

令和 5 年 5 月 24 日現在

機関番号：10101

研究種目：挑戦的研究（萌芽）

研究期間：2021～2022

課題番号：21K19814

研究課題名（和文）論文内の記述と各種科学技術DBを連携させる特定研究グループ向け論文DBの研究

研究課題名（英文）Research on paper database systems for a specific research group that links description in the paper with various scientific database

研究代表者

吉岡 真治（Yoshioka, Masaharu）

北海道大学・情報科学研究院・教授

研究者番号：40290879

交付決定額（研究期間全体）：（直接経費） 4,900,000円

研究成果の概要（和文）：本研究は、特定の分野に興味を持つ研究者が収集する関係分野の論文を対象に、専門用語抽出を行うことにより、用語間の共起関係に基づく分析や、時系列を考慮した研究動向分析を行うデータベースシステムの研究を基礎とし、論文内の記述と科学技術データベースを連携させる手法を提案することで、分野の研究者の研究活動を支援するシステムの構築を目指した。本枠組の有用性を有機化学反応の分野を対象に機械学習による情報抽出と化学物質データベースであるPubChemを活用して実験的に構築し、このようなデータベースが化学物質名の記載などにおける表記の揺れの解消などに貢献できることが確認された。

研究成果の学術的意義や社会的意義

本研究は、多くの論文データベースが幅広く集めた論文を対象に分析を行うのとは異なり、特定の研究グループが興味を持つ論文群に限定してデータベースを構築することで、分野の研究者の研究動向の変化といった解析可能な論文データベースの構築を目指している。また、そこで作成したデータベースを他の科学技術データベースと連携することは、用語のエントリなどをうまく活用することで、分野の研究者のデータベース整備に関わる作業を効率化できるだけでなく、より質の高いデータベースの構築に結びつく。また、論文を読む際に、関連する外部データベースを容易に参照する枠組みを提供することができ、研究者の支援にもつながると考えている。

研究成果の概要（英文）：We have worked on the project to make a paper database system for a specific research group that can analyze the technical terms from the point of view of co-occurrence analysis and trend analysis. In this research, we proposed a system that links the description in the paper with various scientific databases. To discuss the framework of the proposed system, we use research papers that introduce new organic synthesis process by using machine learning based information extraction system and PubChem, which is an open database for the chemical substances. We confirmed that it is useful to use such scientific database to normalize the description used by different authors for co-occurrence analysis and trend analysis.

研究分野：知識工学

キーワード：論文データベース テキストマイニング 科学技術データベース

## 様式 C - 19、F - 19 - 1、Z - 19 (共通)

### 1. 研究開始当初の背景

我々の研究グループでは、特定の専門領域に属する研究グループが収集した論文群が研究グループの興味を表すと考え、その論文群における用語の使われ方を分析することで、論文中からの重要語抽出、研究動向の分析などを行うことができるデータベースの構築を行っていた。この研究では、専門用語の特性を考慮した用語辞書の構築支援などを行なっていく枠組みを提案していたが、その対象が集めた論文に限定されるという問題があった。このような問題に対し、近年のオープンサイエンスの流れに伴って、研究活動にも活用されるようになってきている科学技術データベースと連携させることで、より有用な研究支援が行えるのではないかと考えた。

### 2. 研究の目的

本研究では、これまでに作成してきた特定の研究グループが属する専門領域に関する情報が、そのグループが興味を持つ論文群によって特徴づけられると考えて、論文の収集および分析を行う論文データベースを基礎として、分野の研究者が必要とする情報を抽出する方法について検討するとともに、その情報を科学技術データベースとの連携を行うことでより有用な分析を実現する論文データベースを構築することをその目的とする。

### 3. 研究の方法

本研究では、化学反応データベースである Reaxys や様々な化学物質に関するデータを整理している PubChem などの様々な科学技術データベースが活用できる有機化学反応の解析に関する分野を対象として研究を進めることとした。具体的には、論文中の記述から Reaxys の詳細度のレベルで化学反応情報を抽出するとともに、そこに現れる化学物質などのデータを PubChem などと関連づけて解析する枠組を構築し、その結果に基づいて、これまで構築してきた論文データベースに組み込むことで、その有用性について検討する。

### 4. 研究成果

本研究では、まず、研究分担者である長田の研究背景を考慮し、有機化学反応について、編集部により慎重に実験の再現性が検証された化学反応を収録している学術雑誌である Organic Syntheses (<http://www.orgsyn.org/>)を対象に、化学反応の情報抽出を行うこととした。Organic Syntheses には、多くの新しい有機化合物を作るための方法を提案した論文が採録されているだけでなく、その論文が、人手により作成された html ページとして、関連する化合物に関するメタデータの付加が行われている。しかし、事前に、論文と、その論文に関する化学反応データベース Reaxys におけるデータを照合したところ、必ずしも、十分な情報が付与されているわけではないということが確認された。そのため、機械学習などによる情報抽出をおこない、その情報を付与することが有用であると考えた。このような情報の抽出を目指すために、特許文書を対象として、Reaxys のレベルの化学反応情報を抽出することを試みるワークショップ型のプロジェクトである ChEMU (Cheminformatics Elsevier Melbourne Universities lab <https://chemu2022.eng.unimelb.edu.au/>) というタスクに参加した。本プロジェクトでは、化合物や化学反応に関する手順に関する情報を特許文書から抽出することを目的としており、本研究の目的とも合致している。我々はこのタスクに大規模言語モデルに基づいて情報抽出を行うシステムを構築した。本システムは、他の参加システムの抽出性能を上回る性能を示すことはできなかったが、ほぼ同等の性能を有し、Reaxys に記載しているレベルの情報が精度や再現率も高く抽出できることが確認された。また、本システムで作成した情報抽出システムが、Organic Syntheses においても情報抽出に有用であることを確認した。

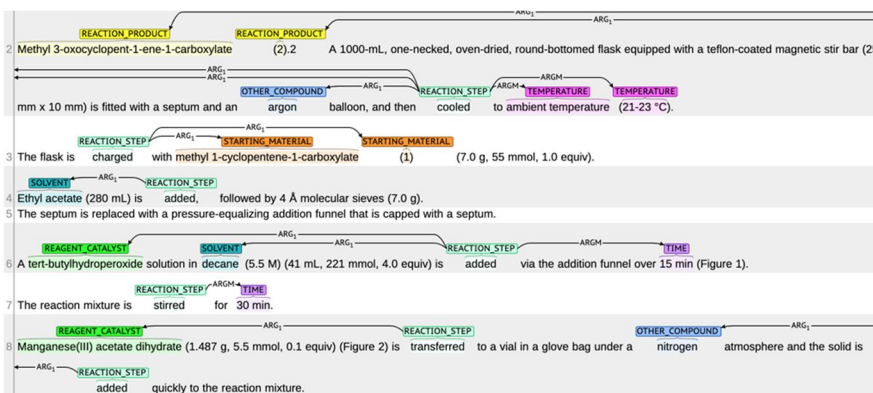


図 1 : Organic Syntheses の論文からの化学反応情報の抽出  
ChEMU のタスクでは、REACTION\_PRODUCT (化学反応に関係する化学物質である最終生成物)

STARTING\_MATERIAL (初期の材料) REAGENT\_CATALYST (触媒) SOLVENT (溶媒) OTHER\_COMPOUND (その他の化合物)に加え、REACTION\_STEP (反応の手順) や WORKUP (反応の結果得られた溶液や粉末などから、目的の化合物を取り出す手順) TEMPERATURE (温度) や TIME (時間) などのパラメータの情報が抽出している。我々が提案してきた論文データベースにおいて、この用語抽出の結果に基づいたメタデータを化学反応のステップに対応する形で付与することで、材料・触媒・溶媒などの共起関係を分析することが可能となる。しかし、TEMPERATURE や TIME などのパラメータについては、定量的な値をうまく扱えないため、データベースの収録の対象外とした。このように化学物質に関する情報抽出を行なった化学反応記述に関する論文データベースシステムの有用性を検討するために、Organic Syntheses の論文のうち、Vol.1(1921)から Vol.97(2020)の 2,988 件の論文を取得し、各々の論文に複数の反応手順が含まれる場合には、それらを複数に分け、4,627 件の反応手順を抽出し、データベースを構築した。

ここで、化学物質を表す REACTION\_PRODUCT, STARTING\_MATERIAL, REAGENT\_CATALYST, SOLVENT, OTHER\_COMPOUND の 5 種類については、適切に抽出が行われていれば、化学物質データベースに対応するデータが存在することが期待される。本システムでは、PubChem を化学物質データベースとして活用し、対応する化学物質の ID が存在するか否かを調べた。

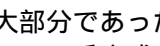
表 1 に観点毎の対応可能な用語の数、対応しない用語の数、また対応する用語の割合をのべと異なりで整理したものを示す。この表から、のべで見ると限りにおいて、REACTION\_PRODUCT を除く多くの用語が PubChem と対応できる用語として抽出されていることが確認された。また、その間違いの多くは、用語の境界の判断を誤っているパターンと、  
、  
などのギリシャ文字や+、-や数字などを用いて構造に関する情報が多様な形で付与されている化合物においてうまく対応ができないパターンが大部分であった。一方、REACTION\_PRODUCT については、その表記の書き方が、として示されている反応式中で示されている A, B といった指示語に対応するものである場合や、product や oil といった照応解析が必要なものが多く存在することが確認された。また、各論文において、生成したい化合物は異なる事が多く、各々の用語の頻度が小さかった。そのため、REACTION\_PRODUCT では、他の分類と異なり、「のべ」の照合結果が「異なり」よりも小さい状況になっている。

表 1 : 論文中の記載と PubChem のデータの照合結果

観点	のべ			異なり		
	対応可能	対応なし	対応可能割合	対応可能	対応なし	対応可能割合
REACTION_PRODUCT	4006	5368	0.43	3569	3755	0.49
STARTING_MATERIAL	6155	2220	0.73	3407	1964	0.63
OTHER_COMPOUND	28251	4845	0.85	2104	1232	0.63
REAGENT_CATALYST	8183	1404	0.85	1687	857	0.66
SOLVENT	5362	429	0.93	205	57	0.78

この分析をしている中で、化学反応データベースにおける同じ物質に対応する複数の表記があることも確認された。具体的には、略称 MTBE と Methyl t-butyl ether、言語表記と化学式 Carbon dioxide と CO<sub>2</sub>、大文字、小文字の違い Carbon dioxide と carbon dioxide などが確認された。上記のような異表記の問題があるため、データの分析を行った際に、本来まとめて扱うことが望ましい同一の化学物質が表記の違いのために、別々の語として扱われるといった問題点があった。この問題を解決するために、同じ化学物質 ID に対応づけられる用語は、同じ化学物質を表す用語として扱うこととし、その代表表記としては、出現頻度が最も多いものを選ぶこととした。このような正規化は、PubChem などの異表記を扱うことができるデータベースが存在することによって可能となるものであり、分野における情報資源に応じた対応が不可欠であることが確認された。

このような問題を解決することで、例えば、初期材料を決めた際にどんな溶媒や触媒が用いられるかといった解析が可能となる。例えば、以下の図は、STARTING\_MATERIAL として、benzaldehyde を選択した場合であり、その時用いられる SOLVENT としては、water が多く、REAGENT\_CATALYST としては、sodium hydroxide や hydrochloric acid が用いられることが多いことが読み取れる。

**Nano Figure**

年/月/日  Caption  検索

REACTION\_PRODUCT STARTING\_MATERIAL REAGENT\_CATALYST SOLVENT OTHER\_COMPOUND REACTION\_STEP WORKUP

benzaldehyde

VolN

REACTION_PRODUCT	STARTING_MATERIAL	REAGENT_CATALYST	SOLVENT	OTHER_COMPOUND	REACTION_STEP	WORKUP	VolN
product (10)	benzaldehyde (65)	sodium hydroxide (8)	water (16)	water (45)	added (55)	washed (53)	9 (3)
oil (2)	sodium cyanide (4)	hydrochloric acid (5)	dichloromethane (10)	ether (19)	stirred (30)	dried (39)	44 (3)
solid (2)	acetic anhydride (3)	butyllithium (3)	ethanol (10)	aqueous (19)	placed (28)	added (25)	86 (3)
title compound (2)	benzyl cyanide (3)	dimethyl sulfate (3)	hexane (5)	hydrochloric acid (18)	charged (27)	extracted (20)	21 (2)
(2,2-Dibromoethenyl)benzene (1)	(S)-BINOL (2)	triethylamine (3)	tetrahydrofuran	ethanol (17)	heated (23)	cooled (17)	77 (2)

図2：論文分析システムのインターフェース

実際の有機化学反応に関する論文から化学反応に関するメタデータを抽出するとともに、化学物質データベースなどに対応づけることにより、初期の用語集の作成といったコールドスタートの問題がなく、大規模な論文データを用いた解析を行うことが可能であることが確認された。

しかし、単純な情報抽出システムでは、照応解析などが必要な解析には不十分であること、今回利用した情報抽出システムでは、用語境界を誤認識する問題や、構造を表す文字の表記のバリエーションの影響を受けて、適切な化学物質 ID と対応づけられないといった問題があることが確認された。前者の単語境界については、部分文字列が用語境界を間違えていない対応可能な用語として認識されていることも多く、化学物質データベースを後処理に活用することで抽出システム自体の精度向上を図ることも検討する必要がある。また、後者の構造を表す文字の影響については、構造を表す文字を簡略化した基本構造を含む化学物質 ID の候補から適切なものを選択するといった手法についても検討する必要がある。

さらに、用語の正規化については、分野の広がりにより表記のバリエーションなどが多く存在することが確認された。これは、従来のナノデバイス開発分野のような研究領域よりも有機化学反応に関する研究領域の広がりが大きく、その表記の手法も多様になる場合があるだけでなく、同じ化合物についての正式な記法についても、いくつかのバリエーションが存在することに起因していると考えている。この問題に対しては、化学物質データベースはこのようなバリエーションに対応する形でデータベースを構築するため、対応可能な正規化が多く存在することが確認できたが、今後、他分野に展開を検討する際には、表記のバリエーションの取り扱いについて、より注意を払う必要があると考えている。

また、ナノデバイス開発の分野においては、特定の国際会議 (MNC: International Microprocesses and Nanotechnology Conference と SSDM: International Conference on Solid State Devices and Materials) について最新の会議録を含む 10 年以上の会議録を収集し、継続的に分析できる環境を構築している。

今後は、この分析で得た知見をもとに、より詳細な化学反応の手順を分析する方法の提案や、類似の研究分野へのデータベースシステムの展開についても検討していきたいと考えている。

## 5. 主な発表論文等

〔雑誌論文〕 計3件（うち査読付論文 3件/うち国際共著 1件/うちオープンアクセス 3件）

1. 著者名 Kojiro Machi and Masaharu Yoshioka	4. 巻 -
2. 論文標題 HUKB at ChEMU 2022 Task 1: Expression-Level Information Extraction.	5. 発行年 2022年
3. 雑誌名 CLEF (Working Notes) 2022	6. 最初と最後の頁 797-807
掲載論文のDOI (デジタルオブジェクト識別子) なし	査読の有無 有
オープンアクセス オープンアクセスとしている (また、その予定である)	国際共著 -

1. 著者名 Yuan Li, Biaoyan Fang, 他22名 (11番目にMasaharu Yoshioka)	4. 巻 -
2. 論文標題 Extended Overview of ChEMU 2022 Evaluation Campaign: Information Extraction in Chemical Patents	5. 発行年 2022年
3. 雑誌名 CLEF (Working Notes) 2022	6. 最初と最後の頁 758-781
掲載論文のDOI (デジタルオブジェクト識別子) なし	査読の有無 有
オープンアクセス オープンアクセスとしている (また、その予定である)	国際共著 該当する

1. 著者名 Kojiro Machi and Masaharu Yoshioka	4. 巻 -
2. 論文標題 HUKB at ChEMU 2021 Task 2: Anaphora Resolution	5. 発行年 2021年
3. 雑誌名 Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - to - 24th, 2021	6. 最初と最後の頁 720-731
掲載論文のDOI (デジタルオブジェクト識別子) なし	査読の有無 有
オープンアクセス オープンアクセスとしている (また、その予定である)	国際共著 -

〔学会発表〕 計6件（うち招待講演 0件/うち国際学会 2件）

1. 発表者名 鈴木 晃, 石井 真史
2. 発表標題 学術論文からの物性データ抽出に向けた文分類技術の開発
3. 学会等名 第69回応用物理学会春季学術講演会
4. 発表年 2022年

1. 発表者名 鈴木 晃, 石井 真史
2. 発表標題 磁石マテリアルズ・インフォマティクスのための物性値データマイニング
3. 学会等名 日本金属学会2022年春季(第170回)講演大会
4. 発表年 2022年

1. 発表者名 鈴木 晃, 石井 真史
2. 発表標題 材料辞書データベースを使った論文からの大量データ抽出：体系的自動タグ付け精度向上の検討
3. 学会等名 第83回応用物理学会秋季学術講演会
4. 発表年 2022年

1. 発表者名 吉岡真治, 町光二郎, 長田裕也
2. 発表標題 化学反応に関する情報抽出システムを用いた論文データベース分析システム
3. 学会等名 2023年度 人工知能学会全国大会 (第37回)
4. 発表年 2023年

1. 発表者名 M. Makino, S. Okuda, D. Goto, W. Jevasuwan, N. Fukata, and S. Hara
2. 発表標題 Effect of SiO <sub>2</sub> Mask on Selective-Area Growth of Ge Nanowires on Si (111) Substrates by Vapor-Liquid-Solid Method
3. 学会等名 35th International Microprocesses and Nanotechnology Conference (MNC 2022) (国際学会)
4. 発表年 2022年

1. 発表者名 M. Makino, S. Okuda, D. Goto, W. Jevasuwan, N. Fukata, S. Hara
2. 発表標題 Annealing Effect on Ge Nanowire Growth on Si (111) Substrates by Vapor-Liquid-Solid Method
3. 学会等名 35th International Microprocesses and Nanotechnology Conference (MNC 2022) (国際学会)
4. 発表年 2022年

〔図書〕 計0件

〔産業財産権〕

〔その他〕

-

6. 研究組織

	氏名 (ローマ字氏名) (研究者番号)	所属研究機関・部局・職 (機関番号)	備考
研究分担者	原 真二郎 (Hara Shinjiro) (50374616)	北海道大学・量子集積エレクトロニクス研究センター・准教授  (10101)	
研究分担者	鈴木 晃 (Suzuki Akira) (50799723)	国立研究開発法人物質・材料研究機構・統合型材料開発・情報基盤部門・NIMS特別研究員  (82108)	
研究分担者	長田 裕也 (Nagata Yuuya) (60512762)	北海道大学・化学反応創成研究拠点・特任准教授  (10101)	

7. 科研費を使用して開催した国際研究集会

〔国際研究集会〕 計0件

8. 本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関
---------	---------