特定の専門分野を対象とした意味役割付きデータ作成手法 ～有機合成手順の抽出を例として～

町 光二郎 [1]　秋山 世治 [2]　長田 裕也 [2]　吉岡 真治 [1,2,3]
[1] 北海道大学 大学院情報科学院　[2] 北海道大学 化学反応創成研究拠点
[3] 北海道大学 大学院情報科学研究院
machi@eis.hokudai.ac.jp　{s.aki, nagata}@icredd.hokudai.ac.jp
yoshioka@ist.hokudai.ac.jp

概要

文書から作業手順などの詳細な情報を抽出するためには、述語の意味役割に注目した解析が役立つ。このようなタスクに活用可能な言語資源として、PropBank などの意味役割付与コーパスが提案されている。しかし、このようなコーパスは、専門分野における手順などの情報を表現するには意味役割のバリエーションが不十分である。本稿では、我々が提案した有機化学反応手順解析のための意味役割付きコーパスを紹介すると共に、特定の専門分野における意味役割付きデータ作成の方針について議論する。

1　はじめに

学術論文や特許文書に報告される化学反応の数は急速に増加している。新しい化学反応は、Reaxys [1] などの化学反応データベースに収集されているが、これは人間の専門家によって行われており、多くの時間と高いコストがかかっている。この負担を軽減するために、自動情報抽出のための手法が研究されている [2, 3]。その多くは化学反応式を構成する材料となる物質、生成物、触媒といった化学物質やパラメータなどの既存の化学反応データベースに収録されるような主な情報に焦点を当てたものであるが、化学反応の手順を文書から抽出しようとした研究もいくつかある [3, 4, 5]。

この化学反応手順の抽出のために、文内の述語と項の意味的な関係 (意味役割) を付与する方法が提案されている。Cheminformatics Elsevier Melbourne Universities (ChEMU) [3] では、意味役割付与のコーパスの１つである Proposition Bank (PropBank) [6] の形式を用いて、特許文書から述語を関係する項とその意味役割とともに抽出する反応情報抽出タスクを提案している。しかし、化学反応式のレベルの情報とそれに関連するパラメータの情報を抽出することを重視した設計になっているため、2.2 節で述べる意味役割の簡略化が行われており、化学反応手順の再現という観点から見ると情報が不十分である。また、PropBank も化学分野を対象とした文書を扱っていないため、化学反応手順を表現するためには不十分である。

このような課題を解決するために、我々は、論文中に記述された化学反応手順に関する詳細な手順を抽出するためのコーパスとして、適切な意味役割を付与した OSPAR[1)] (Organic Synthesis Procedures with Argument Roles、図 1) を提案している [7]。本稿では、この OSPAR の概要について紹介するとともに、コーパス作成に用いた意味役割付きデータ作成手法について述べる。

2　関連研究

2.1　意味役割と PropBank

意味役割は、文内の述語と項の関係を表すことに用いられる。PropBank [6] は、句構造コーパスの Penn Treebank [9] に含まれる文に対して、意味役割を定義したコーパスである。そのため、Penn Treebank と同様に *The Wall Street Journal* (*WSJ*) の記事に含まれる文を元に意味役割が定義されている。PropBank では動詞の用法ごとに意味役割がそれぞれ定義されているため、意味役割のセット (roleset) は、“add.01” のように動詞の後に用法を表すための ID と合わせて表記することで区別される。roleset を統語フレームと結びつけたものは、frameset と呼ばれる。各動詞の意味役割に番号 (Arg0, Arg1, Arg2, ...) が付与されている。Arg0 は一般に述語の動作主 (agent) を表し、Arg1 は被動作主 (patient) を表す。Arg2 以上の他の引数は一貫した用法を持たない。その他にも時

1） https://github.com/mlmachi/OSPAR

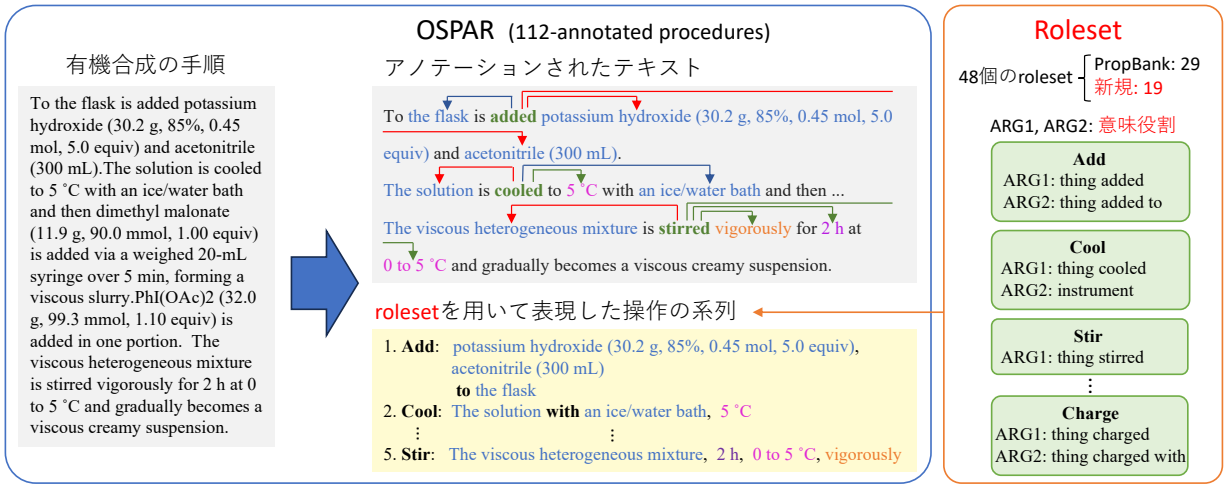

図 1: OSPAR コーパスの概要 (図は文献 [7] を元に和訳したもの。例のテキストは文献 [8] より引用)

間や場所などの付加的な情報は、動詞によらない共通のものとして ARGM というラベルが定義されている。以下に roleset の例を示す。

frameset

add.02 "mathematics"

roleset

- ARG0: adder
- ARG1: thing being added
- ARG2: thing being added to

Example:

[$_{ARG0}$ Judge Curry] added [$_{ARG1}$ an additional $55 million] to [$_{ARG2}$ the commission's calculations].

2.2 ChEMU における意味役割の簡略化

ChEMU では、PropBank に基づく意味役割ラベルが用いられている。しかし、化学反応式のレベルの反応情報の抽出が主な目的となっているため、PropBank の roleset を簡略化して使用していた。例えば、彼らのコーパスでは、A+B として記述される化学反応式があればよいので、"*A is added to B*" という文において PropBank では A を ARG1、B を ARG2 とするのに対し、意味役割を区別せず両方に ARG1 というラベルをつけている。しかし、B に A を加えることと、A に B を加えるということは違う結果をもたらす可能性があり、化学反応の再現に影響を及ぼす。他の動詞に対しても同様に、意味役割を考慮せずに、付加的な情報である ARGM 以外は ARG1 というラベルが用いられており、何らかの関係があるということのみ表現されている。

3 OSPAR コーパス

3.1 化学反応情報抽出のための roleset の整備

PropBank で定義された roleset は、*WSJ* の記事に出現する動詞を対象としているため、化学反応手順を表す roleset が十分に収録されていないと考えられる。そこで我々は、PropBank 形式の roleset を用いて、化学反応手順、特に有機合成の操作を表現するための新しい roleset を定義し、有機合成化学の論文誌の記事を対象として OSPAR コーパスを作成した [7]。その方法として、コーパスに含まれるすべての操作を表す述語について、PropBank に適当な roleset が存在するか確認し、存在した場合はそのまま利用し、そうでない場合には、既存の roleset に修正を加えたり、新しく roleset を作成するなどした。修正の例として、化学分野の専門家にアノテーションを依頼する際に、引数の説明が不適切と判断したものについては、化学分野への適応を考慮した説明文に修正を行なったというものがある。

ARG0 は動作主を示すが、行動主体は明らかに実験者であるため、この作業では基本的にこのラベルを無視した[2)]。PropBank では、汎用的な用法の記述を目指していることもあり、一つの述語に対して、3 つ以上の引数を持つ意味役割が設定されていた。しかし、本研究における roleset は、化学反応手順の情報を十分に記述するという観点から役割を整理した。その結果、$n = 3$ 以上の ARGn を必要とする述語は存在しなかったため、ARG1(被動作主) と ARG2(それ以外) のみを使用することとした。

2) 例外として、"contain" という roleset に対してのみ ARG0 を定義した。

3.2 アノテーションラベル

化学反応の手順の情報を抽出するための用語抽出タスクとしては、ChEMU の定義などを参考に、次の 6 つの用語を抽出することとした。

- REACTION_STEP: 化学反応の操作を表す動詞。
- ENTITY: 化学物質名や実験器具などのエンティティ
- TIME: 時間
- TEMPERATURE: 温度
- TEMP_TARGET: TEMPERATURE が示す場所 (容器の内部の温度か外部の温度か)
- MODIFIER: 上記以外で化学反応の再現に役出つ情報

また、関係抽出の際には、roleset に定義された意味役割に加え、動作に関する修飾詞を ARGM として抽出することとした。

- (ARG0): 動作主
- ARG1: 被動作主
- ARG2: 動作主と被動作主以外
- ARGM: パラメータ (TIME, TEMPERATURE, TEMP_TARGET, MODIFIER)

3.3 コーパスのアノテーションと roleset の追加

アノテーションの方法 コーパスの作成にあたっては、実験の手順がより詳細に記述されている論文のデータを用いることが望ましいと考え、掲載されている論文中の化学反応手順の再現性が編集者らによって確認されている論文誌 *Organic Syntheses* [10] を用いた。具体的には、112 件の論文中に記載されている 112 件の手順 (以下、文書とする) を対象とした。また、REACTION_STEP に関するコーパスのアノテーションの際には、PropBank を参照しながら、必要に応じて、3.1 節で紹介した基準により、roleset の追加や編集を同時に行うこととした。コーパスのアノテーションは、brat アノテーションツール [11] を用いて、2 名の有機化学者 (准教授と助教授) と 1 名の情報科学者 (博士課程の学生) の 3 人の著者によって行われた。

roleset の追加 まず、コーパスの仮アノテーションを行い、被動作主は ARG1、それ以外の必要と思われる項については ARG2 としてラベルを付与した。次に、アノテーションした動詞に対して、情報科学を専門とするアノテータが既存の roleset に適切なものが存在するか確認し、存在する場合にはそのまま roleset の候補リストに追加し、そうでないものについては新たに定義した。その後、アノテータ全員で実際にテキストの例と roleset を見比べながら適宜修正を行い、最終的な roleset を確定した。その中で、化学分野の専門家にアノテーションを依頼する際に、引数の説明が不適切と判断したものについては、化学分野への適応を考慮した説明文に修正を行なった。例えば、PropBank の “mix.01” の “ingredient one” (ARG1) と “ingredient two” (ARG2) は、我々のコーパスでは “ingredient” (ARG1) として集約された。これは、PropBank の意味役割の違いは統語的な観点に由来するものが存在するが、化学的な観点では重要ではなかったためである。PropBank の動詞は用法に応じて複数の roleset を持つことができるが、我々のコーパス内では複数の用法で用いられている動詞は存在しなかった。そのため、roleset を表す際の末尾の番号は省略した。

コーパスの統計情報 アノテーションした 112 件の有機合成手順 (以下、文書とするは、8:1:1 の割合で訓練セット、開発セット、テストセットに分割された。表 1 に各データセットに含まれる文書数、文の数、用語の数、関係の数を示す。

表 1: OSPAR コーパスの統計情報

	訓練	開発	テスト	合計
文書の数	90	11	11	112
文の数	664	86	68	818
用語の数	2142	259	223	2624
関係の数	1621	197	168	1986

3.4 roleset の分析

コーパスのアノテーションを行った結果、48 個の roleset が使用された。そのうち、新しく定義した roleset は、19 個であり、全体の約 40%であった。新しく定義した roleset をさらに分類すると、動詞自体は PropBank に存在したものが 11 個、動詞自体が存在しなかったものが 8 個であった。したがって、PropBank の語彙は化学反応の手順を表現するには不十分であり、roleset を拡張することの必要性が確認された。

図 2 は、OSPAR コーパスにおける roleset の出現頻度の分布である。このグラフから、化学反応を記述するために特有な高頻度の roleset が存在し、そのバリエーションは多くないことを確認した。

3.5 コーパスを用いた学習結果

我々は、文献 [7] において、OSPAR を利用した情報抽出システムを構築し、コーパスの実用性を検証

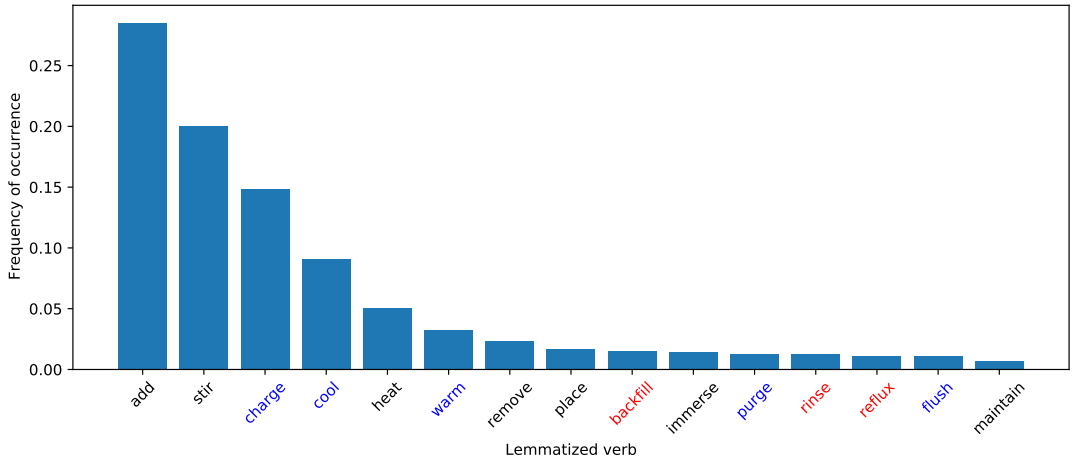

図 2: 出現頻度が上位 15 件の roleset の分布 (図は文献 [7] より引用)。色付きの文字は新規の roleset である。

した。情報抽出は、化学分野の事前学習言語モデルである ChemBERT [12] を用いた用語抽出タスクと関係抽出タスクのパイプラインのシステムを用いた [13]。その結果、テストデータにおいて、用語抽出タスクの F 値は、0.8709 であった。関係抽出タスクの F 値は、用語抽出の予測結果を用いたパイプライン方式の場合で 0.8393、用語抽出の正解データを用いた場合で 0.9436 であった。この差 (約 0.1) は小さくないことから、用語抽出の性能の重要性を確認した。それぞれのタスクの F 値から性能は悪くないと考えられるため、コーパスのアノテーションの一貫性とコーパスの有用性が示唆された結果となっている。訓練データに含まれないまたは数回しか出現しない動詞については、抽出に失敗する傾向があった。一方で、訓練データに含まれない動詞が抽出されている例も見られた。訓練セットで 3 回以上含まれていた動詞は全てテストセットで発見された。

4 特定の専門分野を対象とした意味役割付きデータの作成

本研究では、特定の専門分野における情報抽出を目的として roleset の作成とコーパスの作成を並行して行う方法を提案した。新しい専門分野の文書に対して、roleset を付与する際には、既存の汎用的な roleset を使うだけでは十分でないことを確認し、アノテーション作業を通じて、作成する roleset について議論することが言語学を背景としない専門家と共に作業をする際に有用であることも確認された。また、その作業の中で、抽出したい情報の粒度に応じた roleset の簡単化を行うことが、言語学を背景としないアノテータのアノテーション作業を効率化することに寄与するだけでなく、少数の事例であっても、十分に精度の高い機械学習が行えるようなコーパスの作成につながったと考えている。

このような形で作成した意味役割フレームワークは、PropBank のものと比べ、以下のような違いがある。まず、統語構造に関する分析を目的としていないため、統語構造に関する情報が付与されていない。つまり、frameset ではなく roleset のレベルでの表現を用いている。次に、ARG1 や ARG2 のスパンに前置詞を含まない。これは、作業手順の抽出という観点から見ると、意味役割さえついていればエンティティの役割がわかるためである。

これまで有機合成の文脈で roleset について議論してきた。有機合成と材料科学には共通する作用があるため、材料科学にも応用できると考えている [14, 15]。一方、roleset の種類は材料科学における操作を表現するには不十分であると考えている。隣接分野においてもこのような違いがあることを考慮して roleset を拡張していくことが必要であると考えている。

5 おわりに

本稿では、ニュース記事を対象として作成された PropBank の roleset が専門分野における作業手順を表現するのに不十分であることを確認し、コーパス作成時にコーパスのアノテーションと roleset の作成を並行して行う方法を提案した。この方法により有機合成を対象とした作業手順抽出用コーパスを作成し、特定の専門分野において、専門家の意見を反映させながら、有用な意味役割付きデータの作成が可能であることを確認した。

謝辞

本研究は JSPS 科研費 JP21K19814, JP23K18500 の助成を受けたものである。また、本研究の一部には、JST ERATO JPMJER1903 および、文部科学省世界トップレベル研究拠点プログラム (WPI) により設置された北海道大学化学反応創成研究拠点 (ICReDD) からも支援を受けた。

参考文献

[1] Reaxys. https://www.reaxys.com. (accessed June 20, 2023).

[2] Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong Lu, Robert Leaman, Yanan Lu, Donghong Ji, Daniel M. Lowe, Roger A. Sayle, Riza Theresa Batista-Navarro, Rafal Rak, Torsten Huber, Tim Rocktäschel, Sérgio Matos, David Campos, Buzhou Tang, Hua Xu, Tsendsuren Munkhdalai, Keun Ho Ryu, S. V. Ramanan, Senthil Nathan, Slavko Žitnik, Marko Bajec, Lutz Weber, Matthias Irmer, Saber A. Akhondi, Jan A. Kors, Shuo Xu, Xin An, Utpal Kumar Sikdar, Asif Ekbal, Masaharu Yoshioka, Thaer M. Dieb, Miji Choi, Karin Verspoor, Madian Khabsa, C. Lee Giles, Hongfang Liu, Komandur Elayavilli Ravikumar, Andre Lamurias, Francisco M. Couto, Hong-Jie Dai, Richard Tzong-Han Tsai, Caglar Ata, Tolga Can, Anabel Usié, Rui Alves, Isabel Segura-Bedmar, Paloma Martínez, Julen Oyarzabal, and Alfonso Valencia. The chemdner corpus of chemicals and drugs and its annotation principles. **Journal of Cheminformatics**, Vol. 7, No. 1, p. S2, Jan 2015.

[3] Dat Quoc Nguyen, Zenan Zhai, Hiyori Yoshikawa, Biaoyan Fang, Christian Druckenbrodt, Camilo Thorne, Ralph Hoessel, Saber A. Akhondi, Trevor Cohn, Timothy Baldwin, and Karin Verspoor. Chemu: Named entity recognition and event extraction of chemical reactions from patents. In Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins, editors, **Advances in Information Retrieval**, pp. 572–579, Cham, 2020. Springer International Publishing.

[4] S Hessam M Mehr, Matthew Craven, Artem I Leonov, Graham Keenan, and Leroy Cronin. A universal system for digitization and automatic execution of the chemical synthesis literature. **Science**, Vol. 370, pp. 101–108, 2020.

[5] 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, p. 3601, 2020.

[6] Martha Palmer, Daniel Gildea, and Paul Kingsbury. The proposition bank: An annotated corpus of semantic roles. **Computational linguistics**, Vol. 31, No. 1, pp. 71–106, 2005.

[7] Kojiro Machi, Seiji Akiyama, Yuuya Nagata, and Masaharu Yoshioka. OSPAR: A Corpus for Extraction of Organic Synthesis Procedures with Argument Roles. **Journal of Chemical Information and Modeling**, Vol. 63, No. 21, pp. 6619–6628, 2023.

[8] Sébastien R Goudreau, David Marcoux, and André B Charette. Synthesis of dimethyl 2-phenylcyclopropane-1, 1-dicarboxylate using an iodonium ylide derived from dimethyl malonate. **Organic Syntheses**, Vol. 87, pp. 115–125, 2003.

[9] Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of English: The Penn Treebank. 1993.

[10] *Organic Syntheses*. http://www.orgsyn.org. (accessed October 14, 2021).

[11] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. brat: a web-based tool for NLP-assisted text annotation. In **Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics**, pp. 102–107, Avignon, France, April 2012. Association for Computational Linguistics.

[12] Jiang Guo, A Santiago Ibanez-Lopez, Hanyu Gao, Victor Quach, Connor W Coley, Klavs F Jensen, and Regina Barzilay. Automated chemical reaction extraction from scientific literature. **Journal of Chemical Information and Modeling**, Vol. 62, pp. 2035–2045, 2021.

[13] Kojiro Machi and Masaharu Yoshioka. Hukb at chemu 2022 task 1: Expression-level information extraction. Vol. 3180, pp. 797–807. CEUR-WS.org, 2022.

[14] Sheshera Mysore, Edward Kim, Emma Strubell, Ao Liu, Haw-Shiuan Chang, Srikrishna Kompella, Kevin Huang, Andrew McCallum, and Elsa Olivetti. Automatically extracting action graphs from materials science synthesis procedures. **arXiv**, 11 2017.

[15] Kohei Makino, Fusataka Kuniyoshi, Jun Ozawa, and Makoto Miwa. Extracting and analyzing inorganic material synthesis procedures in the literature. **IEEE Access**, Vol. 10, pp. 31524–31537, 2022.