変数定義抽出におけるテンプレート文を活用したデータ拡張法

永山 航太郎　加藤 祥太　加納 学
京都大学大学院情報学研究科
nagayama.kotaro.63c@st.kyoto-u.ac.jp
{shota, manabu}@human.sys.i.kyoto-u.ac.jp

概要

科学技術論文における変数定義の抽出は，論文の理解や活用に欠かせない．しかし，分野によって変数定義の長さや構成する単語は異なるため，既存の変数定義抽出手法の性能は分野間で差がある．各分野の学習データを用意することは性能向上に効果的だが，高品質な学習データの作成コストは高い．この課題を解決するため，本研究では，変数を定義するテンプレート文と学習データ中の変数-変数定義ペアから新たな定義文を生成する手法を提案する．化学プロセス関連論文からの変数定義抽出において，提案手法で生成した定義文で学習した定義抽出モデルは，山本ら [1] のモデルを上回る正解率89.6%を達成した．

1　はじめに

近年は世界の論文数の増加率が大きくなっている．2020 年は自然科学分野で 190 万報が発表されたが，2021 年には 9.2%増加し 205 万報以上が発表された [2]．これに伴い，文献調査に必要な時間と労力は年々増加している．この負担を軽減するべく，膨大な文献から重要な情報を自動的に抽出，整理，活用する技術の開発が進められている [3]．このような技術開発の恩恵が大きい分野の一つとして，プロセス産業が挙げられる．プロセス産業では，装置設計や運転条件の最適化などの様々な場面で科学原理に基づく物理モデルが活用されている．しかし，対象プロセスの挙動を正確に再現できる物理モデルを構築するには，膨大な量の文献を調査し，モデル構築に必要な情報を統合する必要がある．この多大な労力を要する作業を効率化するために，我々は物理モデル自動構築システム（Automated physical model builder; AutoPMoB）の実現を目指している [4]．AutoPMoB は，（1）文献データベースからの対象プロセス関連文書の収集，（2）収集した文書の形式統一，（3）物理モデル構築に関する情報（数式，変数，実験データなど）の抽出，（4）複数文書から抽出した情報の表記統一，（5）表記統一した情報の統合，を自動で行い所望の物理モデルを構築する．AutoPMoB を実現するには，文献からの正確な変数定義抽出手法が必要である．本研究では化学プロセス関連論文を対象とした変数定義抽出に取り組む．

化学プロセス関連論文からの変数定義抽出において，山本らは正解率 85.6%を達成した [1]．山本らが学習に用いた化学プロセス関連論文データセットは関連研究 [5, 6] で用いられたデータセットのサイズと比較して小さい．そのため，学習データを増やすことで定義抽出性能が向上する可能性がある．ただし，変数定義を構成する単語やその長さは分野によって異なり，変数定義抽出性能も分野によって差がある [7]．そのため，他分野の定義文を増やしても性能向上に寄与する可能性は低い．さらに，適用対象の分野に応じて新たに学習データを作成するには高いコストが伴う．

このような問題を解決するために，本研究では，図 1 に示すような学習データ拡張手法を提案する．提案手法は，“[symbol] is defined as [primary].”のように変数 [symbol] と対応する定義 [primary] を含むテンプレート文を複数用意し，それに学習データ中の変数と変数定義を代入することで新たな定義文を生成する．化学プロセス関連の論文 47 報から作成したデータセットを対象に，山本ら [1] のモデルと提案手法により生成した定義文を用いて学習したモデルの性能を比較する．

2　関連研究

変数定義抽出タスクでは，パターンに合致する名詞句を抽出するルールベース手法 [8] や品詞タグや文中の位置といった特徴を用いた機械学習に基づく手法 [9, 10]，深層学習モデルを用いた手法 [11] など

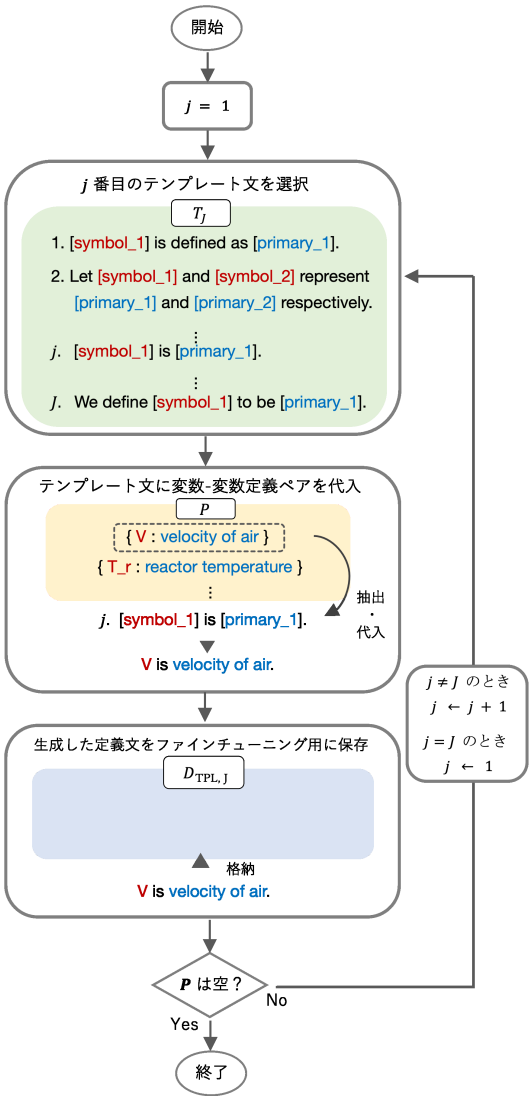

図 1 提案手法の概略図．T_J は J 種類のテンプレート文の集合，P は学習データに含まれる変数-変数定義ペアを格納したリスト，$D_{\mathrm{TPL},J}$ は生成した文を格納するためのリストを表す．

が提案されている．その中で，BERT（Bidirectional Encoder Representations from Transformers）[12] に代表される事前学習モデルを用いた手法は特に高い性能を達成している．Kang らは，SciBERT [6] によって文から専門用語とその定義を同時に抽出する手法を提案し，ACL Anthology [13, 14] に含まれる 50 報の論文からの定義抽出において F1-score 70.8%を達成した [15]．Lee らは，変数と変数定義の対応付けタスクである SemEval 2022 Task 12: Symlink [5] にて，単純なルールに基づく記号用のトークナイザと SciBERT を用いて，変数と変数定義の固有表現認識とそれらの関係抽出を順に実行する方法を提案し，最高性能を達成した [16]．Popovic らは，Symlink に

- Let's define [primary_1] as [symbol_1].
- Let [symbol_1] and [symbol_2] represent [primary_1] and [primary_2] respectively.
- Our system includes [primary_1] [symbol_1], [primary_2] [symbol_2], and [primary_3] [symbol_3].
- We will work with five variables: [symbol_1], [symbol_2], [symbol_3], [symbol_4], and [symbol_5], each representing [primary_1], [primary_2], [primary_3], [primary_4], and [primary_5] respectively.

図 2 テンプレート文の例．テンプレート文には変数-変数定義ペアを 1 つ含むものから 6 つ含むものが存在する．

て，SciBERT を用いて固有表現認識と関係抽出を同時に行う手法を提案し，第 3 位の性能を達成した [7]．山本らは，定義抽出対象の変数が既知であるという条件の下での変数定義抽出タスクを対象とし，変数を特殊トークンに置換した文を BERT モデルに入力することで文中の変数定義の位置を予測する方法を提案した [1]．彼らは，化学プロセスとの関連性が低い 5 分野からなる Symlink データセット [5] と化学プロセス関連論文データセットを順に用いて二段階のファインチューニングを行い定義抽出モデルを構築した．彼らの手法は，化学プロセス関連論文からの変数定義抽出において正解率 85.5%，F1-score 81.6%を達成した．しかし，いずれの手法も AutoPMoB の要素技術としては十分な性能に達していない．

3 提案手法

提案手法の概略図を図 1 に示す．本手法は，変数-変数定義ペアと図 2 のようなテンプレート文を用いて，新たな定義文を生成する．

事前準備として，学習データ中の変数-変数定義ペアを格納したリスト P と，生成した文を格納するための空のリスト $D_{\mathrm{TPL},J}$ を作成する．加えて，J 種類のテンプレート文の集合 T_J を用意し，j 番目のテンプレート文を TPL_j とする．TPL_j は n_j 組の [symbol] と [primary] を含み，P に含まれるペアの数は J よりも十分大きいとする．テンプレート文の作成には ChatGPT（gpt-3.5-turbo-1106）を使用する．プロンプトには出力形式の説明，生成する文中の変数の数の指定，定義文の例が含まれるように設計する．

定義文を生成する手順を以下に示す．

1. $j = 1$ とする．
2. n_j 個の変数-変数定義ペアを P から選択する．

表 1 D_{Process} に含まれる論文数と定義を持つ変数の数

データセット	論文数	定義を持つ変数の数
D_{CRYST}	11	299
D_{CSTR}	10	169
D_{BD}	10	210
D_{CZ}	9	313
D_{STHE}	7	285
D_{Process}	47	1,276

選択したペアは P から削除する．

3. 選択した n_j 個の変数-変数定義ペアを TPL_j 中の [symbol] と [primary] に代入し，定義文を 1 つ生成する．生成した文を $D_{\text{TPL},J}$ に追加する．
4. j に $j+1$ を代入して手順 2 に戻る．ただし，$j = J$ のときは j に 1 を代入して手順 2 に戻る．P が空になった時点で終了する．

4 実験

4.1 タスク設定とデータセット

タスク設定 本研究では次の 2 つの条件を満たす変数の定義を抽出するタスクに取り組む．

1. 文中にその変数に対応する定義が存在する．
2. スペースで囲まれているか数式の左辺にある．

例えば，

> The pressure $P = nRT/V$ is assumed constant, where T and V are the temperature and volume, respectively.

において，P, T, V が定義抽出対象であり，n と R は対象外である．定義抽出対象の変数は既知とする．

化学プロセス関連論文データセット 化学プロセスに関連する計 47 報の論文に登場する変数に対して定義を付与したデータセット D_{Process} を作成した．D_{Process} には晶析プロセス（crystallization process; CRYST），連続槽型反応器（continuous stirred tank reactor; CSTR），バイオディーゼル生産プロセス（biodiesel production process; BD），チョクラルスキープロセス（Czochralski process; CZ），多管式熱交換器（shell and tube heat exchanger; STHE）の 5 つのプロセスに関する論文が含まれる．各プロセスの論文数と定義を持つ変数の数を表 1 に示す．

Symlink データセット Symlink で用いられたデータセット D_{Symlink} は情報科学，生物学，物理学，数学，経済学の 5 分野の合計 101 報の論文からなり，定義を持つ変数は 11,462 個含まれる．

データセットの分割割合 D_{Process} を論文単位で訓練用，検証用，テスト用に分割する．STHE で 2 報，それ以外のプロセスで 3 報をテスト用とする．さらに各プロセスにおいて 1 報を検証用，残り全てを訓練用とする．D_{Symlink} と $D_{\text{TPL},J}$ は訓練用と検証用に 3:1 に分割する．訓練用と検証用のデータセットを用いてモデルをファインチューニングし，テスト用データセットを用いて性能を評価する．

表 2 T_J 中の含まれる変数の数ごとのテンプレート文の数

データセット	テンプレート文に含まれる変数の数					
	1 個	2 個	3 個	4 個	5 個	6 個
T_{20}	5	5	3	3	2	2
T_{100}	24	24	13	13	13	13
T_{600}	140	140	80	80	80	80

4.2 実験設定

山本ら [1] と同じく，ベースモデルには DeBERTa-V3$_{\text{LARGE}}$ [17]，オプティマイザには Adam [18]，GPU には A100 を用いる．バッチサイズは 8，学習係数は 1e-5 とする．また，山本らの二段階ファインチューニングと同様に，D_{symlink}，$D_{\text{TPL},J}$，D_{Process} を順に用いて三段階でモデルをファインチューニングする．エポック数は 3 とし，検証用データに対する交差エントロピー誤差が最小のモデルで性能評価を行う．

テンプレート文の数の影響 テンプレート文の数が定義抽出性能に与える影響を検証するため，テンプレート文の数が異なる 3 つのテンプレート文の集合 T_{20}, T_{100}, T_{600} と対応するデータセット $D_{\text{TPL},20}$，$D_{\text{TPL},100}$，$D_{\text{TPL},600}$ を作成する．T_{100} は，変数の数の分布が均等になるように，T_{600} からランダムに選択し，T_{20} も同様に T_{100} から選択する．T_{20}, T_{100}, T_{600} における変数の数とテンプレート文の数の関係を表 2 に示す．

未知プロセスへの性能検証 提案手法を実際に用いる際には D_{Process} に含まれる 5 つのプロセス以外のプロセスに関する論文も対象にする．そのような場合に提案手法が有効かを検証するために，D_{Process} からプロセス X に関するデータセット D_X を取り除いたデータセット $D_{\text{Process}-X}$ を訓練と検証に用いて，D_X に対する定義抽出性能を検証する．

4.3 評価方法

評価指標として，全ての変数のうち正しい定義の抽出に成功した変数の割合（正解率）を用いる．定義抽出成功の基準として，正解の定義の範囲とモデ

ルの予測した定義の範囲が完全に一致した場合を正解とする基準（full）と，正解の定義の範囲と予測した定義の範囲に重複があれば正解とする基準（partial）の 2 つを採用する．D_{Process} の訓練・検証・テスト用データに含まれる論文の種類によって性能が変動することが予想されるため，4.1 節に従って D_{Process} を 10 通りに分割し，それらから得られる評価指標の平均を比較する．

5 結果と考察

提案手法と山本らの手法による定義抽出結果を表 3 に示す．正解率（full）と正解率（partial）のいずれにおいても提案手法が山本らの手法を上回り，正解率（full）は最高で 2.8 ポイント向上した．また，全ての場合で正解率（partial）は 96%を超え，提案手法は 97%を超えた．いずれの手法でも文中での定義の位置は概ね判定できたが，より正確な位置の判定に提案手法が有効であることが示された．

5.1 テンプレート文の数の影響

表 3 に示すように，テンプレート文の数 J を増やすほど正解率（full）が向上した．また，提案手法（$J = 20$）では定義抽出に失敗したが提案手法（$J = 600$）では成功した事例の中には，T_{600} に含まれないパターンの定義文も存在した．このことから，J の増加はテンプレート文にないパターンの定義文からの定義抽出性能の向上にも寄与していると考えられる．

5.2 未知プロセスへの性能

提案手法と山本らの手法による未知プロセスへの性能検証結果を表 4 に示す．5 つ全てのプロセスにおいて，学習データに定義抽出対象プロセスのデータが含まれるか否かに関わらず，提案手法により生成した定義文で学習したモデルが山本らの手法によるモデルを上回った．提案手法は，学習データに含まれないプロセスに対する変数定義抽出においても有効であることが示された．また，CZ を除く 4 つのプロセスにおいて，$D_{\text{Process}-X}$ を学習データとした提案手法が，D_{Process} を学習データとした既存手法よりも高い正解率（full）を達成した．CZ では，登場する変数定義や定義文のパターンが他の 4 つのプロセスと大きく異なり，学習データから D_{CZ} を削除することによる性能低下が最も大きい．以上のことから，D_{Process} に含まれる 5 つのプロセスと類似したプロセスに関しては，D_{Process} に提案手法を適用することで，そのプロセスのデータを用意せずとも既存手法と同等以上の性能を達成することが期待できる．

表 3 実験結果．太字は各列における最高値を表す．D_{symlink}，$D_{\text{TPL},J'}$，D_{Process} による三段階ファインチューニングを提案手法（$J = J'$）と表記している．

手法	正解率（full）	正解率（partial）
既存手法 [1]	86.8	96.7
提案手法（$J = 20$）	88.3	**97.4**
提案手法（$J = 100$）	89.0	97.2
提案手法（$J = 600$）	**89.6**	**97.4**

表 4 未知プロセスへの性能．各数値は正解率（full）を表す．提案手法は $J = 600$ としている．

学習用	テスト用	既存手法 [1]	提案手法
$D_{\text{Process}-\text{BD}}$	D_{BD}	78.9	80.8
D_{Process}		80.1	82.8
$D_{\text{Process}-\text{CSTR}}$	D_{CSTR}	87.1	89.8
D_{Process}		88.0	90.7
$D_{\text{Process}-\text{CRYST}}$	D_{CRYST}	79.9	83.9
D_{Process}		83.7	87.6
$D_{\text{Process}-\text{STHE}}$	D_{STHE}	89.1	92.4
D_{Process}		91.6	94.0
$D_{\text{Process}-\text{CZ}}$	D_{CZ}	81.3	84.4
D_{Process}		85.1	88.2

6 おわりに

本研究では変数定義抽出における学習データ不足を解決するため，定義文のテンプレート文を複数用意し，各文に学習データ中の変数-変数定義ペアを代入することで学習データを拡張する手法を提案した．既存手法と提案手法を比較した結果，提案手法は正解率（full）において 2.8 ポイント高い 89.6%を達成した．また，テンプレート文の数を増やすほど正解率が向上した．今後は，既存のデータに含まれるプロセスと適用対象のプロセスとの類似度を算出して，提案手法が有効かを事前に判断できるようにする．また，本研究では文中に定義が存在する変数のみを対象としたが，同一文に定義が含まれない変数も存在する．そのような変数に対して定義が存在しないと判断できるように提案手法を拡張する．

謝辞

本研究は，JSPS 科研費 JP23K13595 の支援を受けたものである．

参考文献

[1] 山本蒔志, 加藤祥太, 加納学. 二段階のファインチューニングを行った BERT による変数定義抽出. 言語処理学会第 29 回年次大会発表論文集, pp. 2957–2961, 2023.

[2] 文部科学省科学技術・学術政策研究所. 科学技術指標 2023. 調査資料 328, 2023.

[3] Elsa A. Olivetti, Jacqueline M. Cole, Eun Kim, Olga Kononova, Gerbrand Ceder, Tzu-Ying J. Han, and Anna M. Hiszpanski. Data-driven materials research enabled by natural language processing and information extraction. **Applied Physics Reviews**, Vol. 7, No. 4, 041317, 2020.

[4] Shota Kato and Manabu Kano. Towards an automated physical model builder: CSTR case study. **Computer Aided Chemical Engineering**, Vol. 49, pp. 1669–1674, 2022.

[5] Viet Lai, Amir Pouran Ben Veyseh, Franck Dernoncourt, and Thien Nguyen. Semeval 2022 task 12: Symlink - linking mathematical symbols to their descriptions. In **Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)**, pp. 1671–1678, 2022.

[6] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 3615–3620, 2019.

[7] Nicholas Popovic, Walter Laurito, and Michael Färber. AIFB-WebScience at SemEval-2022 task 12: Relation extraction first - using relation extraction to identify entities. In **Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)**, pp. 1687–1694, 2022.

[8] Robert Pagel and Moritz Schubotz. Mathematical language processing project. In **Joint Proceedings of the MathUI, OpenMath and ThEdu Workshops and Work in Progress track at CICM**, 2014.

[9] Moritz Schubotz, Leonard Krämer, Norman Meuschke, Felix Hamborg, and Bela Gipp. Evaluating and improving the extraction of mathematical identifier definitions. In **Experimental IR Meets Multilinguality, Multimodality, and Interaction**, Vol. 10456 LNCS of **CLEF 2017**, pp. 82–94, 2017.

[10] Jason Lin, Xing Wang, Zelun Wang, Donald Beyette, and Jyh-Charn Liu. Prediction of mathematical expression declarations based on spatial, semantic, and syntactic analysis. In **Proceedings of the ACM Symposium on Document Engineering 2019**, DocEng '19, pp. 1–10, 2019.

[11] Yiannos Stathopoulos, Simon Baker, Marek Rei, and Simone Teufel. Variable typing: Assigning meaning to variables in mathematical text. In **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)**, pp. 303–312, 2018.

[12] Jacob Devin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pretraining of deep bidirectional transformers for language understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, 2019.

[13] Steven Bird, Robert Dale, Bonnie Dorr, Bryan Gibson, Mark Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir Radev, and Yee Fan Tan. The ACL Anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In **Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)**, 2008.

[14] ACL Anthology team. Acl anthology. https://aclanthology.org/. (Accessed on 12/29/2023).

[15] Dongyeop Kang, Andrew Head, Risham Sidhu, Kyle Lo, Daniel S Weld, and Marti A Hearst. Document-level definition detection in scholarly documents: Existing models, error analyses, and future directions. **arXiv preprint arXiv:2010.05129**, 2020.

[16] Sung-Min Lee and Seung-Hoon Na. JBNU-CCLab at SemEval-2022 task 12: Machine reading comprehension and span pair classification for linking mathematical symbols to their descriptions. In **Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)**, pp. 1679–1686, 2022.

[17] Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving DeBERTa using ELECTRA-style pretraining with gradient-disentangled embedding sharing. **arXiv preprint arXiv:2111.09543**, 2021.

[18] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. **arXiv preprint arXiv:1412.6980**, 2014.