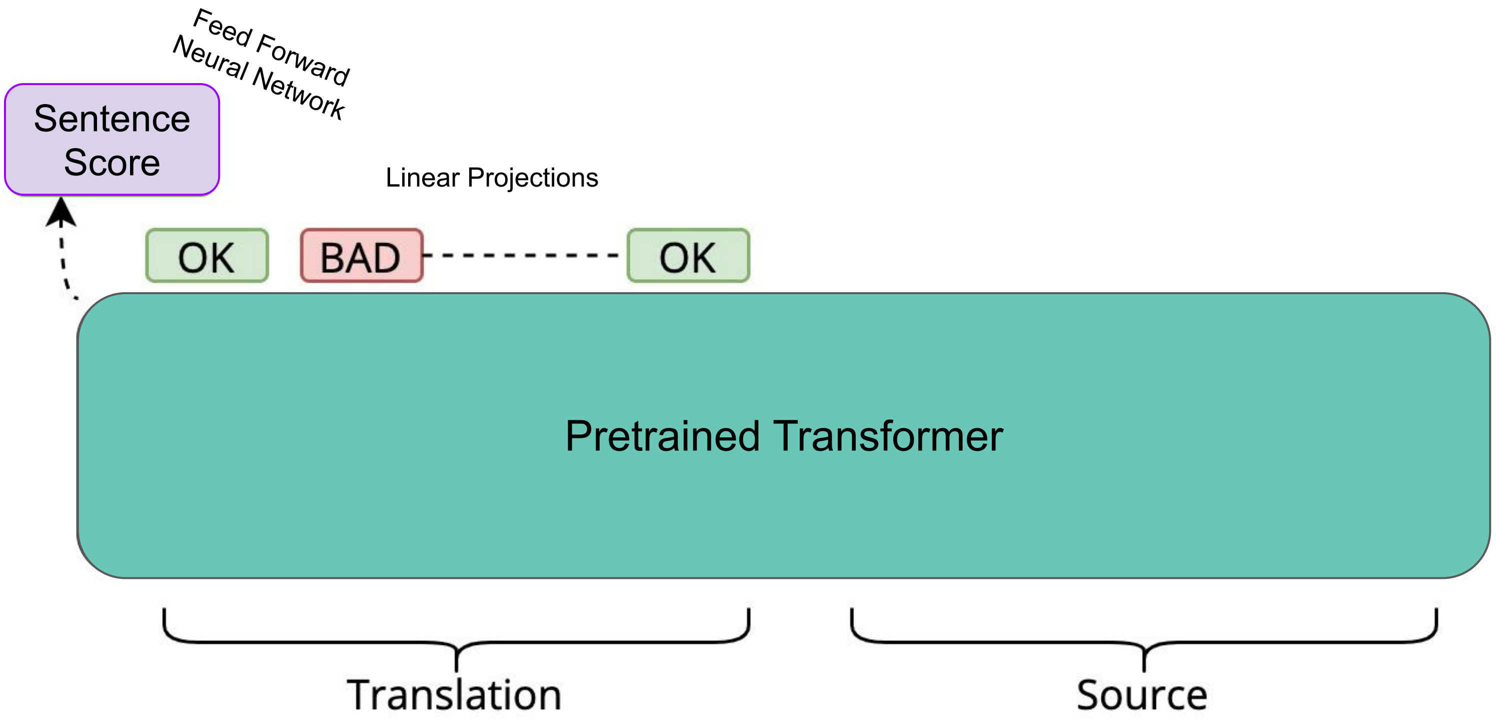

Sentence Score
OK
BAD
OK
Pretrained Transformer
Translation
Source

Estimator:



Feed Forward Neural Network
Sentence Score
Linear Projections
OK
BAD
OK
Pretrained Transformer
Translation
Source

Estimator:


Feed Forward
Neural Network
Sentence Score
Linear Projections
OK
BAD
OK
Pretrained Transformer
Translation
Source

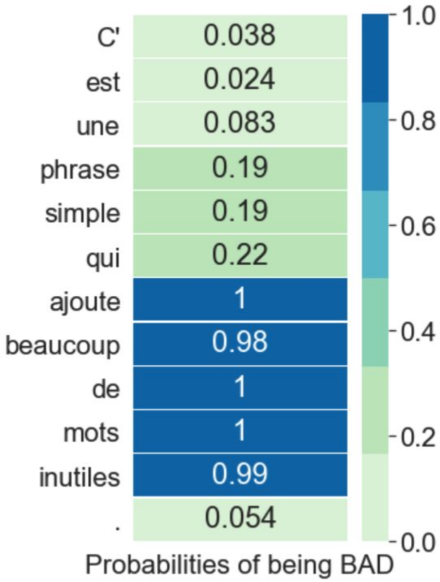
Example:

Source

This is a simple sentence .

MT

C' est une phrase simple qui ajoute beaucoup de mots inutiles .



'sentence_scores': [0.5956864953041077]

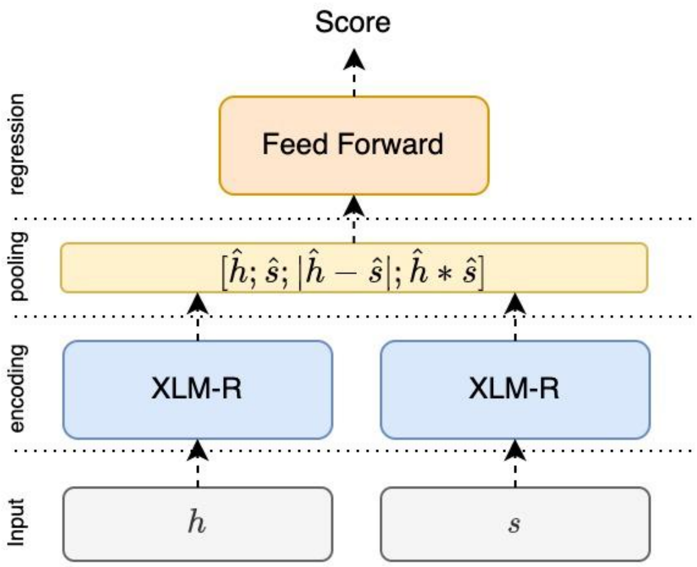
['OK', 'OK', 'OK', 'OK', 'OK', 'OK', 'BAD', 'BAD', 'BAD', 'BAD', 'BAD', 'OK']

MACHINE_TRANSLATION: C' est une phrase simple qui ajoute beaucoup de mots inutiles .

COMET-QE Dual Encoder

COMET* was initially developed for MT evaluation with metric but it has showed promising results in QE

- Sentence Embeddings are created by **Avg. Pooling**
- Along with source and target embeddings we extract the **element-wise difference and dot-product between embeddings**.
- A feed forward is used to predict a quality assessment (MQM or DA)



* COMET: A Neural Framework for MT Evaluation (Rei et al., EMNLP 2020)

Workshop on Machine Translation Evaluation Shared Tasks

Quality Estimation is becoming competitive with 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)

WMT 2022 QE Task:

This year shared task was divided into 3 subtasks:

1) **Quality Prediction**
   a) Sentence-level (DA + MQM)
   b) Word-level (Post edit + MQM tags)

2) **Explainable QE**
   a) DA + MQM explanations

3) **Critical Error Detection**

WMT 2022 QE Task:

This year shared task was divided into 3 subtasks:

1) **Quality Prediction**
   a) Sentence-level (DA + MQM)
   b) Word-level (Post edit + MQM tags)

2) **Explainable QE**
   a) DA + MQM explanations

3) **Critical Error Detection**

WMT 2022 QE Task: Unbabel-IST Submission

Main Challenges:

1) Our systems need to **generalize well to different types of annotations**

2) Our systems have to **generalize for languages for which we have little or no training data**

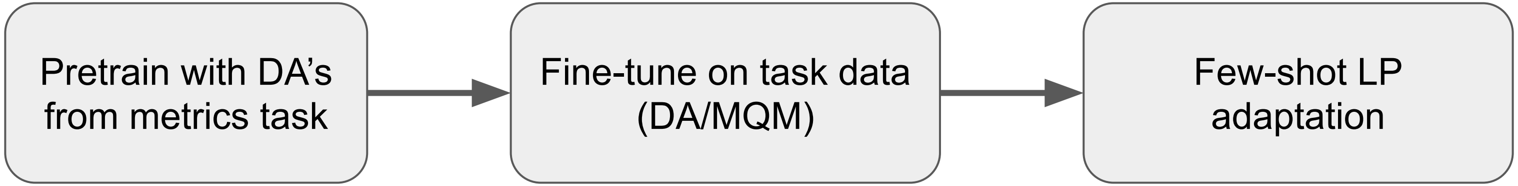
Our submission:

1) We take advantage of the training features from COMET to build models that generalize well!

2) We extend COMET with a **predictor-estimator architecture**

3) We focus on **multilingual models** and we adapt them to **new language pairs with just a few sentences**

WMT 2022 QE Task: Unbabel-IST Submission



WMT 2022 QE Task: Unbabel-IST Submission

Encoder	km-en	ps-en	en-ja	en-cs	en-mr	ru-en	ro-en	en-zh	en-de	et-en	si-en	ne-en	avg.
	Direct Assessments												
	Baseline (Zerva et al., 2021)												
XLM-R	0.615	0.601	0.295	0.535	0.419	0.703	0.828	0.513	0.500	0.806	0.565	0.793	0.598
	Pretrained models												
InfoXLM	0.619	0.603	0.328	0.510	0.462	0.731	0.829	0.554	0.516	0.803	0.561	0.777	0.608
RemBERT	0.600	0.621	0.338	0.525	0.447	0.680	0.818	0.487	0.491	0.810	0.525	0.747	0.591
XLM-R	0.610	0.579	0.325	0.503	0.405	0.715	0.832	0.541	0.514	0.782	0.540	0.740	0.591
	Sentence-level only												
XLM-R	0.628	0.591	0.350	0.531	0.551	0.761	0.859	0.577	0.568	0.800	0.565	0.796	0.631
InfoXLM	0.629	0.623	0.348	0.515	0.574	0.747	0.858	0.586	0.551	0.828	0.568	0.790	0.635
RemBERT	0.633	0.629	0.356	0.565	0.575	0.762	0.854	0.558	0.528	0.833	0.570	0.796	0.638
	Few-shot Language Adaptation												
XLM-R	0.650	0.619	0.352	0.551	0.546	0.753	0.852	0.571	0.554	0.813	0.562	0.798	0.635
InfoXLM	0.641	0.650	0.367	0.549	0.549	0.751	0.855	0.591	0.565	0.824	0.563	0.803	0.642
RemBERT	0.644	0.645	0.356	0.567	0.568	0.759	0.856	0.545	0.552	0.835	0.561	0.804	0.641
	Sentence + word-level training												
InfoXLM	0.617	0.586	0.344	0.532	0.572	0.761	0.865	0.586	0.579	0.829	0.576	0.804	0.637
RemBERT	0.634	0.628	0.356	0.564	0.571	0.762	0.860	0.541	0.553	0.826	0.564	0.799	0.638
	Few-shot Language Adaptation												
InfoXLM	0.643	0.632	0.335	0.557	0.560	0.766	0.860	0.575	0.582	0.833	0.578	0.809	0.644
RemBERT	0.644	0.645	0.356	0.567	0.568	0.759	0.856	0.545	0.552	0.835	0.561	0.804	0.641
	Final Ensemble												
Ensemble 6x	0.664	0.669	0.380	0.591	0.593	0.782	0.871	0.597	0.593	0.845	0.588	0.820	0.666

Table 1: Results for sentence-level QE in terms of Spearman correlation for DA.

WMT 2022 QE Task: Unbabel-IST Submission

	Direct Assessments												
Encoder	**km-en**	**ps-en**	**en-ja**	**en-cs**	**en-mr**	**ru-en**	**ro-en**	**en-zh**	**en-de**	**et-en**	**si-en**	**ne-en**	**avg.**
Baseline (Zerva et al., 2021)													
XLM-R	0.615	0.601	0.295	0.535	0.419	0.703	0.828	0.513	0.500	0.806	0.565	0.793	0.598
Pretrained models													
InfoXLM	0.619	0.603	0.328	0.510	0.462	0.731	0.829	0.554	0.516	0.803	0.561	0.777	0.608
RemBERT	0.600	0.621	0.338	0.525	0.447	0.680	0.818	0.487	0.491	0.810	0.525	0.747	0.591
XLM-R	0.610	0.579	0.325	0.503	0.405	0.715	0.832	0.541	0.514	0.782	0.540	0.740	0.591
Sentence-level only													
XLM-R	0.628	0.591	0.350	0.531	0.551	0.761	0.859	0.577	0.568	0.800	0.565	0.796	0.631
InfoXLM	0.629	0.623	0.348	0.515	0.574	0.747	0.858	0.586	0.551	0.828	0.568	0.790	0.635
RemBERT	0.633	0.629	0.356	0.565	0.575	0.762	0.854	0.558	0.528	0.833	0.570	0.796	0.638
Few-shot Language Adaptation													
XLM-R	0.650	0.619	0.352	0.551	0.546	0.753	0.852	0.571	0.554	0.813	0.562	0.798	0.635
InfoXLM	0.641	0.650	0.367	0.549	0.549	0.751	0.855	0.591	0.565	0.824	0.563	0.803	0.642
RemBERT	0.644	0.645	0.356	0.567	0.568	0.759	0.856	0.545	0.552	0.835	0.561	0.804	0.641
Sentence + word-level training													
InfoXLM	0.617	0.586	0.344	0.532	0.572	0.761	0.865	0.586	0.579	0.829	0.576	0.804	0.637
RemBERT	0.634	0.628	0.356	0.564	0.571	0.762	0.860	0.541	0.553	0.826	0.564	0.799	0.638
Few-shot Language Adaptation													
InfoXLM	0.643	0.632	0.335	0.557	0.560	0.766	0.860	0.575	0.582	0.833	0.578	0.809	0.644
RemBERT	0.644	0.645	0.356	0.567	0.568	0.759	0.856	0.545	0.552	0.835	0.561	0.804	0.641
Final Ensemble													
Ensemble 6x	0.664	0.669	0.380	0.591	0.593	0.782	0.871	0.597	0.593	0.845	0.588	0.820	0.666

Table 1: Results for sentence-level QE in terms of Spearman correlation for DA.

WMT 2022 QE Task: Unbabel-IST Submission

Encoder	km-en	ps-en	en-ja	en-cs	en-mr	ru-en	ro-en	en-zh	en-de	et-en	si-en	ne-en	avg.
							Direct Assessments						
							Baseline (Zerva et al., 2021)						
XLM-R	0.615	0.601	0.295	0.535	0.419	0.703	0.828	0.513	0.500	0.806	0.565	0.793	0.598
							Pretrained models						
InfoXLM	0.619	0.603	0.328	0.510	0.462	0.731	0.829	0.554	0.516	0.803	0.561	0.777	0.608
RemBERT	0.600	0.621	0.338	0.525	0.447	0.680	0.818	0.487	0.491	0.810	0.525	0.747	0.591
XLM-R	0.610	0.579	0.325	0.503	0.405	0.715	0.832	0.541	0.514	0.782	0.540	0.740	0.591
							Sentence-level only						
XLM-R	0.628	0.591	0.350	0.531	0.551	0.761	0.859	0.577	0.568	0.800	0.565	0.796	0.631
InfoXLM	0.629	0.623	0.348	0.515	0.574	0.747	0.858	0.586	0.551	0.828	0.568	0.790	0.635
RemBERT	0.633	0.629	0.356	0.565	0.575	0.762	0.854	0.558	0.528	0.833	0.570	0.796	0.638
							Few-shot Language Adaptation						
XLM-R	0.650	0.619	0.352	0.551	0.546	0.753	0.852	0.571	0.554	0.813	0.562	0.798	0.635
InfoXLM	0.641	0.650	0.367	0.549	0.549	0.751	0.855	0.591	0.565	0.824	0.563	0.803	0.642
RemBERT	0.644	0.645	0.356	0.567	0.568	0.759	0.856	0.545	0.552	0.835	0.561	0.804	0.641
							Sentence + word-level training						
InfoXLM	0.617	0.586	0.344	0.532	0.572	0.761	0.865	0.586	0.579	0.829	0.576	0.804	0.637
RemBERT	0.634	0.628	0.356	0.564	0.571	0.762	0.860	0.541	0.553	0.826	0.564	0.799	0.638
							Few-shot Language Adaptation						
InfoXLM	0.643	0.632	0.335	0.557	0.560	0.766	0.860	0.575	0.582	0.833	0.578	0.809	0.644
RemBERT	0.644	0.645	0.356	0.567	0.568	0.759	0.856	0.545	0.552	0.835	0.561	0.804	0.641
							Final Ensemble						
Ensemble 6x	0.664	0.669	0.380	0.591	0.593	0.782	0.871	0.597	0.593	0.845	0.588	0.820	0.666

Table 1: Results for sentence-level QE in terms of Spearman correlation for DA.

WMT 2022 QE Task:

This year shared task was divided into 3 subtasks:

1) **Quality Prediction**
   a) Sentence-level (DA + MQM)
   b) Word-level (Post edit + MQM tags)

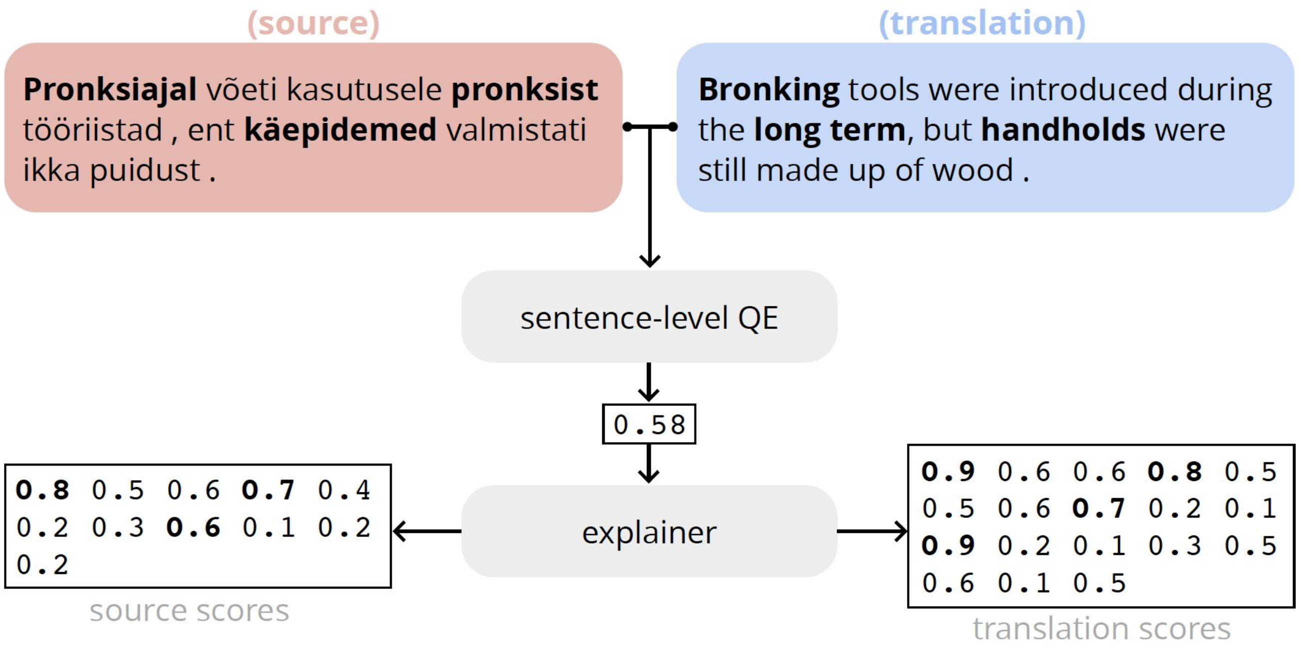
2) **Explainable QE**
   a) DA + MQM explanations

3) **Critical Error Detection**

WMT 2022 QE Task: Unbabel-IST Submission

Explainable QE shared task objective:

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



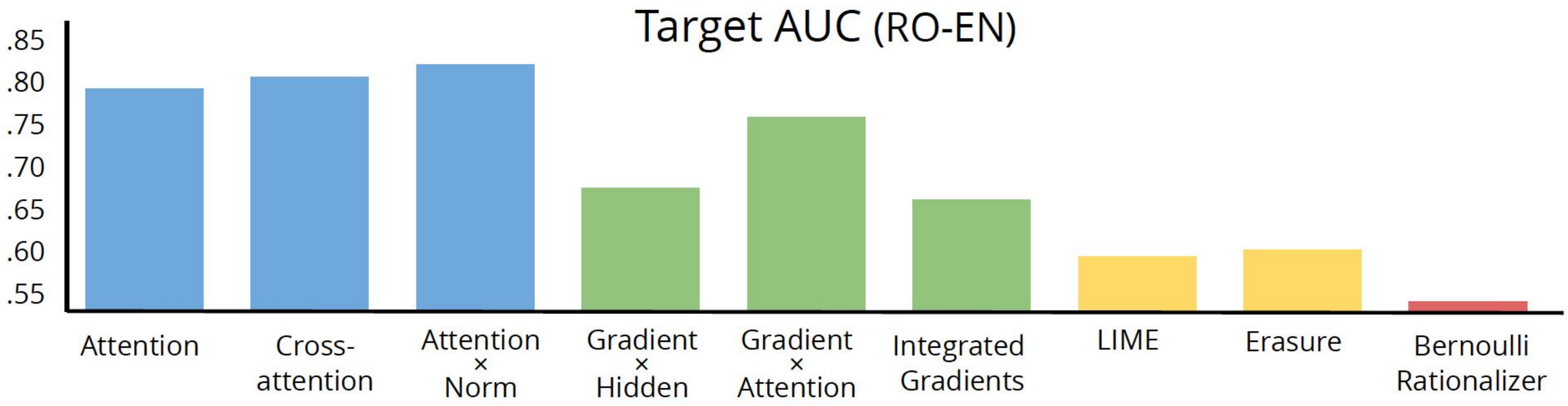
WMT 2022 QE Task: Unbabel-IST Submission

• **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).

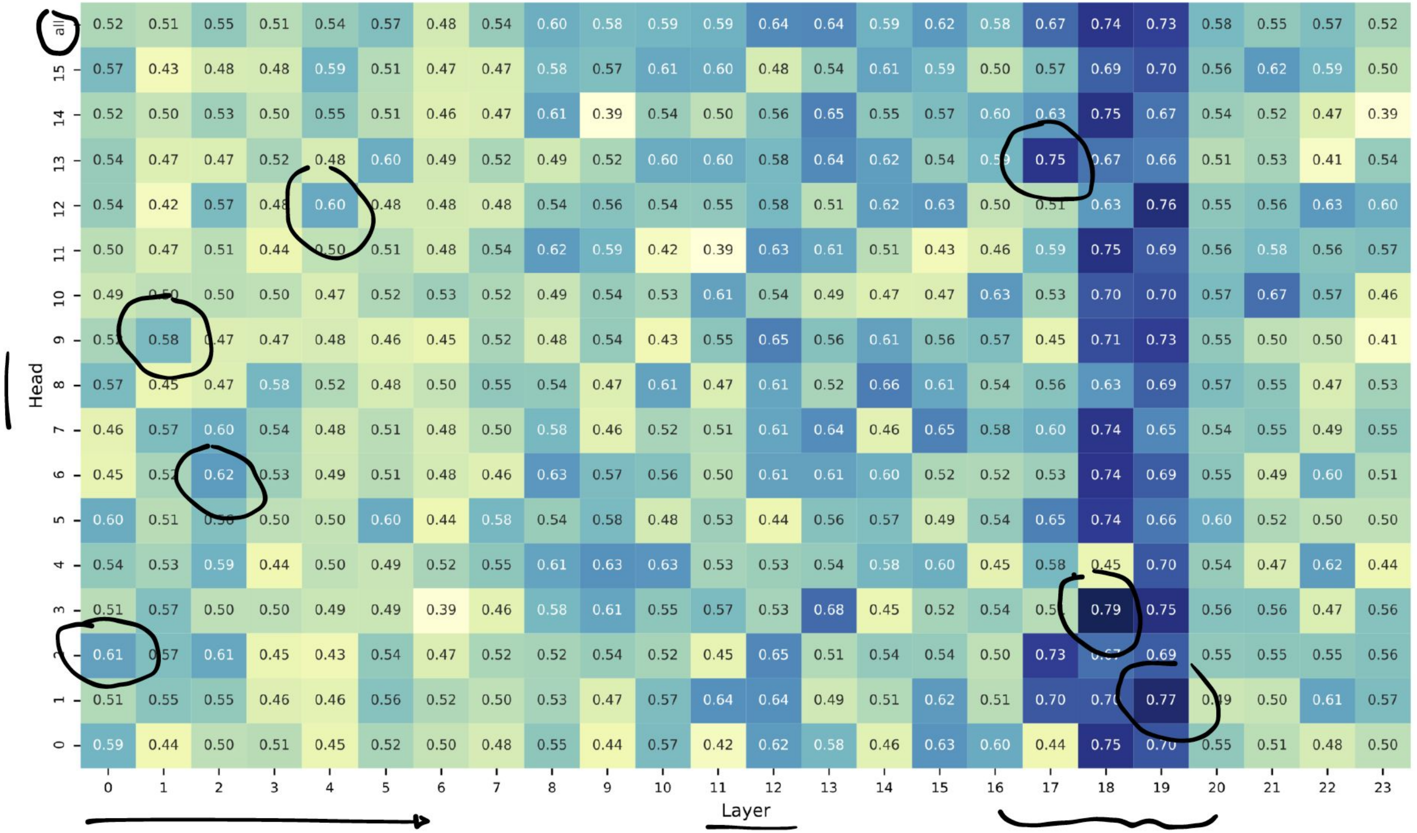
WMT 2022 QE Task: Unbabel-IST Submission

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)

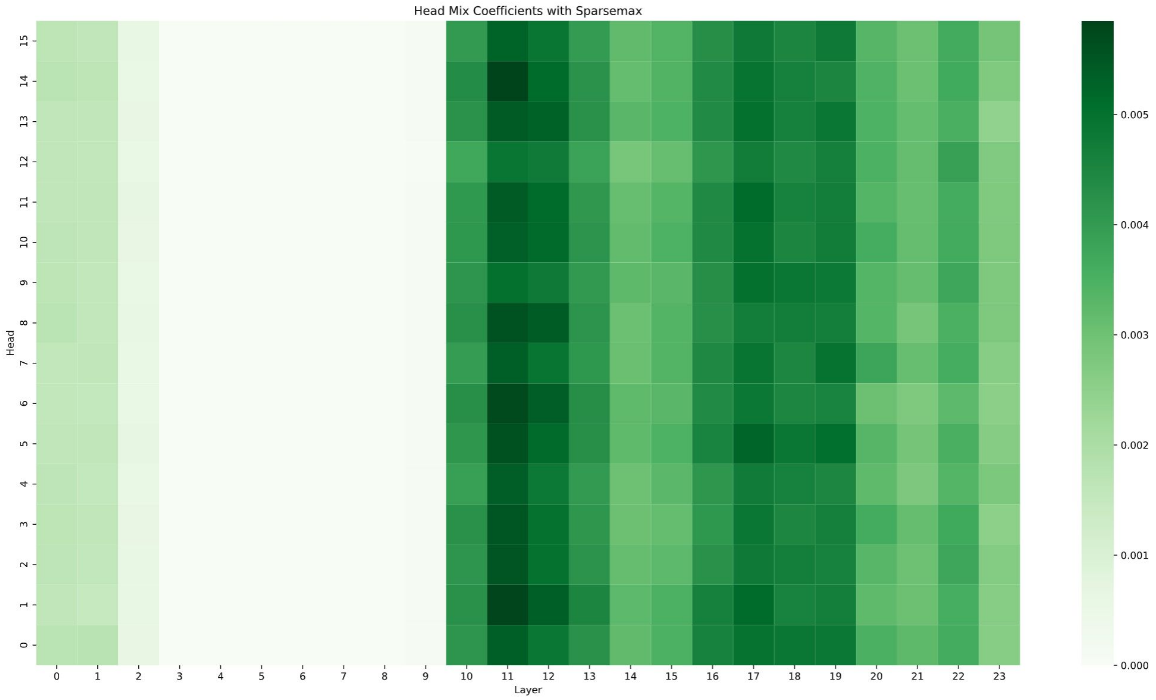
WMT 2022 QE Task: Unbabel-IST Submission

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.



* We are still writing the system submission paper. TBA: WMT 2022

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

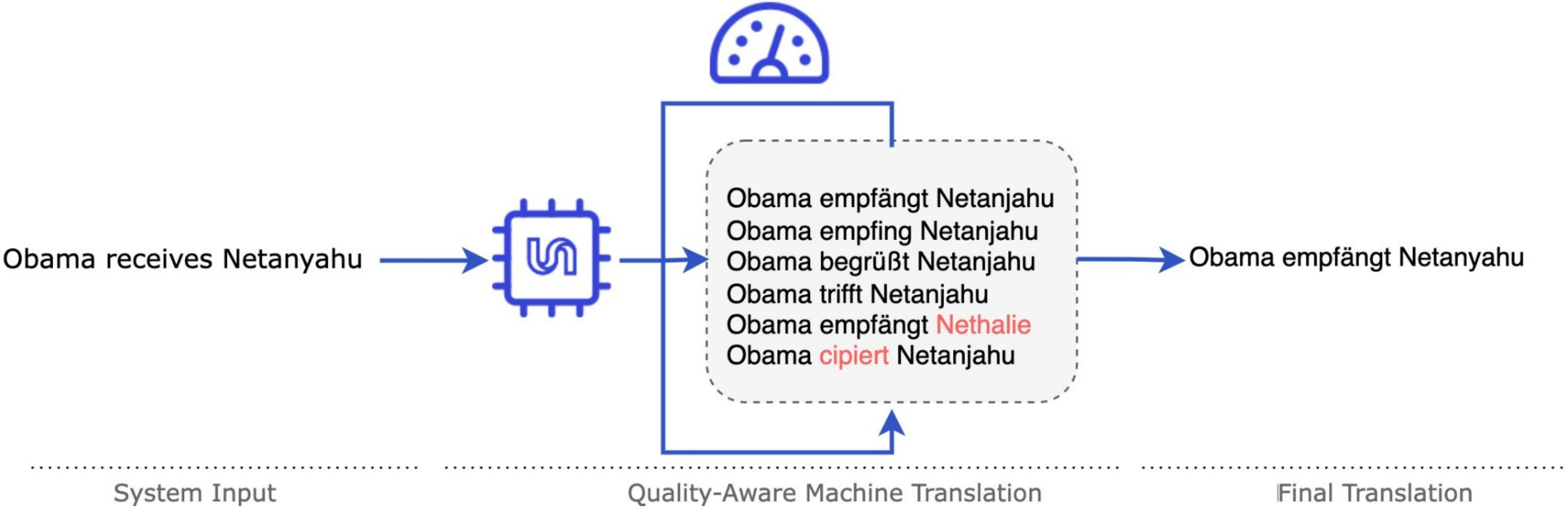
Official results: https://www.statmt.org/wmt22/quality-estimation-task_results.html

	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.

Quality Aware Decoding

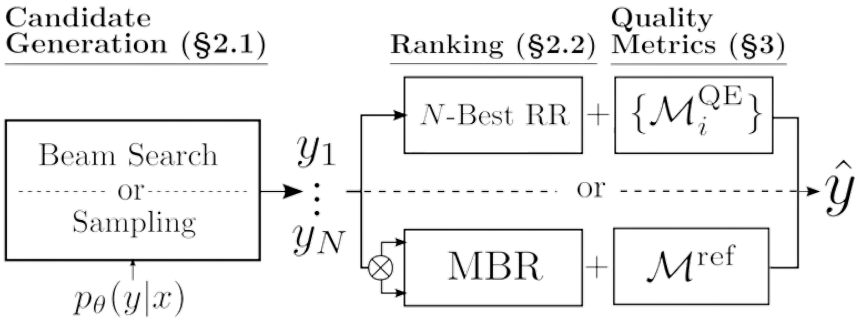
Quality Aware Decoding*:



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

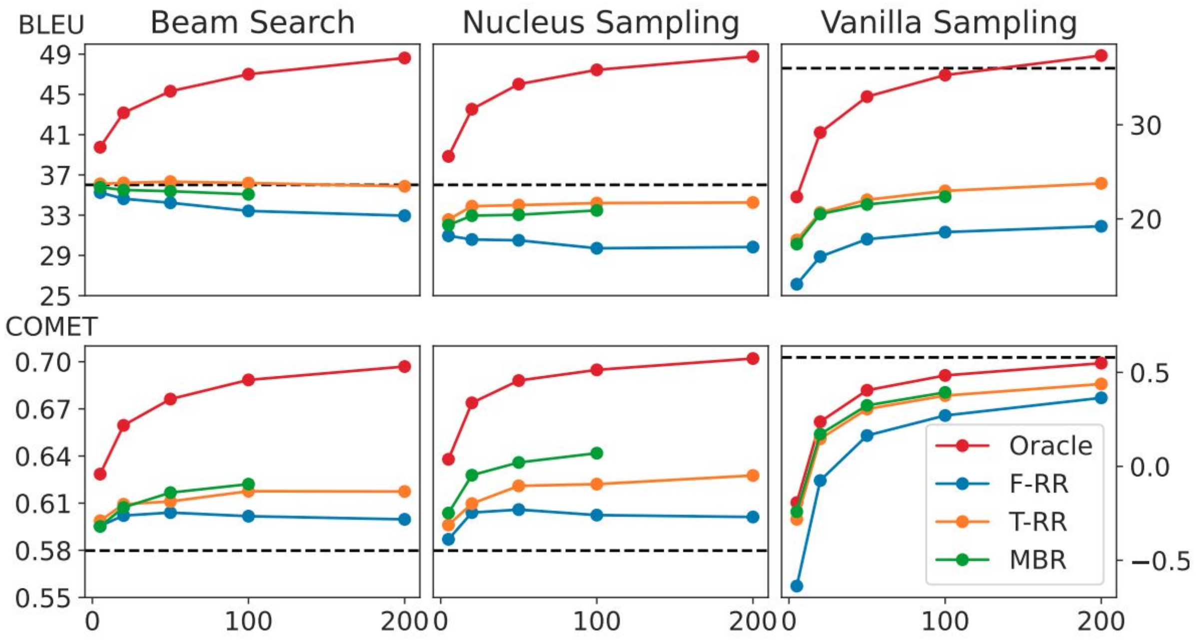
Quality Aware Decoding

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



Values for BLEU (top) and COMET (bottom) for EN → DE as we increase the number of candidates for different generation and ranking procedures, as well as oracles with the respective metrics. Baseline values (with beam size of 5) are marked with a dashed horizontal line.

Quality Aware Decoding:
Impact on different Automatic Metrics

	Large (WMT20)				Small (IWSLT)			
	BLEU	chrF	BLEURT	COMET	BLEU	chrF	BLEURT	COMET
Baseline	**36.01**	63.88	0.7376	0.5795	29.12	56.23	0.6635	0.3028
F-RR w/ COMET-QE	29.83	59.91	<u>0.7457</u>	<u>0.6012</u>	<u>27.38</u>	54.89	<u>0.6848</u>	<u>0.4071</u>
F-RR w/ MBART-QE	<u>32.92</u>	<u>62.71</u>	0.7384	0.5831	27.30	<u>55.62</u>	0.6765	0.3533
F-RR w/ OpenKiwi	30.38	59.56	0.7401	0.5623	25.35	51.53	0.6524	0.2200
F-RR w/ Transquest	31.28	60.94	0.7368	0.5739	26.90	54.46	0.6613	0.2999
T-RR w/ BLEU	<u>35.34</u>	<u>63.82</u>	0.7407	0.5891	**<u>30.51</u>**	**<u>57.73</u>**	0.7077	0.4536
T-RR w/ BLEURT	33.39	62.56	<u>0.7552</u>	0.6217	30.16	57.40	<u>0.7127</u>	<u>0.4741</u>
T-RR w/ COMET	34.26	63.31	0.7546	<u>0.6276</u>	30.16	57.32	0.7124	0.4721
MBR w/ BLEU	<u>34.94</u>	<u>63.21</u>	0.7333	0.5680	29.25	56.36	0.6619	0.3017
MBR w/ BLEURT	32.90	62.34	<u>0.7649</u>	0.6047	28.69	56.28	<u>0.7051</u>	0.3799
MBR w/ COMET	33.04	62.65	0.7477	<u>0.6359</u>	<u>29.43</u>	<u>56.74</u>	0.6882	<u>0.4480</u>
T-RR+MBR w/ BLEU	<u>35.84</u>	**<u>63.96</u>**	0.7395	0.5888	<u>30.23</u>	<u>57.34</u>	0.6913	0.3969
T-RR+MBR w/ BLEURT	33.61	62.95	**<u>0.7658</u>**	0.6165	29.28	56.77	**<u>0.7225</u>**	0.4361
T-RR+MBR w/ COMET	34.20	63.35	0.7526	**<u>0.6418</u>**	29.46	57.13	0.7058	**<u>0.5005</u>**

Quality Aware Decoding:

Impact on different Automatic Metrics

	Large (WMT20)				Small (IWSLT)			
	BLEU	chrF	BLEURT	COMET	BLEU	chrF	BLEURT	COMET
Baseline	**36.01**	63.88	0.7376	0.5795	29.12	56.23	0.6635	0.3028
F-RR w/ COMET-QE	29.83	59.91	<u>0.7457</u>	<u>0.6012</u>	<u>27.38</u>	54.89	<u>0.6848</u>	<u>0.4071</u>
F-RR w/ MBART-QE	<u>32.92</u>	<u>62.71</u>	0.7384	0.5831	27.30	<u>55.62</u>	0.6765	0.3533
F-RR w/ OpenKiwi	30.38	59.56	0.7401	0.5623	25.35	51.53	0.6524	0.2200
F-RR w/ Transquest	31.28	60.94	0.7368	0.5739	26.90	54.46	0.6613	0.2999
T-RR w/ BLEU	<u>35.34</u>	<u>63.82</u>	0.7407	0.5891	**30.51**	**57.73**	0.7077	0.4536
T-RR w/ BLEURT	33.39	62.56	<u>0.7552</u>	0.6217	30.16	57.40	<u>0.7127</u>	<u>0.4741</u>
T-RR w/ COMET	34.26	63.31	0.7546	<u>0.6276</u>	30.16	57.32	0.7124	0.4721
MBR w/ BLEU	<u>34.94</u>	<u>63.21</u>	0.7333	0.5680	29.25	56.36	0.6619	0.3017
MBR w/ BLEURT	32.90	62.34	<u>0.7649</u>	0.6047	28.69	56.28	<u>0.7051</u>	0.3799
MBR w/ COMET	33.04	62.65	0.7477	<u>0.6359</u>	<u>29.43</u>	<u>56.74</u>	0.6882	<u>0.4480</u>
T-RR+MBR w/ BLEU	<u>35.84</u>	**63.96**	0.7395	0.5888	<u>30.23</u>	<u>57.34</u>	0.6913	0.3969
T-RR+MBR w/ BLEURT	33.61	62.95	**0.7658**	0.6165	29.28	56.77	**0.7225**	0.4361
T-RR+MBR w/ COMET	34.20	63.35	0.7526	**0.6418**	29.46	57.13	0.7058	**0.5005**

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**†

Error severity counts and MQM scores for WMT20 (large models). Best overall values are bolded. Methods with † are statistically significantly better than the baseline, with $p < 0.05$.

Take home message

Take home message

- Quality estimation estimates **how good a translation is**

- Predictor-estimator architecture is still the SOTA but today's systems are built on top of Muppet models

- More and more we need to worry about generalization of our QE systems.
  - Generalization for new language pairs
  - Generalization to new domains
  - Robustness to different type of annotations

- QE can be effectively used to improve decoding by ranking translations in a candidate list

Take home message

Some future work directions:

- How to incorporate context into QE (document-level QE)
- How to efficiently incorporate QE into decoding

Questions?

Unbabel

Thank you!