BERT is not The Count: Learning to Match Mathematical Statements with Proofs

Weixian Waylon Li   Yftah Ziser   Maximin Coavoux   Shay B. Cohen

School of Informatics, University of Edinburgh

Overview

- We introduce a task consisting in matching a proof to a given mathematical statement.
- We present a dataset for the task (the MATcH dataset) consisting of over 180k statement-proof pairs extracted from modern mathematical research articles.
- We propose a bilinear similarity model and two decoding methods to match statements to proofs effectively.
- Through a symbol replacement procedure, we analyze the "insights" that pre-trained language models have in such mathematical article analysis and show that while these models perform well on this task with the best performing MRR of 73.7, they follow a relatively shallow symbolic analysis and matching to achieve that performance.

Task Description

Given a collection of mathematical statements $\{s^{(i)}\}_{i \le N}$, and a separate equal-size collection of mathematical proofs $\{p^{(i)}\}_{i \le N}$, we are interested in the task of assigning a proof to each statement.

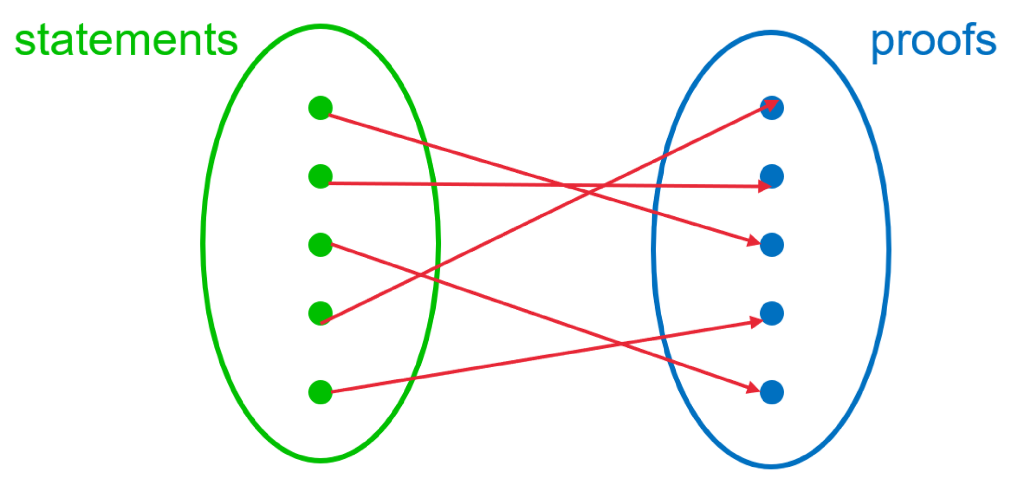

Figure 1. An illustration to the statement-proof matching task.

Dataset Construction

- **Source corpus:** the MREC corpus [1].
- **Statistics:** some statistics about the dataset we collected.

Number of articles in the MREC corpus	439,423
Extracted articles with statement-proof pairs	27,841
Total number of statement-proof pairs	184,094
Number of (primary) categories	(120) 135
Average number of categories per article	1.7

Table 1. Statistics about the dataset.

Symbol Replacement

Motivation:

- It is not realistic for researchers to match the proofs they authored.
- Each person has a unique writing style expressed by unique mathematical jargon and notations.

Symbol Replacement Levels:

$$a_n = a_{n-1} + a_{n-2}$$

Symbol conservation
$$a_n = a_{n-1} + a_{n-2}$$

Partial symbol replacement
$$x_n = x_{n-1} + x_{n-2}$$

Full symbol replacement
$$x_i = x_{i-1} + x_{i-2}$$

Symbol transposition
$$n_a = n_{a-1} + n_{a-2}$$

Figure 2. Four different levels of symbol replacement for the Fibonacci sequence.

Experimental Setup

Dataset: We shuffle the collection of statement-proof pairs before performing a $80\%/10\%/10\%$ train-development-test split.

Encoders: "No Pre-training Encoder" (NPT), ScratchBERT (pre-train BERT from scratch on MATcH) and MathBERT [2].

Decoders: Bilinear Similarity Model.

Bilinear Similarity Model

Trainable Bilinear Similarity Function: Given the encoded representations of a statement $\mathbf{s} = \mathrm{enc}(s)$ and a proof $\mathbf{p} = \mathrm{enc}(p)$:

$$\mathrm{score}(\mathbf{s}, \mathbf{p}) = \mathbf{s}^\top \cdot \mathbf{W} \cdot \mathbf{p} + b,$$

where $\mathbf{W}$ and b are parameters that are learned together with a self-attentive encoder parameters.

Local Decoding: A proof can be one of the candidates of multiple statements.

⇒ We found that 23% of the proofs were assigned to at least two different statements, whereas more than 40% of proofs were assigned to no statement.

Global Decoding: A proof can be assigned only to a single statement.

Local Training: for a single statement s and its gold proof p:

$$\mathcal{L}_{\mathrm{loc}}(s, p, P; \boldsymbol{\theta}) = -\log \mathbb{P}(p | s; \boldsymbol{\theta}),$$

where P is the set of proofs, and $\boldsymbol{\theta}$ are the parameters of the model.

⇒ Can we do even better by matching the hypothesis of global decoding?

Hybrid Local and Global Training: For a set B of n pairs corresponding to matrix M:

$$\mathcal{L}_{\mathrm{glob}}(B; \boldsymbol{\theta}) = \max(0, \Delta(\hat{A}, I) + \mathrm{score}(\hat{A}, M) - \mathrm{score}(I, M)),$$

where $\boldsymbol{\theta}$ is the set of all parameters $\hat{A}$ is the predicted assignment and I is the gold assignment, i.e. the identity matrix.

Main Findings

	Symbol Replacement Level							
	Conservation		Partial		Full		Transposition	
Encoder-Decoder	MRR	Acc	MRR	Acc	MRR	Acc	MRR	Acc
NPT-Local-Local	63.22	56.08	47.19	39.24	40.36	32.52	56.17	48.30
NPT-Local-Global	-	61.89	-	42.55	-	35.43	-	53.49
NPT-Global-Global	-	62.14	-	43.68	-	35.85	-	55.28
ScratchBERT-Local-Local	73.73	67.12	64.79	57.20	60.67	52.54	73.17	66.51
ScratchBERT-Local-Global	-	74.68	-	62.80	-	57.69	-	74.03
ScratchBERT-Global-Global	-	71.38	-	58.06	-	52.31	-	70.32
MathBERT-Local-Local	54.51	46.45	44.31	36.10	38.91	30.62	52.57	44.52
MathBERT-Local-Global	-	49.77	-	37.92	-	32.03	-	47.43
MathBERT-Global-Global	-	45.38	-	33.64	-	28.47	-	43.41

Table 2. The MRR and accuracy scores for different combinations of encoders, decoders, and symbol replacement levels. All the models are trained and tested on the same replacement level.

- Vocabulary is essential for learning from mathematical texts.
- The symbols' order, context, and function within the mathematical text do not play a significant role when the theorem and proof share the same symbols.
- Global decoding substantially improves accuracy.

	Target	Symbol Replacement							
	Source	Conservation		Partial		Full		Transposition	
		MRR	Acc	MRR	Acc	MRR	Acc	MRR	Acc
Mixed	Conservation	73.73	67.12	43.87	36.36	29.74	25.36	69.56	62.23
	Partial	74.21	67.96	64.79	57.20	53.77	45.40	72.13	65.42
	Full	65.26	57.63	63.01	55.13	60.67	52.54	64.59	56.92
	Transposition	73.78	67.40	43.67	36.02	29.76	25.47	73.17	66.51

Table 3. Cross-replacement levels performance for the ScratchBERT-Local-Local model.

- The model developed a strong dependency on exact symbol name matching.
- The model trained on the Partial symbol replacement level demonstrated significant resilience when tested with other symbol replacement levels.

Lemma 3.2. Let M be a module and H a local submodule of M. Then H is a supplement of each proper submodule $K \le M$ with $H + K = M$.
Proof. Since K is a proper submodule of M and $K + H = M$, we have $K \cap H$ is a proper submodule of H. Therefore $K \cap H \ll H$, since H is local. That is, H is a supplement of K in M.

(a) Example - Symbol conservation

Lemma 3.2. Let M be a module and H a local submodule of M. Then H is a supplement of each proper submodule $K \le M$ with $H + K = M$.
Proof. Since K is a proper submodule of M and $K + H = M$, we have $K \cap H$ is a proper submodule of H. Therefore $K \cap H \ll H$, since H is local. That is, H is a supplement of K in M.

(b) Example - Full symbol replacement

Figure 3. LIME visualizations for the model that was trained in the symbol conservation setup and full symbol replacement setup. "match" - orange, "mismatch" - blue.

References

[1] Martin Liška, Petr Sojka, Michal Růžička, and Petr Mravec. Web Interface and Collection for Mathematical Retrieval: WebMIaS and MREC. In Petr Sojka and Thierry Bouche, editors, *Towards a Digital Mathematics Library.*, pages 77–84, Bertinoro, Italy, 2011. Masaryk University.

[2] Jia Tracy Shen, Michiharu Yamashita, Ethan Prihar, Neil T. Heffernan, Xintao Wu, and Dongwon Lee. Mathbert: A pre-trained language model for general NLP tasks in mathematics education. *ArXiv preprint*, abs/2106.07340, 2021.