特許文書を対象とした化学実験構造化のための基礎的検討

作本猛 [1]　邊土名朝飛 [1]　山本雄太 [2]　森楓 [2]　野中尋史 [1]

[1] 長岡技術科学大学大学院 工学研究科
[2] 長岡技術科学大学 工学部
{s183353,s173348,s193378,s203372}@stn.nagaokaut.ac.jp
nonaka@kjs.nagaokaut.ac.jp

1　はじめに

化学情報の構造化は、薬品や材料の開発、低コストな合成経路の探索といった応用のために重要である．化学分野では、新規発明に対する権利保護の観点から特許出願が論文出版より優先される傾向にあり、また、特許明細書は、化合物の性質、評価実験等に関する詳細な記述を含むことから、特許を対象とした情報抽出、構造化の研究が広く行われてきた [1][2]．近年、こうした研究の対象として、固有名や物性といった化学物質に関連する情報に加え、操作や条件といった実験手順に関連する情報が注目されている．

Vaucher ら [3] は、合成実験の自動化に向けて、有機化学の特許明細書中に記載された合成実験の手順を、操作を主体とした一連の機械読解可能な表現に変換することで構造化した．また、W-NUT2020[1)] では、化学実験の手順を対象とした関係抽出タスクの評価型ワークショップ [4] が開催された．ところが、こうした研究は英語を対象としたものが中心であり、日本語を対象とした実験手順の抽出、構造化に関する研究は我々の知る限り存在しない．欧米日中韓 5 か国の特許庁が公開した 2018 年度の統計 [5] によると、日本では 1 年間で 3.3 万件の化学特許が出願されており、これは中国の 9.5 万件、米国の 4.0 万件に次ぐ規模である．このことは、日本語での化学情報構造化の重要性を示している．

そこで、本研究では図 1 に示すように日本語で記述された化学実験手順の構造定義と情報抽出手法の開発を行う．本論文では構造化の基礎として、有機化学分野の特許明細書から、操作を表す動詞とその目的語について、文節単位での特定と、動詞-目的語間の関係抽出を行った．

1）　Workshop on Noisy User-generated Text

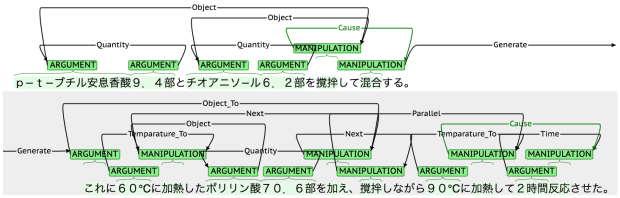

図 1　化学実験手順の構造化を目的とした情報抽出の例

長尾ら [6] は、手続きを表す日本語文の意味解析のために、動詞辞書の構築と動詞-目的語間に存在する文法的特徴のまとめ上げを行った．こうした手法は近年の研究 [7] にも応用されており、手続き的な文章を対象とし、かつドメインを限定している場合、辞書構築と文法的特徴によるナイーブな解析手法は情報抽出においても実用的な性能が期待できる．しかし、特許明細書には造語が頻出するため、動詞や固有名詞の辞書構築がドメインを限定しても現実的ではない．そのため、我々は大規模な辞書構築なしに、実験手続きを表す文の特定と情報の抽出を行うための文法的特徴についてまとめ上げ、その有効性と適用範囲に関する実験、考察を行った．

2　化学実験文書の構造的特徴

2.1　化学実験文書を構成する文

日本語で記述された化学実験文書を構成するのは、大きく分けて以下の 2 種類の文である．

- 実験操作文
  - 混合液を常温まで放置し反応させた
- 実験説明文
  - 生成物の収率は８０．０％であった
  - 温度が常温になるまで昇温した

それぞれの特徴を表 1 に示す．

実験操作文は、操作を表す動詞またはサ変名詞による述語が主体となる文であり、主に操作、操作の

表 1　実験操作文と実験説明文の比較

	実験操作文	実験説明文
主語の省略	○	×
述語の構成	動詞、サ変名詞	動詞、サ変名詞、形容詞、判定詞

目的語または修飾語、操作の目的語に対する修飾語の 3 種類が存在する．以後、操作を表す動詞による文節を**操作**、操作の目的語や修飾語を含む文節を**操作引数**と呼ぶ．

実験説明文は、操作や操作引数の性質、状態、物理量等について説明を行うものであり、判定詞や形容詞による述語が主体となる文である．

これらのうち、本論文では実験操作文を情報抽出の対象とした．

2.2　実験操作文の構造

本節では、実験操作文に頻出する、格助詞や品詞、修飾の形態によって表現される操作引数-操作間の関係構造について、主要な大分類である 3 種類と、それらを構成する小分類や主な関係パターンについて、その実例として代表的なものを提示する．各パターンについて、操作を意味する文節は**太字**で、操作引数を意味する文節は下線付きで表示し、また、そのパターンが各大分類中に占める割合を 4 節で提示するアノテーションデータから算出し、パターンの右側に示した．

2.2.1　主対象

主対象は、操作とその主な対象となる操作引数との関係を表現する．

- ヲ格 + **述語** (93.2%)
  - 溶媒を**留去**した
- ノ格 + **述語** (+ 述語) (1.6%)
  - 溶媒の**留去**を行った
- ガ格 + **述語** (3.0%)
  - 溶媒が**留去**された

2.2.2　間接対象

間接対象は、任意の主対象を取る操作と、それによって間接的に影響を受ける操作引数との関係を表現する．基本的には起点対象、終点対象の 2 種類に分類される．

起点対象　操作、あるいは主対象関係にある操作引数の起点となる対象を表現する．

- カラ格 + **述語** (5.3%)
  - 反応物から不要物を**除去**した

終点対象　操作、あるいは主対象関係にある語の終点となる対象を表現する．

- ニ格 + **述語** (88.4%)
  - 氷上に溶液を**滴下**した

2.2.3　条件

条件は、操作とその定量的あるいは定性的な条件、または操作を促進するためのモノを表す操作引数との関係を表現する．基本的には、基本条件、終点条件、物的条件の 3 種類に分類される．

基本条件　操作、あるいは主対象関係にある操作引数に付随する条件を表現する．

- デ格 + **述語** (11.4%)
  - 室温で**撹拌**する
- 無格 (+ 述語) + **述語** (24.5%)
  - 1 時間**放置**する
  - 1 時間かけて**放置**する
- 形容系の連用 + **述語** (6.5%)
  - ゆっくりと**撹拌**する

終点条件　操作が、その主対象関係にある操作引数に対して、数量や性質を変化させる時、その終点となる条件を表現する．

- ニ/ト格 + **述語** (9.6%)
  - 40 ℃/塩基性に**調整**する
  - 40Torr/塩基性と**する**
- マデ格 + **述語** (5.3%)
  - 常温まで**昇温**した

物的条件　任意の主対象を持つ操作と、それを可能にする、あるいは促進するために用いられる操作引数との関係を表現する．

- デ格 + **述語** (25.2%)
  - 酢酸エチルで**抽出**した
- ニ格 + テ + **述語** (1.5%)
  - 酢酸エチルにて**抽出**した
- カラ格 + **述語** (1.5%)
  - 溶媒から**再結晶**した
- ニ格 + 述語 + **述語** (2.2%)
  - 酢酸エチルによって**抽出**した
- ヲ格 + 述語 + **述語** (2.6%)
  - 酢酸エチルを用いて**抽出**した

3　提案手法

3.1　タスク設計

本論文では、2.2 節でそれぞれ**太字**、下線付きで示した操作、操作引数の 2 種類の役割について、文節単位での特定と、対応する役割ラベルの付与を行う．さらに、各操作引数文節の係り先となる操作文節の特定、2.2 節で示した 3 種類の関係を表すラベルの付与を行う．以下の図 2 に情報抽出例を示す．

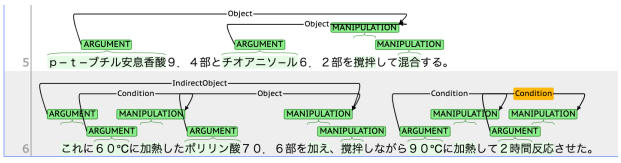

図 2　情報抽出例

3.2　情報抽出のワークフロー

3.2.1　前処理

化合物名の置換　有機化合物の名称は、「1’-(ω-ブロモデシル)-3’,3’-ジメチル-6-ニトロスピロ〔2H-1-ベンゾピラン-2,2’-インドリン〕」のように、官能基名や位置番号等をハイフン、カンマ等の記号で結合した形で表される．これは形態素・構文解析の誤りに繋がるが、日本語有機化合物名は慣用名の使用や省略等によって辞書置換が困難であるため、ハイフン記号に着目した正規表現で単純な置換を行った．

括弧の除去　括弧は表題番号、ラベル、直前の文節の説明等広い用途で使用されるが、係り受け解析時に誤りの原因となりうる．括弧は入れ子構造を取りうるため正規表現は使用せず、スタックを使用して一時的に除去を行った．

形態素・構文解析　Juman++[8]、KNP[9] を使用し、係り受けや格、修飾形態等の情報を持った文節リストを生成した．

実験操作文の抽出　構文解析によって得た文節リストを対象に、ヲ格または述語のない文を除去する．次に、2.1 節で示した実験操作文、実験説明文の特徴により、実験説明文を除去する．また、特定の動詞述語を末尾に持つ文も除去する．以下に除去対象の動詞例を示す．

除去対象の動詞例

なる、示す、有する、説明、確認、例示、向上

3.2.2　文節の役割、係り先、関係の特定

2.2 節で示した文法的特徴に合致する文節の組に対し、操作、操作引数の役割ラベルを各文節に付与し、対応する関係ラベルを操作引数-操作間に付与する．非述語の文節が非述語の文節に係っている場合、以下の例外処理を実行する．

- 文節が並列文節なら、並列の先端文節の係り先述語に係り、同じ関係ラベルを付与
- 文節が無格で数量文節に係るなら、数量文節の代わりにその係り先述語に係り、その格情報で関係ラベルを付与

4　実験

化学分野の実験手順を表す日本語文書として、1991 年から 2000 年の間に出願された、国際特許分類 C07(有機化学) の特許 42 件をランダムに選択し、2.2 節に示した文節の役割ラベルと関係ラベルについてアノテーションを行った．文節の役割ラベルと関係ラベルの個数を表 2 に示す．このデータに対して、3.2 節で示した方法によって文節の役割ラベル、操作引数の係り先操作、操作引数-操作間の関係ラベルの特定を行い、それぞれの性能、また、ワークフローの全体的な性能について評価を行った．ここで、操作または操作引数と予測した文節について、下記の 4 種類を評価に用いる正解条件とした．

条件 1　見出し語がアノテートされたものと一致またはそれを完全に含む
条件 2　役割ラベルがアノテートされたものと一致
条件 3　役割ラベルが操作引数の場合、係り先の操作がアノテートされたものと一致
条件 4　関係ラベルがアノテートされたものと一致

文節の役割ラベル特定は、全データを評価対象とし、条件 1〜2 を満たすものを正解として評価した．操作引数の係り先操作の特定は、独立した性能評価のために、条件 1〜2 を満たすものを評価対象とし、条件 3 を満たすものを正解として評価した．操作引

表 2　ラベルの個数

分類	ラベル	個数
役割	操作	1495
	操作引数	2066
関係	主対象	1010
	間接対象	210
	条件	822

表 3　文節の役割特定の評価結果

		再現率	適合率	F 値
操作		0.803	0.954	0.872
操作引数	主対象	0.882	0.916	0.899
	間接対象	0.843	0.843	0.843
	条件	0.902	0.895	0.899

表 4　操作引数の係り先操作特定の評価結果

	再現率	適合率	F 値
主対象	0.775	0.843	0.808
間接対象	0.786	0.820	0.803
条件	0.686	0.757	0.720

数-操作間の関係ラベル特定も、これと同様にして条件 1〜3 を満たすものを評価対象とし、条件 4 を満たすものを正解として評価した．また、ワークフローの全体的な評価としては、全データを評価対象とし、条件 1〜4 を全て満たすものを正解として評価した．評価指標には再現率、適合率、F 値を使用した．それぞれの評価結果を、表 3〜6 に示す．

表 5 より、関係ラベルはアノテーションデータの 9 割以上を再現できており、F 値も 8 割を超える結果となった．また、表 6 より、F 値で 0.64〜0.77 と教師なしの情報抽出としては十分な性能が示された．

5　考察

表 6 の結果より、ルールベース手法としては全体的に十分な結果が得られたが、間接対象や条件等、一部の関係について評価値が低い結果となった．そこで、ボトルネック要因の分析を行う．

表 3、5 より、これら 2 タスクの性能は十分に高いことがわかる．ただし、関係ラベル特定における間接対象の適合率は比較的低い値であり、表 6 にも類似した傾向が見られるため、関係ラベル特定が間接対象のボトルネック要因であると考えられる．間接対象と誤予測された例を見ると、「常圧に」、「室温に」のように、終点条件であるものを終点対象と予測したものが大半であった．こうした語は「常、高、低」のような接頭辞や、「圧、温」のような語が文節内に存在しやすいため、見出し語中に頻出する上記のようなパターンを取り入れることで解決を図ることができる．

表 4 より、操作引数の係り先操作の特定は、表 3、5 に示す 2 タスクと比較して比較的低めの評価値となっており、これがワークフローの全体的な評価値を落とす主原因であると考えられる．そこで、操作

表 5　操作引数-操作間の関係ラベル特定の評価結果

	再現率	適合率	F 値
主対象	0.931	0.981	0.955
間接対象	0.964	0.726	0.828
条件	0.902	0.895	0.899

表 6　ワークフローの全体的な評価結果

	再現率	適合率	F 値
主対象	0.723	0.827	0.772
間接対象	0.757	0.596	0.667
条件	0.619	0.673	0.645

引数の係り先操作特定に失敗した文を目視で確認し、原因パターンのまとめ上げを行った．代表的な失敗パターンとその割合を以下の表 7 に示す．これらのうち、構文解析器による失敗について述べる．特許には過剰に読点が打たれた文章が多く、「全体を、撹拌しながら２４時間、室温で保持し、次に水５００ｍｌを加えた」のような文章に対して、構文解析器は読点を基準とした並列関係を誤予測しやすい．また、もし不要な読点がなかったとしても、「全体を」が「攪拌」に係るのか、「保持」に係るのかの特定は、文節の意味と文脈に依存するため、誤予測が発生しやすい．こうした例は表層的なパターンのみで対処することは難しく、係り受け例を大量に収集して教師あり学習を行う等のアプローチが必要となる．

表 7　操作引数の係り先操作特定の失敗パターン例

失敗箇所	失敗原因	割合
手法	操作文節の特定失敗	21%
構文解析器	並列関係の解析失敗	47%
	未知語の解析失敗	16%
	入れ子表現の解析失敗	15%

6　おわりに

本研究では、日本語で記載された化学実験手順の構造化を目的とし、その基礎として、化学分野の実験手順文書に存在する文法的特徴のまとめ上げを行った．一部課題点が見られはしたが、基本的には操作引数-操作間の関係は文法的特徴を使用することで高精度な特定が可能であり、一連の情報抽出についても、大量の教師データや辞書の作成なしに、十分な精度で行うことが可能であることが確認できた．今後は、ルールベース手法での限界が見られた点に対し、統計的手法の適用による改善を行う．

参考文献

[1]Daniel Mark Lowe. *Extraction of Chemical Structures and Reactions from the Literature*. Phd thesis, University of Cambridge, 2012.

[2]田中一成, 池田紀子. オープンデータを用いた化学特許情報活用へのアプローチ. *Japio year book*, pp. 206–211, 2017.

[3]Alain C. Vaucher, Federico Zipoli, Joppe Geluykens, Vishnu H. Nair, Philippe Schwaller, and Teodoro Laino. Automated extraction of chemical synthesis actions from experimental procedures. *Nature Communications*, Vol. 11, No. 1, p. 3601, 2020.

[4]Jeniya Tabassum, Sydney Lee, Wei Xu, and Alan Ritter. Wnut-2020 task 1 overview: Extracting entities and relations from wet lab protocols. *arXiv:2010.14576 [cs]*, 2020.

[5]Five IP Offices. Ip5 statistics report 2018 edition, （2021-01 閲 覧）. https://www.fiveipoffices.org/wcm/connect/fiveipoffices/8c519416-173d-4b32-99ed-5387045c46a2/IP5+Statistics+Report+2018_20122019_full.pdf?MOD=AJPERES&CVID=.

[6]長尾真, 辻井潤一, 田中一敏. 意味および文脈情報を用いた日本語文の解析-名詞句・単文の処理. 情報処理, Vol. 17, No. 1, 1976.

[7]豊辻宏旨, 松崎拓也, 佐藤理史. オントロジーに基づく意味解析を用いた「化学」正誤問題の自動解法. 人工知能学会全国大会論文集 第 31 回全国大会 (2017), pp. 3G11–3G11, 2017.

[8]Arseny Tolmachev, Daisuke Kawahara, and Sadao Kurohashi. Juman++: A morphological analysis toolkit for scriptio continua. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 54–59, 2018.

[9]Daisuke Kawahara and Sadao Kurohashi. A fully-lexicalized probabilistic model for japanese syntactic and case structure analysis. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pp. 176–183, 2006.