大規模言語モデルを用いた索引語の自動付与

菊井 玄一郎[1]， 鈴木 慶二[1]， 関根 基樹[1]， 水田 寿雄[1]

[1]国立研究開発法人 科学技術振興機構 情報企画部
{genichiro.kikui, k3suzuki, sekine, mizuta}@jst.go.jp

概要

本稿では大規模言語モデル(LLM)を利用した、科学技術文献に対する索引語の自動付与方法について報告する。本稿で対象とする索引語は全文献に付与される「主索引語」と医薬系文献において主索引語に付加される 26 語の「副索引語」である。前者については、過去の人手索引データで微調整（fine-tuning）した LLM で索引語の候補を生成し、辞書を用いて同義語集約、フィルタリング等を行う。後者についてはエンコーダ型言語モデルの微調整による教師あり二値分類を基本として、正例の少ない一部の索引語については LLM による生成的手法を適用した。評価実験の結果、主索引については人手と遜色がなく、副索引は人手と比較して概ね 8 割以上の性能を達成した。

1 はじめに

JST では内外の科学技術文献（医学分野を含む）の検索・活用のため、これらに対して索引（語）を付与した文献データベースを整備している([1]､[2])。JST の索引語は大きく「主索引語（以下、メインヘディング；MH）」と「副索引語（以下、サブヘディング；SH）」に分けられる。MH は通常の文献キーワードに相当し、基本的に JST の科学技術用語辞書群から付与される統制語索引である。SH は医薬系文献のみを対象に主索引語に付加されるラベルであり、26 語（ラベル）の中から付与される。

索引語を正しく付与するためには文献の主題を把握し、その内容を適切な統制語群にマッピングする専門性の高い作業が必要である。このため、従来は主に訓練を受けた専門家が人手で行なってきた。しかしながら、作業時間の増大や熟練作業者の不足などの課題を受けて、自動化の検討を進めている([3])。

MH の自動付与については「キーワード抽出」などの名称で連綿と研究が行われてきた。伝統的な方法はテキストから名詞（句）的な単語列を抜き出して、統語的・統計的に重要なものを選ぶという「抽出型」の方法である([4]など)。統制語付与を超大規模なマルチラベル分類問題とみなす「分類型」の方法も試みられている([3])。さらに、近年では索引対象テキストを先行文脈として LLM に投入し、索引語を得る「生成型」の手法が注目されている([5]など)。

副索引の付与は文献と 1 つの MH を入力とする 26 ラベルのマルチラベル分類問題[6]（または 26 個の 2 値分類問題）とみなすことができる。テキスト読解を伴う分類問題についても生成型の手法([5])が有望である一方、学習データが一定程度あれば言語モデルの微調整（fine-tuning）による方法([7][8])も自動的な最適化が可能であるという点で検討に値する。

JST では過去に人手で付与された索引データがあり、このうちの一部文献については自動索引の入力側、すなわち、テキストデータ（本文、あるいは、抄録）が利用可能である。そこで、我々はこれらのデータを用いて大規模言語モデル（LLM）、および、そのための指示文を調整することにより、索引語付与の自動化を試みる。

機械学習においてはデータの品質が大きな鍵となる。一貫性の低い学習データは性能の上限を制約する。過去の研究によれば、索引語を人手で付与した場合、その一貫性はある程度にとどまることが知られている([9]など)。そこで、本稿では我々のタスクにおける「人手索引の一貫性」の近似的な評価を行い、自動索引の精度目標の参考とする。

以下、第 2 章で主索引語、第 3 章で副索引語の自動付与方法について説明する。第 4 章で評価実験と結果について述べ、第 5 章でまとめる。

なお、JST では文献を「医薬系」と「医薬系を除く理工系」の 2 つに分けて扱っている。以下では後者を単に「理工系」と呼ぶ。また、本稿で対象としている「文献」は論文、研究報告、会議予稿集などの学術文献、雑誌記事、各種報告書など、書籍以外の広範な科学技術文献である。

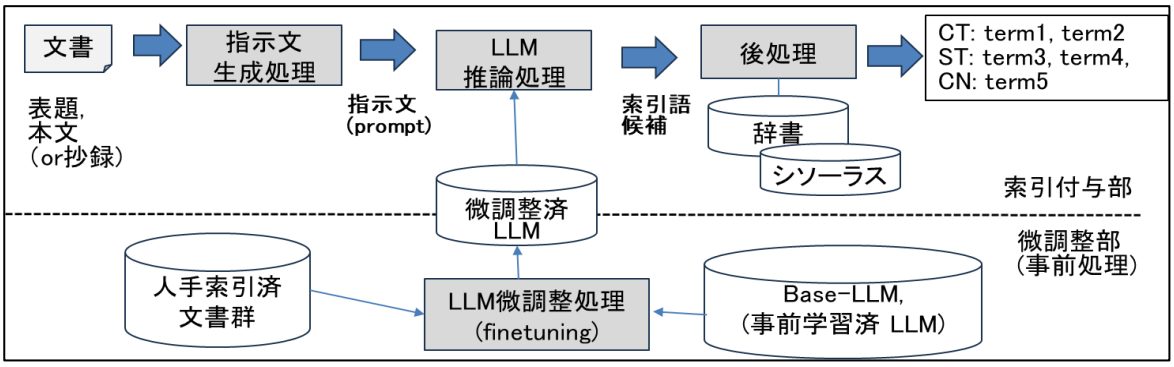

図 1　主索引語自動付与の全体構成を示すブロック図

2 主索引語の自動付与

2.1 主索引語（MH）の概要

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

・（コア）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（約 3.7 万語、2024 年版で 11 万語に拡張)の見出し語から付与される。

・準統制語（ST: Semi-controlled Terms）：統制語以外で記事を特徴づける用語であり、原則として準統制語用の辞書（付録 A）から付与される。この辞書の各見出しには語義を示す「同義語 ID」と「意味カテゴリコード」が付いている（同義語 ID のタイプ数は約 26 万）。なお、辞書に存在しないが文献を特徴づける用語もここに含める。

・物質索引(CN: Chemical Names)：JST の化学物質名用の辞書（付録 A）から付与する。ST と同様、辞書不掲載であるが文献を特徴づける語を含める。

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 語であり、CN は理工系で 0.25 語, 医薬系で 0.63 語 である。

2.2 主索引語の自動付与手法

図 1 に JST における主索引語自動付与の全体構成を示す。全体は点線の上側の索引付与部と下側の微調整(fine-tuning)部とに分けられる。

索引付与（推論）部

索引付与部では、まず、指示文生成処理で文献の表題、本文（または抄録）を入力として、図 2 に例を示すようなテンプレートを用いて指示文（プロンプト）の生成を行う。この例で<title>および<text>にはそれぞれ文献の表題と本文（または抄録）が代入される。なお、この段階では索引種別（CT, ST, CN）は区別しない。

```
次の文章を特徴づけるキーワードのリストをカンマ区切りで出力せよ。
## 表題: <title>
## 本文: <text>
```

図 2　指示文テンプレートの例

次の LLM 推論処理では指示文を先行文脈として回答、すなわち、索引語リストの生成を行う。生成時の確率的なゆらぎに対応するため、一つの指示文に対して、複数個（例：20 個）の回答（=索引語リスト）を生成する。

後処理では、得られた索引語リストを集約する。ある索引語が k 個のリストに出現した場合、その索引語の初期スコアを k とする。次に辞書類を参照して同義語の集約と索引種別の決定を行う。2 つの索引語が同一の同義語 ID を持つ場合は 1 つに集約し、それぞれのスコアの和を新たなスコアとする。さらに準統制語用の辞書のエントリに「同時に索引すべき統制語」が記載されていればそれに従って統制語の追加も行う。もし、LLM から出力された語がどの辞書にも含まれない場合は候補から削除する[i]。最後にスコアが閾値以下の索引語を削除して最終的な結果とする。なお、閾値は検証データの F1 値が最大になるように決定する。

微調整（ファインチューニング）部

主索引語の候補を生成する LLM は汎用の指示チューニング済み LLM を人手索引データで微調整して利用する。訓練サンプルは、推論時と同じ方法で

[i] 辞書類に追加する新語の候補として利用する。

生成した指示文の後に人手の索引語をカンマ区切り等で文字列化した「回答文」を付加したものである。この学習サンプルの指示文部分を含む全トークンを損失計算の対象としてモデルパラメータの更新を行う。指示文を含めるのは指示文中の文献の表題や本文が分野、文体の適応に有効と考えたためであり、予備実験でも有効性を確認している。なお、メモリ使用量と学習コストを低減するために低ランクアダプター(LoRA)を使用した。

3 副索引語（SH）の自動付与

3.1 副索引語の概要

副索引語[ii]（SH）は、PubMed における MeSH[10]の subheading と同様、医薬系文献において主索引語(MH)に付加されて、その MH が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 26 個の SH を定義している（付録 B）。SH ごとに付加可能な MH の「意味カテゴリコード」の範囲が決まっている。なお、SH は利用頻度に応じて（利用頻度）高、中、低の 3 種類に分けている（ラベル数はそれぞれ 11,7,8。付録 B）。

3.2 副索引の自動付与手法

第 1 章で述べたように、副索引語の付与は分類問題と見なすことができる。但し、いくつかの副索引語は正例率が 0.005 未満と極めて低く、学習が難しい。そこで、我々は正例率が一定以上の 24 個の副索引語については教師あり微調整による分類手法、残りの 2 つの副索引語については微調整の適用を断念し、副索引付与マニュアルをもとに人手で作成した指示文により各副索引語の付与の可否を回答させる生成的な手法とした。

テキスト分類手法として、PET（Pattern Exploiting Training）[7]を改良した ADAPET [8]と呼ばれるアルゴリズムを試した。PET は分類問題を空所補充（＝単語予測）問題に変換することで、微調整の効果を高める方法であり、ADAPET は PET の損失関数を改良して学習サンプルの利用効率の向上を図った手法である。空所補充問題への変換のためのテンプレートを付録 C に示す。

4. 評価実験

以上で説明した手法による索引付与の出力性能について評価実験を行った。実験は医薬系分野と理工系分野に分けて行った。

4.1 実験データ

実験用の文献テキストおよび正解索引(ground truth)は JST の文献データベースからサンプリングして準備した。理工系、医薬系とも、2019 年から 2023 年にかけての文献データから、評価での利用に不適当なもの（抄録テキストが自動翻訳結果である文献など）を除いてベースセットとした。理工系文献はベースセットをランダムに訓練用、検証用、評価用に分割した。医薬系についても同様であるが、「重要な副索引」のうち出現頻度が極端に少ない「薬物動力学」と「薬物相互作用」については、評価セットに一定量含まれるようにサンプリングし、残りを検証用、訓練用のデータとした。それぞれのデータ量を表 1 にまとめる。

表 1 評価実験に用いた文献数

	訓練・検証データ	評価データ
理工系	855,517	800
医薬系	273,802	600

4.2 実験手法

主索引語(MH)については、訓練・検証データを用いてベースモデル[iii]を微調整した後、検証データの F1 値（マイクロ平均）が最大となるように後処理の閾値を決定した。

副索引語(SH)のうち、分類モデルによる方法については、日本語版の modern-bert モデル[iv]に対して、索引語（ラベル）ごとに ADAPET による微調整を行い、MH を自動付与した評価データに対して SH 推定を行った。学習データにおいて正例比率が概ね 5%以下の索引語については一つのバッチに正例が１つは入るように負例を間引いてサンプリングした。推定の際には検証データにより F1 値が最大になる尤度閾値を設定して正負を判定した。

[ii] 副標目とも呼ばれる

[iii] https://huggingface.co/mistralai/Mistral-Nemo-Instruct-2407

[iv] SB Institutions（株）が事前学習したもの。https://huggingface.co/sbintuitions/modernbert-ja-310m

副索引のうち、生成的手法による索引付与判定については医薬系ドメイン用の指示チューニング済みモデル[v]を用いた。

4.3 評価尺度

評価尺度は過去に人手で付与した索引語を正解（ground truth）とした場合のシステムの出力の精度(precision)、再現率(recall)、F1 値である（マイクロ平均、付録 D）。なお、ground truth とは独立に、別途人手作業により評価対象文献に付与した索引に対して、ground truth を正解とした場合の評価値を算出し「人手の一貫性（性能）」の目安とした。

4.4 結果と考察

4.4.1 主索引語(MH)

索引種別ごとに精度、再現率、F1 値のマイクロ平均を算出した。表 2 に人手、自動（微調整済みモデル, Auto(FT)）、自動（微調整無しモデル）; Auto(base)）による索引結果の F1 値を示す。微調整済みモデルによる結果は人手より若干落ちるものの、概ね同等の性能であることが分かる。一方、微調整前のモデルの性能はかなり低く、微調整の効果を示している。なお、一部で自動索引が人手を超えているのは、微調整の結果、過去データに良く適応したためと思われる。

表 2 主索引語推定の評価値（F1 値）

分野	種別	人手	Auto (FT)	Auto (base)
医薬系	CT	0.572	0.520	0.305
	ST	0.462	0.476	0.303
	CN	0.799	0.750	0.629
理工系	CT	0.486	0.484	0.305
	ST	0.421	0.440	0.295
	CN	0.578	0.622	0.453

4.3 副索引語(SH)

利用頻度中以上の分類手法（ADAPET）の評価結果を表 3 にまとめる。殆どの SH については人手と同程度か 80%を超える F1 値となったが、正例データの少ない薬物動力学、薬物相互作用については 80 %を下回る結果となった。これらは利用頻度の高い SH でもあることから、LLM による生成手法を適用した。その結果、薬物動力学で F1 値 0.612(人手の 84%), 薬物相互作用で同 0.615（80%）となり、指示文の調整に人手を要するとはいえ、人手の概ね 8 割程度には改善されることが分かった。

表 3 副索引語推定の評価値（F1 値）
*付の SH は利用頻度「高」のもの

副索引語(SH)	人手①	ADAPET②	②/①
治療利用*	0.780	0.702	0.90
有害作用*	0.622	0.551	0.89
薬理学*	0.664	0.641	0.97
薬物動力学*	0.778	0.559	0.72
多剤併用*	0.762	0.728	0.96
薬物相互作用*	0.770	0.396	0.51
化学誘発*	0.869	0.818	0.94
病因	0.418	0.359	0.86
合併症*	0.511	0.58	1.14
薬物療法*	0.826	0.801	0.97
放射線療法	0.752	0.743	0.99
外科的療法	0.809	0.774	0.96
治療*	0.615	0.69	1.12
予防	0.623	0.521	0.84
診断	0.612	0.667	1.09
疫学	0.521	0.579	1.11
分析	0.205	0.194	0.95
血液分析*	0.492	0.702	0.90

5. まとめと今後の課題

JST における索引語付与の自動化手法について解説した。主索引語については過去の人手索引データにより微調整した LLM による「キーワード生成」に辞書類を組み合わせた手法により人手レベルの性能を達成した。副索引語については２つを除く重要索引語について、人手の索引結果を学習データとして LLM を微調整した二値分類手法により、人手より多少劣化する程度の性能を達成した。正例が僅少なため分類学習が厳しい副索引語についてはプロンプト調整により性能を向上させることができた。

今後の課題として、主索引については辞書不掲載であるが索引語とすべき用語の付与方法の検討がある。副索引についてはまだ人手レベルに及ばないものがいくつかあるため、索引マニュアルと実際の索引例を併用する few-shot のアプローチなどが考えられる。

[v] https://huggingface.co/google/medgemma-27b-text-it

謝辞

索引実験、データ整備、指示文調整の作業を分担していただいた跡見真児、真瀬進、加藤凉介の各氏に感謝する。また、株式会社 Preferred Networks の岩澤 諄一郎、吉川 真史両氏との LLM を用いた副索引付与に関する議論は示唆に富むものであった。記して感謝する。

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] 冨永祥平: “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] 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.

[4] C.D.Manning, et al.（原田隆史ほか訳）: “情報検索の基礎”, (Introduction to Information Retrieval, Cambridge University Press,2008) オーム社, 2012.

[5] Byungha Kang and Youhyun Shin: “Empirical Study of Zero-shot Keyphrase Extraction with Large Language Models”, Proc. of Coling-2025, pp, 3670-3686, 2025.

[6] 杉本海人、壹岐太一、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.

[7] Schick, T. & Schütze, H. (2021). Exploiting Cloze Questions for Few Shot Text Classification and Natural Language Inference. EACL 2021.

[8] Tam, D. R., Menon, R. R., Bansal, M., Srivastava, S., & Raffel, C.: “Improving and Simplifying Pattern Exploiting Training”,. arXiv:2103.11955、2021.

[9] 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: 10.1186/s13326-017-0113-5, 2017.

[10] https://www.nlm.nih.gov/mesh/meshhome.html

付録 A：索引用の辞書について

主索引で利用する JST の辞書リソースは、シソーラス辞書、大規模辞書、日本化学物質名 DB（日化辞 DB）の３つである。
・シソーラス辞書（統制語辞書）：1 語義 1 見出しで語間に階層関係が定義されている。
・大規模辞書：シソーラス語や利用頻度の高い化学物質名を含む包括的な用語辞書であり、同義語も見出しとして含んでいる。準統制語として付与されるのは大規模辞書のうちシソーラス語と化学物質名を除いた語彙である。化学物質名は物質索引語（CN）として扱われる。
・日化辞 DB は化学物質名の辞書であり、通常の単語（固有名詞）である慣用名と IUPAC (International Union of Pure and Applied Chemistry)命名法に従って化学構造式を文字列化した「体系名」に分けられる。後者は本文からではなく構造式（図）を参照して索引されることが多いため、テキストからの自動索引では慣用名のみを対象としている。

付録 B　副索引語（ラベル）一覧

末尾に＊、+を付与した語はそれぞれ利用頻度が高、および中の副索引語。

主に薬物・化学物質である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

主に疾患名である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、および、疾患名の両者に付与されるもの

分析＋、血液分析＊

付録 C　副索引語の推定問題を空所補充に変換するためのテンプレート

″〈title〉〈text〉この文章は〈MH〉について〈SH〉の観点で述べていますか？［LBL］″

〈title〉〈text〉はそれぞれ文献の表題と本文（または抄録）に置き換えられる。MH と SH にはそれぞれメインヘディング、判定対象のサブヘディングを代入する。LBL は分類ラベルに対応する言語表現(verbalizer と呼ばれる)が入り、学習時には positive の場合「はい」、negative の場合「いいえ」(negative)を代入し、推論時は空欄（MASK トークン）とする。

付録 D　評価値計算

評価セット（文献集合）の各文献について、システムが出力した索引語数を(SYS)、ground truth の索引語数を REF とし、システムが出力し、かつ、ground truth でもある索引語数を TP とする。全文献に対して SYS, REF, TP それぞれの和を計算し，次の式で精度(precision) と再現率, F1 値を求める。
精度＝TP の和/SYS の和
再現率=TP の和/REF の和
F1 = 2*精度*再現率/(精度+再現率)