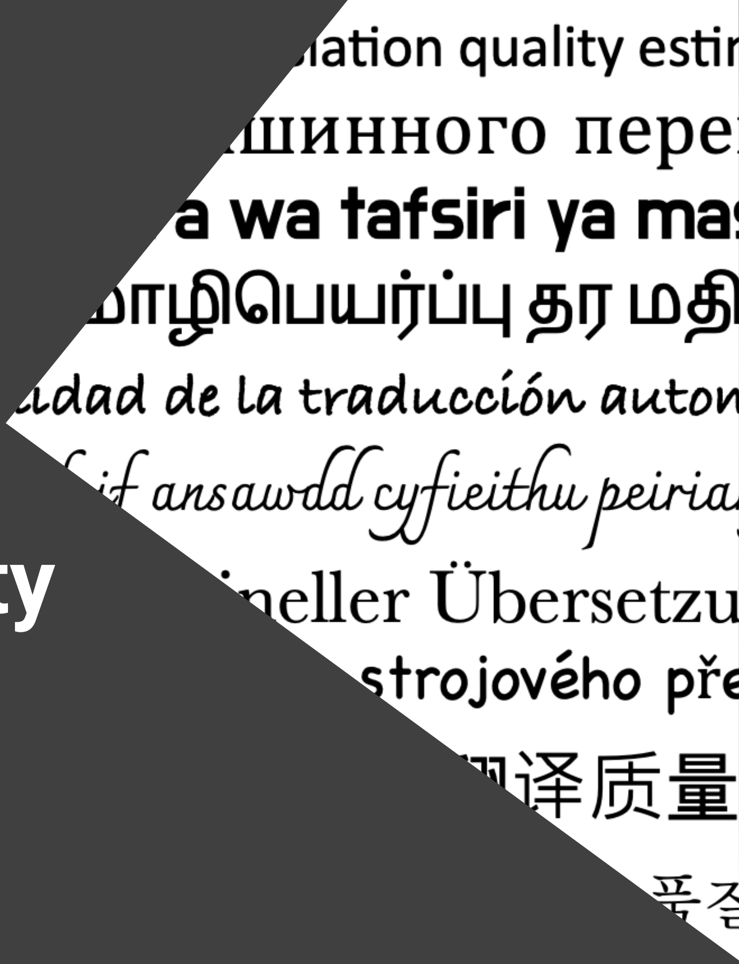
The Alan Turing Institute

Machine Translation Quality Estimation (MTQE)

Jo Knight | Radka Jersakova | James Bishop

26th February 2024, Foundation Models Reading Group



Machine Translation Quality Estimation

MT: e.g., Google Translate

QE: How good is this translation?

SOURCE

Czech

Tři sta třicet tři stříbrných stříkaček stříkalo přes tři sta třicet tři stříbrných střech

TARGET

English

Three hundred and thirty-three silver syringes sprayed over three hundred and thirty-three silver roofs

NO REFERENCE FOR COMPARISON

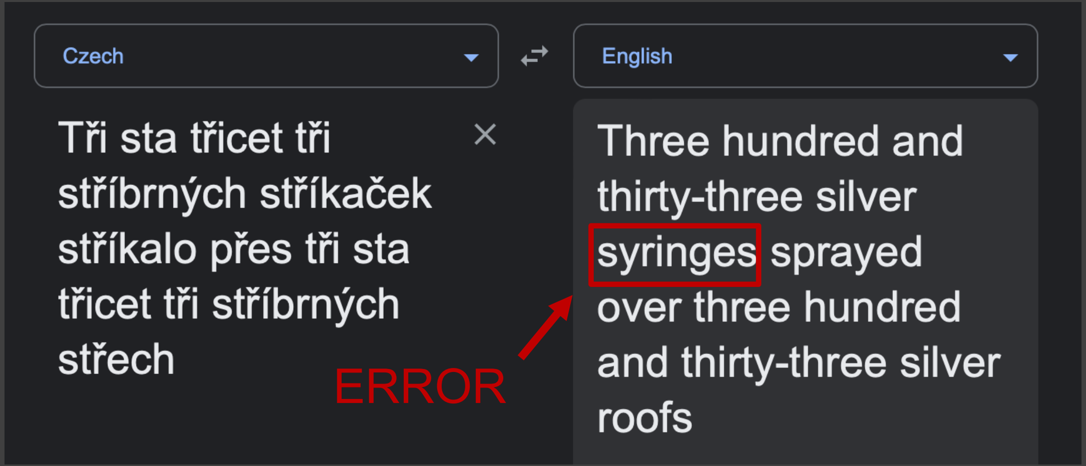
Three hundred and thirty three silver fire engines sprayed over three hundred and thirty three silver roofs

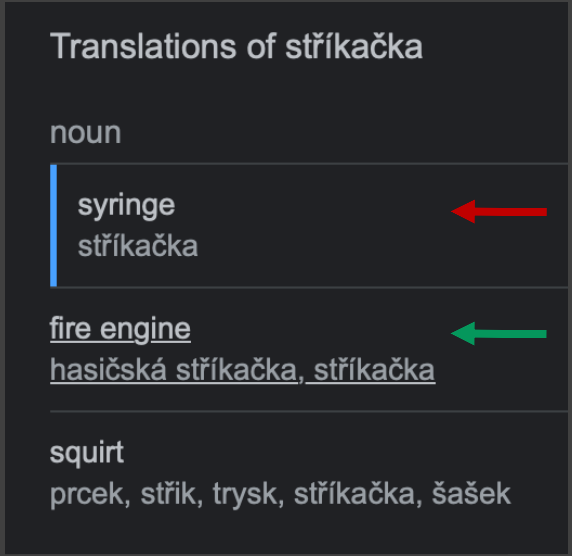
Machine Translation Quality Estimation

MT: e.g., Google Translate

QE: How good is this translation?

SOURCE TARGET





What makes (machine) translation hard?

- Words have **multiple meanings** (e.g., novel, bark)
- **User generated content** (e.g., social media posts)
  - **Idioms** (e.g., it’s raining cats and dogs)
  - **Slang** (e.g., apple and pears = stairs)
  - **Non-standard spelling** (e.g., nvr, tbh, wot, intro)
- **Local conventions** (e.g., DD-MM-YYYY vs MM-DD-YYYY)

How to measure translation quality?

– Not all errors are equal, metrics should reflect this

MINOR	MAJOR	CRITICAL
- spelling	- tense	- antonyms
- punctuation	- word order	- negation

SCORES:
Minor = 1
Major = 5
Critical = 10

Source:
This year's trend for a second Christmas tree in the bedroom sends sales of smaller spruces soaring

sentence score=10

Translation:
Der diesjährige Trend für einen zweiten Weihnachtsbaum in der Schlafzimmer sendet Umsatz von kleineren Fichten steigen

severity: Major
category: Grammar

severity: Major
category: Mistrai

* From WMT QE 2022

Taxonomy of QE models

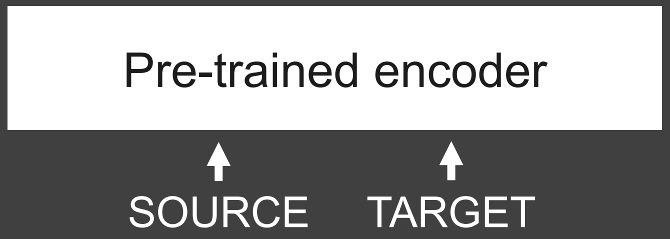
	Black-box only	Black-box and glass-box	Glass-box only
Training required	Traditional approaches & SOTA	~2021	~2020
No training required	LLMs (emerging)		~2020

Glass-box features

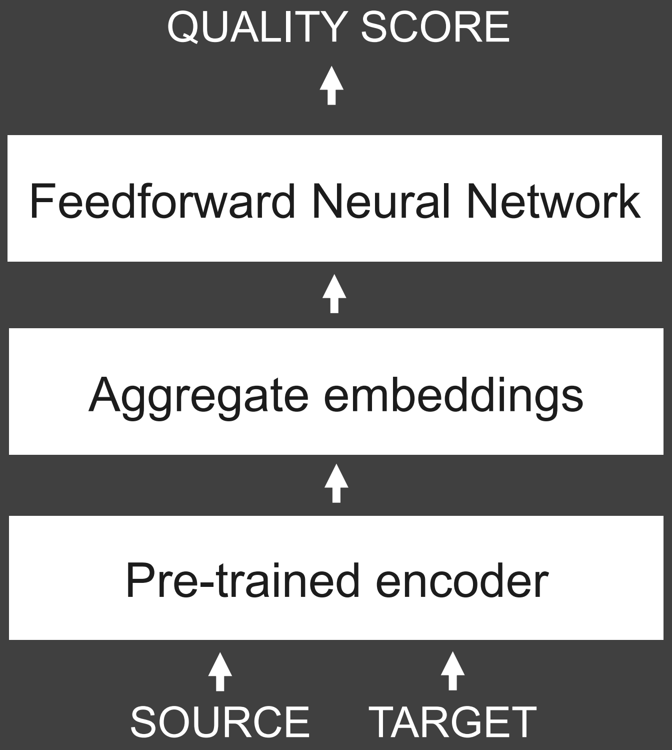
- Features obtained from the underlying MT system:
  - Softmax output distribution
  - Monte-Carlo dropout
    - Run *N* stochastic forward passes through the MT model
    - Calculate the mean of some output over all models
    - e.g., average of softmax output
  - Attention weights

Black-box features

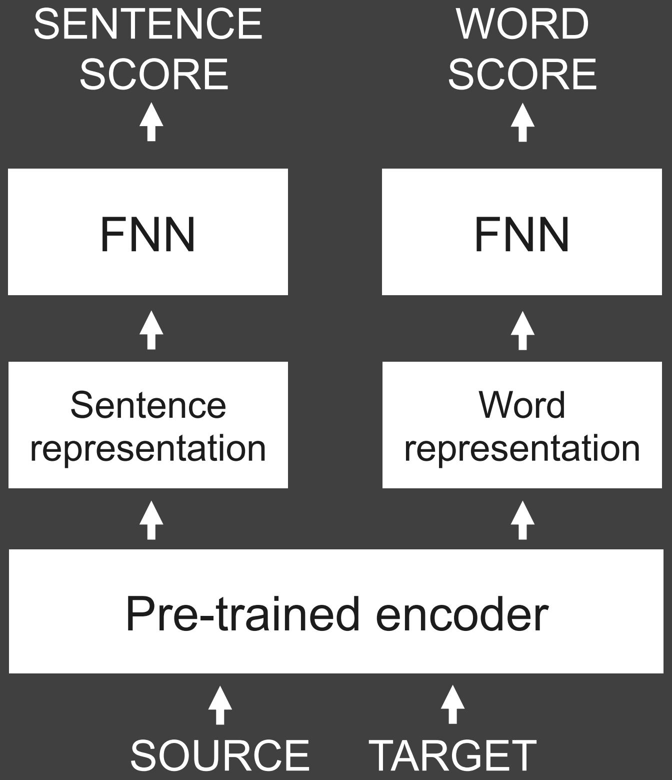
– Generate token embeddings with a pre-trained multilingual encoder (e.g., XLM-R)



Supervised QE models



Supervised QE models



LLMs for QE

– Prompt-based, few-shot approaches
– Aim to mimic human reasoning to calculate a score

````
(System) You are an annotator for the quality of machine translation. Your task is to identify
errors and assess the quality of the translation.

(user) {source_language} source:\n
```{source_segment}```\n
{target_language} translation:\n
```{target_segment}```\n
\n
Based on the source segment and machine translation surrounded with triple backticks, identify
error types in the translation and classify them. The categories of errors are: accuracy
(addition, mistranslation, omission, untranslated text), fluency (character encoding, grammar,
inconsistency, punctuation, register, spelling),
 locale convention (currency, date, name, telephone, or time format)
style (awkward), terminology (inappropriate  for context, inconsistent use), non-translation,
other, or no-error.\n
Each error is classified as one of three categories: critical, major, and minor.
Critical errors inhibit comprehension of the text. Major errors disrupt the flow, but what
the text is trying to say is still understandable. Minor errors are technically errors,
but do not disrupt the flow or hinder comprehension.

(assistant) {observed error classes}
````

Figure 1: The general prompt for GEMBA-MQM omits the gray part which performed subpar on internal data (we include it in GEMBA-locale-MQM). The "(user)" and "(assistant)" section is repeated for each few-shot example.

QE Performance

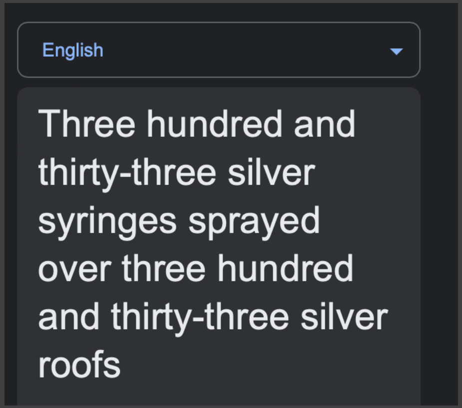
- Improved in recent years
  - In terms of correlations with human scores
  - Scale helps but bigger models not consistently better
- Performance varies by language pair
  - Not always better for high-resource language pairs
  - Possibly due to skew in data (MT too good)

Are we solving the right task?

– The motivation for QE is to filter translations for human review and edit

– How to interpret a sentence-level score?

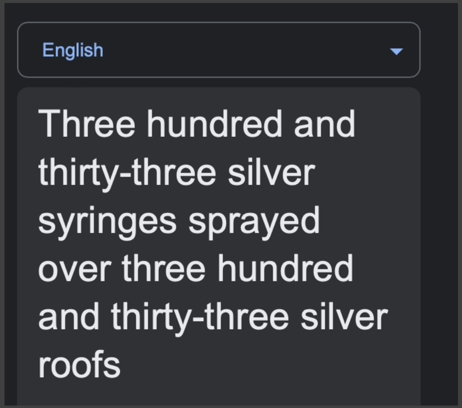
score = 5



Binary QE classification

- Instead of "How good?" ask "Good enough?"
  - Simple
  - Interpretable
  - Underexplored

label = NOT



Explainable QE

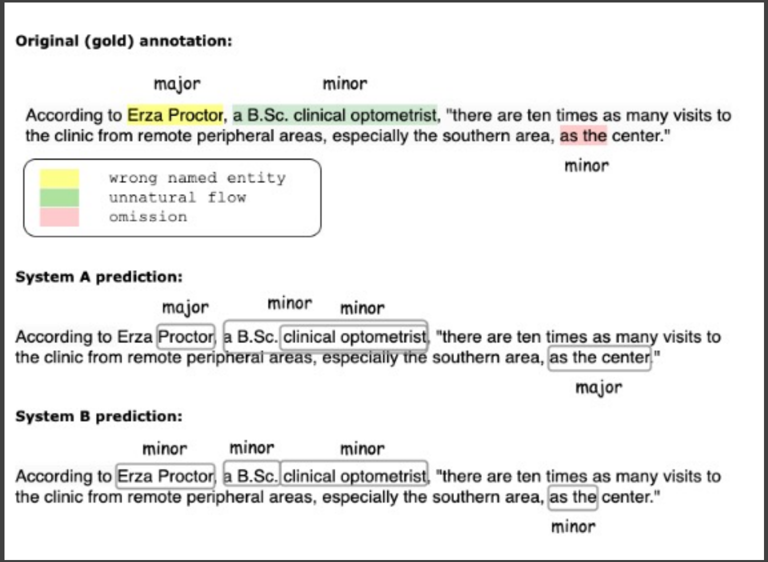
– Models trained to label error span and severity

– Desirable but not mature

– Winning submissions:

  - better recall than precision for spans

  - mostly output major (vs. minor) error labels



* from WMT QE 2023

QE in applied settings

- There is potential for QE to be useful
    - But is it asking the right question?
- Practical considerations:
    - compute vs performance trade-offs

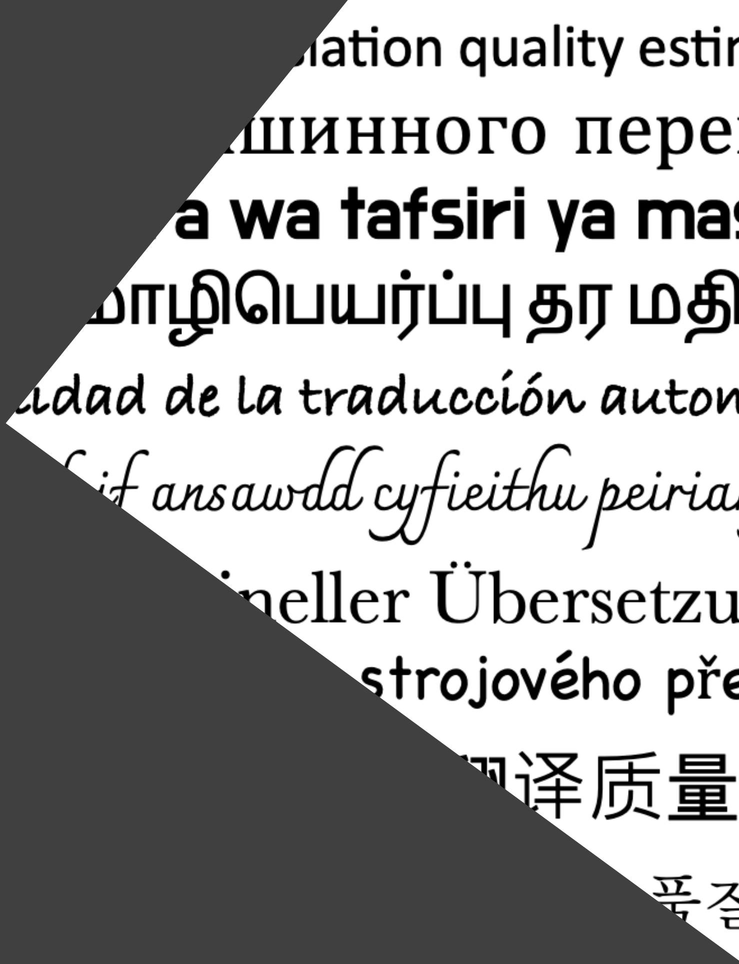
Need to start from a real-life use case:

- specify domain, language pair, errors, output granularity
- will probably find that the right data does not exist

Next steps

- Investigate binary QE classification
  - Critical Error Detection (CED)
- Using:
  - SOTA QE models and LLM(s)
  - Publicly available authentic and synthetic CED datasets

Any Questions?



Critical Error Data

Released by	Type	Data Source	Language pair	Sentences	% Errors
WMT QE 2021	Authentic	Wikipedia comments	English-Chinese	8,859	15.9
			English-Czech	9,476	17.3
			English-German	9,878	28.1
			English-Japanese	9,658	9.3
WMT QE 2022	Synthetic	news	English-German	173,291	5.8
			Portuguese-English	44,862	5.8
Unbabel MQM annotations 2022	Authentic	multi-domain	English-Russian	1,215	7.5
DEMETR dataset	Synthetic	multi-domain	Chinese-English	3,500	28.6
			Czech-English	3,500	28.6
			French-English	3,500	28.6
			German-English	3,500	28.6
			Hindi-English	3,500	28.6
			Italian-English	3,500	28.6
			Japanese-English	3,500	28.6
			Polish-English	3,500	28.6
			Russian-English	3,500	28.6
			Spanish-English	3,500	28.6

Table 7: Overview of datasets with critical error annotations. The data is primarily from WMT 2021 and 2022 CED tasks as well as the DEMETR dataset. Included are also the original Unbabel annotations of the WMT 2022 English-Russian MQM data. The table presents the number of sentences in each dataset that are publicly available and approximately what percentage of those are examples of critical errors. It also indicates the text domain and whether the errors are synthetic or authentic.

QE Data

Language Pair	Number of Sentences Train	Dev	Test	DA	PE	MQM	Data Source	Release
English-German	10,000	-	-	✓	✓		Wikipedia	2021/22
English-Chinese	10,000	-	-	✓	✓		Wikipedia	2021/22
Russian-English	10,000	-	-	✓	✓		Reddit	2021/22
Romanian-English	10,000	-	-	✓	✓		Wikipedia	2021/22
Estonian-English	10,000	-	-	✓	✓		Wikipedia	2021/22
Nepali-English	10,000	-	-	✓	✓		Wikipedia	2021/22
Sinhala-English	10,000	-	-	✓	✓		Wikipedia	2021/22
Pashto-English	2,000	-	-	✓	✓		Wikipedia	2021/22
Khmer-English	2,000	-	-	✓	✓		Wikipedia	2021/22
English-Japanese	2,000	-	-	✓	✓		Wikipedia	2021/22
English-Czech	2,000	-	-	✓	✓		Wikipedia	2021/22
English-Yoruba	1,010	-	-	✓	✓		Wikipedia	2021/22
English-Marathi	27,000	1,000	1,086	✓	✓		multi-domain/corpus	2022/23
English-Hindi	7,000	1,000	1,074	✓			multi-domain/corpus	2023
English-Gujarati	7,000	1,000	1,075	✓			multi-domain/corpus	2023
English-Tamil	7,000	1,000	1,067	✓			multi-domain/corpus	2023
English-Telugu	7,000	1,023	1,000	✓			multi-domain/corpus	2023
English-Farsi	-	-	1,000		✓		news (multi-domain)	2023
English-German	30,425	-	1,897			✓	multi-domain	2021/23
English-Russian	17,144	-	-			✓	multi-domain	2021/22
Chinese-English	36,851	-	1,675			✓	multi-domain	2021/23
Hebrew-English	-	-	1,182			✓	multi-domain	2023

Table 6: Overview of WMT 2023 QE data showing the number of sentences in the train, development and test datasets and the scoring scheme of DA, post-edits (PE) and MQM, adapted from Kocmi et al. (2023).