化学工学分野の論文に含まれる命名法に基づく変数記号および定義の解析

加藤 祥太　加納 学
京都大学大学院情報学研究科
{shota, manabu}@human.sys.i.kyoto-u.ac.jp

概要

プロセス産業においてデジタルツインを実現するために対象プロセスの挙動を正確に模倣できる物理モデルが必要であるが，精緻な物理モデルの構築には膨大な労力を要する．著者らは，物理モデル構築工程を効率化するために，物理モデルを自動で構築する人工知能（AutoPMoB) の実現を目指している．AutoPMoB の実現には文献からの変数定義抽出技術の開発が必要である．本研究ではその一環として，化学工学分野の 42,323 報の論文に含まれる命名法から 549,840 個の変数記号–変数定義ペアを抽出し，解析した．本解析により，化学工学分野では変数記号に ρ, μ, α，変数定義には Reynolds number が最もよく用いられることと，変数定義の単語数が他分野と化学工学分野とでは異なることを明らかにした．

1　はじめに

プロセス産業のデジタルトランスフォーメーション（DX）の実現に向けた最も重要な技術であるデジタルツインを実現するためには，対象プロセスの挙動を正確に再現できる物理モデルが不可欠である．しかし，高精度な物理モデルを構築するには，対象プロセスに関する専門知識を有する技術者や研究者が膨大な文献を調査し，モデル構築に必要な情報を収集し，それらを組み合わせて試行錯誤的に精度向上を行うことが必要とされている．著者らは，この物理モデル構築工程を効率化するために物理モデル自動構築 AI（Automated physical model builder; AutoPMoB）の実現を目指している [1].

物理モデルの表記では数式が重要な役割を果たしており，数式を理解するためには数式を構成する変数の意味を正しく認識する必要がある．しかし，論文中の変数の意味を正しく把握するためには，文中から変数定義を抽出したり別の情報源をもとに

Nomenclature

A_n	area of nozzle, m^2	v^*	friction velocity, m/s
C_A	concentration of CO_2 in solution, mol/m^3	*Greek letters*	
C_{AG}	concentration of CO_2 in gas phase, mol/m^3	Φ	empirical constant

図 1　Nomenclature の例 [2]

定義を予測したりする必要がある．論文には図 1 に示すような Nomenclature が記載されていることがある．Nomenclature とは用語に対する定義の一覧のことであり，一般に命名法と訳される．Nomenclature が記載されている場合には変数記号に対する変数定義の意味を容易かつ正確に抽出できる．本研究は，AutoPMoB の実現に必要な要素技術の 1 つである変数定義抽出手法の開発の一環として，この Nomenclature を複数の論文から抽出して解析する．

2　関連研究

変数定義抽出を目的とした最近の自然言語処理タスクに SymeEval2022 Task 12: Symlink がある [3]. Symlink タスクでは，論文中に存在する変数記号と変数定義を抽出し，さらにそれらの関係も抽出することを目指す．Symlink データセットでは，計算機科学・生物学・物理学・数学・経済学の 5 つの分野の TeX 形式の論文に含まれる変数記号と変数定義が対応付けられている．Symlink データセットには 102 報の論文が用いられ，21,915 個の変数記号に対して 9,556 個の変数定義がアノテーションされた．同様に変数記号に対する変数定義を抽出する手法の開発を目的としたタスクとして，NTCIR-10 の Math understanding subtask がある [4]．このタスクでは 45 報の論文に含まれる 4,323 個の変数記号に変数定義が付与された [5]．分野によって変数定義の記述の仕方は異なるため，我々の最終目的である AutoPMoB の実現に有力な手法を開発するには化学工学関連の文献に含まれる変数記号と変数定義を含

むデータセットが必須である．上述の既存研究ではいずれも arXiv [6] 上の論文を対象としているが，化学工学関連の論文が arXiv に投稿されることは稀である．そのため，これまでに化学工学関連の論文を対象としたデータセットは存在しない．

3　データセット

Elsevier Research Product APIs [7] を用いて，110 の化学工学関連論文誌から XML（Extensible Markup Language）形式の論文を収集した．Elsevier 社が公開している XML 形式の論文では，現在，2001 年と 2002 年に開発された XML DTD (Document Type Definitions)　5 が用いられている [2]．API で収集した論文を確認したところ，2003 年以前の論文には XML DTD 5 に基づくものとそれ以前の XML DTD に基づくものが存在していた．そこで，本研究では 2004 年以降の論文を対象とした．XML DTD 5 において，Nomenclature は `ce:nomenclature` 要素で表され，文書で使用される用語と定義のリストを含む [2]．用語と定義はそれぞれ `def-term` 要素と `def-dsc` 要素で表される [2]．Nomenclature の中には略語（Abbreviation）が含まれることもあるが，本研究では変数のみを対象とする．本稿では，数学的表現を表す `mml:math` 要素が 1 つのみ含まれる `def-term` 要素を変数記号，それと対応する `def-dsc` 要素を変数定義とみなす．2004 年 1 月から 2022 年 3 月までに出版された，変数を含む Nomenclature が存在する論文を収集したところ，42,323 報を得た．さらに各論文の Nomenclature から合計 549,840 個の変数記号-変数定義ペアを抽出した．

4　解析結果

4.1　変数

論文に含まれる変数記号-変数定義ペアの数の度数分布を図 2 に示す．図ではペアの数の最大値を 100 としたが，1 つの論文に含まれる変数記号-変数定義ペアの数の最大値は 168，最小値は 1, 平均値は 13，中央値は 5 だった．図 2 において，変数の数が 1, 2, 3 個の論文は 8,173, 5,293, 3,463 報であった．これらの論文のうち多くは，変数を `mml:math` 要素ではなく，`ce:italic` 要素（斜体表記）としていた．そのため，図 2 に示す変数の数は実際に論文で用いられているよりも少ない．`ce:italic` 要素の変数は他の専門用語や略語との区別が難しいと判断したため，今回の解析では対象としなかった．

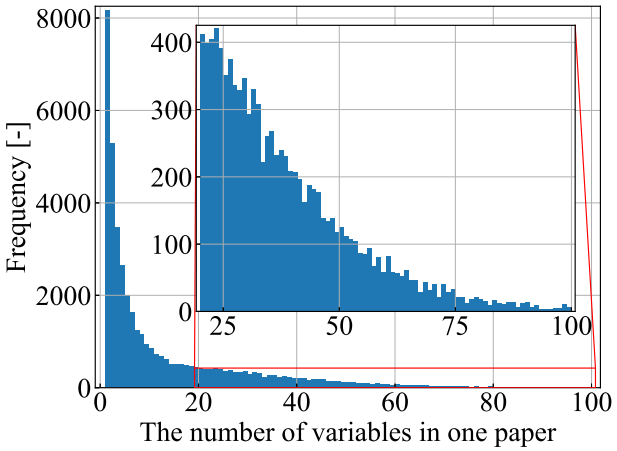

図 2　論文に含まれる変数の数の度数分布

表 1　変数記号の頻度上位 30 個

Symbol	Freq.	Symbol	Freq.	Symbol	Freq.
ρ	7,752	β	3,337	ϕ	2,648
μ	5,281	m'	3,235	g	2,638
α	4,816	h	3,119	v	2,520
T	4,809	L	3,035	A	2,519
ε	3,804	θ	3,003	D	2,458
λ	3,750	η	2,915	f	2,445
σ	3,588	τ	2,806	ν	2,434
t	3,460	R	2,806	γ	2,343
$\dot{m}$	3,370	p	2,734	δ	2,315
k	3,340	P	2,686	n	2,246

4.2　変数記号

変数記号は 143,722 種類であった．変数記号の頻度の上位 30 個を表 1 にまとめる．頻度が最大の変数記号は ρ であり，その数は 7,752 報 (5.4%) であった．

論文間で変数定義が異なるかを確認するため，上位 3 つの変数記号 ρ, μ, α のそれぞれについて，その定義の主辞の頻度（Frequency; Freq.）を表 2 に示す．本稿では，主辞の抽出に ScispaCy [8] を用いた．ρ と μ ではそれぞれ 94.8%, 83.1%の定義の主辞が density, viscosity であった．一方，α の定義の主辞は 19.3%が diffusivity，11.1%が fraction，10.5%が coefficient であった．表 1 に示した各変数について，最も頻度が高い定義が全定義に占める割合を計算したところ，最小値は ϕ の 17%，最大値は ρ の 86%であり，この割合が 50%以上である変数は 9 個だった．

化学工学関連の論文では，T_{R} や T^{e} のように，変数記号に下付き文字や上付き文字などの修飾が付与されることが多い．本稿では，このように修飾さ

表 2　変数記号に対する定義の主辞の頻度

ρ		μ		α	
Head	Freq.	Head	Freq.	Head	Freq.
density	6,668	viscosity	3,504	diffusivity	928
ratio	46	coefficient	184	fraction	535
resistivity	37	ratio	79	coefficient	504
reflectivity	35	potential	62	angle	484
coefficient	25	parameter	43	transfer	196
Total	7,752	Total	5,281	Total	4,816

表 3　主変数記号の頻度上位 30 個

Symbol	Freq.	Symbol	Freq.	Symbol	Freq.
T	20,176	u	10,749	σ	8,998
C	19,045	V	10,722	f	8,990
m	15,620	q	9,863	α	8,714
ρ	14,474	h	9,576	v	8,164
P	13,617	c	9,546	t	8,115
Q	12,084	F	9,462	x	7,992
R	11,758	A	9,423	E	7,927
k	11,441	S	9,378	N	7,815
δ	11,097	μ	9,203	L	7,758
D	10,769	ε	9,128	p	7,417

れる変数記号を主変数記号と呼び，数学的表現を表す mml:math 要素に含まれる最初の mml:mi 要素を主変数記号として抽出した．主変数記号の頻度の上位 30 個を表 3 に示す．表 1 と表 3 を比べると，C や Q などの変数は単独で用いられることが少ないが修飾を付与して用いられる割合が他の変数よりも高いことがわかる．

主変数記号についても表 2 と同様に，3 つの主変数記号 T, C, m に対応する変数定義の主辞の頻度を表 4 に示す．主変数記号も変数記号と同様に，最も頻度が高い定義が全体に占める割合が変数記号の種類によって 30 ポイント程度異なる．以上の結果より，変数記号に対応する変数定義を正確に予測するには，変数記号のみに着目するのではなく，文献のその他の情報を参照する必要があるといえる．

4.3　変数定義

変数定義を構成する単語数の度数分布を図 3 に示す．一部の変数定義には ce:math 要素で表される数学的表現も含まれていたが，今回は ce:math 要素 1 つを 1 単語として数えた．図 3 より，単語数が 3 の変数定義が最も多い 111,732 個（20.3%）であった．また，変数定義の単語数の最大値は 79，平均値は 5，中央値は 4，であった．Lai らが作成した Symlink データセットでは，変数定義の長さは最大 47 で，1–3 の長さの変数定義が大部分を占めていた [3]．化

表 4　主変数記号に対する定義の主辞の頻度

T		C		m	
Head	Freq.	Head	Freq.	Head	Freq.
temperature	11,792	concentration	3,772	rate	6,386
time	292	coefficient	2,654	mass	1,451
period	262	capacity	1,363	flow	562
torque	139	heat	1,103	flux	522
value	124	cost	981	number	241
Total	20,176	Total	19,045	Total	15,620

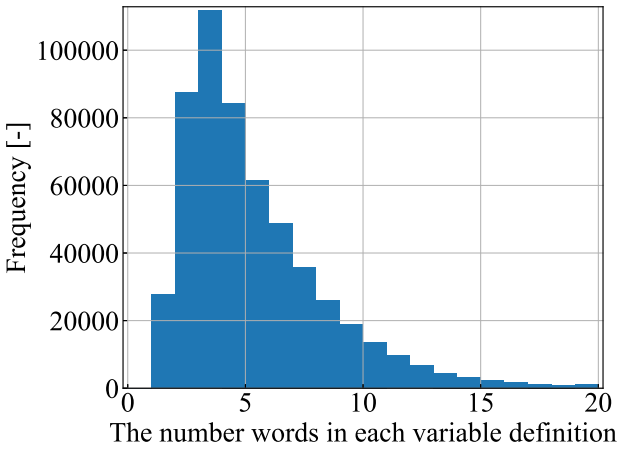

図 3　変数定義に含まれる単語数の度数分布

学工学分野の論文では，Lai らが用いた 5 つの分野よりも変数定義の長さが長く，分野によって長さに差があることが明らかになった．

変数定義と変数定義の主辞の頻度の上位 30 個をそれぞれ表 5 と表 6 に示す．表 5 に示すように，今回対象とした変数定義のいくつかには単位が含まれる．変数定義に単位が含まれるために，temperature (K) と temperature, K のように，単位表記の仕方は異なるが意味が同じ変数定義を複数確認できる．これは数学や情報学の分野に無い工学分野に特有の傾向である．表 6 において，number, rate, coefficient の順に主辞の頻度が高い．表 5 を見ると，number は数の表記に加えて，レイノルズ数（Reynolds number）やプラントル数（Prandtl number）のような専門用語の表記に用いられるために頻度が高いことがわかる．

頻度が高い順に 3 つの変数定義と変数定義の主辞に対する主変数記号を表 7 と表 8 に示す．表 7 より，変数記号に対する変数定義の場合と比べて，変数定義に対する変数記号の使い方は多くの論文で同じであることがわかる．一方で，表 8 が示すように，変数定義の主辞に対する変数記号の頻度はいずれも 30%以下であった．したがって，変数定義から変数記号を予測する場合，本稿で作成したデータ

表 5　変数定義の頻度上位 30 個

Definition	Freq.	Definition	Freq.
Reynolds number	1,125	time	412
mass flow rate (kg/s)	1,049	temperature, K	410
Prandtl number	983	dynamic viscosity	403
density	822	velocity vector	387
Nusselt number	785	dimensionless temperature	371
density (kg/m3)	642	pressure	360
mass flow rate, kg/s	622	mass flow rate [kg/s]	357
density, kg/m3	487	average Nusselt number	346
temperature (K)	479	inlet	346
porosity	478	wall	318
temperature	476	heat flux	317
mass flow rate	472	thermal diffusivity	313
thermal conductivity	441	time, s	311
time (s)	441	volume fraction	290
kinematic viscosity	431	outlet	289

表 6　変数定義の主辞の頻度上位 30 個

Head	Freq.	Head	Freq.	Head	Freq.
number	22,729	fraction	7,448	area	5,708
rate	22,614	concentration	7,207	time	5,689
coefficient	17,965	flux	7,142	length	5,651
temperature	14,451	factor	7,127	angle	5,045
density	13,573	ratio	6,830	value	4,897
velocity	12,852	parameter	6,504	volume	4,327
vector	9,010	diameter	6,075	transfer	4,293
viscosity	8,458	function	6,018	capacity	4,223
pressure	8,282	heat	5,870	force	4,043
constant	7,634	conductivity	5,717	efficiency	3,936

セットをそのままが活用できる可能性がある．

5　おわりに

本研究では，変数定義抽出手法開発のために，化学工学関連論文に含まれる 549,840 個の変数記号と変数定義のペアを解析した．解析の結果，変数記号に対応する変数定義を正しく抽出するには変数記号と変数定義以外の文献の情報が必要であることが示唆された．今後は特徴量作成や言語モデルのチューニングに本データセットを活用し，正確に変数定義を抽出できる手法の開発に取り組む．

謝辞

本研究は JSPS 科研費 JP21K18849 の助成を受けたものです．

参考文献

表 7　変数定義に対する主記号の頻度

Reynolds number		mass flow rate (kg/s)		Prandtl number	
Symbol	Freq.	Symbol	Freq.	Symbol	Freq.
Re	610	m	964	Pr	606
R	328	M	27	P	239
N	6	q	2	σ	15
ρ	5	G	2	ν	6
d	4	W	2	μ	5
Total	1,125	Total	1,049	Total	983

表 8　変数定義の主辞に対する主記号の頻度

number		rate		coefficient	
Symbol	Freq.	Symbol	Freq.	Symbol	Freq.
N	4,056	m	6,386	C	2,654
Nu	2,000	Q	3,318	D	2,251
Re	1,879	V	930	β	1,469
n	1,668	q	928	α	1,021
R	1,387	ε	811	k	907
Total	22,729	Total	22,614	Total	17,965

Symposium on Process Systems Engineering, Vol. 49 of **Computer Aided Chemical Engineering**, pp. 1669–1674. Elsevier, 2022.

[2] Tag by Tag, The Elsevier DTD 5 Family of XML DTDs Version 1.9.5.9. https://www.elsevier.com/__data/assets/pdf_file/0003/58872/ja5_tagbytag5_v1.9.5.pdf, 2016. (Accessed on 01/03/2023).

[3] 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, Seattle, United States, July 2022. Association for Computational Linguistics.

[4] Akiko Aizawa, Michael Kohlhase, and Iadh Ounis. NTCIR-10 math pilot task overview. In **Proceedings of the 10th NTCIR Conference**, pp. 654–661, 2013.

[5] Giovanni Yoko Kristianto, Goran Topić, and Akiko Aizawa. Extracting textual descriptions of mathematical expressions in scientific papers. In **D-Lib Magazine**, Vol. 20. Corporation for National Research Initiatives, 2014.

[6] arxiv.org e-print archive. https://arxiv.org/. (Accessed on 01/13/2023).

[7] Elsevier developer portal. https://dev.elsevier.com/. (Accessed on 01/12/2023).

[8] Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In **Proceedings of the 18th BioNLP Workshop and Shared Task**, pp. 319–327, Florence, Italy, August 2019. Association for Computational Linguistics.