文献からの有機合成手順の自動抽出と専門家によるその結果の編集作業を支援する枠組み

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

文献から有機合成手順を自動抽出するために、ルールベースの手法や生成系大規模言語モデル（GLLM）を用いた方法が提案されている。しかし、自動抽出の結果が正しいとは限らないため、再現性や安全性の観点から、専門家による結果の確認と修正作業は必要不可欠である。本稿では、我々が提案した、専門家による確認と修正作業を支援するための枠組みについて紹介する。本枠組みでは、意味役割を用いた有機合成手順のアノテーションを行うことで視覚的に確認作業を支援し、ルールベースとGLLM ベースの 2 種類の性質の異なるシステムの出力を候補として提示することで修正作業を支援する。

1　はじめに

特許文書や論文などの文献に記述されている有機合成手順を再現可能なレベルで構造化することは、その手順を再利用するために重要である。例えば、自動合成ロボットに構造化された手順を与えて有機合成を行うことが可能となり、効率や再現性が向上するだけでなく、訓練を積んでいない専門家以外でも実験を行うことができる。

このような手順を表現するために、chemical description language（χDL、以下 XDL と記載）という XML 形式で記述された言語が提案されている [1]。XDL では、自動合成ロボットによる自動実験が可能なレベルの詳細な情報を記述することが可能であり、実際に XDL を元に自動実験を行なった例が報告されている [1, 2, 3]。

専門家により文献の有機合成手順を XDL に変換するという作業は、多くの時間と高いコストがかかる。そのため、ルールベースの情報抽出システム [1] や、生成系大規模言語モデル（GLLM）ベースの情報抽出システム [2] が提案されている。しかし、情報抽出の結果が正しいとは限らないため、再現性や安全性の観点から、専門家によるその結果の編集作業が必要不可欠である。

本稿では、専門家による編集作業を支援するために我々が提案した枠組み [4] について紹介する。[1)] 本枠組みでは、テキスト中の手順を表す部分について意味役割を用いてアノテーションすることで、有機合成手順に関する記述を確認することを支援する。さらに、アノテーション結果を用いたルールベースの自動変換手法の出力と、GLLM ベースの手法による出力を専門家に提示し、どちらかに誤りや不十分な点があった時にもう一方を参照することを可能とすることで、修正作業を支援する。

2　関連研究

本節では、GLLM ベースの XDL の自動変換手法として用いた CLAIRify [2] と、手順のアノテーションのスキーマとして用いたコーパス Organic Synthesis Procedures with Argument Roles（OSPAR）[5] について説明する。

2.1　CLAIRify

Yoshikawa らは、有機合成手順が記述されたテキストから XDL に自動変換することを行うシステムとして、GPT [6] を用いた GLLM ベースの CLAIRify を提案した [2]。CLAIRify では、XDL の仕様書と有機合成手順を入力として GPT に与える zero-shot で XDL の生成を行い、出力された XDL をルールベースの verifier にかけ、誤りがあった場合、verifier によるエラーメッセージを入力に加えて再度 XDL の生

1)　提案した編集システムは https://github.com/mlmachi/OSPAR_XDL/ で公開されている。

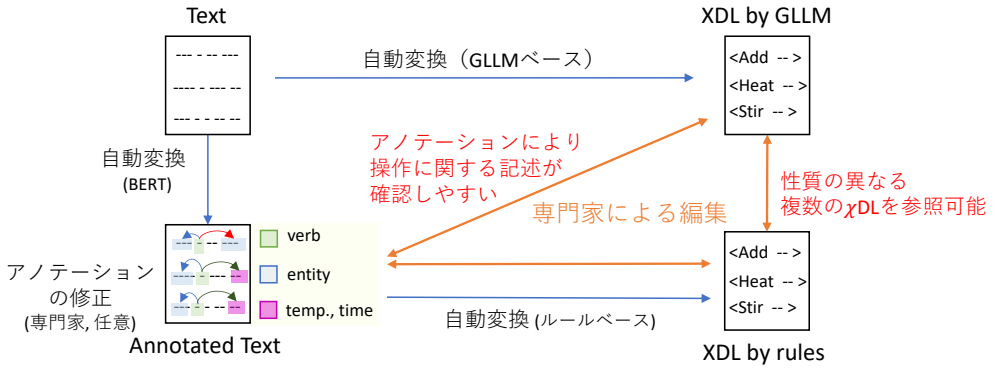

図 1　本枠組みの概要 (文献 [4] を元に再作成したもの)

成を行うということを誤りが無くなるまで一定の回数を上限として繰り返す。Yoshikawa らは、ルールベースの手法である SynthReader [1] と CLAIRify によって生成された XDL の出力のどちらの方がより好ましいかを化学の専門家らによって比較する実験を行った。評価の際に専門家らは再現率をより重視しており、SynthReader よりも適合率は低かったが再現率は高かった CLAIRify の出力が好まれる傾向があった。

2.2　OSPAR コーパス

OSPAR コーパスは、論文に記述された 112 件の有機合成手順に対して、文内の述語と項の関係を表す意味役割を用いることで、手順が構造的にアノテーションされたコーパスである [5]。アノテーションの例は、図 2 の左側に示されている。

OSPAR コーパスでは、操作を表す語、その対象となるもの、関連するパラメータのスパンがアノテーションされている。操作を表す語は REACTION_STEP としてアノテーションされている。化学物質、ガス、器具など、操作の対象となるものは ENTITY としてアノテーションされている。パラメータを表すラベルには、時間を示す TIME、温度を示す TEMPERATURE、その他の付加的な情報を示す MODIFIER などが存在する。

意味役割を表す手法として、ニュース記事を対象に作成されたコーパスである PropBank [7] 形式の意味役割セット (roleset) が用いられており、OSPAR コーパス中の手順を適切に表現する roleset が PropBank に含まれていない場合は、新たに roleset を定義している。OSPAR コーパスの roleset では、ARG1、ARG2、ARGM の 3 つのラベルが用いられている。ARG1 は主に被動作主を表し、ARG2 は必要に応じて roleset ごとに定義された動作主と被動作主以外の関係を表す。ARGM は roleset に依存しない時間や温度などのパラメータを表す。

3　自動抽出結果の編集システム

本節では、図 1 に示す本枠組みを実現する編集システムの概要と、そのシステムで用いるルールベースの自動変換手法について説明する。

3.1　編集システムの概要

本編集システムは、以下の 3 つの主要な機能から構成される。

- ChemBERT を用いたテキストから OSPAR 形式への変換
- ルールに基づく OSPAR 形式から XDL への変換
- CLAIRify を用いたテキストから XDL への変換

図 2 にユーザーインターフェースを示す。ChemBERT により変換された OSPAR 形式のアノテーション結果は、ウェブベースのアノテーションツールである brat [8] によって可視化され、必要に応じてユーザーがアノテーションの修正も行うことができる。各操作の roleset を見たい場合には、カーソルを REACTION_STEP に移動させると、その操作の roleset が表示される。2 つの変換された XDL は編集可能なテキストボックスに表示され、ユーザーは好ましい方をベースの XDL として編集を行い、誤りや不十分な点を発見した場合に、もう一方の XDL を参照することができる。

3.2　ルールベースの変換手法

テキストから OSPAR 形式への変換については、文献 [5] と同様に、ChemBERT [9] を OSPAR コーパスで学習させたシステムを用いた。OSPAR 形式か

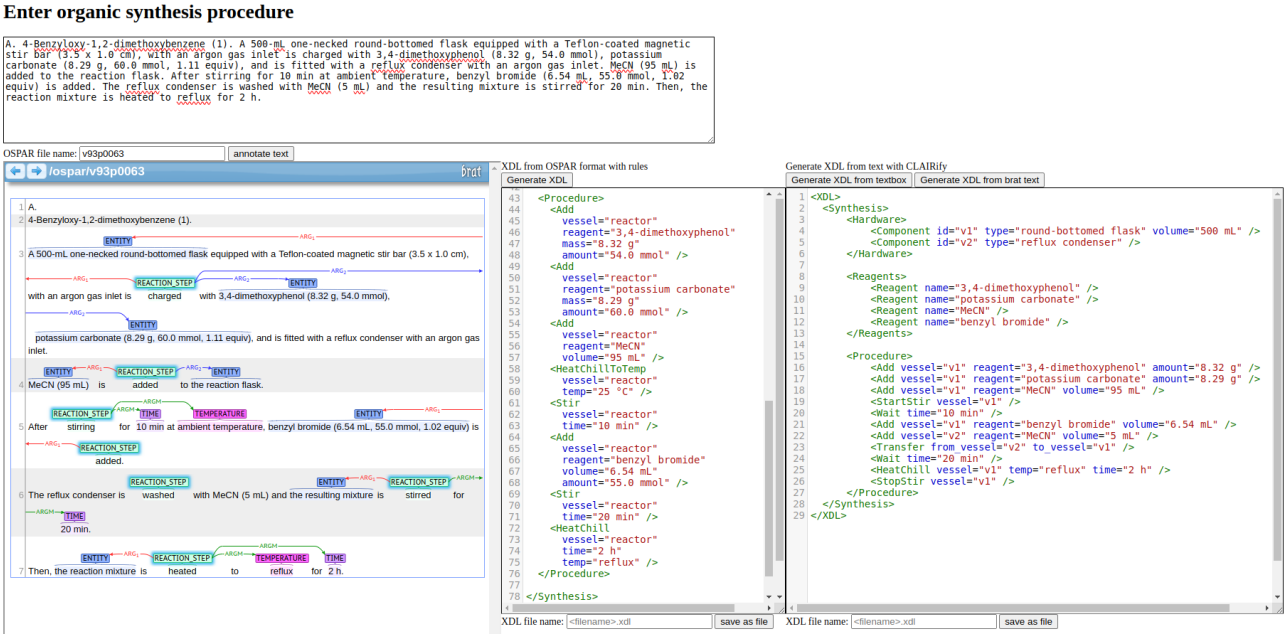

図 2　編集システムのスクリーンショット (文献 [4] より引用)

ら XDL への変換は、各 roleset に含まれる ARG1 や ARG2 を XDL の `Procedure` に含まれる操作（XDL action）に対応づけるルールと、ARGM を XDL の操作に対応づけるルールを作成することで行なった。操作を表す語については、WordNet [10] の WordNetLemmatizer を用いて正規化を行なった。OSPAR 形式の ARG1 や ARG2 の ENTITY は、化学物質名や質量などがまとめて 1 つのものとして扱われているため、ChemicalTagger [11] を使用して詳細な情報を取得した。TEMPERATURE や TIME といった ARGM が存在した場合、`HeatChillToTemp` などの温度を変更する XDL action を生成するために使用するか、roleset と関連する XDL action の属性（`time="1 h"` など）を定義するために使用した。

4　自動変換システムに関する実験

4.1　実験設定

ルールベースの提案手法により出力された XDL と GLLM ベースの CLAIRify により出力された XDL の両方を専門家に提示することの有効性を検証するために、有機合成手順 6 件に対してそれぞれの手法で XDL を作成し、評価を行なった。[2)]

本実験では、提案手法のテキストから XDL への変換までを行うもの（Pipeline）、人手でアノテーションされた OSPAR 形式[3)]から XDL への変換を行うもの（OSPAR2XDL）、および CLAIRify の 3 つの手法を比較した。Pipeline と OSPAR2XDL の 2 つを用いたのは、ChemBERT による OSPAR 形式のアノテーションの正確さが、変換された XDL の質にどの程度影響を与えるのかを調べるためである。CLAIRify の実装には、GitHub で公開されているもの [12] を使用し、CLAIRify で用いる GLLM として、本研究では GPT-4o（gpt-4o-2024-05-13）を使用した。

ルールベースと CLAIRify の 2 つのシステムを独立して実行し、それぞれの XDL action に正解が含まれているかを確認することで、2 つのシステムを組み合わせた際の再現率を調べた。誤りの度合いを評価するための再現率として、以下のような exact recall と action recall を用いた。

- exact recall: 正解データ中の XDL action のうち、属性も含めて完全に一致するものの割合。[4)]
- action recall：正解データ中の XDL action のうち、一つ以上の属性が正しい XDL action の割合。

exact recall と action recall のそれぞれの基準による評価の例は、図 3 に示す通りである。

文献に記述されている有機合成手順には、明示的に書かれていない暗黙的な操作が存在する。[5)] テキストからの情報抽出という観点から考えた時に、こ

2)　正解データは、有機合成化学の専門家を含む著者らが作成した。
3)　OSPAR コーパスに含まれる有機合成手順をアノテーションの正解例として用いた。
4)　ただし、vessel に関しては、それぞれの XDL で定義されているものが正しく利用されていれば良いものとした。
5)　例として、複数の物質を足し合わせる操作が書かれている時、基本的には撹拌する操作も行われるが、撹拌の操作は明示的に書かれない場合がある。

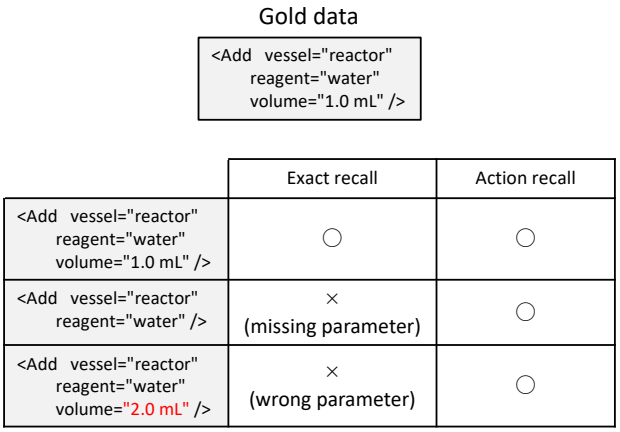

図 3 exact recall と action recall による評価の例 (文献 [4] より引用)

れらの操作は性質が異なるため、それぞれ分けて評価を行った。

4.2 結果・考察

4.2.1 明示的な操作

表 1 に明示的な XDL action の再現率を示す。各

表 1 明示的な XDL action の再現率

	CLAIR	CLAIR +Pipe	CLAIR +O2X	Pipe	O2X
exact	38/65	48/65	49/65	28/65	31/65
action	60/65	60/65	60/65	41/65	50/65

*表の値は、文献 [4] より引用

数字は、(正解と判定された数)/(正解データの数) を表している。CLAIR は CLAIRify、Pipe は Pipeline、O2X は OSPAR2XDL を示している。

ルールベースの提案手法と CLAIRify の結果を組み合わせることで、exact recall が向上することを確認した。一方で、Pipeline と OSPAR2XDL をそれぞれ CLAIRify と組み合わせた結果の action recall は向上しなかった。その理由として、CLAIRify は action recall では 60/65 という高い値を示したが、exact recall では 38/65 まで低下したことが挙げられる。これは、CLAIRify が化学物質の量やパラメータの抽出に失敗していたことが主な原因である。有機合成手順の文書ごとに見ると、6 件のうちの 2 件において一貫してこのような誤りが見られたが、他の 4 件では一貫して変換が行われていた。したがって、プロンプトに含まれる XDL の仕様書の部分を改善し、変換の対象とするものの基準を制御することで、文書ごとに基準が変わることを抑制できる可能性がある。

Pipeline と OSPAR2XDL の比較により、テキストから OSPAR 形式のアノテーションへの自動変換については、改善の余地があることを確認した。主な原因として、ChemBERT が化学物質名の抽出に失敗することがある点が挙げられる。さらに、エンティティの境界の認識ミスによる失敗例も見られた。テキストから OSPAR 形式への情報抽出システムを改善するには、学習データを増やしたり、深層学習モデルを改良する必要がある。そのためには、ユーザーが編集したアノテーション結果を蓄積させることで学習データを増やし、それを学習データとして用いたシステムを作成することで、情報抽出の性能を向上していくというサイクルを回していくことが有効だと考えている。

4.2.2 暗黙的な操作

表 1 に暗黙的な XDL action の再現率を示す。

表 2 暗黙的な XDL action の再現率

	CLAIR	CLAIR +Pipe	CLAIR +O2X	Pipe	O2X
exact	6/12	6/12	6/12	0/12	0/12
action	6/12	6/12	6/12	0/12	0/12

*表の値は、文献 [4] より引用

CLAIRify は暗黙的な操作を補完できる場合があり、暗黙的な操作に対しても GLLM が有望であると考えられる。ルールベースの手法は、明示的な操作のみ考慮したルールを書いていたため、暗黙的な操作を抽出することができなかった。このような操作を考慮するためには、例えば `Add` の後には `Stir` するというようなルールを追加するということが考えられる。実際に専門家が編集を行う際には、明示的に書かれているアノテーションされたテキストだけでなく、暗黙的な操作についても考慮しながら作業をする必要があることがわかった。

5 おわりに

本稿では、我々が提案した、文献からの有機合成手順の自動抽出と専門家によるその結果の編集作業を支援する枠組みについて紹介した。実験を通して、複数の XDL を提示することの有効性を検証したが、6 件という少量の有機合成手順に対してのみの評価であるため、今後はより大規模なデータで検証を行うことが必要である。手順のアノテーションによる効果について、現状は「視認性が高いと手順を理解しやすい」という専門家のコメントによる評価のみであるため、今後は具体的な応用を通してその有効性を検証していく必要がある。

謝辞

本研究は JSPS 科研費 JP23H03810、JP23K18500 および、JST 次世代研究者挑戦的研究プログラム JPMJSP2119 の支援を受けたものである。また、本研究の一部は、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] 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.

[2] Naruki Yoshikawa, Marta Skreta, Kourosh Darvish, Sebastian Arellano-Rubach, Zhi Ji, Lasse Bjørn Kristensen, Andrew Zou Li, Yuchi Zhao, Haoping Xu, Artur Kuramshin, et al. Large language models for chemistry robotics. **Autonomous Robots**, Vol. 47, No. 8, pp. 1057–1086, 2023.

[3] Kristine Laws, Marcus Tze-Kiat Ng, Abhishek Sharma, Yibin Jiang, Alexander JS Hammer, and Leroy Cronin. An autonomous electrochemical discovery robot that utilises probabilistic algorithms: Probing the redox behaviour of inorganic materials. **ChemElectroChem**, Vol. 11, No. 1, p. e202300532, 2024.

[4] Kojiro Machi, Seiji Akiyama, Yuuya Nagata, and Masaharu Yoshioka. A framework for reviewing the results of automated conversion of structured organic synthesis procedures from the literature. **Digital Discovery**, pp. –, 2025.

[5] 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.

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, **Advances in Neural Information Processing Systems**, Vol. 33, pp. 1877–1901. Curran Associates, Inc., 2020.

[7] 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.

[8] 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.

[9] 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.

[10] George A Miller. Wordnet: a lexical database for english. **Communications of the ACM**, Vol. 38, No. 11, pp. 39–41, 1995.

[11] Lezan Hawizy, David M Jessop, Nico Adams, and Peter Murray-Rust. Chemicaltagger: A tool for semantic text-mining in chemistry. **Journal of Cheminformatics**, Vol. 3, pp. 1–13, 2011.

[12] Clairify. https://github.com/ac-rad/xdl-generation. (accessed March 18, 2024).