34

It's *Um* Hard to Process but Not Hard to Evaluate: Written Disfluency Affects Humans and Language Models Differently

Aida Tarighat[1], Patrick Sturt[2], Martin Corley[2]
[1]UKRI CDT in NLP, University of Edinburgh, Edinburgh, United Kingdom. [2]School of PPLS, University of Edinburgh, Edinburgh, United Kingdom

Abstract

Objectives: Written disfluencies, such as *um*, are rare in experimental research yet occur in everyday writing. They are intentional in text and can signal pragmatic cues, including nonliteral meaning such as sarcasm. Prior work has shown effects on human reading times and word predictions. Building on this, we examine their impact on human evaluation of statements and language model predictions.

Design: This study reports two strands of work: a psycholinguistic evaluation task assessing human judgements of written disfluency, and language modeling experiments building on prior cloze-based findings. While conducted separately, both strands use matched stimuli, enabling integration with earlier reading and prediction results.

Methods: In a statement evaluation task (n=101), participants rated sentences varying in fluency (presence vs. absence of *um*) and meaning (literal vs. nonliteral) for sarcasm, offensiveness, formality, and six basic emotions. Bayesian ordinal regression models analyzed human data. Parallel masked and causal language modeling experiments were run with BERTweet and Llama. Outputs were compared using Latent Semantic Analysis (LSA) cosine similarities with previously collected human predictions and model confidence scores.

Findings: Disfluency slightly reduced sarcasm ratings for nonliteral items. Sarcasm increased offensiveness ratings, with a small positive fluency effect and weak interactions. Fluency consistently reduced formality ratings. Emotion ratings showed sarcasm most affected disgust and anger, with disfluency dampening this impact. Models matched human predictions for fluent items but diverged when fluency changed.

Conclusions: This work informs language processing theory and applications in e-mental health and wellbeing services requiring nuanced, context-sensitive human–AI communication.