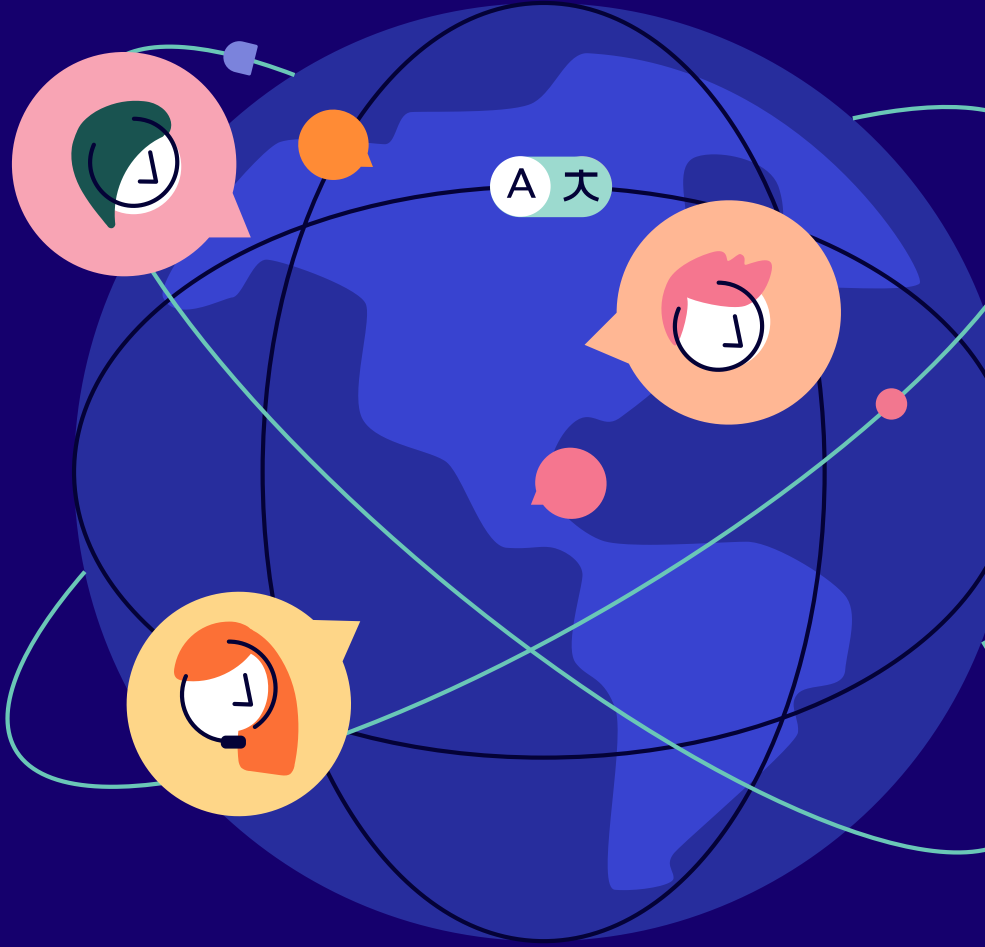
Quality Estimation for Machine Translation

Nuno M Guerreiro
Unbabel AI

August 2023



Unbabel

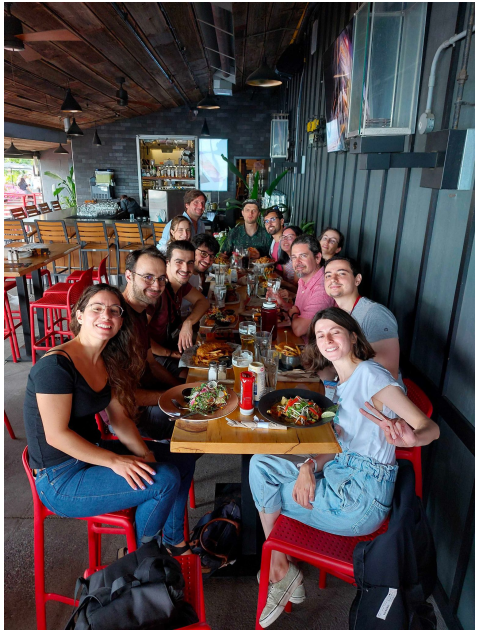
Work with the Research Team


André Martins


Catarina Farinha


José Souza


João Alves




José Pombal


Ricardo Rei


Amin Farajian


Pedro Martins

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

- **We actively collaborate with the Sardine LAB**

Agenda

01

Definition

02

Models

03

COMET for QE

04

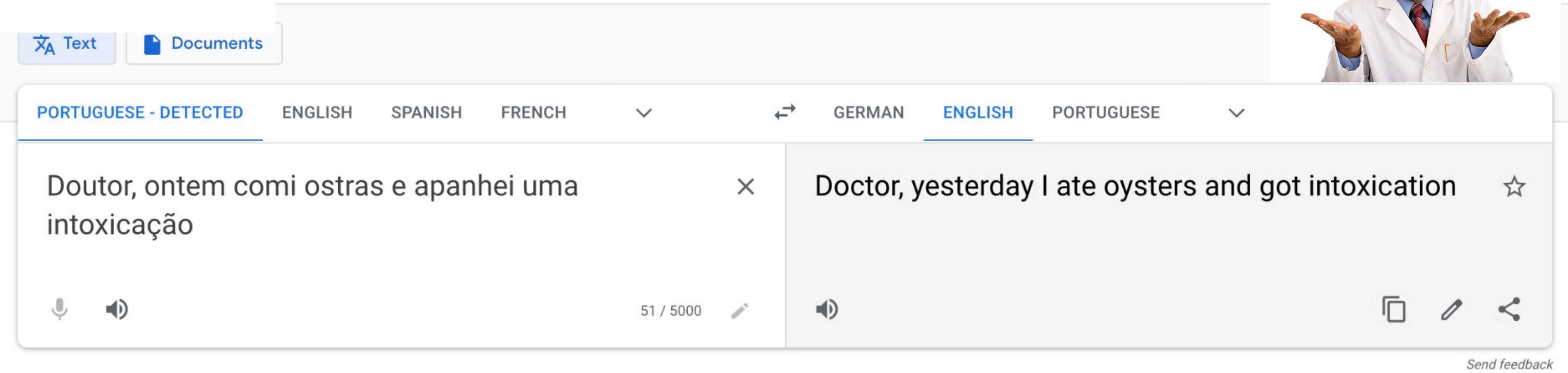
Challenges and Applications

05

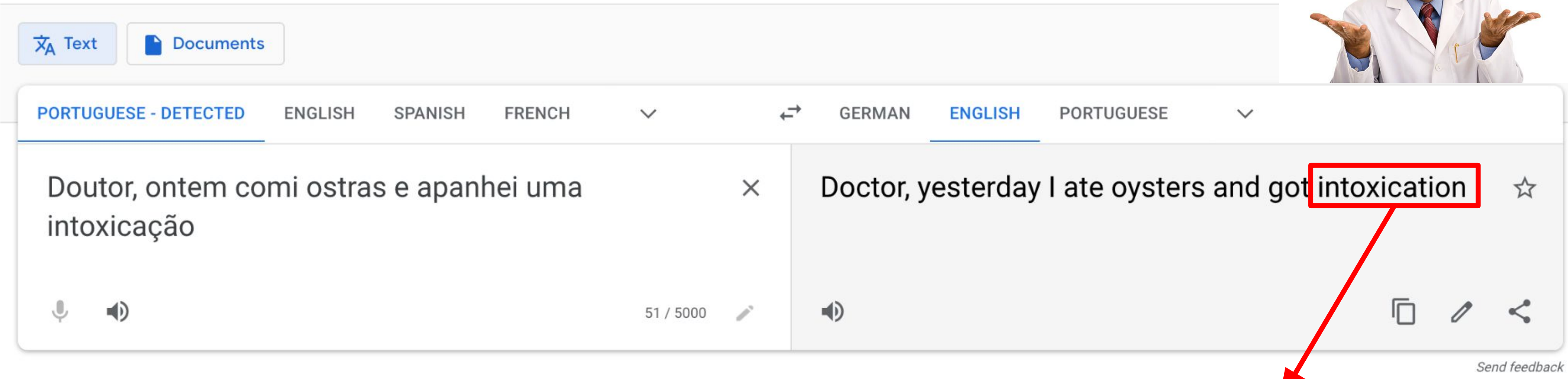
Take home messages

Why Quality Estimation?

Is Machine Translation solved?

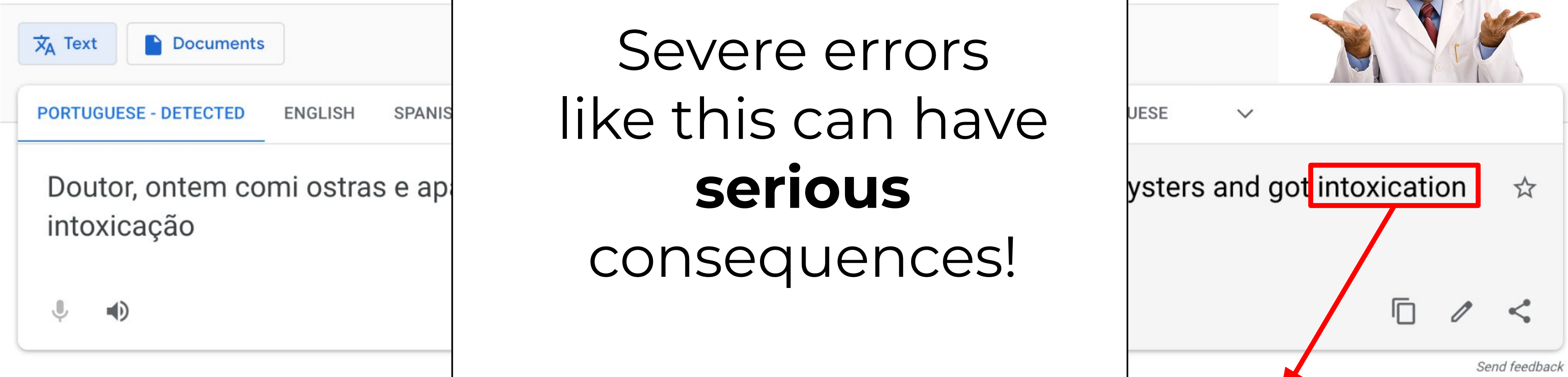


Is Machine Translation solved?



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 know the **quality of a translation**?

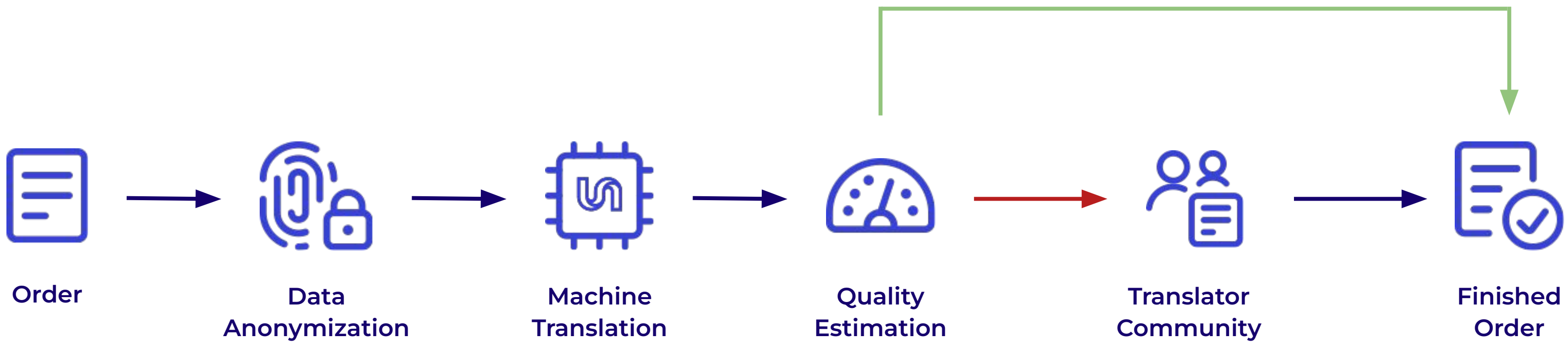
1) If it is good, we can trust it and use it.

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

3) Many other things...

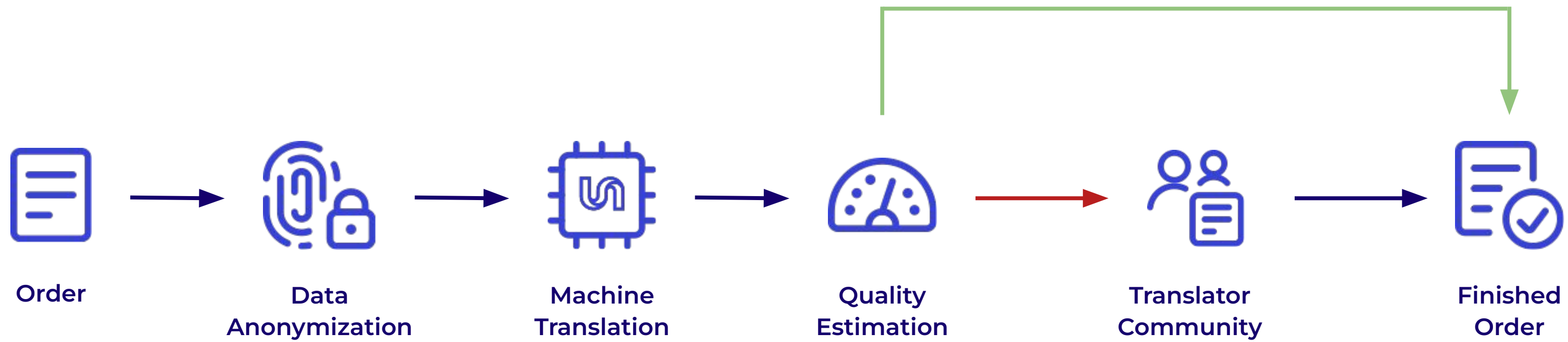
Motivation:

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



Motivation:

What can we do if we know 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	English
Tarde :) Como posso ajudá-lo?	Afternoon :) How may I help you?
Comprei um monitor cardíaco mas não consegui colocar em funcionamento.	I bought a heart monitor but I couldn't get it up and running
Já atualizei o sistema e tetei colocar a recarregar, mas parece que não carrega.	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**

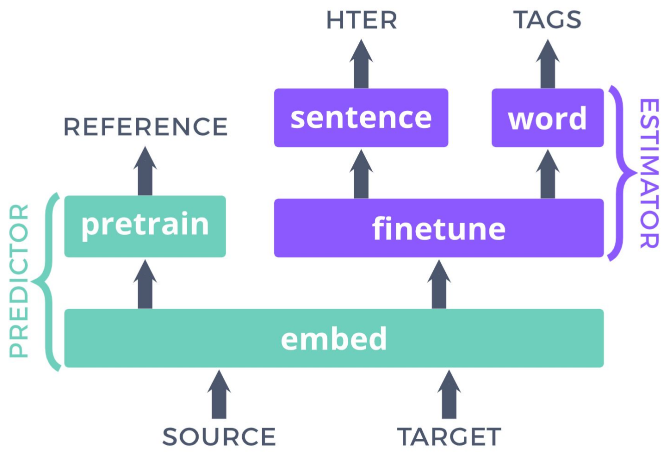
Architecture of QE models

Predictor-Estimator

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

- Deep contextualized language model pretraining

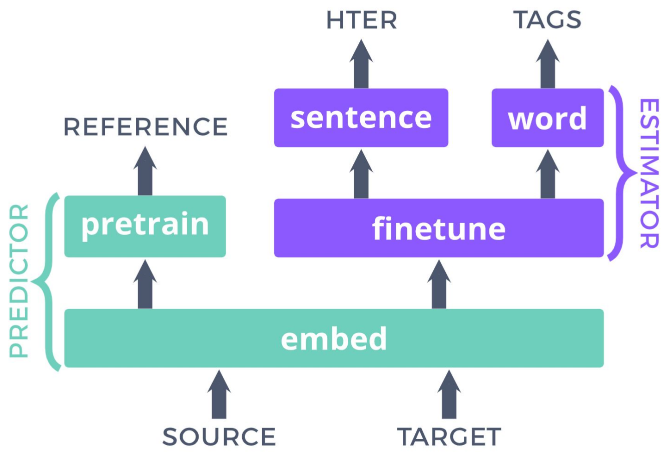
- 1 year ahead of muppet models!



* 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

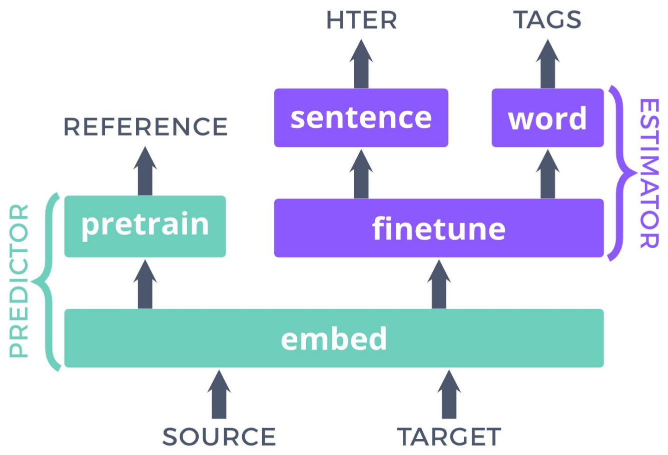


* Predictor-Estimator using Multilevel Task Learning with Stack Propagation for Neural Quality Estimation (Kim et al., 2017)

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.



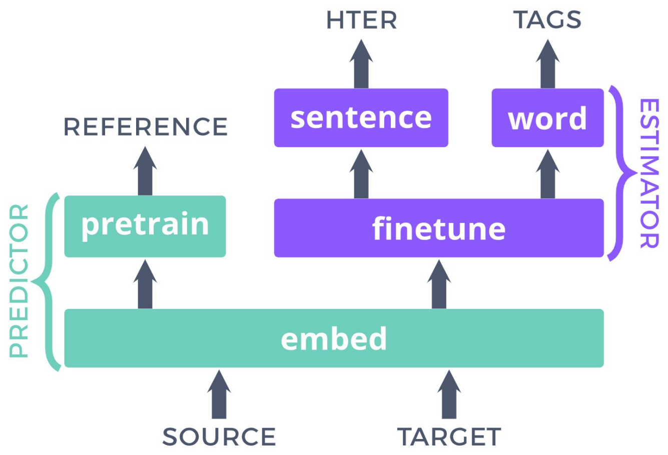
* Predictor-Estimator using Multilevel Task Learning with Stack Propagation for Neural Quality Estimation (Kim et al., 2017)

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.

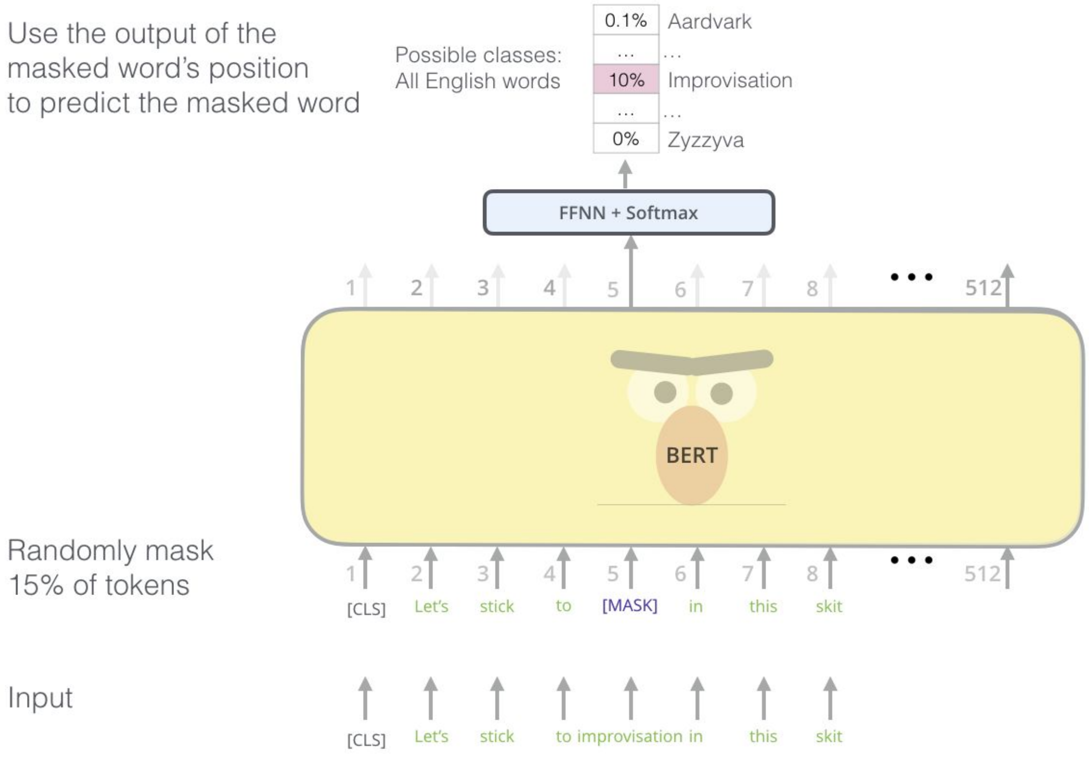
Unbabel's winning participation in WMT19



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

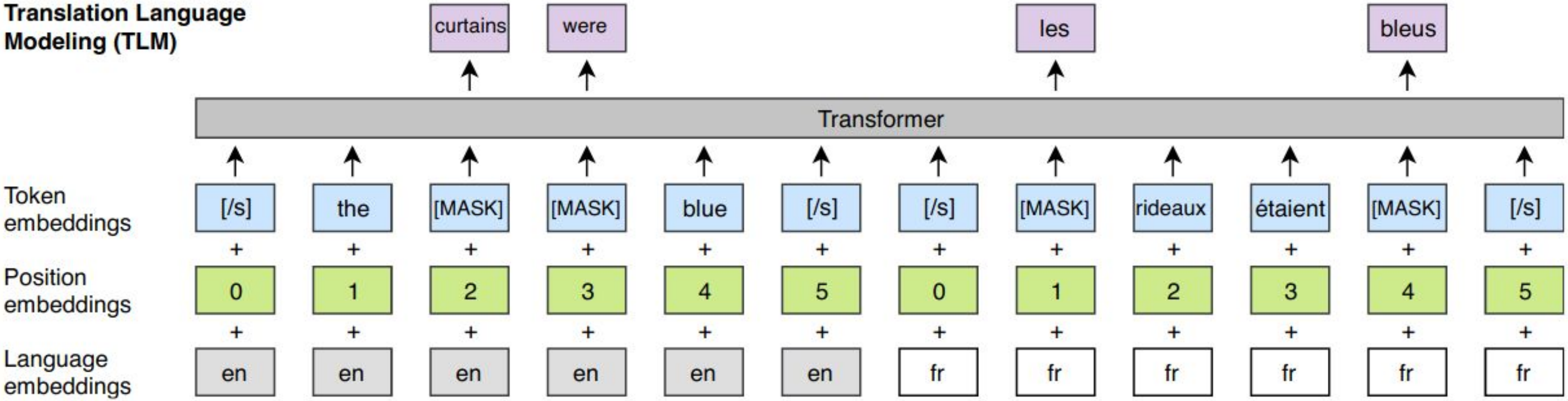
* [TransQuest: Translation Quality Estimation with Cross-lingual Transformers]() (Ranasinghe et al., COLING 2020)

Predictor: BERT & XLM-R

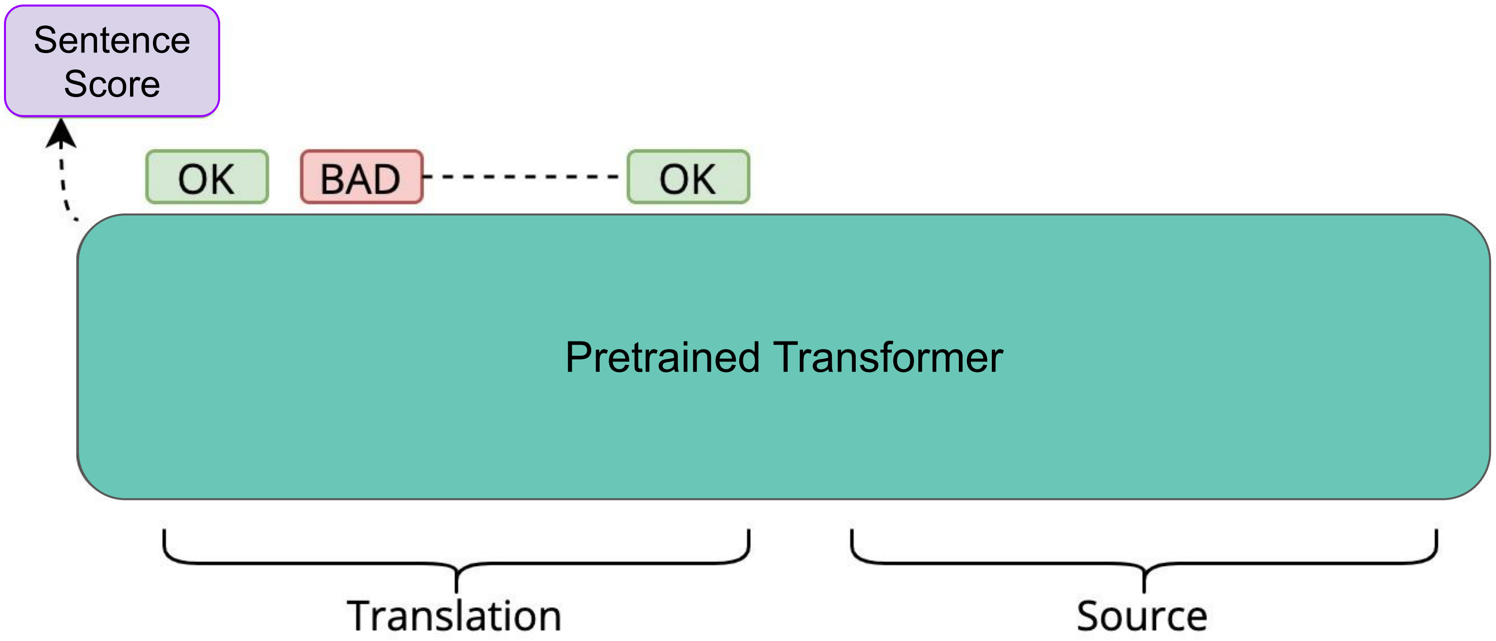


Source: The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning), Jay Alammar, 2019.

Predictor: XLM & InfoXLM

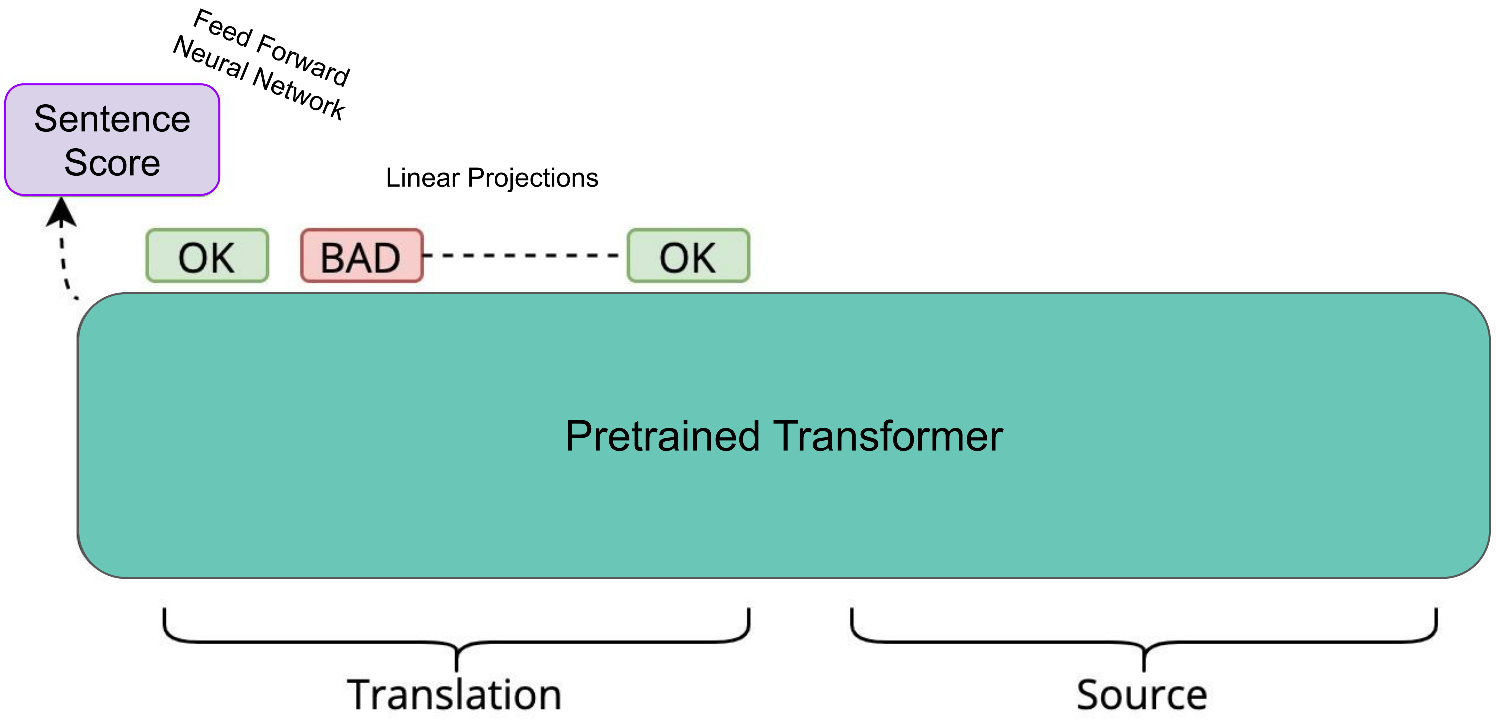


Estimator:

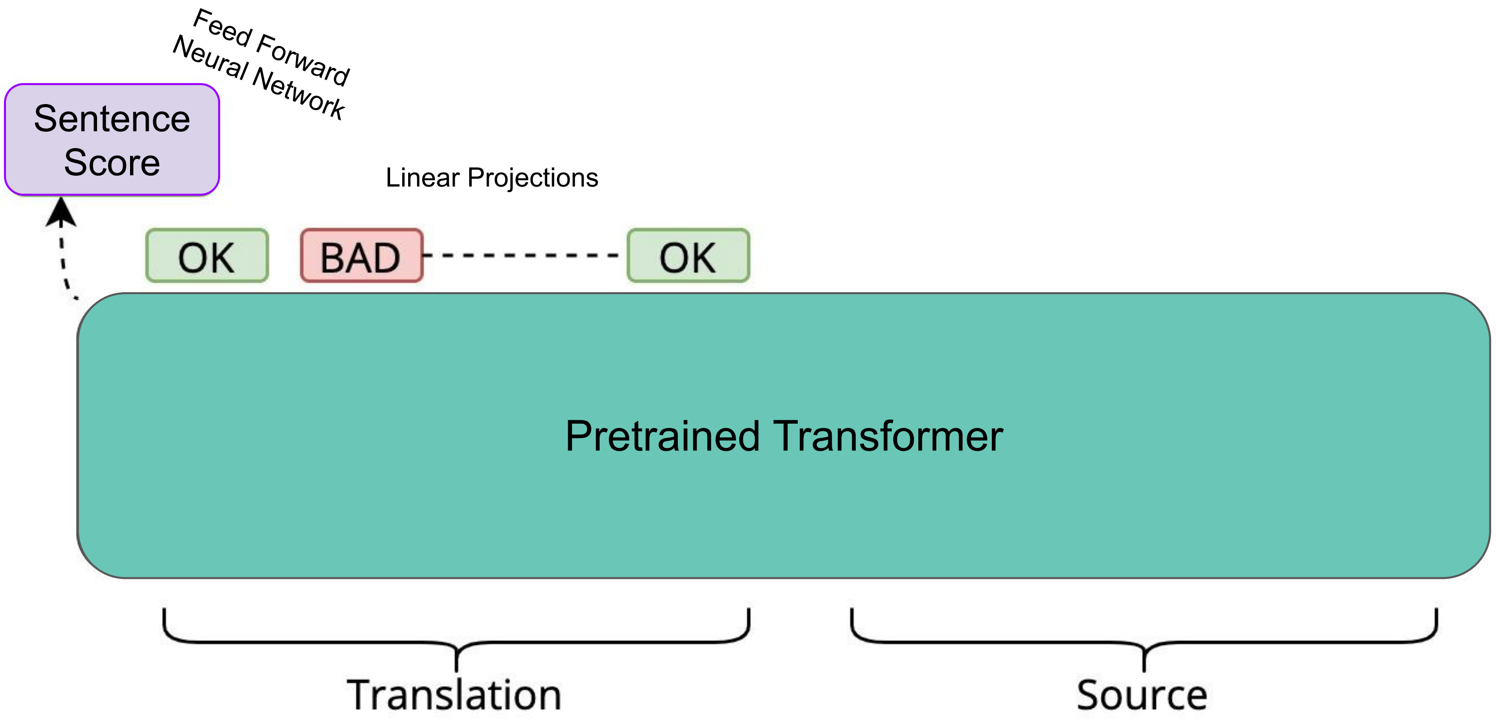


Sentence Score
OK
BAD
OK
Pretrained Transformer
Translation
Source

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

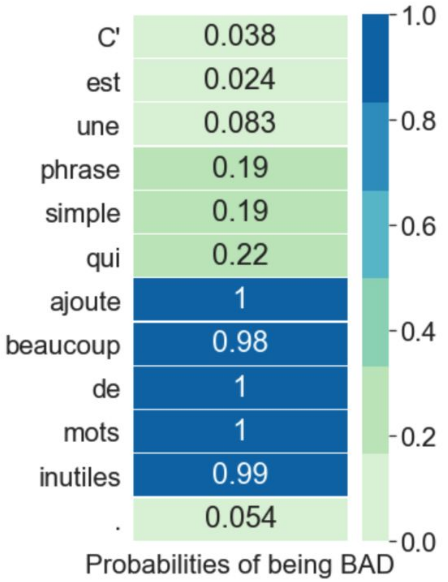
Example:

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']
```

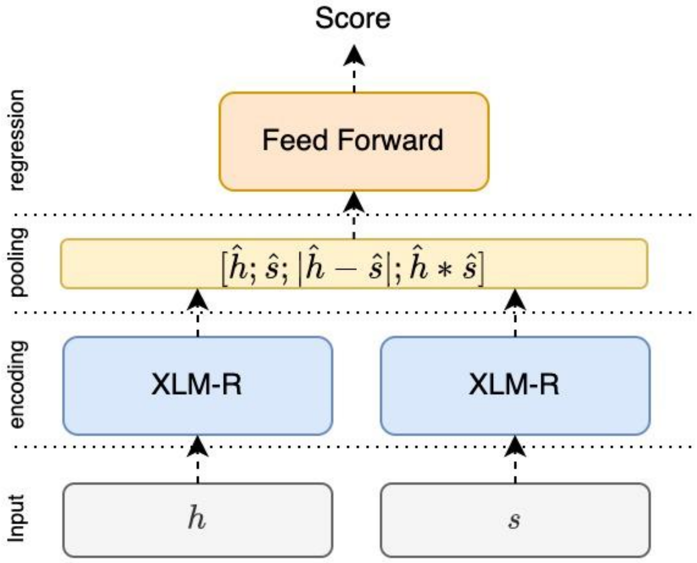
```
MACHINE_TRANSLATION: C' est une phrase simple qui ajoute beaucoup de mots inutiles .
```

COMET for
Quality Estimation

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)

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

Results from the WMT 21 Metrics task

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)

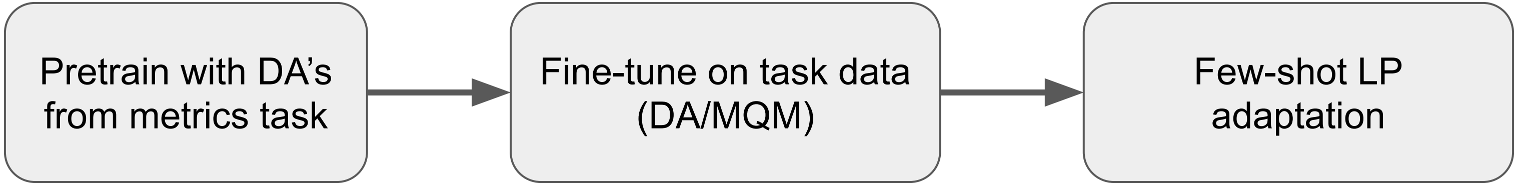
CometKiwi

COMETKIWI:
IST-Unbabel 2022 Submission for the Quality Estimation Shared Task

Ricardo Rei*[1,2,4], Marcos Treviso*[3,4], Nuno M. Guerreiro*[3,4], Chrysoula Zerva*[3,4], Ana C. Farinha[1], Christine Maroti[1], José G. C. de Souza[1], Taisiya Glushkova[3,4], Duarte M. Alves[1,4], Alon Lavie[1], Luisa Coheur[2,4], André F. T. Martins[1,3,4]

CometKiwi follows a "curriculum" during training:

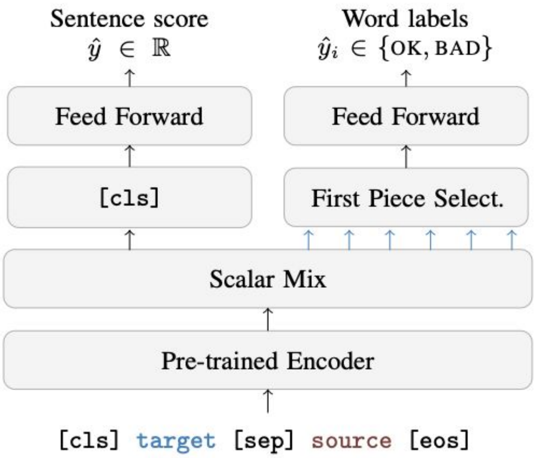
- We first start by training on data with references to obtain a metric
- This serves as initialization to training a QE system (only trained with src + mt)
- We further tuned the model for languages for which training data is very scarce (with up to 500 samples only)



CometKiwi

CometKiwi combines sentence-level and word-level objectives during training.

- We obtain positive transfer from this multi-task objective
- This architecture allows for a **single** model to perform both sentence and word-level quality estimation
- **Winning submission of all tasks in the WMT 2022 QE Shared Task**



$$\mathcal{L}_{\text{sent}}(\theta) = \frac{1}{2}(y - \hat{y}(\theta))^2$$

$$\mathcal{L}_{\text{word}}(\theta) = -\frac{1}{n}\sum_{i=1}^{n} w_{y_i} \log p_\theta(y_i)$$

$$\mathcal{L}(\theta) = \lambda_s \mathcal{L}_{\text{sent}}(\theta) + \lambda_w \mathcal{L}_{\text{word}}(\theta)$$

🤗 `Unbabel/wmt22-cometkiwi-da`

* CometKiwi: IST-Unbabel 2022 Submission for the Quality Estimation Shared Task (Rei et al., WMT 2022)

xCOMET
a more fine-grained system

Looking back at MQM...

English to Portuguese

I am going to present some interesting research works in Tartu, Estonia.	Nós vamos apresentar alguns trabalhos de investigação interessantes em Tallinn, Estónia.	Weak

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

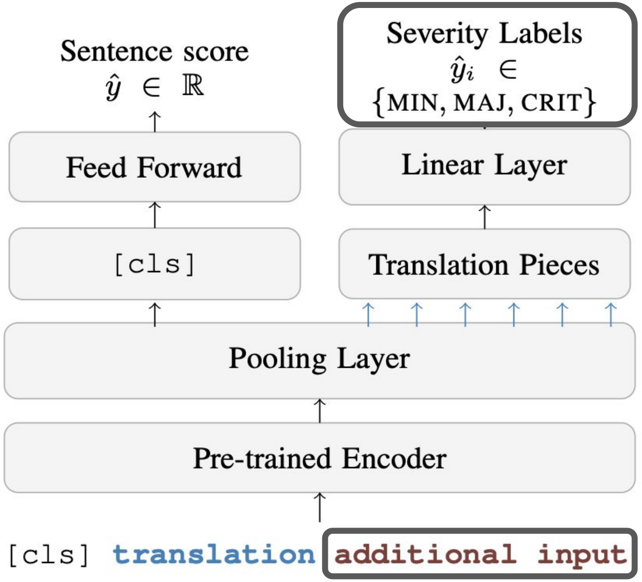
- The MQM severities are way **more fine-grained than OK/BAD**;
- If we predict the severities well, we get **sentence-level scores for free**!

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

xCOMET

xCOMET* adds two different components:

- We design xCOMET as a **single model that can be used as a metric or as a QE system**:
  - Reference-based quality estimation (metrics — ref-only and src+ref)
  - Quality estimation (src-only)
- We now predict fine-grained error severities
- We will be releasing two versions of xCOMET: xCOMET-XL (3.5B) and xCOMET-XXL (10.7B)



* Paper to be released soon.

xCOMET brings more transparency to the scores

xCOMET* brings a finer-grained look at predicted scores

- The sentence-level scores correlate **very strongly** with MQM scores obtained via the error spans
- Given this correlation, when the quality of a translation is low, we can look at what the model flagged as error spans — **additional transparency**!

Pearson correlations between predicted score via sentence-level head and via MQM formula through span predictions

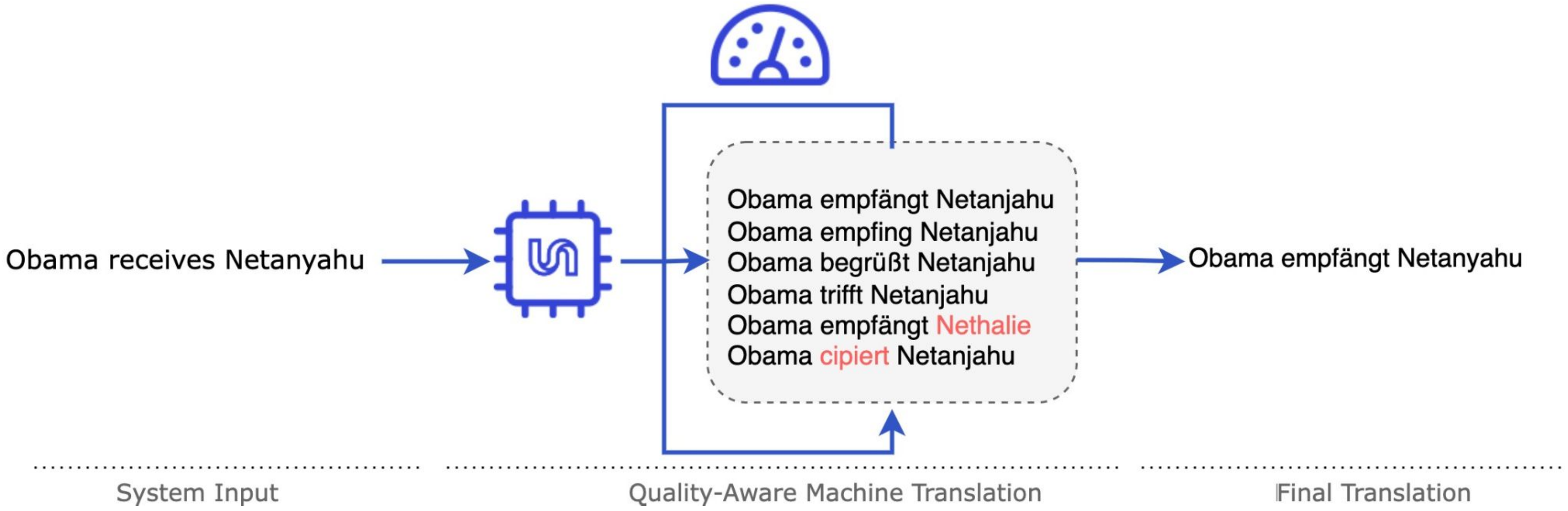
zh-en	en-de	he-en
0.91	0.95	0.90

Source	Translation (English Formal)	Quality
Die Zimmer beziehen, die Fenster mit Aussicht öffnen, tief durchatmen, staunen.	The staff were very friendly and helpful.	Weak
Vielen Dank, Herr Kollege.	Thank you very much, Mr Schroedter.	Weak

* Paper to be released soon.

Challenges and Applications

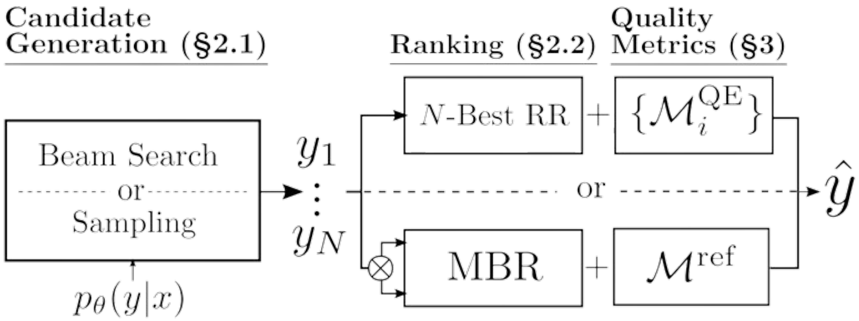
Quality aware decoding leads to consistent gains in translation performance



* 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 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	0.7552	0.6217	30.16	57.40	0.7127	0.4741
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	0.7051	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	**0.7658**	0.6165	29.28	56.77	**0.7225**	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

	EN-DE (WMT20)					EN-RU (WMT20)				
	BLEU	chrF	BLEURT	COMET	Human R.	BLEU	chrF	BLEURT	COMET	Human R.
Reference	-	-	-	-	4.51	-	-	-	-	4.07
Baseline	**36.01**	**63.88**	0.7376	0.5795	4.28	**23.86**	51.16	0.6953	0.5361	3.62
F-RR w/ COMET-QE	29.83	59.91	0.7457	0.6012	4.19	20.32	49.18	0.7130	0.6207	3.25
T-RR w/ COMET	34.26	63.31	**0.7546**	0.6276	**4.33**	22.42	50.91	**0.7243**	0.6441	3.65
MBR w/ COMET	33.04	62.65	0.7477	0.6359	4.27	23.67	51.18	0.7093	0.6242	3.66
T-RR + MBR w/ COMET	34.20	63.35	0.7526	**0.6418**	4.30	23.21	**51.26**	0.7238	**0.6736**	**3.72**†

	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 $p < 0.05$.

Quality aware decoding/training can help uncover biases in the metrics

Identifying Weaknesses in Machine Translation Metrics Through Minimum Bayes Risk Decoding: A Case Study for COMET

Chantal Amrhein[1] and **Rico Sennrich**[1,2]

`src`	Schon drei Jahre nach der Gründung verließ Green die Band **1970**.
`ref`	Green left the band three years after it was formed, in **1970**.
MBR$_{\texttt{chrF++}}$	Already three years after the foundation, Green left the band in **1970**.
MBR$_{\texttt{COMET}}$	Three years after the creation, Green left the band in **1980**.
`src`	[...] **Mahmoud** Guemama's Death - Algeria Loses a Patriot [...], Says President **Tebboune**.
`ref`	[...] **Mahmoud** Guemamas Tod - Algerien verliert einen Patrioten [...], sagt Präsident **Tebboune**.
MBR$_{\texttt{chrF++}}$	[...] **Mahmoud** Guemamas Tod - Algerien verliert einen Patriot [...], sagt Präsident **Tebboune**.
MBR$_{\texttt{COMET}}$	[...] **Mahmud** Guemamas Tod - Algerien verliert einen Patriot [...], sagt Präsident **Tebboene**.

Table 1: Examples of MBR decoding outputs with chrF++ and COMET as utility metrics. The outputs chosen with COMET indicate less sensitivity towards discrepancies in numbers and named entities.

Quality-aware decoding may exacerbate and help uncover biases in the metrics

Quality aware decoding/training can help uncover biases in the metrics

BLEURT Has Universal Translations: An Analysis of Automatic Metrics by Minimum Risk Training

Yiming Yan[1*], Tao Wang[2], Chengqi Zhao[2], Shujian Huang[1†], Jiajun Chen[1], Mingxuan Wang[2]

hypo: Lage vom Hotel war grundsätzlich bestens – Hotelpersonal weitgehend zuvorkommend bzw. ggf. hilfehilfsbereit. Vor allem die Lage des Hotels war gut, Hotelmitarbeiter grundsätzlich äußerst lieb bzw. gegebenenfalls auch durchaus hilfehilfsbereit.

ref: 123	**BLEURT:** 0.8693	**ref:** a	**BLEURT:** 0.8970
ref: May the sunshine always be with you.			**BLEURT:** 0.8341

Optimize BARTScore on De⇒En	141	! Mallorca! Mallorca! Mallorca! Mallorca! Mallorca! Mallorca! Mallorca! Mallorca! Mallorca! Mallorca! Mallorca! Mallorca! Mallorca! Mallorca! Mallorca! Mallorca! Mallorca! Mallorca!
	137	Mallorca! Mallorca! Mallorca! Mallorca! Mallorca! Mallorca! Mallorca! Mallorca! Mallorca! Mallorca! Mallorca! Mallorca!

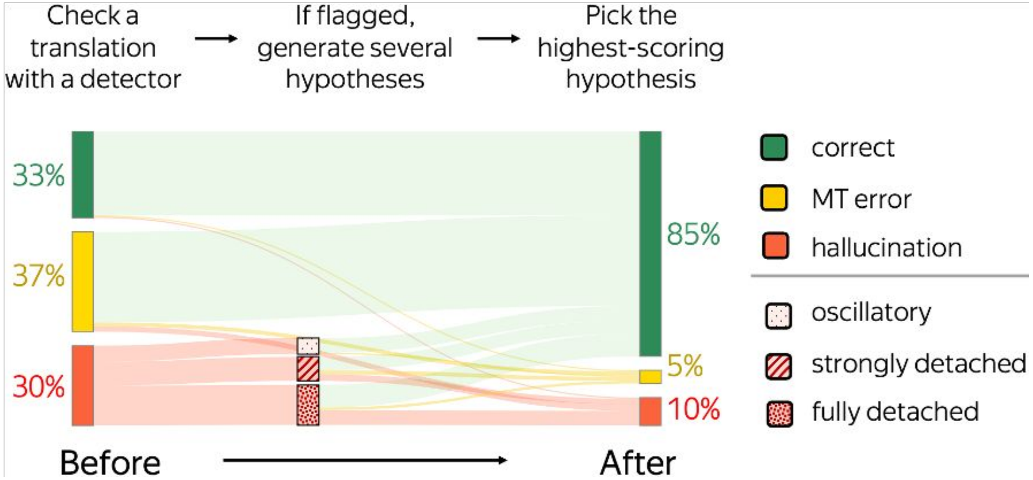
Using MRT to optimize a model for a reward provided by a metric — helps uncover biases in the metrics.

On-the-fly mitigation of hallucinations

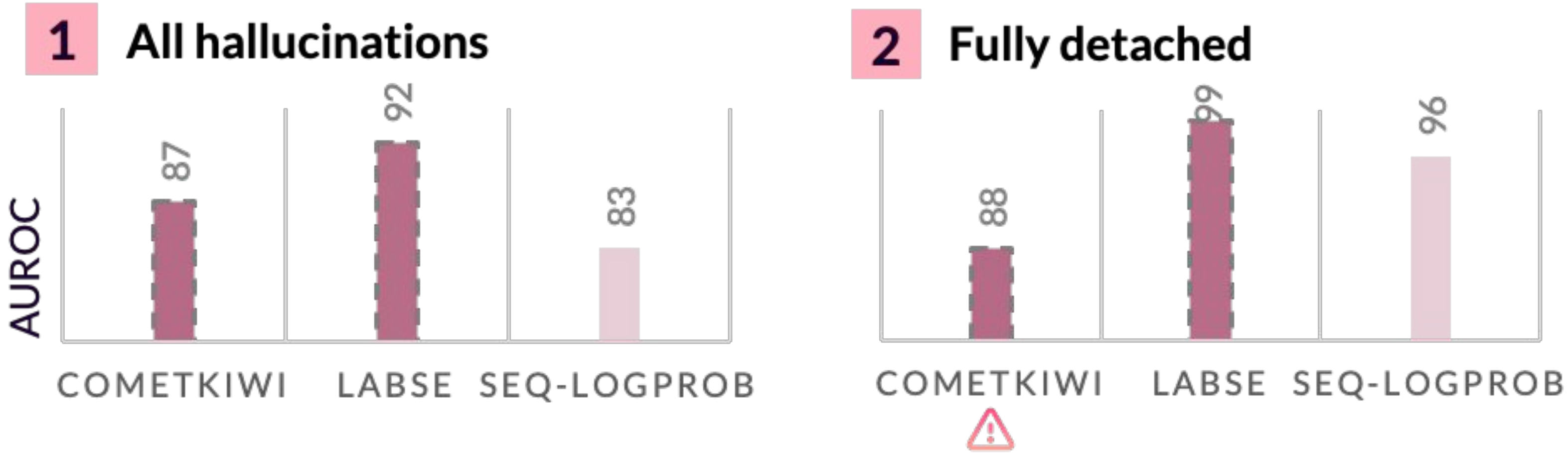
Looking for a Needle in a Haystack:
A Comprehensive Study of Hallucinations in Neural Machine Translation

Nuno M. Guerreiro[1,2] **Elena Voita**[4] **André F. T. Martins**[1,2,3]

w/ COMET-QE



Leverage contrastive losses for making QE systems more robust?



LaBSE is trained with a **translation matching objective** that is very much aligned with hallucination detection; could a similar objective be employed successfully for training more robust and general QE systems?

Analysis of quality estimation systems and neural metrics through different lenses!

Extrinsic Evaluation of Machine Translation Metrics

Nikita Moghe and **Tom Sherborne** and **Mark Steedman** and **Alexandra Birch**

Investigate the correlation between translation quality and translation utility in downstream tasks.

Analysis of quality estimation systems and neural metrics through different lenses!

Extrinsic Evaluation of Machine Translation Metrics

Nikita Moghe and **Tom Sherborne** and **Mark Steedman** and **Alexandra Birch**

Investigate the correlation between translation quality and translation utility in downstream tasks.

The Inside Story: Towards Better Understanding of Machine Translation Neural Evaluation Metrics

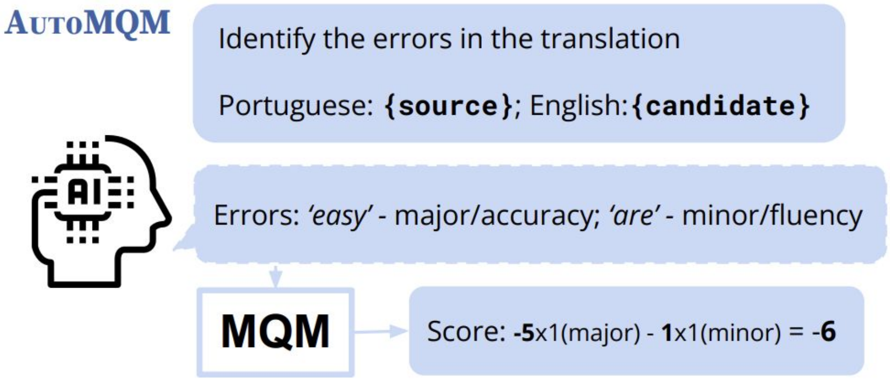
Ricardo Rei[*1,2,4], Nuno M. Guerreiro[*3,4], Marcos Treviso[3,4], Alon Lavie[1], Luisa Coheur[2,4], André F. T. Martins[1,3,4]

Investigate with explainability methods whether salient tokens correlate with errors in MQM annotations.

Generative LLMs can be leveraged for quality estimation

The Devil is in the Errors: Leveraging Large Language Models for Fine-grained Machine Translation Evaluation

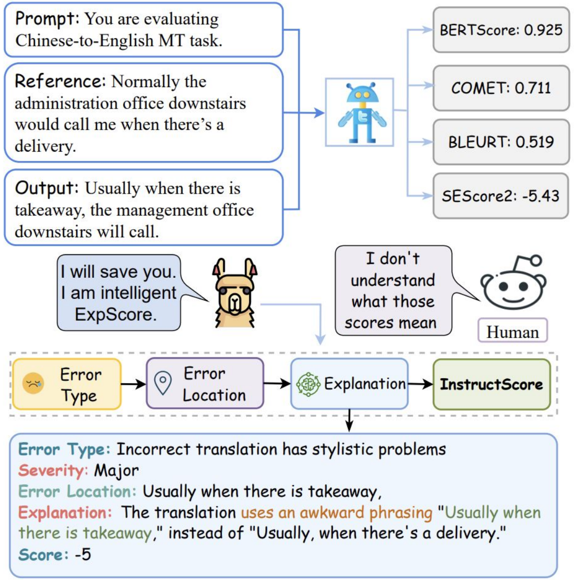
Patrick Fernandes[*,2,3,4] Daniel Deutsch[1] Mara Finkelstein[1] Parker Riley[1]
André F. T. Martins[3,4,5] Graham Neubig[2,6]
Ankush Garg[1] Jonathan H. Clark[1] Markus Freitag[1] Orhan Firat[1]



Generative LLMs can be leveraged for quality estimation

INSTRUCTSCORE: Towards Explainable Text Generation Evaluation with Automatic Feedback

Wenda Xu[¶], Danqing Wang[¶], Liangming Pan[¶], Zhenqiao Song[¶], Markus Freitag[†], William Yang Wang[¶], Lei Li[¶]



Generative LLMs still lag behind dedicated systems in sentence-level quality estimation

		System-Level	Segment-Level					
		All (3 LPs)	EN-DE		ZH-EN		EN-RU	
Model	Ref?	Accuracy	ρ	acc*	ρ	acc*	ρ	acc*
Baselines								
MetricX-XXL	✓	85.0%	0.549	61.1%	0.581	54.6%	0.495	60.6%
COMET-22	✓	83.9%	0.512	60.2%	0.585	54.1%	0.469	57.7%
COMET-QE	✗	78.1%	0.419	56.3%	0.505	48.8%	0.439	53.4%
Prompting								
PaLM 540B	✓	90.1%	0.247	55.4%	0.255	48.5%	0.180	48.6%
PaLM-2 BISON	✓	88.7%	0.394	56.8%	0.322	49.3%	0.322	52.8%
PaLM-2 UNICORN	✓	90.1%	0.401	56.3%	0.349	51.1%	0.352	55.3%
FLAN-PaLM-2 UNICORN	✓	75.9%	0.197	55.6%	0.139	46.1%	0.198	52.0%
PaLM 540B	✗	84.3%	0.239	56.1%	0.270	43.1%	0.300	51.8%
PaLM-2 BISON	✗	85.0%	0.355	57.0%	0.299	48.6%	0.303	53.1%
PaLM-2 UNICORN	✗	84.3%	0.275	56.1%	0.252	48.3%	0.209	49.8%
FLAN-PaLM-2 UNICORN	✗	69.7%	0.116	54.6%	0.112	43.8%	0.156	47.8%

PaLM and PaLM-2

The Devil is in the Errors: Leveraging Large Language Models for Fine-grained Machine Translation Evaluation (Fernandes et al., 2023)

Generative LLMs still lag behind dedicated systems in sentence-level quality estimation

Model	Ref?	System-Level: All (3 LPs) Accuracy	EN-DE ρ	EN-DE acc*	ZH-EN ρ	ZH-EN acc*	EN-RU ρ	EN-RU acc*
Baselines								
MetricX-XXL	✓	85.0%	0.549	61.1%	0.581	54.6%	0.495	60.6%
COMET-22	✓	83.9%	0.512	60.2%	0.585	54.1%	0.469	57.7%
COMET-QE	✗	78.1%	0.419	56.3%	0.505	48.8%	0.439	53.4%
Prompting								
PaLM 540B	✓	90.1%	0.247	55.4%	0.255	48.5%	0.180	48.6%
PaLM-2 Bison	✓	88.7%	0.394	56.8%	0.322	49.3%	0.322	52.8%
PaLM-2 Unicorn	✓	90.1%	0.401	56.3%	0.349	51.1%	0.352	55.3%
FLAN-PaLM-2 Unicorn	✓	75.9%	0.197	55.6%	0.139	46.1%	0.198	52.0%
PaLM 540B	✗	84.3%	0.239	56.1%	0.270	43.1%	0.300	51.8%
PaLM-2 Bison	✗	85.0%	0.355	57.0%	0.299	48.6%	0.303	53.1%
PaLM-2 Unicorn	✗	84.3%	0.275	56.1%	0.252	48.3%	0.209	49.8%
FLAN-PaLM-2 Unicorn	✗	69.7%	0.116	54.6%	0.112	43.8%	0.156	47.8%

PaLM and PaLM-2

The Devil is in the Errors: Leveraging Large Language Models for Fine-grained Machine Translation Evaluation (Fernandes et al., 2023)

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

GPT*

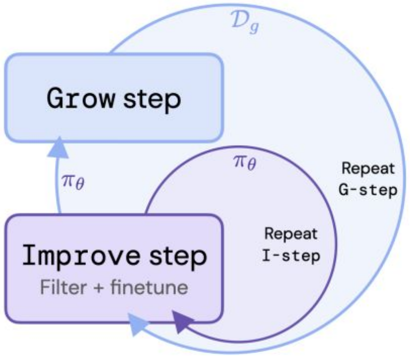
Large Language Models Are State-of-the-Art Evaluators of Translation Quality (Kocmi and Federmann, 2023)

Quality estimation/metrics may act as reward models (different instance of RLHF)

DeepMind

2023-8-22

Reinforced Self-Training (*ReST*) for Language Modeling



Metrics and QE systems can provide alternatives to RLHF, since they are modelled to replicate human preferences.

Take home message

Take home message

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

- Predictor-estimator architecture on top of pre-trained models is still SOTA; this may change soon with the emergence of LLMs.

- More and more we need to worry about generalization, robustness, and defects of our QE systems.

- QE can be used for multiple other applications beyond just sentence-level quality estimation:
    - Discern between systems
    - Hallucination detection
    - Generating translations
    - Training new models

Questions?

Unbabel

Thank you!