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

Waylon Li, Yftah Ziser, Maximin Coavoux, Shay B. Cohen

1. Background

What is the major difference between a general article and a mathematical article?

$$e^{i\pi} + 1 = 0$$

Task & Dataset

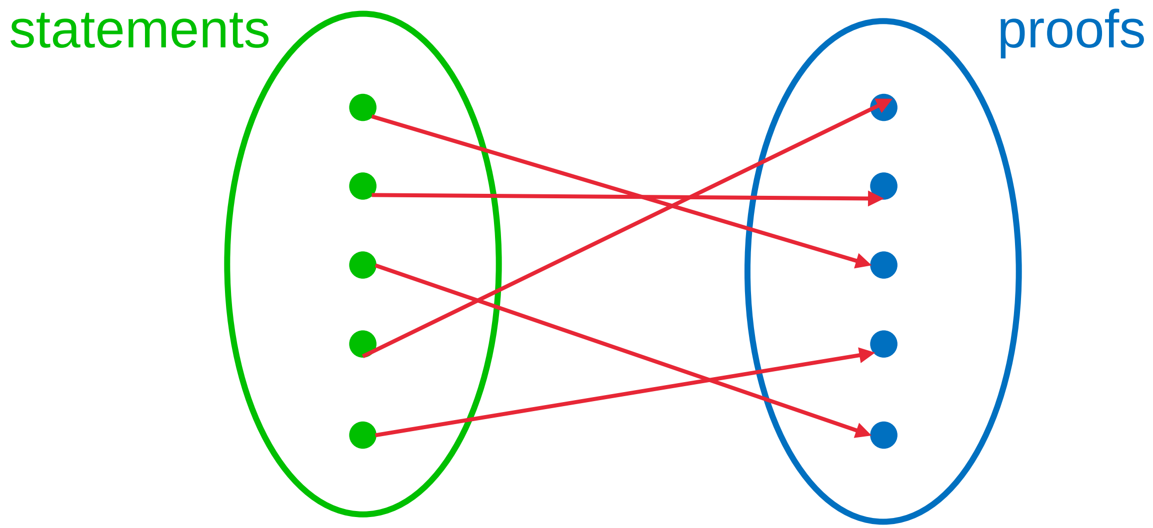
1. Background

Task:

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 problem of assigning a proof to each statement.

Statement. When $m = 0$ we have $E^0_{rg} = \emptyset$, and when $m \neq 0$ we have $E^0_{rg} = E^0$.
Proof. When $m = 0$, the image of r is $\{1\}$. Hence $E^0_{rg} = \emptyset$. When $m \neq 0$, the map r is a surjective proper map. Hence $E^0_{rg} = E^0$.

Figure 1: Example of a statement-proof pair.



THE UNIVERSITY *of* EDINBURGH informatics

Edinburgh NLP University of Edinburgh Natural Language Processing

COHORT NLP LAB

1. Background

Why we designed the task:

- Mathematical research can benefit from NLP
- Prior NLP work on mathematical research articles focused on Mathematical Information Retrieval (MIR) and related tools or data (Zanibbi et al., 2016; Stathopoulos and Teufel, 2016, 2015)
- It may help MIR by serving as a proxy for the search for the existence of a mathematical result
- Learning to match statements and proofs would also benefit computer-assisted theorem proving

2. The MATcH Dataset

Motivations for creating our dataset:

- Related datasets, such as LEANSTEP (Han et al., 2021) and the synthetic dataset of Polu and Sutskever (2020) do not include natural language.

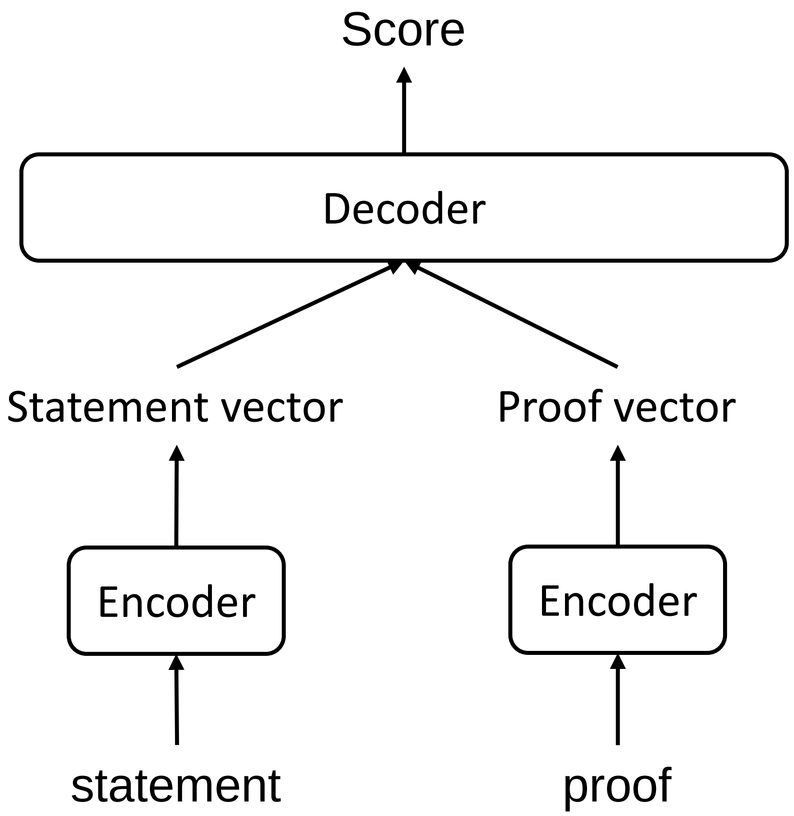
- NaturalProofs (Welleck et al., 2021) , another related dataset, only consists of 32k theorem-proof pairs from ProofWiki, some sub-topics in algebraic geometry and two textbooks.

2. The MATcH Dataset

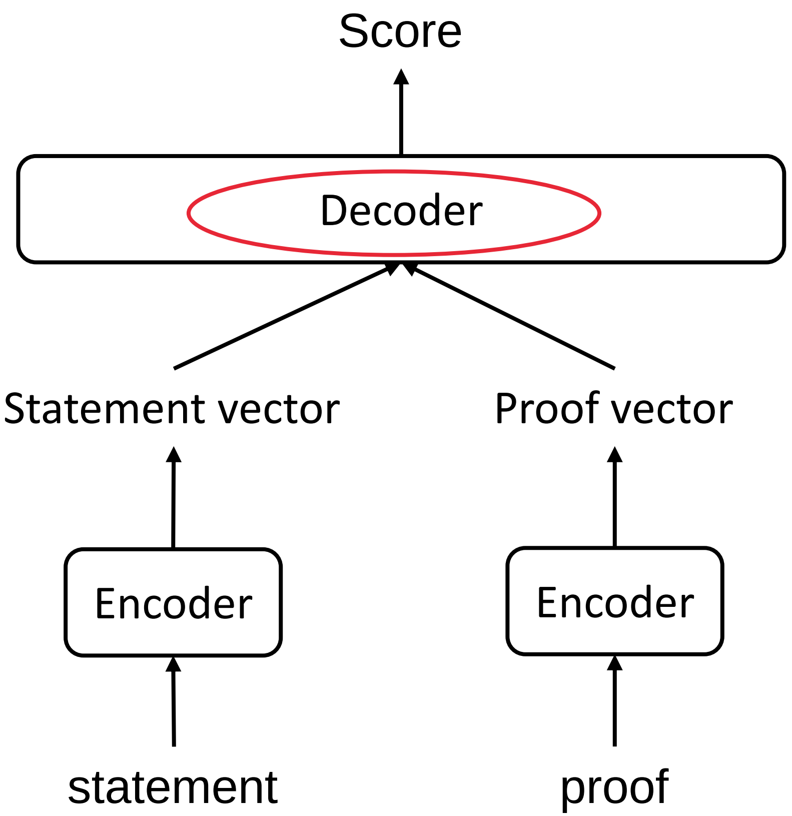
Source corpus: MREC corpus (Liska et al., 2011)
https://mir.fi.muni.cz/MREC/

- Contains around 450k articles from ArxMLiV (Stamerjohanns et al., 2010)

3. Model


Score
Decoder
Statement vector
Proof vector
Encoder
Encoder
statement
proof

3. Model



Bilinear Similarity Model

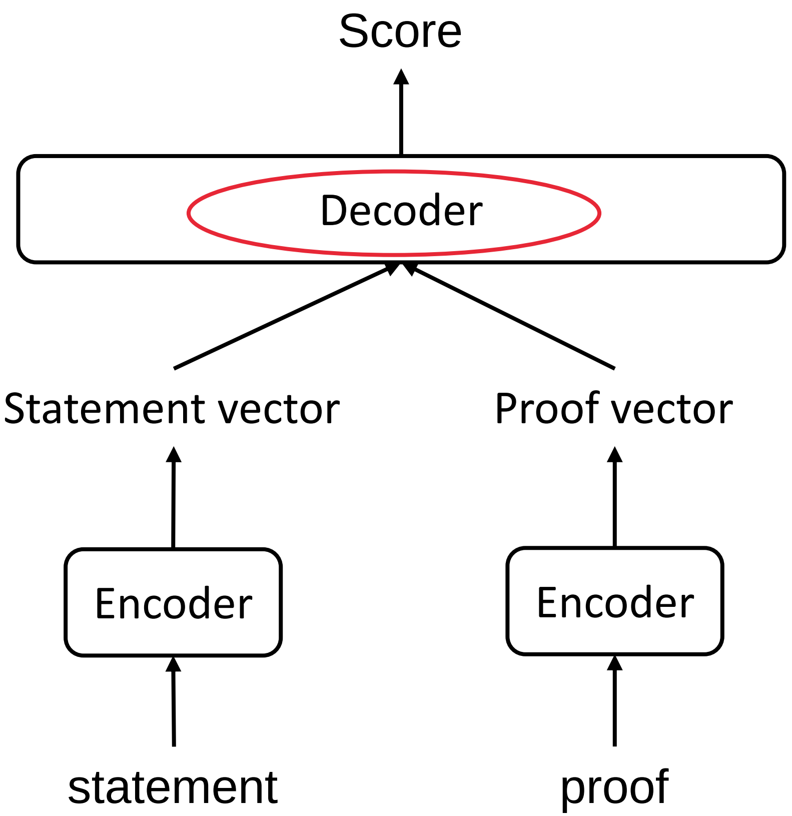
- Trainable Bilinear Similarity Function:

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

statement proof

- Local decoding
- Global decoding

3. Model



Local decoding

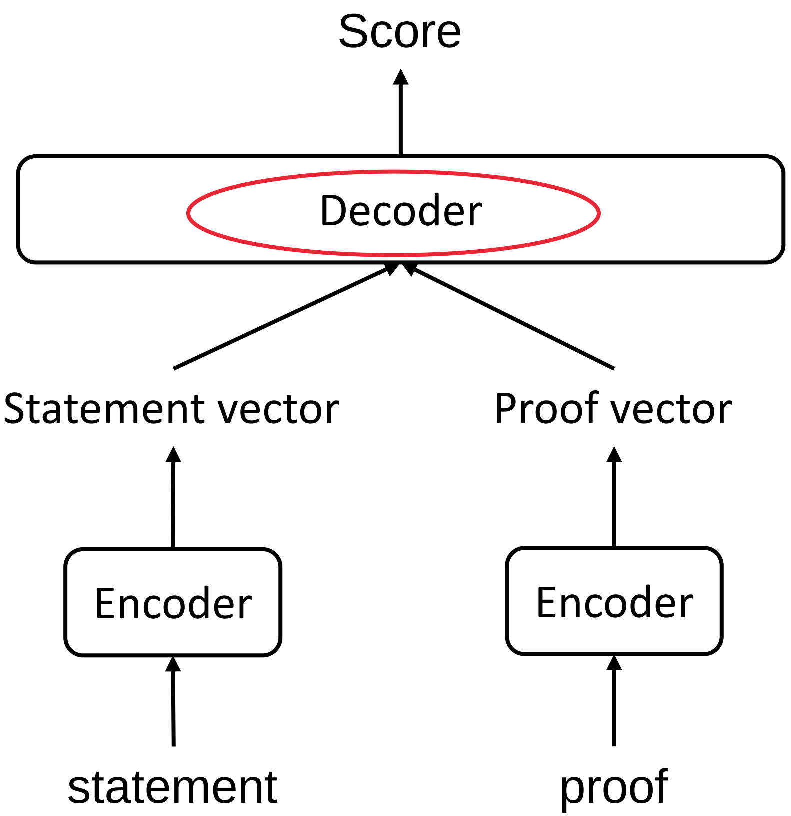
Straightforwardly sort each row by decreasing order and assign the proof ranking to the corresponding statement.

$$\hat{p}^{(i)} = \arg\max_j \; m_{ij}$$

$$m_{ij} = \text{score}(\mathbf{s}^{(i)}, \mathbf{p}^{(j)})$$

statement proof

3. Model



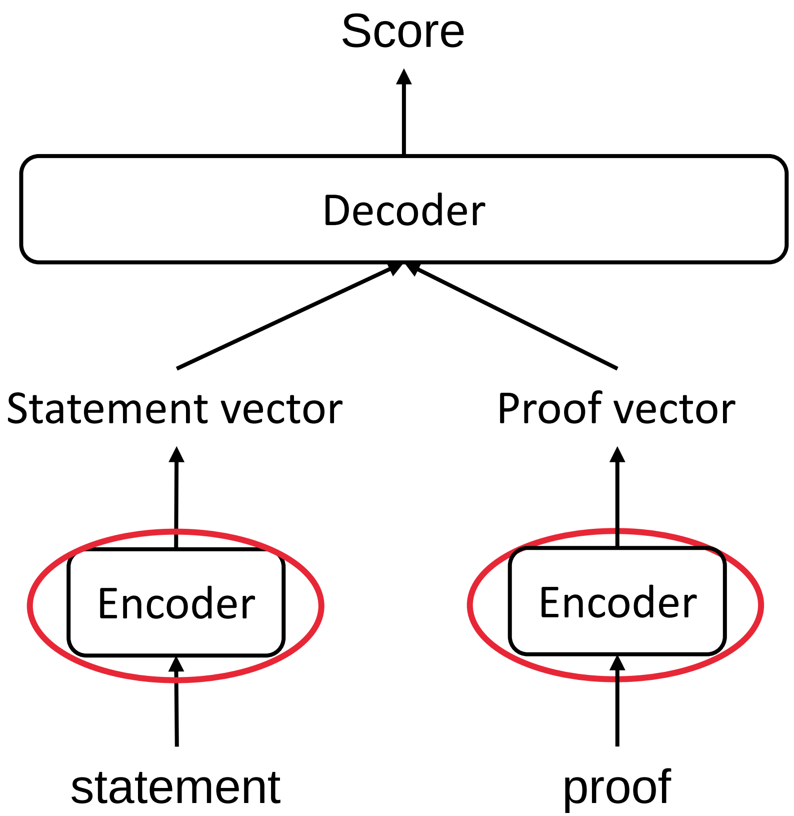
Global decoding

A proof can be assigned only to a single statement, which becomes a Linear Assignment Problem (LAP).

Statements	Proofs	%
≥ 20	7	0.0
≥ 10	80	0.2
≥ 5	1027	1.9
≥ 2	11949	22.6
$= 1$	19531	37.0
< 1	21275	40.3

Table 7: Cumulative distribution of proofs in the development set, by number of statements to which they are assigned with the local decoding method.

3. Model



Encoders:

- No pre-training encoder (NPT)
- ScratchBERT: pre-train BERT from scratch on MATcH
- MathBERT (Shen et al. 2021): a state-of-the-art pre-trained model for mathematical formula understanding

3. Model

Local training:

$$\mathcal{L}_{\text{LOC}}(s, p, P; \boldsymbol{\theta}) = -\log \mathbb{P}(p|s; \boldsymbol{\theta})$$
$$= -\log \left(\frac{\exp(\text{score}(\mathbf{s}, \mathbf{p}))}{\sum_{p' \in P} \exp(\text{score}(\mathbf{s}, \mathbf{p}'))} \right)$$

where P is the set of proofs, and θ are the parameters of the model.

3. Model

Hybrid Local and Global training:

We use the following max-margin objective, for a set B of n pairs corresponding to matrix M:

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

$$\Delta(\hat{A}, I) = \sum_{ij} \max(0, (\hat{A} - I)_{ij})$$

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

PS: this global objective had a slow convergence rate in practice, we use a hybrid local-global objective.

4. Encoders Comparison

- Importance of vocabulary
- Global decoding substantially improves accuracy

Encoder-Decoder	MRR	Acc
NPT-Local-Local	63.22	56.08
NPT-Local-Global	-	61.89
NPT-Global-Global	-	62.14
ScratchBERT-Local-Local	**73.73**	67.12
ScratchBERT-Local-Global	-	**74.68**
ScratchBERT-Global-Global	-	71.38
MathBERT-Local-Local	54.51	46.45
MathBERT-Local-Global	-	49.77
MathBERT-Global-Global	-	45.38

5. Symbol Replacement

$a_n = a_{n-1} + a_{n-2}$	Symbol conservation	All symbols remain intact, so the theorem and the proof overlap.
$x_n = x_{n-1} + x_{n-2}$	Partial symbol replacement	A fraction of α of all the symbols in the proof remain the same, and the rest are changed. In our experiments, we use α = 0.5.
$x_i = x_{i-1} + x_{i-2}$	Full symbol replacement	All symbol names are changed (α = 1.0 as above).
$n_a = n_{a-1} + n_{a-2}$	Symbol transposition	We permute the variables' names such that no symbol remains the same, thus changing their original functionality.

THE UNIVERSITY *of* EDINBURGH informatics

Edinburgh NLP University of Edinburgh Natural Language Processing

COHORT NLP LAB

6. Cross Replacement Results

Source \ Target		Conservation		Partial		Full		Transposition	
		Symbol Replacement							
		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**

- Strong dependency on exact symbol name matching
- Lack of importance of mathematical functionality, order and context
- Significant resilience when trained on partial symbol replacement level

7. Qualitative Analysis: LIME (Ribeiro et al., 2016)

Lemma 3.2. Let M be a module and H a local submodule of M. Then H is a supplement of each proper submodule $K \leq M$ 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.
(https://arxiv.org/pdf/0810.0041.pdf)

(a) Example statement/proof 1 - 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 $K \leq M$ 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.
(https://arxiv.org/pdf/0810.0041.pdf)

(b) Example statement/proof 1 - Full symbol replacement

8. Protected symbols

Symbol	Usage	Articles with usage
P	$P(A)$	Probability measure
E	$E(X)$	Expected value
V	$V(X)$	Variance
σ	$\sigma(X)$	Standard deviation
	$\sigma(X,Y)$	Covariance
ρ	$\rho(X,Y)$	Correlation

Source \ Target	Symbol Replacement			
	Conservation		Partial+P	
	MRR	Acc	MRR	Acc
Conservation	**69.26**	**59.59**	27.9	18.29
Partial	61.36	51.72	54.06	42.67
Partial+P	62.1	51.92	**55.92**	**45.23**
Full	53.63	42.08	52.85	41.4
Full+P	56.27	45.13	**55.92**	44.84

Table 6: Controlled cross-replacement levels performance for the SCRATCHBERT-Local-Local model. Both train and test sets are curated from the probability theory domain. +P next to a symbol replacement method means that Protected symbols are not being replaced.

9. Conclusion & Contribution

- A large dataset (MATcH) for a task focusing on the domain of mathematical research articles
- We proposed two ways to train and do inference with our model and dataset: local matching and global matching
- We assessed the difficulty of the task with several pre-trained encoders, demonstrating the importance of the vocabulary support for these models
- We run further assessment relying on symbol replacement and observe that the model makes a relatively shallow use of the text and formulae to obtain this performance

Thank you!

https://bollin.inf.ed.ac.uk/match.html
https://github.com/waylonli/MATcH