大規模言語モデルと ISA アプローチ

荒井 柚月 [1]　津川 翔 [2]
[1] 筑波大学情報学群情報メディア創成学類　[2] 筑波大学 システム情報系
y-arai@snlab.cs.tsukuba.ac.jp　s-tugawa@cs.tsukuba.ac.jp

概要

人間中心主義的に展開されてきた言語の哲学は、ChatGPT（OpenAI）、Claude（Anthropic）といった人間に比肩する言語的能力を持つとされる大規模言語モデル（Large Language Models, LLMs）の出現によって、脱人間中心主義化を迫られている。従来はその基礎的意味論として分布意味論があてがわれてきた LLM であるが、現在では、LLM の基礎的意味論として分布意味論以外の基礎的意味論を探る研究が続けられている。本発表は、言語の表象性という観点から、言語モデルに最適な基礎的意味論としてロバート・ブランダムの推論的意味論を提案し、推論的意味論の反表象主義性や論理的表出主義性が、LLM の性質や振る舞いを解釈する上で有用であることを示す。

1　はじめに

人間中心主義的に展開されてきた言語の哲学は、ChatGPT（OpenAI）[1)]、Claude（Anthropic）[2)]、Gemini（Google）[3)]、Microsoft Copilot（Microsoft）[4)] といった人間に比肩する言語的能力を持つとされる大規模言語モデル（Large Language Models, LLMs）(Zhao et al. 2023) の出現によって、脱人間中心主義化を迫られている (Cappelen and Dever 2021; Millière and Buckner 2024)。従来は、分布意味論が LLM の基礎的意味論としてあてがわれてきた (Enyan et al. 2024; Grindrod 2024; Havlík 2024; Lenci and Sahlgren 2023)。しかし現在では、LLM の基礎的意味論として分布意味論以外の基礎的意味論を探る研究が続けられている (Grindrod 2024; Mallory 2023)。本発表は、この基礎的意味論の脱人間中心主義化という潮流のうちで、言語の表象性という観点から、言語モデルに最適な基礎的意味論としてロバート・ブランダムの推論的意味論 (Brandom 1994) を提案するものである。言語の本質と意味の形成過程に関する哲学的探究において、真理条件意味論と推論的意味論は二つの主要なアプローチとして対立してきた (Brandom 2010, chap. 5.2)。前者は表象主義を、後者は反表象主義を標榜する意味論であるが、その間の対立は認識論における対立でもあり、言語と世界の関係性について根本的に異なる見解を示すものである。

表象主義は、言語を世界の真理を映す鏡として捉える立場である。表象主義的意味論の筆頭である真理条件意味論は、文の意味をその真理条件、すなわちそれが真となるために必要かつ十分な条件として定義する (Heim and Kratzer 1998)。真理条件意味論の一つであるモンタギュー意味論は、意味の合成性を前提とする形式意味論でもあり、論理学を基盤としていることが特徴である。モンタギュー意味論はモデル論的であり、言語外在的な可能世界や個物、普遍といったカテゴリー内の存在間の対応の集合をモデルとして与える外在主義的意味論でもある (Partee 2016)。一方、反表象主義は、表象主義的な意味論が帰結してしまう「意味に対する懐疑主義」(Kripke 1982) を避けるために、初めから言語外在的な表象を否定する立場であり、その源流はリチャード・ローティの『哲学と自然の鏡』(Rorty 1979) に求められる。反表象主義的意味論の一つである Brandom (1994) の推論的意味論は、意味の意味を言語外在的な特権的表象に求めず、言語使用者の間における規範や推論上の役割に求める意味論である。

本発表は LLM の反表象主義性を指摘し、その上で、ロバート・ブランダムによって提唱された推論的意味論 (ibid.) が、言語モデルの基礎的意味論として従来採用されてきた分布意味論や、現在の言語哲学で優勢である真理条件意味論よりも、LLM の機能や性質を説明する上で有効であると主張する。この主張は、LLM の性質が意味論的外在主義や意味の合成性といった主流の言語哲学上の前提に対して疑問を投げかけるものであるという主張にほかなら

1) https://openai.com/index/chatgpt/
2) https://claude.ai
3) https://gemini.google.com
4) https://copilot.microsoft.com/

ない。もし本発表の主張が成功しているとすれば、反表象主義的な言語観の妥当性が再評価され、言語哲学の新たな展開の可能性が提示されることになるだろう。

2 LLM

OpenAI の GPT 4 や Anthropic の Claude、Meta の Llama 3[5]、InstructGPT (Ouyang et al. 2022) といった LLM は、Vaswani et al. (2017) が提案した Transformer と呼ばれるアーキテクチャに基づいており、特に複数の頭・層を備えた Transformer を多頭・多層 Transformer という[6]。多頭・多層 Transformer は、古典的記号主義と古典的結合主義の性質を併せ持ったものである。記号主義（計算主義）とは、人間の思考を離散的表象の合成や交換の連鎖であると見なす立場を指す。反対に結合主義は、ニューラルネットワークの上で表象が分散的に存在するとする立場を指す。LLM 以前のいわゆる GOFAI（Good Old-Fashioned Artificial Intelligence）は、記号主義的パラダイムに属し、離散的記号の演算によって自然言語の処理を指向するものであった。これに対して言語モデルは、人工的ニューラルネットワークをそのアーキテクチャの基礎とし、分散表象の演算によって言語処理を行う結合主義的パラダイムに属しながらも、高位の層において構文や前方照応などの記号主義的な振る舞いも示すものである[7]。Chalmers (2023) は、このような LLM の性質が「半記号主義（subsymbolism）」(Smolensky 1987) に属するものであると述べている。

3 LLM の基礎的意味論

言語の哲学において、意味論は基礎的意味論と記述的意味論に区別される (Stalnaker 1997, p.535)[8]。基礎的意味論とは、まさに意味の意味を与える理論である。一方、記述的意味論は、与えられた意味論の中で、文などに対して意味論的価値を与える理論を指す。

現在、LLM の基礎的意味論として与えられるのは、アメリカ構造主義言語学の潮流において Firth (1957) や Harris (1954) により建設された、分布意味論である (Grindrod 2024)。分布意味論とは、ある語彙項目の意味が、それが出現する場所の周りの分布により与えられるとする意味の理論である (Grindrod 2023)。現在の LLM（LLM）や分布意味論モデル（DSM）は、この分布意味論を意味の理論として採用したモデルであり、文章における単語間の共起頻度といった言語的要素の統計的共起確率をその根底に措いている。現在の一般的な Transformer ベースの LLM は、この言語的要素の埋め込みベクトルの列を入力とし、埋め込みベクトルの列を返す[9]。

しかしながら分布意味論は、LLM の振る舞いを説明するためには不十分であると考えられる。たとえば Enyan et al. (2024) は、計算言語学における自身らの研究成果を踏まえた上で、分布意味論が LLM の振る舞いを説明するには不十分であり、分布意味論のさらなる洗練が必要であると述べている。また線形表象仮説 (Mikolov, Yih, and Zweig 2013; Park et al. 2024) によれば、言語モデルのなかでは部分空間が概念を表象しているのであり、これは通常概念などの存在を考慮しない分布意味論のさらなる拡張の必要性を示すものと言える。本発表は、分布意味論ではなく推論的意味論を言語モデルの基礎的意味論としてあてがうことによって、言語モデルの振る舞いや性質を説明する上で、どのような部分が上手くいき、どのような部分が上手くいかないかを調べるものである。

4 推論主義

推論主義（inferentialism）とは、ロバート・ブランダムの著作である『Making it Explicit』(Brandom 1994) において提唱された言語の哲学における反表象主義的理論の一つであり、主に推論的意味論や規範的語用論から構成される。なお紙幅の都合上、推論主義自体に関する説明は省略する。推論主義の解説書としては、たとえば Bouche (2020), Weiss and Wanderer (2010), and 白川 (2021) を参照。

5） https://ai.meta.com/blog/meta-llama-3/

6） Transformer アーキテクチャの詳細な数理的説明は、Elhage et al. (2021) を参照。

7） GOFAI と LLM の比較については、Gubelmann (2023) を参照。

8） 基礎的意味論はメタ意味論、記述的意味論は意味論とも呼ばれる (Kaplan 1989, p.573–4)。

9） 最先端のモデルでは、Byte-Level BPE （Byte Pair Encoding） (Wang, Cho, and Gu 2020) という埋め込み手法が用いられる場合がある。この埋め込み手法は、文字や単語、文といった水準ではなく、バイトの水準において、頻出する文字や文字列のペアを繰り返し結合し、新しいトークンとして扱うというものである。意味の形而上学にとってこの埋め込み手法が何を意味するのかは今後の検討課題である。

5　LLM と ISA アプローチ

本節では、LLM と推論主義の ISA（Inferece、Substitution、Anaphora）アプローチを結び付けて論じながら、LLM の性質が推論主義の論理的表出主義および反表象主義性を支持することを示す。

5.1　推論

古典的記号主義においては形式的推論が推論であり、その統語論や意味論は論理学によって記述されるものであった。これに対して Transformer アーキテクチャは、形式的論理ではなく、高い層のヘッド群により捉えられる統計的規則を用いて推論を行う。つまり、LLM は、明示的な論理的推論規則を与えられることなく、訓練データである言語使用のパターンから推論能力を獲得するのである。この特性は実質的推論という概念と親和性が高く、形式的推論に依存する真理条件意味論とは親和性が低いことが分かる。

Sellars (1953) は、「A はリンゴである $\Longrightarrow$ A は果物である」(MI) というような実質的な推論 (material inference) を、われわれの言語的実践に不可欠であるとして擁護している。推論の形式性を重視する論者は、この推論を三段論法（enthymeme）の省略であるとして、生の形では認めることができない。彼らは、推論 MI を次の三段論法の省略形であると考える。

前提 1: A はリンゴである
前提 2: すべての X について、X がリンゴならば X は果物である
結論: A は果物である

言語モデルの推論能力は、訓練データに含まれる推論規則の実質的な使用において訓練され、モデルにおける高層の重み行列に保持されたあと、文字トークンを出力する際にそれらの重み行列が作用することによって発揮されるものである。言語モデルは、形式的推論における省略された三段論法（enthymeme）による分析といった間接的な方法で推論を行っている訳ではない。Sellars (ibid.) が擁護する実質的な推論は、それが創発的に十分であるほど訓練データに含まれている際には統計的なパターンから直接的に導かれるものなのであり、言語モデルにおける推論は形式的推論ではなく実質的推論として考えることが妥当だと言えるだろう。

言語モデルにおける推論が実質的推論であるという性質に加えて、言語モデルにおいて表出する形式的な論理的関係が、言語モデルに直接的にコーディングされたものではないという性質も重要である。言語モデル以前の記号主義に基づく GOFAI では、論理的演算子は言語処理システムに埋め込まれたものであり、我々がふだん使用するような実質的な関係を表現することが難しかった。これに対して言語モデルは、文の間に存在する実質的な論理的関係を大量の学習データから獲得し、その中に現れる論理的関係を模倣して出力する。言語モデルにおける論理的推論はモデルにコーディングされたものではなく、ニューラル・ネットワークの重みの上に随伴する性質であり、古典論理的な推論を完璧に達成するものではない。たとえば、論理的な含意関係が文の間に存在するか否かを正しく認識することができるかを問うタスクである含意関係認識（Text Entailment Recognition）タスクのスコア向上に際して、深層学習を採用する研究者は手を焼いている (Putra, Siahaan, and Saikhu 2024)。つまり、言語モデル上の論理が大量の学習データからボトムアップ的に獲得されたものであり、形式的論理がトップダウンに実装されたものではないという意味で、言語モデルの半記号主義性 (Chalmers 2023) は推論主義の論理的表出主義性に合致するということができる。

5.2　置換

推論主義において単称名辞および述語の定義は、置換推論によって与えられるものであった。しかしながら、LLM は単称名辞や述語といった言語的カテゴリーを設計的に導入しているわけではない。それらのカテゴリーは自動的に獲得され、いわば事後的に確認されるものである[10)]。LLM は、$Qa \Longrightarrow Qb \wedge Qb \Longrightarrow Qa$ ならば a と b は同じ意味を持つ単称名辞である、というような仕方で単称名辞を引き出している訳ではない。述語についても同様である。したがって、置換推論によって単称名辞と述語を抽出するという推論主義のアプローチと、LLM 内での単称名辞と述語の扱いは噛み合わせが悪いと考えられる。

ただし、言語的要素の置換可能性が意味や内容を生み出すための基盤となっていること (Adriaans 2024) と、言語モデルが誘導頭 (induction heads) (Olsson et al. 2022) や抑制頭 (suppress heads)

10)　反対に、形式的意味論であるモンタギュー意味論においては、単称名辞と述語は全く異なる言語的カテゴリーとしてアプリオリに与えられる。

(McDougall et al. 2024) に従ってトークン生成を行っていることの両者は関連していると考えられる。誘導頭と抑制頭はそれぞれ、以前にあるトークンを参照して、参照されたそれらのトークンを複製したり抑制したりする機能を持つ。一方推論主義は、たとえば「このリンゴは緑である」という命題と「このリンゴは赤である」という命題には同時にコミットすることができない（「実質的な非可換性（material incompatibility）」 (Brandom 1994; Sellars 1953)）というような、非可換性に基づく言語の規範的実践に着目している。誘導頭と抑制頭のような置換可能性に関するヘッドの役割は、「意味と内容の規範性」(Glüer, Wikforss, and Ganapini 2024) に関連している可能性がある。置換可能性に関わるヘッドの性質は、推論主義が批判する傾向性主義に陥らない形で言語モデル上の規範性を説明するための鍵となると考えられる。

5.3 前方照応

前方照応によって直示語や指示といった概念を解決する推論的意味論の仕方は、直示語の意味を世界との関係においてではなく言語の内で解決しようとするものであり、言語モデルの反表象主義性に合致するものといえる。推論主義において世界との接続は最終的に与えられるものであり、直示語は文中の語の前方照応という置換推論によって与えられるものであった。これに対して、言語データ以外を引数に持たない非マルチモーダル的な言語モデルは、言語以外のメディアを介して世界と接続することはできないため、直示や指示といった概念は言語内部に閉じていると言える。

Transformer において前方照応は、注意機構や「誘導頭（induction heads）」と呼ばれる注意頭の一種によって実現されていると考えることができる[11]。注意は、あるトークンが他のトークンに対してどれほど依存しているかを示す標準単体である。図 1 は、Transformer のエンコーダモデルの一つである BERT (Devlin et al. 2019) における、文「That pig is grunting, so it must be happy」中の指示語「it」の注意を視覚化したものである[12]。この図から分かる通り「it」は「That pig」との内積（類似度）[13]が大きく、BERT が指示語の前方照応を行っていることが分かる。

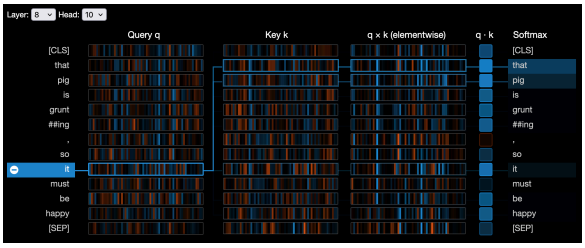

図 1 BertViz による前方照応の Neuron View。「it」と「That pig」の内積が大きく、指示語の前方照応が行われている。

また誘導頭は、「［A］［B］… ［A］」という列が与えられたとき、次トークンとして［B］を選択するといったかたちで、文中で自身より前に生起した列を複写することによって、パターンを完成させる回路である (Elhage et al. 2021; Olsson et al. 2022)。たとえば、「That pig is grunting … it」という単語列があり、注意によって指示語と被指示語は関連付けられている場合、誘導頭はその次に「grunts」という単語を選ぶ確率が高くなる。

注意や誘導頭によって、指示語が被指示語と関連付けられ、それらを囲む文脈もまた関連付けられる。言語モデルは、与えられた言語的情報によってしか世界を想像することができない。「この豚」や「それ」といった直示語の意味は、世界を指差すことによって得られているものではなく、注意による前方照応によって獲得されるものなのである。

6 おわりに

かつて、哲学者のリチャード・ローティは、「解釈学的転回」と呼ばれる概念を提唱した。それは、認識論や存在論といった哲学的な考察もまた世界を解釈するための営為に過ぎないとする道具主義的な考え方である。認識論や存在論といった哲学上の枠組みがこの世界の解釈に貢献してきたことと同様に、本発表における LLM に対する哲学的探究もまた、その性質や振る舞いの解釈に貢献することができると我々は考えている。

11） Transformer による後方参照を防ぐためには、注意マスクと呼ばれる処理（masked attention）が用いられる。Transformer エンコーダモデルである BERT (Devlin et al. 2019) では自己注意が用いられており、後方参照が可能である。反対に、Transformer デコーダモデルである GPT (Yenduri et al. 2023) ではマスク注意が用いられており、後方参照が可能でない。

12） 図は、Vig (2019) による BertViz を用いて作成した。ビューは Neuron View、事前学習モデルは ‘bert-base-uncased’ を使用した。また、図は第 8 層および第 10 頭の注意を表したものである。

13） 注意は、カーネル関数として一般化され (Tsai et al. 2019)、クエリベクトルとキーベクトルの類似度は通常の内積に限らない。たとえば (Chen et al. 2021) は、クエリベクトルとキーベクトルの類似度を、通常の内積ではなくガウシアンカーネル（L2 ノルム）として与えている。

参考文献

Adriaans, Pieter (2024). “Information”. In: **The Stanford Encyclopedia of Philosophy**. Ed. by Edward N. Zalta and Uri Nodelman. Summer 2024. Metaphysics Research Lab, Stanford University.

Bouche, Gilles, ed. (2020). **Reading Brandom: On a Spirit of Trust**. New York: Routledge.

Brandom, Robert (1994). **Making It Explicit. Reasoning, Representing, and Discursive Commitment**. Cambridge, Mass.: Harvard University Press.

— (2010). “Reply to Jerry Fodor and Ernest Lepore’s "Brandom Beleaguered"”. In: **Reading Brandom**. Ed. by Bernhard Weiss and Jeremy Wanderer. 1st ed. Accessed July 4, 2024. Routledge. Chap. 27. URL: https://www.perlego.com/book/1609062.

Cappelen, Herman and Josh Dever (2021). **Making Ai Intelligible: Philosophical Foundations**. New York, USA: Oxford University Press.

Chalmers, David J. (2023). “The Computational and the Representational Language-of-Thought Hypotheses”. In: **Behavioral and Brain Sciences** 46, e269. DOI: 10.1017/s0140525x23001796.

Chen, Yifan et al. (2021). **Skyformer: Remodel Self-Attention with Gaussian Kernel and Nyström Method**. arXiv: 2111.00035 [cs.LG]. URL: https://arxiv.org/abs/2111.00035.

Devlin, Jacob et al. (June 2019). “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, Volume 1 (Long and Short Papers)**. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423.

Elhage, Nelson et al. (2021). “A Mathematical Framework for Transformer Circuits”. In: **Transformer Circuits Thread**. https://transformer-circuits.pub/2021/framework/index.html.

Enyan, Zhang et al. (2024). **Are LLMs Models of Distributional Semantics? A Case Study on Quantifiers**. arXiv: 2410.13984 [cs.CL]. URL: https://arxiv.org/abs/2410.13984.

Firth, J. R. (1957). “A Synopsis of Linguistic Theory”. In: **Studies in Linguistic Analysis**. Blackwell, pp. 1–32.

Glüer, Kathrin, Åsa Wikforss, and Marianna Ganapini (2024). “The Normativity of Meaning and Content”. In: **The Stanford Encyclopedia of Philosophy**. Ed. by Edward N. Zalta and Uri Nodelman. Fall 2024. Metaphysics Research Lab, Stanford University.

Grindrod, Jumbly (2023). “Distributional Theories of Meaning: Experimental Philosophy of Language”. In: **Experimental Philosophy of Language: Perspectives, Methods, and Prospects**. Ed. by David Bordonaba-Plou. Springer Verlag, pp. 75–99.

— (2024). “Large language models and linguistic intentionality”. In: **Synthese** 204.2, p. 71. ISSN: 1573-0964. DOI: 10.1007/s11229-024-04723-8. URL: https://doi.org/10.1007/s11229-024-04723-8.

Gubelmann, Reto (2023). “A Loosely Wittgensteinian Conception of the Linguistic Understanding of Large Language Models Like Bert, Gpt-3, and Chatgpt”. In: **Grazer Philosophische Studien** 99.4, pp. 485–523. DOI: 10.1163/18756735-00000182.

Harris, Zellig S. (1954). “Distributional Structure”. In: **WORD** 10.2-3, pp. 146–162. DOI: 10.1080/00437956.1954.11659520. eprint: https://doi.org/10.1080/00437956.1954.11659520. URL: https://doi.org/10.1080/00437956.1954.11659520.

Havlík, Vladimír (Dec. 23, 2024). “Meaning and understanding in large language models”. In: **Synthese** 205.1, p. 9. ISSN: 1573-0964. DOI: 10.1007/s11229-024-04878-4. URL: https://doi.org/10.1007/s11229-024-04878-4 (visited on 01/08/2025).

Heim, Irene and Angelika Kratzer (1998). **Semantics in Generative Grammar**. Ed. by Angelika Kratzer. Malden, MA: Blackwell. Chap. 1.1.

Kaplan, David (1989). “Afterthoughts”. In: **Themes From Kaplan**. Ed. by Joseph Almog, John Perry, and Howard Wettstein. Oxford University Press, pp. 565–614.

Kripke, Saul A. (1982). **Wittgenstein on Rules and Private Language: An Elementary Exposition**. Cambridge: Harvard University Press.

Lenci, Alessandro and Magnus Sahlgren (2023). **Distributional Semantics**. Studies in Natural Language Processing. Cambridge University Press.

Mallory, Fintan (Nov. 2023). “Fictionalism about Chatbots”. In: **Ergo an Open Access Journal of Philosophy** 10.0. ISSN: 2330-4014. DOI: 10.3998/ergo.4668.

McDougall, Callum Stuart et al. (Nov. 2024). “Copy Suppression: Comprehensively Understanding a Motif in Language Model Attention Heads”. In: **Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP**. Ed. by Yonatan Belinkov et al. Miami, Florida, US: Association for Computational Linguistics, pp. 337–363. DOI: 10.18653/v1/2024.blackboxnlp-1.22. URL: https://aclanthology.org/2024.blackboxnlp-1.22.

Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig (June 2013). “Linguistic Regularities in Continuous Space Word Representations”. In: **Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**. Ed. by Lucy Vanderwende, Hal Daumé III, and Katrin Kirchhoff. Atlanta, Georgia: Association for Computational Linguistics, pp. 746–751. URL: https://aclanthology.org/N13-1090.

Millière, Raphaël and Cameron Buckner (2024). **A Philosophical Introduction to Language Models – Part I: Continuity With Classic Debates**. arXiv: 2401.03910 [cs.CL]. URL: https://arxiv.org/abs/2401.03910.

Olsson, Catherine et al. (2022). “In-context Learning and Induction Heads”. In: **Transformer Circuits Thread**. https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Ouyang, Long et al. (2022). **Training language models to follow instructions with human feedback**. arXiv: 2203.02155 [cs.CL]. URL: https://arxiv.org/abs/2203.02155.

Park, Kiho et al. (2024). **The Geometry of Categorical and Hierarchical Concepts in Large Language Models**. arXiv: 2406.01506 [cs.CL]. URL: https://arxiv.org/abs/2406.01506.

Partee, Barbara H. (2016). “Formal semantics”. In: **The Cambridge Handbook of Formal Semantics**. Ed. by Maria Aloni and PaulEditors Dekker. Cambridge Handbooks in Language and Linguistics. Cambridge University Press, pp. 3–32.

Putra, I Made Suwija, Daniel Siahaan, and Ahmad Saikhu (2024). “Recognizing textual entailment: A review of resources, approaches, applications, and challenges”. In: **ICT Express** 10.1, pp. 132–155. ISSN: 2405-9595. DOI: https://doi.org/10.1016/j.icte.2023.08.012. URL: https://www.sciencedirect.com/science/article/pii/S2405959523001145.

Rorty, Richard (1979). **Philosophy and the Mirror of Nature**. Princeton University Press.

Sellars, Wilfrid (1953). “Inference and Meaning”. In: **Mind** 62.247, pp. 313–338. ISSN: 00264423, 14602113. URL: http://www.jstor.org/stable/2251271 (visited on 12/11/2024).

Smolensky, Paul (June 1987). “Connectionist AI, symbolic AI, and the brain”. In: **Artif. Intell. Rev.** 1.2, pp. 95–109. ISSN: 1573-7462. DOI: 10.1007/BF00130011.

Stalnaker, Robert (1997). “Reference and Necessity”. In: **A Companion to the Philosophy of Language**. Ed. by Bob Hale, Crispin Wright, and Alexander Miller. Wiley-Blackwell, pp. 902–919.

Tsai, Yao-Hung Hubert et al. (2019). **Transformer Dissection: A Unified Understanding of Transformer’s Attention via the Lens of Kernel**. arXiv: 1908.11775 [cs.LG]. URL: https://arxiv.org/abs/1908.11775.

Vaswani, Ashish et al. (2017). “Attention is All you Need”. In: **Advances in Neural Information Processing Systems**. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Vig, Jesse (July 2019). “A Multiscale Visualization of Attention in the Transformer Model”. In: **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations**. Florence, Italy: Association for Computational Linguistics, pp. 37–42. DOI: 10.18653/v1/P19-3007. URL: https://www.aclweb.org/anthology/P19-3007.

Wang, Changhan, Kyunghyun Cho, and Jiatao Gu (Apr. 2020). “Neural Machine Translation with Byte-Level Subwords”. In: **Proceedings of the AAAI Conference on Artificial Intelligence** 34.05, pp. 9154–9160. DOI: 10.1609/aaai.v34i05.6451. URL: https://ojs.aaai.org/index.php/AAAI/article/view/6451.

Weiss, Bernhard and Jeremy Wanderer, eds. (2010). **Reading Brandom: On Making It Explicit**. New York: Routledge.

Yenduri, Gokul et al. (2023). **Generative Pre-trained Transformer: A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions**. arXiv: 2305.10435 [cs.CL]. URL: https://arxiv.org/abs/2305.10435.

Zhao, Wayne Xin et al. (2023). **A Survey of Large Language Models**. arXiv: 2303.18223 [cs.CL]. URL: https://arxiv.org/abs/2303.18223.

白川, 晋太郎 (May 25, 2021). **ブランダム 推論主義の哲学. プラグマティズムの新展開**. 青土社. ISBN: 978-4-7917-7379-4.