# マテリアルズ・インフォマティクスのための材料辞書群の構築

## Materials-dictionary set construction for Materials informatics

国立研究開発法人 物質・材料研究機構[1] ○鈴木 晃[1], 石井 真史[1]

National Institute for Materials Science[1], ○Akira Suzuki[1], [1] Masashi Ishii[1]

E-mail: SUZUKI.Akira3@nims.go.jp

### 1. 背景

マテリアルズ・インフォマティクス(MI)のための学習データの効率的収集を目的に、学術論文からのテキストデータマイニングによる材料辞書セットの自動構築手法を開発している[1,2]。用語の自動抽出には正規表現等を用いたルールベースによる手法と機械学習による自動分類が挙げられ、本研究では、両者の組み合わせによる効率化を検討した。

### 2. 材料辞書セットの構築

#### 2.1 材料辞書セット

材料辞書セットは材料名、試料、手法、理論、物性・単位、略語のカテゴリに分類した辞書セットで、それぞれの辞書では用語に対する類義語、広義概念、狭義概念等の情報を付加する。また、各辞書間のリンクによりどの材料に対しどの手法で測定し、どの物性が得られた等の情報を取得可能にする。

本稿では材料名辞書の構築手法を例として紹介する。

#### 2.2 対象論文

AIP publisher により 2018 年に発行された約 12,000 論文を対象とした。主に、応用物理学、計算科学等が含まれる。各ファイルは Extensible markup (XML)および MathML で記述されているため、タグ情報によるセクション毎の文章抽出が可能であり、文字の上付き、下付き、フォント等の情報も併せて活用できる。

#### 2.3 アーカイブ作成

全文の一律な言語処理では、論文の主旨に沿ったデータ抽出は困難である。そこで、(1)研究トピックの選定、(2)対象セクションの選択によりアーカイブ化を行った。本稿では、拡散データベース [i]の構築を念頭に、(1)“diffusion coefficient”、“activation energy”を含む論文を対象とした。また、材料名が多様過ぎる場合、後述の固有表現抽出（NER）の効率が低下するため、“semiconductor”が含まれる論文に絞った。ただし、論文内には金属、有機材料名等も含まれる。(2)材料名の効率的な取得を考慮し、論文タイトルと“Introduction”とした。選択条件を満たしたものは 12 論文であった。

#### 2.4 アノテーション

アーカイブ内のテキストをスペースや句読点等で分割し（トークン化）各トークンについてタグ付けを行った。材料名辞書に用いるタグとして、材料分類（materialclass）、材料名(material)、構成要素(element)を設定した。

予め既存辞書やルールベースで抽出された用語によりタグ付けを行い、アノテーションツール（WebAnno[ii]）を用いて確認・修正・追加を行った。

#### 2.5 機械学習による辞書の強化

アノテーションデータを段落毎にランダムで訓練データと評価データに分け DeLFT[iii]を使用して BidLSTM-CRF モデルによる NER[3]を行った。訓練データが少ないためか各タグの F 値（精度と再現率の調和平均）は 0.6-0.9 と大きくばらついた。このトレーニング結果を元に新たなアーカーブに対して自動アノテーションを実行し、人による確認・修正を繰り返すことで訓練データを増加させていく。訓練データ量が増えるにつれ学習精度が上がり、より効率的なデータ構築が期待できる。

### 3. まとめ

ルールベースおよび機械学習を組み合わせた材料名辞書の自動構築手法を開発した。少数の訓練データではあるが、効率的な構築が可能であることを確認した。今後、タグセットの増加（結晶構造、相状態等）、他の辞書への適用、対象論文の増加を検討する。

参考文献

[1] A. Suzuki and M. Ishii, Proc. of Third International Workshop on SCIentific DOCument Analysis (SCIDOCA2018) paper 11.
[2] 鈴木 晃, 高山 英紀, 石井 真史, 第 66 回応用物理学会春季学術講演会, 9p-W321-3 (2019).
[3] G. Lample, M. Ballesteros, S Subramanian, K. Kawakami, C. Dyer. Proceedings of NAACL 2016.

[i] https://diffusion.nims.go.jp/
[ii] https://webanno.github.io/webanno/
[iii] https://github.com/kermitt2/delft