# 実験データを持つポリマー論文の分類

## Classification of polymer articles bearing experimental data

物材機構　○岡 博之, 石井 真史
NIMS　○Hiroyuki Oka, Masashi Ishii

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、表からの自動データ抽出が行えるようになった[2]。そのため、多くの材料系論文の表からポリマーデータの自動抽出を進めているが、ただ、この抽出では計算機シミュレーションなどで得られた計算値も抽出しており、MI 用データとしては実験値のみが望ましい。このため、あらかじめ計算機シミュレーションなどが主題の論文(非実験論文)は除外し、実験論文のみを分類しておくのが良いが、これについてはまだ行えていない。また、表からの自動データ抽出ではある程度の誤抽出があるが、金属材料のようにポリマー以外の材料を主題としている論文(非ポリマー論文)ではそれが多い。これはポリマー名以外の材料名を頻繁に誤認識しているためであるが、ポリマー名認識の精度を今以上に向上させるのは簡単ではなく、そのため、非ポリマー論文はあらかじめ除外し、ポリマー論文のみを分類しておくのが良い。これについてもまだ行えておらず、そこで本研究ではこれら2つの論文分類を検討し、実験データを持つポリマー論文を分類することを目指した。それぞれの論文分類を個別に検討したので、その手法ならびに得られた結果を報告する。

2. 実験

ポリマー論文分類では、アメリカ化学会(ACS)の非ポリマー系 6 雑誌から選んだ 180 論文とエルゼビアの非ポリマー系 5 雑誌から選んだ 150 論文を用いて実験を行った。分類方法は、アブストラクト中でポリマー名の認識を行い、略称名以外のポリマー名(poly から始まる IUPAC 名や cellulose などの慣用名)が認識された場合はその論文をポリマー論文と推定した。ポリマー名の認識は以前に報告したルールベース手法によって行った[2]。実験論文の分類では、ACS の Macromolecules から選んだ 400 論文を用いた。分類は、実験論文と非実験論文での特徴的な単語表現を利用して行った。実験論文では、セクションタイトルや図キャプションで”Experiment”や”synthesis” などの表現が頻繁に使われており、一方、非実験論文では、”simulation”や”computation”などの表現が多い。非実験論文では数式が多いことも特徴で、これも利用して行った。これらの表現および数式の認識はルールベース手法によって行った。分類結果の出力は各論文に対して”1”または”0”のラベル付与を行い(”1”は正例を意味し、ポリマー論文および実験論文に対して付与)、事前作成の正解ラベルと比較することで、TP(true positive)、FP(false positive)および FN(false negative)を判定した。そして、これらの数から precision(TP / TP + FP)、recall(TP / TP + FN)および F 値(2 · precision · recall / precision + recall)を算出して分類の評価を行った。

3. 結果と考察

ここではポリマー論文分類についてのみ記載する。分類の評価値は、ACS 論文で precision、recall および F 値がそれぞれ 0.803、0.813 および 0.808、エルゼビア論文で 0.580、1 および 0.734 であった。ACS の方では評価値がすべて 0.8 付近と良い結果であった。一方、エルゼビアは recall が 1 であるが、precision が 0.6 以下と低く、FP が多かった。この原因は、二次電池関連の論文で、無機活物質を主題した非ポリマー論文が多かったが、電極で使用されるバインダーなどのポリマー名がアブストラクト中に記載されていることが多く、それによって誤分類を多くしていた。このことから、論文内容の認識なども必要であることが分かった。

4. まとめ

現在、分類精度を高める方法を検討しており、ルールベース手法以外に、機械学習を利用することも視野に入れている。当日はこれらの結果、また、実験論文分類についての結果も併せて報告する。

[1] PoLyInfo, https://polymer.nims.go.jp/
[2] 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.