Domain Specific Augmentations as Low Cost Teachers for Large Students

A Paper on the Shared Task DEAL

Po-Wei Huang
National University of Singapore
WIESP@AACL-IJCNLP 2022


ClinicalBERT


BioBERT


SciBERT


FinBERT
+80.01%
+0.01%
-0.09%


ContractBERT
CONTRACT

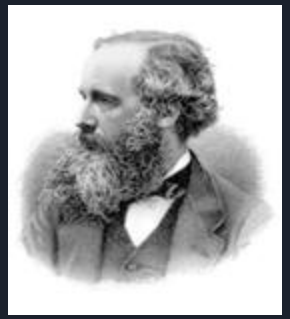

astroBERT

Are we able to achieve similar or better results by finetuning general models larger in size if we can transfer knowledge from pretrained domain-specific models?

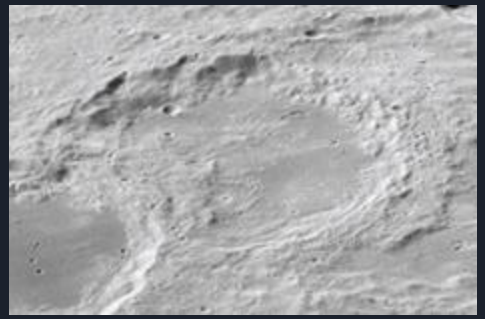
Task Description

Detecting Entities in the Astrophysics Literature

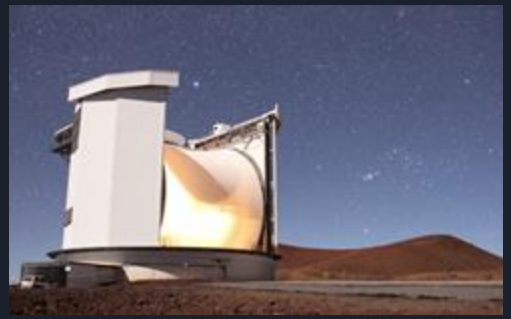
MAXWELL



Physicist



Crater on the Moon



Telescope in Hawaii

$$\nabla \cdot \mathbf{E} = \frac{\rho}{\varepsilon_0}$$

$$\nabla \cdot \mathbf{B} = 0$$

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}$$

$$\nabla \times \mathbf{B} = \mu_0 \mathbf{j} + \frac{1}{c^2}\frac{\partial \mathbf{E}}{\partial t}$$

Set of Equations

Dataset and Evaluation

Labeled:

- Training Dataset: 1753 samples
- Development Dataset: 20 samples

Unlabeled:

- Validation Dataset: 1366 samples
- Testing Dataset: 2505 samples

Evaluation:

- Token-Level: Matthew's Correlation Coefficient
- Entity-Level: seqeval Macro F1

Maxwell's demon was a thought experiment proposed by physicist James Clerk Maxwell.

B-Beginning I-Inside O-Outside

Methodology

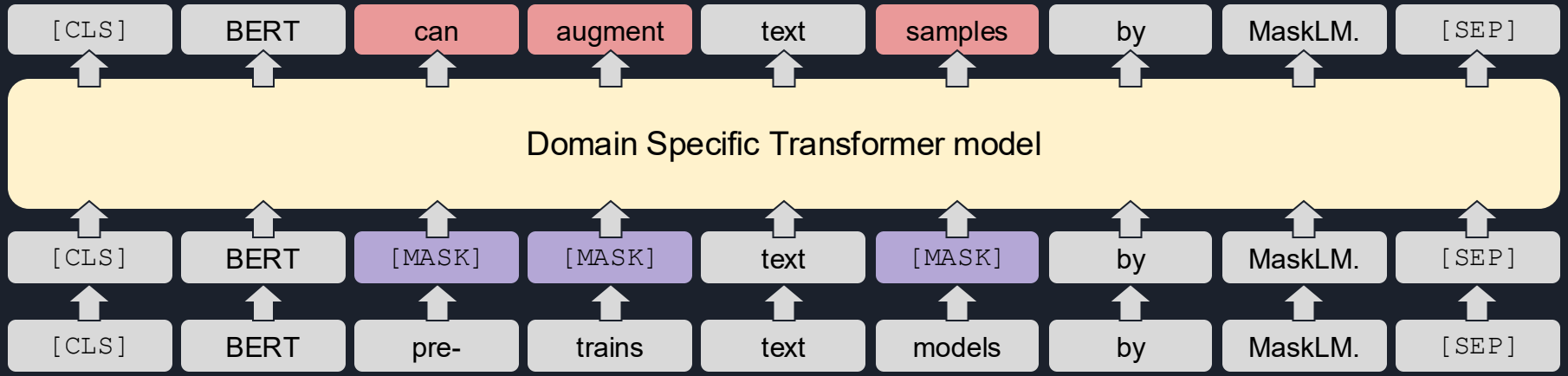
Preprocessing

The input text for the DEAL dataset was long and contained multiple sentences. We tokenize the sentences using regex, filtering end-of-sentence punctuations and breaking the sentences, while ignoring a list of abbreviations such as fig., tab., et al. Capitalization is retained due to its importance in entity recognition.

1. The input text for the DEAL dataset was long and contained multiple sentences.
2. We tokenize the sentences using regex, filtering end-of-sentence punctuations and breaking the sentences, while ignoring a list of abbreviations such as fig., tab., et al.
3. Capitalization is retained due to its importance in entity recognition.

Methodology

Augmentation



Methodology
Augmentation

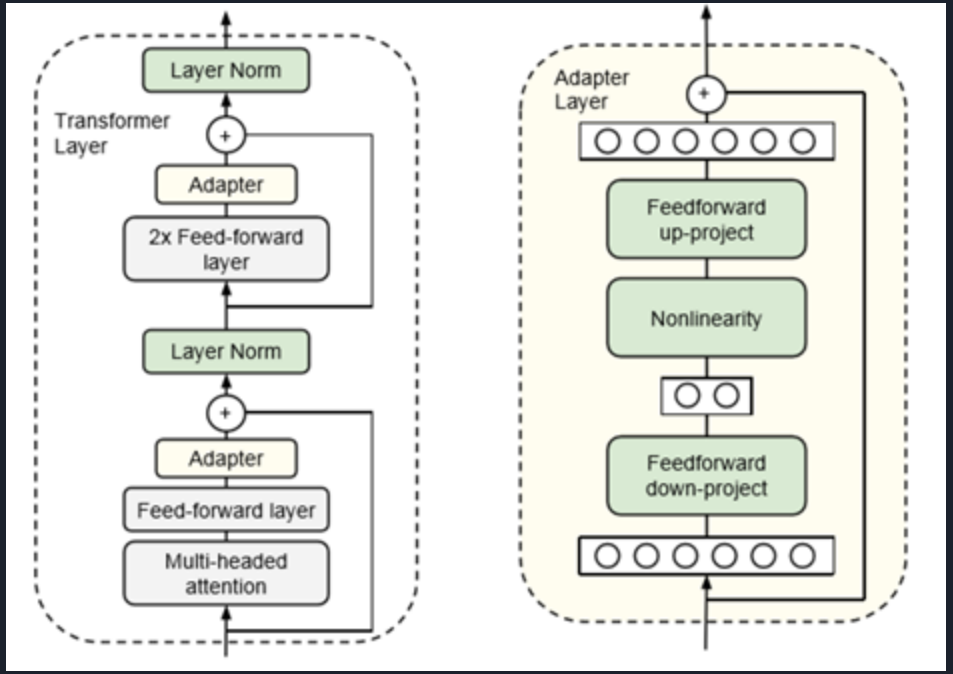
Original:	This research made use of NASA's Astrophysics Data System Bibliographic Services; the SIMBAD data base (Wenger et al. 2000) and VizieR catalogue access tool (Ochsenbein, Bauer Marcout 2000), both operated at CDS, Strasbourg, France; and the Jean-Marie Mariotti Center Aspro2 service 1 .
Augmented:	**The project** made use of NASA's Astrophysics Data System Bibliographic **database**; the SIMBAD data base (Wenger et al. 2000) and VizieR **data** access tool (**Sch,ouin, and** Marcout 2000), which operated at **CNR**, Strasbourg, France; and the Jean-Marie Mariotti Center **Asprox** service 1 .

* Bold text indicates augmented text.
† ulined text indicates named entities.

Table 1: Sample Augmentations by CosmicRoBERTa

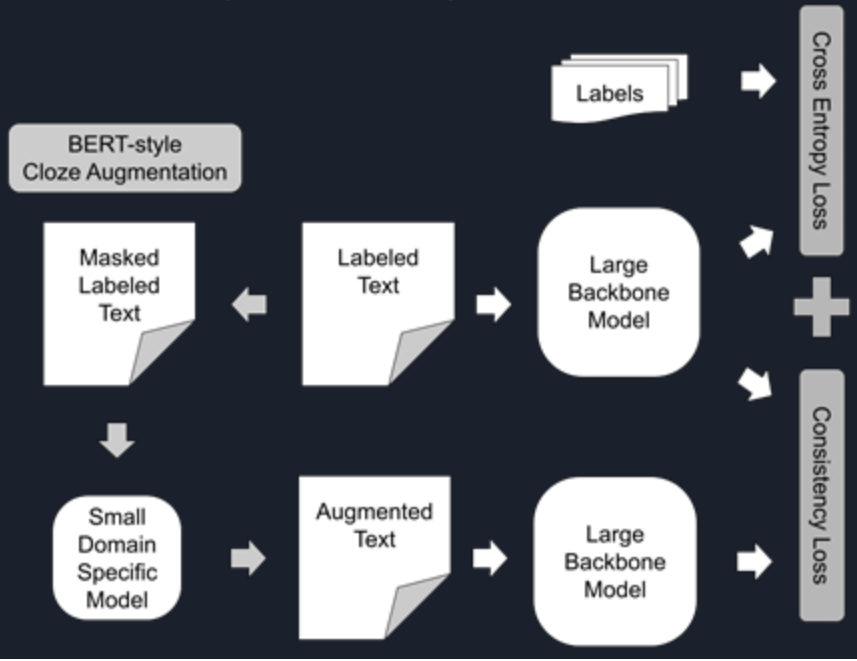
Methodology

Backbone Model Architecture



Methodology

Loss Function Engineering



$$\mathcal{L} = \frac{1}{B}\sum_{i=1}^{B} H(y_b, \hat{y}(x_b)) + D(\hat{y}(\mathcal{A}(x_b))||\hat{y}(x_b))$$

Results

	F1(entity)	MCC(word)
Random	0.0166	0.1089
BERT (Devlin et al., 2019)	0.4738	0.7405
SciBERT (Beltagy et al., 2019)	0.5595	0.8016
astroBERT (Grezes et al., 2021)	0.5781	0.8104
(*Ours*) DeBERTaV3$_{\text{adapter}}$ (He et al., 2021a,b; Houlsby et al., 2019)		
+ SciBERT (Beltagy et al., 2019)	0.7751	0.8898
+ CosmicRoBERTa (Berquand et al., 2021)	*0.7799*	*0.8928*

Table 2: Evaluation Results on Testing Dataset

	F1(entity)	MCC(word)	Accuracy(entity)
astroBERT	0.5781	0.8104	0.9389
DeBERTaV3$_{\text{adapter}}$ (He et al., 2021a,b; Houlsby et al., 2019)	0.7896	0.8987	0.9667
+ SciBERT$_{\text{cased}}$ (Beltagy et al., 2019)	**0.7988**	**0.9063**	**0.9692**
+ RoBERTa (Liu et al., 2019)	0.7970	0.9057	0.9690
+ CosmicRoBERTa (Berquand et al., 2021)	0.7972	0.9050	0.9687
+ SpaceSciBERT$_{\text{uncased}}$ (Berquand et al., 2021)	0.7859	0.9030	0.9680

Table 3: Augmentation Model Comparison on Validation Dataset

Connect with the first author!

huangpowei@comp.nus.edu.sg