Simple Recipes for Assessing Translation Quality

Darek Kłeczek

KLEJ

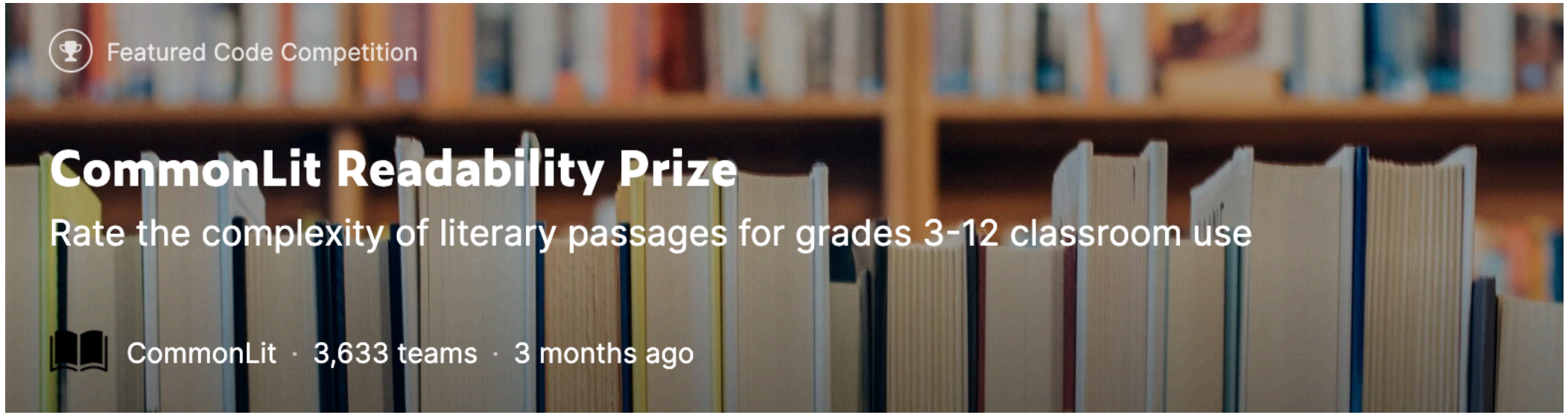
KLEJ

Paper Code Tasks Leaderboard

Leaderboard

Rank	Model	Fine-tuning	Average	NKJP-NER	CDSC-E	CDSC-R	CBD	PolEmo2.0-IN	PolEmo2.0-OUT	DYK	PSC	AR
1	HerBERT (large)	None	88.4	96.4	94.1	94.9	72.0	92.2	81.8	75.8	98.9	89.1
2	XLM-RoBERTa (large) + NKJP	Polish RoBERTa scripts	87.8	94.2	94.2	94.5	72.4	93.1	77.9	77.5	98.6	88.2
3	Polish RoBERTa (large)	Polish RoBERTa scripts	87.8	94.5	93.3	94.9	71.1	92.8	82.4	73.4	98.8	88.8

Kaggle CommonLit Readability Prize



Backbone

Model backbone	RMSE
Herbert base cased	0.323824
Herbert large cased	0.303575
Polish Roberta large	0.320343
Polish Roberta base	0.337123
Polbert base cased	0.325650

Average RMSE from 4 BLIND training runs with different hyperparameter settings. Trained on 80% and evaluated on 20% of development set.

Technique and tricks

- Hugging Face transformers and datasets
- Regression head on top of transformer backbone with MSE loss function
- Set dropout across the model to zero

Training – non blind

- Backbone: HerBERT Large Cased
- Input: automated and reference translations
- Training
  - 6 epochs
  - batch size 8
  - learning rate 2e-5
  - constant schedule
  - sequence length 192
  - weight decay 0.01
- 5-fold CV, early stopping based on validation loss
- 5-fold models average: 0.6137 test-B Pearson score

Training – blind

- Backbone: HerBERT Large Cased
- Input: Automated translation
- Training
  - 3 epochs
  - batch size 8
  - learning rate 1e-5
  - constant schedule
  - sequence length 128
  - weight decay 0.001
- 5-fold CV, early stopping based on validation loss
- Additional 5-fold model trained on back-translation data w/pseudolabels
- 2 * 5-fold models average: 0.4840 test-B Pearson score

Thank You!