Evaluating writing in the age of AI: what’s left to assess?

Rurik Tywoniw

Department of Linguistics

University of Illinois at Urbana-Champaign

Contents

Selected history of writing assessment

- Teaching academic writing
- Direct assessment of writing
- Introducing ChatGPT to academic writing

Exploring academic writing with ChatGPT

- Data and Methods
- Move analysis
- Linear modeling
- Going beyond the automated essay

Academic Writing

Teaching

- Preparing students for academic, research-oriented writing (IMRD formats)
- Shifting to multimodal, intertextual writing formats – “Integrated Writing” (Plakans, 2013)
- Use of corpus resources to break down the writing process (Swales, 1990; Upton & Connor 2001)

Testing

- Templated writing: length ←→ quality (Fleckenstein et al., 2020)
- Shifting away from surface level language to focus on *cohesion, topical structure,* and *information within noun phrases*. (Huot, 1990)
- Moving beyond the essay as the sole measurement of academic literacy (Campbell & Latimer, 2012)

AI and machine learning in writing assessment

Machine learning in computational linguistics

- Research (classification systems, corpus analytics, text similarity)
- Machine Translation
- Natural Language Understanding (Speech to text)
- Automated Evaluation of Writing (Borade & Netak, 2021; Cahill & Madnani, 2018)
- Generation and curation of language test materials (Settles, LaFlair, & Hagiwara, 2020)
- Natural Language Generation (Dale, 2020) – in 2020 explained the emergence and remaining weaknesses of GPT-2

GPT-4, ChatGPT and beyond

- Based on large-scale statistical models and extensive corpus data.
- Openly available (with some limitations)
- Performs complex NLG tasks
  - Stronger with more templated evidence available online (**short essays**, poems, reviews, news stories, etc.)
  - Weaker to novel tasks (gibberish assignments, true dialog), tasks with little open access (certain legal genres, longer form writing), and simple tasks which would not be formally written (describing a room or spatial arrangements)
- Able to beat the Turing Test in certain circumstances (White, 2022)

Reaction to ChatGPT

- Debate on the place of ChatGPT in education (Seaton, 2023)
  - Develop security measures against ChatGPT in education
    - OpenAI tools
    - Turnitin
    - Tian’s GPTZero
  - Utilize ChatGPT in the classroom
    - Revision
    - Drafting/Generation
    - Novel tasks
  - Revisit the constructs of academic writing in light of other bygone dimensions of writing quality:
    - Handwriting – replaced by typesetting and word processing
    - Spelling – greatly aided by various spell checkers
    - Grammar checking – greatly aided by data-driven writing tools (Grammarly, etc.)

Context

UIUC’s English Placement Test (EPT)

- Places undergraduate and graduate students whose primary language is a language other than English into ESL writing courses
- 4 levels of quality (A, B, C, D), with profiles at the B and C level for stronger argumentation (profile 1) or stronger lexico-grammar (profile 2).
- Scores produced based on a single essay writing task.
  - Source-based
  - Structure guides provided (intro-body-conclusion, TEA/PIE)
  - Scoring rubric with descriptors for controlling idea (unity), paragraph structure, and lexico-grammar
- Test secured with Safe-Exam Browser
- Authentic writing?

Research questions

- How are ChatGPT essays evaluated based on the EPT scoring rubric?
  - What qualities lead to this scoring?
- To what degree can ChatGPT-generated and human-generated essays be distinguished by textual features?
  - What features are most predictive?
- What features are present in high-level human-generated essays that are missing from ChatGPT essays?
  - *Forthcoming*

Methods

Mixed Methods study

- 50 essays generated by ChatGPT from 25 recent EPT prompts
- 250 Essays produced by test-takers form the same prompts
- Qualitative analysis of AI-generated essays using the EPT rubric and Move Analysis (Swales, 1990)
- Logistic regression using automatically calculated linguistic features of essays
    - Lexical: CW frequency, Processing Time, Collocational strength
    - Morpho-syntactic: Inflections per word, derivations per word
    - In future studies: number of sentences, words per sentence, type-token ratio

Results: Q1

How are ChatGPT essays evaluated based on the EPT scoring rubric?

- ChatGPT essays were typically scored B2 (30), with some A (7) and C2 (13).
- Lexico-grammar was never an issue (even prompting errors or lower-level language proved impossible)
- Controlling ideas could be weak or absent.
- Explanation, analysis, or connections to introduction and controlling idea could be absent or weak.
- Paragraphs often shorter than human-generated essays
- Source is never acknowledged, and paraphrasing level varied widely

Sample ChatGPT response (using source texts)

“

The use of cashless currency has been an increasingly debated topic, and its proponents argue that it eliminates the costs of printing and transporting cash, facilitates tax compliance, reduces the risk of robbery, and is more efficient. However, others argue that it could create problems, such as connectivity and infrastructure issues, and leave vulnerable populations at risk of financial abuse.

…

A cashless economy could also lead to a reduction in physical crimes, as cash plays a critical role in motivating crime such as robbery and assault. *According to a study by Roubini ThoughtLab*, **consumers currently lose approximately $5.8 billion per year due to cash-related crimes, an average of about $8.40 per person**. *In a cashless economy, people would not carry cash with them, which would decrease the risk of physical assault and robbery.*

…

In conclusion, the move to a cashless currency has its advantages and disadvantages. While it eliminates costs, facilitates tax compliance, and reduces the risk of physical crimes, it could create connectivity and infrastructure issues and leave vulnerable populations at risk of financial abuse. Ultimately, the decision to transition to a cashless economy should be based on a careful evaluation of the costs and benefits and a consideration of the potential impact on all members of society. ”

Sample ChatGPT response (w/o source texts)

“

The world has been transitioning towards cashless currencies for quite some time now, with the proliferation of mobile payments, digital wallets, and cryptocurrencies. While this shift may bring many advantages, it's essential to consider whether the world is truly ready for a cashless currency. In my opinion, while there are clear benefits, there are also significant challenges that must be addressed before a cashless currency can become a reality.

…

Another challenge with cashless currencies is that they require an established regulatory framework. Governments must be able to monitor and regulate digital transactions to prevent money laundering, terrorist financing, and other illegal activities. This means that governments must invest in technology and cybersecurity infrastructure to ensure that digital currencies are secure and reliable.

…

In conclusion, while a cashless currency can offer many advantages, it's essential to consider whether the world is ready for such a transition. The benefits of cashless currencies must be balanced against the challenges of the digital divide, cyberattacks, and privacy concerns. Ultimately, a successful transition to a cashless currency will require careful planning, investment in technology, and the establishment of a robust regulatory framework to ensure that the benefits are fully realized while the risks are minimized. ”

Discussion

What is left to assess?

- ChatGPT’s formulaic, templated written production gives structured rhetoric on just about any topic.
    - Certain aspects of academic writing conventions are still slighted
    - For now, ChatGPT tends to fabricate sources (in rare cases) or over-rely on given ones.
    - ChatGPT cannot generate knowledge, and in its current interface, cannot react to images or non-text data structures.
    - ChatGPT exhibits difficulty connecting topics to a new topic.

Discussion

What is left to assess?

- We can work around the edges of ChatGPT's current capabilities.
    - Focus on summary/note writing from sources and fact checking generated texts.
    - Paragraph completion tasks given the missing argumentative components of ChatGPT.
    - Writing summaries or explanations of data and figures remains a targetable skill.
    - Extending instruction and assessment to evaluate originality, creativity, and style. These types of qualities are not currently considered baseline assessable constructs in writing assessment, but can be investigated for value in specific fields.
    - Ideally, ChatGPT becomes a better writing tool that no longer mimics humans, but writes in a more AI-literate style. This can be approached by incorporating other data besides text likelihood and semantic analysis into the algorithm, including information from eye-tracking.

References

Borade, J. G., & Netak, L. D. (2021). Automated grading of essays: a review. In ***Intelligent Human Computer Interaction: 12th International Conference,*** IHCI 2020, Daegu, South Korea, November 24–26, 2020, Proceedings, Part I 12 (pp. 238-249). Springer International Publishing.

Campbell, K. H., & Latimer, K. (2012). ***Beyond the five-paragraph essay.*** Stenhouse Publishers.

Dale, R. (2020). Natural language generation: The commercial state of the art in 2020. ***Natural Language Engineering, 26***(4), 481-487.

Fleckenstein, J., Meyer, J., Jansen, T., Keller, S., & Köller, O. (2020). Is a long essay always a good essay? The effect of text length on writing assessment. ***Frontiers in psychology***, ***11***, 562462.

Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. ***Review of Educational research, 60***(2), 237-263.

Madnani, N., & Cahill, A. (2018, August). Automated scoring: Beyond natural language processing. In ***Proceedings of the 27th International Conference on Computational Linguistics*** (pp. 1099-1109).

Plakans, L. (2013). Writing integrated items. In ***The Routledge handbook of language testing*** (pp. 263-275). Routledge.

Seaton, B. (2023). Is Penn about to be caught flat-footed by ChatGPT? ***Daily Pennsylvanian.*** Retrieved April 3, 2023, from https://www.thedp.com/article/2023/02/chat-gpt-education-changes-curricula-artificial-intelligence

Settles, B., T. LaFlair, G., & Hagiwara, M. (2020). Machine learning–driven language assessment. ***Transactions of the Association for computational Linguistics, 8***, 247-263.

Swales, J. (1990). ***Genre analysis: English in academic and research settings***. Cambridge university press.

Upton, T. A., & Connor, U. (2001). Using computerized corpus analysis to investigate the textlinguistic discourse moves of a genre. ***English for specific purposes***, ***20***(4), 313-329.

White, L. (2022, November 14). OpenAI's GPT-4 consistently beats The Turing Test. ***Stealth Optional***. https://stealthoptional.com/news/openai-gpt-4-beats-the-turing-test/

UNIVERSITY OF
ILLINOIS
URBANA-CHAMPAIGN