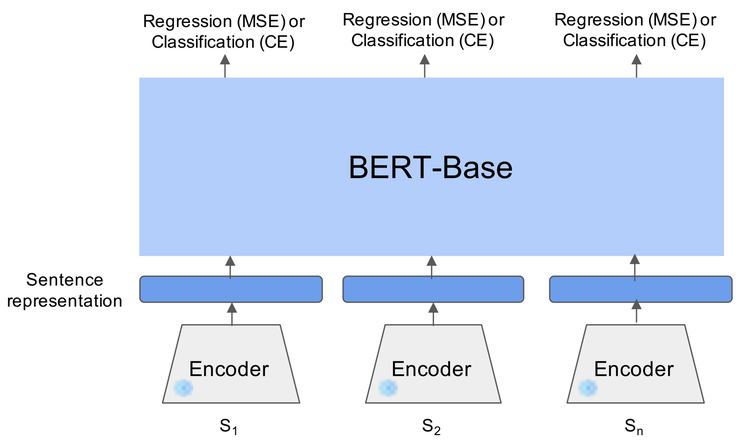
Less Mature is More Adaptable for Sentence-level Language Modeling

Abhilasha Sancheti[1], David Dale[2], Artyom Kozhevnikov[2], and Maha Elbayad[2]
[1] *University of Maryland, College Park,* [2] *FAIR at Meta*
sancheti@umd.edu, {daviddale, artyomko, elbayadm}@meta.com

Research Questions



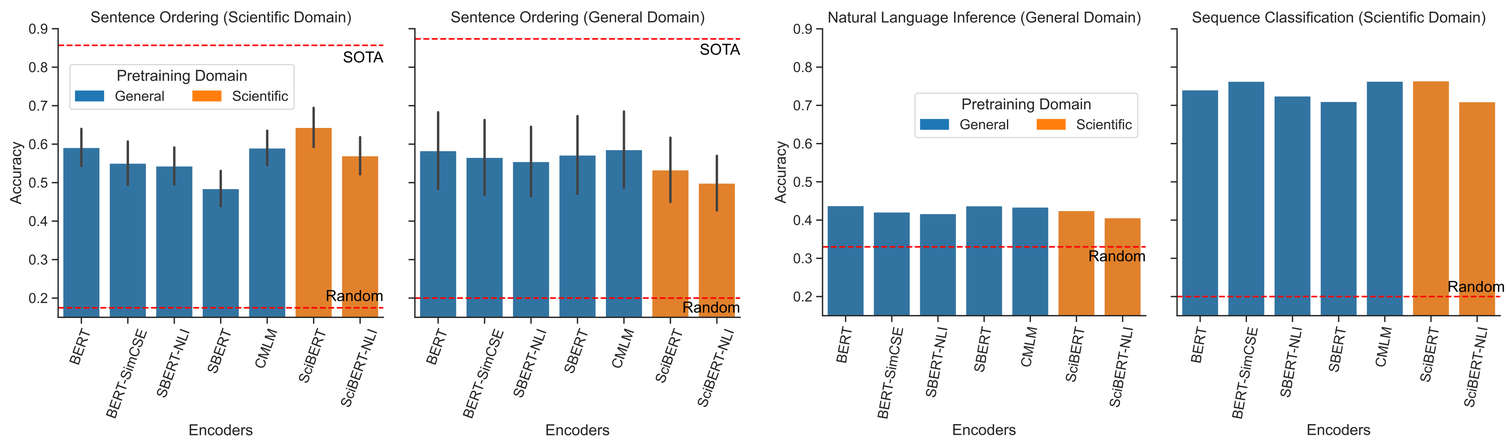
A sentence-level model takes in sentence representations (from a sentence encoder) as input as opposed to tokens that are used in standard token-level models.

- (RQ1): How does a representation learning approach (*e.g.*, fine-tuning, contrastive learning, conditional language modeling) impact the performance of a sentence-level modeling task?
- (RQ2): What properties must be encoded in sentence representations to enable sentence-level modeling tasks?
- (RQ3): What are the advantages of sentence-level models over standard token-level models?

Experimental Setup

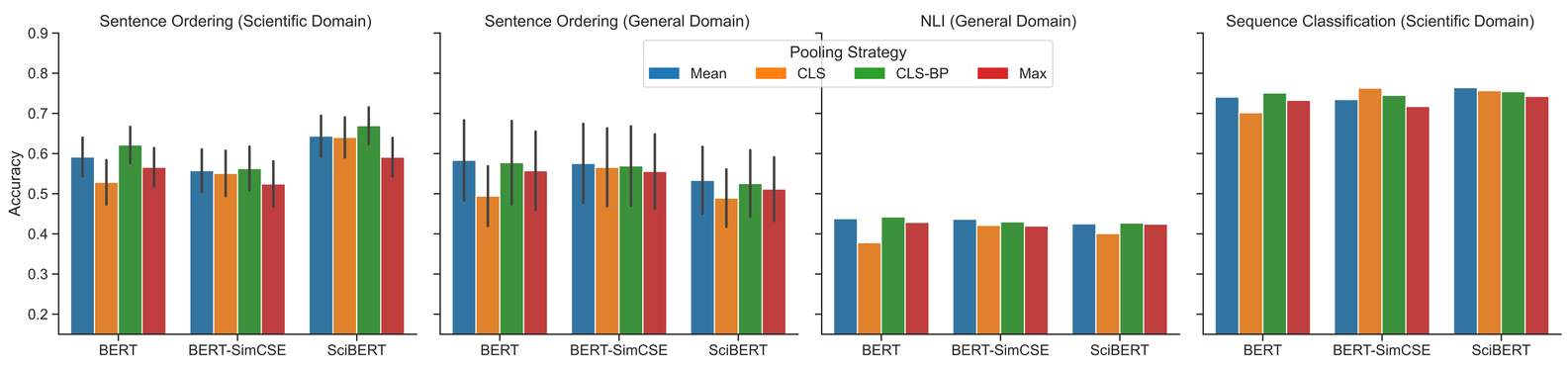
- Experiment with several sentence representations, obtained from a variety of sentence encoders, to build sentence-level models for addressing (RQ1)
- Use existing sentence representation evaluation benchmarks (Conneau et al., 2018; Conneau and Kiela, 2018; Muennighoff et al., 2023; Chen et al., 2019) to assess and correlate the surface-level, syntactic, semantic, and discourse-level properties encoded in embeddings with their downstream task performance to answer (RQ2)
- Compare the downstream task performance of a token-level model with that of sentence-level model to investigate (RQ3)
- Three multi-sentence input tasks requiring coarse-to-finegrained reasoning across two domains: Sentence Ordering, Sequential Sentence Classification, and Natural Language Inference

Finding 1: Less supervised training signals for better generalization



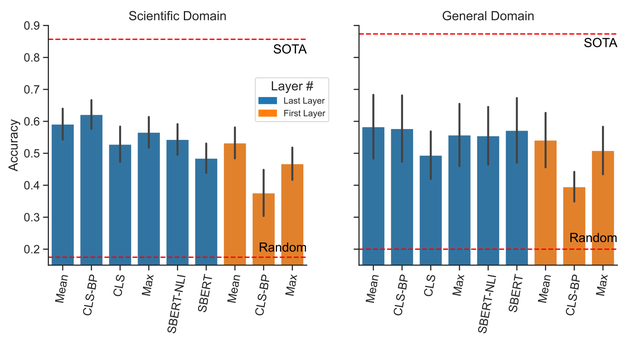
Sentence-level model accuracy using representations from encoders with varying training regimes and pretraining domains.

Finding 2: CLS token representations right before the pooling layer are surprisingly better



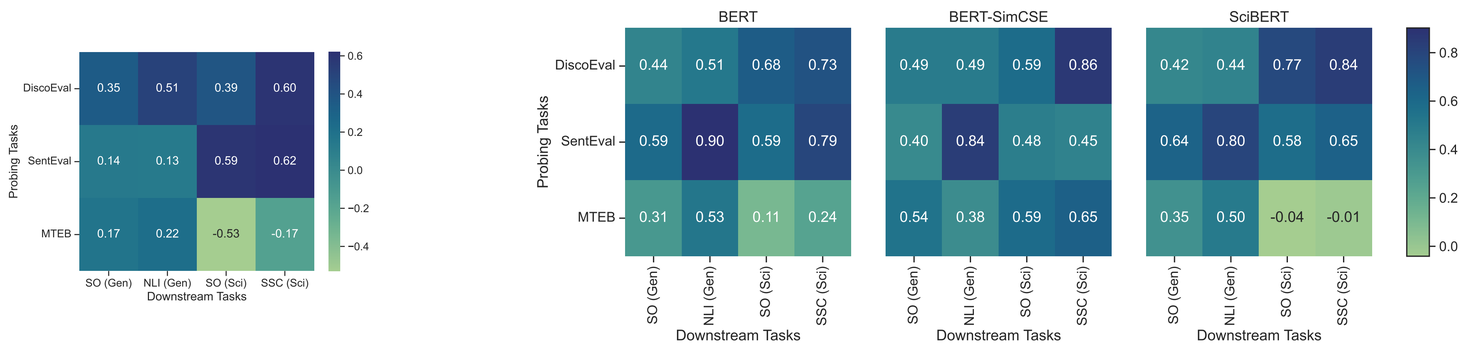
Accuracy for sentence-level models trained using representations from BERT, BERT-SimCSE, and SciBERT obtained via different pooling strategies from the final transformer layer.

Finding 3: First-layer representations $>_{effective}$ specifically trained encoders



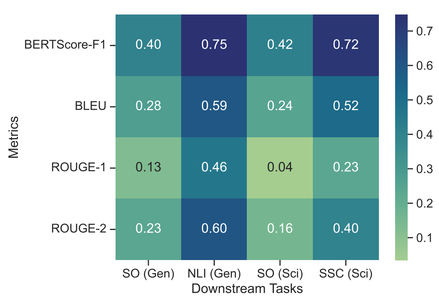
Comparing first/last layer pooling for BERT against SBERT and SBERT-NLI. Variance across sentence ordering datasets is shown.

Analysis 1: Syntactic and discourse properties drive downstream performance, not MTEB



Correlations between probing and downstream task performance for encoders with different training regimes and domains (left) and pooling strategies (right). Gen: General, Sci: Scientific

Analysis 2: Decodability does not always equal downstream success



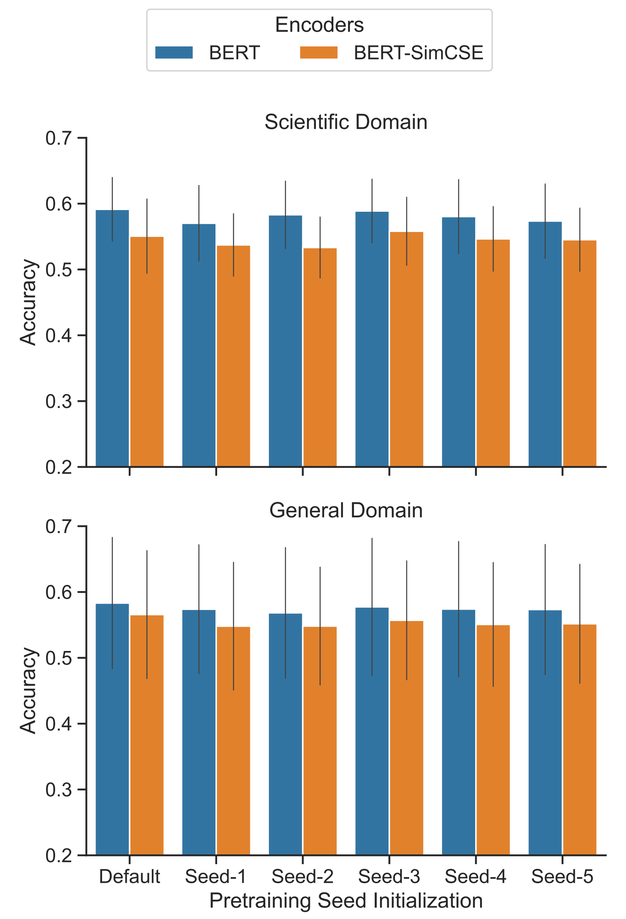
Correlations between decodability (measured via BLEU, ROUGE, and BERTScore) and downstream task performance for various encoders. Gen: General, Sci: Scientific

Analysis 3: Sentence-level models $>_{efficient}$ token-level models

Task	Dataset	Accuracy	
		Token	Sentence
SO	NIPS	36.31*	**54.29**
	AAN	48.55*	**63.98**
	SIND	48.30	**48.33**
	RocStories	61.15	**68.30**
SSC	CSAbstruct	73.76	**74.06**
NLI	ANLI	**47.38**	43.78

Accuracy (%) for different tasks from a token- level and corresponding sentence-level model. ∗: model was trained with 8 permutations per training example

Analysis 4: Our findings are robust across different BERT initializations



Sentence-level model accuracy (BERT and BERT-SimCSE) across different pretraining initializations. "Default" denotes Hugging Face models. Variance is across sentence ordering datasets.