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：物性値取得精度向上の検討

Automatic data extraction from scientific articles using the materials-dictionary database: Improvement of accuracy for obtaining physical properties

物材機構　°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)用学習データを効率的に収集するための技術を構築している。これまで大量の学術論文から材料用語を抽出した材料辞書データベース（MDDB）を構築し、論文内用語に対し体系的に自動タグ付けとタグ間の関連付けを行なう手法を開発した [1,2]。本手法による材料物性値および付随情報の取得精度向上を目的に、MDDB への用語および用語間関係性の追加・修正、およびそれらを効率的に実施するための手法を検討した。

2. 材料辞書 DB の更新

2.1 語彙の追加・修正

MDDB では主に出現頻度を基に用語を抽出しているため、複雑な表記（途中に括弧を含む、カンマ区切りによる元素記号や数値の羅列等）や語数の多い複合語等の抽出に対応しきれていなかった。また、"diffusion coefficient”のみならず、“diffusivity of”、“value of ”、“that of”といった代替表現や“D”、“E_a”のような記号を使った表現等もアノテーションの対象とする必要があることが分かった。そこで、論文内用語の手動アノテーションによる抽出、および正規表現マッチングにより類似表記用語を抽出し MDDB へ追加登録した。

2.2 アノテーションラベルの修正

本 MDDB で付与されたアノテーションラベルについて、修正が必要なものもいくつか見られた。例えば“0 ℃”（0 は数値を正規化）では物性値として数値＋単位（ラベル名：unit）が付与されていたが、“at 0 ℃”となった場合、測定条件（ラベル名：condition）のほうがふさわしいとした。

2.3 関係性付与

“activation energy”と“0 eV”のように頻繁に共起される用語を紐づけし、MDDB に登録することで、自動的に関係性付与をできるようにした。これらの関係性については外部 DB からの流用も検討している。

2.1～2.3 の修正は人手で実施しているが、一度 MDDB に登録すれば他の分野にも流用でき効率化が図れる。

3. 結果とまとめ

図 1 に MDDB 更新前後の自動アノテーション結果の比較を示す。複数値の関係性付与など検討事項は残るが、物性値取得精度の向上が認められた。現在、辞書の拡張、および自動アノテーションの精度の定量評価手法を検討している。

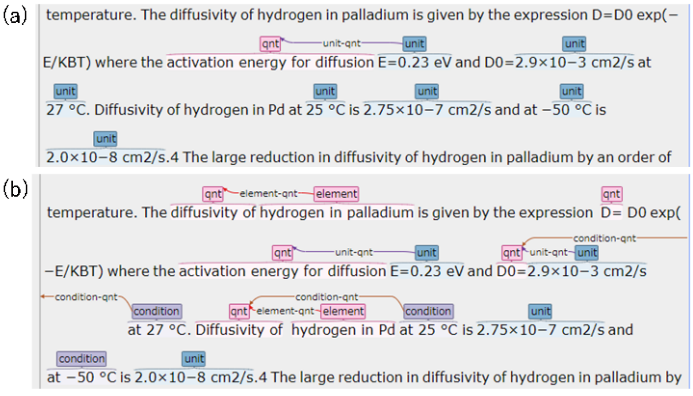

Fig. 1 Comparison of an automatic annotation result: (a) before modification, (b) after modification of Materials dictionary database (MDDB).

【謝辞】

本研究の一部は文部科学省「元素戦略磁性材研究拠点」委託業務の一環として実施されました。

参考文献

[1] 鈴木 晃，石井 真史，第 81 回応用物理学会秋季学術講演会, 9p-Z09-17(2020).

[2] 鈴木 晃，石井 真史，第 68 回応用物理学会春季学術講演会, 19p-Z32-11(2021).