When the Gold Standard Isn't Necessarily Standard: Challenges of Evaluating the Translation of User-Generated Content

Lydia Nishimwe, Benoît Sagot, Rachel Bawden
Inria, France
{lydia.nishimwe, benoit.sagot, rachel.bawden}@inria.fr

his TOOOO funny!!

- A: Il est trop drôle !
- B: il est TROOOOP drôle !!
- C: il et TROOOOP drôle !!

Key Takeaways

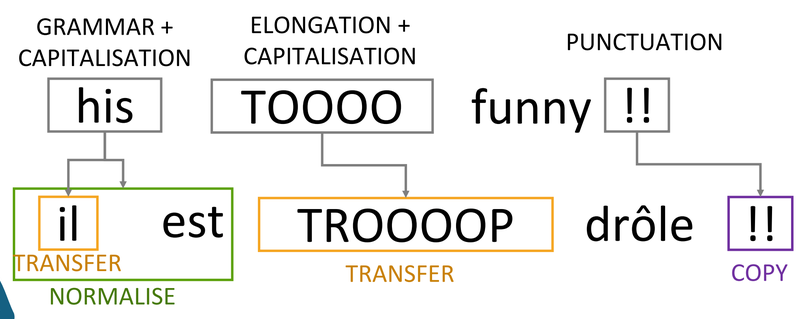
There is **no single gold standard** for UGC translation.

LLMs can be **steered** toward **different translation styles**.

Fair evaluation requires **guideline-aware metrics**.

RQ1 What is a "good" translation of user-generated content (UGC)?

Translation Guidelines



Method

Defining a **taxonomy of phenomena and actions** from translation guidelines of UGC datasets

Finding

Different datasets expect **different levels of standardness in the gold** translations

	Phenomenon	RoCS-MT	FooTweets	MMTC	PFSMB
1	Grammar	Normalise	Normalise	Normalise	Normalise
2	Spelling	Normalise	Normalise	Normalise	Normalise
3	Word elongation *(e.g., goooooaaaalllll)*	Normalise	Transfer	Transfer	Transfer
4	Capitalisation *(e.g., NOPE, SoRry)*	Normalise	Transfer	Transfer	Transfer
5	Informal abbreviations *(e.g., gonna, u)*	Normalise	Normalise	Normalise	Transfer
6	Informal acronyms *(e.g., LOL, TBH)*	Normalise	Normalise	Transfer	Transfer
7	Hashtags and subreddits	Copy	Copy	Transfer	Transfer
8	URLs, user IDs, and retweet marks *(RT)*	Copy	Copy	Copy	Copy
9	Emoticons and emojis	Copy	Copy	Copy	Copy
10	Atypical punctuation	Normalise	Copy	Copy	Copy
11	Overt profanity *(e.g., fuck)*	Transfer	Transfer	Transfer	Transfer
12	Self-censored profanity *(e.g., f*ck)*	Normalise	Normalise	Normalise	Transfer

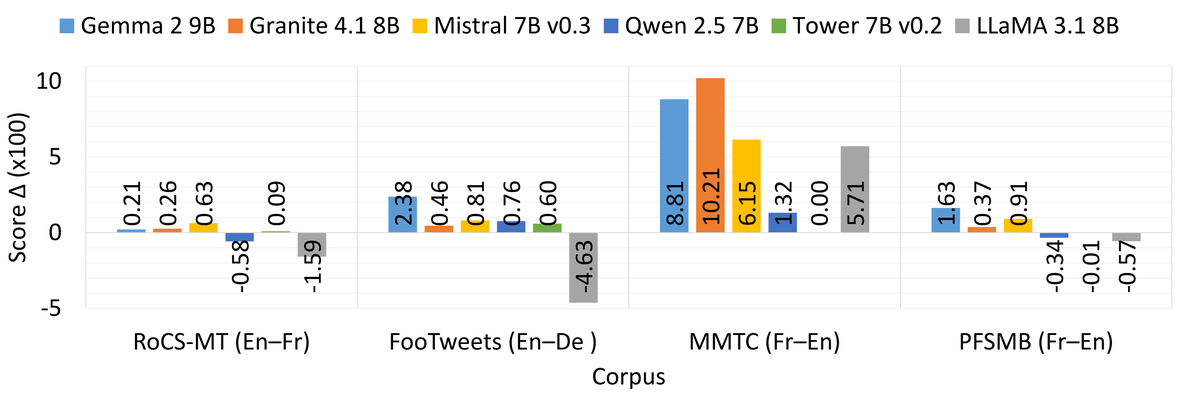
Highly standardising dataset ← → *Minimally standardising dataset*

RQ2 How can we ensure **fair evaluation** of **UGC translation**?

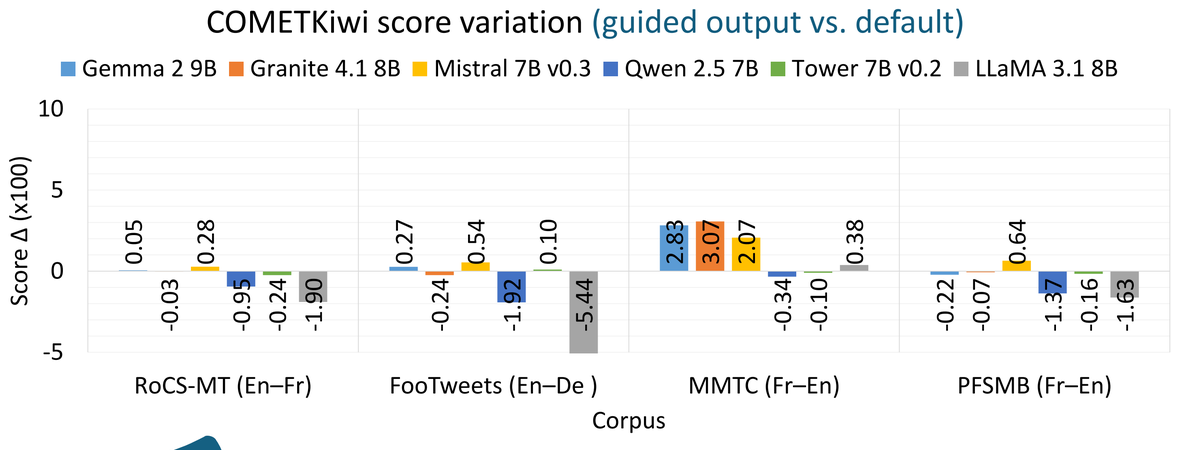
Method

Prompting LLMs to translate with the **same guidelines** as the UGC datasets

Reference-based Automatic Evaluation



Reference-less Automatic Evaluation



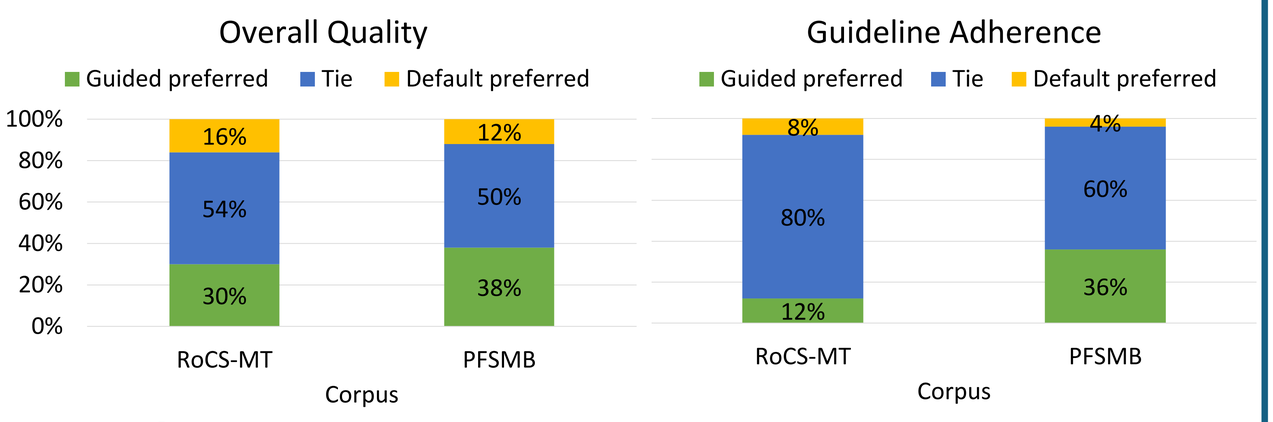
Qualitative Analysis *(Example from MMTC)*

Source	**@JulieTom62** même pas **besoins** de **regardé** le match 😂
Normalised source	***@JulieTom62*** *même pas* ***besoin*** *de* ***regarder*** *le match* 😂
Reference	**@JulieTom62** don't even need to watch the match 😂
Gemma's default output	Ø No need to even watch the game 😂
+ guidelines	**@JulieTom62** no even need to watch the game 😂

COMET
78.44
91.64 ↗

NORMALISE TRANSFER COPY OMIT (Ø)

Human Evaluation



Findings

- Most **LLMs follow** the UGC **guidelines** (Qwen follows them but hallucinates, Tower ignores them, LLaMA refuses to translate)
- Matching **guidelines** generally **improve** UGC **translation** but can introduce errors
- **COMET** aligns better with **guideline adherence**
- **COMETKiwi** aligns better with **quality judgments**

Future Work

- **LLM-as-a-judge** evaluation with **guideline-specific rubrics**
- More **compositional guidelines** and **prompting** strategies

References:

- *RoCS-MT: Robustness Challenge Set for Machine Translation* (Bawden & Sagot, WMT 2023)
- *FooTweets: A Bilingual Parallel Corpus of World Cup Tweets* (Sluyter-Gäthje et al., LREC 2018)
- [MMTC] *The Multilingual Microblog Translation Corpus: Improving and Evaluating Translation of User-Generated Text* (McNamee & Duh, LREC 2022)
- [PFSMB] *Comparison between NMT and PBSMT Performance for Translating Noisy User-Generated Content* (Rosales Núñez et al., NoDaLiDa 2019)
- *COMET-22: Unbabel-IST 2022 Submission for the Metrics Shared Task* (Rei et al., WMT 2022)
- *CometKiwi: IST-Unbabel 2022 Submission for the Quality Estimation Shared Task* (Rei et al., WMT 2022)

Full paper here →