



Title	A Framework for Extraction of Structured Organic Synthesis Procedures from the Literature [an abstract of dissertation and a summary of dissertation review]
Author(s)	町, 光二郎
Degree Grantor	北海道大学
Degree Name	博士(情報科学)
Dissertation Number	甲第16503号
Issue Date	2025-03-25
Doc URL	https://hdl.handle.net/2115/95008
Rights(URL)	https://creativecommons.org/licenses/by/4.0/
Type	doctoral thesis
File Information	Kojiro_Machi_abstract.pdf, 論文内容の要旨



学 位 論 文 内 容 の 要 旨

博士の専攻分野の名称 博士（情報科学） 氏名 町 光二郎

学 位 論 文 題 名

A Framework for Extraction of Structured Organic Synthesis Procedures from the Literature

（文献から構造化された有機合成手順を抽出するための枠組み）

有機合成化学の文献において、有機合成手順（有機化合物の化学反応手順）は、既存のデータを活用したデータ駆動型の化学研究への応用には適していない散文（非構造化データ）の形式で公開されている。有機合成手順を構造化して再利用・共有するために利用可能な技術として、機械可読な化学反応手順の記述言語である χ DL (Chemical Description Language) が提案されている。さらに、テキストから χ DL への自動変換手法として、ルールベースの手法や生成 AI ベースの手法が提案されている。しかし、これらの手法は完全ではなく、情報が不十分であったり、誤った情報を含む χ DL が生成されてしまうことがある。そのため、より正確な χ DL を得るためには、自動変換を行なった後に、人間の専門家による χ DL の確認と修正の作業が必要不可欠である。

本研究の目的は、文献から自動変換により構造化された有機合成手順に対する専門家による確認と修正の作業を支援することである。そのために、確認と修正の作業それぞれを支援する方法を提案し、それらを実現する編集環境の開発を行う。確認作業を支援するためには、テキスト中の有機合成手順を表す部分を抽出し可視化する方法を提案する。具体的には、意味役割セットを付与した OSPAR (Organic Synthesis Procedures with Argument Roles) コーパスを提案するとともに、本コーパスを利用した深層学習ベースの自動アノテーションを行うシステムを作成する。修正作業を支援するためには、異なる特性を持つ複数の手法による自動変換結果を並べて専門家に提示する方法を提案する。具体的には、異なる特性を持つ手法として、既存の生成 AI ベースの手法と、本研究で提案するアノテーション結果を利用したルールベースの手法の 2 つを利用する。

本論文は、全 6 章で構成される。第 1 章では、本研究の背景と目的、本論文の構成について述べた。第 2 章では、本研究における有機合成手順の可視化のためのアノテーションのスキーマを考案するための基礎となった、特許文書からの化学反応情報抽出を行うためのタスクである ChEMU タスクで用いられたコーパスのスキーマについて説明した。さらに、化学反応情報抽出に関連する既存手法について述べるとともに、本研究で有機合成手順を記述するために使用した χ DL について説明した。

第 3 章では、本研究における有機合成手順の抽出のための基盤となる ChEMU のタスクを解くために開発したシステムについて説明した。本システムでは、化学分野向けの深層学習ベースの言語モデルである ChemBERT に対して、ChEMU コーパスを訓練データとして用い、固有表現抽出と関係抽出を行う手法を提案した。

第 4 章では、テキスト中の有機合成手順を表す部分を抽出し可視化する方法を提案した。そのために、有機合成化学の専門家 2 名とともに、意味役割セットを用いたアノテーションのスキーマの定義とそれに基づいた OSPAR コーパスの作成を行った。OSPAR コーパスでは、意味役割セットのベースとしてニュース記事を対象とした既存のコーパスである Proposition Bank を用いたため、有機合成を中心とした化学反応に関連する語彙に関する意味役割セットの追加もコーパスの作成

と合わせて行うこととした。論文誌 *Organic Syntheses* に掲載された 112 件の有機合成手順をアノテーションの対象として、新たに定義した 19 個を含む 48 個の意味役割セットを利用したコーパスを作成した。OSPAR コーパスを用いて ChemBERT の学習と評価を行った結果、評価用データに対して、固有表現抽出の F 値が 0.8709 となり、固有表現抽出の結果を用いた関係抽出の F 値が 0.8393 となった。この値は、改善の余地はあるものの、このシステムを以降の枠組みに用いることとした。

第 5 章では、専門家による有機合成手順の自動変換結果の確認と修正の作業負担を軽減する枠組みを提案した。修正作業を支援するために、 χ DL の編集作業においては変換されていない手順を追加するコストが不要な手順を削除するコストよりも高いことに注目し、一定以上の精度を有する異なる特性を持つ手法による自動変換結果を並べて専門家に提示する方法を提案した。異なる特性を持つ手法として、既存の生成 AI ベースの手法と、本研究で提案するアノテーション結果を利用したルールベースの手法の 2 つを利用した。実験においては、6 件の有機合成手順に対して χ DL を作成し、2 つのシステムそれぞれから得られた χ DL を比較し、それぞれの χ DL で欠損している情報がもう一方に含まれている事例があることを確認することで、複数の手法で作成した χ DL を表示することの有効性を確認した。また、作成した自動変換結果の確認作業を支援するために、第 4 章で提案したアノテーションを利用することで、テキスト中の有機合成手順を表す語とそれらの意味的な関係をあわせて可視化する方法を提案した。また、本枠組みに基づく χ DL の確認と修正の作業を行えるようなインターフェースを開発し、オープンソースソフトウェアとして公開した。

第 6 章では、本研究の成果を総括し、今後の展望について述べた。

以上のように、本研究は、今後の有機合成化学における有機合成手順の構造化を通じた再利用・共有を目的としたテキストからの有機合成手順の抽出において、専門家による確認と修正の作業が必要である点について指摘し、その作業の負担を軽減しながら構造化された有機合成手順である χ DL を記述するための枠組みとして、アノテーションされたテキストと、自動生成された χ DL を組み合わせて提示する枠組みを提案した。本研究の成果は、有機合成化学の有機合成手順に関するデータを効率的に収集する方法について、情報科学の観点から提案するものであり、有機合成化学の分野の発展ならびに情報科学と化学分野の連携を深めることに貢献する。