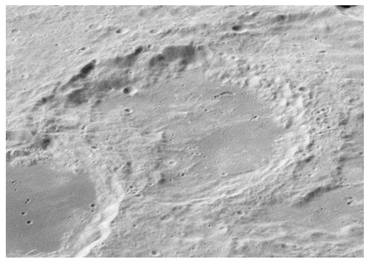
Domain Specific Augmentations as Low Cost Teachers for Large Students

Po-Wei Huang

NUS National University of Singapore | School of Computing

Task Description

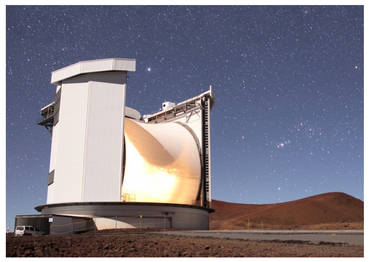
James Clerk Maxwell



Maxwell Crater (Moon)

$$\nabla \cdot \mathbf{E} = \frac{\rho}{\varepsilon_0}$$
$$\nabla \cdot \mathbf{B} = 0$$
$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}$$
$$\nabla \times \mathbf{B} = \mu_0 \mathbf{j} + \frac{1}{c^2}\frac{\partial \mathbf{E}}{\partial t}$$

Maxwell's Equations



James Clerk Maxwell Telescope

- Task: Recognize and extract 31 types of pre-defined categories of entities in astrophysics text.
- Current scientific document processing place a lot of emphasis on pretraining models on domain specific text, but is often restricted in model size and accuracy due to technical limitations
- *Aim: Discuss whether we can achieve similar or better results with finetuning general models larger in size whilst transferring knowledge from such pretrained scientific models to increase robustness.*

Dataset and Evaluation

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

B-Beginning I-Inside O-Outside

Labeled:

- Training: 1753 samples Development: 20 samples

Unlabeled:

- Validation : 1366 samples Testing : 2505 samples

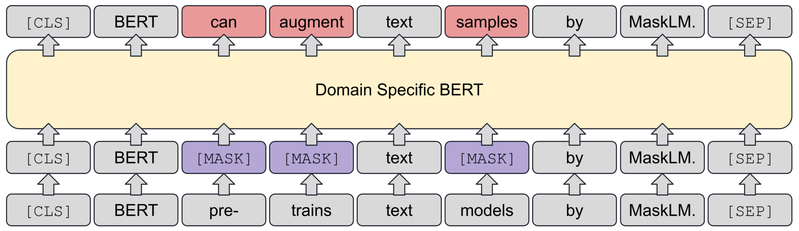
Evaluation:

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

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.

Augmentation



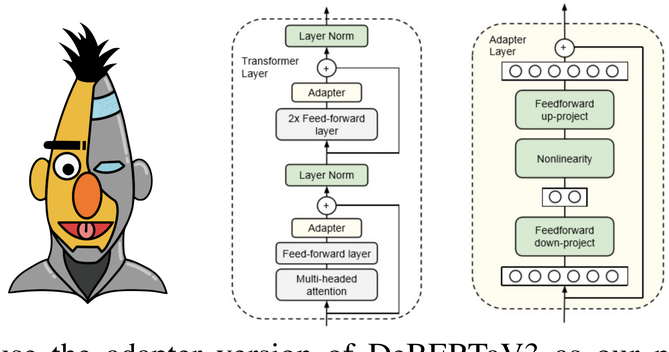
We augment the text with a domain specific transformer model via MaskLM to utilize data augmentation as a low-cost teacher model in order to transfer domain specific knowledge to our main model. For the DEAL task, i.e., astrophysical text, we experiment with SciBERT, and SpaceTransformers.

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 .

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 .

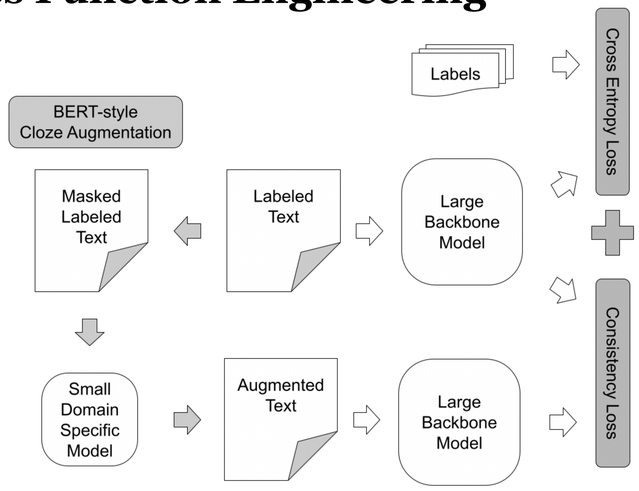
Sample Augmentation by CosmicRoBERTa

Backbone Model



We use the adapter version of DeBERTaV3 as our main backbone model to decrease computational costs, while obtaining similar results to finetuning the full model itself.

Loss Function Engineering



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

We employ the usage of the augmented text by adding an extra consistency loss term that computes the divergence between the predictions of the original text and the augmented text.

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$_{adapter}$ (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$_{adapter}$ (He et al., 2021a,b; Houlsby et al., 2019)	0.7896	0.8987	0.9667
+ SciBERT$_{cased}$ (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$_{uncased}$ (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

Scan to read the full paper!