無機材料文献からの合成プロセス抽出のための関係抽出

牧野 晃平[*1*2]　國吉 房貴[*2*3]　小澤 順[*2*3]　三輪 誠[*1*2]
[*1]豊田工業大学　[*2]国立研究開発法人産業技術総合研究所
[*3]パナソニック株式会社
{kohei.makino, kuniyoshi.fusataka, ozawa.jun, makoto.miwa}@aist.go.jp

1　はじめに

材料開発の分野では，膨大な文献に記述されている合成プロセスの解析により，新材料の探索や開発にかかる時間を短縮する技術が求められている．合成プロセスは，一連の実験操作を示すための手順であり，複数の文に渡って記述されるのが一般的である．このような複数文にまたがる関係を考慮して，文献に直接タグ付けしたコーパスが提案されている [1, 2]．本研究では，このうち，文献中に記載された合成プロセスをフローグラフとして文献上に直接タグ付けした，國吉らのコーパス [2] を対象とする．

このようなコーパスを対象とした関係抽出には，文間の関係を対象とした関係抽出手法が必要とされる．文間を考慮した関係抽出手法としては，データベースからの遠距離教師あり学習を対象に，文献レベルでの用語ペアを抽出する手法がほとんどであるが，近年，深層学習を利用した文書に紐付いた記述レベルの用語ペアも利用できる手法が提案されている [3, 4]．

本論文では，無機材料分野における文間の関係抽出手法について検討する．そのために，深層学習モデルとルールに基づくモデルの 2 つの関係抽出器を提案し，評価を行う．深層学習モデルは，文間を考慮した関係抽出において高い性能を示している BRAN [3] を参考に，近年の言語処理タスクの性能を大きく向上させた BERT [5] を用いて，関係抽出モデルを実現した．ルールに基づくモデルは，データからの知見をもとに単純なルールを提案する．

実験では 2 つのモデルを國吉らのコーパスで評価した．深層学習モデルでは合成プロセス抽出の精度が十分に得られず，一方で，構築したルールに基づくモデルの方が高精度に関係を抽出できるとわかった．

2　合成プロセスコーパス

本研究では，合成プロセスに関連する用語と，プロセスの進行などのプロセス内での用語間の関係がタグ付けされている國吉らのコーパス [2] を拡張し，関係抽出手法の検討を行う．

このコーパスでは,用語は,材料を示す Material･操作を示す Operation・付加条件を示す Property の 3 種で定義され, Property についてはサブラベルとして, 時間条件を表す Property-Time･温度条件を表す Property-Temp・回転速度条件を表す Property-Rot･圧力条件を表す Property-Press, その他の条件 Property-Others の 6 種で定義されている．関係はプロセスの進行を表す Next と条件付けを表す Condition の 2 種の有向関係で定義される．

本研究では，上記の Material について，出発材料を表す Material-Start・中間生成物を表す Material-Intermedium・最終生成物を表す Material-Final･溶媒を表す Material-Solvent･その他材料を表す Material-Others の 5 種のサブラベルについて追加のタグ付けを行った．合成プロセスの例を図 1に示す．

3　利用する関係抽出手法

本節では対象とする関係抽出モデルである深層学習モデルとルールに基づいたモデルについて述べる．

3.1　深層学習モデル

深層学習モデルの関係抽出手法として，文間を考慮した関係抽出手法の中で高い性能を示している深層学習モデル BRAN [3] を参考にモデルを構築する．BRAN は遠距離教師学習の手法で，文をまたぐ関係を抽出することを目的に提案された，Transformer [6] を用いたモデルである．作成したモデルは，BRAN の Transformer 部を近年の様々な言語処理タスクで高い性能を示して

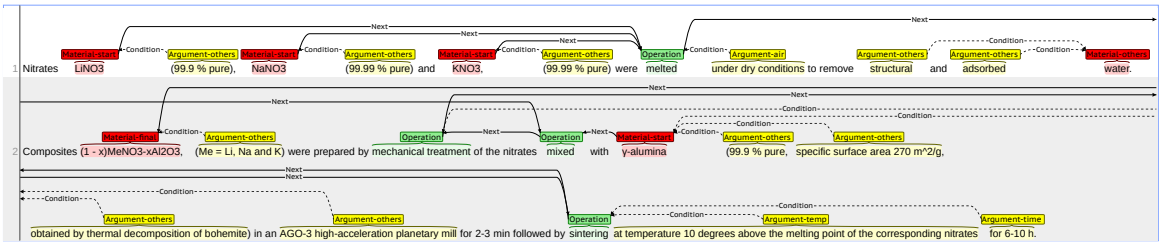

図1　合成プロセスコーパスの例

いる BERT [5] に置き換える．また，遠距離教師学習ではなく教師あり学習を行い，用語ラベルは入力に付加し，用語抽出は行わない．

まず，事前学習モデル BERT によって文章をサブワード単位のトークンに分割し，そのトークンを BERT で特徴量化する．そしてそのトークン毎の特徴量に対して位置埋め込みを結合し，用語を構成するトークンの特徴量に最大値プーリングをして，トークン毎の特徴から用語毎の特徴にまとめる．このときその特徴量に対して，用語ラベルの情報を付与するため，用語ラベルの埋め込みを結合する．その後用語が関係ラベルごとに Head になる場合と Tail になる場合[*1]のそれぞれの特徴量を 2 層の全結合層（FC）によって計算する．このとき，Head と Tail それぞれの特徴量 $\boldsymbol{e}_l^{(head|tail)} \in \mathbb{R}^d$ を計算するための全結合層 $\mathrm{FC}^{(head|tail)}$ は，

$$\boldsymbol{e}_l^{head} = \mathrm{FC}^{head}(\boldsymbol{e}_l) \tag{1}$$
$$\boldsymbol{e}_l^{tail} = \mathrm{FC}^{tail}(\boldsymbol{e}_l) \tag{2}$$

と表される．ただし，用語毎の特徴量ベクトルは次元数を d として $\boldsymbol{e}_l \in \mathbb{R}^d$ であり，l は用語の出現順番を表し，$l = 1, \ldots, L$ であり，L は文献中の用語の数である．この Head と Tail になる場合の特徴量から，「関係なし」を含むそれぞれの関係のスコアを計算する．$l = l1$ が Head，$l = l2$ が Tail が関係 r となるときのスコア $s_{l1,r,l2}$ は，

$$s_{l1,r,l2} = \boldsymbol{e}_{l1}^{head} \boldsymbol{A}_r \boldsymbol{e}_{l2}^{tail} \tag{3}$$

で表される．ただし，$\boldsymbol{A}_r \in \mathbb{R}^{d \times d}$ は関係ラベルごとに用意した行列である．予測するクラスは

$$\arg\max_r s_{l1,r,l2} \tag{4}$$

とし，スコアが最も高いものを選択する．このモデルを学習するための損失関数には交差エントロピー損失を用いる．損失 $\mathcal{L}$ は，

$$\mathcal{L} = -\sum_{l1=1}^{L}\sum_{l2=1}^{L}\sum_{r} t_{l1,r,l2} \frac{s_{l1,r,l2}}{\sum_r s_{l1,r,l2}} \tag{5}$$

と表せる．ただし，$t_{l1,r,l2}$ は教師ラベルで，$l1$,$l2$ 間に r の関係がある場合に 1，それ以外は 0 となる．

3.2　ルールに基づいたモデル

ルールに基づいたモデルでは，合成プロセスコーパスにおいて用語のペアが与えられたとき，その用語クラスに対して定義したルールによって関係を抽出する．ルールは，訓練用データセットを観察し，関係の種類ごとに用語間の距離と用語が出現した順序に基づいて定義する．ここでの用語間の距離は単語をスペース区切りで分割した場合の単語数である．作成したルールは以下の通りである．

Operation → Operation (O-O): Operation は操作する順に記述されていると仮定し，出現した順番に関係をつなぐ．

Material → Operation (M-O): Material-Start・Material-Solvent から文内で最近傍の Operation に対して関係を接続する．前後で同一距離に Operation が存在する場合は，前方に出現した Material を優先する．文内に Operation が存在しない場合，以降の文で Operation を探索し，最初に出現した Operation に対して関係をつなぐ．ただし，対象となる Operation が括弧で括られている場合，その Operation は括弧の直前の Material に対して係っていると仮定し，括弧の直前の語のみをその Operation に接続するようにする．例えば，“Samples were prepared from H_3BO_3, AL_2O_3, SiO_2 and either Li_2CO_3 (dried at 200 degC).” という文が存在した場合，“Li_2CO_3” のみを “dried” に接続し，ほかの Material は “dried” を無視して次の文を探索する．

Operation → Material (O-M): すべての操作が終了したときに最終生成物が生成すると仮定し

[*1] 関係の始点となる用語を Head, 終点を Tail と呼ぶ．

て，最後に記述されている Operation からすべての Material-Final に対して関係をつなぐ．

Property-Others → Operation or Material (Po-OM): Property-Others から文内で最も近傍の Material もしくは Operation の用語に関係をつなぐ．ただし，Property-Others が括弧で囲まれている場合は，前方へ係ると仮定して前方への探索のみを行う．文内に Material もしくは Operation が存在しない場合は，その Property-Others は関係を持たない用語とする．

Property → Operation (P-O): 英語論文では条件が後置されることが多いため，Property-Others を除いた Property（Property-Time, Property-Temp, Property-Rot, Property-Press）から前方へ探索し，文内・文間を問わずに最も近くに存在する Operation に対して関係をつける．

4 関係抽出手法の性能評価

まず 3 章で述べた深層学習モデルとルールに基づくモデルの比較を行った．コーパスは國吉ら [2] の分割と同様になるように分割し，訓練用 145 件，テスト用 49 件，開発用 49 件に分割した．関係ラベルごとの評価指標は F 値を用い，全体の評価値としてそれぞれの F 値を平均したマクロ F 値を用いた．評価はテスト用と開発用を合わせた 98 件で行った．深層学習モデルについては 10 回学習させたときの平均値をスコアとし，誤差は標準偏差を用いる．深層学習モデルの訓練は訓練用データで訓練し，十分性能が飽和する 200 エポック学習した中で開発データに対するマクロ F 値が最大となったときのパラメータを最終的なパラメータとして扱った．最適化手法は Adam [7] を用いて，学習率を 0.0005，減衰率を 0.01 として学習した．

評価した結果を表 1（上部）に示す．結果から，ルールに基づくモデルと比較して深層学習モデルは十分な抽出性能が得られていないとわかる．さらに，ルールそれぞれについて，予測時に適用されたルールの回数の割合を示すカバレッジと精度を評価し，結果を表 2に示した．この結果からほとんどのルールが関係抽出の性能に寄与するように働いていることが確認できる．しかし，O-M のルールについてはカバレッジが低く，精度も低い．これは最後に述べられた Operation から Material-Final が生じるという仮定が誤っている場合があり，実際の文献には倒置や薄膜の合成のような合成プロセス後に記述されるプロセスが存在していることが要因であると考えられる．

次に，高い性能を記録したルールに基づくモデルについて，國吉らの用語抽出器 [2] を用いたパイプラインでの抽出を評価した結果を表 1に示す．この結果より，ルールに基づくモデルのパイプラインでの抽出性能はマクロ F 値 0.499 であった．他のモデルと比べてマクロ F 値が低い理由は，用語抽出器による用語の予測の誤りが関係抽出器に悪影響を及ぼすことが原因である．例えば，“An amount of LiNi0.8Co0.2O2 powder was mixed with ground CoSO4*7H2O.” という文に対して，國吉らの用語抽出器では “ground” を Property-Others, 正解は Operation となる．このとき Material-Start の “CoSO4*7H2O” に着目して Operation の “mixed ” について考えると，パイプラインでは “ground” から “CoSO4*7H2O” に対する Condition・“CoSO4*7H2O” から “mixed” に対する Next が予測され, 正解を与えると “CoSO4*7H2O” から “ground” に対する Next のみが正しく予測される．

5 抽出結果の解析

ルールに基づくモデルと深層学習モデルについて，実際に関係抽出器が出力した結果を比較しその結果を解析した．図 1と同一の文章に対し，深層学習モデルとルールに基づくモデルそれぞれを用いて関係抽出した結果を，図 2と図 3に示した．深層学習モデルは実験した 10 回のうち，最も高いマクロ F 値を記録したモデルの結果を示す．

表1 それぞれの関係抽出器の抽出性能

	Next	Condition	マクロ F
深層学習モデル	0.478 ± 0.008	0.614 ± 0.007	0.546 ± 0.006
ルール	0.860	0.916	0.888
パイプライン（ルール）	0.544	0.454	0.499

表2 ルールそれぞれの評価

ルール	カバレッジ	精度
O-O	0.219	0.811
M-O	0.160	0.811
O-M	0.046	0.489
Po-OM	0.322	0.853
P-O	0.254	0.951

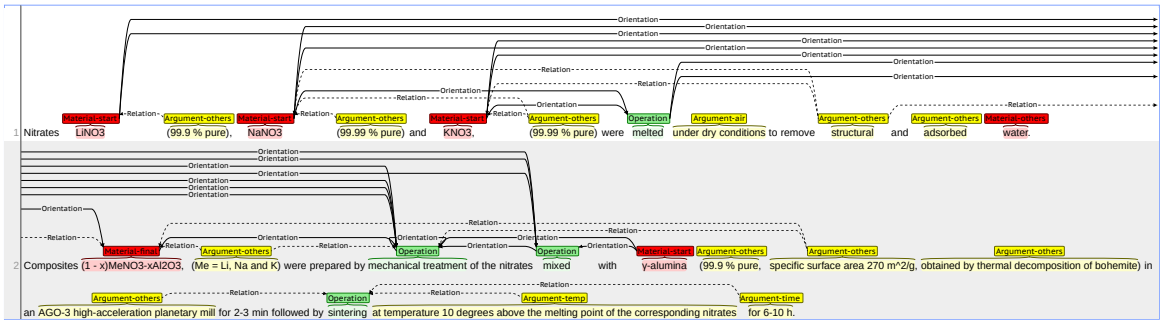

図2　深層学習モデルの出力

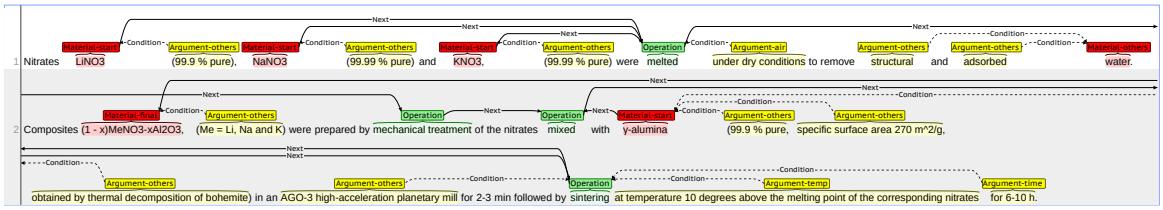

図3　ルールに基づくモデルの出力

結果を比べると，ルールに基づくモデルはほぼ正解を再現できていることがわかる．ただし，OPERATION の “mechanical treatment” と “mixed” に注目すると，「混合」された γ-alumina が「機械的操作」によって準備された，となるため，“mixed” から “mechanical treatment” への NEXT の関係が正解となる．ルールに基づくモデルでは言語的情報を利用していないため “mechanical treatment” から “mixed” への NEXT を出力して誤っているが，深層学習モデルでは正解の関係を出力できている．このように，ルールでは誤っているが深層学習が正解できている関係もあり，現在のルールで利用している情報では不十分であり，網羅的にルールを記述するのは難しいことから，ルールと深層学習を組み合わせて利用する方法についても検討する必要があると考えられる．

6　おわりに

本論文では合成プロセスコーパスに対する関係抽出において深層学習モデルとルールに基づくモデルを提案し，比較を行った．ルールに基づく関係抽出器が深層学習を用いた関係抽出と比較して高い性能を示し，マクロ F 値 0.888 を記録した．しかし，ルールに基づくモデルと深層学習モデルの出力を比較すると，ルールベースの関係抽出には抽出できる情報の限界があり，深層学習の方が正解できている点があることもわかった．今後の課題として，高性能なルールと言語的情報を扱える深層学習を組み合わせて，倒置やプロセスの分岐などの言語的情報が必要となる問題が解けるよう，それぞれの利点を活かして関係抽出を行うシステムの構築を行う．

参考文献

[1] Mysore et al. The materials science procedural text corpus: Annotating materials synthesis procedures with shallow semantic structures. In *Linguistic Annotation Workshop*, 2019.

[2] 國吉ら．学術文献からの無機材料合成プロセス抽出のためのグラフ表現．Technical report, 第 15 回テキストアナリティクス・シンポジウム, 2019.

[3] Verga et. al. Simultaneously self-attending to all mentions for full-abstract biological relation extraction. In *NAACL-HLT*, 2018.

[4] Christopoulou et. al. Connecting the dots: Document-level neural relation extraction with edge-oriented graphs. In *EMNLP-IJCNLP*, 2019.

[5] Devlin et. al. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.

[6] Vaswani et. al. Attention is all you need. In *NIPS*. 2017.

[7] Kingma et. al. Adam: A method for stochastic optimization. In *ICLR*, 2015.