

材料レシピ：構文解析を使った論文からのプロセスの自動抽出と構造化

Materials Recipe: Automatic extraction and structuring of material processes from academic papers using syntax parsing

物材機構 高山英紀, °石井真史

NIMS, Eiki Takayama and °Masashi Ishii

E-mail: ISHII.Masashi@nims.go.jp

【序】 マテリアルズ・インフォマティクス (MI) が材料開発における新しい手法として注目される中で、その基盤としてのデータの重要性が高まっている。多くの MI では、材料の原子レベルの化学構造と特性の間で、何らかの法則性 (モデル) を機械学習などから見出し、そのモデルを使ってより高性能の未知材料を予測するが、実際の材料は巨視から微視まで様々なスケールの構造に加え、添加物あるいは混合状態などが特性を決定しており、これらの多種多様な要素を制御するプロセスが材料開発においてきわめて重要になる。従って、実用的な MI は、プロセスを含めたデータ駆動が必須となろう。こうした背景の下、本研究「材料レシピ」では、学術論文からプロセス情報の自動抽出と、その機械可読化のためのデータ構造設計を行う。

【手法】 今回は高分子を対象として、「材料レシピ」の自動作成を試みた。入力は、学術論文 (American Chemical Society, “Macromolecules” (2016) 973 論文) または NIMS にて構築・運用している PoLyInfo [1]で自由記述されたプロセス情報である。「材料レシピ」自動作成のフローは図 1 の通りである。概要としては

- (1) プロセスに関する受動態形動詞辞書 (自作) を使い、対象の文章を文書内から抽出
- (2) (1)の文章を単文に分割
- (3) 自然言語処理オープンソース・ライブラリ spaCy[2]を使った係り受け解析により、(1)の受動態形動詞に紐づく前置詞を特定
- (4) 前置詞の目的句からプロセスパラメータをルールベースで抽出
- (5) (4)を機械可読化した 13 の主要なプロセス素過程のスキーマに当てはめ、xml (eXtensible Markup Language) 形式でデータ構造化

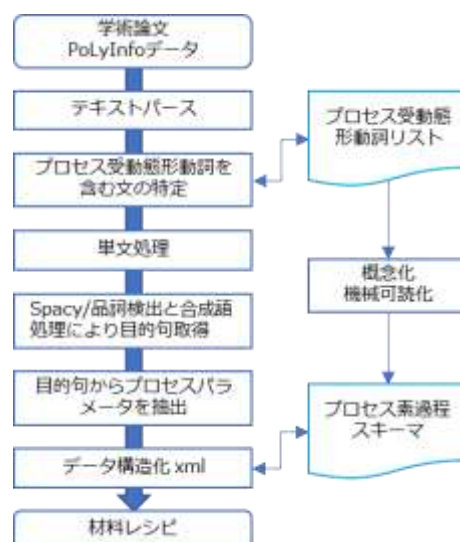


Fig.1 Automatic formation procedure of “Materials recipe”

となる。

【結果】 プロセスを示す文章” Polymer powder was dried at 40 ° C under reduced pressure for 48 h “を本手法でレシピ化し、表形式で表すと

Verb	Temp	Time	Pres
dry	at 40 ° C		under reduced pressure
dry		for 48 h	

となる。第一行は at の目的句が” under reduced pressure”までを含んでいる事を示す。現在のところ、PoLyInfo に収録されている、事前に編纂され短文化されたプロセスの記述は、ほぼ完全な構造化が可能であるのに対し、原著論文からの直接的な構造化は、表現の多様性から依然多くの課題が残っている。当日は、機械可読化した 13 のプロセス素過程のデータ構造を含めて、「材料レシピ」自動作成の全体を俯瞰する。

【参考文献】

[1] <https://polymer.nims.go.jp/>

[2] <https://spacy.io/>