(Machine) Translation Evaluation — Human Evaluation

Multilingual NLP — lecture n°4

Guillaume Wisniewski

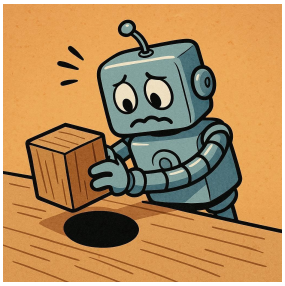
guillaume.wisniewski@u-paris.fr

UPCité Master in Computational Linguistics

October 15th, 2025

Part I

Warning



Evaluation is hard
Evaluation is (a big part of) your job

No language left behind

(NLLB Team et al. 2022)

- translation models and dataset for 200 languages
- a true *tour de force*
  ↪ "Driving inclusion through the power of AI translation"
  ↪ "Translating Wikipedia for everyone"
- `https://ai.meta.com/research/no-language-left-behind/`

Science Left Behind

"It is also a scientifically dubious work. In this article, I demonstrate that many of Meta AI claims made in NLLB are: unfounded, misleading, and the result of a deeply flawed evaluation. I will also show that, following Meta AI evaluation methodology, it is very easy to obtain even higher numbers than what they have reported."

`http://blog.benjaminmarie.com/science-left-behind.html`

But: no science without measuring

"When you can measure what you are speaking about, and express it in numbers, you know something about it, when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind; it may be the beginning of knowledge, but you have scarely, in your thoughts advanced to the stage of science."

Lord Kelvin, 1883

Why Evaluating (Machine) Translation Quality?

Practitioners and Users

- how to compare alternative systems, choices for building and customizing off-the-shelves systems?
- quantify the effectiveness of using MT

Researchers

- evaluate models
- train our systems!

⇒ intrinsic evaluation ≠ extrinsic evaluation

Usage scenario

Offline "benchmark" testing of MT engine performance

- Typical evaluation scenario is AI/ML: reproduce a reference output
  - ↪ imitation game
- sample representative test documents with reference human translation available
- measure "similarity" between output & expected output
  - ↪ not so easy for MT

Operational quality assessment at runtime

- MT engine is translating new source material
- Is the output "good enough" for the underlying application?
- reference-less evaluation / confidence estimation

Common usage of reference-based evaluation

On the same dataset

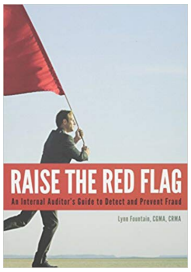
- compare two MT engines
- compare two versions of the same engine
  - ↪ before and after customizing the engine
  - ↪ before and after incremental development of the engine
  - ↪ during training

On different datasets

- compare MT engine performance
  - ↪ across domains or types of input data
  - ↪ on different sentence types, linguistic structures

Common usage of confidence evaluation

- identify / flag poorly translated segment during MT engine operation
  - ↪ can the translation be used as-it?
  - ↪ is it worth correcting the translation or is it better to translate it from scratch?
- Intuitive, promising approach
  - ↪ To my knowledge: no work shows practical value
  - ↪ Useful = saves translators' time, eases workload



Be careful

User expectations on MT depends on their knowledge of:

- the source language
- the target language

But...

Evaluation is also useful to:

- data filtering: filter large-scale parallel corpora with automatic quality assesments
- model alignment: optimize MT systems with quality metrics as reward/preference models
- quality award decoding: develop decoding strategies that incorporate quality assessments

An open question

- Can we use the same methods to evaluate human and machine translations?

↪ A natural idea

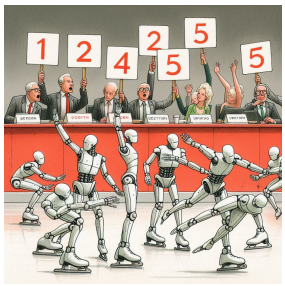
- However:
  1. Two separate communities that politely ignored each other for a long time
  2. MT evaluation developed when systems were so poor that evaluating them like humans made no sense — no longer true!
  3. Evaluation methods have been converging in recent years
- Open question: do humans and machines make the same types of errors?

Part II

Human Evaluation of MT

Early Evaluation of MT Sytems

- early days of MT: manual evaluation
- Two criteria:
  - Fluency: how natural and grammatical the translation is in the target language.
  - Adequacy: how well the translation preserves the meaning of the source text.
- both evaluated on a $[\![1, 5]\!]$ scale
- used in the ALPAC Report (1966)
- formalised in the 1990s for the DARPA and NIST MT evaluations (White, O'Connell, and O'Mara 1994)



- A simple idea, a complex implementation
- Bilingual or non-bilingual annotators
- Providing a reference translation
- Defining appropriate scales
- Better to rank translation hypothesis or to score them?
- ...

Fluency and Adequacy Scales

Adequacy	
1	no meaning
2	little meaning
3	much meaning
4	most meaning
5	all meaning

Fluency	
1	incomprehensible
2	disfluent English
3	non-native English
4	good English
5	flawless English

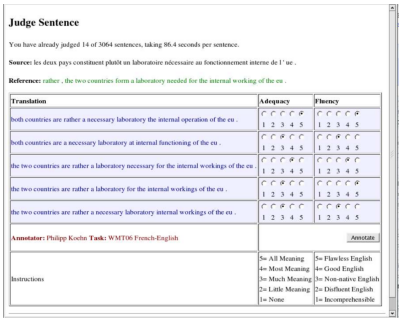
An Important — and Often Overlooked — Question

Who produces the annotations?

- MT experts
- (Professional) translators
- Crowd-workers

⇒ huge impact on evaluation quality (more on that latter)

With a picture:



Judge Sentence

You have already judged 14 of 3064 sentences, taking 86.4 seconds per sentence.

Source: les deux pays constituent plutôt un laboratoire nécessaire au fonctionnement interne de l ' ue .

Reference: rather , the two countries form a laboratory needed for the internal working of the eu .

Translation	Adequacy	Fluency
both countries are rather a necessary laboratory the internal operation of the eu .	1 2 3 4 5	1 2 3 4 5
both countries are a necessary laboratory at internal functioning of the eu .	1 2 3 4 5	1 2 3 4 5
the two countries are rather a laboratory necessary for the internal workings of the eu	1 2 3 4 5	1 2 3 4 5
the two countries are rather a laboratory for the internal workings of the eu .	1 2 3 4 5	1 2 3 4 5
the two countries are rather a necessary laboratory internal workings of the eu .	1 2 3 4 5	1 2 3 4 5
Annotator: Philipp Koehn **Task:** WMT06 French-English		Annotate
Instructions	5= All Meaning 4= Most Meaning 3= Much Meaning 2= Little Meaning 1= None	5= Flawless English 4= Good English 3= Non-native English 2= Disfluent English 1= Incomprehensible

Let's try it!

- **Source:** N'y aurait-il pas comme une vague hypocrisie de votre part?
- **Reference:** Is there not an element of hypocrisy on your part?
- **System 1:** Would it not as a wave of hypocrisy on your part?
- **System 2:** Is there would be no hypocrisy like a wave of your hand?
- **System 3:** Is there not as a wave of hypocrisy from you?

Measuring Annotator Agreement

The problem

- **Agreement:** the proportion of annotations on which two annotators agree.
  - ↪ "Agree" means they give the same label to the same item.
- Problem (from a statistician's point of view): even if two people label completely at random, they will sometimes agree by chance.
  - ↪ We therefore need a more precise measure of agreement.

Measuring Annotator Agreement

Cohen's κ

- Quantifies the extent to which two annotators agree on how a set of data items should be labelled.
  - ↪ Pre-defined set of labels ⊕ set of items to be annotated.
  - ↪ Each item is labelled independently by both annotators.
- Measures the proportion of items assigned the same label by both annotators.
- Adjusts for the fact that some level of agreement may occur purely by chance.

$$\kappa = \frac{P_a - P_e}{1 - P_e}$$

- ↪ P_a: observed proportion of agreement (actual agreement).
- ↪ P_e: expected proportion of agreement by chance.

The Landis and Koch 1977 interpretation scale

κ	interpretation
< 0.00	Poor agreement
0.00–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–1.00	Almost perfect

- The scale everyone uses — basically because it's the only one that exists.
- In short: "good enough" when $\kappa > 0.5$... if we're feeling optimistic.

Example (I)

- First annotator:

```
["B", "B", "B", "A", "C", "C", "B", "A",
 "A", "C", "C", "B", "B", "A", "C", "B",
 "C", "C", "C", "A"]
```

- Second annotator:

```
["B", "B", "C", "A", "C", "C", "C", "A",
 "A", "B", "C", "B", "B", "A", "C", "A",
 "B", "C", "C", "A"]
```

Example (2)

Contingency table:

	A	B	C
A	5	1	0
B	0	4	2
C	0	2	6

↪ how often the two annotators assigned the same or different labels to the same items.

Example (3)

$$P_a = \frac{5+4+6}{5+1+4+2+2+5} = \frac{15}{20} = 0.75$$

$$P_e = \sum_{c \in \{A,B,C\}} p_1(c) \times p_2(c)$$

After a rather tedious round of calculations:

$$P_e = 0.34$$
$$\kappa \simeq 0.62$$

Extensions & Challenges

Variants of agreement measures

- **Weighted** κ: takes into account the degree of disagreement.
- **Fleiss'** κ: generalises Cohen's κ to multiple annotators.
- **ICC (Intraclass Correlation Coefficient)**: suitable for continuous or ordinal data.

Limitations and practical issues

- High cost: requires multiple annotations per item, which may reduce the total number of annotated items.
- Interpretation can be difficult: what level of κ is "good enough" depends on the context.
- What to do when results are poor? (e.g., low agreement)
  - Improve annotation guidelines or training.
  - Refine label definitions.
  - Reconsider task complexity.

Consistency of judgments of MT

all figures are coming from WMT Evaluation Campaign reports

Inter-rater agreement	
	κ
2011	0.40
2012	0.33
2013	0.26
2014	0.37

Intra-rater agreement	
	κ
2011	0.58
2012	0.41
2013	0.48
2014	0.52

↪ very low in both cases

↪ Would the result be the same if the translations were better?

More Broadly: Limitations of Adequacy & Fluency Evaluation

1. Subjectivity and inter-annotator variance

↪ annotators interpret criteria differently

2. Scale bias

↪ some annotators are stricter or more generous

3. High correlation between fluency and adequacy

↪ redundant, inefficient measures

4. Poor sensitivity to small differences

↪ discrete scales fail to capture fine quality gaps

↪ unstable system rankings

Why "Good Translation" Is a Problematic Notion

- There is no single, universal definition of translation quality.
- Quality can refer to:
  - **Semantic**: preservation of meaning.
  - **Stylistic**: tone and register appropriateness.
  - **Functional**: effectiveness in achieving communicative goals.
- What counts as "good" depends on:
  - The **purpose** of the translation (legal, literary, technical...).
  - The **target audience** and its expectations.
  - The **socio-cultural context**.

⇒ Quality in translation is **context-dependent**, not absolute.

Translation Studies: Function Over Form

- From linguistic equivalence to **functional approaches**.
- **Skopos theory** (Vermeer 1978):
  - A translation is "good" if it fulfils its communicative function.
  - Multiple translations can be equally valid.
- Other influential frameworks:
  - Reiss and Vermeer 1984 — text type and function.
  - House 2015 — functional-pragmatic equivalence.

Key idea

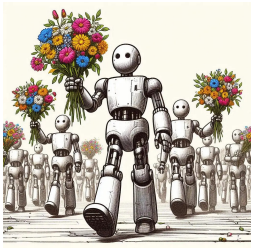
Quality ≠ fidelity to one "true" source text; it is evaluated against **purpose, context and audience expectations**.

Part III

Human Evaluation in the 21st Century

Direct Assessment (DA)

- introduced by Graham, Baldwin, Moffat, et al. 2013 and Graham, Baldwin, and Mathur 2015
- one simple, "global" question
  To what extent does this translation adequately convey the meaning of the source text?
- ↪ combines fluency and adequacy
- continuous evaluation on a 1–100 scale
- ↪ the annotator moves a slider to a value reflecting the perceived quality
- collect several independent judgements for each segment (typically 3 to 5)
- ↪ aggregate scores (median or mean)



The Revolution of Flowers (in MT)

Two Core Ideas

KISS

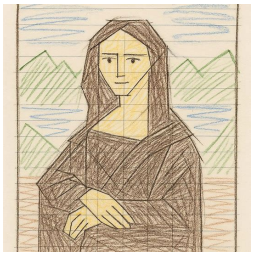
- keep it simple — no overcomplicating things
- give annotators as much freedom as possible

The Unreasonable Effectiveness of Data

- Alon Halevy, Peter Norvig, and Fernando Pereira
- more data > trying to control data quality
- simple idea: law of large numbers
- ↪ the average of many ratings = a better estimate of the true score

Because Things Shouldn't Be Too Simple

- Control annotation quality with:
  1. "trap" sentences (where the correct translation is obvious)
  2. perfect reference sentences inserted into the sample
- normalise scores per annotator to correct individual scale biases
- ↪ typically, a z-score is sufficient



Why DA Is the Answer to Everything

if only E. M. had studied machine translation...

1. Scale bias correction
   - ↪ scores are normalised per annotator (z-scoring)
   - ↪ neutralises strict vs. generous raters
   - ↪ enables cross-annotator comparability
   - ↪ improves aggregate robustness
2. Less inter-annotator variance
   - ↪ combining fluency and adequacy into one intuitive rating reduces criterion divergence
   - ↪ significant increase in κ (≥ 0.6)
3. Continuous scale ⇒ higher sensitivity
   - ↪ captures subtle quality differences
   - ↪ allows statistical smoothing
4. Alignment with real human perception
   - ↪ DA simplifies evaluation → one global, intuitive score

Limits

- no information about the nature of errors
- très souvent complété par
  - des annotations d'erreurs (MQM),
  - des comparaisons par paires

Part IV

Diagnostic-Based Evaluation

A New Source of Inspiration

- How are human translations evaluated?
- ↪ But surely: humans don't make mistakes!
- Core idea: annotate errors
- ↪ identify the error span (words involved)
- ↪ label describing the error type
- No numerical evaluation
- ↪ not of interest to translators!
- ↪ goal = understand errors to improve

Expériences de traduction de la parole vers le texte **LA-TL-INS, LA-TL-ICS** pour les journaux télévisés **LA-TL-ING, LA-TL-INS**

Nous présentons S2TT, un système intégré de traduction de la parole vers le texte **LA-TL-INS, LA-TL-ICS** basé sur POCKETSPHINX et MOSES. Il est comparé à différentes lignes de base **LA-TL-INS, TR-SI-TL** basées sur **LA-ST-AW, LA-ST-TA** ANTS — le système de transcription de journaux télévisés **LA-TL-ING, LA-TL-INS** développés par le groupe Speech du LORIA, MOSES et les outils de traduction de Google. Un petit corpus de transcriptions de référence de journaux télévisés **LA-TL-ING, LA-TL-IND** provenant de la campagne d'évaluation ESTER2 a été traduit par des experts humains pour l'évaluation. Le taux d'erreur sur les mots (WER) de l'étape de reconnaissance des deux systèmes est évalué, et BLEU est utilisé pour noter **LA-TL-NT, TR-TI-TT, LA-TL-INS** les traductions. En outre, les transcriptions de référence sont traduites automatiquement à l'aide de **LA-ST-AW** MOSES et de GOOGLE afin d'évaluer l'impact des erreurs de reconnaissance sur la qualité de la traduction.

A Key Tool: Error Typology

- List of possible errors
- Often organised hierarchically

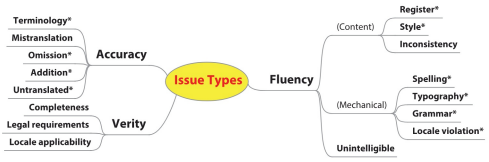
MQM

Principle

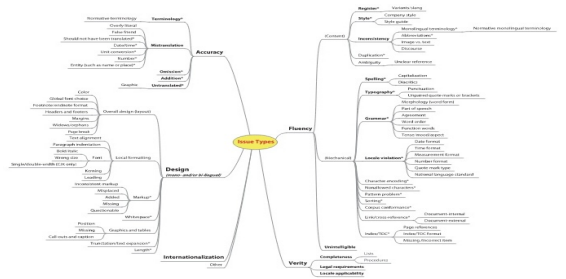
- Annotators do not give a global score,
- ↪ they identify and classify errors in the translation.
- Each error is annotated with:
  1. a category (e.g. Accuracy, Fluency, Terminology...)
  2. a more specific subtype (e.g. Omission, Mistranslation, Grammar...)
  3. a severity level (Minor / Major / Critical)
  4. the exact span in the translation (and often in the source)
- Annotations are then aggregated to calculate scores, often weighted by severity.
- (Lommel, Burchardt, and Uszkoreit 2014)

⇒ A diagnostic rather than purely evaluative approach.

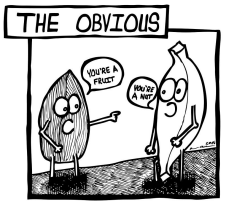
The MQM (Core) Typology



The MQM (Full) Typology



Stating the Obvious



- Error annotation requires experts trained in (i) translation and (ii) the use of this typology
- Time-consuming

Cost and Comparison with Direct Assessment

	DA	MQM
Type of judgement	Global score (0–100)	Detailed error annotation
Time per segment	15–30 s	2–5 min (or more for longer sentences)
Annotator training	Simple	Complex (categories, severity, calibration)
Analysis	Statistical (means, z-scores)	Qualitative + quantitative
Granularity	Global	Fine (specific error categories)
Cost	Low to medium	High

Example of Annotation with MQM

A Large-Scale Study of Human Evaluation for Machine Translation

(Freitag et al. 2021)

Research Question

- Observation: increasingly difficult to evaluate MT output
  ↪ higher quality ⇒ errors are finer and subtler
- Core idea: all evaluation relies on error identification
  ↪ error-based evaluation
  ↪ evaluate evaluation using MQM

Methodology

- Manual MQM annotation of WMT'20 English→German and Chinese→English system outputs
  ↪ largest MQM dataset ever collected
- Compare expert MQM rankings with crowd rankings (WMT) and automatic metrics

Main Findings

1. Expert MQM rankings can substantially differ from traditional crowd rankings (notably restoring a preference for human translation).
2. Some modern automatic metrics (especially embedding-based) align more closely with expert judgements than crowd evaluations.
3. Public release of a large MQM-annotated corpus to support comparison, supervised metric training, and further research.
4. In-depth analysis of error type distribution, inter-annotator disparities, contextual effects, and more.

Part V

Conclusion

The 3 steps of Human Evaluation in MT

Overview of evolution

Three steps:

1. 1990 → 2015: fluency & adequacy

↪ collected in various forms

2. 2015 → ... DA
3. 2020 → ... MQM

An important note

- same evolution in different (sub)fields of NLP
- increasing expertise in translation

Part VI

References

- Freitag, Markus et al. (2021). "Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation". In: *Transactions of the Association for Computational Linguistics* 9. Ed. by Brian Roark and Ani Nenkova, pp. 1460–1474.
- Graham, Yvette, Timothy Baldwin, and Nitika Mathur (May 2015). "Accurate Evaluation of Segment-level Machine Translation Metrics". In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Rada Mihalcea, Joyce Chai, and Anoop Sarkar. Denver, Colorado: Association for Computational Linguistics, pp. 1183–1191.
- Graham, Yvette, Timothy Baldwin, Alistair Moffat, et al. (Aug. 2013). "Continuous Measurement Scales in Human Evaluation of Machine Translation". In: *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*. Ed. by Antonio Pareja-Lora, Maria Liakata, and Stefanie Dipper. Sofia, Bulgaria: Association for Computational Linguistics, pp. 33–41.
- House, Juliane (2015). *Translation Quality Assessment: Past and Present*. Routledge.

- Landis, J. Richard and Gary G. Koch (1977). "The Measurement of Observer Agreement for Categorical Data". In: *Biometrics* 33.1, pp. 159–174.
- Lommel, Arle Richard, Aljoscha Burchardt, and Hans Uszkoreit (Dec. 2014). "Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics". In: *Tradumàtica: tecnologies de la traducció* 12. Ed. by Attila Görög and Pilar Sánchez-Gijón, pp. 455–463.
- NLLB Team et al. (2022). *No Language Left Behind: Scaling Human-Centered Machine Translation*.
- Reiss, Katharina and Hans J. Vermeer (1984). *Grundlegung einer allgemeinen Translationstheorie*. Niemeyer.
- Vermeer, Hans J. (1978). "Ein Rahmen für eine allgemeine Translationstheorie". In: *Lebende Sprachen* 23.3, pp. 99–102.
- White, John S., Theresa A. O'Connell, and Francis E. O'Mara (Oct. 1994). "The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches". In:

Proceedings of the First Conference of the Association for Machine Translation in the Americas. Columbia, Maryland, USA.