Lessons from a User Experience Evaluation of NLP Interfaces

Eduardo Calò, Lydia Penkert, Saad Mahamood

e.calo@uu.nl

NAACL 2025 (Findings)

Universiteit Utrecht

Imagine having to go through this...

Motivations

- How **human evaluation** is **presented impacts** the quality of the **data collected** [3, 6]
- However this **aspect** is often **overlooked** by researchers
- Can result in **unreliable annotations** $\Rightarrow$ One factor **hindering reproducibility** [7]

What we do

- **Bridge** natural language processing (**NLP**) and human-centered design (**HCD**)
- **Evaluate** user interfaces (**UIs**) used in past NLP human evaluations based on **HCD principles** [4]
- Draw **convenient recommendations** for designing UIs for human evaluations

HCD Interaction Principles

Suitability for the user's tasks: *the UI supports the users in the completion of their tasks.*
Self-descriptiveness: *appropriate information is presented in the UI to make its capabilities and use immediately obvious.*
Conformity with user expectations: *the UI's behavior is predictable based on the context of use and commonly accepted conventions in that context.*
Learnability: *the UI supports the discovery of its capabilities, allows exploration, provides support, and minimizes the need for learning.*
Controllability: *the user maintains control of the UI and the interactions' speed, sequence, and individualization.*
Use error robustness: *the UI tolerates and assists the user in avoiding and recovering from errors.*
User engagement: *functions and information are presented in an inviting and motivating manner.*

UIs and Participants

Four UIs from papers in ReproHum:

1. "It's not Rocket Science: Interpreting Figurative Language in Narratives" [2] (**FL**)
2. "Data-to-text Generation with Macro Planning" [5] - study 1 (**MLBF**)
3. "Data-to-text Generation with Macro Planning" [5] - study 2 (**MLBC**)
4. "NeuralREG: An end-to-end approach to referring expression generation" [1] (**REG**)

Three user experience (**UX**) **experts** (7-16 years of professional expertise)

Evaluation Procedure

INTERFACE: `Link to the interface`

Are the following principles met?

Suitability: *Not met*, *Partially met*, *Met*

If you answered *Not met* or *Partially met*, why do you think the principle is not (fully) met?

Self-descriptiveness: *Not met*, *Partially met*, *Met*

If you answered *Not met* or *Partially met*, why do you think the principle is not (fully) met?

...

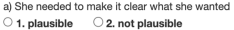
Results

IAA: from low to moderate (overall $\alpha = 0.339$) $\Rightarrow$ not surprising (highly subjective)

Ranking: REG $>$ FL $>$ MLBC $>$ MLBF

Examples of Flaws

MLBF - Controllability issue



FL - Self-descriptiveness issue

MLBC - Self-descriptiveness issue

Recommendations

Principle	Recommendations
Self-descriptiveness	• Avoid confusing/subjective/judgmental/technical/redundant language • Avoid long instructions, but if needed explain/present them properly • Explain any part that may turn out to be unclear
Conformity	• Ensure uniformity in layout (e.g., length of the input fields) • Use proper/consistent colors (e.g., brightness, palette, etc.) • Organize/structure and position text in the right way • Use the appropriate type of question based on the data you want to collect
Learnability	• Provide the right amount of examples • Explain the terminology • Give feedback • Explain how to interact with the system
Controllability	• Provide users with the ability to revisit the instructions • Enable empty state revert
Robustness	• Clearly mark mandatory information • Provide proper error messages (e.g., not too early, not persistent, not generic) • Check input data in the backend • Check if unwanted interactions with UI/text may occur • Avoid default answers that may be misleading (e.g., default value of a slider)
Engagement	• Add a progress bar • Do not use aggressive language (e.g., all-caps) • Avoid heavy text/content/tables • Give positive feedback after completion

Takeaway

RETHINK YOUR HUMAN EVALUATION INTERFACE!!!

References I

[1] Thiago Castro Ferreira et al. "NeuralREG: An end-to-end approach to referring expression generation". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Iryna Gurevych and Yusuke Miyao. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 1959–1969. DOI: 10.18653/v1/P18-1182. URL: https://aclanthology.org/P18-1182/.

[2] Tuhin Chakrabarty, Yejin Choi, and Vered Shwartz. "It's not rocket science: Interpreting figurative language in narratives". In: *Transactions of the Association for Computational Linguistics* 10 (2022), pp. 589–606.

[3] Jessica Huynh, Jeffrey P. Bigham, and Maxine Eskénazi. "A Survey of NLP-Related Crowdsourcing HITs: what works and what does not". In: *CoRR* abs/2111.05241 (2021). arXiv: 2111.05241. URL: https://arxiv.org/abs/2111.05241.

[4] ISO-9241-110. *Ergonomics of human-system interaction — Part 110: Interaction principles*. May 2020. URL: https://www.iso.org/obp/ui/en/#iso:std:iso:9241:-110:ed-2:v1:en.

References II

[5] Ratish Puduppully and Mirella Lapata. "Data-to-text Generation with Macro Planning". In: *Transactions of the Association for Computational Linguistics* 9 (May 2021), pp. 510–527. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00381. eprint: https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00381/1924176/tacl_a_00381.pdf. URL: https://doi.org/10.1162/tacl%5C_a%5C_00381.

[6] Jamar Sullivan Jr. et al. "Explaining Why: How Instructions and User Interfaces Impact Annotator Rationales When Labeling Text Data". In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz. Seattle, United States: Association for Computational Linguistics, July 2022, pp. 521–531. DOI: 10.18653/v1/2022.naacl-main.38. URL: https://aclanthology.org/2022.naacl-main.38/.

References III

[7] Craig Thomson, Ehud Reiter, and Anya Belz. "Common Flaws in Running Human Evaluation Experiments in NLP". In: *Computational Linguistics* (Mar. 2024), pp. 1–11. ISSN: 0891-2017. DOI: 10.1162/coli_a_00508. eprint: https://direct.mit.edu/coli/article-pdf/doi/10.1162/coli_a_00508/2348458/coli_a_00508.pdf. URL: https://doi.org/10.1162/coli%5C_a%5C_00508.