Measuring translation quality

Gábor Bessenyei

Globalese

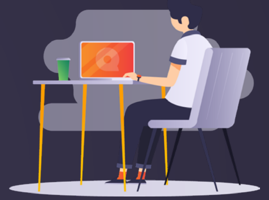
memoqfest

What is a good translation?

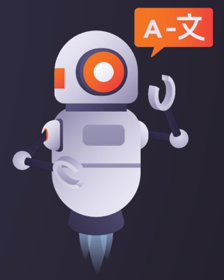
A good translation is any translation accepted by the customer

Reasons for measuring

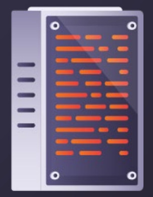
- Evaluating human translators



- Evaluating Machine Translation output



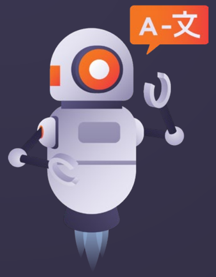
- Evaluating large training data sets



Measuring options

- Human evaluation

- Automated evaluation

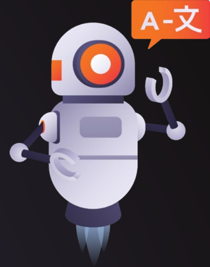


Human evaluation

- Performed manually by humans
  - Crowdsourcing
  - Professional translators
  - Subject matter experts

Human evaluation

- Frequently performed on a limited inspection lot
- Scoring is typically combining different weighted error types (terminology, style, correctness etc.)
- Implementation is typically highly company-dependent
- Rarely used due to time and budget restrictions



Automated evaluation

- MT quality evaluation
  - Based on references
  - Based on automated metrics like BLEU, COMET, Chrf, Meteor etc.
- MT quality estimation
  - No reference translation available
  - Typically based on language models, combined with metrics like MQM

MT quality evaluation – automated metrics

COMET

NIST

Bleu

HLEPOR

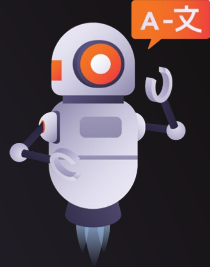
TER

RIBES

Levenshtein

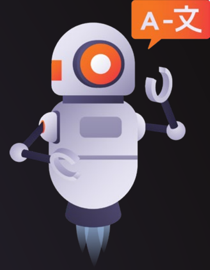
ChrF

F-Measure



MT quality evaluation – automated metrics

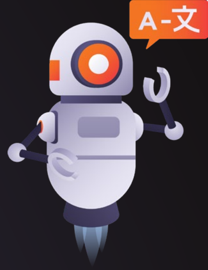
- The problems with automated metrics
  - They are only measuring the similarity to the reference
  - The do not tell anything about translation quality
  - They are not reflecting the effort required to spot and correct the translation issues
  - They are not standardized



MT quality evaluation – automated metrics

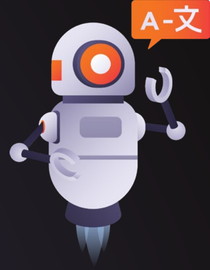
Source	In the event of such termination, Seller will immediately stop all work.
Reference (100%)	Im Falle einer solchen Kündigung hat der Verkäufer unverzüglich alle Arbeiten einzustellen.
NMT1 (84,69 chrF)	Im Falle einer solchen Kündigung wird der Verkäufer unverzüglich alle Arbeiten einstellen.
NMT2 (75,97 chrF)	Im Falle einer solchen Kündigung wird der Verkäufer alle Arbeiten unverzüglich einstellen.

memoqfest



MT quality evaluation – automated metrics

Source	Private use of a vehicle provided to an employee is a benefit in kind.
Reference (100%)	Die private Nutzung eines Fahrzeugs durch einen Mitarbeiter ist eine Sachleistung.
NMT1 (67,76 chrF)	Die private Nutzung eines Fahrzeugs, das einem Mitarbeiter zur Verfügung gestellt wird, ist ein Sachleistung.
NMT2 (78,40 chrF)	Die private Nutzung eines an einen Mitarbeiter bereitgestellten Fahrzeugs ist eine Sachleistung.

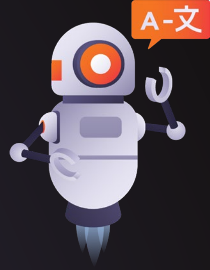


MT quality evaluation – chrF++ score: 98.41

You may not (i) use the Services in a way that infringes, misappropriates or violates any person's rights; (ii) reverse assemble, reverse compile, decompile, translate or otherwise attempt to discover the source code or underlying components of models, algorithms, and systems of the Services (except to the extent such restrictions are contrary to applicable law); (iii) use output from the Services to develop models that compete with OpenAI; (iv) except as permitted through the API, use any automated or programmatic method to extract data or output from the Services, including scraping, web harvesting, or web data extraction; (v) represent that output from the Services was human-generated when it is not or otherwise violate our Usage Policies; (vi) buy, sell, or transfer API keys without our prior consent; or (vii), send us any personal information of children under 13 or the applicable age of digital consent.

You may (i) use the Services in a way that infringes, misappropriates or violates any person's rights; (ii) reverse assemble, reverse compile, decompile, translate or otherwise attempt to discover the source code or underlying components of models, algorithms, and systems of the Services (except to the extent such restrictions are contrary to applicable law); (iii) use output from the Services to develop models that compete with OpenAI; (iv) except as permitted through the API, use any automated or programmatic method to extract data or output from the Services, including scraping, web harvesting, or web data extraction; (v) represent that output from the Services was human-generated when it is not or otherwise violate our Usage Policies; (vi) buy, sell, or transfer API keys without our prior consent; or (vii), send us any personal information of children over 13 or the applicable age of digital consent.

Source: https://openai.com/policies/terms-of-use



MT quality evaluation

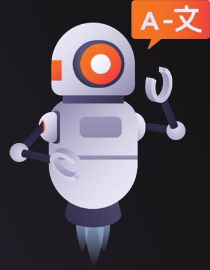
In drei Jahren möchte ich ein Arzt werden.

In three years, I want to be a doctor.

⇅

In three years, I want to be a translator.

memoqfest

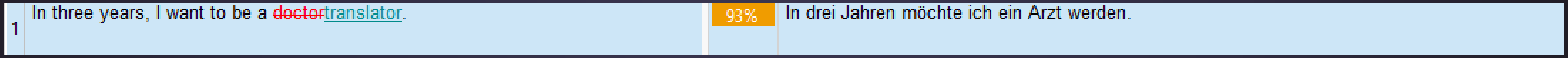


MT quality evaluation

In three years, I want to be a doctor.

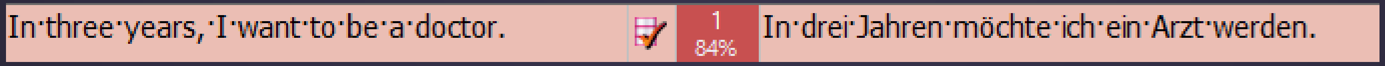
In three years, I want to be a translator.

93%



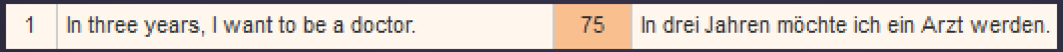
1 In three years, I want to be a ~~doctor~~translator. 93% In drei Jahren möchte ich ein Arzt werden.

84%

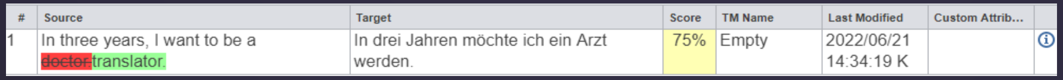


In·three·years,·I·want·to·be·a·doctor. 1 84% In·drei·Jahren·möchte·ich·ein·Arzt·werden.

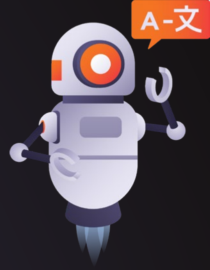
75%



1 In three years, I want to be a doctor. 75 In drei Jahren möchte ich ein Arzt werden.

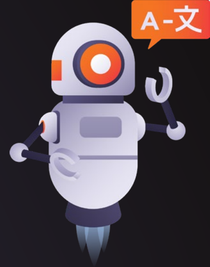


#	Source	Target	Score	TM Name	Last Modified	Custom Attrib...
1	In three years, I want to be a ~~doctor~~translator.	In drei Jahren möchte ich ein Arzt werden.	75%	Empty	2022/06/21 14:34:19 K	



MT quality estimation – use cases

- Automated training data evaluation
- Automated MT scoring (no Post Editing vs full Post Editing)



MT quality estimation

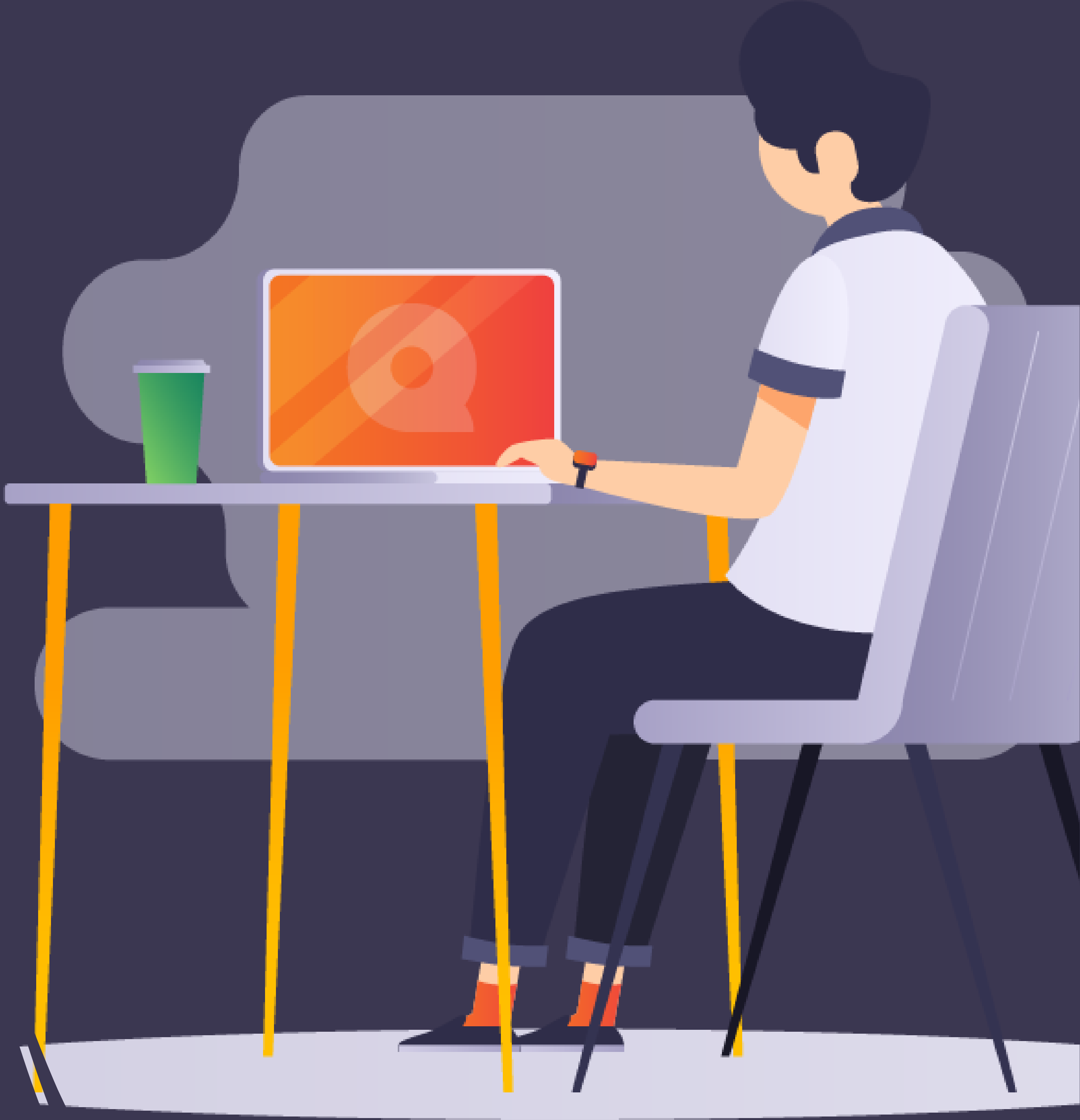
- Similar to quality evaluation but without any reference
  - If we could do good QE, we could do it already on MT level
  - QE models are telling only about the linguistic quality, but not really about the translation quality
  - Even if they are combined with other metrics like MQM, they are far from perfect
  - Currently we have a real problem that more and more machine generated content is flowing into the models

The problem

- The web and TMs are more and more polluted with raw neural MT
- Automated metrics are not working well
- Filtering out incorrect neural MT translations is currently nearly impossible
- We have a real problem that an increasing number of machine generated content is flowing into the models

The problem

- TMs are typically noisy
- Terminology is typically noisy
- QA on human reference is not on the radar



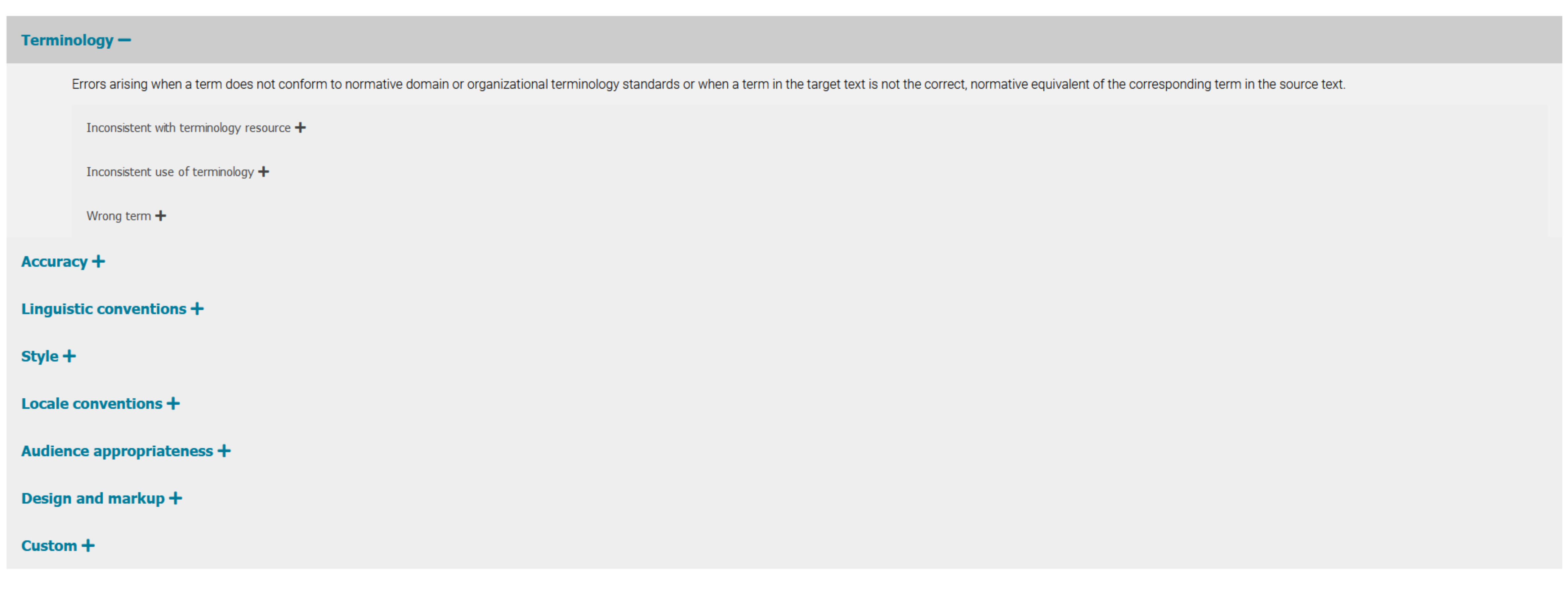
Measuring methodology

MQM (Multidimensional Quality Metrics)

- Multidimensional Quality Metrics (MQM) is a framework for analytic Translation Quality Evaluation (TQE). It can be applied to both human translation and machine translation.
- https://themqm.org/

memoqfest

MQM (Multidimensional Quality Metrics)



Terminology —

Errors arising when a term does not conform to normative domain or organizational terminology standards or when a term in the target text is not the correct, normative equivalent of the corresponding term in the source text.

- Inconsistent with terminology resource +
- Inconsistent use of terminology +
- Wrong term +

Accuracy +

Linguistic conventions +

Style +

Locale conventions +

Audience appropriateness +

Design and markup +

Custom +

memoqfest

More Quality Assurance, please!

Use the time and the money saved by MT for human QA

- On training data
- On glossaries
- With special focus on reference

memoqfest

Thank you!

Any questions?

gabor.bessenyei@globalese.ai