proceedings

MDPI

Proceeding Paper

Advanced NLP Procedures as Premises for the Reconstruction of the Idea of Knowledge †

Rafal Maciag

Institute of Information Studies, Jagiellonian University, 30-348 Kraków, Poland; rafal.maciag@uj.edu.pl;
Tel.: +48-602-289-144

† Presented at the Conference on Theoretical and Foundational Problems in Information Studies, IS4SI Summit 2021, online, 12–19 September 2021.

Abstract: This paper evaluates artificial texts produced by the GPT 2 and GPT 3 language models and proposes to use hermeneutic tools for their analysis, putting them on an equal footing with man-made texts. This decision is due to the intelligibility of these texts and the implicit knowledge on which they are based. Since the base of these language models is always the language corpus, it can be assumed that it is the source of this knowledge and that tools such as discourse and, in particular, the theory of discursive space can be used for its analysis.

Keywords: knowledge; NLP; GPT; hermeneutics; language model; discourse

check for updates

Citation: Maciag, R. Advanced NLP Procedures as Premises for the Reconstruction of the Idea of Knowledge. *Proceedings* **2022**, *81*, 105. https://doi.org/10.3390/proceedings2022081105

Academic Editor: Mark Burgin

Published: 1 April 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Deng and Liu date the beginning natural language processing to 1950, when Turing proposed his famous test [1]. They divide this long period of development into three waves. We live in the third wave, which began around 2010 and is related to the so-called deep learning, i.e., the use of artificial neural networks. Their most advanced solution so far is a language model called GPT 3 (Generative Pre-trained Transformer 3), which is based mostly on the previous model, called GPT 2 and differs from its predecessor mainly in terms of the amount of data used to train this model [2,3].

According to the definition given by Jurafsky and Martin, language models are models "that assign probabilities to sequences of words" [4] (p. 30). Aggarwal explains that "language models [are] used to create a probabilistic representation of the text" [5] (p. 4). Charniak writes that "a language model is a probability distribution over all strings in a language" [6] (p. 71), and Goodfellow et al. stated that: „A language model defines a probability distribution over sequences of tokens in a natural language" [7] (p. 461).

According to these definitions, the solutions proposed in the GPT 2 and GPT 3 models constitute the-state-of-the-art of natural language processing, because they propose the most advanced version of a language model so far. The basic embedding technique used in them [8] is supplemented by procedures in the area of the so-called recurrent neural networks, and in particular, their variant called. LSTM (Long Short-Term Memory) [9] and the so-called attention procedure [10].

The basis for the model is a linguistic corpus that is also used to train it. It is a set of texts that, as the authors of the model write, should represent "as large and diverse a dataset as possible in order to collect natural language demonstrations of tasks in as varied of domains and contexts as possible" [3]. For both models, these sets are sourced from the Internet, and in particular, are the result of the Common Crawl web crawler. It can be said that in this way the early idea of Pierre Lévy of collective intelligence is renewed [11], which is the result of combining the knowledge and activity of individual people through the network.

In the case of the GPT 2 model, the authors relied on a set of links inside the Reddit social platform and used only those that were rated at least 3 on the internal scale of the platform (the so-called karma), which caused only texts that have been curated/filtered by humans to emerge. Karma "reflects how much good the user has done for the Reddit community. The best way to gain karma is to submit links that other people like and vote for, though you won't get karma for self posts" [12]. The resulting set, WebText, contained 45 million links, which, after additional cleaning, deduplicating, and the removal of texts from Wikipedia due to their possible problems during the testing procedure, allowed to collect approx. eight million documents for a total of 40 GB of text.

The corpus of GPT 3 model was also based on a set of texts from the Internet provided mostly by Common Crawl, but this resource in order "to improve the average quality" was filtered and fuzzily deduplicated "i.e., removed documents with high overlap with other documents". Filtering took place using a classifier trained on the original WebText, Wikipedia, and web books corpus (used also later) as positive examples. Unfiltered Common Crawl was used as a set of negative examples. The resulting collection of texts was supplemented by "several curated high-quality datasets": an expanded version of the WebText, two internet-based books corpora, and English-language Wikipedia. The volume of the non-abundant Common Crawl collection from 2016-2019 was 45TB, and the filtered volume was 570GB, which was about 78% of the final set of tokens for training. The rest fell on the remaining collections. The sets of texts were treated unequally during the training of the model. Some of them (Common Crawl, one of the collections of books) were used also in sampling during the training significantly less frequently. Wikipedia was the most used. This was because authors viewed these datasets as higher-quality [2].

The result of the work of the GPT models are texts that are correctly constructed in terms of syntax and semantics, as well as making a specific cognitive contribution, i.e., based on certain, immanent knowledge. The latter issue is also in line with the theoretical assumptions regarding the so-called distributed connectionist models of artificial intelligence, i.e., artificial neural networks [13]. In this situation, these texts should be considered as full-fledged texts: correct and intelligible, and thus equal to man-made texts. However, there is a question about the source of this knowledge.

Despite many tests to which IT solutions in the field of artificial neural networks are subjected, which were also used in the GPT models, in such tasks as language modeling, cloze and completion, translation, Winogradow-style, common sense reasoning, reading comprehension, etc. there is no test to confirm or deny the equivalence of artificial and human texts, and in particular to confirm or deny their intelligibility based on explicit and exhaustive criteria. The only such procedure, which rather confirms the fact of this lack, is the so-called Turing test [14]. In the case of intelligibility, not only the so-called intelligence but also knowledge has a crucial meaning, although it should be noted that both categories do not have precise denotations. Rather, they meet Wittgenstein's definition, i.e., they are rather the result of a specific and continuous interpretive game. There is no doubt, however, that knowledge as a subject of reflection gains a new research plane in the case of artificially generated texts.

Some help in this regard can be provided by the fact that text is one of the oldest objects of philological and philosophical reflection. It can be said without exaggeration that it is the foundation of Western civilization, as well as many others. At the same time, this role should be considered in the context of other linguistic materializations, such as speech (oral expression) or printing. Each of them introduces its specificity and has its literature on the subject. It should be noticed that although in the case of the GPT models only the text in electronic form appears, this should not prevent its analysis as analogous to human products.

This type of analysis is usually associated with the so-called hermeneutic research. Porter and Robinson recall the legendary source of hermeneutics and emphasize the "in-between" relation based on the figure of Hermes who mediated between gods and humans [15]. Hermeneutics can be juxtaposed with philosophy as the search and production of meaning through interpretation. In this juxtaposition, the text appears as the world's equivalent, and the analytical procedures can be extended accordingly [16]. As

Bleicher writes “hermeneutics can loosely be defined as the theory or philosophy of the interpretation of meaning” [17] (p. 1). Although hermeneutics has appeared at least since ancient times, its self-conscious (philosophical) version has been developing intensively since the mid-19th century, and its numerous representatives include philosophers such as Frederick Nietzsche, and later Martin Heidegger and Hans-Georg Gadamer, as well as the so-called poststructuralists: Michel Foucault, Jacques Derrida, Gilles Deleuze, and Richard Rorty. Other similar schools are the English analytical philosophers, e.g., Bertrand Russell, Gilbert Ryle, and John Longshaw Austin, or the so-called Vienna Circle, with its most famous representative Ludwik Wittgenstein. It seems that hermeneutics is probably one of the effective strategies that can be used to indirectly evaluate the intelligibility of artificially generated texts.

Recognizing artificial texts as equal to human ones makes important the issue of the language context, which takes the form of the raw text in in various concepts and models appearing in the field of NLP, e.g., [1,4,7]. It is a procedure resulting from a generally semantic and in this sense qualitative attitude in the proposed solutions. These solutions, by necessity, also activate a description level higher than the text instance itself, i.e., language level. This type of problem is also justified in the extensive hermeneutic reflection mentioned above, i.e., present in the texts by Foucault, Derrida, Austin, etc. Sowa gives a very rich and concise description of this issue [18].

In this situation, the question about the component, which is knowledge, which is a feature of an intelligible text, in addition to semantic and syntactic correctness, finds theoretical support. It is the discourse theory, in which the concept of Michel Foucault should be distinguished. This concept interprets discourse as a linguistic instance of knowledge, which is the result of social, historical, and collective activity, as a result of which a specific and autonomous entity appears called the discourse [19–21]. This concept developed intensively in the following years, creating a separate, multi-threaded discourse theory.

One of such theories is the approach proposed by the author, entitled discursive space. Discursive space is an n-dimensional dynamical space in which discourses, as autonomous instances of knowledge, run in time trajectories describing the real state of knowledge in the subject they concern. The category of space can be extrapolated to the manifold [22,23]. The corpus of the texts is by no means a homogeneous and stable mass, but a dynamic structure. Thanks to the idea of discourse space it reveals this complex structure. It seems fibrous (in time), where particular fibers are created from the individual trajectories of discourses.

Discourse is not the simple sum of utterances, but the effect of their dynamical interdependent interactions dependent also on the local social, cultural, racial, etc., circumstances. These utterances may take various material (symbolic) forms, although the text form is the easiest available. The aforementioned interactions at the moment of articulation in the form of an utterance reveal themselves as temporarily stable relationships of various orders and types, in which there is a concise knowledge of the content of the utterance, and this temporality may have a very different duration. The discursive space makes it possible to illustrate the aforementioned dynamics of interaction. This is important because it shows the lack of this property in the case of GPT models.

2. Results

In the absence of any convincing arguments excluding the achievement of a sufficient level of cognitive quality and linguistic correctness of artificial texts, texts created by GPT models should be considered full-fledged texts, which should then be subjected to hermeneutical analysis on par with man-made texts, looking for a possible falsification of this assumption.

Artificial texts open up a new perspective in the discussion of knowledge. If we assume that these texts show the presence of their immanent and implicit knowledge, the question of the type and source of this knowledge becomes legitimate. Since the only semantic resource on which the algorithms that create these texts are based is language, or more precisely a set of real, spontaneous texts, it should be considered that this is the only source of knowledge for these algorithms.

This conclusion provides support for the idea of knowledge for which language is a carrier. This type of concept is already present and its expanded variant represents the concept of discourse, which acts as a knowledge container. The issue of placing the phenomenon at the language level is a separate problem because actually the models are based on sets of texts. However, such a generalization, which appears at the level of the very definitions of the linguistic model, is legitimate, as it follows from the justifications provided by hermeneutics.

It should also be added that despite the reference to the language category, both models do not use linguistic tools in practice, but propose their own techniques, such as tokenization, LSTM, or attention. These are specific and effective techniques for manipulating the text.

Artificial texts, treated on a par with human texts, can be considered as one of the direct pieces of evidence of the existence of non-human knowledge, i.e., knowledge not directly related to human cognitive or mental competencies.

3. Discussion

The GPT 2 and GPT 3 models presented make it possible to interpret a very important relationship that exists between knowledge and language, which appears directly as massive sets of texts. There is an extensive theory of this relationship based on the idea of discourse. On the other hand, there is also an extensive theory of the text as an object of (semantic) interpretation, i.e., hermeneutics. However, it is necessary to point out some circumstances that hinder the research procedure:

1. Hermeneutics as a set of philosophical reflections is entirely qualitative. Its procedures were also not designed to evaluate texts, but to interpret them. However, on the other hand, the thread underlying hermeneutics is important, as it makes the text completely autonomous and detaches it from the author's instance. In the original version, this autonomy meant the divine origin of the text undergoing exegesis.
2. Textual datasets used as corpora cannot be considered the final representation of the resources. They are subject to internal processes of continuous construction. They refer also directly to the Internet, which is a very unstable source of text.
3. Internet text resources are not homogeneous. They are scattered among various communities, including national communities, and do not constitute a coherent whole. Since we are dealing with a complex phenomenon, the deterministic delimitation of the separate areas is very difficult.
4. The datasets underlying the linguistic model are arbitrary or even random. Model authors often use vague selection criteria, such as quality, without defining this quality in any way and adopting it intuitively. Their general assumption is that the maximum richness of texts is only a loose postulate. Each interference in the corpus increases the degree of its randomness, which does not necessarily mean reaching the level of statistical representativeness of knowledge, language, etc.
5. Each of the procedures (algorithms) used within GPT models has an unknown impact on the semantic structure, starting with tokenization of the text corpus using Byte Pair Encoding (BPE). The techniques of recurrence and attention, as well as other, less innovative techniques, such as softmax, etc. also have such an impact on the formation of the final semantic structure. It is also difficult to imagine the possibility of a semantic evaluation of these procedures by a method other than trial and error due to the distributed and implicit nature of this semantics.

4. Conclusions

The GPT 2 and GPT 3 models can be treated as generators of autonomous text equal to human texts, which has further consequences in the form of understanding the knowledge on which they are based, which can also be treated in this way.

These models confirm the existence of knowledge nested in language, although they do not make it possible to understand the nature of this existence due to its implicit nature. Procedures based on the linguistic nature of knowledge, such as those based on the idea of discourse, such as the theory of discursive space, may apply to such knowledge.

The border between the quantitative and qualitative approaches, and appropriately formal (mathematical, syntactic) and content (semantic) approaches within digital technology seems to be practically overcome, but the assumptions used and the choices made by the creators raise further problems and questions.

Funding: This work was supported by the National Science Centre, Poland (grant no 2018/29/B/HS1/01882).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data sharing not applicable.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Deng, L.; Liu, Y. (Eds.) *Deep Learning in Natural Language Processing*; Springer: Singapore, 2018; ISBN 978-981-10-5208-8.
2. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models Are Few-Shot Learners. *arXiv* **2020**, arXiv:2005.14165.
3. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language Models Are Unsupervised Multitask Learners. *OpenAI blog* **2019**, *1*, 9.
4. Jurafsky, D.; Martin, J.H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 3rd ed.; Prentice Hall: Upper Saddle River, NJ, USA, 2020.
5. Aggarwal, C.C. *Machine Learning for Text*; Springer International Publishing: Cham, Switzerland, 2018.
6. Charniak, E. *Introduction to Deep Learning*; The MIT Press: Cambridge, MA, USA, 2019; ISBN 978-0-262-03951-2.
7. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
8. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; Dean, J. Distributed Representations of Words and Phrases and Their Compositionality. *arXiv* **2013**, arXiv:1310.4546.
9. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
10. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762.
11. Lévy, P. *Collective Intelligence, Mankind's Emerging World in Cyberspace*; Perseus Books: Cambridge, MA, USA, 1999.
12. Reddit Karma-Reddit.com. Available online: https://www.reddit.com/wiki/karma (accessed on 25 September 2021).
13. Flasiński, M. *Introduction to Artificial Intelligence*; Springer: Cham, Switzerland, 2016; ISBN 978-3-319-40020-4.
14. Turing, A.M. Computing Machinery and Intelligence. *Mind* **1950**, *236*, 433–460. [CrossRef]
15. Porter, S.E.; Robinson, J.C. *Hermeneutics. An Introduction to Interpretative Theory*; William B. Erdmans Pulishing Company: Grand Rapids, MI, USA; Cambridge, UK, 2011.
16. Malpas, J.; Gander, H.-H. (Eds.) *The Routledge Companion to Hermeneutics*; Routledge: Abingdon, UK, 2015.
17. Bleicher, J. *Contemporary Hermeneutics: Hermeneutics as Method, Philosophy and Critique*; Reprint, 1982 editon; Routledge & Kegan Paul: London, UK; Boston, MA, USA, 1980; ISBN 978-0-7100-0552-6.
18. Sowa, J.F. The Role of Logic and Ontology in Language and Reasoning. In *Theory and Applications of Ontology: Philosophical Perspectives*; Poli, R., Seibt, J., Eds.; Springer: Dordrecht, The Netherlands, 2010; pp. 231–263. ISBN 978-90-481-8845-1.
19. Foucault, M. *Les Mots et Les Choses. Unearchéologie des Sciences Humaines*; Gallimard: Paris, France, 1966; ISBN 978-2-07-022484-5.
20. Foucault, M. *L'archéologie du savoir*; Gallimard: Paris, France, 1969; ISBN 978-2-07-026999-0.
21. Foucault, M. *L'ordre du Discours: Leçoninaugurale au Collège de France Prononcée le 2 Décembre 1970*; Gallimard: Paris, France, 1971.
22. Maciag, R. Discursive Space and Its Consequences for Understanding Knowledge and Information. *Philosophies* **2018**, *3*, 34. [CrossRef]
23. Maciag, R. Ontological Basis of Knowledge in the Theory of Discursive Space and Its Consequences. *Proceedings* **2020**, *47*, 11. [CrossRef]