材料系文書の自然言語処理用 MaterialBERT の作成

MaterialBERT for Natural Language Processing of Materials Science Texts

物材機構[1]，RidgeLinez[2]　○吉武 道子[1]，佐藤 文孝[1,2]，河野 洋行[1,2]，寺岡 潤[1,2]

NIMS-MaDIS[1], RidgeLinez[2], ○Michiko Yoshitake[1], Fumitaka Sato[2], Hiroyuki Kawano[2], Hiroshi Teraoka[2]

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

数値データではなく文字データをコンピュータにより機械学習させるには、文字を数値に変換する必要がある。文字を数値に変換する方法として単語をベクトル化するのが一般的である。Google が 2013 年にベクトル化手法 Word2Vec を発表し、ベクトルの加減で単語の意味が分かると大きな反響を呼び、同じベクトル化手法を材料系論文のアブストラクトをコーパスとして適用した Mat2Vec が作成され[1]学習モデルが公開されている[2]。Google はその後 2018 年に様々な自然言語処理に適用可能な pre-training モデル BERT を発表し[3]、翻訳や要約など様々なタスクに利用されるようになった。我々は、材料系文書の自然言語処理に役立てる目的で、BERT 手法を材料系論文をコーパスとして適用した MaterialBERT の作成を試みた。

NIMS が出版社から購入した XML 形式の論文ファイルのうち、本文が XML で提供されている論文約 75 万報のアブストラクトと本文をコーパスとした。用いた論文は ACS、AIP, APS, ELSEVIER, IOP, JJAP, RSC, SPRINGER, WILEY の 9 社から出版されている学術論文である。語彙数は約 3×10^9 語で、Google が発表した BERT の作成に用いられた語彙数とほぼ同等である。モデルとしては全てを小文字化した１２層のものを用いた。

図に示したのが、pre-trainig 時の学習による損失の変化である。ほぼ順調に低下しており、問題なく学習がされていることが分かる。CoLA(The Corpus of Linguistic Acceptability, https://nyu-mll.github.io/CoLA/)を用いて、文章がもっともらしいかどうかの判断タスクを行わせたところ、BERT の Large と呼ばれる精密な学習モデルのスコアが 60.5 % (https://arxiv.org/pdf/1810.04805.pdf) なのに対し、今回作成した MaterialBERT のスコアは 62.5%となり、上記の損失カーブと合わせて、学習モデルが適切に生成されていると考えられる。今後、材料に特化した学習モデルの利点を示す具体例を計算して論文化し、論文の発行後に、学習モデルを NIMS-MDR（material data repository）にて公開して誰でも利用できるようにする予定である。

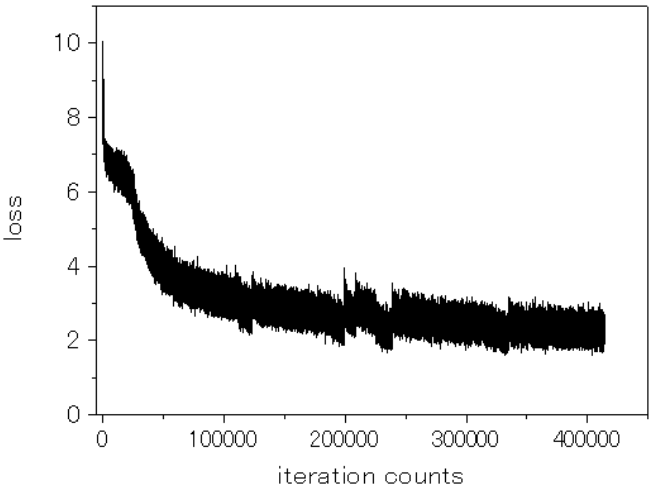

図　MaterialBERT の pre-training 時の学習による損失の変化

[1] Vahe Tshitoyan, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A. Persson, Gerbrand Ceder, Anubhav Jain, Nature, 571, 95 (2019).
[2] https://github.com/materialsintelligence/mat2vec
[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, https://arxiv.org/pdf/1810.04805.pdf

© 2021年 応用物理学会