Journal of Artificial Intelligence Research 71 (2021) 175-189 Submitted 12/2020; published 06/2021

Viewpoint: AI as Author – Bridging the Gap Between Machine Learning and Literary Theory

Imke van Heerden ivanheerden@ku.edu.tr *Department of Comparative Literature CSSH, Ko¸c University, Istanbul, Turkey*

Anil Bas anil.bas@marmara.edu.tr *Department of Computer Engineering Faculty of Technology, Marmara University, Istanbul, Turkey*

Abstract

Anticipating the rise in Artificial Intelligence's ability to produce original works of literature, this study suggests that literariness, or that which constitutes a text as liter ary, is understudied in relation to text generation. From a computational perspective, literature is particularly challenging because it typically employs figurative and ambigu ous language. Literary expertise would be beneficial to understanding how meaning and emotion are conveyed in this art form but is often overlooked. We propose placing experts from two dissimilar disciplines – machine learning and literary studies – in conversation to improve the quality of AI writing. Concentrating on evaluation as a vital stage in the text generation process, the study demonstrates that benefit could be derived from literary theoretical perspectives. This knowledge would improve algorithm design and enable a deeper understanding of how AI learns and generates.

1. Introduction

The surge in Artificial Intelligence (AI) research in the world today is remarkable. Machine learning experts predict that AI would have 'written' a New York Times best seller by the year 2049 (Grace et al., 2018; Hall, 2018). The field of computational creativity has been identified as the next frontier in AI research (Colton & Wiggins, 2012) and holds intriguing implications for the literary industry. Algorithms capable of generating natural language (Gatt & Krahmer, 2018) could potentially transform the way we sell, read and review books.

Studies in computational creativity concentrate on identifying the core elements of cre ative forms (such as literature, visual art and music) from an algorithmic perspective, with the aim of replicating or stimulating human creativity (Turner, 2014; Besold et al., 2015; Veale et al., 2019). Similarly, natural language generation is interested in the production of realistic text (Reiter & Dale, 1997) Within this field, the generation of writing that might be considered literary (such as poetry, prose and

drama) is an active research do main, ranging from human-in-the-loop (K¨obis & Mossink, 2021) and machine-in-the-loop approaches (Clark et al., 2018) to fully automated systems. The aim is to generate creative texts that are indistinguishable from those written by humans (Singh et al., 2017; Xu et al., 2018; Chandu et al., 2019; Gero & Chilton, 2019; He et al., 2019; Chakrabarty et al., 2020; Moreno-Jim´enez et al., 2020; Zhai et al., 2020).

In this article, we argue for the inclusion of literary scholars in the development of machine learning models to improve the quality of generated text. To be clear, we are specifically interested in recent deep learning approaches that do not draw on literary ex pertise, including those of Jain et al. (2017), Li et al. (2018), Loller-Andersen and Gamb¨ack (2018), Wei et al. (2018), Xu et al. (2018), Yang et al. (2018), Yi et al. (2018), Chen et al. (2019), Jhamtani et al. (2019), Liu et al. (2019), Yeh et al. (2019), Zugarini et al. (2019), Agarwal and Kann (2020). Although some studies ask human judges (with varying back grounds in literature) to evaluate their output, these evaluators are not involved in the development of evaluation criteria nor the models in general.

This paper's concern with evaluation is an illustration of possible insights that may arise through such interdisciplinary collaboration. Specifically, we rethink the conceptualisation of weakness in creative language models, which is addressed by way of an example of the state-of-the-art language model GPT-2 (Radford et al., 2019a). Furthermore, we examine the role of fluency, coherence and readability in the evaluation of generated poetry, touching on other features such as ambiguity, figurative language and originality as well. This article suggests that a network of researchers from literary studies and machine learning could work together to create a shared language between disciplines with vastly different methodologies. According to Beatie (1979), "[o]nly when computer people learn, for example, to write readable prose and literary critics learn to understand the language of measurement can the 'computer revolution' in literary studies really begin".

2. Literary Theory

For this work, we define literature simply and traditionally as imaginative or invented writing, like fiction (Eagleton, 2011). Literature is a broad term that involves many different genres (or categories) of texts: poems, short stories, novels, dramatic works, letters, essays, film scripts and speeches (Todorov & Lyons, 2007). A common misconception is that literary interpretation is purely subjective or merely based on a person's intuition (Hirsch, 1967; Wolfgang, 1978; Richards, 2017). Every good work of literature follows a certain logic and uses specific techniques to communicate meaning to readers. As an example, Shakespearean texts are so complex that the bard's use of literary devices (or tools of writing) are examined and debated to this day (Powell, 1980; Vickers, 1995; Purcell, 2010). Although every reader will interpret a story in a slightly different way because of their unique background, certain elements in the text point in the same direction.

Not to be dismissed as personal and unscientific – though this is a topic of much dis cussion (Beatie, 1979) – scholarly deliberations on the substance of literature have a long standing history. Literary theory[1]is a well-established field that includes various competing scholarly approaches, each with its theoretical positions and commitments (Culler, 1997). It involves the systematic examination of literary texts to understand how they work (i.e. con vey meaning and give rise to particular interpretations) and why they are deemed literary, valuable or 'good'.

1. Note that this article concentrates on literary theory. The nature of the relationship between linguistics and poetics is subject to scholarly debate (Jakobson, 1960). Although literature may be studied lin guistically (Fabb, 1997), literary studies and linguistics should not be conflated (Cameron, 2011). For example, to understand “the work of literature” (Attridge, 2015) one must consider artistic merit, which requires more than a linguistic approach to literature.

As there are various types of methods to train machine learning models, there are various types, or schools, of literature and literary theory. Each theory is an arrangement of principles or a collection of ideas that helps define and explain particular categories of writing. Within the field of literary studies one might, for example, examine genre conventions or the ‘rules’ of these categories, asking which features define a book as narrative nonfiction, or how does one structure a detective novel, or what makes a Shakespearean sonnet successful.

3. Beyond Form

Several types of texts have been generated by deep neural networks, including but not lim ited to financial data (Plachouras et al., 2016), news content (Carlson, 2015), advertisements (Wang et al., 2019), film scripts (Sharp & Goodwin, 2016) and lyrics (Potash et al., 2015). Demonstrating what AI has managed to achieve, Deep-speare (Lau et al., 2018) captured the attention of the press, giving rise to the brainteaser in Table 1, “AI or not AI: that is the question” (Firth, 2018). (Capitalisation and end-of-line punctuation were removed.)

Stanza 1	Stanza 2
with joyous gambols gay and still array let those who are in favour with their stars no longer when he twas, while in his day of public honour and proud titles boast at first to pass in all delightful ways whilst I, whom fortune of such triumph bars around him, charming and of all his days unlook’d for joy in that I honour most	
Stanza 3	Stanza 4
o, call not me to justify the wrong shall i behold him in his cloudy state that thy unkindness lays upon my heart for just but tempteth me to stop and pray wound me not with thine eye but with thy tongue a cry: if it will drag me, find no way use power with power, and slay me not by art from pardon to him, who will stand and wait	
Shakespeare: Stanzas 2 and 3; Deep-speare: Stanzas 1 and 4	

Table 1: Excerpts from sonnets by Shakespeare and Deep-speare (Lau et al., 2018).

Readers not familiar with Elizabethan English found the automatic compositions nearly indistinguishable from their human-written counterparts. At a surface level, the quatrain structure and vocabulary do ‘look’ Shakespearean. The texts are written in iambic pentame ter and have a discernible rhyme scheme (though not strictly that of a sonnet). Presumably given the rhyme scheme as well as grammatical errors, such as “he twas” in Stanza 1, the texts were unable to convince an English literature expert (Lau et al., 2018). Form aside, the expert (co-author of the related study) mentioned that he was able to clearly distinguish the output because of its low emotional impact and readability (Lau et al., 2020).

Lau et al. (2018) show the importance of expert evaluation for poetry generation, and argue that future work should focus on moving beyond form. Instead, it could attend to

the complex interconnection of form and feeling (Freeman, 2009). According to Brooks and Warren (1976), a poem is not an assembly of mechanically combined elements (e.g. rhyme and meter) like a wall composed of bricks. Rather than concentrating on an element in isolation, we should concentrate on the relation between elements, on how they work

together to communicate meaning and emotion to the reader, i.e. create a poetic effect (Brooks & Warren, 1976). In other words, we should move beyond generating texts that 'look' literary (by mimicking formal properties) to texts that are literary (Veale, 2013).

In the next two sections, we extend this idea through a problematisation of priorities in text generation. First, centring on evaluation as an example of an important stage in the creative text generation process and, then, exploring possible avenues of collaboration, we seek to demonstrate that benefit could be derived from literary theoretical perspectives.

4. Example Gap: Evaluation

To improve the quality of computer-generated literature, we suggest combining tools and insights from various text-centred (rather than biographical, cultural or socio-historical) ap proaches in literary theory. Involving true expertise on what literature is and how it works would strengthen current research on learning-based text generation systems. The issues currently faced in machine-generated writing could be addressed more effectively by ap plying literary theoretical understandings of creativity, originality, ambiguity and emotion, among others.

However, efforts at collaboration between the "two cultures" face serious obstacles (Ham mond et al., 2013). The greatest challenge lies in explaining key devices, theories and techniques in both fields clearly enough so that AI experts and literary theorists could understand, without oversimplifying the complexity of the research. This type of communi cation would require the summarisation of literary concepts like the aesthetic and emotive qualities of poetry, which could lead to a loss of meaning. Scholars need to find more ways for the two disciplines to talk to one another, without losing essential information. It is difficult to incorporate qualitative results into algorithm design, and the primary question arises of how to combine essential literary concepts with data.

Equally, it is challenging to evaluate machine-generated text. How are creativity and originality to be measured? Evaluation is a crucial practical tool in the development process (Jordanous, 2012) but challenging in computational creativity "given the subjectivity and the lack of a 'right answer' to be achieved by creative systems" (Jordanous, 2017).

4.1 Quantitative vs Qualitative Evaluation

It might seem *de facto* to assess a language model quantitatively (Manurung et al., 2012; Jain et al., 2017; Alikaniotis & Raheja, 2019) or get a credibility score from surveys (Xie et al., 2017; Solaiman et al., 2019; Jhamtani et al., 2019). However, Da (2019) identifies a "fundamental mismatch between the statistical tools that are used and the objects to

which they are applied". Indeed, the question of creativity may be obscure and frustrating (Bown, 2014) and requires qualitative research to evaluate the evaluation process itself (H¨am¨al¨ainen & Alnajjar, 2019).

Qualitative evaluation, in comparison to quantitative methods, has been viewed as in consistent, unsystematic and therefore less effective (Lawrence, 1993). However, the differ ence between the two methods lies in that "the logic of qualitative evaluation is grounded in a willingness to accept ambiguity, rather than being wedded to a 'horse race' mentality in which the [approach] with the highest gain score is the winner" (McLeod, 2011). This makes it the ideal approach to literature, given its propensity for resisting straightforward

explanation. In fact, literary language has long been considered in respect of "its deviations from or distortions of ordinary language" (Bennett & Royle, 2016). Gross (1997) empha sises the importance of "novel uses of language" to literature and explains that "the kinds of insight [literary texts] provide are qualitatively different from those of pragmatic texts".

4.2 Ordinary vs Literary Language

Whereas developers have succeeded in training deep networks to produce coherent text, they are tested by the complex meaning that is characteristic of literature. Why is literature a challenging medium? The language used in, for instance, a newspaper report or an instruction manual is clear and simple. Every word or sentence is factual and generally has only one meaning. The language used in literature, on the other hand, can be very different and requires suitable evaluation criteria.

4.2.1 Departure from Norms

The difference between literary and ordinary language is a central theoretical concern (Le ung & Durant, 2018). Arguments have been made in favour of the distinctiveness of literary language (Fabb, 2010). From a formalist perspective, deviate or deformed language "makes poetry poetry and not a weather report" (Rivkin & Ryan, 2017). Jakobson (1923) fa mously described literary language as "organised violence committed on ordinary speech". In the modernist sense, poetry may be thought of as a language laboratory, a space of experimentation with language itself (Korg, 1979).

Gruber (1988) views originality as a constituent of creativity, which he associates with a deliberate departure from norms. Literary techniques include asyntactic structure (showing no syntactical rules or regularity), anastrophe (deliberate change of word order), anadiplosis (repetition for special effect) and ambiguity (discussed in 4.2.3). Considering, for instance, stream of consciousness and surrealist techniques, further examples are the absence of formal features (such as rhyme, meter and punctuation) and typographic experiments – see "l(a" (Cummings, 1991) and "In a Station of the Metro" (Pound, 1913). The argument that creative endeavours "must deal not with the predictable and repeatable – the stuff of normal science – but with the unique and unrepeatable" (Gruber, 1988) presents an occasion to rethink conceptualisations of weakness in language models, returning us to the vital question of evaluation.

4.2.2 Rethinking Weakness

OpenAI provides an example of weakness in their language model GPT-2 (Radford et al., 2019a): at times, it generates failures such as *fires happening under water* (Radford et al., 2019b). If the aim is to produce clear, informative text, this topic would be unsuitable. However, if read figuratively, the notion of fires happening under water is rather intriguing from a literary perspective. The primary point is: what might be considered a weakness in a standard factual text could be considered a strength in a creative text. If poetry, or any other literary form, is thought to typically bend or even break the rules of ordinary speech, what appears to be rule-breaking in AI-generated writing is not necessarily a failure.

4.2.3 On Poetry

Poetry is a typically dense and polysemous form of literature that may employ ambiguous and abstract language and, as a result, offer interpretive difficulties (Fabb, 2010). Simply put, because of figures of speech (such as metaphor), a poem may say one thing but mean another (Riffaterre, 1978). Literary scholars frequently pay attention to ambiguity in texts (Bennett & Royle, 2016). Empson (2004) defines ambiguity as “any verbal nuance, however slight, which gives room for alternative reactions to the same piece of language”, stating that its “machinations [...] are among the very roots of poetry”.

Poetic texts have been described as ambiguous, confusing, elusive, inaccurate, incorrect, peculiar, unreliable, unclear and uncertain (Bennett & Royle, 2016). Fabb (2015) explains that “[s]ome poetry, including traditional poetry, is for social or aesthetic reasons intended to be difficult”. Moreover, “[d]ifficulty can be part of the aesthetic of the text, either because it must be solved or in some cases because it is unsolvable, and this produces its own effects”.

As a compelling example, Hopkins and Kiela (2017) state that evaluators found their generated poems to be more humanlike than those actually written by humans. The study succeeded in generating high-quality rhythmic verse. To evaluate their results, the re searchers conducted an indistinguishability test. In the selection of human-written texts, prosodic elements were favoured. The findings underscore the importance of rethinking current evaluation criteria: although this is not explored in their study, half of the group of human evaluators misjudged the writing of Dickinson, Dryden, Tennyson and Shakespeare as AI-generated.

Shakespeare’s “A Fairy Song” received the lowest human likeness score, which could be related to unfamiliarity with Shakespearean English. However, as another instance, Dickin son’s “I’m Nobody” was misjudged as well. Whether the evaluation results would coincide if judges were presented with only contemporary literary works, i.e. written in present-day English, is open to discussion. Nonetheless, the results might suggest that participants mistook difficulties and peculiarities as flaws, i.e. an indication of AI. (Metaphor, which is common in poetry, could also be read as an error if interpreted literally.) Moreover, it reveals a misunderstanding of the nature, workings and purpose of poetry. Literary perspec tives could be useful in investigating the functions of ambiguity, peculiarity and complexity regarding text generation and evaluation.

A recent review of human evaluation criteria in natural language generation identifies prevalent categories of evaluation: fluency, coherence and readability, among others (Van der Lee et al., 2021). These categories are appropriate for standard factual text generation. Creative text, on the other hand, may have different aims than informative writing. In the former, as has been suggested, the purpose could be to depart from norms and defamiliarise (Shklovsky, 1917), to make difficult and strange, i.e. take readers out of

their comfort zones and inspire new insights and emotions. Human poets use various tools and techniques to do so, which may be erroneously read as a sign of AI. We agree that fluency, coherence and readability are important, however using strict adherence to rules as an indication of human likeness is not necessarily effective. It follows that the prioritisation of these evaluation categories in creative text generation might be counter-productive if it loses sight of ambiguity, complexity, peculiarity and polysemy. Perhaps, these are qualities that

current evaluation frameworks seek to eliminate. We believe that their presence in especially generated poetry is not problematic but essential. Understanding that literary language may at times “tremble on the edge of meaning” (Bennett & Royle, 2016) and pose a deliberate challenge to interpretation also poses a challenge to evaluation in creative text generation.

5. Plans to Bridge the Gap

Following our discussion of how evaluation could benefit from literary expertise, this section explores general opportunities for collaboration.

First, bridging the gap involves a conversation on strengthening text generation algo rithms using literary theory. Specifically, it requires the development of methods for distill ing abstract literary concepts and ideas into a practical technical register for use in language models. Instrumental theories in literary studies should be identified, systematically ren dered and tested with the ultimate aim of improving the quality of machine-generated literature.

Second, we suggest establishing theoretical principles to examine and evaluate computer generated literary texts, which would be useful as the quality of AI writing advances. These principles could improve the capabilities of creative AI. The results would be of benefit to literary scholars as well, given the likelihood of seeing, in the coming decades, literature authored by AI on bookstore shelves – which, presumably, will be studied in the future.

Third, we need to consider possible ways in which the automation of creativity might transform the literary industry as well as the academic study of literature, and how scholars and professionals might contribute to or prepare themselves for these changes. Each of the following questions on the impact of AI on the literary industry has the potential to develop into fully-fledged conversations:

- Will we see the rise of AI publishing houses or AI departments within publishing houses? How would these operate and what legal and ethical challenges would they face (for example, potential plagiarism and copyright infringement)? Van der Weel (2015) highlights the need for new definitions of authorship and intellectual property rights concerning technological development.
- How will the job market be affected (Zanzotto, 2019)?
- How will AI literature be monetised? Will the developer receive royalties – or will AI literature transform the industry by making books freely available?
- To what extent will AI and human creators collaborate and what shape will this assume? Examples of literary works that have already been co-created by humans and algorithms include a theatrical play by THEaiTRE (Rosa et al., 2020), the horror story generator ShelleyAI (Yanardag et al., 2017), a poetry collection that reimagines the classics (Hsieh, 2019), the novella *The Day a Computer Writes a*

Novel (Sato, 2016) as well as an experimental emulation of Jack Kerouac's *On the Road* titled *1 the Road* (Goodwin, 2018).

- How and to whom will credit be given? For instance, if a work of AI writing is awarded a literary prize, will the developer be the one accepting it? Similarly, will the developer be held accountable for possible expressions of hate speech?

Scholars have already flagged up AI's tendency to generate text with bias (Caliskan et al., 2017), including gender bias (Bolukbasi et al., 2016; Hendricks et al., 2018), racial prejudice (Schlesinger et al., 2018) as well as anti-Semitic language and discrimination against people with disabilities (Guo et al., 2019). Jones (2018) provides an overview of

legal responses to algorithmically generated defamatory and hate speech content.

Fourth, concerning higher education, we need to reflect on the effects of AI literature on literary studies as a discipline, including questions of authorship, literature and its role in society. Significant questions are as follows:

- What impact will AI literature have on definitions of originality and creativity? Ac cording to Klebanov and Madnani (2020), there is currently no operational scoring system that prioritises originality in generated text, and "once various indicators of originality can be successfully measured, additional work may be necessary to incorpo rate these measurements into scoring ecosystems". Traits and measurement criteria of originality have yet to be determined in computational linguistics (Klebanov & Mad nani, 2020) and have, outside this context, been considered unachievable (Gruber, 1988). In a literary context, Gross (1997) explains that attempts at categorisation may fail to do justice to the uniqueness and power of poetry and, therefore, require great dexterity.

- Will we judge this kind of literature by an entirely different set of criteria? What new theoretical perspectives can we expect as AI writing increases in sophistication?

- Will AI-authored texts be seen as inferior? Will it ever be taken seriously? According to Colton and Wiggins (2012), "[i]t seems that people allow their beliefs that machines can't possibly be creative to bias their judgement on such issues".

- Will AI writing always be read comparatively, i.e. in comparison to human writing?

- Will AI-generated text appeal to and, in actual fact, be read by readers? Regarding the ultimate marginality of the hypertext novel, Mangen and Van der Weel (2017) identify a "mismatch between theorists' predictions and readers' neglect". It is important to keep this mind as it could easily happen to AI as well.

- What will the impact be on disciplinary boundaries?

Will we see the development of more humanities-computer science courses? (For example, creative writing courses could start teaching the implementation of AI writing tools.)

6. Conclusion

Ng (2017) states that AI is the new electricity. Collaboration between areas of knowledge on this subject is inevitable. According to Potter (1991),

> we [computer scholars] must connect ourselves – through theory – to the larger world of thought that we live in. As a developing discipline, we have for too long lived in a tinkering "let's-try-this-and-see-what-happens" mode. [. . .]. Brute analysis without an elegantly and elaborately structured sense of why we are doing what we are doing leads to assertions that *do not matter*.

This study recommends drawing on expertise in the humanities, primarily literary the ory, to contribute to the development of computer science, specifically AI writing. To achieve human-level creativity, machine-generated literature has to overcome various obsta cles, such as ambiguity, emotional impact, poetic effect and storytelling. Engaging with the scholars that specialise in the building blocks of imaginative writing – literature's codes, if you will – would allow AI researchers to better determine the present shortcomings of machine-generated literature and explore how structural elements jointly convey meaning and emotion. Bridging the gap between machine learning techniques and literary theory could guide future analyses toward developments that *matter*.