テキストからの物性間関係性抽出への Mat2Vec＆BERT 適用の試み

Mat2Vec & BERT for Extraction of Materials Property Relations from Texts

物材機構[1]　富士通総研[2]　○吉武 道子[1]，佐藤 文孝[2]，河野 洋行[2]

NIMS[1], Fujitsu Research Institute[2], ○Michiko Yoshitake[1], Fumitaka Sato[2], Hiroyuki Kawano[2]

E-mail: yoshitake.michiko@nims.go.jp

マテリアルインフォマティクスの領域では、材料の密度や屈折率などの数値データを機械学習する例がほとんどを占める。言語（自然言語）を対象としたインフォマティクスは、特許文献検索への適用や、科学技術文献などから材料名と数値データをペアにして抽出して数値データの機械学習用の数値データベースを作成する試みなどが少数例ながら存在する。一方、ビジネスの世界では、画像判別のような数値データのみでなく、チャットボットや PRA のように自然言語を対象としたインフォマティクスも盛んである。

Google は、2013 年に単語をベクトル化する Word2Vec を公開し、コンピュータが king-man+woman=queen という答えを出す＝意味がベクトル化されていることを示した。公開されたパラメータファイルは、Google News dataset (約 1000 億語)を用いたもので、一般的なニュースの表現から成る。それに対し、google が公開したアルゴリズムを用い、材料科学分野の論文アブストラクトを学習データとしたパラメータファイル Mat2Vec が作成され、材料科学的文章に適用するには Mat2Vec の方が google 版よりも良いことが示された[1]。

一方、google は Word2Vec をはるかにしのぐ、BERT という「非常に大量に存在する公開された電子テキスト情報を使って pre-training を行うと、fine-tuning と呼ばれる、質疑応答システムや感情分析といった個々のタスクについての学習は少量の学習データでもかなり良い予測エンジンが作成できる」というアルゴリズムを 2018 年に公開した。BERT は Wiki データをベースにパラメータファイルが作成されており、そのパラメータファイルを誰でも利用して、各自のタスクに利用することができる。

著者らは、材料科学関係の教科書的な書籍の自然言語処理により、密度や屈折率、バンドギャップなどの物性の間に存在する関係性を抽出してネットワーク構造としてデータベース化し、それを探索するシステム、マテリアルキュレーション支援システムを開発している。その際の最大の技術的問題点は、コンピュータによる自動抽出では「イオン」のような、「物性を表さない単語」も抽出してしまうことである。今回、コンピュータによる自動抽出結果から手作業で「物性を表さない単語」を除くクレンジング作業を行った結果を正解データとして、Mat2Vec や BERT のパラメータファイルを用いてコンピュータによるクレンジング作業を試みた。

[1] Vahe Tshitoyan, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A. Persson, Gerbrand Ceder, Anubhav Jain, Nature, 571, 95–98(2019).