CO32

LINGUISTS' PERCEPTIONS OF HUMAN- AND MACHINE-TRANSLATED CLINICAL OUTCOME ASSESSMENT (COA) WORDING: A MIXED METHODS STUDY

Oliver Delgaram-Nejad, PhD; Tim Poepsel, PhD; Payton Ramsey, MPH; Chryso Hadjidemetriou, PhD; Rebecca Israel, MS; Allyson Nolde, MPP; Rachael Browning, BA

AIMS:

The role of machine translation (MT) in linguistic validation (LV) is an emerging discussion (Vanmassenhove et al., 2019). Because LV is concerned with ensuring cultural and conceptual clarity, MT applications must appreciate linguistic and cultural nuances. Available evidence from our previous qualitative survey research with linguists and LV professionals suggests that MT applications are presently unsuited to COA translation contexts. In this study, we extended the qualitative results with an experimental task examining linguists' ability to distinguish machine-translated (MT) and human-translated (HT) phrases in a clinical outcome assessment (COA) context.

METHODS:

Participants reviewed human translated (HT; using LV methodology) and MT phrases outside of instrument context, based on low-complexity, culturally neutral source English COA phrases. These were randomized, balanced, and length-matched (see Table 1). Participants were also given text prompts: 'explain your choices' and 'list any relevant linguistic / cultural factors' and follow-up questions about 'the role of AI in LV' and 'performance in the experiment' (see Table 2).

Table 1: Example human- and MT-generated phrases

Source English	Human-Generated	Machine-Generated	Language
Please answer all questions.	يرجى الإجابة على جميع الأسئلة.		Arabic
I had problems with my sleep		كان لدي مشاكل مع نومي	Arabic
I felt unsupported by people	شعرت بأنني غير مدعوم من قبل الناس		Arabic
I felt I had nothing to look forward to		شعرت أنه ليس لدي ما أتطلع إليه	Arabic
Indicate the score that best fits with the patient status.	指出与患者状态最相符的分数。		Chinese-Simplifed
Rate the function independently from the nature of the signs.		根据标志的性质对功能进行评级。	Chinese-Simplifed
Mildly affected. No difficulties being understood.	轻度受影响。别人理解没有困难。		Chinese-Simplifed
Occasional food aspiration with choking more than once a week.		偶尔的食物渴望与窒息每周超过一次。	Chinese-Simplifed

Table 2: Text prompts given at follow-up

Question
Q1. Based on your experience of the exper-iment, what are your views on the role of AI in linguistic validation?
Q2. How do you feel about your performance on the experiment?

Table 3: Mean correct answers by language and respondent count

	Language	Respondents (n)	Mean Correct Score	SD
1	Arabic	64	44%	15%
2	Chinese	42	54%	15%
3	Finnish	22	45%	18%
4	German	50	52%	17%
5	Greek	18	49%	16%
6	Hindi	11	49%	16%
7	Russian	58	58%	16%
8	Spanish	98	42%	14%
9	Swahili	3	43%	12%
10	Turkish	35	43%	15%

References:

Vanmassenhove, E., Shterionov, D. and Way, A., 2019. Lost in translation: Loss and decay of linguistic richness in machine translation. arXiv preprint arXiv:1906.12068.

Table 4: Mean correct scores for highest- and lowest-difficulty questions

Low Accuracy Responses	Phrase Length	Mean Score	High Accuracy Responses	Phrase Length	Mean Score
Arabic – Q6	High	13%	Arabic – Q2	Low	75%
Arabic Q7	High	17%	Arabic – Q10	Low	78%
Chinese – Q3	Moderate	12%	Chinese – Q8	Moderate	79%
Chinese – Q6	Moderate	19%	Chinese – Q4	Moderate	98%
Finnish – Q6	Low	14%	Finnish – Q5	High	85%
Finnish – Q7	Low	14%	Finnish – Q4	High	95%
German – Q6	High	14%	German – Q10	High	80%
German – Q2	Moderate	28%	German – Q4	High	84%
Greek – Q1	Moderate	17%	Greek – Q4	Moderate	78%
Greek – Q7	Moderate	28%	Greek – Q3	High	83%
Hindi – Q9	High	9%	Hindi – Q7	Moderate	73%
Hindi – Q10	Moderate	36%	Hindi – Q2	Moderate	82%
Russian – Q1	Moderate	25%	Russian – Q8	Moderate	95%
Russian – Q5	Moderate	25%	Russian – Q2	Low	95%
Spanish – Q4	Moderate	4%	Spanish – Q5	Low	63%
Spanish – Q9	Moderate	8%	Spanish - Q8	Moderate	88%
Swahili – Q2	Moderate	0%	Swahili – Q3	Moderate	100%
Swahili – Q4	Moderate	0%	Swahili – Q7	High	100%
Turkish – Q1	Moderate	9%	Turkish – Q6	High	74%
Turkish – Q4	Low	11%	Turkish – Q10	Moderate	86%

Table 5: Summary of qualitative findings 1/2

Please explain your choices	Please list any relevant linguistic / cultural factors
• Fluency suggests HT • Very literal translation suggests MT • Syntax differentiates HT and MT • Grammar, tone, style, spelling, and punctuation errors might suggest MT	• MT lacks cultural and idiomatic insight • Language-specific linguistic factors sometimes ignored by MT

Table 6: Summary of qualitative findings 2/2

What are your thoughts on the role of AI in LV?	How do you feel about your performance in the experiment?
• LV requirements exceed MT capability • Humans required to post-edit/check MT output • MT may speed up routine translation tasks • MT may actually delay translation • Human translations provided in the experiment were at times too literal	• Task was difficult • Phrases were too short/simple for a valid test • Surprised by own (in)accuracy • Task instructions were unclear • Longer sentences were easier to distinguish

RESULTS:

In the post-experiment, participants expected fluency and naturalness to signal HT. They also expected over-literal translations, lack of idiomaticity, and technical (e.g., grammatical) errors to signal MT. Yet participants' (n=401, 10 languages) ability to distinguish MT from HT experimentally was variable and mixed (43-58%, SD:15-18%). It is notable, though, that accuracy increased for longer phrases, suggesting that accuracy depends on source-related factors (such as linguistic content, the presence of technical language or jargon, or terms that have an idiomatic basis in English that machine translation cannot adequately account for). Follow-up responses emphasized task difficulty as dependent on phrase simplicity (4% of n = 176) and the importance of human oversight of MT applications (58% of n = 176).This was the main result from the qualitative work. Participants were keen to emphasise that while MT may have a place in the LV process, it will need to be supervised and subjected to human quality control.

CONCLUSIONS

Qualitative feedback from linguists identified many linguistic factors that may distinguish MT and HT, while experimental task performance showed variable success in distinguishing low-complexity, culturally neutral MT and HT phrases. Higher success in some cases, and overall performance variability, signal that phrase content, length, and language identity may impact distinguishability. Linguists' assumption that technical errors signal MT may have caused over-literal HT to be mistaken for MT, and index both underlying distrust of MT and variable MT quality across languages. Further work with full instruments and more culturally specific COA content is planned. These findings underscore the need for caution with machine COA translation use cases.