THE PERFECT POET?

How humans evaluate AI-generated poetry

7-papers assignment by Anna Sivera van der Sluijs

With the rapid advances in Artificial Intelligence (AI), it is still generally believed that computers cannot be creative. It is said creating art is an inherent human quality or even the thing that makes us human. One of these forms of art, which depends heavily on human emotion and expression, is poetry. However, creating artificial poets is exactly what researchers in the field of computational creativity attempt to achieve. Poetry generation is a difficult type of creative text generation, therefore it may be hard to imagine that these machine-generated poems could come close to the passionate works of art of us humans. Surprisingly, evidence suggests that this is already the case (Gunser et al., 2022; Hopkins & Kiela, 2017; Köbis & Mossink, 2021; Lau et al., 2018; Wang et al., 2021; Wang et al., 2016; Wu et al., 2020). Moreover, poems created by an AI author are surpassing professional human poets regarding human assessment (Hopkins & Kiela, 2017; Wang et al., 2016).

Poetry is a form of literary art that requires a specific and complex semantic structure (depending on the genre). Besides this, the author needs to understand the symbolic meaning of language and be able to use it in creative ways, which makes it difficult to generate artificially (Gunser et al., 2022). Poetry follows the rules of a language, using spelling and grammar to fluently write lines and verses. In addition, it uses rhythm, meter and a consistent theme to elicit affection (Wang et al., 2016).

In an attempt to create high-quality generated poetry, researchers use AI-based text generation. This AI uses deep learning models that are trained on an existing dataset to generate original outputs. Depending on the selected training material, it can recreate anything from contemporary English poetry (Köbis & Mossink, 2021) to traditional Chinese poetry (Wang et al., 2016). Adaptions or additions to the models can be made to adhere to specific rules for genres of poetry like limericks (Wang et al., 2021) or sonnets (Lau et al., 2018). Within the research field of poetry generation, the output of the AI is often intrinsically tested or compared with other text generation models. With the fast improvement of these models, it is increasingly more relevant to involve human evaluators and researchers are paying more attention to this in their research (Wu et al., 2020).

It is interesting to consider if an AI could ever reach the level of a professional human poet. This can be tested with (a variant of) the Turing Test. The test is based on the assumption that if a human continuously cannot differentiate human-written poems from generated poems, the AI has reached the human level of creating poetry. It might be surprising that with the current quality of AI-generated poems, humans already are incapable of reliably distinguishing professional human-written poems from generated poems (Gunser et al., 2022; Hopkins & Kiela, 2017; Köbis & Mossink, 2021; Lau et al., 2018; Wang et al., 2021; Wang et al., 2016; Wu et al., 2020). Often, poems by AI authors are falsely classified as human poems (Gunser et al., 2022; Hopkins & Kiela, 2017). In one study by Hopkins & Kiela (2017) a generated poem was even rated most humanlike, in another study by Wang et al. (2016) a generated poem ranked highest in quality among all human-written and AI-generated poems evaluated.

However, there are certain variables that influence the ability of humans to differentiate the AI from the human author. Lau et al. (2018) found that non-experts (i.e. crowd workers) often relied on rhyme and could not distinguish the poems, but a literature expert easily could. In the research carried out by Köbis & Mossink (2021) participants in a variant of the Turing Test could not differentiate between the AI or human author when humans selected the best poems out of a group of AI-generated poems. However, when random chosen generated poems were presented next to human poems, humans were successful in determining which was which. The kind of human poems used in the test matters too. If generated poems were shown besides classic poems, they were more often rightly classified relative to a comparison with newly written poems for the study, although professionals wrote these new poems (Gunser et al., 2022).

Despite the clear results from the Turing Tests, humans state they prefer human-written poems over generated poems and do not entirely approve of the computer as an artist (Köbis & Mossink, 2021). Furthermore, the quality of AI-generated poetry may not entirely be up to par compared to the work of human poets. AI-generated poems generally underperform in readability, emotion (in particular humour), overall quality and competence of the writer (Hopkins & Kiela, 2017; Lau et al., 2018; Wang et al., 2021). Interestingly though, these assessments can differ between cultures. In a study with American and Chinese subjects, the Chinese participants rated AI-generated poetry higher in imaginativeness, empathy, quality and competence than human-written poems (Wu et al., 2020).

It seems it is not too difficult for aspiring AI poets to master basic language skills and the semantic rules of poetry (Lau et al., 2018), but conveying emotion proves to be a challenge (Wu et al., 2020). For now, computers are still reflecting the given input without actually understanding what poem they are creating or its meaning. Regarding this, it is hard to imagine computers as truly creative. As humans, we can create poems that poetically describe our personal experiences, feelings and passions, which computers cannot do… yet.

References

Gunser, V. E., Gottschling, S., Brucker, B., Richter, S., Çakir, D., Gerjets, P., Gunser, V. E., Gottschling, S., Brucker, B., Richter, S., Çakir, D., & Gerjets, P. (2022). The pure poet: How good is the subjective credibility and stylistic quality of literary short texts written with an artificial intelligence tool as compared to texts written by human authors? *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*. https://escholarship.org/uc/item/1wx3983m

Hopkins, J., & Kiela, D. (2017). Automatically generating rhythmic verse with neural networks. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, *1*, 168–178. https://doi.org/https://doi.org/10.18653/V1/P17-1016

Köbis, N., & Mossink, L. D. (2021). Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. *Computers in Human Behavior*, *114*, Article 106553. https://doi.org/10.1016/J.CHB.2020.106553

Lau, J. H., Cohn, T., Baldwin, T., Brooke, J., & Hammond, A. (2018). Deep-speare: A joint neural model of poetic language, meter and rhyme. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, *1*, 1948–1958. https://doi.org/https://doi.org/10.18653/v1/P18-1181

Wang, J., Zhang, X., Zhou, Y., Suh, C., & Rudin, C. (2021). There once was a really bad poet, it was automated but you didn't know it. *Transactions of the Association for Computational Linguistics*, *9*, 605–620. https://doi.org/https://doi.org/10.1162/tacl_a_00387

Wang, Q., Luo, T., & Wang, D. (2016). Can machine generate traditional Chinese poetry? A Feigenbaum Test. *Advances in Brain Inspired Cognitive Systems*, *10023*, 34–46. https://doi.org/https://doi.org/10.1007/978-3-319-49685-6_4

Wu, Y., Mou, Y., Li, Z., & Xu, K. (2020). Investigating American and Chinese subjects' explicit and implicit perceptions of AI-generated artistic work. *Computers in Human Behavior*, *104,* Article 106186. https://doi.org/https://doi.org/10.1016/J.CHB.2019.106186