２Ｇ０６

生成 AI 領域の形成過程（1）：先導的研究体制の構築の国際比較

○三浦崇寛（文部科学省），林隆之（政策研究大学院大学）
takahiro-miura@mext.go.jp

1. はじめに

科学技術・イノベーション政策において、データに基づいて研究領域の現状を把握することが重要であることは論を俟たない。内閣府エビデンスシステム（e-CSTI）やサイエンスマップ[1]に代表される大規模な学術情報分析は、各領域の現状を客観的に捉え、それに応じた適切な施策を打つために広く貢献している。その中でも、特に生成 AI に代表される新興領域に対する社会的関心は高く、どのように新興領域の研究体制が構築されてきたかを把握し、そのさらなる発展のために効果的・効率的な政策を立案することが求められている。一方で、こうした新興領域に対する分析は、データが蓄積されるまでの時間が不足していることや、出版プロセスの違いなど、既に確立された分野に対する分析方法をそのまま適用することが適切でないこともあり、発展プロセスの把握が難しい。本研究は、我が国の生成 AI 領域の形成過程を書誌学的観点から分析する手法を開発するとともに、日本の先導的研究体制の構築が他国と比較してどのような特徴を持っていたかを明らかにすることを目的とする。

2. 手法

2-1. データ

本研究では、OurResearch 社が無料公開している書誌データベースである OpenAlex を用いる。書誌分析で広く用いられる Scopus や Web of science と比較して、OpenAlex には生成 AI 領域の分析に 2 つの理由で適している。1 つ目に、OpenAlex は 1 ヶ月に一度データの更新が行われており、直近数年で出版された論文のカバー率が高いため[2]、短時間で爆発的に論文数の増えた新興領域の分析に適している。2 つ目に、生成 AI を含む情報領域は、arXiv に代表されるプレプリントサーバや各学会のプロシーディングスの動きを捉えることが肝要であるが[3]、Scopus や Web of science ではそのカバー率が高くない。本研究では、2023 年 10 月に取得した OpenAlex のスナップショット約 2.4 億レコードの中から、生成 AI 領域に関連する論文群を抽出した。

2-2. 生成 AI 論文の取得

書誌データから特定の領域を抽出する方法には、メタデータに含まれるタグを用いる手法や分野を代表するジャーナルを選択する方法があげられるが、生成 AI 領域は AlphaFold に代表されるように情報領域以外にも広く影響を与えており、まとまった分野として明示的に取得することが難しい。そこで本研究では、生成 AI の発展に多大な影響を与えた論文である”Attention is All You Need” (Vaswani et al. 2017.06.12 公開)を中心論文（p）とし、pから後方引用で連なる論文群を生成AI領域とした。pを引用する論文群を Core set[1]（$\mathcal{P}_c$）、$\mathcal{P}_c$のうち 1 本でも引用する論文群を Marginal set（$\mathcal{P}_m \notin \mathcal{P}_c$）、$\mathcal{P}_m$のうち 1 本でも引用する論文群を Frontier set（$\mathcal{P}_f \notin \mathcal{P}_c, \mathcal{P}_m$）とする。$\mathcal{P}_c, \mathcal{P}_m, \mathcal{P}_f$ はそれぞれ 18,746 件、66,802 件、126,120 件であり、これらを合わせて 211,668 本の生成 AI 論文群（$\mathcal{P}$）を取得した。$\mathcal{P}$は BERT や Vision Transformer, GPT-3 といった主要な生成 AI に関連する論文群を含む。

図 1 は、生成 AI 論文群に十分なテーマが含まれているかを確認するため、各論文のタイトルを入力として BERTopic[4]で分類を行ったものを示している。Transformer がはじめに活用された自然言語処理を中心に、画像・動画・音声といった他ドメインへの適用や、強化学習・マルチタスク・最適化といったモデルの発展など、生成 AI 領域が捉える幅広い研究領域を取得できていると考えられる。

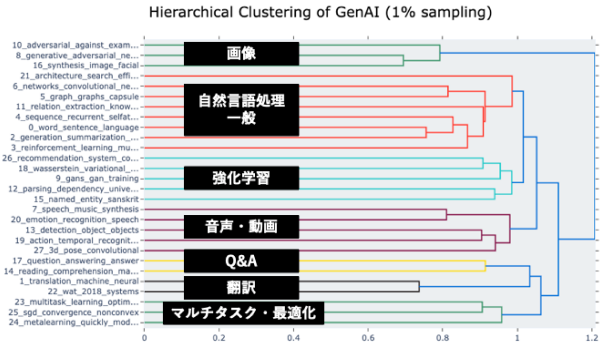

図 1. BERTopic による対象データの分類

[1] 自然言語処理の生成 AI 領域への影響が大きかった PaLM, LLaMa, GPT-4 の代表論文について、$\mathcal{P}_c$に含まれなかったため、分析の網羅性のために Core set として追加している。

2-3. 日本の生成 AI 領域の発展の国際比較

新興領域に対する研究体制を考える上で、Transformer のような革新的技術が外部で生まれたときに、その研究を素早く取り込んで次の研究に活かすことは極めて重要である。本研究では、そうした能力を”研究吸収力(research absorptive capacity)”と名付ける。企業経営においても、内部の研究開発活動を通じて形成される外部技術の吸収力(absorptive capacity[5])は重要とされており、素早く変化する外部環境にいち早く順応し、自社技術と組み合わせることで競争力を高めるために必要な能力である。これと同様に、研究吸収力が発揮される研究環境であれば、革新的な研究が生み出されたときに素早くその知見を研究者自身の知見と組み合わせ、研究の独自性や競争力を高めることに寄与すると考えられる。

本研究では、生成 AI 領域が現れたときに、日本がどの程度素早く、インパクトのある研究を生み出すことができていたかを国際比較するために、$\mathcal{P}$に含まれる各論文著者の所属機関とその国籍を取得し、論文出版日ごとに整数カウントで論文数を積み上げることで、いち早く生成 AI 領域に参画した国・組織を明らかにする。

3. 結果

3-1. 生成 AI 領域における研究吸収力の比較

図 2、図 3 は国別の生成 AI 論文数、高被引用論文数（被引用数$c > 50$）の日次推移を示したものである。論文数ではアメリカと中国の二ヵ国が抜けて多く、2023 年には中国がアメリカを抜いて最も多くの論文を出しているが、高被引用論文数ではアメリカが常にトップとなっている。イギリス・ドイツも 2020 年から他国より多く論文を出しており、2021 年以降インドがそれに続く形で論文数を伸ばしている。

2023 年 10 月時点の日本の累積論文数は 4,220 本で、中国(46,954)、アメリカ(36,580)、イギリス(11,101)、ドイツ(8,643)、インド(8,024)、カナダ(5,442)、オーストラリア(5,335)、韓国(4,869)、イタリア(4,265)に続いて 10 番目となっている。高被引用論文数では、日本は 62 本で、アメリカ(1,118)、中国(685)、イギリス(292)、ドイツ(187)、カナダ(145)、イスラエル(131)、シンガポール(117)・スイス(94)・韓国(82)・イタリア(80)・フランス(80)・スペイン(63)に続いて 13 番目である。

一方で、研究吸収力の素早さを見るために、$\mathcal{P}$の投稿後 1 年以内（2017.6.12~2018.6.11）までの論文数で比較すると、日本はアメリカ(445)・中国(160)・イギリス(86)・ドイツ(57)・カナダ(49)・イスラエル(41)に次ぐ 40 本で 7 番目、高被引用論文数ではアメリカ(94)・中国(20)・イギリス(13)・イスラエル(11)・カナダ(7)・ドイツ(6)と並んで 6 本で 6 番目となっていた。このことから、日本は初期に相対的に研究吸収力があったものの、2018 年以降の各国の成長に比べて相対的に伸びていないことが分かった。

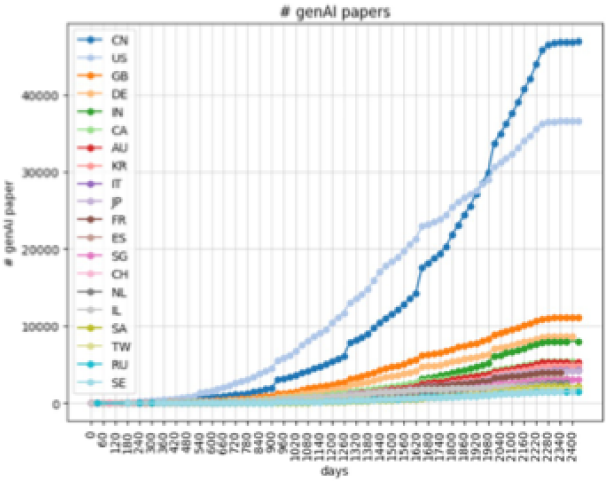

図 2.国別生成 AI 論文数推移

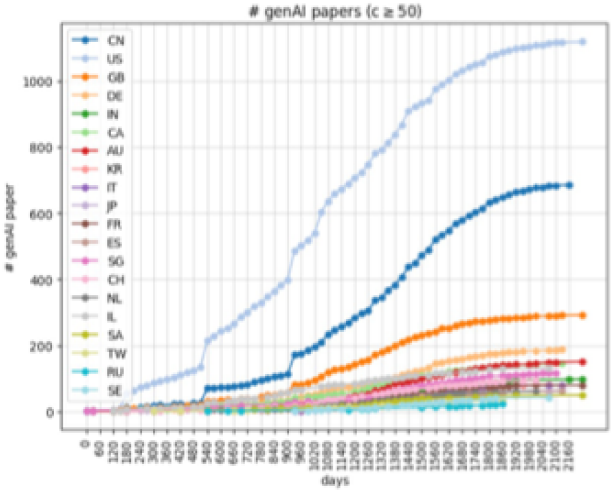

図 3. 国別生成 AI 論文数推移（$c > 50$）

同様の分析を組織別に行ったものが図 4,図 5 である。$\mathcal{P}$の投稿直後 1 年の論文数においては、Google(86), Carnegie Mellon University(56), Microsoft(30), Facebook(28), Stanford University(28), Washington University(27), UC Berkeley(24)をはじめとするアメリカの大学・企業群が大きく牽引し、高被引用論文でも同様の傾向が見られた。一方で、2022 年以降中国の組織が大幅に論文数を伸ばしており、中国科学院・清華大学・浙江大学・上海交通大学・北京大学などで多くの生成 AI 論文が出ている。しかし、高被引用論文数においては中国組織の論文数はあまり多くなかった。

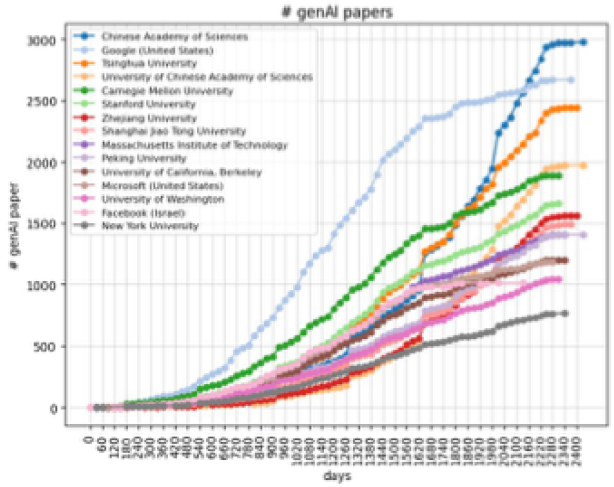

図 4. 組織別生成 AI 論文数推移

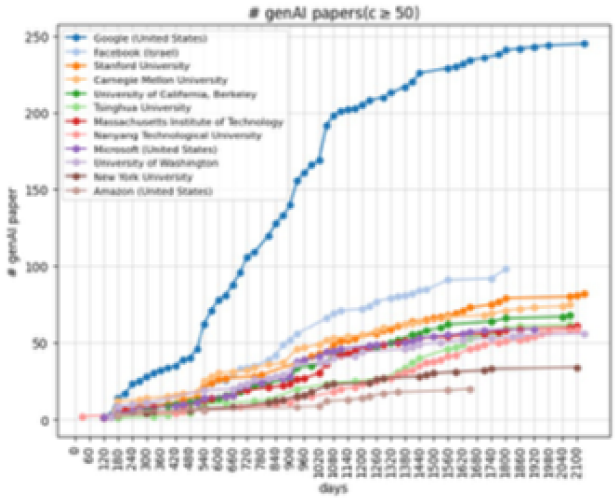

図 5. 組織別生成 AI 論文数推移（$c > 50$）

図 6. 日本の組織別生成 AI 論文数推移

図 6 は、より日本の研究組織に焦点を絞るために、日本の組織別生成 AI 論文数、高被引用論文数の多い組織における論文数の推移を示したものである。現時点の論文数では、東京大学(600)、京都大学(239)、東京工業大学(233)、大阪大学(215)、早稲田大学(213)、国立情報学研究所(185)と続くが、pの投稿後 1 年以内に絞ると、最も多いのは情報通信研究機構(11)で、次いで京都大学(7)、東京大学(6)、大阪大学(4)、奈良先端科学技術大学院大学(4)、東京都立大学(3)、Preferred Networks(3)であり、現時点で論文数が多い機関とは異なっていた。このことから、素早く生成 AI 領域にキャッチアップした研究機関と、その後論文数を伸ばしている機関は異なっていたといえる。

この要因をさらに深ぼるために、研究者レベルでも論文数を算出し、現時点の論文数 Top10 の研究者と、pの投稿後 1 年以内の論文数 Top10 の研究者について推移を示したものが図 7 である。論文数の多い研究者の所属を見ると、現時点で組織として多くの生成 AI 論文を出す東京大学、京都大学、東京工業大学に所属する研究者は多くなく、国立情報学研究所の山岸順一教授、東京都立大学の小町守特任教授、東北大学の乾健太郎教授などの名前が挙がっている。このことから、日本の生成 AI 研究においては、各大学に散らばる個別の研究者の取組によってキャッチアップされていたものが、数年経過した後に規模の大きい大学によってフォローアップされる形で進んできたことが分かる。

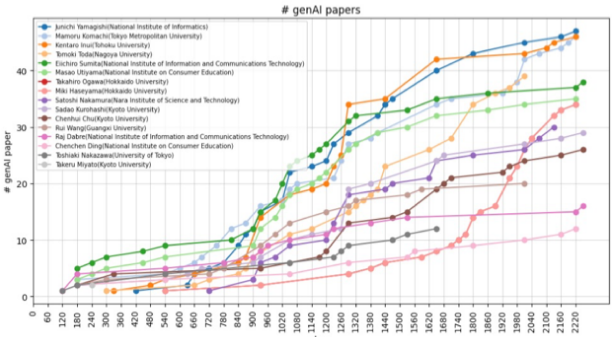

図 7. 日本の研究者別生成 AI 論文数推移
（所属は最後に出版した論文より算出）

3-2. 生成 AI 領域における研究吸収力の比較

3-1 の結果より、日本の生成 AI 研究は初期に各地の研究者がキャッチアップしていたものの、そのリソースを活用しきれずに資源が分散してしまっていたために、発展速度が相対的に低下してしまったのではないかと考えられる。そこで、各国の中で生成 AI 領域に取り組む組織の分散度を比較し、日本の生成 AI 領域における研究体制がどの程度の組織に分散していたかを分析した。分散度の指標として、データ全体の分散を考慮するジニ係数（$Gini = 1 - 2\int P(x)dx$, $P(x)$は組織別累

積論文割合分布）と、特に一部の研究機関に論文が集まっているかを示すハーフィンダール係数（$HF = \sum p^2$, pは各組織の論文数割合）を用いた。これにより、各国の組織別論文集中度を示したものが図 8 である。ジニ係数が高く、ハーフィンダール係数も高い国は、ごく少数の組織が国内のほとんどの生成 AI 論文を占めている構造を示しておりシンガポールがあてはまる。アメリカ、イギリス、オーストラリアは、ジニ係数が高くハーフィンダール係数が低いため、論文数の多い大学が一定数存在しており、少数の機関に極端に集まっていないことを示す。その中で日本は、ジニ係数が 0.845、ハーフィンダール係数が 0.026 で全体の中でも中程度であり、広く様々な大学から生成 AI 論文が生み出されていることが分かった。

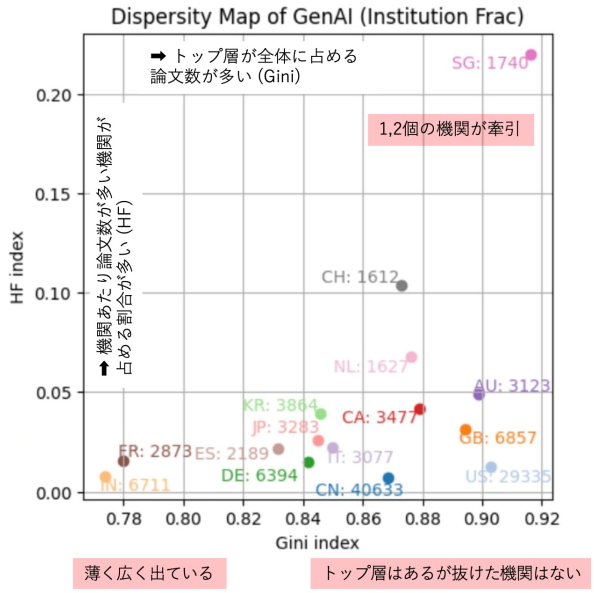

図 8. 各国の組織別論文集中度
（ラベル内の数字は論文数（分数カウント））

4. 考察

本研究では、OpenAlex を用いることで生成 AI 領域の分析するための手法を開発した。その結果、日本の生成 AI 研究は、初期の頃各大学に点在する個別の研究者の取組に依るところが多く、国際的に見ても 6 番目程度のインパクトを発揮していたが、その後規模の大きい大学が参入するにつれて、相対的にその影響力を落としていることが示された。これは、日本の生成 AI 研究において、研究者の所属と研究機関のリソースにミスマッチが起きていた可能性があることを示唆している。これまでの情報領域の研究では実験室レベルの CPU、GPU で行う研究が一般的であったが、生成 AI 研究では大量の計算資源を必要とすることが多いため、いち早く領域にキャッチアップしていた研究者に適切なリソース配分をすることがより一層重要であった可能性がある。逆にイギリス・アメリカ・シンガポールなどでは、一部の研究機関が生成 AI 論文を多く出す構造となっているため、投資の集中効果だけでなく、組織の中の情報交換などを通じて、研究吸収力を高めていた可能性も考えられる。

これらの結果から、日本の先導的研究体制構築において研究吸収力を高めるためには、いち早く当該領域に着目している研究者を特定し、所属機関から十分な支援を受けられない場合には積極的に支援をしていくことが肝要である。本研究で開発した OpenAlex を活用した新興領域の分析手法は、データセットの更新頻度が十分に高いため、今後他の新興領域に対しても応用可能な手法であると考えられる[2]。こうした書誌学的分析を、過去事例の整理のために用いるだけでなく、生まれつつある新興領域に対して適用し、支援を必要とする研究者に対して必要な支援を行なっていくことは、国全体としての研究吸収力を高めていくためにも重要であると考えられる。

謝辞

本研究は、内閣府委託事業「エビデンスに基づく重要科学技術領域の調査分析」からの支援を受けたものである。なお、本研究は第一著者が東京大学博士課程在学中に、政策研究大学院大学のリサーチアシスタントとして実施したものである。

参考文献

[1] 文部科学省 科学技術・学術政策研究所, サイエンスマップ 2020, NISTEP REPORT No.196, 2023 年 3 月

[2] Culbert, J., Hobert, A., Jahn, N., Haupka, N., Schmidt, M., Donner, P., & Mayr, P. (2024). Reference coverage analysis of openalex compared to web of science and Scopus. arXiv preprint arXiv:2401.16359.

[3] 林 和弘, 小柴 等 「arXiv に着目したプレプリントの分析」, NISTEP DISCUSSION PAPER, No.187, 文部科学省科学技術・学術政策研究所. DOI: https://doi.org/10.15108/dp187.

[4] Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794.

[5] Cohen, W. M., & Levinthal, D. A. (1990). Absorptive capacity: A new perspective on learning and innovation. Administrative science quarterly, 35(1), 128-152.

[2] 著者名・組織名の名寄せ精度等において、必ずしも OpenAlex が他のデータベースよりも優れているとは限らないため留意が必要である。