

Field	Task	Dataset	SOTA	BERT-Base		SCI BERT	
				Frozen	Finetune	Frozen	Finetune
Bio	NER	BC5CDR (Li et al., 2016)	88.85 ⁷	85.08	86.72	88.73	90.01
		JNLPBA (Collier and Kim, 2004)	78.58	74.05	76.09	75.77	77.28
		NCBI-disease (Dogan et al., 2014)	89.36	84.06	86.88	86.39	88.57
	PICO	EBM-NLP (Nye et al., 2018)	66.30	61.44	71.53	68.30	72.28
	DEP	GENIA (Kim et al., 2003) - LAS	91.92	90.22	90.33	90.36	90.43
		GENIA (Kim et al., 2003) - UAS	92.84	91.84	91.89	92.00	91.99
	REL	ChemProt (Kringelum et al., 2016)	76.68	68.21	79.14	75.03	83.64
CS	NER	SciERC (Luan et al., 2018)	64.20	63.58	65.24	65.77	67.57
	REL	SciERC (Luan et al., 2018)	n/a	72.74	78.71	75.25	79.97
	CLS	ACL-ARC (Jurgens et al., 2018)	67.9	62.04	63.91	60.74	70.98
Multi	CLS	Paper Field	n/a	63.64	65.37	64.38	65.71
		SciCite (Cohan et al., 2019)	84.0	84.31	84.85	85.42	85.49
Average				73.58	77.16	76.01	79.27

Table 1: Test performances of all BERT variants on all tasks and datasets. **Bold** indicates the SOTA result (multiple results bolded if difference within 95% bootstrap confidence interval). Keeping with past work, we report macro F1 scores for NER (span-level), macro F1 scores for REL and CLS (sentence-level), and macro F1 for PICO (token-level), and micro F1 for ChemProt specifically. For DEP, we report labeled (LAS) and unlabeled (UAS) attachment scores (excluding punctuation) for the same model with hyperparameters tuned for LAS. All results are the average of multiple runs with different random seeds.

Task	Dataset	BIOBERT	SCI BERT
NER	BC5CDR	88.85	90.01
	JNLPBA	77.59	77.28
	NCBI-disease	89.36	88.57
REL	ChemProt	76.68	83.64

Table 2: Comparing SCI BERT with the reported BIOBERT results on biomedical datasets.

BC5CDR and ChemProt, and performs similarly on JNLPBA despite being trained on a substantially smaller biomedical corpus.

4.2 Computer Science Domain

We observe that SCI BERT outperforms BERT-Base on computer science tasks (+3.55 F1 with finetuning and +1.13 F1 without). In addition, SCI BERT achieves new SOTA results on ACL-ARC (Cohan et al., 2019), and the NER part of SciERC (Luan et al., 2018). For relations in SciERC, our results are not comparable with those in Luan et al. (2018) because we are performing relation classification given gold entities, while they perform joint entity and relation extraction.

4.3 Multiple Domains

We observe that SCI BERT outperforms BERT-Base on the multidomain tasks (+0.49 F1 with finetuning and +0.93 F1 without). In addition, SCI BERT outperforms the SOTA on Sci-

Cite (Cohan et al., 2019). No prior published SOTA results exist for the Paper Field dataset.

5 Discussion

5.1 Effect of Finetuning

We observe improved results via BERT finetuning rather than task-specific architectures atop frozen embeddings (+3.25 F1 with SCI BERT and +3.58 with BERT-Base, on average). For each scientific domain, we observe the largest effects of finetuning on the computer science (+5.59 F1 with SCI BERT and +3.17 F1 with BERT-Base) and biomedical tasks (+2.94 F1 with SCI BERT and +4.61 F1 with BERT-Base), and the smallest effect on multidomain tasks (+0.7 F1 with SCI BERT and +1.14 F1 with BERT-Base). On every dataset except BC5CDR and SciCite, BERT-Base with finetuning outperforms (or performs similarly to) a model using frozen SCI BERT embeddings.

5.2 Effect of SCIVOCAB

We assess the importance of an in-domain scientific vocabulary by repeating the finetuning experiments for SCI BERT with BASEVOCAB. We find the optimal hyperparameters for SCI BERT-BASEVOCAB often coincide with those of SCI BERT-SCIVOCAB.

Averaged across datasets, we observe +0.60 F1 when using SCIVOCAB. For each scientific do-