学術論文からのポリマーの溶解性に関わる溶媒名の自動抽出

Automatic Extraction of Solvent-names on Polymer's Solubility from Academic Articles

物材機構¹, 奈良先端大², 理研 AIP³

○岡 博之¹, 吉澤 篤志¹, 進藤 裕之^{2,3}, 松本 裕治^{2,3}, 石井 真史¹

NIMS¹, NAIST², RIKEN AIP³

○Hiroyuki Oka¹, Atsushi Yoshizawa¹, Hiroyuki Shindo^{2,3}, Yuji Matsumoto^{2,3}, Masashi Ishii¹

E-mail: OKA.Hiroyuki@nims.go.jp

1. 序

NIMS では以前からポリマーデータを学術論文から収集しており、それらをデータベースで管理・公開し^[1]、また Materials Informatics に活用している。データ抽出は人手で行っているが、最近の出版論文数の増加、テキストマイニング技術の向上などから、コンピュータを用いた自動抽出も試みている。ポリマーデータの多くは論文中の図や表にまとめられていることが多いが、ポリマーの溶解性を示す溶媒については本文に記載されていることが多い。そこで本研究では、その溶媒名を自動抽出するために、本文中のポリマー名と溶媒名の関係抽出を行う方法を検討した。これらの固有表現を含む文章には”dissolve”や”soluble”などの溶解性を示唆する単語が記載されていることが多く、また、ポリマーを溶媒に溶解させることを前提とした作業(粘度測定やフィルム作成など)に関する文章では、良溶媒名を特定することができる。本研究では、ポリマー論文に見られるこのような文表現傾向を考慮してルールベースによる関係抽出を行った結果について報告する。

2. 実験

2-1. 使用論文およびプログラム言語

論文は英論文である Macromolecules (出版社: American Chemical Society, 2016 年分)を使用した。ファイルは xml 形式のものを使用した。プログラム言語は Python 3 を使用した。

2-2. 評価用データの作成

自動抽出精度を調べるために、評価用データを人手でアノテーションすることで作成した。文章中のポリマー名とその溶解性に関わる溶媒名にラベル付けを行い、さらに、これらの間を関係づけるラベル付けも行った。これらの作業では HTMLAnno^[2]を用いた。アノテーションは溶解性に関する内容が記載されている段落のみに行い、合計で 199 段落分(ポリマー名と溶媒名を関係づけた数は 392)の評価用データを作成した。

2-3. ルールベースによるポリマー名 - 溶媒名の関係抽出

本研究では、ポリマー名および溶媒名の固有

表現認識の検討は行わず、評価用データのラベルを用いて、2つの固有表現間の関係を自動で取ることを主に行った。これをルールベースで行うルールとして、1 文内にポリマー名および溶媒名を含み、かつ、ポリマーの溶解性に関連する単語を含んでいる場合に、その 1 文内のポリマー名と溶媒名はすべて関係がある、とした。なお、1 文の範囲はピリオドからピリオドまでとし、また、前述の”dissolve”や”soluble”に加え、”spin-coat”や”viscosity”など、ポリマーの溶解に関わる作業を表す語をあらかじめリスト化して溶解性に関連する単語として用いた(合計で 18 単語)。

3. 結果と考察

評価用データを用いて、ルールベースで関係抽出を行った結果、Precision、Recall および F 値はそれぞれ 0.577、0.913 および 0.707 であった。Recall が高く、Precision が低いことから、False Negative (FN)は少ないが、False Positive (FP)は多い結果であった。FPが多くなった原因の一つとして、ピリオドでの文区切りが良くないと考えられた。文途中で”and”や”but”などの接続詞が入ると、それ以降、溶解性に関する文章でなくなっている場合があり、このようなとき、その部分に記載の溶媒名は誤って抽出されていた。

4. まとめ

学術論文からポリマーの溶解性を示す溶媒名を自動抽出するために、本文中のポリマー名と溶媒名の関係抽出を行った。溶解性に関して、ポリマー論文でよく見られる文表現傾向を考慮して、ルールベースで関係抽出を行ったところ、F 値が約 0.7 の結果が得られた。Precision は 0.6 以下と低い値であったが、文区切りなどのルールの改善により、さらに良い精度で抽出が行えると考えられた。現在、その改善を行っており、発表当日はその結果も含めて報告を行う予定である。

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

[2] NAIST 開発アノテーションツール, <https://github.com/paperai/htmlanno>