2021 IEEE International Conference on Big Data (Big Data) | 978-1-6654-3902-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/BigData52589.2021.9671697

Fine-Tuning BERT Model for Materials Named Entity Recognition

Xintong Zhao
Information Science
Drexel University
Philadelphia, USA
xz485@drexel.edu

Jane Greenberg
Information Science
Drexel University
Philadelphia, USA
jg3243@drexel.edu

Yuan An
Information Science
Drexel University
Philadelphia, USA
ya45@drexel.edu

Xiaohua Tony Hu
Information Science
Drexel University
Philadelphia, USA
xh29@drexel.edu

***Abstract*—Scientific literature presents a wellspring of cutting-edge knowledge for materials science, including valuable data (e.g., numerical data from experiment results, material properties and structure). These data are critical for accelerating materials discovery by data-driven machine learning (ML) methods. The challenge is, it is impossible for humans to manually extract and retain this knowledge due to the extensive and growing volume of publications.**

To this end, we explore a fine-tuned BERT model for extracting knowledge. Our preliminary results show that our fine-tuned Bert model reaches an f-score of 85% for the materials named entity recognition task. The paper covers background, related work, methodology including tuning parameters, and our overall performance evaluation. Our discussion offers insights into our results, and points to directions for next steps.

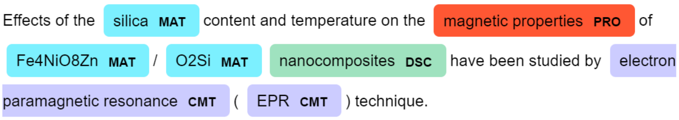
***Index Terms*—named entity recognition, materials science, natural language processing, BERT**

I. Introduction

Materials research outputs show an increased interest in data-driven approaches, drawing on recent advances in machine learning (ML). The enthusiasm in high-throughput computational materials study is hampered by the limited structured materials data for machine learning algorithms. Weston et al. [3] note that although there are increasingly examples of successful computationally designed materials, available structured data is still not sufficient for computational experiments. One potential solution to address this challenge is through automatic extraction of material data from peer-reviewed materials literature by natural language processing (NLP) techniques.

As a major knowledge source for researchers, materials literature seems like an under-explored gold mine. Scholarly literature contains extensive valuable materials knowledge (e.g., synthesis methods, numerical data from experiment results, material structure and properties information) stored in unstructured texts. Even so, materials researchers are only able to manually search and read a small proportion of knowledge recorded in previously established literature due to its extreme volume. According to [4], there are at over nine millions of English scientific literature in the material science area. In this case, manually finding important data in selected literature becomes time-consuming and less efficient. To accelerate the materials knowledge discovery process from literature, we propose automatic knowledge extraction as a candidate solution.

In this study, we focus on extracting materials named entities from literature. As an initial step, we build an end-to-end named entity extraction model by fine-tuning existing Bidirectional Encoder Representations from Transformer (BERT) pretrained language models and analyzed its potential in materials text extraction. Our early phase experiments show promising result: the fine-tuned BERT model achieved an overall f-score of 85.0% in entity extraction. This indicates that the BERT model has significant potential for materials text mining tasks, and has motivated us to build a new BERT model specifically for materials science as next step.



Effects of the silica MAT content and temperature on the magnetic properties PRO of Fe4NiO8Zn MAT / O2Si MAT nanocomposites DSC have been studied by electron paramagnetic resonance CMT (EPR CMT) technique.

Fig. 1. A Visualized Demonstration of Named Entity Recognition task in materials science. Different entity types are highlighted to different colors, where MAT stands for Materials, PRO stands for Material Property, DSC is Descriptor and CMT is Characterization method. The goal of NER is to automatically detect entities that fall into these pre-defined semantic types. This example is from [3] and visualized by SpaCy Python library [16].

II. Background

As the first step toward extracting structured knowledge from unstructured text data, named entity recognition (NER) aims to recognize entity units from input texts to predefined semantic types - its output serves as the foundation for relation extraction, as well as many downstream natural language processing tasks such as information retrieval [17], question answering [18] and knowledge graph construction [19].

Figure 1 is an example of a material entity extraction task. For input sentences from materials literature, our goal is to build an NER model which automatically detects key entities and assigns specific types to each of them. Essentially, the manual reading is replaced by automatic text extraction, expediting the knowledge search process.

This research is funded by the U.S. National Science Foundation, Office of Advanced Cyberinfrastructure (OAC): Grant: #1940239

978-1-6654-3902-2/21/$31.00 ©2021 IEEE

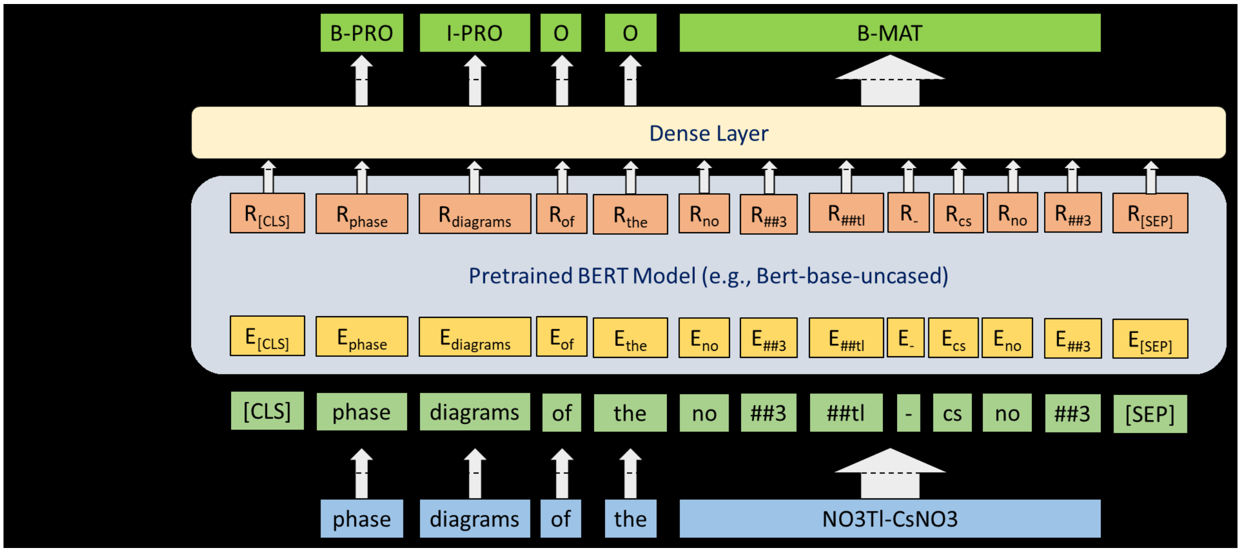

B-PRO
I-PRO
O
O
B-MAT
Dense Layer
$R_{[CLS]}$ R_{phase} $R_{diagrams}$ R_{of} R_{the} R_{no} $R_{\#\#3}$ $R_{\#\#tl}$ R_{-} R_{cs} R_{no} $R_{\#\#3}$ $R_{[SEP]}$
Pretrained BERT Model (e.g., Bert-base-uncased)
$E_{[CLS]}$ E_{phase} $E_{diagrams}$ E_{of} E_{the} E_{no} $E_{\#\#3}$ $E_{\#\#tl}$ E_{-} E_{cs} E_{no} $E_{\#\#3}$ $E_{[SEP]}$
[CLS] phase diagrams of the no ##3 ##tl - cs no ##3 [SEP]
phase
diagrams
of
the
NO3Tl-CsNO3

Fig. 2. The model architecture of fine-tuned BERT for named entity recognition.

Various approaches have been developed for NER tasks in the past two decades. The state-of-the-art has progressed through multiple phases, starting with rule-based approach, followed by ML using handcrafted features, to task-specific deep learning models with word embedding. Currently, deep learning with transformer-based pretrained language models (e.g., BERT) achieves superior performance and hence stands as the new state-of-the-art.

Although the most recent two approaches all involve deep learning, transformer-based language models have several advantages over task specific models. BERT models usually have better adaptability to different NLP tasks with minimal needs for model modification. This is because BERT can transfer contextual knowledge learnt from its large training corpora to solve new tasks. For the same reason, it can achieve a descent performance with less labeled data compared to task specific models.

III. Related Work

Over the past 5 years, machine learning enhanced computational methods have been proven to be effective in various material research topics. Ward et al. [14] show the high potential of ML algorithms in materials discovery by applying multiple ML algorithms (e.g., decision tree, LASSO) to predict properties of general inorganic materials. [9] use ML methods to guide high-throughput experiments for new metallic glass discovery in the Co-V-Zr ternary. ML methods not only show their potential in inorganic materials design, but organic materials too. In 2018, [15] find that deep learning can help synthesis planning for organic materials as well. A key motivator for applying ML in materials science is the Material Genome Initiative [10]. Overall, the application of ML algorithms in materials design is an innovative topic in the materials science community.

However, machine learning algorithms applied in materials research are data-driven in general – namely, they all require a significant amount of structured data as input to achieve their best performance. This is especially true for deep learning models, which heavily rely on large amounts of high-quality labeled data. In 2017, [1] emphasize the limitation of data for ML-enhanced studies. To address this limitation for data-driven approaches, many materials researchers try to extract structured knowledge from unstructured scholarly literature. Various studies develop rule-based methods that could involve regular expression, word dependency information, and text matching to extract either keywords or entity pairs from materials literature [5]–[7]. Weston et al. [3] built a manually labeled corpus containing eight-hundred abstracts and applied a BiLSTM-CRF deep learning model to extract named entities. The use of ontology is also discussed by materials researchers [8].

The research studies noted above successfully demonstrate that datasets can be extracted from literature, but their methods are not without limitations. Two key limitations are: (1) rule-based approaches usually have very little adaptability – their rules are designed for specific topics and scenarios, which restricts their adaptability to other contexts; and (2) many of existing word embeddings for material text mining are context-independent (e.g., Word2Vec, Fasttext), which can limit on the prediction accuracy. We believe the BERT model has the potential to address the above limitations.

To the best of our knowledge, there are still no BERT models trained for materials text mining purposes publicly available. In this case, we take our initial step to explore existing BERT models and analyze their potential toward materials knowledge extraction from text.

IV. Methodology

A. Problem Formulation

As the first step toward automatic knowledge extraction, we formulate named-entity-recognition (NER) as a sequence-to-sequence labeling problem.

Formally, given a natural sentence S consisting of a sequence of N tokens $S = < t_1, t_2, t_3, ..., t_N >$ and their own corresponding labels $L = < l_1, l_2, l_3, ..., l_N >$, the goal is to predict the list of tuples $< t_i, l_i >$ from input sentences, where entity types are not given.

B. Fine-Tuning BERT for NER Tasks

We used two existing BERT pretrained language models in this study: original BERT base model [2] and SciBert [12]. We fine-tuned both models on the training and developing set from the benchmark dataset (see description in section V). The original Bert model is trained on 11,308 novel books and English widipedia content; SciBert is trained on a random sample of scholarly articles, which consists of 18% computer science papers and 82% of biomedical research articles.

The overall model architecture is shown in figure 2. In brief, the model takes natural language sentences as sequences of tokens as input - each sequence of tokens will be further divided into sub-tokens such as "no", "##3", "##Tl"(which are originally from the token NO3Tl), then feed into the BERT transformers architecture for generating context-aware language representations. A dense layer is added on top of word representations to predict the entity type (label) of each input word. The model architecture is built by Pytorch [13], Python 3.

C. Evaluation Methods

In this study, we use precision, recall and f-score to evaluate the NER performance of fine-tuned BERT models. These measures are defined in the equations below.

$$\text{precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \tag{1}$$

$$\text{recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \tag{2}$$

$$\text{f-score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{3}$$

V. Experiment Details

A. Dataset

In this study, we select the MatScholar dataset released by [3], which is a manually annotated corpus consisting of 800 abstracts from literature of inorganic material studies. The MatScholar corpus have seven(7) predefined entity types shown in table I below. We lowercased the entire dataset during our fine-tuning process.

TABLE I
A Glance at MatScholar Dataset

Entity Type	**Example(s)**	**Num**
Inorganic Material (MAT)	Fe4NiO8Zn, NiZn ferrite	682
Symmetry / Phase Label (SPL)	cotunnite phases	75
Sample Descriptor (DSC)	nanocomposites, nanotubes surface	437
Material Property (PRO)	magnetic properties	772
Material Application (APL)	ethanol sensor	170
Synthesis Method (SMT)	acid-assisted hydrothermal method, sputtering	171
Characterization Method (CMT)	electron paramagnetic resonance	195

B. Hyperparameters

We use the following model hyperparameters values suggested by [2] during fine-tuning process: we set our model learning rate to 5e-5, warm-up proportion as 0.1, maximum sequence length as 128. We use Adam as the optimizer and we enabled GPU acceleration with batch size 16.

C. Performance Evaluation

The detailed performance report is shown in table II. Overall, both fine-tuned models have achieved total accuracy over 80%, fine-tuned SciBert model outperforms original Bert model by 2.1% and reached 85.0%.

TABLE II
Performance Evaluation on Two Fine-Tuned Bert NER Models

Evaluation Results on Test Set (original Bert/SciBert)				
Type	Precision	Recall	F-score	Support
MAT	0.875 / **0.880**	0.915 / **0.915**	0.895 / **0.897**	682
SMT	0.760 / **0.770**	0.813 / **0.842**	0.785 / **0.805**	171
APL	0.817 / **0.859**	0.788 / **0.824**	0.802 / **0.841**	170
DSC	**0.874** / 0.849	**0.922** / 0.911	**0.898** / 0.879	437
PRO	0.767 / **0.813**	0.792 / **0.811**	0.779 / **0.812**	772
CMT	0.747 / **0.809**	0.846 / 0.867	0.793 / **0.837**	195
SPL	0.762 / **0.785**	**0.853** / 0.827	0.805 / 0.805	75
Total	0.816 / **0.837**	0.855 / **0.865**	0.835 / **0.850**	2502

VI. Discussion and Future Work

This research prompts us to ask: *What makes the performance of BERT models different?* One significant factor could be their different training corpora. As stated in earlier section, the original BERT model was trained on text in generic domains such as novels, whereas SciBert was mainly trained on scientific biomedical research articles. The language used in corpora from different domains could be very different, which results in a difference in the vocabularies learnt by model. As mentioned in [12], there are only 30% common vocabularies between SciBert and original Bert model.

During our fine-tuning process, we observed that original Bert model was unable to find a small set of sub-tokens in its vocabulary (hence labeled as [UNK]); we did not notice any unknown sub-tokens marked by Scibert model. SciBert model outperform the original Bert model in most of entity types on the highly domain specific Matscholar Dataset. This result is consistent with our above assessment.

As the result of our early-phase experiment, the fine-tuned SciBert model result reaches 85% f-score. This is a solid result, particularly given that neither SciBert nor original BERT were not trained on materials-related corpora. Based on our experiment result, we believe a BERT model trained on solely materials scholarly literature is likely to achieve even higher performance, and has great potential to contribute to various NLP tasks toward knowledge discovery from materials literature. We are aware that previous research [3] using task-specific model with Word2Vec embedding trained on materials articles achieved a slightly higher performance (87%). However, BERT models are more advantageous than task-specific models, since they can be fine-tuned to adapt different NLP tasks, such as relation extraction and question-answering [2]. Our early phase work confirms the usefulness of the approach underlying BERT. A logical next step is to build a brand new BERT model trained solely on materials research corpora, which could serve as a foundation of future materials text mining tasks.

VII. Conclusion

Overall, the research presented here analyzes the potential of BERT models in materials text mining tasks. We performed our initial exploration on materials named entity recognition. The fine-tuned Bert model achieved a solid performance without knowing much materials-related contextual knowledge. Given Bert model is designed to adapt a range of different NLP tasks instead of focusing on a specific task, we believe this transformer-based pretrained language model is a promising direction toward more sophisticated NLP system for materials knowledge extraction from literature.

References

[1] E. Kim, K. Huang, A. Saunders, A. McCallum, G. Ceder, and E. Olivetti, ”Materials synthesis insights from scientific literature via text extraction and machine learning,” Chemistry of Materials, vol. 29, no. 21, pp. 9436-9444, 2017.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, ”Bert: Pre-training of deep bidirectional transformers for language understanding,” arXiv preprint arXiv:1810.04805, 2018.

[3] L. Weston et al., ”Named entity recognition and normalization applied to large-scale information extraction from the materials science literature,” Journal of chemical information and modeling, vol. 59, no. 9, pp. 3692-3702, 2019.

[4] X. Zhao, S. Lopez, S. K. Saikin, X. Hu, and J. Greenberg, ”Text to Insight: Accelerating Organic Materials Knowledge Extraction via Deep Learning,” arXiv preprint arXiv:2109.12758 , 2021.

[5] S. Shah, D. Vora, B. Gautham, and S. Reddy, ”A relation aware search engine for materials science,” Integrating Materials and Manufacturing Innovation, vol. 7, no. 1, pp. 1-11, 2018.

[6] S. Huang and J. M. Cole, ”A database of battery materials auto-generated using ChemDataExtractor,” Scientific Data, vol. 7, no. 1, pp. 1-13, 2020.

[7] C. J. Court and J. M. Cole, ”Auto-generated materials database of Curie and Néel temperatures via semi-supervised relationship extraction,” Scientific data, vol. 5, no. 1, pp. 1-12, 2018.

[8] E. Kim, K. Huang, O. Kononova, G. Ceder, and E. Olivetti, ”Distilling a materials synthesis ontology,” Matter, vol. 1, no. 1, pp. 8-12, 2019.

[9] F. Ren et al., ”Accelerated discovery of metallic glasses through iteration of machine learning and high-throughput experiments,” Science advances, vol. 4, no. 4, p. eaaq1566, 2018.

[10] A. Jain, K. A. Persson, and G. Ceder, ”Research Update: The materials genome initiative: Data sharing and the impact of collaborative ab initio databases,” APL Materials, vol. 4, no. 5, p. 053102, 2016.

[11] Z. Jensen et al., ”A machine learning approach to zeolite synthesis enabled by automatic literature data extraction,” ACS central science, vol. 5, no. 5, pp. 892-899, 2019.

[12] I. Beltagy, K. Lo, and A. Cohan, ”Scibert: A pretrained language model for scientific text,” arXiv preprint arXiv:1903.10676, 2019.

[13] A. Paszke et al., ”Pytorch: An imperative style, high-performance deep learning library,” Advances in neural information processing systems, vol. 32, pp. 8026-8037, 2019.

[14] L. Ward, A. Agrawal, A. Choudhary, and C. Wolverton, ”A general-purpose machine learning framework for predicting properties of inorganic materials,” npj Computational Materials, vol. 2, no. 1, pp. 1-7, 2016.

[15] M. H. Segler, M. Preuss, and M. P. Waller, ”Planning chemical syntheses with deep neural networks and symbolic AI,” Nature, vol. 555, no. 7698, pp. 604-610, 2018.

[16] M. Honnibal and I. Montani, ”Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing,” Unpublished software application. https://spacy. io, 2017.

[17] J. Guo, G. Xu, X. Cheng, and H. Li, ”Named entity recognition in query,” in Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, 2009, pp. 267-274.

[18] D. M. Aliod, M. van Zaanen, and D. Smith, ”Named Entity Recognition for Question Answering,” in ALTA, 2006, pp. 51-58.

[19] O. Etzioni et al., ”Unsupervised named-entity extraction from the web: An experimental study,” Artificial intelligence, vol. 165, no. 1, pp. 91-134, 2005.