第81回応用物理学会秋季学術講演会 講演予稿集 (2020 オンライン開催)

文章からのポリマー良溶媒の抽出

Extraction of Polymer Good Solvents from Sentences

物材機構[1], 奈良先端大[2], 理研 AIP[3]
○岡 博之[1], 佐藤 義貴[2], 近藤 修平[2], 進藤 裕之[2,3], 松本 裕治[3], 石井 真史[1]
NIMS[1], NAIST[2], RIKEN AIP[3]
○Hiroyuki Oka[1], Yoshitaka Sato[2], Shuhei Kondo[2], Hiroyuki Shindo[2,3], Yuji Matsumoto[3], Masashi Ishii[1]
E-mail: OKA.Hiroyuki@nims.go.jp

1 序

NIMS では長年にわたりポリマーデータを学術論文から収集し、データベース[1]として公開するとともにマテリアルズ・インフォマティクス(MI)への活用を進めている。これまでデータ収集は人手で行ってきたが、近年機械学習などによる自動収集を研究しており、その一環として文章からのポリマー良溶媒の抽出を試みている。既報では論文中でポリマーの良溶媒を含んでいる文(以下、「ポリマー良溶媒文」)を、機械学習による自然言語処理(BERT)で、F 値 0.8 以上の良い精度で分類できた[2]。ここでは、soluble in ～などの直接的な表現を持つ文に加え、GPC 測定に関する実験手順の説明等、使用した試薬が良溶媒であることを間接的に示す文も分類できている。今回は、ポリマー良溶媒文からのポリマー名と良溶媒名の抽出を議論する。実際の文では具体的なポリマー名を省略する場合が多く、本研究では、文に加えて関連する表の解析も併用して精度の向上を試みた。

2 実験

データ抽出には Macromolecules(2002-2007 年)の 100 報(PDF 形式)を用いた。この論文からポリマー名とその良溶媒名の組合せを人手により 380 セット選び出し、正解データとした。これらを含むポリマー良溶媒について、以下に示すアルゴリズムを適用し、抽出結果と正解データを比較した。

ポリマー良溶媒文は表 1 に示す 7 つのタイプ(S1～S7)に分類できる。S1 は直接的な表現であるために良溶媒を比較的容易に特定できる。S2 の文では良溶媒と貧溶媒の区別を伴い、S3～S7 の文では表のポリマー名を認識する複合的な手順を必要とする。表の解析については、表キャプションの文字列マッチングや PDF 中の文字座標などを利用して行った。ポリマー良溶媒文中にポリマー名があれば、それを優先的に認識した。ポリマー名の認識は既報のルールベースまたはポリマー名を学習させた機械学習によって行った[3]。溶媒名については、独自に編集した溶媒名辞書を用いた文字列マッチングで行った。ポリマー名と良溶媒名の抽出では、同じポリマー良溶媒文から認識できた両者の組合せすべてを抽出した。尚、S2 の文では、from や into などの前置詞および良溶媒が先行することを利用して、貧溶媒との区別を行った。

表 1 ポリマー良溶媒文のタイプ分け

	文タイプ	表現例
S1	直接的な溶解性表現	be soluble in, dissolve ～ in, good solubility, good solvent
S2	再沈殿	precipitated from ～ into
S3	GPC 測定	GPC, SEC
S4	NMR 測定	NMR
S5	粘度測定	The intrinsic viscosity was ～
S6	フィルム作成	cast from, spin-coat
S7	ブレンド作成	Blends were prepared from

3 結果と考察

本提案手法を評価したところ、Precision、Recall、F 値はそれぞれ 0.16、0.23、0.19 であり、総じてかなり低い値であった。Recall の低さはポリマー名認識精度の低さが主な原因であった。特に複数語や代名詞でポリマー名が記述された文では認識率が低く、改善が必要である。また重文や複文により、複数のポリマー名と溶媒名が記載された場合の誤抽出が多く、単文への事前分割が精度を改善することが示唆された。

4 まとめ

現時点での抽出精度はかなり低いが、単文への自動分割などの前処理の追加を検討しており、当日はその効果をアルゴリズムの詳細と共に報告する。

[1] PoLyInfo, https://polymer.nims.go.jp/
[2] 学術論文中のポリマー良溶媒文の分類, 岡博之他, 応用物理学会第 67 回春季学術講演会, 上智大学(四谷, 東京), 2020 年 3 月 12-15 日.
[3] H.Oka et al., “Automatic extraction of polymer data from tables in xml”, Third International Workshop on SCIentific DOCument Analysis (SCIDOCA2018), 慶応義塾大学(日吉, 横浜市), 2018. 11. 12-13.