超伝導材料に関する固有表現コーパスの構築

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]rasahi@mosk.tytlabs.co.jp

1 はじめに

近年，物質科学分野においてマテリアルズインフォマティクス (Mateirals Infomatics: MI) の研究が活発化している．この背景として，多くの材料研究者が機械学習を活用した新規材料探索の効率化を目指していることが挙げられる．新規材料探索に際しては，材料組成や適切なドーピング，プロセス条件を選択する必要があり，コストが大きく多大な時間を要するという問題点がある．こうした探索を手探りで行っていた従来の開発プロセスを MI により変革することが期待されているが，現状として利用可能なデータセットはごく少数に限られている [8].

物質科学において超伝導材料は重要な材料の 1 つである．銅酸化物系の高温超伝導体 [1] 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 [6, 4]．材料探索においては，研究方針を決定するために専門家が大量の文献を調査している状況であり，超伝導材料探索の効率化を妨げる要因となっている．

本研究では，超伝導に関する学術文献からの固有表現抽出を行うための文献コーパスを作成し，解析を行った．また，作成したコーパスを元に固有表現認識 (Named Entity Recognition; NER) モデルの学習を行った結果，F 値が 77% となり人手に迫る精度であることが明らかとなった．

2 用語ラベルの定義

学術文献に対して注釈付けを行うにあたり，物質科学の専門家の意見を参考に用語ラベルを検討した．元となる用語ラベルとしてレタージャーナル *scripta materialia* のキーワードリスト[*1]を参照し，専門家がそれを修正することで最終的な用語ラベルを定義した．元のキーワードリストは Synthesis/Processing, Characterization, Material Type, Property/Phenomena, Theory/Computer Simulation/Modeling の 5 つの用語ラベルで構成される．今回はここから修正を加え，以下に示す 7 つの用語ラベルを定義した．

■`Characterization`: 本ラベルは “X-ray diffraction (XRD)” や “scanning electron microscope (SEM)” などの分析手法に関する用語を対象とする．この情報は `Element` や `Property` との関係を見る上で重要である．

■`Process`: 本ラベルは “sol-gel” や “calcination”, “AC/DC sputtering” などの合成・プロセスに関する用語を対象とする．これは密度汎関数理論 (Density Function Theory; DFT) などの理論シミュレーションからは得ることが困難かつ物質科学に重要な情報である．

■`Property`: 本ラベルは “electrical conductivity”, “mechanical hardness”, “electron-phonon coupling”, “Fermi surface” などの材料特性や物性理論に関する用語を対象とする．元のキーワードリストにおいて，材料特性と理論はそれぞれ別の用語ラベルとして分類されているが，場合によってはこれらの識別が難しいため，本研究ではこれらをまとめて `Property` とした．

■`Material`: 本ラベルは “tetragonal crystal symmetry” や “bulk/film”, “grain boundary” などの構造に関する用語を対象とする．これらは `Element` に関する付加情報を得る上で重要である．

■`Element`: 本ラベルは “Ti”, “oxygen”, “$YBa_2Cu_3O_7$” といった元素や化合物を対象とす

[*1] https://www.elsevier.com/__data/promis_misc/SMM%20Keywords.pdf

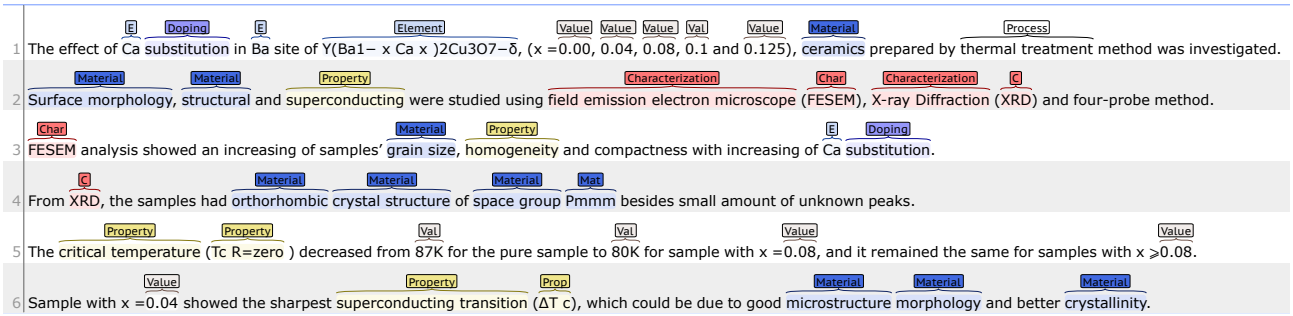

図1　注釈付けの例

る．本ラベルに属するエンティティは化学式で表現できるため，Material と厳密に区別される．

■Doping:　本ラベルは “doping”, “substitute”, “addition” などのドーピング操作を表す用語を対象とする．ドーピング操作は材料特性に与える影響が大きく重要な情報である．

■Value:　本ラベルはドープ量や転移温度といった定量情報を対象としている．

3　コーパス作成の流れ

ここでは，文献データ収集から注釈付けまでの流れについて説明する．本研究では注釈付け作業を 2 回に分けて実施した．1 回目の作業時は実験的に 200 件の抄録を収集し，専門家による注釈付けを行いながら用語ラベルの妥当性を検証した．2 回目の作業時はより大きなデータを扱うため，800 件の追加抄録に対して注釈付けを行った．この際，作業者への負担を考慮し，既にラベルが明らかとなっている用語についてはパターンマッチングに基づく注釈付けを実施し，その後に人手による注釈付けを行った．図1に注釈付けの例を示す．

4　コーパス解析

ここでは作成したコーパスの概要情報と作業者間一致率について説明する．

4.1　コーパス概要

表1に作成したコーパスの概要を示す．下表から Material, Property, Element の用語が頻出しており，Process と Doping, Value についても一定数取れていることが分かる．一方で Characterization (Char.) に関しては比較的出現回数が少ない結果となったが，5.3節で述べる通り，モデルを学習する上では問題ない数である．

文献数	1,000
文数	6,639
トークン数	204,884
1 文献当たりの平均文数	6.64
1 文当たりの平均トークン数	30.9
用語ラベル	**出現回数**
Char.	1,789
Material	6,953
Property	15,129
Element	9,526
Doping	2,565
Process	2,173
Value	4,202

表1　コーパス概要

ラベル	適合率 (%)	再現率 (%)	F 値 (%)
Char.	84.4 ± 8.3	88.3 ± 6.2	86.3 ± 7.2
Material	82.0 ± 6.4	65.6 ± 7.0	72.9 ± 6.8
Property	74.5 ± 3.0	71.8 ± 5.2	73.0 ± 3.2
Element	82.3 ± 1.0	79.8 ± 3.2	81.0 ± 1.6
Doping	97.0 ± 0.4	75.2 ± 9.7	84.4 ± 6.2
Process	81.0 ± 7.1	78.4 ± 5.0	79.7 ± 6.0
Value	73.8 ± 3.7	89.1 ± 6.5	80.7 ± 4.9
全ラベル	79.8 ± 2.1	75.3 ± 2.0	77.5 ± 1.9

表2　注釈付け一致率

4.2　作業者間一致率

2 回目の注釈付けについて，作業者間での注釈付けの一致率を評価した．同一の 10 件の抄録について，4 人の作業者にそれぞれ相談なしで注釈付けを行ってもらい，また 1 回目の注釈付けを行った専門家も同様の

ラベル	適合率 (%)	再現率 (%)	F 値 (%)
Char.	78.0 ± 4.2	77.1 ± 4.2	77.4 ± 3.3
Material	75.2 ± 6.8	73.8 ± 2.6	74.3 ± 4.2
Property	73.4 ± 3.8	68.9 ± 1.2	71.1 ± 2.1
Element	84.2 ± 2.4	86.9 ± 2.0	85.5 ± 2.1
Doping	96.0 ± 2.8	96.4 ± 1.5	96.1 ± 1.8
Process	73.9 ± 8.7	72.9 ± 5.3	73.3 ± 6.7
Value	80.7 ± 3.9	63.5 ± 7.8	70.9 ± 5.6
全ラベル	78.4 ± 3.7	75.4 ± 1.8	76.8 ± 2.5

表3　SciBERT NER の精度

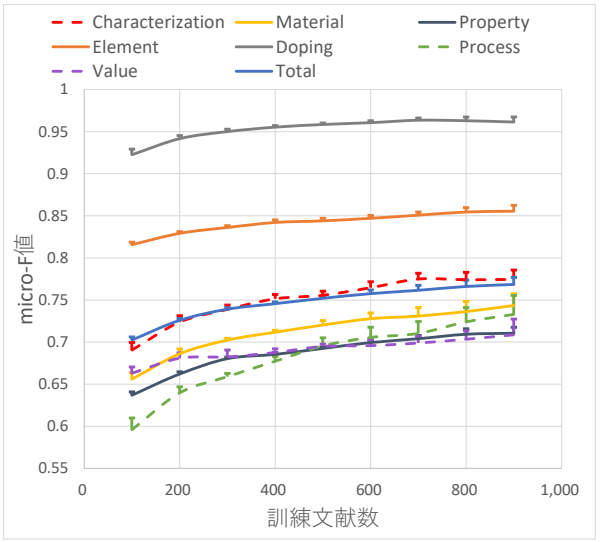

図2　用語ラベル別の学習曲線

10 件に対して注釈付けを行った．結果を表2に示す．スコアは各作業者と専門家の間での一致率を計測したもので，用語のスパンとラベルが完全に一致しているかを基準としている．結果から作業者間でのバラつきと比較すると，用語ラベル毎の一致率にバラつきが出ていることが分かる．具体的には，Characterization と Doping には高い一致率がある一方で，Material と Property については区別が難しく注釈付けが曖昧になる場合があるため，一致率が比較的低い傾向がある．識別が難しい用語の例として “anisotropic” と “order” などが挙げられる．こうした場合，作業者は分野の知識に基づいて文脈を考慮する必要がある．

更に評価指標として Fleiss の κ 係数を計算し，専門家の注釈とは独立した 4 人の注釈者間の一致率を測定した．スコアは BIO ラベルに基づくトークンレベルで算出し，4 人の作業者の全員のペアの組み合わせが考慮されている．結果として κ 係数は 83.1% となり，比較的高い一致率であることが分かった．

2 回目の注釈付け作業の後，800 件の注釈付けについて作業者全員で再検討し，必要に応じて修正を加えた．このため最終的なコーパスの注釈付け一致率は表2のスコアよりも高くなっている点に留意されたい．

5　NER モデルの構築

5.1　実験設定

近年，*Bidirectional Encoder Representations from Transformers (BERT)* [3] が様々な自然言語処理タスクで使用されており，高い精度を出している．本研究では科学分野の大規模な文献で学習された BERT モデルである SciBERT [2] をベースとした NER モデル[*2]を使用した．今回実験に用いたモデルは SciBERT の研究チームが提供している NER モデルで，SciBERT の単語埋め込みベクトルを入力とし，Bi-LSTM→CRF と続くモデル構成となっている．CRF 層における出力は IOB2 フォーマットであり，ハイパーパラメータは既定で設定されているものを適用した．またデータの前処理として，ネストした用語ラベルを取り除く処理を施した．

5.2　NER モデルの精度

10 分割交差検証により，モデルの評価を行った．モデル訓練時には，SciBERT モデルパラメータは固定とし，Bi-LSTM と CRF のみで学習を行った．評価結果を表3に示す．表中のスコアは学習モデルの 10 回平均と標準偏差を表している．全ラベルを通した一致率については，人手による注釈付けの一致率に匹敵する精度であり，実用レベルの精度であることが確認できた．

5.3　学習曲線

今回使用した訓練データの数が十分であるか確認するため，訓練データの数を 100 件から 900 件まで 100 件ずつ変化させた際のモデル精度の変化を調査した．結果を図2に示す．図中のプロットは各学習件数における 10 回平均と標準誤差を表しており，訓練件数が増えるに従ってスコアが上昇していることが分かる．900 件で学習した場合，Doping, Characterization, Property

[*2] https://github.com/allenai/scibert

のスコアが飽和している一方で，Process をはじめとするその他の用語ラベルについてはスコアが上昇し続ける結果となった．また Characterization については，4.1節で述べたようにコーパス全体での出現回数は少ないが，スコアが飽和しているため問題ないと言える．しかしその一方で，人間の作業者の一致率を大きく下回っており，モデル自体の性能については改善の余地がある．

6 関連研究

物質科学分野における NER コーパスとして，[9, 7, 5] がある．文献 [9] では，材料名，材料特性など，本研究と類似する用語ラベルを対象としている一方で，一般的な無機材料を対象としている点，定量情報を扱っていない点が本研究と異なる．参考までに作業者間の注釈付け一致率は 87.4% であった．文献 [7] は，材料物質の合成プロセスをイベントとして抽出するコーパスである．ここでは，論文中の実験の章を対象としているため，超伝導に関連する広範囲を扱う本研究の目的には適していない．作業者間の一致率は 20.5–97.1% であり，これは用語ラベルに大きく依存している．文献 [5] は本研究と同様の超伝導材料に関する情報抽出を目的としている．この研究では 5 つの論文に対して材料名と転移温度に注釈付けを行い，これらの用語ペアを取得している．しかしながら，本研究の実験結果から 1,000 件程度の文献データが必要であることが明らかとなり，データ数が十分とは言えない．

7 おわりに

本研究では，超伝導に関する固有表現抽出を目的とした注釈付きコーパスの構築を行った．本コーパスは基板材料やプロセス情報だけでなく，ドーピングや定量情報などを対象としており，より実用的な内容となっている．コーパスの解析を行ったところ，作業者間一致率は約 75–85% となり，MI 関連の他のコーパス [9, 7] と比較的類似する結果となった．また，SciBERT をベースとする NER モデルを構築した結果，F 値が約 77% となり，人手による注釈付け一致率に迫る精度を実現した．さらに，訓練データ数を変化させた学習曲線から，コーパスサイズが十分であることが明らかとなった．今後の研究方針として，関係抽出や文書要約などでの活用，新規超伝導材料探索を支援する用語およびドキュメント検索システムの構築を検討している．

参考文献

[1] 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.

[2] Beltagy et al. SciBERT: A pretrained language model for scientific text. In *Proceedings of EMNLP-IJCNLP*, pp. 3613–3618, 2019.

[3] Devlin et al. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*, pp. 4171–4186, Minneapolis, Minnesota, 2019.

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

[5] Foppiano et al. Proposal for automatic extraction framework of superconductors related information from scientific literature. *Letters and Technology News*, Vol. 119, No. 66, pp. 1–5, 2019.

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

[7] Mysore et al. The materials science procedural text corpus: Annotating materials synthesis procedures with shallow semantic structures.

[8] Ramprasad et al. Machine learning in materials informatics: recent applications and prospects. *npj Comput. Mater.*, Vol. 3, p. 54, 2017.

[9] Weston et al. 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.