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

EMNLP 2024

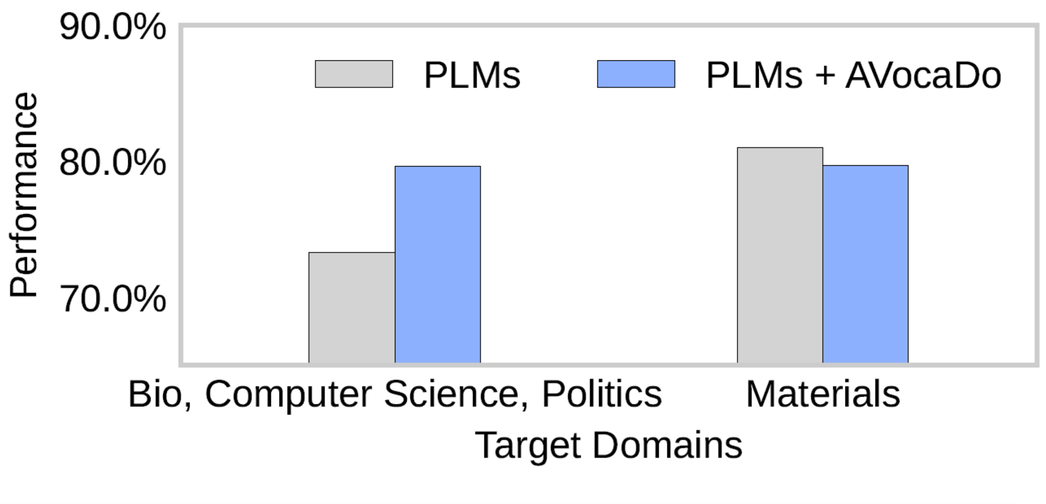
Yeachan Kim, Jun-Hyung Park, SungHo Kim, Juhyeong Park, Sangyun Kim, SangKeun Lee
Korea University

KOREA UNIVERSITY

Summary & Motivation

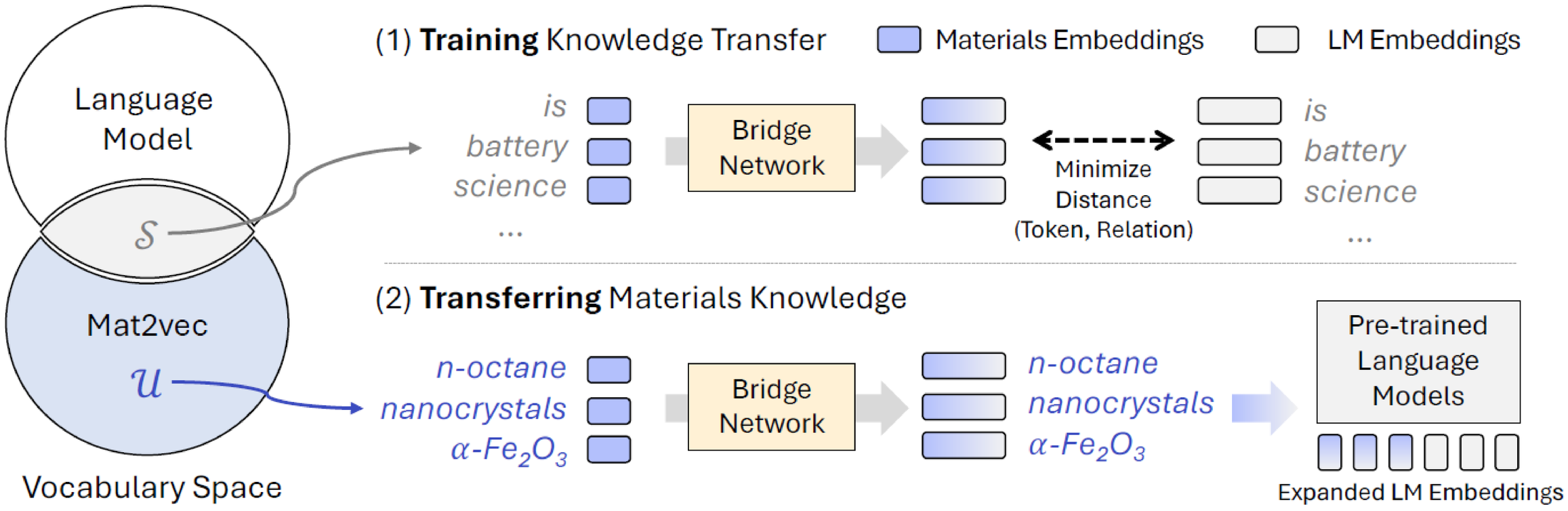
*"**Frequency-based** vocabulary expansion for LMs failed on the domain of materials science"*

*"We introduce **SEED**, a vocabulary expansion method by transferring embedding knowledge (w/o **frequency** information)"*



Semantic Knowledge Transfer from Mat2vec

"Learning bridge networks between Mat2vec and LMs!"



Evaluation Results and Analysis

- **Evaluation Results on materials-related tasks**
  - Applying SEED to the off-the-shelf pre-trained LMs

Method	$SOFC_{SF}$ dev	$SOFC_{SF}$ test	$SOFC_{NER}$ dev	$SOFC_{NER}$ test	MatScholar dev	MatScholar test	Glass Science dev	Glass Science 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

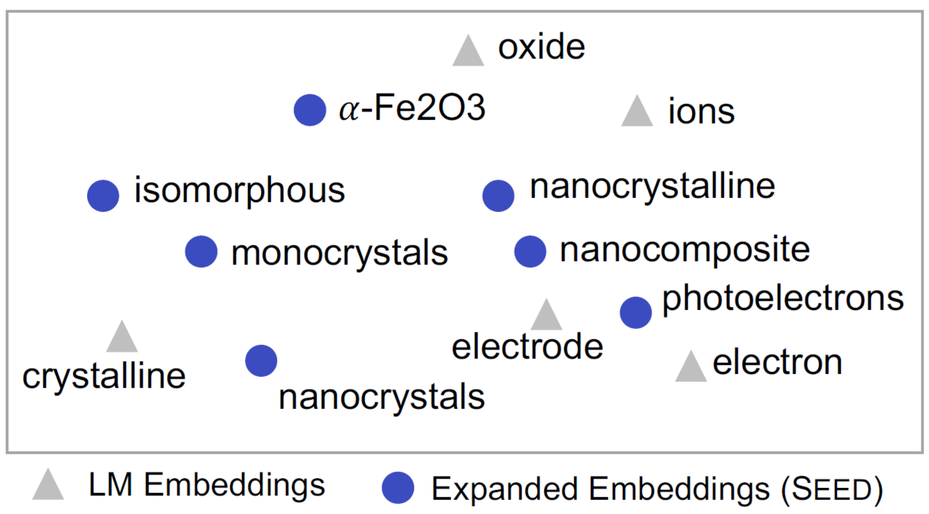
▲ Evaluation on BERT

Method	$SOFC_{SF}$ dev	$SOFC_{SF}$ test	$SOFC_{NER}$ dev	$SOFC_{NER}$ test	MatScholar dev	MatScholar test	Glass Science dev	Glass Science 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**

▲ Evaluation on SciBERT

- ✓ Baselines show poor performance on materials tasks
- ✓ Only SEED (ours) improves the performance on almost all settings

- Expanded embeddings by **SEED**
  - t-SNE visualization on embeddings



Conclusion & Future work

- ✓ We propose **SEED**, a novel vocabulary expansion method by transferring the knowledge of materials embeddings into LMs
- ✓ We verify that adopting SEED leads to the improved performance on materials tasks
- ✓ We plan to develop efficient knowledge expansion methods for **decoder models**

[1] Hong et al., AVocaDo: Strategy for Adapting Vocabulary to Downstream Domain, EMNLP 2021
[2] Tshitoyan et al., Unsupervised Word Embeddings Capture Latent Knowledge from Materials Science Literature, Nature (2019)