*computer sciences and mathematics forum*

MDPI

*Proceeding Paper*

# Pretrained Language Models as Containers of the Discursive Knowledge †

**Rafal Maciag**

Institute of Information Studies, Jagiellonian University, 30-348 Cracow, Poland; rafal.maciag@uj.edu.pl; Tel.: +48-602-28-91-44

† Presented at the 2023 Summit of the International Society for the Study of Information (IS4SI 2023), Beijing, China, 14–16 August 2023.

**Abstract:** Discourses can be treated as instances of knowledge. The dynamic space in which the trajectories of these discourses are described can be regarded as a model of knowledge. Such a space is called a discursive space. Its scope is defined by a set of discourses. The procedure of constructing such a space is a serious problem, and so far, the only solution has been to identify the dimensions of this space through the qualitative analysis of texts on the basis of the discourses that were identified. This paper proposes a solution by using an extended variant of the embedding technique, which is the basis of neural language models (pre-trained language models and large language models) in the field of natural language processing (NLP). This technique makes it possible to create a semantic model of the language in the form of a multidimensional space. The solution proposed in this article is to repeat the embedding technique but at a higher level of abstraction, that is, the discursive level. First, the discourses would be isolated from the prepared corpus of texts, preserving their order. Then, from these discourses, identified by names, a sequence of names would be created, which would be a kind of supertext. A language model would be trained on this supertext. This model would be a multidimensional space. This space would be a discursive space constructed for one moment in time. The described steps repeated in time would allow one to construct the assumed dynamic space of discourses, i.e., discursive space.

**Keywords:** natural language processing; neural language models; knowledge; discourse; discursive space



**Citation:** Maciag, R. Pretrained Language Models as Containers of the Discursive Knowledge. *Comput. Sci. Math. Forum* **2023**, *8*, 93. https://doi.org/10.3390/cmsf2023008093

Academic Editors: Zhongzhi Shi and Wolfgang Hofkirchner

Published: 12 January 2024



## 1. Introduction

In recent years and even months, NLP solutions have been rapidly developed. Although research in this area has been going on since at least the 1960s, only recent solutions have provided spectacular achievements, such as GPTChat by OpenAI (https://openai.com/blog/chatgpt/, accessed on 11 January 2024), but also include the newest solutions, in particular GPT-4 (which is the basis of GPTChat) [1]. They are based on transformer technology, which uses the attention technique [2]. This technology allows one to define deeper semantic inferences in the text and is currently the most advanced solution in the field of NLP, the so-called state of the art.

The solution based on transformers technology allows one to create semantic language models based on the so-called vector semantics, which "instantiates this linguistic embeddings hypothesis (distributional hypothesis) by learning representations of the meaning of words, called embeddings, directly from their distributions in texts" [3] (p. 103). The analysis of the probability of occurrence of a word in the context of the nearest words allows one to construct a multidimensional space in which the coordinates of these words are treated as vectors [4]. The revolutionary feature of this technology is the ability to use text, taken from real sources and collected in large corpora, as "implicitly supervised training data" [3] (p. 120). Mikolov et al. [5] described a breakthrough solution in this area.

Language models trained on large-text corpora allow one to generate artificial text. This text is created as a response to the so-called prompt: a text. Spectacularly advanced and intelligible artificial texts have revived the discussion on the cognitive competence of these algorithmic solutions. However, what seems to be better justified, and therefore more important, is that neural language models (NLMs) are containers of knowledge.

The relationship between language and knowledge has been extensively reflected in research in the form of the so-called discourse analysis. Based on Michel Foucault's concept of discourse, especially the text from 1971 (Foucault, 1971), a knowledge model was proposed, named discursive space, in which discourses are instances of knowledge travel trajectories in a multidimensional dynamic space. This model is based on a qualitative procedure to construct the dimensions of this space [6].

The qualitative procedure used so far to construct the dimensions of discursive space was a derivative of discourse analysis itself. The description of the discourse characteristics was based on the identification of a number of features. These features, or more precisely, the degree of their impact on various aspects, became the basis for the dimensions based on them. Determining this impact is the standard type of analysis of social phenomena. The coordinates were constructed by arbitrarily scaling this impact. Therefore, dimensions and coordinates were constructed on the basis of the same set of data used to analyze the discourse itself.

## 2. Results

The procedure for constructing the output semantic space in the NLM creation process is at a very basic level, similar to the construction of the discursive space. It consists of calculating the values of the coordinates of semantic units, the so-called tokens, based on words that allow the construction of a space representing the semantic relations between these tokens. Therefore, this space can be treated as a formal semantic model of language. In the process of further manipulation of vectors, further significant semantic effects are achieved. At the current stage of development, they are sufficient to achieve spectacular effects in the form of generating an artificial text that is practically equivalent to human text.

However, it can be assumed that it is possible to isolate semantic structures that are more complex than the semantic units so far, i.e., tokens, which are based on words and their relationships in sentences. Such structures are discourses, i.e., linguistic (semantic) structures with a higher degree of abstraction than the sentences they consist of. Therefore, one should search for higher-order units (discourses) composed of lower-order semantic units (words) and their relationships in sentences. This would be a repetition of the embedding technique, but transferred to a higher semantic level, the aim of which is to create a set of vectors describing discourses as semantic units of a higher order.

The embedding discourse technique would consist of constructing discourse coordinates, which would be based on the assumption that discourses as semantic units of a higher order also fulfill the distributional hypothesis, i.e., discourses related to semantic relations (relevant) are grouped together. Due to their semantic basis (sentences), they can also overlap (the same sentences can belong to different discourses). These discourses would be identified as sets of sentences (fragments of texts), constituting a discourse related to certain concepts (words), represented as tokens at a lower level of embedding. The strength of the semantic range in the text, determining the size of these sets, would be determined by the relevance coefficient, calculated on the basis of a model built at a lower level of embedding (token embedding). This model would also be the basis for the selection of qualified tokens as the basis of discourse. By analyzing the mutual position of the indicated discourses in the corpus of texts, a discursive linguistic model would be created. The introduction of a time variable, i.e., the construction of a dynamic discursive model, would fulfill the assumptions of discursive space. The introduction of this variable would be the next step in the analysis, which, however, does not appear at the level of the current embedding technique.

## 3. Discussion

An advanced research reflection on language in the context of knowledge is provided by discourse theory. Since its beginning, it has combined two key approaches. The first, formal, deals with the internal analysis of various types of statements, which are instances of discourse. The second approach of a social nature, deals with the study of the broad contexts of discourse as a phenomenon of language, inheriting the field of research from the latter. Fairclough describes it as follows: "I see discourses as ways of representing aspects of the world—the processes, relations, and structures of the material world, the 'mental world' of thoughts, feelings, beliefs, etc., and the social world." [7] (p. 124). Jørgensen and Phillips explicitly interpret the diversity mentioned by Fairclough as a problem of different instances of knowledge: "the struggle between different knowledge claims could be understood and empirically explored as a struggle between different discourses which represent different ways of understanding aspects of the world and construct different identities for speakers" [8] (p. 2). The topic of knowledge in the context of discourse gained a specific summary in the theory of discourse studies, which was created on the initiative of the outstanding expert on the subject, Teun van Dijk, who wrote: "Discourse presupposes (semantic) situational models of events talks about, as well as (pragmatic) context models of the communicative situation, both construed by the application of general, socially shared knowledge of the epistemic community" [9] (p. 601).

In particular, an advanced analysis of discourses as instances of knowledge was carried out by Michel Foucault in a number of his publications, e.g., [10–12], which then became the basis for further extensive research [7] (p. 2), [8] (p. 12). According to Foucault, "[a] group of elements [form and rigor, objects, statement types, notions, strategies—author's note], formed in a regular manner by a discursive practice, and which are indispensable to the constitution of a science, although they are not necessarily destined to give rise to one, can be called knowledge. (. . .) there is no knowledge without a particular discursive practice; and any discursive practice may be defined by the knowledge that it forms" [13] (pp. 182–183). Describing the specificity of the discourse, Foucault proposes the use of a method based on four rules (orig. *principes, règles*): reversal (*principe de renversement*), discontinuity (*principe de discontinuité*), specificity (*principe de spécificité*), and exteriority (*règle de l'extériorité*), which also define this specificity. These rules can be transferred to the phenomenon of knowledge, which acquires the unusual form of a set that is numerous, mobile, internally variable, and related, and at the same time elusive directly and observable only indirectly. These features allow this set to meet the conditions of a complex system [14] (p. 92).

Based on this observation and Foucault's concept, a theory of knowledge can be constructed as a dynamic space in which the trajectories of instances of this knowledge, i.e., discourses, can be described in time. This theory has been proposed in the following texts: [6,15,16]. According to its definition "Discursive space is an n-dimensional dynamical space in which discourses, which are autonomous instances of knowledge, run in time trajectories describing the real state of knowledge in the subject that they concern" [6] (p. 6). According to the idea of a manifold given in 1854 by Bernhard Riemann, the concept of space can be extended. This allows one to formulate the following definition of knowledge: "knowledge is a set of discourses contained in an n-dimensional manifold that can be interpreted locally as a discursive space" [6] (p. 7).

A serious problem faced by the presented theory of discursive space is the method used to construct the dimensions of this space. In the existing version, they are constructed qualitatively through the analysis of the researched discourse. This method was presented as an example of a discourse on the phenomenon of the Internet [17]. It was inspired by the approach of Byrne and Callaghan [18]. The quantitative approach necessary to construct the dynamic space has been complemented by a qualitative approach that provides a way to determine the dimensions through the analysis of the discourses studied. As discourses are studied, elements of their characteristics (features and qualities) appear, which can be treated as dimensions because their relevance changes over time. The latter property applies to all social phenomena whose features, such as significance, degree of social

involvement and interest, degree of influence on social processes and individual attitudes, etc., can be determined during the analysis of the discourse concerning these phenomena. As they are variable, they can be used as a basis for an arbitrary numerical scale.

Byrne and Callaghan refer to the idea of a topological space, which provides a similar approach in contrast to a metric space. Their approach finds support in the concept of topological geometry introduced by Henri Poincaré, abandoning the level of direct computation in favor of a higher level of abstraction analysis. Byrne and Callaghan describe this approach as follows: "Actually what we have in this example is not a set of models calibrated against real data in terms of initial inputs but rather a modelling process which establishes its correspondence to reality through a qualitative appreciation of how things are working out in reality" [18] (p. 162).

This approach can be replaced by the application of the solution proposed in the NLM technology, which is based on the construction of semantic space. In these models, the construction of the semantic space takes place at the level of tokens, i.e., semantic units located at the level between words and letters. The technique that allows one to determine tokens effectively and successfully in the case of the GPT-4 language model is probably the so-called byte-pair encoding that was used in GPT-2 and repeated in GPT-3 [19,20].

The construction of the discursive space based on the embedding technique would consist of isolating the discourses present in the text and then calculating their location in the abstract space, analogically to tokens, i.e., based on the probability of their occurrence in the context of other discourses. This would implement the so-called distributional hypothesis, as described by Juraffsky and Martin cited above, but would move to a higher level of semantic order.

These discourses would form a sequence analogous to the text that would be their source. A discourse is created around a specific issue, which is also the source of the name describing the subject of this discourse and the knowledge it contains. Thus, these names (concepts, words) would form a sequence of words (concepts) in the order resulting from the source text, repeating the order of the discourses. The resulting structure can be interpreted as a kind of supertext, which could then become the basis for training the language model, analogous to the token embedding technique. Due to a procedure that could be called discourse embedding, a linguistic model based directly on knowledge instances would be created because discourses as its base would be interpreted in this way. Therefore, this model would necessarily be a model of knowledge.

The described discourse embedding technique would be the first step towards a complete discursive space that represents a set of states that a particular discourse assumes at a particular time. This situation can be extended to a set of discourses, which would lead to the construction of a general model of knowledge, as opposed to knowledge limited to a single discourse. So, step two would be to build a series of language models over time. Then, we would be dealing with a set containing (ordered) sets of discourses (instances of knowledge) appropriate for specific moments in time. This would make it possible to plot the trajectories of relevant discourses in time and would implement the source model of dynamical space, which is the basis of the discursive space.

## 4. Conclusions

This paper indicates a direction for research that can lead to a formally interpreted and justified model of knowledge, which has been proposed in the theory of discursive space as a model of discursive space. This model interprets knowledge from the perspective of the existing discourse theory, which considers discourse as the articulation/retention of knowledge. The model presents these discourses as the trajectories of their instances in a multidimensional space, built on the basis of the analysis of real discourses identified in various types of utterance, primarily of a linguistic nature and textual in practice. Therefore, it is hybrid in nature, i.e., it combines a quantitative and qualitative approach, which already has precedents in existing analytical procedures specific to the social sciences.

NLM technology (neural language models) introduced a method of data analysis in the form of extensive text corpora called embedding, which allows one to determine

the coordinates of semantic units, the so-called tokens, representing words in a sentence directly by analyzing the distribution of occurrences of these units. This solution allows one in the next step to reconstruct a space representing semantic dependencies, leading to the further construction of a semantic, formal model of the language.

Based on significant analogies, primarily concerning the same subject of research, which is language, manifested as real, empirical text corpora, the paper proposes to use a technology analogous to token embeddings, but identifies hypothetical semantic units of a higher order than tokens. These units, i.e., discourses, have been presented and justified in discourse theory and formalized in discursive space theory. Such a technology could be called discursive embedding. Constructing the space of embeddings for discourses as semantic units of a higher level would solve the problem of constructing the dimensions of the discursive space and would allow it to be completely formalized. Although this paper does not propose a formal solution, it indicates a promising direction for the development of neural language models.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Data sharing is not applicable.

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

## References

1. OpenAI. GPT-4 Technical Report 2023. Available online: https://cdn.openai.com/papers/gpt-4.pdf (accessed on 11 January 2024).
2. 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.
3. Jurafsky, D.; Martin, J.H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 3rd ed.; (draft). 2023. Available online: https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf (accessed on 11 January 2024).
4. Kornai, A. *Vector Semantics*; Springer Nature: Singapore, 2023.
5. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; Dean, J. Distributed Representations of Words and Phrases and Their Compositionality. *arXiv* **2013**, arXiv:1310.4546.
6. Maciag, R. Theory of Knowledge Based on the Idea of the Discursive Space. *Philosophies* **2022**, *7*, 72. [CrossRef]
7. Fairclough, N. *Analysing Discourse: Textual Analysis for Social Research*; Routledge: London, UK, 2003.
8. Jørgensen, M.; Phillips, L. *Discourse Analysis as Theory and Method*; Sage Publications: London, UK; Thousand Oaks, CA, USA, 2002.
9. van Dijk, T.A. Discourse and Knowledge. In *The Routledge Handbook of Discourse Analysis*; Routledge: London, UK, 2013; pp. 587–603.
10. Foucault, M. *Les Mots et Les Choses, Une Archéologie des Sciences Humaines*; Gallimard: Paris, France, 1966.
11. Foucault, M. *L'archéologie du Savoir*; Gallimard: Paris, France, 1969.
12. Foucault, M. *L'ordre du Discours: Leçon Inaugurale au Collège de France Prononcée le 2 Décembre 1970*; Gallimard: Paris, France, 1971.
13. Foucault, M. *The Archaeology of Knowledge and the Discourse of Language*; Pantheon Books: New York, NY, USA, 1972.
14. Maciag, R. Advanced NLP Procedures as Premises for the Reconstruction of the Idea of Knowledge. *Proceedings* **2022**, *81*, 105. [CrossRef]
15. Maciag, R. Discursive Space and Its Consequences for Understanding Knowledge and Information. *Philosophies* **2018**, *3*, 34. [CrossRef]
16. Maciag, R. Ontological Basis of Knowledge in the Theory of Discursive Space and Its Consequences. *Proceedings* **2020**, *47*, 11.
17. Maciag, R. The Analysis of the Internet Development Based on the Complex Model of the Discursive Space. *Information* **2018**, *9*, 7. [CrossRef]
18. Byrne, D.S.; Callaghan, G. *Complexity Theory and the Social Sciences: The State of the Art*; Routledge, Taylor & Francis Group: New York, NY, USA, 2014.
19. 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.
20. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. *Language Models Are Unsupervised Multitask Learners*; OpenAI Blog: San Francisco, CA, USA, 2019.