

材料辞書データベースを使った論文からの大量データ抽出： 材料用語の階層化による体系的自動タグ付け

Automatic data extraction from scientific articles using the materials-dictionary database:
Automated systematically tagging using graph-structuring of technical terms

物材料機構 [○]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

本研究では、マテリアルズ・インフォマティクス(MI)用学習データの効率的収集を目的に、学術論文からの自動用語抽出による材料辞書を構築している[1-4]。本辞書は RDF (Resource Description Framework) を用いたデータベース化により、活用、追加、更新が容易となる設計とした。本稿では、辞書内用語の自動階層構造化による自動論文アノテーションについて検討した。

2. 材料辞書の構築

2.1 材料辞書の自動分類

材料辞書は学術論文から抽出した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[4]。拡散関連 (“diffusion coefficient”、“activation energy”が含まれる)論文よりルールベースでの複合語抽出を行った。まずは、複合語の最後の単語(以下、base と記述)を基にしてカテゴリ分類(材料名、物性名、手法等の独自分類)した。また、複合語を数語に分割し、それらが他の複合語にも含まれているものを親カテゴリとした。例えば、“lattice and GB diffusion coefficient”であれば、base は coefficient で物性名として分類、“GB diffusion coefficient”、“lattice diffusion coefficient”、“GB diffusion”等は他に抽出された複合語にも含まれるためこれらを親クラスとした。他の複合語についても同様の処理を行うことで、Fig.1 に示すようなグラフ構造を自動構築した。

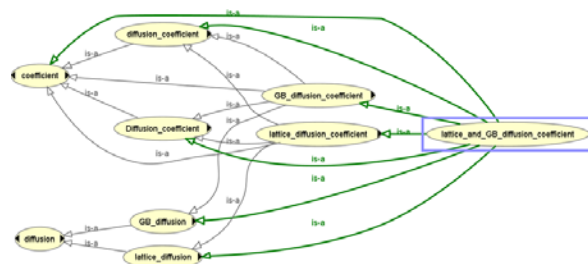


Fig.1 Automatic graph-structure construction for technical terms in scientific articles.

2.2 自動アノテーション

機械学習を行う際、教師データとしてアノテーション(論文内語彙へのラベル付け、および語彙間の関連付け)が必要となるが、論文内表現の曖昧性や目的に応じたガイドラインの調整も必要であるため非常にコストのかかる作業である。そこで、構築した材料辞書データベースによる自動アノテーションを試みた。論文内用語と辞書内用語のマッチング(長い複合語を優先)により URI (Uniform Resource Identifier) と紐づけることにより、目的に応じて親クラス、ラベル種類等を変更した自動アノテーションが可能であることを確認した (Fig.2)。

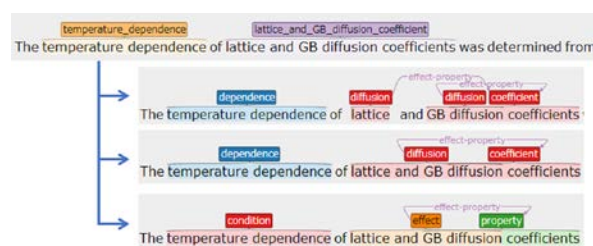


Fig. 2 Automatic corpus construction of the sequential labeling.

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。本辞書の活用により MI 用学習データの自動生成が可能になると考える。

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] 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).
- [4] 鈴木 晃, 石井 真史, 第 67 回応用物理学会春季学術講演会, 15a-A205-4 (2020).