

科学技術文献に対する メタデータ付与の自動化

Automating Metadata Assignment for Science and Technology Literature

国立研究開発法人科学技術振興機構 情報事業部

菊井 玄一郎

NTT 研究所、岡山県立大学、農研機構（農業情報研究センター）を経て 2022 年 4 月より現職。博士（情報学）
専門は自然言語処理とその周辺分野。https://researchmap.jp/kikuig

国立研究開発法人科学技術振興機構 情報事業部

鈴木 慶二

機構に入職後、文献情報データの人手索引業務、研究開発戦略立案業務等を経て、機械処理システムの実行可能性調査に従事。現在は文献情報データ事業計画の策定等を担当。

国立研究開発法人科学技術振興機構 情報事業部

関根 基樹

機構に入職後、文献情報データの作成業務、研究者データベースの運営等を経て、現在は機械処理システムの運営に従事。

国立研究開発法人科学技術振興機構 情報事業部

水田 寿雄

機構の基礎研究・技術移転のファンディング業務等を経て、現在情報資産の提供や企画、文庫 DB 提供サービス等に従事。

✉ genichiro.kikui@jst.go.jp

1 はじめに

科学技術振興機構（JST）では 1960 年代から内外の科学技術文献（医薬系を含む）の書誌情報データベースを整備し、文献検索等の活用に供してきた。このデータベースを本稿では「JST 文献 DB」と呼ぶ。JST 文献 DB の収録対象は、論文、学術会議予稿集等の学術文献のほか、試験研究機関等の発行するレポートや商業雑誌など、単行本以外¹の広範な文献が含まれ、現在 6000 万文献（記事）を超える規模となっている。

メタデータの内容は広義の書誌情報であり、表題や著者名、発行年など、出版元から得られるデータに加え、

索引語^{[1][2]}、分野分類コード^[3]など JST の基準に基づいて付与されるものも含まれる。特に後者については、論文等の先端的な文献の主題を把握した上で JST の分類体系やシソーラス等に基づいて作成するきわめて専門性の高いものであることから、各分野の専門家が一定の経験を積んだ後に作成作業を行ってきた。しかしながら、このような作業には専門家の多大な労力が必要であり、近年の爆発的な文献の量に対応しきれなくなっている。また、熟練作業者を確保することも深刻な課題である。

このような状況のもと、JST では近年進展の目覚ましい大規模言語モデル (LLM: Large Language Models) を活用して、メタデータ付与の自動化に向けた検討を進めてきた^{[4][5]}。

本稿では JST 文献 DB におけるメタデータのうち、

1 周知の通り単行本のメタデータは国立国会図書館で作成・管理されている。

文献の内容分類を示す「分類コード」、および、文献内容を表す複数種類の「索引語」を対象として、JSTで進めている自動付与の試みについて紹介する。

この自動化における大きな課題はJSTの付与基準を極力反映したメタデータをいかにして自動作成するかということである。LLMを利用してこれを行う方法の一つは、付与基準（マニュアル等）を生成モデルに補助文脈として直接与え、指示文（プロンプト）によって、この基準に従った回答の生成を促す文脈内学習（in-context learning）である（「演繹的なアプローチ」）。もう一つは、当該基準に従って付与されたデータ（文献とメタデータのペア）が一定量得られる場合に、これらのデータを「付与基準が具現化された知識源」と考えて、LLMを学習（微調整：ファインチューニング）する方法である（「帰納的なアプローチ」）。JSTでは人手で付与されたメタデータを多数保有していること、データに基づいた最適化が容易であることなどから、今回は後者のアプローチをとることにした²。

以下、2章では本稿で対象とする4つの文献メタデータの概要について述べ、3章から6章まででメタデータごとに自動付与方法、および、性能評価について述べる。7章はまとめと今後の展望である。

2 本稿で対象とする文献メタデータ

本稿で自動付与の対象とするJSTの文献メタデータは文献の読解を経て付与すべきものであり、具体的には「分類コード」と3種類の「索引語」である。後者は全文献に付与される「主索引語（メインヘディング）」と医薬系文献にのみ付与される「チェックタグ」および「副索引語（サブヘディング）」である。以下ではこれらの概要について説明する。

2.1 分類コード（JST分類コード）

JSTでは4階層からなる分類コード体系を整備し、各文献に対して1つ以上、10個以下の分類コードを付与している。なお、2018年以降の文献DBにおいて、1文献あたり平均付与数は1.8個である。分類コード

は先頭から、2桁の英字と5桁の数字から構成され³、先頭から1,2,4,7桁目まででそれぞれ、1～4階層の分類を表す。表1に階層ごとの桁数、分類クラス数、コード文字列の例を示す。なお、コード体系の詳細は^[3]を、他の体系との比較は^[6]を参照されたい。

表1 JST分類コード体系の概要

階層	桁数	クラス数	コード例（日本語ラベル）
1	1	24	N（電気工学）
2	2	155	NC（電子工学）
3	4	786	NC05（電子回路）
4	7	3,209	NC05040（増幅回路）

2.2 主索引語（MH: Main Headings）

主索引語は当該文献を特徴づける用語（キーワード）であり、次の3つの「索引種別」に分けられる。

- ・**統制語（CT: Controlled Terms）**：文献を特徴づける基本的な用語であり、JSTシソーラス（現在約11万語）の見出し語（概念）から付与される。

- ・**準統制語（ST: Semi-controlled Terms）**：統制語以外で文献を特徴づける副次的な用語であり、原則として「JST大規模辞書（大規模辞書）」から付与される。なお、大規模辞書には存在しないが文献を特徴づける用語（主に新語）も準統制語として扱う。

ここで「JST大規模辞書」とは見出し語数80万語を超える科学技術用語集であり、各見出しには語義を示す「同義語ID」が付与され、各同義語IDには「代表表記」や「意味カテゴリコード」などの属性情報が紐付けられている。なお大規模辞書には統制語も収録されている。

- ・**物質索引（CN: Chemical Names）**：JSTの化学物質名用の辞書である日本化学物質名辞書（日化辞）から付与する。なお、日化辞のうち一般性のある用語は大規模辞書にも収録されている。

文献あたりの平均索引語数は理工系、医薬系ともにCTが10.2語、STが4.0語である。CNは理工系で0.25語、医薬系で0.63語であるが、詳細分野ごとの偏りが大きい（化学分野では多く、数学分野では殆ど出現しないなど）。

2 プロンプトを自動的に最適化する方法も最近研究されており、今後の検討課題と考えている。

3 実際のコードには8桁目に誤り検出のためのチェック桁（英文字）がある。

2.3 チェックタグ

医薬系の文献については、統制語（CT）のうち図1に示す13語を「チェックタグ」と呼び、検索結果の絞り込みに利用している。具体的には、検索インタフェースにおいてチェックタグの有無をチェックボックスで指定することにより、文献を容易に絞り込めるようにしている。

症例報告, ヒト, 男性, 女性, 胎児, 新生児, 乳児, 幼児, 子供, 青少年, 成人, 老人, 超高齢者

図1 チェックタグ一覧

2.4 副索引語 (SH: Sub-headings)

副索引語（SH）は「副題目」とも呼ばれ、医薬系文献に対してのみ付与される。SHは、MeSH^[7]のsubheadingsと同様、各文献の主索引語（MH）に付加されて、そのMHが当該文献中でどのような側面から言及されているかを示す「ラベル」である。JSTでは表2に示す26個のSHを定義している。

表2 サブヘディング一覧

薬物・化学物質を表す主索引に付与されるもの
治療利用*, 有害作用*, 薬理学*, 薬物動力学*, 多剤併用*, 薬物相互作用*, 内因性
疾患名を表す主索引に付与されるもの
化学誘発*, 病因+, 合併症*, 転移性, 薬物療法*, 放射線療法+, 食事療法, 外科的療法+, 治療*, リハビリテーション, 予防, 診断+, 病理, 疫学+, 遺伝学, 予後, 死亡率
薬物・化学物質, または, 疾患名のいずれか表す主索引に付与されるもの
分析+, 血液分析*

各SHは付加可能な主索引語（MH）の意味の範囲が決まっており、たとえば、「治療利用」というSHは薬物・化学物質を表す主索引語に対して付与可能である。表2はこの観点に基づいてSHを大きく3つに分けている。実際には各SHに対して、付与可能なMHの意味カテゴリコードの集合が定義されている。

ここで、意味カテゴリコードの制約を満たすMHであっても、実際にそのSHが当該MHに付与される比率は必ずしも高くないことに注意されたい。一部のサブヘディングについては意味的に付与可能なMHに対して実際に（すなわち文献の内容から判断して）付与されている比率が0.5%以下である（不均衡データ）。

なお、SHは過去の利用実績に基づいて利用頻度高・

中・低の3つに分類している。表2において*,+を付加したものがそれぞれ高・中に対応し、一定程度の正確性を保つ必要がある。

2.5 データの「揺らぎ」

ここで、各種メタデータの「揺らぎ」について注記する。分類コードにしても索引語にしても、これらの付与作業においては、文献の主題を認識し、それぞれの付与基準を考慮した判断が必要であることから、熟練者が作業する場合でも、通常、一定の「揺らぎ」は避けられない。例えばMeSHにおける自動索引に関する報告（たとえば^[8]）でも、hooper score⁴と呼ばれる尺度で評価されたディスクリプタの一致率は0.55程度であり、我々のデータでも同程度である。したがって、性能評価（正確性の評価）においては、この人手の揺らぎが基準となる。

また、過去に人手で付与されたメタデータを教師データとして使う場合は、教師データ自体に「揺らぎ」が存在するため、このことが性能の制約（上限）となることも考慮する必要がある。これらの点を勘案して、以下で述べる評価実験では、人手の性能を1とした相対値で結果を示すこととした。

3 分類コードの自動付与

3.1 分類コードの自動推定方法

テキストに対してその内容を表す複数の分類コードを付与するタスクは「マルチラベル・テキスト分類」と呼ばれている。分類クラス数が数千以上に及ぶものは、特に「大規模マルチラベル・テキスト分類」と呼ばれる。我々は、階層構造をなす分類クラスを扱うことができ比較的シンプルな「MATCH」と呼ばれる手法^[9]を利用した。

この手法はエンコーダモデル（BERT）を用いた通常の（マルチラベル）テキスト分類を次の3つの点で拡張したものである。

- 1) テキストの埋め込み空間に書誌情報（雑誌名など）の内容を含めたこと。
- 2) 文献全体を表す埋め込みベクトルの次元⁵を中間層

4 F1値に類似した一致性のスコア。F1値と同様0から1の実数値であり、0.1の場合以外はF1より低い値となる。

5 BERTにおける（先頭の）CLSトークンの数

の次元数の数倍（例 8 倍）にして情報量を増やしたこと。
3) 各クラスのパラメータがクラス間の階層関係を反映するように微調整用の目的関数を改変したこと。

3. 2 分類コード推定の評価実験

評価指標として、階層ごとに F1 値を算出し、人手の F1 値を 1 とした場合の相対値（相対 F1 値）を計算した（表 3）。ここで、人手の F1 値とは、専門作業者が付与した「正解（Gold Standard）」とは独立に、別の専門作業者が付与した分類コードを「推定値」とみなした時の F1 値である。自動推定の結果は概ね人手並の精度を達成していると言える。

表 3 分類コードの階層別推定性能

階層レベル	1	2	3	4
相対F1(医薬系)	1.003	0.975	0.970	1.023
相対F1（非医薬系）	1.009	1.027	1.041	1.153

（相対 F 値とは各階層において、人手付与結果の F1 値を 1 としたときの自動推定の F 値の割合のこと）

4 主索引の自動付与

4. 1 主索引の自動付与方法

主索引の自動付与方法の詳細は文献 [5] に譲り、ここでは概要と評価結果のみ説明する。

図 2 に今回開発した主索引語自動推定処理の全体構成を示す。文献の表題、本文（または抄録）を入力として、簡単なテンプレート埋め込みにより、「この文献のキーワード（索引語）のリストを生成せよ」という趣旨の指示文（プロンプト）を生成する。なお、この段階では索引種別（CT、ST、CN）は区別しない。次の LLM 推論処理では、この指示文を文脈として回答（＝索引語リスト）を生成させる。LLM の生成は確率的なゆらぎ

を伴うため、生成処理の呼び出しパラメータ設定により、1 つの指示文に対して複数個（例：20 個）の索引語リストを生成し、生成頻度の少ない索引語はノイズと考えて後処理で削除する。後処理では、得られた索引語リストの各語に対して各種辞書引きを行い、同義語の集約と索引種別の決定を行う。どの辞書にも含まれない語は候補から削除する。

LLM 推論処理で使用する LLM は、公開パラメータモデルに対して、JST で作成した索引語を正解データとして微調整（ファインチューニング）したものである。具体的には、推論と同様の方法で作成した指示文と、当該文献に対して人手で作成した索引語集合（Gold Standard）のペアを教師データとして、教師あり微調整（SFT: Supervised Fine-Tuning）を行った。

4. 2 主索引推定の評価実験

既に人手により索引語が付与されている文献に対して自動推定の結果を評価した。ここでも既存の人手の索引（「正解」とは独立に専門作業者による索引を別途作成し「正解」との一致性（F1 値等）を基準性能として、自動索引の性能を相対的に評価した。表 4 に結果を示す。微調整済み LLM を用いた結果は医薬系で若干人手に及ばないが概ね人手に近いスコアである。なお、実験

表 4 主索引語推定の評価結果（[5] の表 2 を調整）

分野	索引種別	相対F1値	
		微調整済み	微調整なし
医薬系	CT	0.909	0.587
	ST	1.030	0.637
	CN	0.939	0.839
非医薬系	CT	0.996	0.630
	ST	1.045	0.670
	CN	1.076	0.728

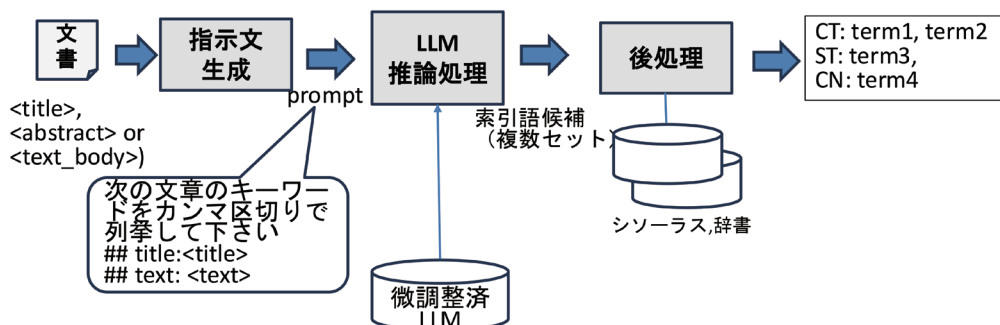


図 2 主索引語自動推定の全体図

の詳細は文献^[5]を参照されたい。

また、相対 F1 値による評価に加えて、専門家に人手の索引と自動索引の結果をブラインドで優劣評価させたところ、図 3（主索引語全体のブラインド評価）に示す通り、両者の間に有意な差は見られなかったという結果となった。

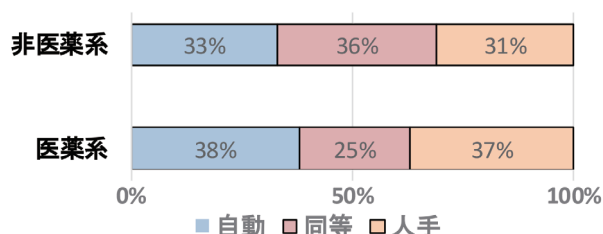


図 3 主索引語全体のブラインド評価
(自動索引と人手索引のどちらが好ましいか)

5 チェックタグの自動推定

5.1 チェックタグの自動推定方法

チェックタグは全てシソーラス語 (CT) であるから、その推定処理は主索引の自動推定処理に含まれる。しかしながら、チェックタグは他の主索引語と比べてより高い性能が望ましいため、チェックタグごとに専用のテキスト分類処理を構築した。分類モデルへの入力チェックタグ (例：高齢者) と文献テキストを併置したものであり、出力は当該文献に当該チェックタグを付与するかどうかである (二値分類)。ベースモデルとして、長い入力を効率的に扱える「ModernBERT」を採用し⁶、過去に人手で付与されたチェックタグを正解データとして微調整 (ファインチューニング) を行った。

5.2 チェックタグ推定の評価実験

各チェックタグについて、その性能を人手の F1 値に対する相対値で評価した。正解率 (accuracy) でなく F1 値を用いたのは、チェックタグの多くが正例と負例に偏りのある不均衡データであるためである。表 5 に示す通り、人手と遜色ない結果が得られた。

表 5 チェックタグ推定の評価結果

索引語	相対F1	索引語	相対F1
ヒト	1.01	幼児	1.139
症例報告	1.006	子供	1.219
男性	1.007	青少年	1.061
女性	1.015	成人	1.058
胎児	1.217	老人	1.048
新生児	1.22	超高齢者	0.996
乳児	1.112		

6 副索引の自動推定

6.1 副索引の自動付与方法

副索引語は、「文献 (テキスト)」と「メインヘディング (主索引語)」の組に対して、特定の SH (副索引語) が付与できるかどうかを判別する二値分類問題と見なすことができる。ただし、2.4 で述べたように、いくつかの副索引語は正例率が 0.5% 未満と極めて低い「極端に不均衡な判別問題」となっており、データからの学習には配慮が必要である。

この課題に対し、我々は、教師ありテキスト分類処理を基本とし、データが不均衡な少数の SH については索引付とマニュアルを指示文に加えることによる文脈内学習手法を併用することを提案した [5]。しかしながら、前者の方法において、入力列構成の調整、正例重みの制御、学習時の最適化指標を F1 値から AUPRC (Area Under the Precision-Recall Curve) に変更したことなどにより、不均衡データに対する性能が向上したため、今回の報告では全て分類学習 (ファインチューニング) による方法で推定を行った。ここで、SH は全部で 26 語 (ラベル) と多いため、一度に学習・推論可能なマルチラベル分類の活用^[9]も効率面で有望であるが、全パラメータを特定の SH 向けに微調整できることから、本研究では SH ごとに個別に微調整した。

6.2 副索引自動付与の評価実験

他のメタデータと同様に、評価セットに対して精度・再現率・F1 値による評価を行った。正負判定時の閾値等のパラメータは評価セットとは別のデータで F1 値が最大になるように調整した。利用頻度が中以上の SH に対する相対 F1 値を表 6 に示す。ほとんどの SH で

6 ModernBERT で得られた埋め込みベクトル列をプーリングにより集約後、線形層で 2 値にマッピングする。

人手と遜色ないスコアであるが、薬物動力学、薬物相互作用が人手の85%程度でありさらなるモデル改善が必要である。なお、これら2つについては当面、再現率高めに出力して、付与されたSHに対して人手のチェックを行う予定である。

表6 サブヘディング推定の評価結果

SH名	相対F1	SH名	相対F1
治療利用*	0.974	薬物療法*	1.003
有害作用*	1.027	放射線療法+	0.917
薬理学*	0.974	外科的療法+	0.978
薬物動力学*	0.865	治療*	1.145
多剤併用*	0.913	予防+	1.013
薬物相互作用*	0.869	診断+	1.086
化学誘発*	0.929	疫学+	0.967
病因+	1.021	分析+	0.946
合併症*	1.149	血液分析*	1.051

7 おわりに

JSTにおける文献メタデータ作成の自動化に向けた取り組みについて述べた。分類コードおよび3種類の索引の全てにおいてエンコーダモデルを含むLLMを活用し、主索引語付与においては生成的手法、そのほかについては分類手法を利用した。その結果、平均性能において人手と同等、あるいはそれに迫る高い精度を確認することができた。しかしながら、個別に結果を精査すると、必ずしも適切とは言えない出力も依然として存在する。このような一部の誤りを自動的に検出・訂正する仕組みが必要である。

また、今回の取り組みでは入力文献のテキストデータを対象としたが、今後は図表を含めた文書理解への拡張が求められる。たとえば、有機化学や薬学の分野ではキーとなる化学物質が構造型（図）として提示されており、テキストでは言及されていないことがしばしばある。このようなケースでは、マルチモーダルLLMなどを用いて、画像情報から化学物質に対する索引を（IUPAC⁷命名法やSMILES⁸記法のようなテキスト表現で）付与することが有用であろう。

マルチモーダルの問題とも密接に関係するが、文献の

電子化も大きな課題である。JSTでは特に「J-STAGE」において、人間が読むためのPDFに加え、機械可読なXML形式での記事公開を促進しており、今後の普及が期待される。当面の間は既存のPDFやテキストの混在環境に対応しつつ、自動付与システムのさらなる高精度化と運用コストの最適化を進めていく予定である。

8 謝辞

本研究開発の一部は国立研究開発法人産業技術総合研究所および株式会社AIST Solutionsが提供するABC13.0「開発加速利用」の支援を受けて実施した。

参考文献

- [1] 堀内美穂、中村徹、永井賢吉：JSTDB 検索における索引の有効性と索引作業の重要性-JSTPlus ファイル, JMEDPlus ファイルにおける索引語の分析-, 情報管理, vol.52, No.1, 2009.
- [2] 富永祥平：“索引作業におけるより高度な支援辞書の利用～JST 抄録・索引支援システム「NAISS」について～”, 情報管理, vol. 50, no. 4, pp. 210-217, 2007.
- [3] https://dbs.g-search.or.jp/jdsub/bunrui/bunrui_index.htm (2026.6.23 アクセス)
- [4] 菊井玄一郎、鈴木慶二、関根基樹、水田寿雄：“ラベル依存の注視機構を用いた医薬系文献に対する統制語付与”, 第29回言語処理学会年次大会, pp.146-150, 2023.
- [5] 菊井玄一郎、鈴木慶二、関根基樹、水田寿雄：“大規模言語モデルを用いた索引語の自動付与”, 第32回言語処理学会年次大会, pp.746-751, 2026.
- [6] 小野寺夏生：“論文データベースにおける主題分類—情報分析への利用の視点から—”, 情報の科学と技術, 66巻6号, pp.272-276, 2016.,
- [7] <https://www.nlm.nih.gov/mesh/meshhome.html>
- [8] James Mork, Alan Aronson Dina Demner-Fushman：“12 years on - Is the NLM medical text indexer still useful and relevant?”, Journal of Biomedical Semantics, 2017 Feb 23;8(1):8. doi:

7 International Union of Pure and Applied Chemistry

8 Simplified Molecular Input Line Entry System

10.1186/s13326-017-0113-5, 2017.

[9]MATCH: Yu Zhang, Zhihong Shen, Yuxiao Dong, KuansanWang, Jiawei Han: “Metadata-Aware Text Classification in A Large Hierarchy” , ArXiv, arXiv.2102.07349, 2021.

[10] 杉本海人、壺岐太一、知田悠生、金沢輝一、相澤彰子 : “JMedRoBERTa: 日本語の医学論文に基づいた事前学習済み言語モデルの構築と評価” , 第 29 回言語処理学会年次大会 , pp.707-712, 2023.

