言語処理学会を外部から調査するための 共著者ネットワークを用いた発表予稿の自動地図作成の試み

中原龍一[1] 竹内孔一[2] 片岡裕雄[3] 品川政太朗[3] 高橋康[4]
笠井聡[5] 長谷井嬢[6] 二川摩周[7] 大戸彰三[8] 加藤直人[9] 鎌倉英嗣[9]
[1]岡山大学大学院医歯薬学総合研究科 運動器地域健康推進講座 [2]岡山大学大学院環境生命自然科学研究科
[3]SB Intuitions 株式会社 [4]日本電気株式会社 [5]新潟医療福祉大学医療技術学部診療放射線学科
[6]岡山大学学術研究院医歯薬学域 医療情報化診療支援技術開発講座,[7]岡山大学医学部医学科 臨床遺伝子遺伝学
[9]株式会社両備 システムズヘルスケアソリューションカンパニー メディカルAI推進室 [9]旭化成ファーマ株式会社 メディカル・アフェアーズ部
[1]r-nakahara@okayama-u.ac.jp, [2]takeuc-k@okayama-u.ac.jp [3]{hirokatsu.kataoka, seitaro.shinagawa}@sbintuitions.co.jp
[4]yas.takahashi@nec.com [5]satoshi-kasai@nuhw.ac.jp [6]py3g9rcw@s.okayama-u.ac.jp, [7]mfutagawa@okayama-u.ac.jp
[8]ooto@ryobi.co.jp [9]kato.nv@om.asahi-kasei.co.jp kamakura.ec@om.asahi-kasei.co.jp

概要

大規模言語モデルや基盤モデルの出現に伴い、医療 AI 領域においても言語処理技術の重要性がますます高まっている。その結果、外部の研究者や企業関係者による言語処理学会に対する網羅的調査のニーズという新しい課題が生まれている。

我々はこの問題を**「知識が少ないドメイン外研究者による、学会の網羅的調査」**という新しいタスクとして提案する。また、その解決手法として全論文の共著関係を利用した共著ネットワーク地図の有用性を検討したので報告する。

1 はじめに

深層学習の出現により様々な医療 AI アプリケーションの研究開発が可能となった。初期の医療 AI アプリケーションは、画像 AI 技術が中心であったため、CVPR（Conference on Computer Vision and Pattern Recognition）などの画像 AI 学会の論文を網羅的に学習し、知識を共有してきた。しかし ChatGPT に代表される大規模言語モデルや基盤モデルの出現に伴い画像 AI 技術に閉じた学習では足りなくなり、言語処理学会などで扱われている言語 AI の知識が必要となった。

我々は言語処理学会に入会し、学会に現地参加し、公開された予稿を元に勉強会を開催した。画像 AI 研究と同様の手法で調査が可能かと思われたが、言語処理知識の不足に起因する様々な問題に遭遇した。

本質的には我々の知識不足に起因する問題であるが、**「知識が少ないドメイン外研究者による、学会の網羅的調査」**という新しいタスクの定式化研究とみなせるのではないかと考え、自らの知識不足を前向きにとらえることにした。

ブレイクスルーとなったのは、発表論文すべての共著関係のネットワーク地図を創ったことである。これによって以下の 3 種類の問題がある程度解決された（図 1）。

(A) 直接共著関係を利用した研究グループ検出
(B) 共著距離の近さを利用した研究グループ検出
(C) 研究グループ周囲の検索漏れ論文検出

まだまだ課題が山積しているが、本報告では共著ネットワークの作成過程や有用性などの研究結果を報告する。

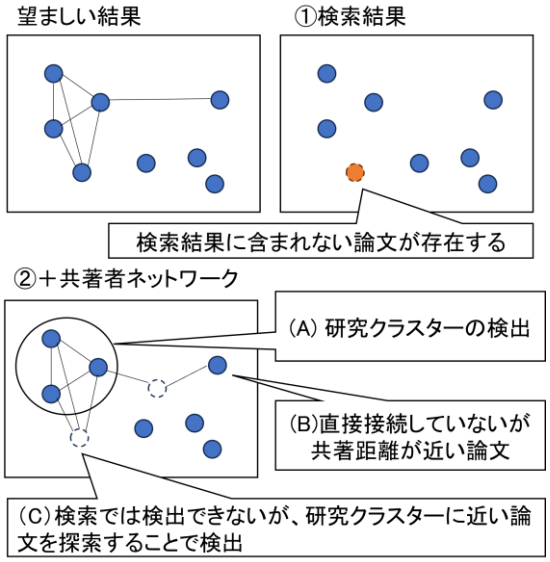

図 1 共著ネットワークを利用した学会調査

2 Related-Work

2.1 AI 関連学会の網羅的紹介プロジェクト

国内において網羅的紹介プロジェクトが多く実施されている国際学会として、CVPR が挙げられる。CVPR はコンピュータビジョンとパターン認識分野における世界最大級の国際学会であり、毎年重要な研究が発表されており、学術面だけでなくビジネス面からの注目度が高いことから、様々な網羅的紹介プロジェクトが存在する。

2.2 cvpaper.challenge

本プロジェクトは片岡裕雄らによって開始された[1]。千人以上のボランティアによって運用されており、メンバーで分業し全論文の要約を作成し、メンバーで話し合ってまとめを作成する点が特徴である。発表媒体は PDF 形式である。近年、CVPR 論文数の増加に伴い、初期のように全論文を紹介することは難しくなっているが、ボランティアベースで運営されている点が他のプロジェクトと異なる特徴である。

2.3 Sony

Sony の nnabla チームは、毎年 CVPR の網羅的紹介を YouTube で公開している[2]。発表媒体が動画形式であるため、視聴しながら学習しやすい点が特徴である。冒頭で論文タイトルを言語解析し、研究トレンドの変化を分析した上で注目分野をピックアップし、その後、分野ごとに注目論文を紹介している点が特徴である。

2.4 日本ディープラーニング協会

日本ディープラーニング協会が主催する CVPR 技術報告会[3]。2020 年から毎年開催されており、Sony と同様に YouTube で動画として報告されている。冒頭で論文タイトルを言語解析し、研究トレンドを分析した後、分野ごとに論文を紹介している。。

2.5 学会の網羅的調査の傾向

学会の網羅的調査に対するニーズは年々高まっており、学会調査を専業とする企業も出現している。その代表例として、ResearchPort [4]が挙げられる。CVPR だけでなく、他の主要な国際 AI 学会のまとめも行っている。論文タイトルの言語解析に加えて、引用数ランキングなどのさまざまな統計量を示している点が特徴である。発表媒体は Web ページである。

このように、さまざまな国際 AI 学会の網羅的紹介プロジェクトが存在するが、解析手法は論文タイトルの言語解析や引用数解析などの従来手法が中心であり、共著者ネットワークのような論文間のネットワーク構造に注目した調査は少ない。

3 方法

言語処理学会のホームページから人力でテキストを取得し、自作した Matlab のプログラムを用いて Execel ファイルに変換を行った。タイトルの検索を用いて注目するキーワードを抽出し、Cytoscape[5]（Version: 3.10.2）を用いてネットワークの可視化を行った（図 2）。

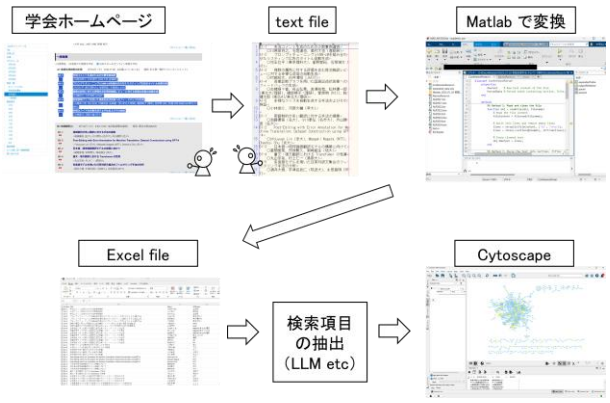

図２　ネットワークの可視化手順

4 結果

4.1 共著者ネットワーク

2021 年から 2024 年にかけての共著者ネットワークを示す（図３）。タイトルと著者をノードとし、エッジで接続している。異なるタイトルが同じ著者を持つ場合、共著者ネットワークを形成する。どの年においても大きなネットワーク構造と、孤立した小さなネットワークに分離されていることがわかる。

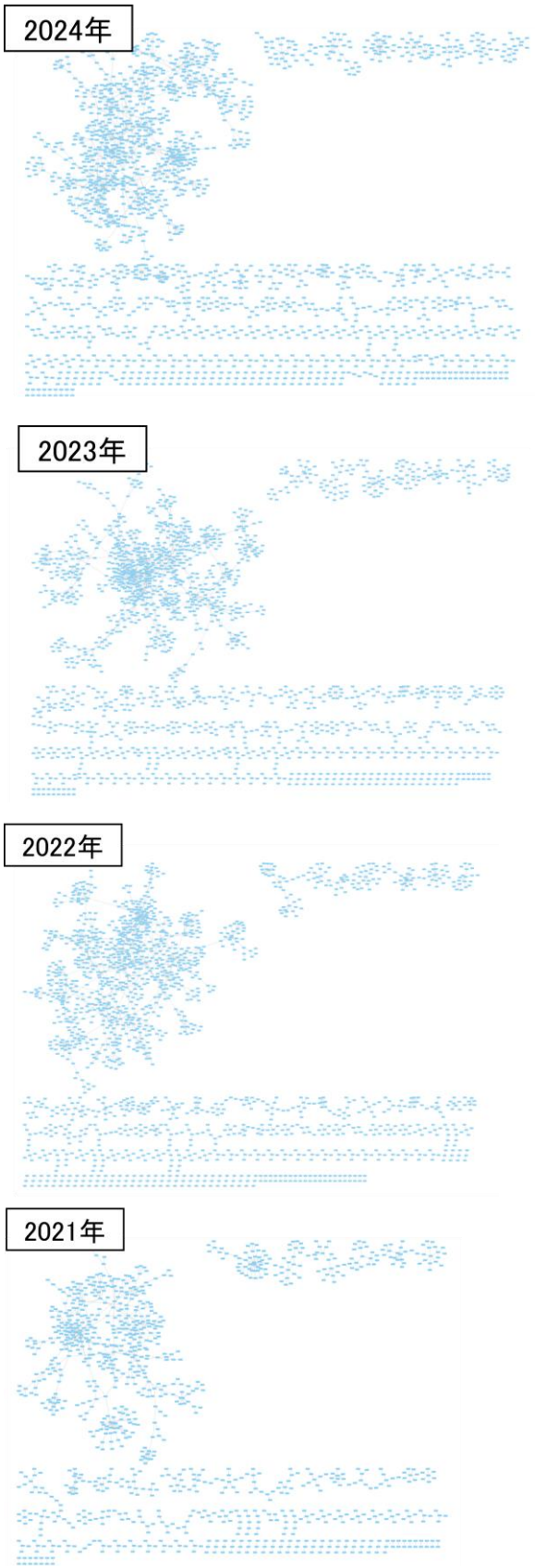

図３　共著者ネットワークの推移

共著者ネットワークの年度ごとの特徴を表１に示す。意外なことに大域構造に大きな変化はない。そのため、年度ごとの変化を捉えるためには、論文タイトルの検索を組み合わせる必要がある。

Year	2024	2023	2022	2021
Number of nodes	1893	1772	1765	1106
Number of edges	2151	2036	2036	1211
Avg. number of neighbors	2.721	2.763	2.685	2.623
Network diameter	24	28	26	24
Network radius	12	15	14	13
Characteristic path length	9.609	10.78	10.695	10.53
Network density	0.003	0.003	0.003	0.005
Network heterogeneity	0.831	0.864	0.823	0.813
Network centralization	0.022	0.031	0.022	0.045
Connected components	178	145	145	99

表１　共著者ネットワークの特徴

4.2 検索＋共著者ネットワーク

図４に共著者ネットワーク上で、大規模言語モデル関係タイトル（「大規模言語モデル」あるいは「LLM」を含む）のハイライトを示す。大規模言語モデル関連の論文が出現したのは 2023 年以降であったため、2023 年と 2024 年のみ示している。

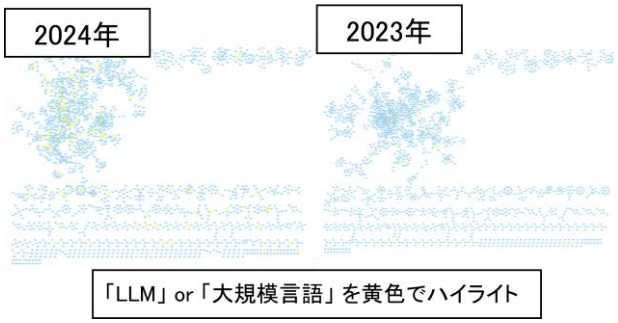

図４　大規模言語モデル関係のハイライト

ネットワーク図に検索された論文をハイライトすることで、（A）直接共著関係による研究ネットワークを検出し、（B）共著距離の近さを利用して研究グループメンバーを検出し、（C）検出された研究グループの周囲を探索することで、検索漏れの論文を低コストに検索することができる（図５）。

またこのような研究ネットワーク検索を勉強会メンバーが個別に行い、その結果を共有することで、学会全体を構造的に把握できることが判明した。

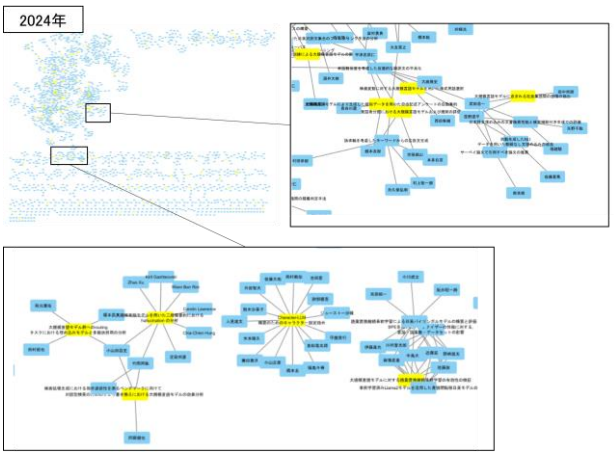

図５　共著者ネットワーク＋検索ハイライト

5 考察

5.1 外部における価値を高めるために

CVPR の h5-index が科学誌『Science』を超えたことが話題になっている。NeurIPS や ICCV などの国際 AI 学会のランキングも上昇している。これらの学会の地位向上は、もちろん良質な論文によるものであるが、論文が無料公開されている点も影響していると考えられる。

国際 AI 学会では、学会論文の無料公開が主流となっている一方、国内で論文（予稿）をすべて公開している学会は少ない。言語処理学会は、その数少ない論文を無料公開している学会の一つであり、我々もたびたび活用してきた。言語処理学会参加者にとっては当たり前の技術や知識であっても、我々のようなドメイン外の研究者や企業関係者にとっては重要な情報となる。言語処理学会内部の人にとっては意外に思える論文が、我々にとって大きな価値をもたらすことが多いため、我々にとっての有用性を示すことには価値があると思われる。

言語処理学会の開催期間中には、学会 Slack が常設され、学会発表内容に対する活発な議論や、発表者による最新スライドの共有などが行われている。これらの非同期コミュニケーション技術が言語処理学会の隆盛を支えているように見える。

共著ネットワークによる可視化は非同期コミュニケーションの一種ととらえることができる。これらの研究を通じて、外部における言語処理学会の価値を高めていくことが我々の目的である。言語処理学会の論文群を外部から見た価値は、外部の人間にしか分からないため、積極的に貢献していきたいと考えている。そのためにも、**「知識が少ないドメイン外研究者による学会の網羅的調査」**というタスクの定式化を基盤とし、研究を進めていく予定である。

6 本手法のリミテーション

本手法には多くのリミテーションがあると言わざるを得ない。

本手法は共著者関係のみを利用したネットワーク地図であるため、論文の重要度や概念的価値を判定することはできない。本当に知りたいのは共著関係ではなく、技術や概念の関連性である。そのため、タイトルや論文内容に対する言語処理を通じた拡張が次の段階として必要となる。また、本手法に対する数値的評価も行われていない。今後は、どの程度人的コストが削減されるかを計測し、可視化を行う予定である。

謝辞

本研究は、言語処理学会の網羅的調査に向けて多くの先生方とともに勉強を重ねる中で得られた貴重なインサイトに基づいています。この場をお借りして、貴重なご助言とご支援を賜りました諸先生方に深く感謝申し上げます。

また、予稿を公開してくださった言語処理学会の皆様にも厚く御礼申し上げます。

参考文献

1. cvpaper.challenge https://xpaperchallenge.org/cv/
2. nnabla ディープラーニングチャンネル https://www.youtube.com/c/nnabla
3. 日本ディープラーニング協会 https://www.youtube.com/@JDLA2017
4. ResearchPort https://research-p.com/
5. ***Cytoscape: a software environment for integrated models of biomolecular interaction networks.*** **Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T., Genome Res, 13:11 (2498-504). 2003 Nov. PubMed ID: 14597658.**