ABCI 3.0開発加速利用 （2025年度）成果概要（公開用）

課題名	実施時期：令和7年度(2025.5-2026.3)
科学技術文献に対するメタデータ自動付与に関する研究開発	所属機関名：国立研究開発法人　科学技術振興機構 代表者氏名：水田寿雄

成果概要：
実施機関(JST)の保有する科学技術文献データを教師データとしてオープンパラメータモデルを継続学習および教師ありファインチューニング(supervised fine-tuning)することにより、文献に対する索引語（主索引、副索引）および分野分類コード（約3000カテゴリ）の自動付与方法を開発し、人手の80%-100%の精度を達成した。

成果のポイント：

概要と目的：科学技術文献のアクセス支援に向けて、各論文に対してその内容に適合したメタデータ（具体的には索引語、および、分野分類コード）を自動付与する手法を研究開発する。

アプローチ：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、大規模言語モデルを微調整（fine-tuning、以下「生成モデル」という）して精度向上を目指す。

得られた成果：メタデータの種類ごとに手法と評価結果について示す。詳細は下記文献に記載。

主索引語（JSTの用語辞書、約21万概念）：過去索引データで微調整した生成モデルにより索引語の候補を生成し、辞書を用いて表記揺れの整理と用語の絞り込みを行う（図1）。主索引語のうち文献の中心的な概念を表すシソーラス語[*注1]の評価結果（正解に対するF1値）を表1に示す。微調整により大幅に評価値が向上し、人手[*注2]の90%以上を達成していることがわかる。

副索引ラベル（医薬系分野）：原則としてラベルごとにエンコーダモデルによる二値分類を行う。正例の希少な一部ラベルについては生成モデルに索引付与基準を指示文として与え付与の可否を出力（生成）させる。利用頻度の高い17個のラベルについてF1値で人手の80 %以上を達成した。

(JST)分野分類コード：分野分類コード体系が大規模かつ階層構造を持つことから、エンコーダモデルによるテキスト分類をこのような分類カテゴリ構造に拡張した手法(Zhang+2021)を利用した。階層ごとのF1値（人手のF1値の相対値）表2に示す。

＜ABCiの利用について＞上記の各処理における、LLMのファインチューニング、予備実験等に活用し、研究開発の加速に大きく寄与しました。

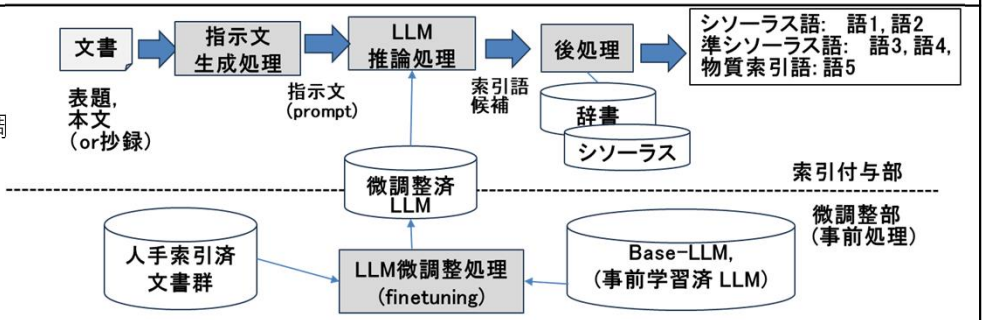

図１：主索引語の自動付与の全体構成

表1: 主索引自動付与の評価結果

	微調整前	微調整後
医薬系	0.533	0.909
非医薬系	0.628	0.996

数値は人手のF1値を１とした場合のF1値の相対値

表2: 分類コード自動付与の評価結果

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

数値は人手のF1値を１としたF1値の相対値

*注1: JSTでは索引語をシソーラス語、準シソーラス語、物質索引語の３種類に分けている。詳しくは文献1)2)を参照
*注2：人手で付与した正解とは独立に、別の専門家によって付与したデータがどの程度正解と一致しているかをF1尺度により評価した値

[Zhang+2021] Zhang et al. , “MATCH: Metadata-Aware Text Classification in A Large Hierarchy”, arXiv: 2102.07349v2.

成果についてより詳細な情報を提供しているWebページ、発表論文などの情報：

１）菊井玄一郎、鈴木慶二、関根基樹、水田寿雄：ＬＬＭを利用した索引語の自動推定、言語処理学会第32回年次大会、pp. 794-799, 2026.
２）菊井玄一郎、鈴木慶二、関根基樹、水田寿雄：科学技術文献に対するメタデータ付与の自動化、Japio YEAR BOOK, 2026年版（投稿予定）