SEED: Semantic Knowledge Transfer for Language Model Adaptation to Materials Science

Yeachan Kim[1] **Jun-Hyung Park**[2] **SungHo Kim**[1]
Juhyeong Park[1] **Sangyun Kim**[1] **SangKeun Lee**[1,3]
[1]Department of Artificial Intelligence, Korea University, Seoul, Republic of Korea
[2]Division of Language & AI, Hankuk University of Foreign Studies, Seoul, Republic of Korea
[3]Department of Computer Science and Engineering, Korea University, Seoul, Republic of Korea
{yeachan, sungho3268, johnida, silky, yalphy}@korea.ac.kr, jhp@hufs.ac.kr

Abstract

Materials science is an interdisciplinary field focused on studying and discovering materials around us. However, due to the vast space of materials, downstream datasets in this field are typically scarce and have limited coverage. This inherent limitation poses challenges when adapting pre-trained language models (PLMs) to materials science, as existing methods rely heavily on frequency information from these limited datasets. In this paper, we propose **Se**mantic Knowl**ed**ge Transfer (SEED), a novel vocabulary expansion method designed to adapt pre-trained language models (PLMs) for materials science. The core strategy of SEED is to transfer materials knowledge from lightweight embeddings into PLMs. To achieve this, we introduce knowledge bridge networks, which learn to transfer the latent knowledge embedded in materials-specific embeddings into representations compatible with PLMs. By expanding the embedding layer of PLMs with these transformed embeddings, the models can comprehensively understand the complex terminology associated with materials science. We conduct extensive experiments across a broad range of materials-related benchmarks. The comprehensive evaluation results convincingly demonstrate that SEED mitigates the limitations of previous adaptation methods, showcasing the efficacy of embedding knowledge transfer into PLMs.[1]

1 Introduction

The pre-training and fine-tuning paradigm of language models is widely adopted in natural language processing (NLP). However, since pre-training is typically performed on general-domain corpora, such as Wikipedia, the adaptability of pre-trained language models (PLMs) is limited when the target domains differ significantly from the pre-training domains. This limitation presents a particular challenge in the field of materials science, which encompasses a wide range of domain-specific jargon and complex chemical formulas (e.g., $(La_{0.8}Sr_{0.2})_{0.97}MnO_3$).

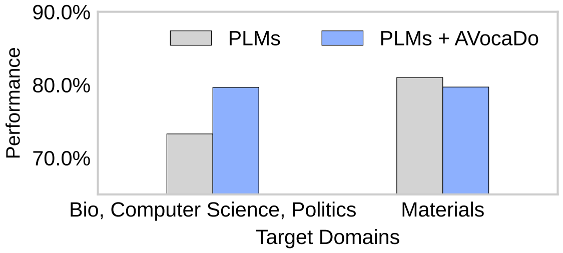

Figure 1: Adaptation performance of the state-of-the-art method (Hong et al., 2021) across different domains: non-materials domains (Biology, Computer Science, Politics) and materials domains. Detailed performance results can be found in the Appendix.

One promising approach to enhance the adaptability of PLMs is to expand the coverage of vocabulary. For example, previous works have expanded the vocabulary of PLMs by considering the frequency information of downstream datasets (Hong et al., 2021; Yao et al., 2021). However, such a frequency-based approach can be suboptimal in materials science, as downstream datasets in this domain are typically scarce and limited in coverage (Song et al., 2023). Indeed, we experimentally observe that a state-of-the-art optimization method (i.e., AVocaDo (Hong et al., 2021)) rather degrades the performance of the original model[2]. Figure 1 illustrates that AVocaDo yields poor adaptation results in materials science, while significantly enhancing the performance of PLMs in other domains (e.g., biomedical, computer science, politics), underscoring the unique challenges of adaptation to the materials science domain.

[1] Our code is available at https://github.com/yeachan-kr/seed

[2] We also show that other vocabulary expansion methods fail in adapting PLMs to materials science (Section 4).

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 421–428
November 12-16, 2024 ©2024 Association for Computational Linguistics

In response, we propose **Se**mantic Knowl**ed**ge Transfer (SEED), a novel method designed to optimize the vocabulary and embedding layer of PLMs for materials science. Specifically, unlike prior works that rely on frequency information from downstream datasets, SEED utilizes latent knowledge within the materials science corpus to adapt the PLM's vocabulary. Given that pre-training models on a large corpus incurs significant adaptation costs, SEED leverages *Mat2vec* (Tshitoyan et al., 2019), lightweight word embeddings trained on materials science journals. To bridge these two distinct types of knowledge representations, we introduce bridge networks that transfer the materials knowledge from *Mat2vec* into PLMs. With the transferred knowledge from *Mat2vec*, PLMs can be effectively adapted to materials science domains.

To verify the efficacy of SEED, we conduct extensive experiments across diverse benchmarks in materials science, including materials entity recognition, slot filling, and glass classification, using various PLM backbones. The evaluation results demonstrate that SEED effectively mitigates the inherent limitations in adapting PLMs for materials science. Additionally, we observe that the transferred embeddings are closely aligned with the original embeddings in PLMs, confirming the successful knowledge transfer achieved by SEED . In summary, the contributions of this paper include the following:

- We discover that existing adaptation methods fail in the field of materials science due to the distinct challenges of materials science.
- We propose SEED, a novel vocabulary expansion method by transferring the latent knowledge of external materials embeddings.
- We demonstrate that SEED outperforms the existing methods, underscoring the efficacy of the knowledge transfer approach in adapting PLMs for materials science.

2 Related Work

2.1 NLP for Materials Science

The growing number of textual datasets in materials science, such as scientific papers and patents, has facilitated the use of NLP-based approaches to address materials-related downstream tasks, spanning relation classification (Mysore et al., 2019; Mullick et al., 2024) and materials entity extraction (Weston et al., 2019; Friedrich et al., 2020). For instance, Weston et al. (2019) performed named entity tagging for materials science tetrahedron by learning a bidirectional LSTM tagger. In exploring unsupervised approaches to materials science, Tshitoyan et al. (2019) demonstrated promising results with a word2vec approach (Mikolov et al., 2013) for understanding chemical properties and broader chemistry knowledge. Trewartha et al. (2022) introduced language models pre-trained on materials science journals using the BERT framework (Devlin et al., 2019). Similarly, Gupta et al. (2022) and Huang and Cole (2022) adapted SciBERT (Beltagy et al., 2019) and BERT (Devlin et al., 2019), respectively, for use in general materials science and battery-focused downstream tasks.

2.2 Vocabulary Expansion of PLMs

Expanding the original vocabulary with domain-specific words has been getting significant attention, as it enables the efficient adaptation of PLMs without the non-trivial costs associated with pre-training on domain-specific corpora (Tai et al., 2020; Zhang et al., 2020; Yao et al., 2021). For example, Tai et al. (2020) extended the vocabulary of PLMs to biomedical domains by learning a new WordPiece (Wu et al., 2016) on biomedical corpus. Similarly, Hong et al. (2021) selected the additional subwords from the downstream datasets and fine-tuned the added embeddings with contrastive learning. Yao et al. (2021) and Kajiura et al. (2023) also adopted the same approach to vocabulary expansion, where the frequency information in the downstream datasets is leveraged to expand the vocabulary. However, given that downstream datasets in materials science are typically limited and scarce (Song et al., 2023), relying solely on frequency information of these datasets can result in sub-optimal adaptation of PLMs.

3 SEED: Semantic Knowledge Transfer

In this work, we elaborate on Semantic Knowledge Transfer (SEED). The key strategy of SEED involves transferring the knowledge from external materials embeddings into PLMs. To achieve this, we begin with the words shared between the vocabularies of materials embeddings and PLMs (§3.1). We then train bridge networks to ensure that the semantic relations of the shared words are transferred to PLMs (§3.2). After training, we transfer the materials knowledge only existed in the materi-

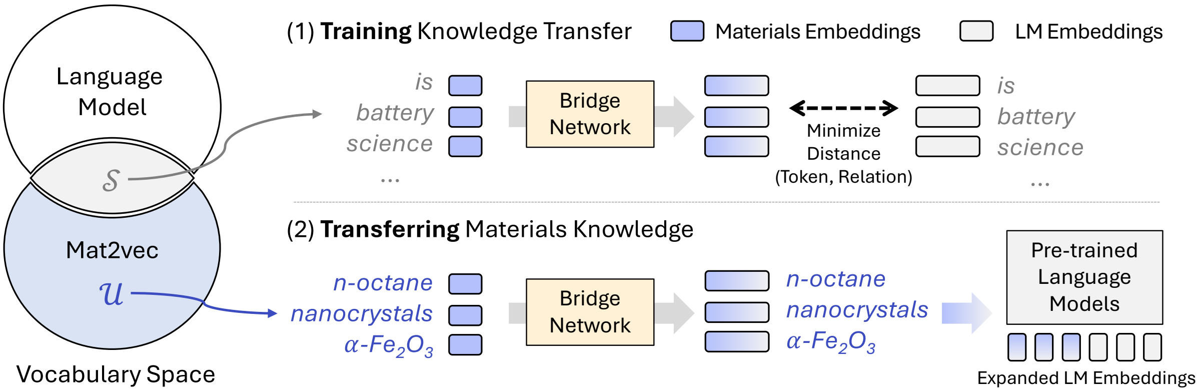

Figure 2: Overall adaptation process of PLMs with SEED. Starting from the shared vocabulary ($\mathcal{S}$) between PLMs embeddings and materials ones, we train the bridge network to transform *Mat2vec* into compatible representations with PLM's embeddings. After converged, words in the unique vocabulary ($\mathcal{U}$) are transformed through the bridge networks. The transferred embeddings are then interleaved to the embedding layer of PLMs.

als embeddings into the PLMs through the learned bridge network (§3.3). The overall procedures of SEED are illustrated in the Figure 2.

3.1 Vocabulary Alignment between Materials Embeddings and PLMs

Unlike the previous works that solely rely on frequency information of the downstream datasets (Yao et al., 2021; Hong et al., 2021), we leverage *Mat2vec* (Tshitoyan et al., 2019) to expand the knowledge of the PLMs. Specifically, we use the skip-gram version (Mikolov et al., 2013) of *Mat2vec* trained on scientific papers, which includes 200-dimension vectors for 500k words[3].

To transfer the knowledge of *Mat2vec*, we first decompose the vocabulary of the materials embeddings into two disjoint sets: a shared set $\mathcal{S}$ and a unique set $\mathcal{U}$. The words in $\mathcal{S}$ appear in both the materials embeddings and the PLMs' vocabularies, while the words in $\mathcal{U}$ only appear in the materials embeddings. We target the transfer of unique materials knowledge without disrupting the existing knowledge structure of the PLMs. Additionally, to mitigate the negative impact of over-expansion, we only consider target words that are originally tokenized into more than four tokens.

3.2 Bridge Networks for Knowledge Transfer

Bridge Networks To transfer the knowledge of the materials embeddings, we introduce bridge networks that learn to transform these embeddings into ones compatible with the PLMs. Let the embeddings in *Mat2vec* and PLMs be denoted as E_{M} and E_{P}, respectively, we first transform the E_{M} as follows:

$$E_{\mathrm{M}\rightarrow\mathrm{P}}(w) = \alpha(E_{\mathrm{M}}(w)), \forall w \in \mathcal{S} \qquad (1)$$

where α represents the bridge networks, which consists of two-layer feed-forward networks, and $E_{\mathrm{M}\rightarrow\mathrm{P}}$ indicates the transformed representations from the materials embeddings. The input and output dimensions of the bridge network α are aligned with the materials embeddings and those in PLMs.

Optimizing Bridge Networks To optimize the bridge network such that the transformed embeddings are compatible with the PLMs, we optimize the bridge networks through the following reconstruction loss as follows:

$$\mathcal{L}_{\text{recon}} = \|E_{\mathrm{P}}(w) - E_{\mathrm{M}\rightarrow\mathrm{P}}(w)\|_2^2, \forall w \in \mathcal{S} \qquad (2)$$

However, we empirically observe that optimizing the parameters solely based on the aforementioned reconstruction loss leads to sub-optimal transformation of materials embeddings. Inspired by relational knowledge distillation (Park et al., 2019), we also introduce additional objectives to consider the relations with other words. Specifically, let the distance function of the embeddings x and y be denoted as $\psi(x, y)$[4], the loss function to inject the relations between words is as follows:

$$\begin{aligned}\mathcal{L}_{\text{rel}} = \delta(\psi(E_{\mathrm{P}}(w_i), E_{\mathrm{P}}(w_j)), \qquad (3)\\ \psi(E_{\mathrm{M}\rightarrow\mathrm{P}}(w_i), E_{\mathrm{M}\rightarrow\mathrm{P}}(w_j))),\end{aligned}$$

[3] Details for *Mat2vec* is described in the Appendix.

[4] While we have a number of design choices, we used the l2 distance function in this work. Exploration on diverse distance metrics and more relations can be a promising future work.

Algorithm 1 Semantic Knowledge Transfer

```
Input: Materials Vocabulary $V_M$ and Embeddings $E_M$, PLMs' Vocabulary $V_P$ and Embeddings $E_P$, Bridge network $\alpha$, L2 distance $L_2$, Relational distance $L_R$
$\mathcal{S} \leftarrow$ JointVocab($V_M$, $V_P$)
$\mathcal{U} \leftarrow$ DisjointVocab($V_M$, $V_P$)
# Learning bridge networks
for word $w$ in $\mathcal{S}$ do
    $E_{M\rightarrow P}(w) \leftarrow \alpha(E_M(w))$
    $\mathcal{L}_{\text{recon}} \leftarrow L_2(E_{M\rightarrow P}(w), E_P(w))$
    $\mathcal{L}_{\text{rel}} \leftarrow L_R(E_{M\rightarrow P}(w), E_P(w))$
    Optimize $\alpha$ based on $\mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{rel}}$
end for
# Transferring knowledge of $V_M$ to $V_P$
for word $w$ in $\mathcal{U}$ do
    $E_{M\rightarrow P}(w) \leftarrow \alpha(E_M(w))$
end for
return Transferred embeddings $E_{M\rightarrow P}$
```

where w_i and w_j are randomly selected materials in batch, and $\delta(x, y)$ is Huber loss (Huber, 1992) that is defined as follows:

$$\delta(x,y) = \begin{cases} \frac{1}{2}(x-y)^2 & \text{if } |x-y| \leq 1, \\ |x-y| - \frac{1}{2}, & \text{otherwise.} \end{cases} \quad (4)$$

By combining the two loss functions (i.e., $\mathcal{L}_{\text{recon}}$, $\mathcal{L}_{\text{rel}}$), the bridge network learns the mapping function between the knowledge of materials embeddings and PLMs.

3.3 Adapting SEED to Downstream Tasks

Transfer Knowledge Selection After the optimization of the bridge networks converges, we transfer knowledge from the unique set $\mathcal{U}$ absent in the PLMs by feeding their embeddings into the bridge networks and placing them into the embedding layer of PLMs. However, since the vocabulary size of the materials' embeddings is substantially larger than that of the PLMs, transferring all words would require significant memory overhead in the PLMs. Following previous work (Hong et al., 2021), we selectively transfer the knowledge of the words in a task-specific manner. Specifically, we extract all words from the training set and expand this list by searching for similar words using the materials embeddings to identify these similar terms.

$$\mathcal{U} \leftarrow \{\text{TopK}(w) \mid w \in D, w \notin \mathcal{S}\} \quad (5)$$

where TopK(w) indicates the function that returns k words that are most similar to the given word w, the similarity measure is the cosine similarity based on the materials embeddings, and D is the word list in the downstream dataset. After narrowing down the unique set based on the downstream dataset, we transfer the knowledge of materials embeddings to the PLMs through the trained bridge network. With the expanded embeddings and vocabulary, the PLMs are adapted to downstream tasks through a typical fine-tuning process.

Optimization Following the previous work by (Hong et al., 2021), we introduce a contrastive regularization term that encourages representations derived from expanded embeddings not to deviate from the original embeddings. The overall algorithm of SEED is described in Algorithm 1.

4 Experiments

In this section, we experimentally demonstrate the efficacy of SEED in adapting PLMs to downstream tasks. Specifically, we mainly focus on whether SEED mitigates the limitations of vocabulary expansion methods in materials science.

4.1 Experimental Setups

Baselines The goal of SEED is to effectively adapt the PLMs to the downstream tasks in materials science by optimizing vocabulary and its embeddings. To confirm the effectiveness, we mainly compare ours with the three strong baselines with the backbone: **AdaLM** (Yao et al., 2021), **AVocaDo** (Hong et al., 2021), **Replace** (Kajiura et al., 2023). **AdaLM** adapts the PLMs to specific domains by expanding the vocabulary based on the frequency of the subwords. While this method includes the distillation phase to train the smaller domain expert model, we only apply the vocabulary expansion algorithm to fairly compare the effectiveness of the vocabulary expansion. Similarly, **AVocaDo** considers the frequency information of subwords in the downstream datasets with the contrastive learning designed to stabilize the training. **Replace** selects frequent words in downstream datasets, and the less frequent words in vocabulary are replaced with the new frequent words. For the setups of SEED, we list the selected parameters and search space in the Appendix.

Downstream Tasks and Datasets To demonstrate the diverse aspects, we evaluate each method

Table 1: Evaluation results on four materials benchmarks based on BERT (Devlin et al., 2019). For SOFC and MatScholar, the reported performances are Macro-F1 scores. For the Glass Science dataset, we report accuracy scores for each baseline. The best and the second best results are highlighted in **boldface** and underline, respectively.

Method	SOFC$_{\text{SF}}$		SOFC$_{\text{NER}}$		MatScholar		Glass Science	
	dev	test	dev	test	dev	test	dev	test
BERT (Devlin et al., 2019)	0.652	0.569	0.808	0.787	0.848	0.844	0.932	**0.938**
AdaLM (Yao et al., 2021)	0.637	0.566	0.792	0.793	0.837	0.841	0.935	0.937
AVocaDo (Hong et al., 2021)	0.629	0.579	0.787	0.777	0.844	0.841	0.928	0.935
Replace (Kajiura et al., 2023)	0.656	0.576	0.810	0.790	0.846	0.839	0.935	0.936
SEED (ours)	**0.661**	**0.594**	**0.811**	**0.807**	**0.859**	**0.853**	**0.944**	0.937

Table 2: Vocabulary statistics of *Mat2vec* (Tshitoyan et al., 2019) and two different PLMs (BERT (Devlin et al., 2019) and SciBERT (Beltagy et al., 2019)).

Models	# Words	# Overlap to *Mat2vec*
Mat2vec (E_{M})	529,686	-
BERT (E_{P})	30,522	17,261 (56.5%)
SciBERT (E_{P})	31,090	15,123 (48.6%)

on the four different tasks in materials science. These tasks include materials entity recognition, paragraph classification, and slot filling.

For the materials entity recognition tasks, we use two widely used datasets, MatScholar (Weston et al., 2019) and SOFC (Friedrich et al., 2020), in which the model is required to recognize entities including materials, descriptors, materials properties, and applications from materials science text. For the paragraph classification task, we use the glass paragraph dataset (Venugopal et al., 2021) which requires the model to determine whether a given paragraph is related to glass science. The slot-filling task is to extract slot fillers from particular sentences based on a pre-defined set of semantically meaningful entities, and we use the SOFC (Friedrich et al., 2020) dataset.

Backbones To verify the general applicability of the proposed method, we apply our method to two different backbone models which are SciBERT (Beltagy et al., 2019) and BERT (Devlin et al., 2019). SciBERT (Beltagy et al., 2019) is the encoder-based model trained on 1.14M scientific corpus, and BERT (Devlin et al., 2019) is the encoder-based model trained on general English corpus (Wikipedia and BookCorpus). The statistics of each embedding are presented in Table 2.

4.2 Main Results

Results on BERT Table 1 presents the overall performance results on the four materials benchmarks using the BERT (Devlin et al., 2019) backbone. As mentioned earlier, the existing vocabulary expansion baselines show limited performance improvements across various materials-domain tasks, highlighting the unique challenges in the field of materials science[5]. However, we find that the proposed method, SEED, significantly enhances performance in almost all settings. This improvement underscores the efficacy of knowledge transfer from the materials embeddings in the PLMs. One of the key factors in this superior performance also lies in embedding initialization, as existing methods focus primarily on tokenization and less on the initialization of the added tokens. Overall results confirm that SEED can effectively adapt the PLMs to materials science and effectively mitigates the limitations of the existing methods.

Results on SciBERT To verify the general applicability of SEED and confirm whether the PLMs pre-trained on the scientific corpus can achieve a performance improvement, we apply SEED to a different backbone that is pre-trained on the scientific corpus (i.e., SciBERT(Beltagy et al., 2019)). Table 3 shows the results on the four benchmark datasets. The results show a consistent trend to the results with BERT. While the performance improvement from the existing vocabulary expansion methods is limited, the adaptation performances are boosted when adapting PLMs with the proposed method. These results underscore the general applicability of SEED and show that the PLMs pre-trained on

[5] To demonstrate the effectiveness of the existing baselines in other domains, we adapted each baseline to the fields of biomedical and computer science. Please refer to the Appendix for a more detailed analysis.

Table 3: Evaluation results on four materials benchmarks based on the SciBERT (Beltagy et al., 2019). For SOFC and MatScholar, the reported performances are Macro-F1 scores. For the Glass Science dataset, we report accuracy scores for each baseline. The best and the second best results are highlighted in **boldface** and underline, respectively.

Method	SOFC_{SF}		SOFC_{NER}		MatScholar		Glass Science	
	dev	test	dev	test	dev	test	dev	test
SciBERT (Devlin et al., 2019)	**0.683**	**0.602**	0.824	0.810	0.875	0.856	0.937	0.938
AdaLM (Yao et al., 2021)	0.669	0.580	0.808	0.800	0.865	0.847	0.931	0.940
AVocaDo (Hong et al., 2021)	0.675	0.596	0.796	0.786	0.873	0.849	0.940	0.941
Replace (Kajiura et al., 2023)	0.682	0.597	0.818	0.806	0.869	0.838	0.937	0.937
SEED (ours)	0.673	0.586	**0.839**	**0.818**	**0.886**	**0.861**	**0.947**	**0.943**

Table 4: Ablation results of the training objectives for the bridge network on the two representative datasets. Here, we use the BERT (Devlin et al., 2019) backbone and evaluation results on the test set for each dataset.

Method	SOFC_{NER}	MatScholar
SEED (ours)	0.807	0.853
w/o Relation	0.801	0.844
w/o Reconstruction	0.798	0.840

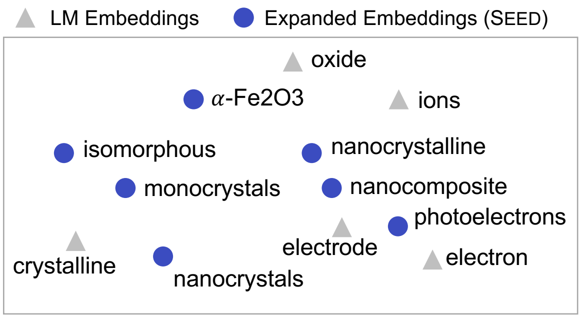

Figure 3: t-SNE visualization of the examples about the PLMs embeddings with the transferred ones.

scientific corpus achieve the benefit from the SEED.

4.3 Ablation Study

To confirm whether each component in SEED is indeed effective in adapting the pre-trained language models to materials science, we perform the ablation studies. Specifically, we evaluate the contributions of the training objectives in training the bridge networks, i.e., reconstruction loss $\mathcal{L}_{\text{recon}}$ and relation loss $\mathcal{L}_{\text{rel}}$. Table 4 presents the ablation results on the two representative datasets. We first observe that omitting each component from the proposed method consistently leads to performance degradation, demonstrating the effectiveness of each component. In particular, we observe that relation loss plays a significant role in effectively training the bridge network. These results empirically justify the contributions of each component in SEED.

4.4 Visualization of the Expanded Vocabulary

Lastly, we visualize the expanded vocabulary to confirm whether the added words from SEED are indeed semantically related to original embeddings in PLMs. Figure 3 shows the examples of the 2D projected embeddings by t-SNE (Van der Maaten and Hinton, 2008). Interestingly, in the vicinity of *electrode, electron* embeddings, semantically related words are closely located. For example, *nanocrystals* and *nanocrystalline*, which are the neighboring words of *electrode* and *crystalline*, play a crucial role in advancing the performance and durability of *electrode* materials used in various energy storage and conversion technologies. Moreover, the chemical formula α-Fe2O3, which has desirable electrochemical properties for *electrodes*, is also closely placed with *electrode* and *oxide*. This result demonstrates that SEED can expand the knowledge of PLMs by augmenting the original embeddings with semantically related words.

5 Conclusion

In this work, we have proposed **Se**mantic Knowl**ed**ge Transfer (SEED), a novel vocabulary expansion method aimed at adapting PLMs to materials science. Specifically, we have leveraged *Mat2vec* to expand the knowledge of the PLMs, which are lightweight embeddings trained on large-scale scientific papers. The knowledge in the materials embeddings is subsequently transferred to the PLMs through the learned bridge networks which serve as a mapping function between two different knowledge representations. We have performed extensive experiments to verify the efficacy of the proposed method across diverse benchmarks and various architectures. Comprehensive results have convincingly demonstrated that adapting the PLMs with SEED leads to substantial improvements in performance across diverse materials-related tasks compared to existing vocabulary expansion methods, highlighting the broad value of SEED in materials science.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No.RS-2024-00415812 and No.2021R1A2C3010430) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2024-00439328, Karma: Towards Knowledge Augmentation for Complex Reasoning (SW Starlab)).

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3613–3618.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Annemarie Friedrich, Heike Adel, Federico Tomazic, Johannes Hingerl, Renou Benteau, Anika Marusczyk, and Lukas Lange. 2020. The sofc-exp corpus and neural approaches to information extraction in the materials science domain. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1255–1268.

Tanishq Gupta, Mohd Zaki, N. M. Anoop Krishnan, and Mausam. 2022. MatSciBERT: A materials domain language model for text mining and information extraction. *npj Computational Materials*, 8(1):102.

Jimin Hong, Taehee Kim, Hyesu Lim, and Jaegul Choo. 2021. Avocado: Strategy for adapting vocabulary to downstream domain. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4692–4700.

Shu Huang and Jacqueline M. Cole. 2022. Batterybert: A pretrained language model for battery database enhancement. *Journal of chemical information and modeling*, 62(24):6365–6377.

Peter J Huber. 1992. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pages 492–518. Springer.

Teruno Kajiura, Shiho Takano, Tatsuya Hiraoka, and Kimio Kuramitsu. 2023. Vocabulary replacement in sentencepiece for domain adaptation. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 645–652.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Ankan Mullick, Akash Ghosh, G Sai Chaitanya, Samir Ghui, Tapas Nayak, Seung-Cheol Lee, Satadeep Bhattacharjee, and Pawan Goyal. 2024. Matscire: Leveraging pointer networks to automate entity and relation extraction for material science knowledge-base construction. *Computational Materials Science*, 233:112659.

Sheshera Mysore, Zachary Jensen, Edward Kim, Kevin Huang, Haw-Shiuan Chang, Emma Strubell, Jeffrey Flanigan, Andrew Mccallum, and Elsa Olivetti. 2019. The materials science procedural text corpus: Annotating materials synthesis procedures with shallow semantic structures. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 56–64.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. 2019. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3967–3976.

Yu Song, Santiago Miret, and Bang Liu. 2023. Matsci-nlp: Evaluating scientific language models on materials science language tasks using text-to-schema modeling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 3621–3639.

Wen Tai, H. T. Kung, Xin Dong, Marcus Z. Comiter, and Chang-Fu Kuo. 2020. exbert: Extending pre-trained models with domain-specific vocabulary under constrained training resources. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1433–1439.

Amalie Trewartha, Nicholas Walker, Haoyan Huo, Sanghoon Lee, Kevin Cruse, John Dagdelen, Alexander Dunn, Kristin A Persson, Gerbrand Ceder, and Anubhav Jain. 2022. Quantifying the advantage of domain-specific pre-training on named entity recognition tasks in materials science. *Patterns*, 3(4).

Vahe Tshitoyan, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A Persson, Gerbrand Ceder, and Anubhav Jain. 2019. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, 571(7763):95–98.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11).

Vineeth Venugopal, Sourav Sahoo, Mohd Zaki, Manish Agarwal, Nitya Nand Gosvami, and NM Anoop Krishnan. 2021. Looking through glass: Knowledge discovery from materials science literature using natural language processing. *Patterns*, 2(7).

Leigh Weston, Vahe Tshitoyan, John Dagdelen, Olga Kononova, Amalie Trewartha, Kristin A Persson, Gerbrand Ceder, and Anubhav Jain. 2019. Named entity recognition and normalization applied to large-scale information extraction from the materials science literature. *Journal of Chemical Information and Modeling*, 59(9):3692–3702.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Yunzhi Yao, Shaohan Huang, Wenhui Wang, Li Dong, and Furu Wei. 2021. Adapt-and-distill: Developing small, fast and effective pretrained language models for domains. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021*, pages 460–470.

Rong Zhang, Revanth Gangi Reddy, Md Arafat Sultan, Vittorio Castelli, Anthony Ferritto, Radu Florian, Efsun Sarioglu Kayi, Salim Roukos, Avi Sil, and Todd Ward. 2020. Multi-stage pre-training for low-resource domain adaptation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 5461–5468.

Appendix

A Baselines on the non-materials domains

We replicate the state-of-the-art vocabulary expansion method (Hong et al., 2021) to confirm whether this method works well on domains that are widely used in previous adaptation papers. Table 5 shows that the expansion method significantly improves the performance of the PLMs on almost all datasets and domains. These results confirm the distinct challenges of adaptation methods for materials science, where each baseline shows degraded performance even after adaptation.

Table 5: Evaluation results on three different datasets with different domains. Macro-F1 score for ACL-ARC (Computer Science) and Hyperpartisan News (News), micro-F1 score for ChemProt (Biomedical).

Models	ChemProt	ACL-ARC	Hyperpartisan News
BERT	0.797	0.568	0.834
AVocaDo	0.812 (+0.015)	0.688 (+0.120)	0.889 (+0.055)

B Hyper-parameter setups of SEED

We follow the fine-tuning strategy of previous works (Hong et al., 2021). For the SEED method, we optimize the bridge networks using a learning rate of 1e-3 with the Adam optimizer and a batch size of 32. To select the unique sets $\mathcal{U}$ from each downstream datasets, we search for the best Top-k values ranging from 10 to 100 (with the step size of 10). We also apply several heuristics for the selection. To use new embeddings only for complex terms, we set a minimum number of split tokens. In other words, we include words that are originally split into more than four tokens. We conduct all experiments on two NVIDIA RTX A6000 GPUs.

C Implementation details of *Mat2vec*

To obtain the materials embeddings (*Mat2vec*), we trained skip-gram word embeddings on scientific journals. We followed the overall procedures of the original work (Tshitoyan et al., 2019), but increased the number of journals to 4.5 million (the paper utilized roughly 3 million scientific journals) to cover recent publications and expand the scope of materials. The overall training process takes 7 hours in the setup of Intel Xeon Gold 6230R CPUs.