GAN and GPT-2 neural networks,
worn words and creativity, namely literary second-hand

Inez Okulska

ORCID: 000-0002-1452-9840

> ‘Everyone writes, but nobody reads’, ‘if you want to write well, first you have to learn the classics, see what the others have created’; these or similar statements have been with writing enthusiasts for a long time. And when I say “old”, I mean really long centuries, because when literary translation did not emancipate itself into a particular genre, it often served as the basis for creative activity. Among the classics, one could call on Jan Kochanowski, Łukasz Górnicki or Adam Mickiewicz, for example, who, inspired by reading others, sat down to create their own works by more or less literally drawing on foreign, already existing works[1].

Then, how does creativity relate to it?

Without a doubt, we learn about existing works to get inspired; that is, to soak ourselves in patterns, ready-made components, even of the slightest measure, and make new ones out of them. We get to know the existing pieces in order not to euphorically discover something which has been already long ago discovered long ago. Then, finally, we explore existing works in a specific critical context, placed in trends and tendencies, in order to gain taste and distinguish between a work of art and a wreck, between creativity and creature. In this type of reflection, there is always a hopeful question about the possibility of creating something completely new, about the definitions and boundaries of this innovation and, of course, about connections with other artworks.

In the era of art with new and even more recent media, Julia Kristeva’s intertextuality is now replaced by interpoetics, where not only the text, but also the structure, mechanism and algorithm can be a quotation. But even on the level of “usual” human writing, where the material

[1] This has already been mentioned by many, but recently also by Małgorzata Łukaszewicz in her book. Małgorzata Łukasiewicz, *Pięć razy o przekładzie* (Kraków: Karakter, 2018)

is language, its words, phrases, syntax, Mikhail Bakhtin has dispelled all hope: almost every word in the work is its own incarnation, the sum of contexts and meanings it has gained earlier, in other pieces, on different lips. A masterpiece of literature is therefore most frequently an exceptional mixture of old elements, not to say 'hackneyed' ones:

> Each statement is associated with "one language". (with centripetal forces and tendencies) and at the same time with social and historical plurilingualism (centripetal forces, which stratify language). It is the language of the day, of the epoch, of the social group, of the species, of the direction, etc. (...) Every concrete word (statement) finds the object to which it is directed, always, so to speak, already debated, discussed, judged, hidden in a distant fog or, on the contrary, approximated by the light of the words spoken about it earlier. This object is entangled and permeated by common thoughts, views, other people's judgments and accents. A word turned to this object integrates into this dialogically activated and moved environment of other people's words, evaluations and accents, integrates into their complex mutual relations, resembles one another, contrasts with the other, and even crosses with the other[2].

If the substance, from which we are to create, is limited and still in use, then it appears that 'creativity is not an integral activity, but a relative one. We can be creative within our culture and within our frame of reference[3]', as the famous mathematician Marcus de Sautoy wrote in his reflections on digital art. In the 1970s, Stanley Fish already pointed to the limitation of creativity associated with immersing oneself in a particular set of cultural references, speaking of interpretative communities[4] that can be explained as follows:

> We are what the school, peer group, local authorities, lectures given to us, views recognized by influential representatives of our profession, etc., will make us. Therefore, the protocols of our individual interpretation strategies are always written by the hand of the interpretation community with which we identify ourselves[5].

It is true that we are only referring to interpretation, i.e. reception, but the same applies to the process of creation, because when we create, we constantly refer to poetics already taught (and expected). The relation, the position taken towards these poetics - approval, contraposition, reduction, contamination, transformation, etc.; that is, all the 'complicated mutual relations' that Bakhtin wrote about - is basically creative.

GAN (Generative Adversarial Networks[6]) is an ideal digital interpretation of creativity defined as creative movement in a specific area of patterns and expectations. The architecture of this type of network implies the existence of a creative duo, from which one network, like an internal child, free of any predefined framework of assumptions, starts to create. This act means a few stages

[2] Michaił Bachtin, *Problemy literatury i estetyki*, transl. Wincenty Grajewski (Warszawa: Czytelnik, 1982): 102–103

[3] 'Creativity is not an absolute but a relative activity. We are creative within our culture and frame of reference' Marcus du Sautoy, *The creativity code. How AI is learning to write, paint and think* (London: Fourth Estate, 2019): 13.

[4] Stanley Fish, *Interpretacja, retoryka, polityka*, ed. Andrzej Szahaj, introduc. Richard Rorty, transl. A. Szahaj (Kraków: Universitas 2002): 63

[5] Leszek Drong, 'Od konwencjonalizmu do normatywizmu. Kilka uwag o ewolucji poglądów teoretycznoliterackich Stanleya Fisha', *Er(r)go: Teoria-Literatura-Kultura*, Vol.12, nr 1(2006): 25-37.

[6] Find more about architecture of these texts in: Rohith Ghandi, "Generative Adversarial Networks – Explained", access 8.05.2019, https://towardsdatascience.com/generative-adversarial-networks-explained-34472718707a

of filtering and arranging random elements (elements of a random vector z), which result in the creation of a work of art. It is interesting to wonder, what would Witkacy himself say about such a pure form? Unfortunately, this purity is only temporary and, for the recipient of the final effect of the network operation, unavailable, because, as I have already mentioned, GAN is a duo. The generator creates, but the Discriminator criticizes, and this with the power that many critics could only dream of: the feedback from the twin censor has an immediate impact on the dynamic process of improving the original work. This procedure is repeated as long as the critic is content.

This approach to the creative process involves two sensitive issues: first, the artist is a slave to his critic. Secondly, the critic (and ultimately also the creator) is a slave of the existing, expected patterns, because the censorship network bases its expertise on the 'knowledge' gained from the collection that it learns, i.e. millions of samples of existing "works" (these can be texts, photographs or music or video, depending on what the network is supposed to be specialized in). And the constant struggle between the two networks is that the Generator is trying to create something new that is convincing enough to deceive a critic who will consider them 'real', i.e. most highly probable. A discriminator, therefore, accepts only those works that fit into the patterns known to him. In other words, works that are nothing more than a centenary, a work composed entirely of creatively selected and interconnected quotations, elements that had already existed somewhere in the past. Although the resolution here is slightly different, because the features distinguished in the vector of z (these "quotations" from taught works) are almost inaccessible to a man (the logic of dividing and separating is far from the natural human categorization into words, objects, colours or shapes), the structural assumption is the same.

The process of creation based solely on new elements and new juxtapositions of already existing elements and referring only to a limited set of expected results may, of course, be rather fatal in the long run, due to the almost exponential development of 'inbreeding'. If networks for generating, editing (such as external Grammarly, built-in functions in text editors) and translating texts are taught on certain corpuses of the texts that meet standards of correctness and style, and are increasingly used by human authors, they can lead to an increasing number of stylistically 'standardised' texts. The release of these texts, however, will increase the probability of their inclusion in the body of teaching data for subsequent networks, and so, step by step, the percentage of participation of this norm in the pool of inspiration will rise, translating into an impoverishment of the range of stylistic or lexical diversity, creating a horrible vision of looping Bakhtin's dialogism[7].

Fortunately, however, the generation of texts based on this mechanism is not the only form of textual expression, neither machine nor human. Cognitive activist Margaret Boden has distinguished three types of creativity[8]:

1. exploratory, i.e. one that makes it possible to explore existing creations in search of border alternatives while not breaking the rules that have been adopted;

[7] It refers to the modernist idea of "literary language" as correct, normative, sometimes transparent, in the fear of collapse. Such an attitude was visible in the long tradition of translations, which raised the register and smoothed out the original roughness of classical works, which entered into the workflow of inspiration in such an unnecessarily impoverished, normalised form.

[8] Margaret A. Boden, *The Creative Mind: Myths and Mechanisms*, (London: Routledge, 2004): xxx

2. combinatory, which allows you to combine elements previously considered to be absolutely irrelevant to each other;

3. transformational, i.e. one which results in real transformations and when the elements considered previously as indisputable change - materials, tools, styles of reception.

The last type seems to be a mixture of juxtaposition and exploration while crossing borders and breaking rules. The key to a creative success powered by this type of creative behaviour is, however, as du Sautoy[9] points out, preparation for failure, because a real breakthrough is usually the last stage of a long journey of completely wrong ideas. Emotional connection with the artwork, the unquantifiable cost of the creative process for a human being often prevents such a cold, almost calculating approach to his own failure, which makes him treat it only as a source of direct information about the course to be taken in the next iteration. Such an approach is what Artificial Intelligence adaptive systems are specialized in.

Since neural network learning for natural language processing is usually an extremely time-consuming process of great processing complexity (often impossible on a regular PC), but also because of the direct benefits of an extended set of learners, pre-trained models are available on the network. Under intriguing names, (such as Transformer, BERT, ELMo, Flair and others) there are models that have already been designed with the right architecture, parameters and huge text corpuses (e.g. current information from all over Wikipedia). Some are designed for a specific task (e.g. sentiment analysis of a text, i.e. whether a statement is positive, neutral or negative); others, such as Google's BERT hybrid, are made for multi-tasking.

Especially in the last case, it is interesting that the teaching sets designed to solve one problem can be used for another, and considering that the problem of prompting words, translating or generating whole texts uses the knowledge of the context and possible combinations distilled from millions of samples of existing texts, it will turn out that quotations from one type of text are successfully used to process another type of text. In other words, following a strictly literary metaphor, it is like a writer writing a fragment of prose using words, expressions, relations and structures derived from poetry, drama or other genres. Which, of course, is a natural mechanism of creation in contemporary literature, where one talks about 'listening' to idiolects, 'street' speech, references to the poetics of new media and the like. Many authors and their critics intuitively felt long ago that such interpoetics, an open and inviting gesture of the artist, bring us closer to good literature and promise a text of good quality. And the AI with its action only proven measurable advantages of this opening to others - models trained on non-specific data sets (i.e. thematically or stylistically corresponding to a given task) generate clearly better results. The transfer of structures can take place not only at the level of the text on which the network learns to create, but also at the level of the architecture of the network itself, so not only the material is borrowed[10].

[9] Du Sautoy, 39

[10] Marius Pompescu and Radu Tudor Ionescu wrote about the applications of machine learning solutions from the area of image analysis to automatic text analysis; *Knowledge Transfer between Computer Vision and Text Mining*, Bucharest: Springer, 2015, Springer 2015; on the issue of method transfer within the different tasks of natural language processing Lili Mou and al., 'How Transferable are neural networks in NLP applications', Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2016): 478-489.

The GPT-2 model developed by OpenAI was designed to impress with the quality of its work, which focused primarily on generating text completion, machine translation, and question answering. The last task consists mainly of parsing the text, extracting information and reasoning, and will not be crucial for our deliberations, while both the generation of text and the generation of translations are tasks based primarily on intelligent and creative use of billions of quotations, their fragmenting and creative re-composition.

GPT-2 must first identify the poetics of the given fragment and then, in the middle of the "read" (previously learned), identify the poetics of the given fragment and find possibly equivalent poetics, from which it creates a new work corresponding to the style and content of the first paragraph of a text written by a human being. And this model is extremely 'well-read', as it uses as much as 40 GB of text data from a total of 45 billion pages (and subpages) published on the Internet. This includes all text genres, from blog posts and portal articles to comments, recipes, consumer reviews and manuals, with the exception of Wikipedia, which has been completely removed from the collection (as it has been too often used by other models). The action of the whole model (whose creators protect, because - as they brag – it is so good that it could be used for some unwanted aims, such as creating very convincing fake news) is really impressive.

The actual entry[11] on the GitHub profile, a social network platform for exchanging files, projects and publishing the code, here from the help file commenting on the configuration of the development environment, has been loaded into the model. The task was to complete this text as much realistically as possible. The result of the algorithm below (original man-made text in blue, added by AI in black):

> In Cygwin on Windows I found that I could not set the permissions of my /.ssh/ folder to be 0600, as is required for ssh to allow you to use keys. The symptom I had was that I no matter what I did, it always modified the owner and the group, making these files read-only. I tried just starting ssh and using a program like passwd which modifies the owner/group to 0644 permissions, but it didn't work. It was annoying but not nearly as bad as it could be, just annoying to have to fiddle with permissions. So I decided to do this as the reverse: For each .ssh folder in your Cygwin directory, set the permission of each file and directory to 2644. Then start ssh with: ssh-add /.ssh/[12]

The creators of the model claim that it is able to continue the text in a chosen style and indeed, if it was not for the colour of the text, the whole thing would have calmly passed the Turing test, and the place of the fusion would have completely escaped the human eye and ear. As you can see, the algorithm easily 'felt' the author, further enclosing his personal self; it also used the words 'annoying' and 'nearly', which appear in the original document and probably gained priority importance as use-worthy stylistic markers of the author.

[11] https://github.com/lelandbatey/configDebDev/blob/master/helpFiles.md

[12] Alec Radford i in, "Language Models are Unsupervised Multitask Learners", technical report, access 20.07.2019, https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf

To create this text, however, it was not enough to create a single sample of a given writing, nor to create many works by its author, but this synthetic work consisted of hundreds of thousands of entries from other people, most probably also from the GitHub platform and the absolutely necessary knowledge based in the 'programming self-help' mode, i.e. StackOverFlow. It is a platform where users answer each other's questions about specific problems, often describing their own experiences in an informal and emotionally distinctive way (after all, there are moments when, after many hours of trials, the program becomes jammed and keyboards fly on the wings of frustration).

GPT-2 is also great at dealing with fiction, generating text whose style and imaginativeness could embarrass not only graphomaniacs:

> In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English. The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science. (...) Dr. Jorge Perez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. (...) Dr. Perez believes that the unicorns may have originated in Argentina, where the animals were believed to be descendants of a lost race of people who lived there before the arrival of humans in those parts of South America. (...) However, Perez also pointed out that it is likely that the only way of knowing for sure if unicorns are indeed the descendants of a lost alien race is through DNA. 'But they seem to be able to communicate in English quite well, which I believe is a sign of evolution, or at least a change in social organization,' said the scientist[13].

The text uses reported speech and independent speech, there are characters and relationships between them, there are authorities and difficult words, there is magic and mystery, and the world is shown in a colorful way like the rainbow land of unicorns. In the last sentence, there is also a good substitute for humour: a structurally convincing punchline based on a surprising, not evident (and therefore creative) juxtaposition of the development of speech in unicorns with a social organization, which is smuggled in under the cover of a cause and effect conclusion, so suitable for an expert quotation from an archetypical researcher. The network apparently topped the unicorns, because it not only spoke with perfect English, but also with British humour.

The same model was also used to generate a text to be a continuation of the first sentence from Orwell's novel *198414*. The network was intelligent enough to 'sense' a subtly futuristic style:

> It was a bright cold day in April, and the clocks were striking thirteen.
>
> I was in my car on my way to a new job in Seattle. I put the gas in, put the key in, and then I let it run. I just imagined what the day would be like. A hundred years from now. In 2045, I was a teacher in some school in a poor part of rural China. I started with Chinese history and history of science.

[13] Radford

[14] 'New AI fake text generator may be too dangerous to release, say creators', *The Guardian*, online article dated 14.02.2019, access 20.07.2019, https://www.theguardian.com/technology/2019/feb/14/elon-musk-backed-ai-writes-convincing-news-fiction

Ivana Kajić, a PhD candidate in IT from Waterloo University in Ontario, decided to test the officially available, simplified GPT-2 model on song lyrics[15]. She gave the network, among other things, the first two verses of Ed Sheeran's song *Shape of you*:

> The club isn't the best place to find a lover
> So the bar is where I go
> Me and my friends at the table doing shots
> Drinking fast, and then we talk slow
>
> And you come over and start up a conversation with just me
> And trust me, I'll give it a chance now
> Take my hand, stop
> Put "Van the Man" on the jukebox
> And then we start to dance
> And now I'm singing like

In the scene there is an indication of a bold interaction between men and women, but there is still a jukebox playing and a desire to dance; the meeting of bodies will be non-literal, with a hint of Irish datedness. In the further part of the original, yes, there are more and more direct images, but within the framework of the assonances and metaphors. The fragment added by the web introduces a completely new poetics, shorter, with a more rhythmic phrase and a stronger subject, and although rhymes are not to be found here in any way, the power of expression (including the surprising 'you better get dressed', suggesting that the situation was more rapidly-developing than we initially thought) is more similar to the lyrics of rap songs:

> You know what I mean
> That's the way it is
> So please, I know it's weird
> But I know it's right
> Because I just tell it
> And I say 'Hi, nice to meet you
> You better get dressed'

The text inspirations resulting from the mentioned wide, non-specialized 'erudition' of the web were thus reflected in the work. This model is a reduced version of his genius brother and this reduction manifests itself most clearly in a greater tendency to follow the patterns learned earlier, while at the same time being less able to 'feel' the style of the given fragment, which eventually results in an interesting mix of poetics and a creative exploration of broader possibilities (one step ahead of Chomsky's furious dream of green ideas).

[15] Ivana Kajić, 'AIternate endings: Lyrics completion using GPT-2', access 15.07.2019, http://www.ivanakajic.me/blog/2019/03/31/aiternate-lyrics

***

Generative approach, juggling with quotes, unrestrained style of using various lexicons - brave representatives of OuLiPo, the French school of automatic literature from the 60s, apparently have quite worthy descendants, although it will be rather difficult to describe the life vicissitudes of machines, in vain search of the impulse of heart, broken nerves, bad choices, happiness and everything else that is part of the creative process.

And the real artwork really pleases with the potential of this process. I do not know what the artist felt when he was creating, but how nice it is to assume that he could have felt something at all. Art created by artificial intelligence still appears as "artificial", because it is deprived of this potential; in a word - completely inhuman.

translated by Agnieszka Kocznur

Bibliography

Michaił Bachtin, *Problemy literatury i estetyki*, translated byW. Grajewski, Warszawa: Czytelnik, 1982

Margaret A. Boden, *The Creative Mind: Myths and Mechanisms*, London: Routledge, 2004

Leszek Drong, „Od konwencjonalizmu do normatywizmu. Kilka uwag o ewolucji poglądów teoretyczno-literackich Stanleya Fisha", *Er(r)go: Teoria-Literatura-Kultura*, Vol.12, nr 1(2006): 25-37.

Stanley Fish, *Interpretacja, retoryka, polityka*, editorial and foreword Andrzej Szahaj, introduction by Richard Rorty, translated by Andrzej Szahaj, Kraków: Universitas, 2002

Rohith Ghandi, "Generative Adversarial Networks – Explained", access 8.05.2019, https://towardsdatascience.com/generative-adversarial-networks-explained-34472718707a

Ivana Kajić, "AIternate endings: Lyrics completion using GPT-2"**,** access 15.07.2019, http://www.ivanakajic.me/blog/2019/03/31/aiternate-lyrics

Agata Kazimierska, „Prawdziwe kłamstwa", *Tygodnik Powszechny*, online article dated 15.04.2019, access 20.07.2019, https://www.tygodnikpowszechny.pl/prawdziwe-klamstwa-158508

Małgorzata Łukaszewicz, *Pięć razy o przekładzie*, Kraków: Karakter, 2017

Lili Mou i in, "How Transferable are neural networks in NLP applications", *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2016): 478-489

"New AI fake text generator may be too dangerous to release, say creators", *The Guardian*, online article dated 14.02.2019, access 20.07.2019,

https://www.theguardian.com/technology/2019/feb/14/elon-musk-backed-ai-writes-convincing-news-fiction

Marius Pompescu i Radu Tudor Ionescu, *Knowledge Transfer between Computer Vision and Text Mining*, Bucharest: Springer, 2015

Alec Radford i in, "Language Models are Unsupervised Multitask Learners", technical report, access 20.07.2019, https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf

Marcus du Sautoy, *The creativity code. How AI is learning to write, paint and think,* London: Fourth Estate, 2019

KEYWORDS

literary criticism

ARTIFICIAL INTELLIGENCE

machine learning

Abstract:

Is creativity only a human domain? Can a neural network, even the most sophisticated architecture, fed with material created and chosen by man, be creative, and even if it is not a work of art secondary to human beings? Or maybe, as Bakhtin, and behind him Kristeva, wanted, each of our expressions is still destined to be secondary, because this is the nature of language? What is creativity, what can artificial intelligence do, what critical literary reflections can its work induce, especially in the context of intertextual and interpoetic relations? In the article I am searching for answers on the example of functioning of neural networks type GAN and GPT-2 model. Apart from fragments of analyzed texts and references to the theory of literature, there is also an introduction to the structure and essence of the analyzed technological solutions.

CREATIVITY

creative writing

INTERPOETICS

INTERTEXTUALITY

Note on the Author:

Inez Okulska - doctor of humanities in the field of literary studies. After going through a colorful humanistic journey (which involved, among others, linguistics, literary comparative studies, cultural studies, philosophy), which ended with a postdoctoral internship at Harvard University, she took up a master's degree in Automatics and Robotics at the Warsaw University of Technology. Currently she is a PhD student at the TIB Doctoral School of the Polish Academy of Sciences in the field of technical computer science. Artificial intelligence methods, and particulary methods of natural language processing, which are currently being scientifically researched at the NASK State Research Institute, perfectly combine these distant fields, especially since these methods are most frequently applied from the analysis of literary material.

She published in "Przekładaniec", "Literatura na Świecie", "Czas Kultury", "Poznański Studiach Polonistycznych", "Forum Poetyki", and has also presented the results of her research at several national and foreign conferences, both humanities and technical.