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], 高山 英紀 [1], 石井 真史 [1]

National Institute for Materials Science [1], ◦Akira Suzuki[1], Eiki Takayama[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)では学習データの効率的収集が課題となっており、その有望な手法として学術論文からのテキストデータマイニングが挙げられる。ここでは、マイニングに重要な材料辞書セットの自動作成に関する最近の取り組みを紹介する。

学術論文はある程度定型化されているため、ルールベースでの用語抽出が比較的容易である。そこで本研究ではルールベースによる訓練データ作成と機械学習による強化を組み合わせた固有表現抽出（NER）により、辞書の構築と強化を図る。

2. アーカイブ作成

2.1 対象論文

Journal of Applied Physics (JAP) vol 97, 98 (2005)の 3,490 論文を対象とした。各ファイルは Extensible markup (XML)および MathML で記述されており、これらに含まれるタグを活用して本文、図表キャプションを分類した。

2.2 アーカイブ作成

全文の一律な自然言語処理では、論文の主旨に沿ったデータ抽出は困難である。そこで論文のカギとなる箇所のアーカイブ化をまず行い、そこからのデータ抽出を試みた（図 1）。ここでは、アーカイブの対象を「単位を含む文」と「図表のキャプション」とした場合を紹介する。

2.3 単位を含む文からの RDB 構築

単位を含む文は、論文内で重要な物性名とその数値データを含む可能性が高い。そこで、Math ML タグを使って、数値と単位を含む文章を抽出し[1]アーカイブ化した。これらの文章内の頻出単語および共起語は、興味深いことに物性名を高い確率で示す。これを自動抽出し、単位と紐づけされた物性名としてリレーショナルデータベース（RDB）化する。更に上位階層でデータを分類するため、物性名を基本ベクトルとした約 2,000 次元の空間で単位を表現した。ベクトル類似度を基に物理的に意味が近い単位をクラスタリングすることができる。実際、“kJ⁄mol”と“kcal⁄mol”、あるいは“W⁄m⁄K”と“W m^{-1} K^{-1}”といった表記の異なる物理的同義単位をグループ化することに成功した。

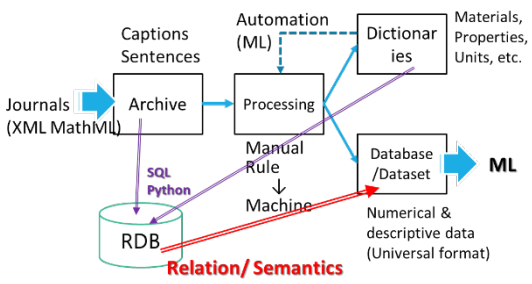

図 1 MI 用データ自動抽出フロー　アーカイブの利用

2.4 図表のキャプションからの RDB 構築

アーカイブの対象を、図表キャプションにすることも有効である。図表は論文の要を簡潔にまとめている可能性が高い。ここでは正規表現や辞書により、キャプションから材料名、特性名、装置名、測定条件等の抽出を行った。ここで、同一図表のキャプションから得られた項目は関連する確率が高く、RDB の自動構築に有効であることが分かった。

3. 機械学習による RDB 強化

2 のルールベースで作成された RDB を基に、文章に機械学習用のタグ（B、I、O）をつけた。このタグに対する NER を実施することにより、ルールベースで関連付けた物性名の妥当性が自動的に判断され、RDB が強化される。こうしてできた材料辞書セットが、最終的に数値の抽出に用いられ、MI 用論文データのマイニングは完結する。現在は、技術シーズが整ったところであり、精度を高めるための文献リソースの増量や対象分野の拡大を進めている。

4. まとめ

MI 用データの効率的収集を目的とした材料辞書セットの自動構築法を検討した。ルールベースで論文から抽出した単位、物性名、試料名、測定条件等を訓練データとした NER により、多様な文章表現への対応が可能となった。

参考文献

[1] A. Suzuki and M. Ishii, Proc. of Third International Workshop on SCIentific DOCument Analysis (SCIDOCA2018) paper 11.