

学術論文中のポリマー良溶媒文の分類

Sentence Classification for Polymer's Good Solvents in Academic Articles

物材機構¹, 奈良先端大², 理研 AIP³

○岡 博之¹, 佐藤 義貴², 近藤 修平^{2,3}, 進藤 裕之^{2,3}, 松本 裕治^{2,3}, 石井 真史¹

NIMS¹, NAIST², RIKEN AIP³

○Hiroyuki Oka¹, Yoshitaka Sato², Shuhei Kondo^{2,3}, 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、データベースでの管理^[1]およびマテリアルズ・インフォマティクス(MI)^[2]への活用を進めている。データ収集は人手で行っているが、最近の出版論文数の増加や MI での大量データの必要性から、コンピュータによる自動収集も検討している。数値データについては論文中、図・表にまとめられていることが多いが、ポリマー良溶媒(または貧溶媒)については、本文に記載されていることが多く、これらを自動収集しデータベース化するには、本文中のポリマー名と溶媒名の関係抽出が必要である。この関係抽出を以前にルールベースで検討したところ、F 値で 0.71 まで達成できた^[3]。ただ、このときは問題を簡単にするために、1 文内にポリマー名と溶媒名が記載されていることを前提としていた。実際の人手によるデータ収集では、溶媒が記載されている文で“the polymer”のような代名詞が使われていても、別の文からポリマー名を特定し収録している。また、ポリマー試料の平均分子量を得る際の GPC や NMR などの測定に使われた溶媒を良溶媒と判断して収録する場合もある(含意による収録)。これらのことから、人手と同様のデータ収集を行うには、ポリマー名の記載がなくても良溶媒を記載した文を分類することが必要である。そこで本研究では、この分類手法としてルールベースと機械学習の両方を検討した。ルールベースの手法はポリマー論文で頻出する溶解性に関する表現をリスト化することで行える。一方、機械学習の手法は学習次第で多様な表現が高精度で分類できることから、自動データ収集を加速すると期待される。

2. 実験

実験用データは高分子データベースの人手収集で利用された文を正例文として用い、それに負例文を追加して作成した。それぞれ 99 および 181 文はいずれも Macromolecules 誌(2005-2009 年)を対象とした。ルールベースによる文分類の詳細は既報^[3]の通りであるが、大まかには次の 3 つのタイプの文をルールにより分類した:(1)直接溶解性を示す文(“soluble”, “dissolve”などの語を含

む)、(2)フィルム作成に関する文(“cast from”, “spin-coat”などの語を含む)、(3)分子量測定に関する文(“GPC”, “NMR”および粘度測定に関する語を含む)。機械学習では BERT^[4]を文分類に使用し、データを 5 分割した交差検定で行った。文分類後、True Positive (TP)、False Positive (FP) および False Negative (FN)を調べ、それらの数から Precision (TP/TP+FP)、Recall (TP/TP+FN) および F 値 (2 precision recall/precision+recall)を算出した。

3. 結果と考察

ルールベースによる文分類では、Precision、Recall、F 値がそれぞれ 0.956、0.657、0.778 となった。一方、機械学習では 0.663、0.636、0.649 であった。Recall は両者でほぼ同じであったが、Precision はルールベースで 0.95 以上となり、ルールが上手く適応されていた。一方、機械学習では 0.66 と低めの値であり、結果を詳細に見てみると、どのタイプの文が FP しやすいなどの傾向はなかった。そのため、実験データ数が少なかったために学習が不十分なことが考えられ、データ増強によっては精度向上の可能性があると考えられた。

4. まとめ

良溶媒文の分類がルールベースまたは機械学習によって F 値 0.65-0.78 で行えることが分かった。現在実験データの増強を行っており、当日はその効果も併せて報告する。

- [1] PoLyInfo, <https://polymer.nims.go.jp/>
- [2] K. Terakura, I. Takeuchi, Focus on materials genome and informatics. Sci. Technol. Adv. Mater., 2017, **18**, 1-2.
- [3] 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, 岡博之他, 応用物理学会第 66 回春季学術講演会, 東京工業大学(大岡山, 東京), 2018 年 3 月 9-12 日.
- [4] J. Devlin et al., BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.