applied sciences

MDPI

Article

Knowing Knowledge: Epistemological Study of Knowledge in Transformers

Leonardo Ranaldi [1,2,*] and Giulia Pucci [3]

[1] Department of Innovation and Information Engineering, Guglielmo Marconi University, Via Plinio 44, 00193 Rome, Italy
[2] Department of Enterprise Engineering, University of Rome Tor Vergata, Viale del Politecnico, 1, 00133 Rome, Italy
[3] Department of History, Cultural Heritage, Education and Society, University of Rome Tor Vergata, Viale del Politecnico, 1, 00133 Rome, Italy
* Correspondence: l.ranaldi@unimarconi.it

Abstract: Statistical learners are leading towards auto-epistemic logic, but is it the right way to progress in artificial intelligence (AI)? Ways to discover AI fit the senses and the intellect. The structure of symbols–the operations by which the intellectual solution is realized–and the search for strategic reference points evoke essential issues in the analysis of AI. Studying how knowledge can be represented through methods of theoretical generalization and empirical observation is only the latest step in a long process of evolution. In this paper, we try to outline the origin of knowledge and how modern artificial minds have inherited it.

Keywords: artificial intelligence; interpretable AI; human-centric AI

1. Introduction

For a long time, the study of symbolic syntactic interpretations has been the cornerstone of natural language understanding, which has been studied by the forerunners of intelligence deriving from solid starting points, e.g., universal grammar [1], and fundamental cognitive mechanisms to represent domains such as that of physical objects [2].

Artificial intelligence (AI) support is essential to speed up the analysis and discovery process in new paradigms. However, AI "minds" are nowadays dominated by non-symbolic, hidden statistical learners who have canceled symbols in their controlling strategies [3]. Moreover, these new technologies work hidden and complex; it is challenging to understand statistical models using human knowledge, placing them in an awkward position. Studying the headwater of knowledge could eviscerate many basic concepts by improving the position of statistical models.

The theories learned for centuries on the headwater of knowledge reminisce the contemporary scenarios: the current **empiric** trend of AI and Chomsky [1,4,5] *nativist* theory. Empiricist models yield no attention to symbols and grammatical rules because they have no primary symbolic structure embedded in a different insight for nativists, who underline form to appreciate substance. Nativists swear that there are innate structures from the moment of childbirth.

The current trend of AI agrees with the view of Aristotle, returned and shared by the more recent Locke. Known as **empiricism**, this theory states that no innateness is required and that learning and experience are essentially all that is needed to develop intelligence, which comes from patterns of sensory experience and interactions with the world. A current example of an empiricist model is the Transformers-models [6], which have no specific a priori knowledge about space, time, or objects besides what is represented in the training corpus. This fundamental idea is the antithesis of Chomsky, who explained the phenomenon of innate linguistics through universal grammar [1].

check for updates

Citation: Ranaldi, L.; Pucci, G. Knowing Knowledge: Epistemological Study of Knowledge in Transformers. *Appl. Sci.* **2023**, *13*, 677. https://doi.org/10.3390/app13020677

Academic Editor: Zeng Nianyin

Received: 1 December 2022
Revised: 23 December 2022
Accepted: 28 December 2022
Published: 4 January 2023

Someone, (un)knowingly, has tested the empiristic theory by building huge Neural Networks and catching what they learn. One example is GPT-model [7]. Based on Transformers-model, it has not specified a priori knowledge about space, time, or objects beyond what is represented in the training corpus. The authors of GPT intend to make a machine converse like a human. However, if these models could converse and reason, the ideas claimed by LeCun et al. [8] might have a solid ground. Unfortunately, things often need to be revised. In essence, the GPT-model, like its Transformers-model relatives, has been a straightforward experiment in the empiricist theory, but so faraway, it has not performed because with enormous computing and massive datasets, the knowledge that it gains is **superficial** and **unreliable** [9,10].

Somewhat than supporting the blank-slate view, Transformers-models appears to be accidental counter-evidence to that view. Moreover, it seems like it could be better news for the symbol-free thought-vector view. Vector-based systems can predict word categories, but they need to embody thoughts in a sufficiently reliable way. Current systems can regurgitate knowledge, but they cannot understand in a developing story; however, they adhere to the hook that emerges from the [3,11–13]. A quarter of a century has passed since Elman et al. [14] tried to use Neural Networks to rethink *innateness*; however, the problems remain more or less the same as they have always been.

Today, artificial intelligence plays a crucial role in different areas of human life and focuses on creating a computer with the same human capabilities [15]. Researchers believe that in the place of highly specialized artificial intelligence will arrive the universal AI system. Therefore, to create an advanced intelligent system, it is necessary to find an answer to questions such as the one about the role of knowledge in AI systems and the underlying methodologies. Our purpose is to study how knowledge can be represented and to show how AI is strongly linked to the mechanisms of human mind and learning. In order to verify the truthfulness of our hypothesis, which states that "the truth lies in the middle", we test two types of architectures and then we try to merge them in the aim of strengthening the theory. In this way we want to prove that, despite the experience acquired by Transformers, there is still need for an innate part that does not seem to be captured by these models.

2. Subject and Methods

The domain of philosophy of knowledge places the human mind face to face with two major theories of knowledge acquisition: nativism and empiricism. In this paper, we try to outline the issues that together constitute a thematic field of the article:

- gnoseological nature of human knowledge;
- origin and encoding of human knowledge in Artificial Neural Networks;
- specificity of artificial knowledge based on symbols and numbers;
- relationship between Human Neural Networks and Artificial Neural Networks;

In order to produce an answer, it is necessary to use the advantages of methods such as theoretical generalization and empirical observation.

Theoretical generalization will be amply covered in the following section, pursued by empirical observations on models spirited to the theories of innatist and empiristic thinking.

3. Obtained Results

Before the formation of AI, the only compromise of studying was the human mind dominated by symbols. Instead, with modern AI technologies, everything has been canceled. In order to study the origins and future developments of Machine Learning and Natural Language Processing, the following sections will examine the nativist and empiricist viewpoints, the representations through which these theories manifest themselves, and a point of intersection.

3.1. Innateness

The theories of innateness originated with Plato, who, in his works, defined the theory of inborn knowledge and gave birth to a current of thought on which many thinkers have based themselves. Symbolic in this sense is the platonic dialogue *Meno*, where the main character, i.e., Socrates, poses a mathematical problem (known as *The Learner's Paradox*) to an enslaved person. Although he did not know the principles of geometry, he managed to answer them. In this way, Plato demonstrated that concepts and ideas are present in the human mind before birth. Therefore, *"seeking and learning is nothing but recollection"*. Plato tried to explain the inactivity of knowledge in the human brain by defining it as a *"receptacle of all that comes to be"* [16]. In this space, matter takes form, and symbols take on meaning thanks to the ideas and thoughts innately embedded in the human brain.

Centuries later, with the same fundamental ideas from the study of symbols in the form of linguistic phenomena and following nativist theories, we came to realize that although languages vary, they share many universal structural properties [4,17]. According to Fodor [18], the mind is divided into modules, i.e., separate and specialized mental faculties. There is a module for sensation and perception, another for volition, another for memory, and one for language.

The faculty of language, which for Chomsky is innate [5], resides in the language module and is constantly evolving. Chomsky argued that human children could not acquire human language unless they were born with a "language acquisition device" [1], or what Steven Pinker [2] called a "language instinct". Moreover, Chomsky in "Review of B. F. Skinner's Verbal Behavior" [19], claimed that the theory of the stimulus-response mechanism used by behaviorism could theoretically only explain the capacity to reproduce sentences that had already been heard, but not the capacity to produce new sentences. Therefore, to mitigate the hard view of innateness, we can assert that humans–unlike animals–are predisposed to learn, starting with the innate mechanism with *which they are equipped from birth*. Even if there is a part of the experience in processing linguistic knowledge, the underlying structures can be defined by recurrent and universal patterns [20]. Moreover, it seems that specific linguistic properties are innate because they appear universally in the absence of corresponding experience [21].

In the following years, these works were made computable and continued by Cristianini et al. [22], Moschitti [23] and Collins et al. [24]. They had the common goal of studying the relationships between the symbols that govern transcribed language and the underlying syntactic structures.

The long studies on syntactic theories over the years have produced perfect functional symbolic representations because they are very light and comprehensible. In the following years, several methodologies have been developed to improve the type of representations and make them computationally lightweight [25–28]. Although accurate to reality, these representations are only a different encoding of the input, which, given the lack of learning by experience, does not change. In a parallel way, there are several techniques for learning representations of entities and relations in knowledge graphs, the although they are of great help, they are still very strict [29], and learning is very challenging [30].

3.2. No-Innateness

On the other hand, there is the theory derived from the thoughts of Aristotle, who can be classed as a *"tabula rasa"* empiricist, for he rejected the claim that we have innate ideas or principles of reasoning.

The deductive mechanism that enables human minds to learn works only in the presence of an inductive mechanism, as stated by Aristotle in the second book of the "Posterior Analytics"[31] and in the "Physics"[32]. Aristotle explained knowledge as a deductive matter that needs a solid inductive basis to function correctly. Matter takes shape through experience, just as symbols take on meaning after acquisition.

Aristotle's "Logic" [33] guarantees the deductive process as long as there are well-defined assumptions. This mechanism makes us think of an idea close to that of AI, more precisely, Machine Learning. Aristotle interpreted the study of thought as the basis of knowledge; thinking about the processes of forming concrete proofs, he developed a non-formalized system of syllogisms and used it to design proof procedures. Aristotle's ideas have been the founding pillars for studying the formal axiomatization of logical reasoning, which, added to a *"tabula rasa"* knowledge, allows the human being to think and be seen as a physical system that takes shape through substance. Like nativist theories, many have investigated the mechanisms of evolution, placing an essential millstone on experience [34].

Nowadays, the long linguistic study about the evolution of human language has brought significant achievements also in the field of artificial intelligence. More to the point, the union of LMs and the recursiveness of RNNs into a single architecture that can be adapted specifically to each submodel: Transformers-models [6]. They are only based on experience and seem to be achieving state-of-the-art results in many downstream tasks. The knowledge of the Transformers-models comes only from the experience they learn by looking at huge symbolic corpora. This way of working follows the empiricist theories widely studied since Aristotle. Thus, the Transformers-models are the proven proof that these theories can work, as they claim that knowledge can be constructed by experience alone. All this is great, if and only if things worked out.

Unfortunately, this is not always the case, considering that these architectures, although very good learners, tend to adapt very much to shallow heuristics [11]. The knowledge they learn is very superficial [9,10], in fact, in hostile contexts - where an acception can totally change the meaning of a sentence - they do not work; however, they perform well in long contexts, only in the presence of many resources [35]. These surveys stem from the fact that the only important thing for neural network is to maximize or minimize a cost function, regardless of the task on which they work. Thus, with Transformers-models we have a representation ready to be used by neural networks; it seems to encode—even if in unclear ways–syntactic and semantic information [36]; moreover, they are very computationally expensive.

3.3. The Truth Lies in the Middle

All thinkers of every time would recognize that genius and experience are not separate but work together. As the nativists, empiricists would not doubt that we were born with a specific biological machinery that enables us to learn.

Instead, the original motivation of the universal grammars - that initialize the stream of thought of the modern innateness theory–arises because the language has been shaped to fit the human brain, rather than vice versa. Following Darwin, we view language itself as a complex and interdependent "organism", which evolves under selectional pressures from human learning and processing mechanisms. That is, languages themselves are shaped by severe selectional pressure from each generation of language users and learners. This suggests that apparently arbitrary aspects of linguistic structure may result from general learning and processing biases deriving from the structure of thought processes, perception factors, cognitive limitations, and pragmatics [37].

Indeed, Chomsky's famous *"Language Acquisition Device"* [1], in another view, should be seen precisely as an **innate learning mechanism** which takes shape from experiences. In the final analysis, it has been supported the stance for a significant part of the innate knowledge provided to humans consists of learning mechanisms, i. e. forms of innateness that enable learning [2,38,39]. From these insights, it is clear how the two theories are related to human minds.

An interesting question relates to the basic argument of this article: do artificially intelligent systems need to be equipped with significant amounts of innate machinery, or is it sufficient for such systems, given the powerful Machine Learning systems recently developed, to work in *"tabula rasa"* mode? An answer to this question has been investigated for years by researchers in various fields: from psychology to neuroscience, as well as traditional linguistics and today's Natural Language Processing.

There are studies that claim that children are endowed early in life with a "knowledge base", as stated by Spelke [40]. Indeed, children have some ability to trace and reason about objects, which is unlikely to arise through associative learning. It seems that the brain of an 8-month-old child can learn and identify abstract syntactic rules with only a few minutes of exposure to artificial grammar [41]. Another work has suggested that deaf children can invent language without any direct model [42], and that a language can be selectively impaired even in children with normal cognitive function [43].

Humans are not precise machines, so many questions remain unanswered. We do not know if at the moment of birth we are equipped with Chomsky's machine or if we have a machine set up for learning language. On the other hand, even if we could demonstrate the evidence of innatism, it would not mean that there is no learning.

Probably, the presence of innatism means that any learning takes place against a background of certain mechanisms that precede learning. The other side of the coin is: can the *"innate machinery"* be an ingredient for a human-like artificial intelligence?

4. Experiments

In order to test the "The truth lies in the middle" hypothesis introduced in Section 3.3, we set up a series of experiments where we test the union of universal syntactic knowledge in KERMIT and knowledge derived totally from experience embedded in BERT. Figure 1 shows the proposed architectures: on the left with the green block is KERMIT, on the right with the yellow block is BERT, and finally, the entire architecture with BERT+KERMIT.

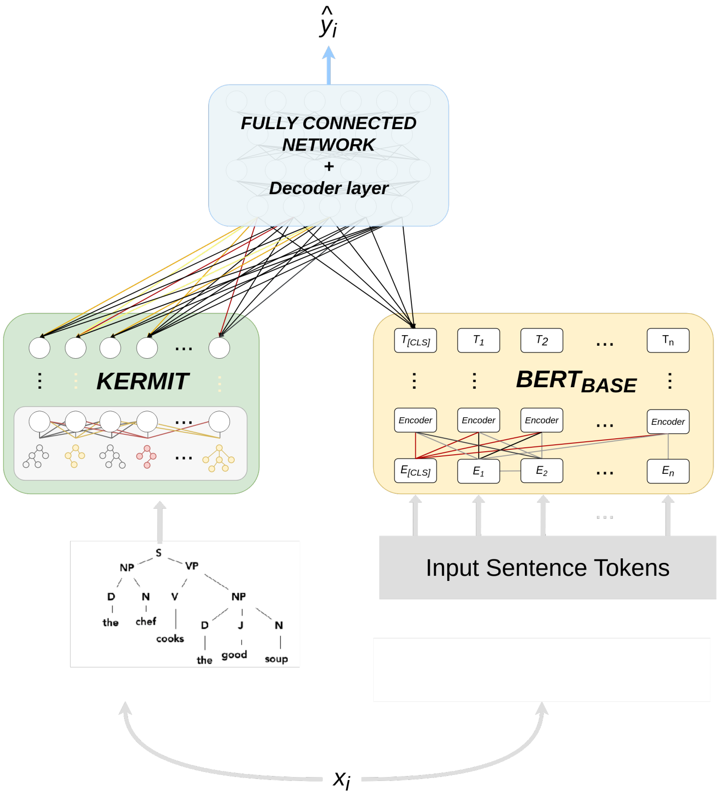

Figure 1. Proposed architecture formed by the modules of BERT and KERMIT.

4.1. KERMIT

To study the role of universal syntactic knowledge, the Kernel-inspired Encoder with Recursive Mechanism for Interpretable Trees (KERMIT) [13,28] was used. This model proposes a mechanism for incorporating syntactic representations of parse trees. The embeddings produced are ready for use in neural networks.

The version used in the experiments encodes parse trees into vectors of $4,000$ dimensions. KERMIT exploits parse trees produced by the CoreNLP parser [44].

The universal syntactic representations produced by KERMIT were given as input to a Feed Forward Neural Network (FFNN) consisting of two hidden layers of size 4000 and 2000, respectively. Finally, the output layer's size equals the number of classes. ReLU activation function and a dropout of 0.1 is used between each layer to evade overfitting the training data.

4.2. Transformer

Transformer-based architectures reach state-of-the-art in many downstream text classification tasks.

In general, the Transformer model is based on the encoder-decoder architecture [6]. The encoder is responsible for scrolling through the time steps of the input and encoding the entire sequence into a fixed-length vector called the context vector. The decoder is responsible for scrolling the output time steps by reading the context vector. The fundamental mechanism of these architectures is self-attention. This mechanism allows output to focus attention on input while producing output. In contrast, the self-attention model allows inputs to interact with each other (i.e., calculate the attention of all other inputs wrt one input. The first step is multiplying each of the encoder input vectors with three weights matrices $(W(Q), W(K), W(V))$ that we trained during the training process. This matrix multiplication will give us three vectors for each input vector: the key vector, the query vector, and the value vector. The second step in calculating self-attention is multiplying the Query vector of the current input with the key vectors from other inputs. In the third step, we will divide the score by the square root of the dimensions of the key vector (dk). For example, in the paper, the dimension of the key vector is 64, so that will be 8. The reason behind that is if the dot products become large, this causes some self-attention scores to be minimal after we apply the softmax function in the future. In the fourth step, we will apply the softmax function on all self-attention scores we calculated wrt the query word (here, first word). In the fifth step, we multiply the value vector on the vector we calculated in the previous step. In the final step, we sum up the weighted value vectors we got in the previous step, giving us the self-attention output for the given the word. The above procedure is applied to all the input sequences. Mathematically, the self-attention matrix for input matrices (Q, K, V) is calculated as:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

Q, K, and V are the concatenation of query, key, and value vectors. In the attention paper [6], the authors proposed another type of attention mechanism called multi-headed attention. This mechanism is based on a distributed variant of attention. The transformer architecture uses the attention model uses multi-headed attention in three steps:

1. encoder-decoder attention layers, in this type of layer, the queries come from the previous decoder layer while the keys and values come from the encoder output. This allows each position in the decoder to give attention to all the positions of the input sequence.
2. self-attention layer contained in the encoder receives key, value, and query input from the output of the previous encoder layer. Each position in the encoder can get an attention score from every position in the previous encoder layer.
3. self-attention in the decoder, this is similar to self-attention in the encoder where all queries, keys, and values come from the previous layer. The self-attention decoder allows each position to attend each position up to and including that position. The future values are masked with (-Inf). This is known as masked-self attention.

BERT

In this paper, one of the most well-known Transformers-based models was tested: Bidirectional Encoder Representations from Transformers (BERT) [45]. The BERT architecture is trained on the BooksCorpus [46] and the English Wikipedia. The BERT model was implemented using the Transformers library of Huggingface [47]. The output of each encoder was decoded by a Feed-Forward Neural Network (FFNN) to a single output layer with an input size equal to 768 and output equal to the number of classes with the classifier function. For the fine-tuning phase, the number of epochs was set to 5, and the optimizer Adam [48] was used.

4.3. Experimental Set-Up

Using the earlier architectures, we tested the hypothesis introduced in Section 3.3 on four downstream tasks: Corpus of Linguistic Acceptability (CoLA) [49], the Semantic Textual Similarity Benchmark (STS-B) [50], Quora Question Pairs (QQP) [51] and Winograd NLI (WNLI) [52] datasets.

The respective versions of train, validation, and test for each task were downloaded directly from Huggingface [47]. Then, the evaluation was done by extracting the accuracy of the classification results, defined as the number of correct predictions divided by the number of total predictions.

4.4. Discussion

Transformer-based models are greatly influencing the world of NLP. Transformers are adept at solving semantic, syntactic, and even stylistic tasks. Transformer-based models, in general, are adapted to the specific task with an appropriate fine-tuning step [53]. These steps seem to work very well, and transformers solve downstream tasks optimally, even when it comes to purely stylistic tasks [12]. The key to success would be pre-training on massive corpora with substantial resources. This learning mode is reminiscent of the concepts defined in Section 3.2. In these experiments, we took a group of tasks from the GLUE family to see if knowledge derived totally from experience and innate derived knowledge can coexist. Specifically, we combined these two types of knowledge by observing downstream performance on four previously introduced tasks. From the results in Table 1, we can observe that the experience-based model, in this case, *BERT*, generally performs well. In contrast, KERMIT only sometimes does well. The key is enshrined in the combination of the two models in *BERT+KERMIT*. In this case, we can observe that three times out of four, *BERT+KERMIT* gets the best results.We can infer that the "The truth lies in the middle" hypothesis introduced in Section 3.3 is true for this set-up of experiments.

Table 1. Comparison of different performances on GLUE task. The experiments were run five different times with different seeds.

Model	CoLa	STSB	QQP	WNLI
BERT	67.5(±0.8)	71.2(±1.2)	82.3(±1.3)	86.3(±0.6)
KERMIT	64.3(±0.7)	68.3(±0.9)	76.5(±0.7)	83.4(±1.3)
BERT+KERMIT	68.6(±0.9)	69.6(±1.3)	83.4(±1.5)	86.5(±1.6)

5. Conclusions

Thanks to the involvement of several branches of knowledge, it is possible to look at the problem and analyze it more profoundly to maximize the results and reduce the margin of error. More specifically, in this paper–which moves from analyzing theories on the origin and development of human knowledge to understanding how modern artificial minds have inherited this–we have resorted to disciplines distant from the world of artificial intelligence, such as philosophy and psycholinguistics. The first of the analogies detected and analyzed concerns the closeness–in terms of concept– between Plato's nativist theory, the study of the symbolic-syntactic structures of artificial minds, and the innate presence in the human mind of a language acquisition device–which can be linked to the theory of modular mind. On the other hand, the cornerstones of the second analogy are the empiricist theory of Aristotle, the vector representations of artificial minds with their related experiments, namely Transformer-models, and the constructivist theory.

The study of the various nuances regarding the analogy between human and artificial intelligence and knowledge, as well as the dialectical movements present in both, leads us to the formulation of a global synthesis related to Cognitive Modeling and Psycholinguistics: probably, the presence of innatism means that any learning occurs in the presence of mechanisms that precede learning itself; but on the other hand, can the "innate mechanism" be a component of "human-like artificial intelligence"?

So, in support of the multi-disciplinary approach and the interpenetration of various fields of knowledge as an optimal method of research and analysis, we consider that in order to understand and solve a problem sincerely, it is necessary to look at it from afar– to have multiple perspectives on the same.

"The distance that makes objects smaller to the eye enlarges them at the thought".

Author Contributions: Conceptualization, L.R. and G.P.; methodology, L.R.; software, L.R.; validation, G.P.; formal analysis, G.P.; investigation, L.R.; resources, L.R.; data curation, L.R.; writing—original draft preparation, G.P.; writing—review and editing, L.R.; funding acquisition, L.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data are available on the Huggingface library.

Conflicts of Interest: The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Chomsky, N. *Aspects of the Theory of Syntax*; The MIT Press: Cambridge, UK, 1965.
2. Pinker, S.; Jackendoff, R. The faculty of language: What's special about it? *Cognition* **2005**, *95*, 201–236. [CrossRef]
3. Ranaldi, L.; Fallucchi, F.; Zanzotto, F.M. Dis-Cover AI Minds to Preserve Human Knowledge. *Future Internet* **2022**, *14*, 10. [CrossRef]
4. Chomsky, N. *Syntactic Structures*; Cambridge Center for Behavioral Studies (CCBS): Littleton, MA, USA, 1957.
5. Chomsky, N. On certain formal properties of grammars. *Inf. Control* **1959**, *2*, 137–167. [CrossRef]
6. 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:cs.CL/1706.03762.
7. 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:cs.CL/2005.14165.
8. LeCun, Y.; Bengio, Y.; Hinton, G.E. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
9. Sinha, K.; Sodhani, S.; Dong, J.; Pineau, J.; Hamilton, W.L. CLUTRR: A Diagnostic Benchmark for Inductive Reasoning from Text. *arXiv* **2019**, arXiv:cs.LG/1908.06177.
10. Talmor, A.; Elazar, Y.; Goldberg, Y.; Berant, J. oLMpics–On what Language Model Pre-training Captures. *arXiv* **2020**, arXiv:cs.CL/1912.13283.
11. McCoy, T.; Pavlick, E.; Linzen, T. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; Association for Computational Linguistics; pp. 3428–3448. [CrossRef]
12. Ranaldi, L.; Ranaldi, F.; Fallucchi, F.; Zanzotto, F.M. Shedding Light on the Dark Web: Authorship Attribution in Radical Forums. *Information* **2022**, *13*, 435. [CrossRef]
13. Ranaldi, L., Fallucchi, F., Santilli, A.; Zanzotto, F. KERMITviz: Visualizing Neural Network Activations on Syntactic Trees. *Metadata Semant. Res.*. **2022**, 139–147.
14. Elman, J.L. Finding structure in time. *Cogn. Sci.* **1990**, *14*, 179–211. [CrossRef]
15. Jiang, Y.; Li, X.; Luo, H.; Yin, S.; Kaynak, O. Quo vadis artificial intelligence? *Discov. Artif. Intell.* **2022**, *2*, 1–19. [CrossRef]
16. Zeyl, D.; Sattler, B. Plato's Timaeus. In *The Stanford Encyclopedia of Philosophy*; Summer 2019 ed.; Zalta, E.N., Ed.; Metaphysics Research Lab, Stanford University: Stanford, CA, USA, 2019.

17. Newmeyer, F.J. Explaining language universals. *J. Linguist.* **1990**, *26*, 203–222. [CrossRef]
18. Fodor, J.A. *The Modularity of Mind: An Essay on Faculty Psychology*; MIT Press: Cambridge, MA, USA, 1983.
19. Chomsky, N. A Review of B. F. Skinner's Verbal Behavior. In *Readings in the Psychology of Language*; Jakobovits, L.A., Miron, M.S., Eds.; Prentice-Hall: Englewood Cliffs, NJ, USA, 1967.
20. White, L. Second Language Acquisition and Universal Grammar. *Stud. Second Lang. Acquis.* **1990**, *12*, 121–133. [CrossRef]
21. Lowenthal, F. Logic and language acquisition. *Behav. Brain Sci.* **1991**, *14*, 626–627. [CrossRef]
22. Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*; Cambridge University Press: Cambridge, UK, 2000.
23. Moschitti, A. Making Tree Kernels practical for Natural Language Learning. In In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics EACL'06, Trento, Italy, 3–7 April 2006.
24. Collins, M.; Duffy, N. New Ranking Algorithms for Parsing and Tagging: Kernels over Discrete Structures, and the Voted Perceptron. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, PA, USA, 6–12 July 2002.
25. Culotta, A.; Sorensen, J. Dependency Tree Kernels for Relation Extraction. In Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, Stroudsburg, PA, USA, 21–26 July 2004; Association for Computational Linguistics, ACL'04; p. 423-es. [CrossRef]
26. Pighin, D.; Moschitti, A. On Reverse Feature Engineering of Syntactic Tree Kernels. In Proceedings of the Fourteenth Conference on Computational Natural Language Learning; Association for Computational Linguistics, Uppsala, Sweden, 15–16 July 2010; pp. 223–233.
27. Zanzotto, F.M.; Dell'Arciprete, L. Distributed Tree Kernels. *arXiv* **2012**, arXiv:cs.LG/1206.4607.
28. Zanzotto, F.M.; Santilli, A.; Ranaldi, L.; Onorati, D.; Tommasino, P.; Fallucchi, F. KERMIT: Complementing Transformer Architectures with Encoders of Explicit Syntactic Interpretations. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; Association for Computational Linguistics; pp. 256–267.
29. Bonner, S.; Barrett, I.P.; Ye, C.; Swiers, R.; Engkvist, O.; Hoyt, C.T.; Hamilton, W.L. Understanding the performance of knowledge graph embeddings in drug discovery. *Artif. Intell. Life Sci.* **2022**, *2*, 100036. [CrossRef]
30. Chen, S.; Liu, X.; Gao, J.; Jiao, J.; Zhang, R.; Ji, Y. HittER: Hierarchical Transformers for Knowledge Graph Embeddings. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), Punta Cana, Dominican Republic, 7–11 November 2021; Association for Computational Linguistics.
31. Pellegrin, P., Definition in Aristotle's Posterior Analytics. In *Being, Nature, and Life in Aristotle: Essays in Honor of Allan Gotthelf*; Lennox, J.G., Bolton, R., Eds.; Cambridge University Press: Cambridge, UK, 2010; pp. 122–146. [CrossRef]
32. Bodnar, I. Aristotle's Natural Philosophy. In *The Stanford Encyclopedia of Philosophy*; Spring 2018 ed.; Zalta, E.N., Ed.; Metaphysics Research Lab, Stanford University: Stanford, CA, USA, 2018.
33. Smith, R. Aristotle's Logic. In *The Stanford Encyclopedia of Philosophy*; Fall 2020 ed.; Zalta, E.N., Ed.; Metaphysics Research Lab, Stanford University: Stanford, CA, USA, 2020.
34. Pinker, S.; Bloom, P. Natural language and natural selection. *Behav. Brain Sci.* **1990**, *13*, 707–727. [CrossRef]
35. Dai, Z.; Yang, Z.; Yang, Y.; Carbonell, J.; Le, Q.V.; Salakhutdinov, R. Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context. *arXiv* **2019**, arXiv:cs.LG/1901.02860.
36. Hewitt, J.; Manning, C.D. A Structural Probe for Finding Syntax in Word Representations. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; Volume 1 (Long and Short Papers), Association for Computational Linguistics; pp. 4129–4138. [CrossRef]
37. Christiansen, M.H.; Chater, N. Language as shaped by the brain. *Behav. Brain Sci.* **2008**, *31*, 489–509. [CrossRef]
38. Marler, P. Innateness and the instinct to learn. *An. Acad. Bras. Ciênc.* **2004**, *76*, 189–200. [CrossRef]
39. Marcus, G. Innateness, AlphaZero, and Artificial Intelligence. *arXiv* **2018**, arXiv:cs.AI/1801.05667.
40. Spelke, E.S.; Kinzler, K.D. Core knowledge. *Dev. Sci.* **2007**, *10*, 89–96. [CrossRef] [PubMed]
41. Gervain, J.; Berent, I.; Werker, J.F. Binding at Birth: The Newborn Brain Detects Identity Relations and Sequential Position in Speech. *J. Cogn. Neurosci.* **2012**, *24*, 564–574. [CrossRef] [PubMed]
42. Senghas, A.; Kita, S.; Özyürek, A. Children Creating Core Properties of Language: Evidence from an Emerging Sign Language in Nicaragua. *Science* **2004**, *305*, 1779–1782. [CrossRef] [PubMed]
43. Lely, H.; Pinker, S. The biological basis of language: Insight from developmental grammatical impairments. *Trends Cogn. Sci.* **2014**, *18*, 586–595. [CrossRef]
44. Zhu, M.; Zhang, Y.; Chen, W.; Zhang, M.; Zhu, J. Fast and Accurate Shift-Reduce Constituent Parsing. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria, 4–9 August 2013; Volume 1, Long Papers, Association for Computational Linguistics; pp. 434–443.
45. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; Volume 1, (Long and Short Papers), Association for Computational Linguistics; pp. 4171–4186. [CrossRef]
46. Zhu, Y.; Kiros, R.; Zemel, R.; Salakhutdinov, R.; Urtasun, R.; Torralba, A.; Fidler, S. Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books. *arXiv* **2015**, arXiv:cs.CV/1506.06724.

47. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv* **2019**, arXiv:abs/1910.03771.
48. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *CoRR* **2015**, arXiv:abs/1412.6980.
49. Warstadt, A.; Singh, A.; Bowman, S.R. Neural Network Acceptability Judgments. *Trans. Assoc. Comput. Linguist.* **2019**, *7*, 625–641. [CrossRef]
50. Cer, D.; Diab, M.; Agirre, E.; Lopez-Gazpio, I.; Specia, L. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Vancouver, BC, Canada, 3–4 August 2017; Association for Computational Linguistics; pp. 1–14. [CrossRef]
51. Sharma, L.; Graesser, L.; Nangia, N.; Evci, U. Natural Language Understanding with the Quora Question Pairs Dataset. *arXiv* **2019**, arXiv:abs/1907.01041.
52. Levesque, H.J.; Davis, E.; Morgenstern, L. The Winograd Schema Challenge. In Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning, Rome, Italy, 10–14 June 2012; AAAI Press, 2012; KR'12; pp. 552–561.
53. Podkorytov, M.; Biś, D.; Liu, X. How Can the [MASK] Know? The Sources and Limitations of Knowledge in BERT. In Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN), Shenzhen, China, 18–22 July 2021; pp. 1–8. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.