変数の記号と定義に関する情報を活用した変数定義抽出手法

沼本 真幸　加藤 祥太　加納 学
京都大学大学院情報学研究科
{numoto.masaki.32z, katou.shouta.23v}@st.kyoto-u.ac.jp
manabu@human.sys.i.kyoto-u.ac.jp

概要

プロセス産業では様々な問題解決にプロセスの物理モデルが活用されている．物理モデルを構築するためには，複数の文献に含まれる数式と変数を把握し，統合する必要があり，多大な労力を要する．我々の目標はこの工程の自動化である．本研究ではその一環として化学プロセス関連文献からの変数定義抽出手法の開発に取り組む．化学プロセス関連文献では変数の記号と定義の用い方に特徴があるが，従来手法ではこの特徴を活用できない．我々はこの点に注目し，新たな変数定義抽出手法を提案する．3 つの化学プロセスに関する論文計 28 報を対象に提案手法と従来手法の性能を検証した結果，2 つのプロセスで提案手法がより高い正解率を達成した．

1　はじめに

プロセス産業では，プロセスの設計や運転に関わる様々な問題の解決にあたって，物理モデルに基づいて検討を加えるというアプローチが広く利用されている．しかし，現実の現象を正確に表現できる物理モデルの構築は多大な労力を要するため，その効率化が望まれている．我々は，複数の文献から必要な情報を抽出，統合し，物理モデルを自動構築する人工知能 (Automated physical model builder; AutoPMoB) の開発に取り組んでいる．文献から抽出する重要な情報の一つに数式があるが，複数の文献から抽出した数式を統合するためには，数式中の変数の定義の把握が不可欠である．そこで本研究では，AutoPMoB の要素技術として，変数の定義を文献から自動的に抽出する技術の開発に取り組む．

変数定義抽出に関連する先行研究はいくつか存在する．Kristianto らは，サポートベクトルマシン (SVM) を用いて文献中の数学的表現の説明を抽出する手法を提案した [1]．Pagel と Schubotz は，文献中の数式に含まれる identifier の定義を抽出するランキングベースの手法を提案した [2]．以上の研究では，抽出対象となる定義あるいは説明の周辺の文章に関する情報の活用が主軸になっている．Stathopoulos らによる研究 [3] では，変数の記号や定義自体に関する情報を活用しているが，定義となる語彙が事前に判明していることが前提とされている．いずれの手法も，問題設定に大きな差異がある，抽出性能が不十分であるなどの理由により，AutoPMoB の要素技術としての採用が難しい．

本研究で対象とする化学プロセス関連文献の間では，変数の記号と定義の用い方に多くの共通点がある．この点に注目し，変数の記号と定義に関する情報を活用した変数定義抽出手法を提案する．また，性能の評価指標を整備し，化学プロセスに関する論文 28 報から作成したデータセットを用いて，Kristianto らによる手法 [1] と提案手法の性能を比較する．

2　問題設定

本研究では文献中に単独で存在する変数を扱い，数式中にのみ現れる変数は対象外とする．例えば，

$$x + y = z, \text{ where } x \text{ is X and } y \text{ is Y.}$$

においては，x, y が定義抽出対象であり，z は対象外である．文献中の記号列から変数を判別する方法を付録 A に示す．単一の文献中では同じ変数が異なる記号で表されることはないと仮定する．

本研究では，文献中に定義が存在する変数に対してはその定義を抽出し，それ以外の変数に対しては何も抽出しない，というタスクに取り組む．ただし，ある変数の定義が単一の文献中に複数存在する場合，そのいずれかを抽出すればよいとする．

データセット　3 つの化学プロセスに関する英語論文計 28 報（連続槽型反応器 [Continuous Stirred Tank Reactor; CSTR] 10 報，晶析プロセス [Crystallization; CRYST] 11 報，多管式熱交換器 [Shell and Tube Heat Exchanger; STHE] 7 報）を用いて，文献中に定義が

存在する変数に対してその定義をアノテーションしたデータセットを作成した．各プロセスのデータセットをそれぞれ $\mathcal{D}_{\mathrm{CSTR}}, \mathcal{D}_{\mathrm{CRYST}}, \mathcal{D}_{\mathrm{STHE}}$ と表す．データセットに含まれる変数の数は，$\mathcal{D}_{\mathrm{CSTR}}$：121 個，$\mathcal{D}_{\mathrm{CRYST}}$：205 個，$\mathcal{D}_{\mathrm{STHE}}$：215 個である．データセットは訓練用とテスト用に分割する．分割方法は 4.2 節で述べる．

3 提案手法

本手法では，ある変数 v の定義の候補を抽出したのち，抽出した候補が正しい定義かどうかを判定する．v の定義の候補となるのは，v と同じ文中に存在する全ての名詞句である．ただし，v を含む名詞句には次の処理を施す．

1. v が名詞句の末尾にのみ存在する場合，v を削除した名詞句を定義の候補とする．
2. それ以外の場合，定義の候補としない．

これは，例えば “temperature T” が T の定義の候補として抽出された時，“temperature” を新たに定義の候補とするための処理である．本研究では，文分割には Stanza [4]，名詞句抽出には Stanford Parser [5] を使用する．次に，v の定義の候補それぞれについて 3 つの特徴量 $x_1, x_2, x_3 \in [0, 1]$ を生成し，式 (1) により $score \in [0, 1]$ を計算する．

$$score = \frac{m_1 x_1 + m_2 x_2 + m_3 x_3}{m_1 + m_2 + m_3} \tag{1}$$

m_1, m_2, m_3 は特徴量の重みを表す．$score$ が最も高く，かつ閾値 t を超えている候補を正しい定義と判定する．ただし，複数の候補が正しい定義と判定された場合，主辞が一致する別の候補に含まれる候補は除外する．主辞の抽出方法は付録 B に示す．以上の処理を文献の本文中に現れる全ての変数に対して行う．

特徴量 1：定義らしさ　定義の候補の中には定義として自然な名詞句と，不自然な名詞句が存在する．この「定義らしさ」を特徴量化するため，まずは次の手順で辞書 D_1 を生成する．

1. 訓練用データセットの正例から変数の定義を取得し，主辞を抽出する．
2. 抽出した全ての主辞を見出し語化する．
3. 得られた見出し語群から重複を取り除いたものを辞書 D_1 とする．

手順 2 について，定義の主辞の語形 (単数形か複数形か) の区別は定義らしさの評価に不要であるため，見出し語化する．見出し語化には Stanza [4] を用いる．辞書 D_1 には，“temperature”，“concentration”，“volume” といった見出し語が集められる．ある変数 v の定義の候補 c に対して，特徴量 x_1 を次の手順で生成する．

1. c の主辞を抽出し，その見出し語 h を得る．
2. $h \in D_1$ のとき $x_1 = 1$，それ以外は $x_1 = 0$ とする．

特徴量 2：慣習的な変数記号の選び方か否か　化学プロセスに関する文献では，変数記号の選び方に，temperature には T を，concentration には C を対応させるといった特定の慣習がある．変数とその定義の候補との間にこのような慣習的な関係が成り立つ場合，その候補が正しい可能性は高いと推測できる．この「慣習的な変数記号の選び方か否か」を特徴量化するため，まずは次の手順で辞書 D_2 を生成する．

1. 訓練用データセットの正例から，変数の記号と定義の組を取得する．
2. 変数の記号から上付き文字や下付き文字等の修飾的な記号を除去した記号 (主記号) を取得する．
3. 変数の定義の主辞を抽出し，見出し語化する．
4. 2 で得られる記号と，3 で得られる主辞の見出し語の組を作る．
5. 得られた全ての組から重複を取り除いたものを辞書 D_2 とする．

手順 3 について，定義の主辞の語形は対応する主記号の選び方に影響しないと仮定し，見出し語化する．辞書 D_2 には，“T: temperature”，“C: concentration”，“V: volume” など，慣習的な変数記号の選び方を表す組が集められる．以降ではこのような組を「主記号-主辞ペア」と呼ぶ．ある変数 v とその定義の候補 c の組に対して，特徴量 x_2 を次の手順で生成する．

1. v と c から主記号-主辞ペア p を生成する．
2. $p \in D_2$ のとき $x_2 = 1$，それ以外は $x_2 = 0$ とする．

特徴量 3：変数と定義の位置関係　変数とその定義の候補が文献中の近い位置に存在するとき，その組が正しい組である可能性は高いと考えられる．しかし，変数とその定義の候補の間のトークンの数を特徴量として採用するには問題がある．例えば，変数 v_1, v_2 の定義がそれぞれ DEF_1, DEF_2 であることが以下のように表現されているとする．

$$v_1 \text{ is } DEF_1 \text{ and } v_2 \text{ is } DEF_2.$$

このとき，v_2 と DEF_2 の間には 1 つのトークン “is” が存在する．一方，v_2 と DEF_1 の間にも 1 つのトークン “and” が存在する．そのため，トークンの数のみではどちらの組が正しいか判別できない．そこで，パターンマッチングを取り入れてこの問題を緩和する．

まずは次の手順で辞書 D_3 を生成する．

1. 訓練用データセットの正例から変数とその定義の組を取得し，訓練用データセットの本文からその組の間の文字列を抽出する．
2. 出現頻度の高い文字列上位 N 個を取り出し，辞書 D_3 とする．

辞書 D_3 には，“is”，“represents” といった，変数とその定義の間によく出現する表現が集められる．ある変数 v とその定義の候補 c の組に対して，特徴量 x_3 を次の手順で生成する．

1. v と c の間の文字列 e を取得する．
2. v と c の間のトークンの数を d とする．ただし，$e \in D_3$ のとき $d = 0$ とする．
3. 特徴量 x_3 を式 (2) で計算する．

$$x_3 = f(\Delta) = \exp\left(-\frac{1}{2}\frac{\Delta^2 - 1}{\sigma^2}\right) \quad \Delta = d + 1 \tag{2}$$

式 (2) は Δ を正規化する式であり，Pagel と Schubotz により提案されたものである [2]．σ は関数 f の形を決定するパラメータである．

4 実験

4.1 評価指標

真陽性 (TP)，偽陽性 (FP)，偽陰性 (FN)，真陰性 (TN) を以下のように定義する．ただし，FP については二種類定義する．

- TP ： 文献中に定義が存在する変数に対して，正しい定義のみを抽出した場合
- $FP^{(1)}$ ： 文献中に定義が存在する変数に対して，誤った定義を一つでも抽出した場合
- $FP^{(2)}$ ： 文献中に定義が存在しない変数に対して，誤った定義を抽出した場合
- FN ： 文献中に定義が存在する変数に対して，何も抽出しなかった場合
- TN ： 文献中に定義が存在しない変数に対して，何も抽出しなかった場合

定義を抽出した変数のうち正しく定義を抽出した割合 (*Pre.*)，文献中に定義が存在する変数のうち正しく定義を抽出した割合 (*Rec.*)，*Pre.* と *Rec.* の調和平均 (*F*1)，正解率 (*Acc.*) を以下のように定義する．

$$Pre. = \frac{TP}{TP + FP^{(1)} + FP^{(2)}} \tag{3}$$

$$Rec. = \frac{TP}{TP + FP^{(1)} + FN} \tag{4}$$

$$F1 = 2 \times \frac{Pre. \times Rec.}{Pre. + Rec.} \tag{5}$$

$$Acc. = \frac{TP + TN}{TP + FP^{(1)} + FP^{(2)} + FN + TN} \tag{6}$$

手法の性能の優劣は *Acc.* により判断する．

4.2 実験設定

性能比較対象となる従来手法として，Kristianto らによる SVM を用いた手法 [1] を採用する．変数とその定義の候補の組を提案手法と同じ手順で生成し，各組の正誤を SVM で判定する．SVM のパラメータは Kristianto らと同じ値を用いる．

提案手法のパラメータは，$m_1 = 1, m_2 = 1, m_3 = 4, \sigma = 2, t = 0.6, N = 10$ に設定する．これは，予備実験によって定めた組み合わせである．

$\mathfrak{D}_{\mathrm{CSTR}}, \mathfrak{D}_{\mathrm{CRYST}}, \mathfrak{D}_{\mathrm{STHE}}$ をそれぞれ訓練用とテスト用に論文単位で分割する．$\mathfrak{D}_{\mathrm{CSTR}}$ と $\mathfrak{D}_{\mathrm{CRYST}}$ からはそれぞれ 3 報，$\mathfrak{D}_{\mathrm{STHE}}$ からは 2 報をテスト用とする．データセットの分割はランダムに 10 通り行い，各評価指標の平均を比較する．

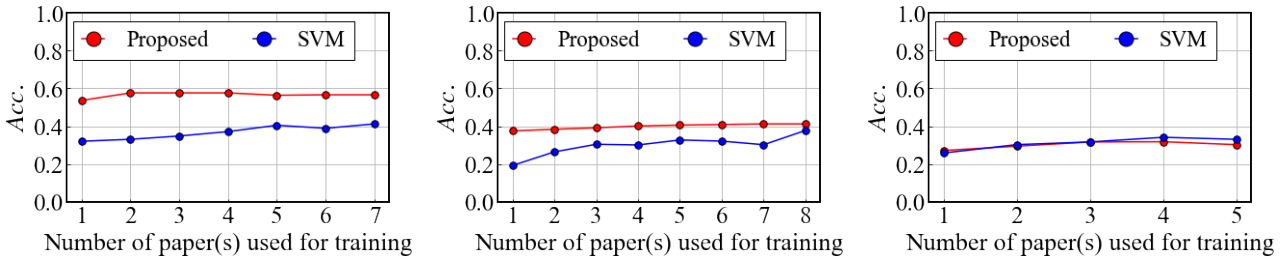
4.3 結果と考察

実験結果 表 1 に実験結果を示す．従来手法 (SVM) に比べて提案手法 (Proposed) は *Rec.* が高い傾向があり，特に $\mathfrak{D}_{\mathrm{CSTR}}$ では，提案手法が従来手法を 17.5%上回った．*Pre.* に注目すると，$\mathfrak{D}_{\mathrm{CSTR}}$ では提案手法が従来手法を 4.3%上回ったが，$\mathfrak{D}_{\mathrm{STHE}}$ では従来手法の方が 17.3%高かった．*Acc.* に注目すると，$\mathfrak{D}_{\mathrm{CSTR}}$ では提案手法が従来手法を 15.4%上回ったが，$\mathfrak{D}_{\mathrm{STHE}}$ では従来手法の方が 2.8%高かった．

名詞句抽出の性能 従来手法と提案手法はともに，3 節で述べた方法によって抽出した名詞句が定

表 1: 実験結果 (表示は%)

データセット	手法	*Pre.*	*Rec.*	*F*1	*Acc.*
$\mathfrak{D}_{\mathrm{CSTR}}$	SVM	63.4	39.3	47.8	41.2
	Proposed	67.7	56.8	61.5	**56.6**
$\mathfrak{D}_{\mathrm{CRYST}}$	SVM	48.9	30.8	37.7	37.7
	Proposed	46.5	39.0	42.4	**41.2**
$\mathfrak{D}_{\mathrm{STHE}}$	SVM	54.7	28.2	37.0	**33.0**
	Proposed	37.4	29.7	32.9	30.2



(a) CSTR (b) CRYST (c) STHE

図 1: 訓練に使用した論文数と *Acc.* の関係

表 2: c_{TP} と c_{FP} の数と *Acc.* [%] の X_0 からの変動

	$\mathcal{D}_{\mathrm{CSTR}}$			$\mathcal{D}_{\mathrm{CRYST}}$			$\mathcal{D}_{\mathrm{STHE}}$		
	c_{TP}	c_{FP}	*Acc.*	c_{TP}	c_{FP}	*Acc.*	c_{TP}	c_{FP}	*Acc.*
X_1	+1.1	−3.6	+5.4	±0.0	−13.5	+7.3	−2.4	−11.7	+7.0
X_2	+1.0	−6.1	+8.3	+0.5	−13.1	+6.7	−0.5	−8.0	+8.1
$X_{1,2}$	+1.1	−5.2	**+8.8**	−0.4	−16.0	**+8.2**	−2.4	−14.6	**+10.9**

義の候補となる．しかし，名詞句抽出は完璧ではないため，定義が文献中に存在していても候補として抽出されない場合があった．$\mathcal{D}_{\mathrm{CSTR}}, \mathcal{D}_{\mathrm{CRYST}}, \mathcal{D}_{\mathrm{STHE}}$ のテスト用データセットにおいて，定義が文献中に存在する変数のうち定義を候補に含む変数の割合はそれぞれ 82.5%，76.3%，57.9% であった．$\mathcal{D}_{\mathrm{STHE}}$ に対して従来手法と提案手法の *Rec.* が低くなる一因は，名詞句抽出の性能の低さにあると考えられる．

特徴量 1 と 2 の効果　提案手法により正しいと判定された定義の候補のうち，実際に正しい候補を c_{TP}，実際は誤っている候補を c_{FP} とする．次の 4 つの場合について，c_{TP} と c_{FP} の数および *Acc.* を比較することで，x_1 と x_2 が抽出性能にもたらす効果を調べた．

X_0 ： x_3 のみを用いる ($x_1 = x_2 = 0$ とする) 場合
X_1 ： x_1 と x_3 を用いる ($x_2 = 0$ とする) 場合
X_2 ： x_2 と x_3 を用いる ($x_1 = 0$ とする) 場合
$X_{1,2}$ ： x_1, x_2, x_3 を用いる場合

$X_1, X_2, X_{1,2}$ における c_{TP} と c_{FP} の数と *Acc.* の，X_0 からの変動を表 2 に示す．X_1, X_2 では，X_0 より c_{FP} の数が少なく，x_1 と x_2 が c_{FP} を減少させることがわかる．ただし，x_1 は「定義らしさ」を備えた候補の *score* を一様に増加させてしまうために，新たに c_{FP} を発生させる場合もある．c_{FP} を減少させる効果に比べると，x_1 と x_2 の c_{TP} を増加させる効果は小さく，$\mathcal{D}_{\mathrm{STHE}}$ ではかえって c_{TP} を減少させている．また，どのデータセットにおいても，$X_{1,2}$ の *Acc.* は X_1 と X_2 の *Acc.* よりも高く，x_1 と x_2 を合わせて用いることが性能向上に有効であることが示された．

訓練用データセットのサイズと性能の関係　訓練用データセットに含まれる論文数を変化させて *Acc.* を算出した結果を図 1 に示す．従来手法では，訓練に用いる論文数を増やすと *Acc.* が増加する傾向があったが，提案手法では，2 報以上を訓練に用いたときの *Acc.* は論文数によってほとんど変化しなかった．この違いは，従来手法では正例と負例の両方を訓練に用いるために大きな訓練用データセットが必要となる一方で，提案手法では正例のみを訓練に用いるために，小さな訓練用データセットで抽出規則を学習できることに起因すると考えられる．

5　おわりに

本研究では，変数の記号と定義に関する情報を活用した変数定義抽出手法を提案した．連続槽型反応器，晶析プロセス，熱交換器に関する論文計 28 報を用いて従来手法と提案手法の性能を比較した結果，提案手法の正解率が連続槽型反応器に対しては 15.4%，晶析プロセスに対しては 3.5%従来手法を上回った．一方，熱交換器に対しては従来手法の正解率が提案手法より 2.8%高かった．今後の主な課題を以下に挙げる．

パラメータの最適化　今回の実験では，提案手法のパラメータは最適化されていない．パラメータの自動的な最適化を手法に組み込むことで，性能の改善が見込まれる．

名詞句抽出の改善　熱交換器に関する論文中の 4 割程度の変数については，名詞句抽出の性能が低いために定義抽出に失敗している．Lin らは構文解析を数学的表現が含まれる文章に対応させることが有効であることを示している [6]．同様に構文解析器を改善し，名詞句抽出の性能を向上させることで変数定義抽出性能の向上が見込まれる．

謝辞

本研究は JSPS 科研費 JP21K18849 および JST 次世代研究者挑戦的研究プログラム JPMJSP2110 の助成を受けたものです.

参考文献

[1] Giovanni Yoko Kristianto, Goran Topic, and Akiko Aizawa. Extracting textual descriptions of mathematical expressions in scientific papers. *D-Lib Magazine*, Vol. 20, No. 11, p. 9, 2014.

[2] 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*, No. 1186, Aachen, 2014.

[3] 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, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[4] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2020.

[5] Dan Klein and Christopher D Manning. Accurate unlexicalized parsing. In *Proceedings of the 41st annual meeting of the association for computational linguistics*, pp. 423–430, 2003.

[6] 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 Sonja Schimmler and Uwe M. Borghoff, editors, *Proceedings of the ACM Symposium on Document Engineering 2019, Berlin, Germany, September 23-26, 2019*. ACM, 2019.

A　定義抽出対象となる変数

数学的表現が Mathematical Markup Language (MathML) 形式で記載された HTML 形式の文献から，`math` タグが付与された要素の一覧 $\varepsilon_{\text{math}}$ を取得する．$\varepsilon_{\text{math}}$ の各要素のうち，次のいずれかに該当するものを定義抽出対象となる変数と定義する．

- 複数の記号が連続した文字列 (e.g. $\Delta E, \delta t$)
- 1 つの記号に上付き文字，下付き文字，上ルビ，下ルビのいずれか，あるいはその組み合わせが付与された文字列

B　主辞抽出方法

名詞句の主辞は，以下の手順で抽出する．

1. 名詞句に対して係り受け分析を行い，係り受けグラフを生成する．
2. 係り受けグラフの始点となっている単語が名詞であれば，それを主辞とする．

係り受け分析には Stanford Parser [5] を使用する．