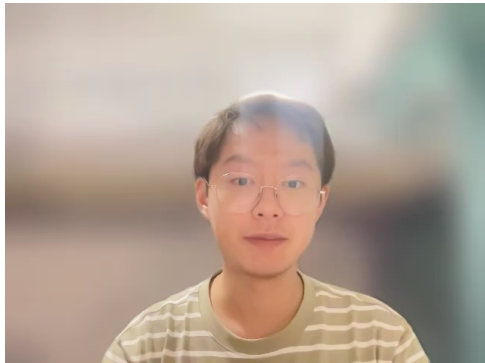
香港科技大學
THE HONG KONG UNIVERSITY OF
SCIENCE AND TECHNOLOGY

Doolittle: Benchmarks and Corpora for Academic Writing Formalization

EMNLP 2023

Shizhe Diao[*], Yongyu Lei[*], Liangming Pan, Tianqing Fang,
Wangchunshu Zhou, Sedrick Scott Keh, Min-Yen Kan, Tong Zhang

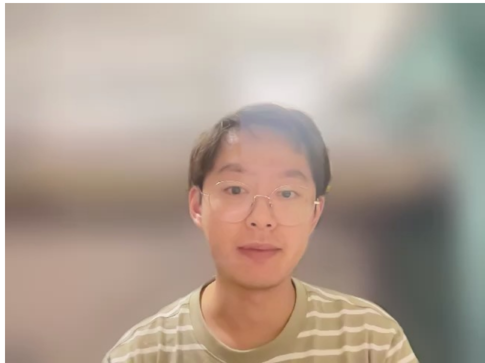


Introduction

- Grammatical Error Correction (GEC)
- Academic Writing Formalization (AWF)
  - grammar correction
  - word refinement
  - structural modification

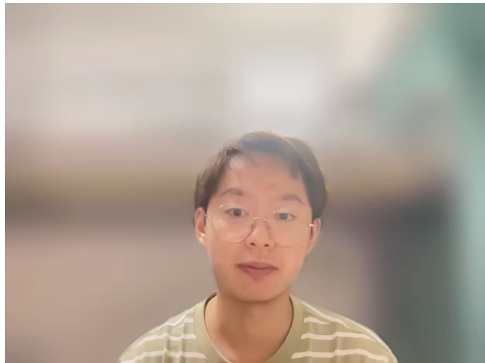
[S]: We propose more sophisticated hierarchical model to include geographical *informations*. [T]: We propose *a* more sophisticated hierarchical model to include geographical *information*.
[S]: This is because the teaching and learning on *science* domain relies *much* on the ability of reasoning and computation, which directly utilizes the *advantage of computer*. [T]: This is because the teaching and learning on *a scientific* domain relies *considerably* on the ability of reasoning and computation, which directly utilizes the *advantages of computers*.
[S]: METEOR is another n-gram overlap measure initially designed for evaluating machine translation systems. ROUGE-L is a commonly-adopted metric for text summarization. [T]: Both METEOR and ROUGE-L rely on n-gram overlaps for machine translation and text summarization evaluation, respectively.

Table 1: Informal-academic paragraphs with formal-academic rewrites, denoted S and T, respectively. The refined form is highlighted blue, the original in red.



Dataset Construction

- Data Source: Semantic Scholar Open Research Corpus (S2ORC)
- Academic Formality Annotation
  - Annotation task: score each paragraph from 1 (sounds informal-academic) to 5 (sounds formal-academic).
  - Publishing: Amazon Mechanical Turk (AMT)
  - Quality Control: time, variance, discrepancy
- Test Set Construction: human rewrites

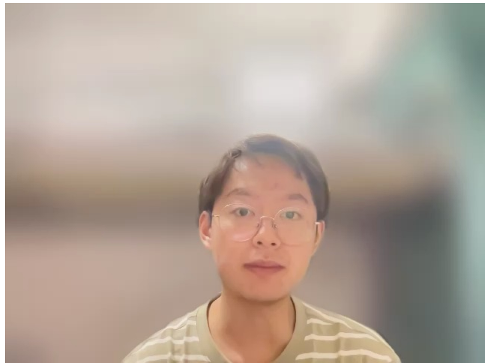


Data Analysis

- Transfer Accuracy
- Fluency
- Semantic Similarity
- BARTScore

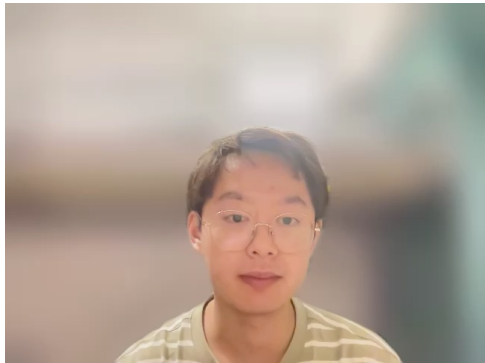
		P#	S#	V#	Avg. Words	Avg. Sent.	ACC-cola	ACC-aesw	PPL	SIM	ED	BARTS
Train	FA	55.6K	172.8K	84.3K	51.42	3.11	97.56	79.64	24.44	-	-	
	IFA	13.0K	41.3K	38.9K	52.17	3.17	95.81	68.51	32.56	-	-	
Dev	FA	465	1359	5.2K	47.33	2.92	98.49	78.27	31.19	98.75	11.03	-1.19
	IFA	465	1362	5.3K	47.79	2.92	95.69	72.04	33.07			
Test	FA	415	927	4.4K	42.52	2.23	98.31	77.83	33.18	98.09	10.87	-1.24
	IFA	415	910	4.5K	43.08	2.19	95.66	69.64	35.97			

Table 2: The statistics of the DOOLITTLE dataset, where P#, S#, V#, Avg. Words, and Avg. Sents. refer to the number of paragraphs, number of sentences, vocabulary size, average words per paragraph, and average sentences per paragraph, respectively. We also report the transfer accuracy (ACC), perplexity (PPL), Semantic Similarity (SIM), char-level edit distance (ED), and BARTScore (BARTS). FA and IFA denote formal-academic and informal-academic, respectively.



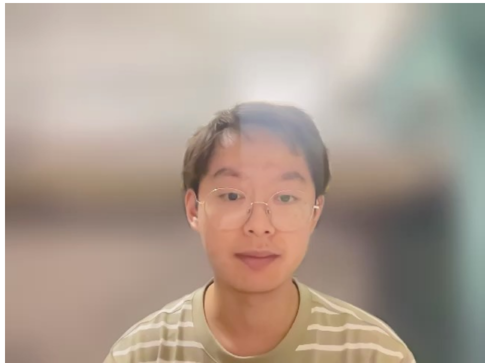
Proposed Methods

- **Metric-Oriented Reinforcement Learning (MORL)**
- Step 1: Train a policy model.
- Step 2: Select metrics that can accurately evaluate the quality. Build a reward model that can score a given policy model’s output with a scalar.
- Step 3: Optimize the policy against the reward model using reinforcement learning with the proximal policy optimization (PPO) algorithm.



Proposed Methods

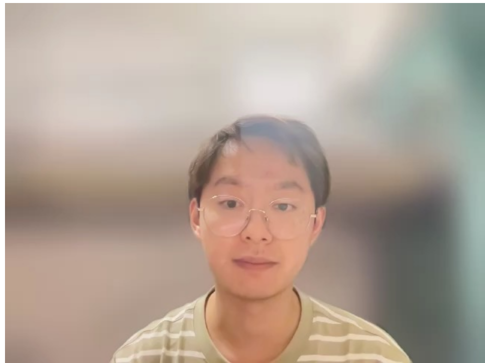
- Policy Models:
  - Galactica-1.3B
  - BART-Large
- Reward Model:
  - Transfer accuracy
  - PPL
  - SIM-input
  - BARTScore



Experimental Results

	Academic Formality					Fluency		Similarity			BARTS
Metric	ACC-cola	ACC-aesw	SARI	GLEU	GPT-4	PPL	GPT-4	SIM-input	SIM-gold	GPT-4	BARTS
Input	95.66	69.64	-	-	4.32	35.97	4.55	-	98.09	-	-
	Style Transfer Models										
ControlledGen	92.77	48.19	48.59	54.54	3.87	60.87	4.13	95.21	93.62	4.20	-1.64
DeepLatentSequence	84.81	50.36	37.46	50.40	3.55	68.45	4.15	90.45	88.97	3.78	-2.06
StyleTransformer	85.30	56.63	38.46	50.87	3.96	66.87	4.38	90.27	88.79	3.64	-2.19
DeleteAndRetrieve	66.50	66.02	7.98	1.07	2.91	34.11	3.36	21.12	20.27	2.22	-5.90
	GEC Models										
SequentialTransfer	94.70	70.36	49.17	71.30	4.32	41.19	4.45	96.80	95.55	4.26	-2.30
BART-GEC	95.90	70.12	69.10	74.72	4.40	35.83	4.66	99.01	97.24	4.94	-2.14
	Instruction Tuned Models										
ChatGPT	**99.20**	**82.56**	48.84	70.21	4.58	**28.84**	4.81	94.58	94.87	**4.73**	-1.62
MORL-BARTLarge	97.83	78.80	55.74	75.75	4.57	35.65	4.78	98.49	97.45	4.35	**-1.32**
MORL-Galactica1.3B	97.83	80.24	**63.79**	**78.37**	**4.60**	34.50	**4.86**	**98.72**	**98.30**	4.70	-1.34
Native Rewrite	98.31	77.83	-	-	4.59	33.18	4.89	98.09	-	4.95	-1.24

Table 3: Results of models on DOOLITTLE test paragraphs. Automatic evaluation and GPT-4 judgments of academic formality, fluency, and meaning preservation are reported. The highest scores of each metric among three instruction-tuned models are **bolded**. Some metrics are not applicable for Input and Native Rewrite as they are derived from comparison against these two sets, which are marked by '-'.



Take away messages

- Propose a new setting *Academic Writing Formalization (AWF)*.
- Contribute a new dataset *Doolittle*.
- Introduce a new method *metric-oriented reinforcement learning (MORL)*.
- MORL with 1.3B Galactica outperforms ChatGPT on AWF.

Thanks for your watching!