文献抄録中の主題材料に着目した超伝導材料に関する情報抽出

山口 京佑 [1]　旭 良司 [2]　佐々木 裕 [1]
[1] 豊田工業大学
[2] 株式会社 豊田中央研究所
[1]{sd19453, yutaka.sasaki}@toyota-ti.ac.jp,
[2]ryoji.asahi@chem.material.nagoya-u.ac.jp

1 はじめに

近年物質科学の分野では，機械学習をはじめとする情報科学のアプローチを物質科学に適用することで，属人的な知見に依存しない効率的な材料探索を目指すマテリアルズインフォマティクス (Materials Infomatics; MI) の研究が活発化している [1]．MI においては所望の材料および物質特性に関する大量のデータが必要となるが，現状ではその多くが関係データベースのような形で構造化されておらず，開発を進める上でのボトルネックとなっている．

超伝導材料は医療機器やリニア新幹線といった幅広い分野へ応用されている材料であり，物質科学において重要な材料の一つである．銅酸化物系の高温超伝導体 [2] が発見されて久しいが，実応用に向けては転移温度や臨界磁場がより高く，加工しやすい材料が必要とされている [3, 4]．

本研究では，文献抄録から超伝導材料に関する情報をスロット抽出するシステムの構築を目指す．本研究の概要を図 1 に示す．システム構築に際して，まずは抄録 1,000 件に対して抽出したい超伝導材料情報を注釈付けし，教師データを作成する．提案システムは固有表現・関係・イベント抽出モデルと主題材料分類モデルの２つのニューラルネットワークと，これらの予測結果を統合して最終的なスロット抽出を行うルールベースのモジュールで構成される．作成した教師データを基に２つのモデルを独立に学習し，予測時にはパイプラインでつなげて運用する．

2 超伝導情報の注釈付け

我々の先行研究 [5] において，超伝導材料に関する文献抄録 1,000 件に対して７つの固有表現クラスを注釈付けした．本研究ではこれを拡張する形で，

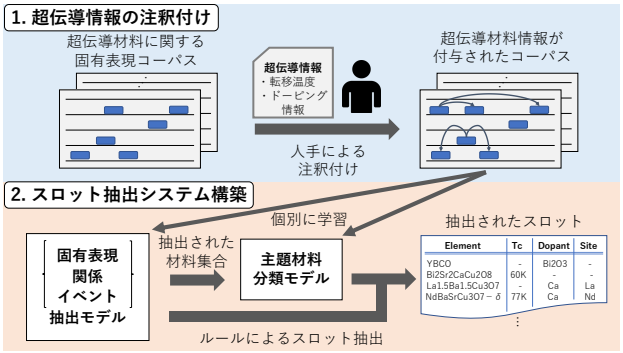

図 1　本研究の概要

超伝導材料に固有の情報である転移温度とドーピング情報を関係付けし，これらを対応する材料組成に紐づけることでスロット抽出を行う．抄録 1 件に注釈付けした例を図 A.1 に示す．本研究で特に重要な固有表現クラスは以下の３つである．その他のクラスに関しては表 A.1 を参照されたい．

Element　“Ti”, “oxygen”, “$YBa_2Cu_3O_7$” といった元素名や化合物名を対象とする．2.1 節と 2.2 節で述べる転移温度・ドーピング情報の紐づけ対象である．

Doping:　“doping”, “substitute”, “addition” などのドーピング操作を表すエンティティを対象とする．本研究ではイベントトリガーとして再定義される．

Value:　“45%” や “95K” といった単位まで含めた定量表現全般を対象とする．

超伝導材料に固有の情報を抽出するにあたり，本研究では上記に加えて新たに SC クラスを定義した．

SC:　超伝導を意味する “superconductivity” や超伝導への転移温度を表す “Tc” など，超伝導特性に関するエンティティを対象とする．

2.1 転移温度の注釈付け定義

物質科学における定量情報は多くの場合，属性名と属性値の関係で整理できる．これらの間の等価性を表す関係クラスとして Equivalent を定義した．

ここで Equivalent は属性名から属性値に対して注釈付けされ，文内の関係のみを対象とする．転移温度の場合，文献中では “$T_c = 95K$” のように記述されることが多く，抽出したい属性名は固有表現クラス SC に，属性値は Value に属する．

2.2 ドーピング情報の注釈付け定義

文献中でのドーピングに関する情報はイベントの枠組みで定義可能である．例として “Zn doping into the CuO_2 plane” を仮定すると，この場合 “doping” がイベント発生を表すトリガーとなり，イベント引数として “Zn” が材料へ添加物であるドーパント，“CuO_2” がドーパントの受け入れ先であるサイトと解釈できる．これを踏まえ，本研究では先行研究 [5] で固有表現クラスとして定義した Doping をイベントトリガークラスとして再定義し，イベントロールクラスとして Dopant と Site を新たに定義した．

2.3 超伝導材料情報の注釈付け定義

本研究では固有表現クラス Element に属する材料組成に対して転移温度・ドーピング情報を紐づける形でスロット抽出を行う．これらを紐づけるための関係クラスとして Target を定義した．ただし Target は転移温度・ドーピング情報から材料組成に対して注釈付けされ，文内のみを対象とする．

ここで問題となるのが，材料組成と転移温度・ドーピング情報が文を跨いで言及される場合であり，文内の関係を対象とする Target による紐づけのみでは情報の取りこぼしが発生してしまう．本研究では，抄録において文献の主題となる材料に関する事実や実験結果が集約して述べられている場合が多い事実に着目し，抄録中の主題材料を特定した上で転移温度・ドーピング情報を紐づける方法を提案する．

主題材料の特定は Element エンティティを対象にそれぞれが文献の主題であるか否かを判定する二値分類問題として設定し，注釈付けは Element エンティティの内，主題材料であるものを新たに定義した固有表現クラス Main で上書きする形で行う．なお，転移温度・ドーピング情報の主題材料への紐づけは 3.3 節で述べるルールに従って行われる．

抄録中では材料組成に対するドーピングの結果として転移温度が記述される場合がある．この場合ドーピングが転移温度を条件付けしていると解釈でき，この時成り立つ関係クラスとして Condition を

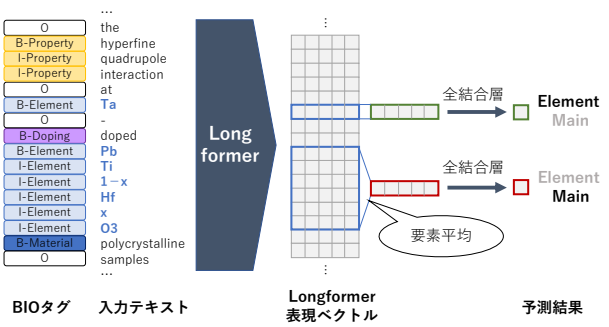

図 2 主題材料分類モデルの概要

定義した．ただし Condition は転移温度からドーピングのイベントトリガーに向かって注釈付けされ，文内の関係のみを対象とする．

3 提案システム

提案システムは固有表現・関係・イベントを抽出するモデルと主題材料分類モデルの 2 つのニューラルネットワークと，これらの結果を統合しスロット抽出するルールベースのモジュールで構成される．以下ではそれぞれについて詳しく述べる．

3.1 固有表現・関係・イベント抽出

本研究では固有表現・関係・イベントの同時抽出手法である DyGIE++ [6] を用いる．DyGIE++は各タスクを解く手掛かりとなる情報を相互にグラフ伝搬する機構により，それぞれのエンドタスクで高い抽出性能を実現している．原著論文ではこれら 3 つのタスクの補助タスクとして共参照解析も同時に解いているが，本研究では共参照を注釈付けしていないため，このタスクは対象外とした．

3.2 主題材料分類

DyGIE++で抽出された Element エンティティを対象に，それぞれが主題材料であるか否かの二値分類問題を解く．提案モデルの概要を図 2 に示す．

主題材料を特定する上では抄録全体の文脈を考慮する必要があるため，モデル入力は文単位ではなく抄録全体とする．単語埋め込み手法は BERT [7] を長い系列長が扱えるように改良した Longformer [8] を用いる．入力系列を $T = \{x_1, x_2, \ldots, x_L\}$ とすると，Longformer の表現ベクトルは以下で表される．

$$\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_L = Longformer(x_1, x_2, .., x_L) \quad (1)$$

ただし，$\boldsymbol{x}_i \in \mathbb{R}^d$ である．

次に抄録中の各 Element エンティティについて，スパンを構成する全トークンの表現ベクトルの要素平均を以下のように計算し，スパン表現 $\boldsymbol{g}_i$ を得る．

$$\boldsymbol{g}_i = \frac{1}{n-m+1}\sum_{j=m}^{n}\boldsymbol{x}_j \tag{2}$$

ただし，m, n はそれぞれスパン $\boldsymbol{g}_i$ の先頭トークンインデックス，末尾トークンインデックスであり，$0 \leq m \leq n \leq L$ である．

さらにスパン表現 $\boldsymbol{g}_i$ を以下の式で計算することで，スパンが主題材料である確率 p_i を得る．p_i が 0.5 よりも小さい場合には「主題材料でない」，大きい場合には「主題材料である」としてモデルの予測結果が得られる．

$$p_i = Sigmoid(\boldsymbol{w}^t(ReLU(\boldsymbol{g}_i))) \tag{3}$$

ただし，$\boldsymbol{w} \in \mathbb{R}^d$ である．

学習時は損失関数としてバイナリ交差エントロピー損失 (Binary Cross Entoropy Loss; BCELoss) を用い，パラメータ更新は Longformer を含めたネットワーク全体に対して抄録単位で行う．

3.3 ルールによるスロット抽出

DyGIE++と主題材料分類モデルにより抽出された情報は事前に定義したルールにより，材料組成・ドーパント・サイト・転移温度の 4 つからなるスロットとして抽出される．大まかなルールは以下の通りである．

1. Condition クラスで関係付けされた転移温度とドーピング情報をスロットに埋める
2. Target クラスで関係付けされた材料組成と転移温度・ドーピング情報をスロットに埋める
3. 2 で紐づけされなかった転移温度・ドーピング情報について，最も近い Main エンティティを対応する材料組成とみなしてスロットに埋める

4 実験

作成した注釈付きコーパスを 10 分割して訓練データ 800 件，検証データ 100 件，テストデータ 100 件の組み合わせを 10 パターン用意し，提案システムの学習と評価を行った．コーパス全体の解析結果を表 A.2, 表 A.3, 表 A.4 に示す．なお，以下で述べる表中の評価スコアは 10 個のモデルの平均および標準偏差である．

4.1 固有表現・関係・イベント抽出の評価

実験は著者実装のモデル[1)]を用いて行った．実験設定の詳細を表 A.5 に示す．各タスクの評価結果はそれぞれ表 1, 表 2, 表 3 の通りである．

表 1 より本研究で重要な 3 つの固有表現クラス Element, Value, SC は高い精度で抽出できていることが分かる．表 2 の関係クラスの評価はエンティティペアの固有表現クラスと関係クラスの予測結果が正しい場合のみ真陽性とした．スロット抽出を行う上で特に重要な Target は F 値が 77.1%という結果となった．表 3 の Total はイベント全体の評価であり，トリガーと引数の固有表現クラス，トリガーと引数間のイベントロールが全て正しい場合を真陽性としてカウントした．Total のスコアが高いのは全体としてドーパントがサイトの約 4.5 倍の出現回数あり，ドーパントの精度に引っ張られていることに起因している．

4.2 主題材料分類の評価

ここではモデルへの入力として人手で注釈付けした正しい Element を与えた場合と，DyGIE++が予測した Element を与えた場合の 2 通りで評価した．実験設定を表 A.6 に，評価結果を表 4 に示す．

人手による正解の Element 情報を与えた場合は，F 値が平均で 83.9%という結果となった (Gold)．この結果が主題材料分類モデルの本来の分類性能である．DyGIE++による予測結果の Element 情報を与えた場合には，F 値が平均で 75.7%という結果となった (DyGIE++)．表 1 から DyGIE++の Element の予測精度は 89.4%であるため，この分の誤差が含まれたスコアとなっている．

4.3 End-to-End でのスロット抽出の評価

ここではまず，注釈付きコーパスをルールで変換した際の主題材料への紐づけ精度の検証を行った．ランダム選択した 200 件を対象にルールによる主題材料への紐づけで抽出されたスロットが正しいかを著者一人が検証したところ，F 値が 97.3%で正確な変換であることが分かった．これを踏まえ，ここでは注釈付きコーパスに対してルールを適用して得られたスロットを正解として扱うこととした．得られたスロットの解析結果を表 A.7 に示す．

上記の正解スロットに対して，材料組成への明

1) https://github.com/dwadden/dygiepp

示的な紐づけを行う DyGIE++のみを用いて得られたスロットの抽出精度 (Rel. only) と，これと併せて材料組成への暗黙的な紐づけを行う主題材料分類モデルを用いて得られたスロットの抽出精度 (E2E: End-to-End) を表 5 に示す．

Rel. only と E2E を比較すると再現率が 2 倍近く向上しており，主題材料を用いた暗黙的な紐づけが効果的であることが明らかとなった．一方で適合率は，Rel. only よりも E2E の方が若干低下しており，主題材料へ暗黙的に紐づけする場合よりも明示的な紐づけを行う場合の方が正確性が高くなるだろうという直感に従う結果が得られた．E2E の F 値は 64.7%となり，これが提案システムにおける最終的なスロット抽出精度である．

5 関連研究

Tshitoyan ら [9] は，無機材料全般に関する文献抄録 330 万件を用いて word2vec [10] を学習し，単語間のベクトル表現の類似度が高いもの同士を調査することで文献中に埋め込まれた知識を獲得するアプローチを提案している．

Weston ら [11] は無機材料に関する抄録を，山口ら [5] は超伝導材料に関する抄録を対象に，材料名や分析手法といった材料全般に共通する固有表現抽出の取り組みを報告している．

大西ら [12] は学術文献から材料設計で重要となるプロセス・構造・特性に関するエンティティとその相関関係を遠距離教師あり学習の枠組みで抽出する手法を提案している．抽出されたエンティティとそれらの相関関係は知識グラフとして整理される．

ChemDataExtractor [13] は論文中の見出し・パラグラフ・キャプション・表を対象として，化学組成とそれに紐づく属性情報を抽出する手法である．具体的には，まず品詞情報と固有表現を機械学習を用いて取り出し，それらの情報を基にテキストを句構造解析することで所望の情報を抽出する流れとなっている．句構造解析では人手で定義したルールが用いられている．

超伝導に関する学術文献に ChemDataExtractor を適用した研究として [14] が報告されている．ここでは磁性・超伝導特性を発現する材料とその転移温度を抽出し，それを元に機械学習を用いた転移温度の予測モデルが提案されている．また，別のアプローチで同様の超伝導材料とその転移温度のペア情報を抽出する研究として [15] が報告されている．これらの先行研究では，本研究が抽出対象としているドーピングに関する情報は扱われていない．

6 おわりに

本研究では文献抄録中で記述される超伝導材料に固有の情報として，定量情報である転移温度と実験操作であるドーピング情報をそれぞれ関係抽出とイベント抽出の枠組みでタスク化し，これらを関係抽出と主題材料情報を用いて対応する材料組成に紐づけることにより，スロット抽出システムを構築した．End-to-End のスロットの抽出精度として F 値約 64%が得られることを世界で初めて明らかにした．今後の主な課題として，スロット情報の拡張と文書レベルでのシステム構築が挙げられる．

表 1 固有表現クラスの抽出精度

クラス	適合率	再現率	F 値
Char.	0.806 ± 0.054	0.804 ± 0.031	0.804 ± 0.034
Proc.	0.737 ± 0.084	0.747 ± 0.052	0.740 ± 0.065
Prop.	0.746 ± 0.061	0.741 ± 0.027	0.743 ± 0.038
Mat.	0.792 ± 0.076	0.789 ± 0.026	0.787 ± 0.034
Elem.	0.900 ± 0.022	0.889 ± 0.030	0.894 ± 0.024
Value	0.949 ± 0.019	0.947 ± 0.025	0.948 ± 0.020
SC	0.907 ± 0.032	0.940 ± 0.021	0.923 ± 0.025

表 2 関係クラスの抽出精度

クラス	適合率	再現率	F 値
Equiv.	0.772 ± 0.105	0.843 ± 0.089	0.802 ± 0.084
Target	0.791 ± 0.058	0.753 ± 0.058	0.771 ± 0.054
Cond.	0.700 ± 0.192	0.836 ± 0.153	0.729 ± 0.111

表 3 Doping イベントの抽出精度

クラス	適合率	再現率	F 値
Doping	0.960 ± 0.015	0.975 ± 0.021	0.967 ± 0.014
Dopant	0.875 ± 0.022	0.861 ± 0.024	0.868 ± 0.019
Site	0.840 ± 0.107	0.754 ± 0.068	0.789 ± 0.046
Total	0.820 ± 0.023	0.837 ± 0.018	0.828 ± 0.019

表 4 主題材料分類の評価結果

入力情報	適合率	再現率	F 値
Gold	0.833±0.034	0.845±0.033	0.839±0.025
DyGIE++	0.784±0.044	0.734±0.041	0.757±0.031

表 5 スロットの抽出精度

項目	適合率	再現率	F 値
Rel. only	0.714±0.047	0.325±0.056	0.444±0.052
E2E	0.655±0.041	0.639±0.039	0.647±0.035

参考文献

[1]Rampi Ramprasad, Rohit Batra, Ghanshyam Pilania, Arun Mannodi-Kanakkithodi, and Chiho Kim. Machine learning in materials informatics: recent applications and prospects. *npj Computational Materials*, Vol. 3, No. 1, pp. 1–13, 2017.

[2]J.G Bednorz and K.A. Müller. Possible hightc superconductivity in the ba-la-cu-o system. *Zeitschrift für Physik B Condensed Matter*, Vol. 64, No. 2, pp. 189–193, 1986.

[3]Malozemoff et al. High-temperature cuprate superconductors get to work. *Physics Today*, Vol. April, pp. 41–47, 2005.

[4]Foltyn et al. Materials science challenges for high-temperature superconducting wire. *Nature Mater.*, Vol. 6, pp. 631–642, 2007.

[5]Kyosuke Yamaguchi, Ryoji Asahi, and Yutaka Sasaki. SCCoMIcs: A superconductivity corpus for materials informatics. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 6753–6760, Marseille, France, May 2020. European Language Resources Association.

[6]David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. Entity, relation, and event extraction with contextualized span representations. In *EMNLP/IJCNLP*, 2019.

[7]Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, 2019.

[8]Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv:2004.05150*, 2020.

[9]Vahe Tshitoyan, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A Persson, Gerbrand Ceder, and Anubhav Jain. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, Vol. 571, No. 7763, pp. 95–98, 2019.

[10]Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, Vol. 26, pp. 3111–3119. Curran Associates, Inc., 2013.

[11]L. Weston, V. Tshitoyan, J. Dagdelen, O. Kononova, A. Trewartha, K. A. Persson, G. Ceder, and A. Jain. Named entity recognition and normalization applied to large-scale information extraction from the materials science literature. *Journal of Chemical Information and Modeling*, Vol. 59, No. 9, pp. 3692–3702, 2019. PMID: 31361962.

[12]Takeshi Onishi, Takuya Kadohira, and Ikumu Watanabe. Relation extraction with weakly supervised learning based on process-structure-property-performance reciprocity. *Science and technology of advanced materials*, Vol. 19, No. 1, pp. 649–659, 2018.

[13]Matthew C Swain and Jacqueline M Cole. Chemdataextractor: a toolkit for automated extraction of chemical information from the scientific literature. *Journal of chemical information and modeling*, Vol. 56, No. 10, pp. 1894–1904, 2016.

[14]Callum J Court and Jacqueline M Cole. Magnetic and superconducting phase diagrams and transition temperatures predicted using text mining and machine learning. *npj Computational Materials*, Vol. 6, No. 1, pp. 1–9, 2020.

[15]Luca Foppiano, Thaer Dieb, Akira Suzuki, and Masashi Ishii. Proposal for automatic extraction framework of superconductors related information from scientific literature. 電子情報通信学会サービスコンピューティング研究会 2019 年度第一回研究会, 第 43 回 MaDIS 研究交流会合同研究会, 2019.

A 付録

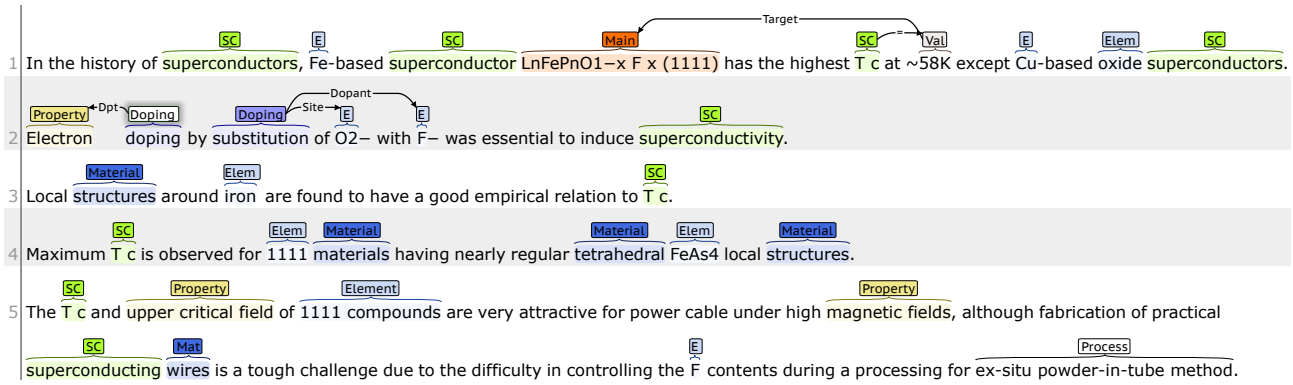

図 A.1 抄録 1 件に注釈付けした例

表 A.1 その他の固有表現クラス

クラス	定義	例
Characterization	分析手法に関する用語	X-ray diffraction, SEM
Process	合成プロセスに関する用語	sol-gel, calcination, sputtering
Property	物質特性・理論に関する用語	electrical, cryogenic, magnetic fields
Material	構造・記述子に関する用語	tetragonal, P4/nmm, bulk, film, grain

表 A.2 固有表現クラスの解析結果

クラス	出現回数
Characterization	1,793
Process	2,142
Property	11,018
Material	6,859
Element (Main)	9,666 (1,157)
Value	4,609
SC	4,105

表 A.3 関係クラスの解析結果

クラス	出現回数
Equivalent	655
Target	1,644
Condition	173

表 A.4 Doping イベントの解析結果

項目	出現回数
Doping (Trigger)	2,778
Dopant (Role)	2,037
Site (Role)	470

表 A.5 DyGIE++の実験設定

エポック	50
最適化手法	AdamW
学習率	5e-4
ターゲットタスク	関係抽出
スパン長さ	10

表 A.6 主題材料分類モデルの実験設定

エポック	30
最適化手法	AdamW
学習率	2e-5

表 A.7 ルールにより抽出されたスロットの解析結果

項目	関係のみ	主題材料のみ	合計
材料組成	1,432	1,625	3,057
転移温度	336	402	738
ドーパント	1,158	1,266	2,424
サイト	273	313	586
全スロット	1,432	1,625	3,057