学術情報のテキスト解析と生成 AI を用いた専門用語抽出

庄　金鳴 [1]　張　馨雲 [1,2]　成　凱 [1]
九州産業大学 理工学部情報科学科
[2] 早稲田大学大学院 情報生産システム研究科
chengk@is.kyusan-u.ac.jp

概要

学術情報の電子化・大規模化に伴い，学術データが大量に蓄積され，学術ビッグデータと呼ばれるようになっている．学術ビッグデータの解析は，専門分野の動向を把握・予測し，新しい課題を開拓することが期待されている．しかし，専門性の高い学術情報を解析するにはその分野の特有な表現（専門用語やトピック）を抽出することが難しいとされている．本発表では，学術情報の自然言語解析による用語抽出を検証しつつ，複数生成 AI の統合による専門用語抽出を考案し，実験による評価結果を報告する．

1　はじめに

近年，学術情報の電子化・大規模化に伴い，学術データが大量に蓄積され，学術ビッグデータ（Big Scholarly Data）と知られるようになっている [1][2]．学術データには研究者，研究組織，学術論文，論文誌や会議録等の出版物に関する情報を含み，分野別に特有な表現が使われる等，専門性の高いことが特徴である．学術ビッグデータ解析には，研究者ネットワーク解析，引用関係解析，影響力評価等のタスクが知られている．学術ビッグデータの解析は，専門分野の動向を把握・予測し，新しい課題を開拓することが期待されている．

一方，専門性の高い学術情報を解析する際には，その分野の特有な専門用語やトピックを効率よく抽出することやトレンドを適切に解析すること等，様々な課題が存在している．

近年，生成 AI の進化は，自然言語処理技術を活用した自動キーワード抽出を可能にした．生成 AI は膨大なテキストデータを基に文書の要点を抽出し，人間が行ってきた手作業のプロセスを大幅に効率化している．しかし，その結果がどの程度文書の主題を的確に反映しているか，またユーザーが期待する重要性や関連性を満たしているかについては，十分な検証がなされていないのが現状である．

表 1　生体医学ウェブ辞書に登録された専門用語例

番号	専門用語	説明
001	BRS	英語・固有名詞
	生体吸収性スキャフォールド	複合名詞
002	Judkins カテーテル	複合名詞
003	f-TUL	英語・固有名詞
	r-TUL	英語・固有名詞
	経尿道的尿管砕石術	複合名詞
004	iPS 細胞	固有名詞
	人工多能性幹細胞	固有名詞
006	アナログ電気回路	複合名詞
028	活動電位伝播	複合名詞
	興奮伝播	複合名詞
029	株化細胞	複合名詞
030	冠動脈形成術	複合名詞
032	間葉系幹細胞	固有名詞

生成 AI は膨大なデータを基に学習し，自動的にキーワードを抽出する能力を持つ．しかし，その抽出結果が文書の主題や内容をどの程度的確に反映しているかについては，未だ明確ではない．特に，生成 AI が抽出したキーワードが，人間の判断や専門家の分析基準とどの程度一致しているかを示す定量的な指標が不足している．また，客観的な基準を用いて評価した場合に，生成 AI のパフォーマンスがどのような傾向を示すかについても十分な議論が行われていない．

本研究は，生体医工学の学術情報を対象とし，学術情報の自然言語解析による用語抽出を検証しつつ，複数生成 AI の統合による専門用語抽出を考案し，実験による評価結果を報告する．

2　学術情報における専門用語

専門用語とは，ある特定の職業に従事する者や，ある特定の学問の分野，業界等の間でのみ使用され，通用する言葉・用語群である．「テクニカルターム」とも言われる．学会等の学術的グループでの専門用語は特に「学術用語」と呼ばれる．英語では，Terminology, Technical Terminology, Technical Term 等の表現がある．

専門用語と一般語の区別は明確ではなく，各分野の専門家であれば明らかに専門用語であると判断できるが，新しい技術に関する用語が出てきた場合，共通して専門性を認識するまでには時間を要する．また，ある分野の初心者によってその用語が専門用語だとわかったとしても，どの程度の専門性であるのか，その分野における基礎的・必須用語なのか，分からない場合が多い [10][11].

専門用語は，(1) 言語学的な構造 (Linguistical Structure)，(2) 単位性 (Unithood), (3) 用語性 (Termhood) により決められている [3][5]．さらに，専門用語に，専門性 (Specificity, Technicality)，基礎性 (Fundamentality)，先端性 (Advancedness) という特徴がある [5][?][6].

専門用語は，単名詞だけではなく複合名詞になることが多い．複合名詞の例として，Judkins カテーテル，3 次元造影，人工多能性幹細胞などがあげられる．テキストから，複合名詞を識別するため，複合名詞の文法的構造，特に品詞のつながり方を示す必要がある．

表 1 は，「生体医工学ウェブ辞典」に掲載された用語の例である [9]．この辞書は，医学と工学の学際領域である生体医工学分野に関わる基本的な用語およびそれに関わる知識を分かりやすく提供することを目的として，日本生体医学会によって編纂されたものである．専門用語は，複数の単名詞からなる**複合名詞**が基本として，そのうち辞書によって**固有名詞**と識別されるものもある．また，英語表記及びその略称となるものもある．

3　専門用語の抽出

専門用語抽出に関して，自然言語処理をはじめ，複数の分野で多くの手法が提案されている．既存の手法は，(1) 品詞と接続情報を用いた複合語抽出 [3][12] と，(2) 共起関係ネットワークに基づくキーフレーズ解析 [8] の二つに大分される．しかし，これらの手法は，複合語抽出には有効とされているが，しかし，抽出結果には一般語と専門用語が区別されず，専門分野に特有な専門用語の抽出にはまだ不十分と知られている [5][7].

3.1　接続情報に基づく用語抽出

専門用語は複合語の形になることが多く，テキストから複合語を適切に抽出することが専門用語抽出に必要である．複合語の重要度スコアは単名詞の左右にほかの単語と接続する．中川らはこの接続情報を利用して複合語を抽出する方法を提案した [?][12].

表 2　複合語重要度の計算例

単名詞	左方接続頻度 LN	右方接続頻度 RN
腫瘍	1	2
識別	2	3
器	1	1

まず，単名詞 N の左方接続頻度 LN と右方接続頻度 RN を求める．そして，単名詞の左右の接続頻度に基づき，複合語のスコアを計算する．単名詞 $N_1, N_2, \cdots, N_L$ が この順で連接した複合名詞を CN とする．CN のスコアとして各単名詞の左右のスコアの平均（ここでは相乗平均を採用する）を取り，CN の長 さに依存しないスコア $LR(CN)$ を式 (1) のように定義できる．

$$LR(CN) = \left(\prod_{i=1}^{L}(LN(N_i)+1)(RN(N_i)+1)\right)^{\frac{1}{2L}} \quad (1)$$

さらに，TF-IDF スコアと同様に，用語候補である単名詞あるいは複合名詞が単独で出現した頻度を考慮すべく, 式 1 を補正し，式 2 のように $FLR(CN)$ を定義する. また，「異なり数」，「パープレキシティ」のような単名詞の連接情報は，連接した単語の種類をカウントする方法（異なり数），（パープレキシティ）の方法がある．

$$FLR(CN) = f(CN) \times LR(CN) \quad (2)$$

「腫瘍識別器」における複合語抽出を考える．この語は 3 つの単名詞「腫瘍」「識別」「器」に分割できる．この際，表 2 に示すように，それぞれの単名詞が他の単名詞とどれだけ結びつくか統計的に分かっているとする．式 (1) によれば，CN =「腫瘍識別」のとき，$LR(CN) = \{(1+1)(2+1)(2+1)(3+1)\}^{1/4} = 2.913$, CN =「識別器」のとき, $LR(CN) = \{(2+1)(3+1)(1+1)(1+1)\}^{1/4} = 2.632$, CN =「腫瘍識別器」のとき，$LR(CN) = 2.569$ となる．

3.2 共起関係グラフに基づくキーフレーズ抽出

キーフレーズ抽出は，文章からその主題を良く表現している句を抽出する技術であり，統計ベース，グラフベース，機械学習ベースの手法に大別され，キーワード抽出ともいえる．キーフレーズは「人工|呼吸|器」のように複数単語の連続を抽出するので，単語を意味する「ワード」ではなく，句を意味する「フレーズ」が使われる．

グラフベースのキーフレーズ抽出手法が PageRank を根幹にして単語の共起関係ネットワークから，重要度を算出し，重要なフレーズを抽出している．本研究では，グラフベースのキーフレーズ抽出手法として MultipartieRank アルゴリズムを利用する．

専門用語抽出は，これまで紹介してきた技術である程度実現可能だが，完璧に行うためには専門分野の深い知識と理解が必要不可欠である．また，抽出結果は，形態素解析時に使用された辞書にも左右される．また，固有名詞の識別は辞書以外，固有表現識別技術も必要であろう．

3.3 生成 AI による専門用語抽出

生成 AI が抽出したキーワードの中から，TF-IDF を用いて重要なキーワードを特定し，それらが文書の主題をどの程度反映しているかを評価すること．これにより，生成 AI によるキーワード抽出の有用性を定量的に明らかにするとともに，使用者が早めに全文の内容を理解できる．TF-IDF は特定の単語が文書や文書集合内でどの程度重要であるかを定量的に評価するための指標である．

$$A_1 = (a_{1,1}, a_{1,2}, \cdots, a_{1,n_1}),$$
$$\cdots,$$
$$A_t = (a_{2,1}, a_{2,2}, \cdots, a_{t,n_t})$$

複数の質問セット $(Q_1, Q_2, \cdots, Q_t)$ が用意され，それぞれの質問が異なる AI エージェント $(AI_1, AI_2, \cdots, AI_t)$ に送信される．AI エージェントからの回答結果が $(A_1, A_2, \cdots, A_t)$ 各エージェントは，得意分野やモデル特性に基づいて回答を生成し，個別の専門用語リストとそれぞれのスコア（重要度）を出力する．図 1 に示すようにこれらの結果が統合プロセスを通じてまとめられる．統合されたリストは，全てのエージェントが提供したスコアを考慮しながら，専門用語の重要度順に並べられている．

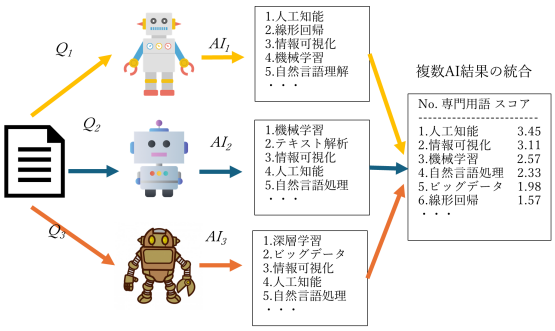

図 1 複数生成 AI 結果の統合による専門用語抽出

表 3 生体医工学会論文題目例

巻	号	論文題目
48	1	α 波を生理指標とした覚醒度と身体動揺との関係
48	1	UVB 照射によるマウス皮膚微小血管床における急性炎症反応に関する研究
49	1	腫瘍識別器の Leave-One-Out による性能評価結果の信頼性に関する考察

$W = (w_1, w_2, \ldots, w_k)$ は結果に含まれたすべての用語であり用語総数 K とする．$a_{i,j} \in W, n_i$ は用語 i の上記のリストにおける出現回数 $(tf(w_i))$ を表す．σ_i は TFIDF スコアである．用語の重要語スコアは，以下のように計算する．

$$\{(w_i, \sigma_i) | w_i \in \cup_{i=1}^{t} A_i, \sigma_i = tf(w_i) * idf(w_i)\}$$

4 評価実験

4.1 実験データ

本研究では，日本生体医学会において 2005 から 2023 年の間に掲載された学術論文 5,755 件を著者リスト，題目，出版年，巻，号を収集した．表 3 に収集されたデータの例を示している．今回の解析対象は論文題目だが，巻号，著者リスト，年度，ページ数等の情報も収集された．論文題目は日本語だけではなく，英語，記号等も含まれているが，解析は主に日本語を対象とした．

また，生成 AI を用いた用語抽出では，本学部卒業論文を対象として，論文本文 PDF からテキスト抽出を行い，目次や付録等不要部分を手作業で削除して，複数の生成 AI に専門用語抽出の指示をし，返された結果を利用した．

表 4　Termextract による専門用語抽出例

専門用語	重要度スコア	出現頻度
fNIRS	7.416	14
モデリング	5.099	5
タンパク質	4.899	4
ヘルスケア	4.472	6
数理モデル	4.000	16
マイクロ波	3.742	3
自律神経系	3.162	5
プログラム	3.000	10
周波数特性	2.449	3
有限要素法	2.000	3
生体インピーダンス法	1.000	3
定常状態視覚誘発電位	1.000	4
ニューラルネットワーク	3.742	6
心臓リハビリテーション	1.000	3
コラーゲンゲルチューブ	1.000	3
レギュラトリーサイエンス	3.464	4
ダブルルーメンカテーテル	1.414	4

表 5　spacy+pke によるキーフレーズ抽出例

キーフレーズ	重要度スコア
ヒト iPS 細胞 由来 心筋 細胞	0.003591
細胞 外電	0.003574
最適 設計	0.003573
ウェアラブル 心電図 計測 技術	0.003531
容量 結合 方式	0.003498
多元 計算 解剖 モデル	0.003474
空間的 配置 条件	0.003470
走査 方法	0.003444
超音波 音源	0.003424
内視鏡 シミュレーター ロボット	0.001887
心房 細動	0.001884
カラー オフセット csk	0.001882
VR 認知症 体験	0.001880
血液 浄化 療法	0.001870
機能的 MRI データ	0.001865
臓器 モデル	0.001858
血管 内皮 細胞	0.001856

4.2　テキストデータの前処理

前処理で最も重要な作業は，形態素解析である．使う形態素解析エンジンによって品詞体系や結果が異なり，また，使用する辞書によっても結果が左右される．本研究では形態素解析エンジン MeCab を使用し，新語・固有表現に強く，語彙数が多いシステム辞書 Mecab-ipadic-NEologd を使用する．Mecab-ipadic-NEologd は，多数の Web 上の言語資源から得た新語を追加することでカスタマイズした MeCab 用のシステム辞書であり，標準のシステム辞書では正しく分割できない固有表現を多数提供している．

4.3　自然言語解析による専門用語抽出

今回はキーフレーズ抽出手法を実装したライブラリとして pke[1)] を使うことにした．pke は英語・フランス語等の欧米の言語の対応が基本で分かち書きや品詞推定には spaCy が用いられている．日本語対応のため，GiNZA ライブラリを利用する．表 5 は，MultipartieRank によって抽出されたキーフレーズの例である．

4.4　複数生成 AI による専門用語抽出

二つの生成 AI, ChatGPT と Claude を利用して同じ論文について「専門用語を抽出してください」と指示した．二つの生成 AI の回答結果を，提案手法で評価し，スコアを算出した．

生成 AI の結果は，「潜在的な観光地」，「新型コロナウイルス感染症」，「ちずったー」，「観光資源の発掘」に対して，IDF 重みを計算する際に様々な分野の論文と比較した場合に，統合した結果に「SNS データ」，「潜在的な観光地」，「バックエンドサーバー」等が上位に表示された．IDF 重みの計算に使われる文書集合によって上位用語が大きく変わることがわかったが，紙面の制限で詳しい説明を省く．

5　終わりに

本研究では，専門性の高い学術情報解析のための専門用語抽出について自然言語解析と生成 AI 利用の手法を検討し評価実験を行った．まず，生体医工学会の学術論文題目を対象に専門用語抽出を試みた．次に複数の生成 AI に同じ文章から専門用語抽出の指示のもと，回答結果を TFIDF スコア算出のもとで統合し最終結果を求めた．さらなる評価等が今後の課題である．

1）　https://github.com/boudinfl/pke

参考文献

[1] C. Lee Giles. 2013. Scholarly big data: information extraction and data mining. In Proceedings of the 22nd ACM international conference on Information & Knowledge Management (CIKM '13). Association for Computing Machinery, New York, NY, USA, 1–2. https://doi.org/10.1145/2505515.2527109

[2] F. Xia, W. Wang, T. M. Bekele and H. Liu, "Big Scholarly Data: A Survey," in IEEE Transactions on Big Data, vol. 3, no. 1, pp. 18-35, 1 March 2017, doi: 10.1109/TBDATA.2016.2641460.

[3] Frantzi K, Ananiadou S, and Mima H. *Automatic recognition of multiword terms: the c-value/nc-value method*, International Journal on Digital Libraries, 2000, 3(2):115–130

[4] Justeson J S, and Katz S M. *Technical terminology: some linguistic properties and an algorithm for identification in text*, Natural Language Engineering, 1995, 1(1): 9–27.

[5] Fumimaro Odakura et al, *Active Learning for Extracting Technical Terms Covering Multiword Phrases*, iiWAS2021, pp. 311-318, https://doi.org/10.1145/3487664.3487706

[6] Nisha Ingrid Simon and Vlado Keselj. *Automatic term extraction in technical domain using part-of-speech and common-word features*, In Proceedings of the ACM Symposium on Document Engineering 2018. 1-4.

[7] Anna Hatty, et al, *Predicting Degrees of Technicality in Automatic Terminology Extraction*, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 2883-2889, July 5 - 10, 2020

[8] Florian Boudin. 2018. Unsupervised Keyphrase Extraction with Multipartite Graphs. In Proc. of the 2018 Conference of the North American Chapter, Volume 2, p. 667–672, 2018.

[9] 日本生体医学会, 生体医工学ウェブ辞典, http://jsmbe.org/cyclopedia.html, 2022.

[10] 内山 清子, 専門用語の専門性判定に関する一考察, Japio YEAR BOOK, 2010. pp.152-153

[11] 内山 清子, 専門分野における用語の分野基礎性に関する研究, 言語処理学会 第 17 回年次大会 発表論文集 (2011 年 3 月)

[12] 中川 裕志 等, 出現頻度と連接頻度に基づく専門用語抽出, 自然言語処理, 第 10 巻第 1 号, p.27-45, 2003 年. https://doi.org/10.5715/jnlp.10.27

A　生成 AI 回答結果統合による専門用語抽出

各生成 AI が抽出した専門用語は表 6 である．表 7 はこれらの回答結果を TFIDF スコアによって統合した結果である。

表 6　各 AI の専門用語抽出の回答結果

AI 回答結果 A_1	AI 回答結果 A_2
SNS 投稿	SNS 投稿
潜在的な観光地	潜在的な観光地
新型コロナウイルス感染症	ちずったー
観光資源の発掘	福岡県
福岡県	観光地発見
観光イメージ	フロントエンドアプリ
藤野研究室	バックエンドサーバー
位置情報	React
自動化システム	Express
福岡県観光地	データベース
都道府県魅力度ランキング	地図表示機能
世界遺産	検索機能
文化遺産	絞り込み機能
温泉地	現在位置表示機能
絶景スポット	距離表示機能
観光イメージ（自然，歴史，文化）	SNS データ
視覚化	アンケート調査
ちずったー	4 段階評価
Web アプリケーション	JavaScript
観光地情報	Node.js
検索機能	React
現在位置表示	Leaflet
距離表示	OpenStreetMap
地図表示（OpenStreetMap，Leaflet）	検索・絞り込み
バックエンドサーバー	SNS 投稿内容表示
データベース（tweets，users，places，categories）	ハッシュタグランキング

表 7　TFIDF スコアによる統合した専門用語抽出結果

keyword	score	keyword	score
SNS データ	10.5357	温泉地	5.2679
潜在的な観光地	9.1494	藤野研究室	4.5747
バックエンドサーバー	9.1494	距離表示機能	4.5747
ちずったー	9.1494	絶景スポット	4.5747
SNS 投稿	9.1494	OpenStreetMap	4.5747
距離表示	9.1494	地図表示	4.5747
React	6.6439	文化遺産	4.5747
観光地情報	5.2679	世界遺産	4.5747
Leaflet	5.2679	検索機能	4.4467
観光地発見	5.2679	位置情報	4.1796
メッシュ ID	5.2679	福岡県	4.1796
SNS 投稿内容表示	5.2679	Web アプリケーション	4.1692
観光資源の発掘	5.2679	フロントエンドアプリ	4.1692
福岡県観光地	5.2679	Node.js	3.3219
API 設計	5.2679	Express	3.1884
観光イメージ	5.2679	新型コロナウイルス感染症	2.7029