科学学術論文の表を対象としたポリマーの物性情報の抽出

Information Extraction of Polymer Properties from Tables in Scientific Papers

奈良先端大[1]，理研 AIP[2]，物材機構[3] ○進藤 裕之[1,2]，岡 博之[3]，石井 真史[3]，松本 裕治[1,2]

NAIST[1], RIKEN AIP[2], NIMS[3], ○Hiroyuki Shindo[1,2], Hiroyuki Oka[3], Masashi Ishii[3], Yuji Matsumoto[1,2]

E-mail: shindo@is.naist.jp

情報科学を活用した効率的な材料開発には，科学技術論文から過去の研究における実験データを収集し，体系化されたデータベースとして情報を蓄積していくことが重要である．高分子材料分野では，ポリマーの物性，化学構造，測定方法などの情報を学術論文から人手で収集したデータベースとして PoLyInfo[1]が公開されている．しかしながら，日々出版される論文数は膨大であり，網羅的なデータベースを人手で構築・管理するには膨大な時間と手間を要する．そこで本研究では，画像処理および自然言語処理技術を用いて，PDF の論文から表を同定し，表に含まれるポリマーの物性値を自動収集することを試みる．PDF は非構造化データであり，図表の位置や構造に関する情報を保持していないため，統計的に推定する必要がある．

具体的な手順は以下の通りである．まず，深層学習に基づく画像認識技術を用いて，PDF の論文から表の領域を同定する．画像認識モデルは，ResNeXt-101 を TableBank データ[1]で学習したものを用いた．次に，表の領域に含まれる文字と位置情報を PDF から抽出し，文字の相対的な位置関係から行と列を決定して，表の構造化を行う．図 1 に表の領域認識と構造解析の全体の流れを示す．本手法により，論文の PDF を入力として，PDF に含まれる表の XML データを得ることができる．

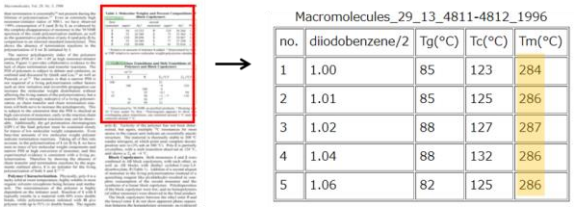

図 1　表の領域認識と構造解析

Figure 1　Table recognition and parsing

本手法の有効性を確認するために，111 本の論文誌（Macromolecules）から，ポリマーの Tm（融点）の情報抽出を行った．具体的には，上記の手法で論文から表の領域抽出と構造化を行った後，“Tm”という文字列を含む行または列に含まれる全ての数値を抽出結果とした．評価は，PoLyInfo を正解として，抽出結果の正答率と再現率を計算した．表 1 に示されるように，抽出結果の約 60%が PoLyInfo に含まれるという結果であった．詳細を確認すると，残りの 40%のうち，表の解析エラーは 1 件しかなく，PoLyInfo が論文中の全ての Tm 値を掲載しているわけではないことが主な要因であった．したがって，実際の情報抽出の性能は極めて高いという結果であった．

表 1　Tm の抽出結果

Table 1　Experimental results of Tm extraction

	正答率	再現率
Tm の抽出性能	0.604	0.731

参考文献

[1] Li, Minghao et al., “TableBank: Table Benchmark for Image-based Table Detection and Recognition”, arXiv preprint arXiv:1903.01949, 2019.

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