

マテリアルズ・インフォマティクスのための材料辞書の構築とデータベース化検討

Materials-dictionary database construction for Materials Informatics

国立研究開発法人 物質・材料研究機構 ○鈴木 晃, 石井 真史

National Institute for Materials Science, ○Akira Suzuki, Masashi Ishii

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)の発展に伴い、学習データの効率的収集が課題となっている。本研究では、学術論文からキーワードマッチングにより学習データを自動取得する手法を検討しており、そのためのキーワード(専門用語)をまとめた材料辞書を構築している[1-3]。本稿では、辞書の開発手法、およびRDF (Resource Description Framework) によるデータベース化と利用例について紹介する。

2. 材料辞書の構築

2.1 材料辞書

材料辞書は学術論文から抽出した専門用語を論文分類や物性データ抽出等の目的に応じて階層分類(辞書化)したものである。上位階層では材料名、試料、手法、理論、物性・単位、略語の辞書に分類し、各辞書の下層に、材料名辞書であれば、材料クラス(semiconductor、oxides等)、材料名(名称、化学組成等)、構成要素(electron、vacancy等)を配置した分類構造とした。上記分類とは別に、各用語について研究トピック(拡散、超伝導等)への関連付け、類似語(表記ゆれ)のグループ化も行った。

2.2 専門用語抽出

物性機構の所有するXML論文約77万本(9出版社)よりテキストおよびXMLタグを抽出した。テキストをトークン化(XMLタグ、スペース、記号等による分割)、および結合(gensim Phrasesⁱ)処理し、単語および複合語リストを作成した。本稿では、まず、複合語の辞書化について試行した。単語リストの辞書化についても別途手法検討中である。

2.3 複合語分類

複合語は最後の単語(以下、baseと記述)がカテゴリ(“activation energy”、“fermi energy”等は“energy”カテゴリ)を表すことが多いため、これを活用して階層化を行った。具体的には、人手によりbaseを抽出し、それらを上述の分類構造に紐づけ、baseを含む複合語をその下層に配置することで階層分類を行い辞書化した。また、例えば拡散分野の場合、“diffusion coefficient”、“activation energy”が含まれる論文を抽出し、それらに含まれる複合語を同一研究

トピックとして関連付けた。さらに、表記ゆれ検索による類似語抽出を行いグループ化した。

2.4 データベース(DB)化

分類された複合語をRDF形式によるトリプル(主語、述語、目的語)で記述した(Fig. 1)。各用語はユニークなURI (Uniform Resource Identifier) により識別される。同一用語に対して複数の情報を羅列できるため、上位階層(subClassOf)、類似語グループ(domain)、さらには人による分類修正等の情報も追加した。また、外部DBとの連携も可能であり、外部URIと紐づけられた用語がどのような分類・定義であるかを参照できる。作成したDBのURIをラベルとして利用することで、論文内の用語に対して上位・下位概念や用語間の関係性等を付与できる。それらの情報をMI用学習データの自動取得に活用できると考える。

```
<rdf:Description rdf:about="http://age.nims.go.jp/mat_dic_base02#f15f13cd92181bf12d258f97e7c7d52852b53df">
  <rdf:type rdf:type="http://www.w3.org/2001/XMLSchema#string">semiconductor_2</rdf:type>
  <rdf:subClassOf rdf:resource="http://age.nims.go.jp/mat_dic_base02#semiconductor"/>
  <rdf:label rdf:datatype="http://www.w3.org/2001/XMLSchema#string">wide-bandgap_semiconductors</rdf:label>
</rdf:Description>

<rdf:Description rdf:about="http://age.nims.go.jp/mat_dic_base02#f15f13cd92181bf12d258f97e7c7d52852b53df">
  <rdf:type rdf:type="http://www.w3.org/2001/XMLSchema#string">wide-band-gap_semiconductor</rdf:type>
  <rdf:subClassOf rdf:resource="http://age.nims.go.jp/mat_dic_base02#semiconductor"/>
  <rdf:label rdf:datatype="http://www.w3.org/2001/XMLSchema#string">semiconductor_2</rdf:label>
</rdf:Description>
```

赤: URI、黒太字: 専門用語、緑: 上位階層、青: 類似語グループ

Fig. 1 An example of RDF.

3. まとめ

大量学術論文から抽出した複合語を階層分類することで材料辞書として構築し、RDFによりデータベース化した。本辞書は、複数の機械可読な情報を有するラベルとしての利用が可能である。今後、ラベル情報の活用によるMI用学習データの自動取得を検討していく。

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] 鈴木 晃, 石井 真史, 第80回応用物理学会秋季学術講演会, 19a-B01-6 (2019).

ⁱ <https://radimrehurek.com/gensim/models/phrases.html>