学術論文からのポリマー・溶媒の固有表現および溶解性の自動抽出

山口 泰弘 [1]　　進藤 裕之 [1]　　松本 裕治 [1,2]
[1] 奈良先端科学技術大学院大学 先端科学技術研究科
[2] 理化学研究所 革新知能統合研究センター
{yamaguchi.yasuhiro.yw2, shindo, matsu}@is.naist.jp

1 はじめに

物質化学の分野では学術論文からポリマーのデータを収集し，データベースとして管理･公開する試みが行われている[1]．論文からのデータ抽出は人手で行われているが，最近の論文出版数の増加に伴い，論文中から自動的にポリマーデータを抽出する取り組みが行われている [4]．データベースに収められるポリマーデータのうち多くは数値データであり，論文中では表にまとめられている場合が多い．一方，ポリマーの溶解性に関する情報は本文中にテキストとして記述される場合が多い (例えば，“Py-PC4MA was soluble in toluen” など)．こうしたテキスト中から情報を自動で抽出するために，自然言語処理技術の応用が期待できる．

本研究では学術論文のテキスト中からポリマーと溶媒の関係を機械学習モデルを用いて自動抽出することを試みた．抽出の過程は固有表現認識タスクと関係抽出タスクの 2 つに分けられる．固有表現認識タスクでは，与えられた文中からポリマーと溶媒のスパンを予測する．現在，固有表現認識のタスクにおいて高い精度を達成しているモデルのひとつに BiLSTM-CRF[3] がある．本研究では BiLSTM-CRF に基づく固有表現認識モデルの作成と評価を行った．一方，関係抽出タスクでは，文とポリマー・溶媒のスパンが与えられたもとで，ポリマーと溶媒の間の溶解性に関わる関係を予測する．ポリマーデータに含まれる溶媒には良溶媒と貧溶媒の 2 種類が存在するが，ここでは簡単のために良溶媒のみを対象に関係抽出を行った．

岡ら [4] は論文中のテキストからポリマーと溶媒の関係を自動抽出するためにルールに基づく手法を提案した．岡らの手法では，溶媒の溶解性に関わるキーワード (“soluble” や “dissol”) を予め決めておき，それらを含む文に出現するポリマーと溶媒をすべて関係づけるというアプローチで関係抽出を行っている．この手法は高い Recall を達成する一方で，“Py-PC1A was soluble in THF, DMF, and DMSO but not in toluene” のように 1 文の中に良溶媒と貧溶媒の両方が含まれる文に対して正しい予測を行うことができない問題がある．本研究ではこの問題を解決するために，文中に含まれるポリマーと溶媒のすべてのペアについて関係を予測する機械学習モデルの作成を試みた．さらに，岡らのルールベースの予測を特徴量としてモデルに取り入れて学習することで関係抽出モデルの予測精度の改善を行った．

また，本研究では科学論文中のテキストで学習された BERT モデルである SciBERT[1] を固有表現認識と関係抽出タスクの両方で利用することを試みた．比較的に似たドメインのテキストで学習されたモデルを使って学習することで，固有表現認識と関係抽出の両方のタスクにおいて予測精度の改善が見られた．

2 実験

2.1 データセット

データセットは岡ら [4] の作成したものを利用した．このデータセットは英語論文である Macromolecules（出版社：American Chemical Society）のうち 2016 年分の 63 論文についてアノテーションが行われている．ポリマーと溶媒に関する記述のある文は全部で 599 文あり，ポリマーと溶媒の固有表現，および溶解性についてアノテーションされている．

固有表現認識のためのデータセットには，各テキスト中の単語に対して BIO タグ付けを行った．タグの種類はポリマー (polymer) と溶媒 (solvent) のそれぞれがあり，次の 5 つになる: B-polymer, B-solvent, I-polymer, I-solvent, O.

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

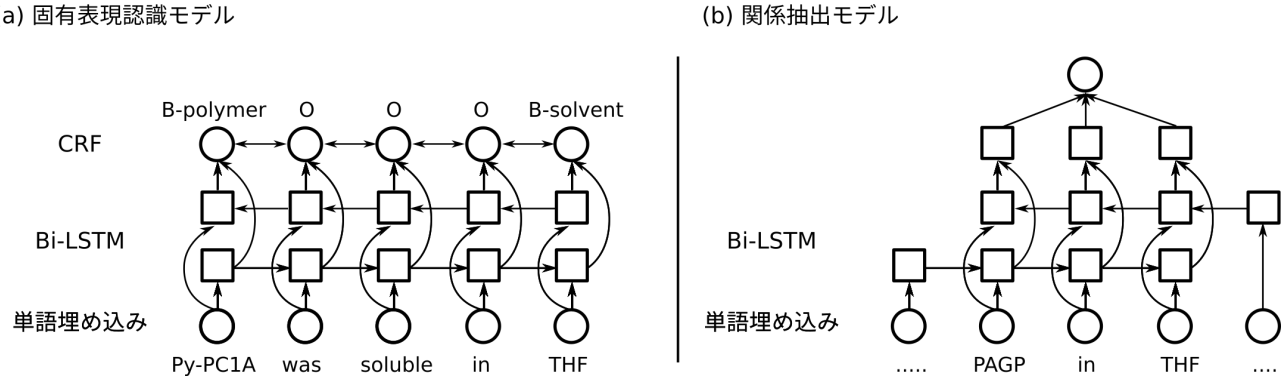

図 1: (a) 固有表現認識モデル．単語埋め込みを Bi-LSTM でエンコードした後，CRF によりタグ系列を予測する．(b) 関係抽出モデル．Bi-LSTM を用いて系列をエンコードした後，ポリマー (“PAGP”) と溶媒 (“THF”) の間のベクトルに基づいて溶解性の記述の有無を予測する．

関係抽出タスクでは，1 文中にポリマーと溶媒を含む文を取り出し，文，ポリマー・溶媒のスパン，ポリマー-溶媒の関係からなるデータセットを作成した．ポリマーと溶媒の間の関係は，ポリマーが溶媒に溶ける (良溶媒) か否かの 2 値である．

2.2 単語埋め込み

固有表現認識と関係抽出の両方のタスクにおいて，char-CNN と，学習済み BERT モデルである BERT-Base[2] と SciBERT[1] を用いて各埋め込みごとにモデルの性能を評価した．

char-CNN は文字レベルの特徴量を計算する手法の一つである．ここでは文字埋め込みを 1 次元 CNN で畳み込み，max-pooling を行うことで単語埋め込みを計算した．文字埋め込みはタスクの訓練時に学習する．

BERT-Base は 12 層の双方向 Transformer から構成されるモデルである．ここでは Wikipedia 等の英語コーパスで訓練された学習済み BERT モデルである BERT-Base[2]を利用した．BERT モデルから単語埋め込みを得るために，各層の Transformer の出力から単語に対応するタイムステップのベクトルを取り出し，それらの重み付け和のベクトルを単語埋め込みとして扱う．各層のベクトルに割り当てられる重み係数は学習パラメータとして訓練時に調整される．また，BERT-Base は単語を WordPiece によりサブワードに分割して学習しているが，ここでは単語埋め込みとして単語を構成するサブワードのうち先頭のサブワードの埋め込みを単語埋め込みとして利用した．

SciBERT[3]は生物医学とコンピュータサイエンス分野の学術論文中のテキストを用いて学習した BERT モデルである．SciBERT のアーキテクチャは BERT-Base と同じである．したがって，SciBERT から単語埋め込みを計算する手順も BERT-Base の場合と同様の手法を用いた．

2.3 固有表現認識

固有表現認識タスクでは，与えられた文からそこに含まれるポリマーと溶媒のスパンを予測する．具体的には，文中の各単語に割り当てられるべき BIO タグを予測する．

固有表現認識モデルには BiLSTM-CRF モデルを利用する．モデルの概要を図 1-a に示す．単語埋め込み $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T$ を 2 層の Bi-LSTM を通した後，CRF を用いてタグ系列 $\boldsymbol{y} = (y_1, \cdots, y_T)$ の確率を計算する．

$$(\boldsymbol{h}_1, \cdots, \boldsymbol{h}_T) = \mathrm{BiLSTM}(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T) \tag{1}$$

$$p(\boldsymbol{y}|\boldsymbol{h}_1, \cdots, \boldsymbol{h}_T) \propto \exp\left(\sum_{t=2}^{T} \theta_{y_{t-1}, y_t} + \sum_{t=1}^{T} \boldsymbol{w}_{y_t}^{\mathrm{T}} \boldsymbol{h}_t\right) \tag{2}$$

中間層 $\boldsymbol{h}_1, \cdots, \boldsymbol{h}_T$ の次元は 50 次元とした．$\theta_{y_{t-1}, y_t}, \boldsymbol{w}_{y_t}$ はそれぞれ CRF の学習パラメータである．学習時には以下の負対数尤度を最小化するように Bi-LSTM と CRF のパラメータを調整する．

$$\mathcal{L}_{\mathrm{NER}} = -\sum_i \log p(\boldsymbol{y}|\boldsymbol{h}_1, \cdots, \boldsymbol{h}_T) \tag{3}$$

[2] https://github.com/google-research/bert

[3] https://github.com/allenai/scibert

2.4 関係抽出

関係抽出タスクでは文とポリマー・溶媒のスパンが1つずつ与えられ，ポリマーと溶媒の溶解性を予測する．ここでは良溶媒のみを対象とするため，予測する関係はポリマーが溶媒の良溶媒か否かの2値である．

関係抽出モデルを図 1-b に示す．単語埋め込み $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T$ を1層の Bi-LSTM に通した後，ポリマーと溶媒の間の系列に対応するベクトル $\boldsymbol{h}_l, \cdots, \boldsymbol{h}_r$ を取り出す．ここで l, r はポリマーと溶媒を含む最小のスパンの両端のインデックスを表す．このポリマー-溶媒間の各系列 $\boldsymbol{h}_t$ を，ベクトル $\boldsymbol{w}$ との内積によって得られる重み a_t に基づき重み付き和 $\boldsymbol{c}$ を計算する．最後に2層の順伝播型ニューラルネットワーク FFN を用いてベクトル $\boldsymbol{c}$ から予測スコアを求める．

$$(\boldsymbol{h}_1, \cdots, \boldsymbol{h}_T) = \mathrm{BiLSTM}(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T) \tag{4}$$

$$a_t = \boldsymbol{w}^{\mathrm{T}} \boldsymbol{h}_t \tag{5}$$

$$\boldsymbol{c} = \sum_{t=l}^{r} a_t \boldsymbol{h}_t \tag{6}$$

$$score = \mathrm{FFN}(\boldsymbol{c}) \tag{7}$$

ここで，中間層 $\boldsymbol{h}_1, \cdots, \boldsymbol{h}_T$ の次元は 50 次元とした．またベクトル $\boldsymbol{w}$ はモデルの学習パラメータである．

以下の負対数尤度を最小化するようにモデルの学習を行う．

$$p = p(y = 1 | \boldsymbol{x}_1, \cdots, \boldsymbol{x}_T, l, r) = \frac{1}{1 + \exp(-score)} \tag{8}$$

$$\mathcal{L}_{\mathrm{RE}} = -\left\{ \sum_i y_i \log p_i + (1 - y_i) \log(1 - p_i) \right\} \tag{9}$$

y は正解ラベルであり，ポリマーと溶媒の間に溶解性がある場合に 1，そうでない場合に 0 となる．

1文の中に複数のポリマーと溶媒に関する記述を含む場合は，可能なすべてのポリマー・溶媒のペアについてそれぞれ溶解性の有無を予測する．

さらに，岡ら [4] のルールに基づく予測も特徴量として利用することを考えた．

$$score = \mathrm{FFN}([\boldsymbol{c}; \hat{y}_{\mathrm{rule}}])$$

ポリマーと溶媒のペアのベクトル $\boldsymbol{c}$ にルールベースの予測 $\hat{y}_{\mathrm{rule}} \in \{0, 1\}$ を結合して予測を行う．$\hat{y}_{\mathrm{rule}}$ はルールベースの予測が溶解性ありの場合に 1，そうでない場合に 0 とする．

3 結果と考察

3.1 固有表現認識

ポリマーと溶媒のスパンがアノテーションされた 599 文を用いて 5 分割交差検証を行い，F1，Precision，Recall を算出した結果を表 1 に示す．SciBERT と char-CNN の単語埋め込みを結合したモデルが最も高い F1 スコア (91.6) となった．BERT-Base，SciBERT のどちらでも，char-CNN を利用したモデルのほうが F1，Precision，Recall すべてで性能の改善が見られる．ポリマーや溶媒は “Py-PC1MA” や “CHCl3” など一般的な単語に比べて異質な文字列になる場合が多いため，char-CNN による文字列の特徴量が予測に有用であると考えられる．また，BERT-Base と SciBERT を比べると SciBERT モデルを用いたモデルがより高い精度になっている．SciBERT は学術論文コーパスで事前学習を行っているため，物質化学分野のテキストにも適合しやすいと考えられる．例えば “poly” を含む単語は，BERT-Base の辞書中には 7 個存在するのに対し，SciBERT の辞書中には 37 個含まれている．事前学習モデルが似た語彙を含んでいることで，予測に有用な単語埋め込みが得られると考えられる．

3.2 関係抽出

文中にポリマーと溶媒の両方のスパンを含む 161 文を用いて 10 分割交差検証を行い，F1，Precision，Recall を算出した結果を表 2 に示す．単語埋め込みに SciBERT を用いてルールベースの予測を特徴量に追加したモデルが最も精度がよく，F1 スコア 82.1 となった．BERT-Base と SciBERT のいずれのモデルにおいても，ルールベースの予測を付け加えることで Recall の改善が見られた．また，関係抽出タスクにおいても固有表現認識タスクと同様に SciBERT を用いたモデルのほうが BERT-Base に比べて高い精度が得られるという結果になった．予測結果を確認すると，“The more polar Py-PC1A was soluble in THF, DMF, and DMSO but not in toluene” のように良溶媒と貧溶媒の両方を含む文においても正しい予測が可能であることがわかった．しかし，“The polymers rP-1 and sP-1 are readily dissolved in apolar solvents like CHCl3 and THF, while sP-3 to sP-5 are soluble in polar solvents like DMF.” のように，長い文で複数の節にそれぞれポリマーと溶媒の関係が記述される

	char-CNN	BERT-Base	BERT-Base + char-CNN	SciBERT	SciBERT + char-CNN
F1	87.1 ± 1.67	88.7 ± 3.47	89.0 ± 3.35	91.1 ± 3.02	**91.6 ± 2.80**
Precision	88.8 ± 1.18	86.9 ± 4.24	87.4 ± 4.01	90.2 ± 3.51	**91.4 ± 3.87**
Recall	85.5 ± 2.92	90.6 ± 2.75	90.6 ± 2.68	91.2 ± 3.14	**91.8 ± 2.31**

表 1: 固有表現認識モデルの実験結果．“+char-CNN” は単語埋め込みに char-CNN から得られた単語ベクトルを結合して学習したモデル．

	rule-based	BERT-Base	BERT-Base + rule-based	SciBERT	SciBERT + rule-based
F1	72.8 ± 11.2	77.2 ± 14.2	81.9 ± 10.2	77.9 ± 10.2	**82.1 ± 9.70**
Precision	60.5 ± 13.8	79.4 ± 11.5	**80.3 ± 11.2**	79.6 ± 12.3	78.7 ± 10.4
Recall	**93.8 ± 5.25**	79.4 ± 11.6	84.5 ± 12.6	82.5 ± 11.6	86.3 ± 10.9

表 2: 関係抽出モデルの実験結果．“rule-based” は岡ら [4] の手法．“+ rule-based” はルールベースの予測を特徴量に用いたモデル．

ような複雑な例については依然として正しい予測は困難であると考えられる．

また，関係抽出タスクにおいては char-CNN の使用は予測精度の改善に寄与しなかった．SciBERT + char-CNN + rule-based モデルで 10 分割交差検証を行ったところ，F1 スコアは 77.5 ± 14.9 となった．関係抽出タスクではポリマーや溶媒の具体的な文字列よりも，“soluble” や “dissolve” などのキーワードがより重要になると考えられる．

4 おわりに

物質化学分野の学術論文からポリマー・溶媒の固有表現と溶解性に関する関係の自動抽出を行った．固有表現認識と関係抽出のどちらのタスクにおいても SciBERT を用いたモデルで高い精度が得られた．関係抽出タスクにおいては，複数の節を含む文における予測性能に課題が見られた．また，溶解性に関する記述の中には “All these three polymers are soluble in DMF.” のように代名詞を含む例が複数存在する．こうした文から情報抽出を行うためには，“these three polymers” が何を示すのかを明らかにす必要がある．代名詞を含む文からもポリマーデータを自動で抽出するために，今後は共参照解析タスクにも取り組む予定である．

謝辞

国立研究開発法人物質・材料研究機構 (NIMS) にはアノテーションデータをご提供いただきました．ここに記して、謝意をあらわしたいと思います．

参考文献

[1] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3613–3618, Hong Kong, China, November 2019. Association for Computational Linguistics.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[3] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 260–270, San Diego, California, June 2016. Association for Computational Linguistics.

[4] 岡博之, 吉澤篤志, 進藤裕之, 松本裕治, 石井真史. 学術論文からのポリマー溶解性データの自動抽出. 言語処理学会 第 25 回年次大会. 言語処理学会, 2019.