IsarStep: a Benchmark for High-level Mathematical Reasoning

Wenda Li

University of Cambridge

wl302@cam.ac.uk

March 26, 2021

Joint work with Lei Yu, Yuhuai Wu, and Larry Paulson

Overview

- A task on synthesising intermediate propositions in non-trivial mathematical derivations.
- A non-synthetic dataset from the largest repository of mechanised proofs: 204K lemmas from 375 contributors.
- A hierarchical transformer that outperforms the transformer baseline.
- A test suite for checking the correctness of types and the validity of the generated propositions.
- The dataset and model implementation are available at `https://github.com/Wenda302/IsarStep`.

Proof of the irrationality of $\sqrt{2}$.

Assume $\sqrt{2}$ is rational. Then there exists a pair of coprime integers a and b such that $\sqrt{2} = a/b$, and it follows that $2 = a^2/b^2$ and then $2b^2 = a^2$. Hence a is even. Thus there exists an integer c such that $a = 2c$, which combined with $2b^2 = a^2$ yields $2c^2 = b^2$: hence b is also even. So a and b are both even although they are coprime, contradiction. $\square$

Proof of the irrationality of $\sqrt{2}$.

Assume $\sqrt{2}$ is rational. Then there exists a pair of coprime integers a and b such that $\sqrt{2} = a/b$, and it follows that $2 = a^2/b^2$ and then $2b^2 = a^2$. Hence a is even. Thus there exists an integer c such that $a = 2c$, which combined with $2b^2 = a^2$ yields $2c^2 = b^2$: hence b is also even. So a and b are both even although they are coprime, contradiction. □

An informal (high-level) proof is, mostly, a sequence of intermediate propositions connected with logical relations (e.g. then, thus, yield).

Informal vs. (Mechanised) Declarative Proofs

Proof of the irrationality of $\sqrt{2}$.

Assume $\sqrt{2}$ is rational. Then there exists a pair of coprime integers a and b such that $\sqrt{2} = a/b$, and it follows that $2 = a^2/b^2$ and then $2b^2 = a^2$. Hence a is even. Thus there exists an integer c such that $a = 2c$, which combined with $2b^2 = a^2$ yields $2c^2 = b^2$: hence b is also even. So a and b are both even although they are coprime, contradiction. □

```
theorem "sqrt 2 ∉ ℚ"
proof
  assume "sqrt 2 ∈ ℚ"
  then obtain a b :: int where
      "sqrt 2 = a/b" "coprime a b"
    \<proof>
  then have "2 = a² / b²" \<proof>
  then have *:"2*b² = a²"
    \<proof>
  then have "even a" \<proof>
  then obtain c::int where "a=2*c"
    \<proof>
  with * have "b² = 2*c²" \<proof>
  then have "even b" \<proof>
  with ‹even a› ‹coprime a b›
  show False \<proof>
qed
```

Proving the irrationality of $\sqrt{2}$ in Isabelle/HOL

```
theorem "sqrt 2 ∉ ℚ"
proof
  assume "sqrt 2 ∈ ℚ"
  then obtain a b :: int where "sqrt 2 = a/b" "coprime a b"
    by (metis Rat_cases Rats_def imageE normalize_stable of_rat_divide
          of_rat_of_int_eq quotient_of_Fract quotient_of_div)
  then have "2 = a² / b²" by (smt of_int_power power_divide real_sqrt_pow2)
  then have *:"2*b² = a²"
    by (cases "b=0",auto simp:field_simps,use of_int_eq_iff in fastforce)
  then have "even a" by (metis dvd_triv_left even_mult_iff power2_eq_square)
  then obtain c::int where "a=2*c" by blast
  with * have "b² = 2*c²" by simp
  then have "even b" by (metis dvd_triv_left even_mult_iff power2_eq_square)
  with ‹even a› ‹coprime a b› show False by auto
qed
```

The task

Part of the derivation in the previous example is as follows:

$$\underbrace{2b^2 = a^2}_{(1)} \Rightarrow \underbrace{a \text{ is even}}_{(2)} \Rightarrow \underbrace{\exists c \in \mathbb{Z}.\, a = 2c}_{(3)}.$$

where the proposition (2) bridges the gap between (1) and (3).

In IsarStep, we want to synthesise (2) given (1) and (3).

Dataset statistics

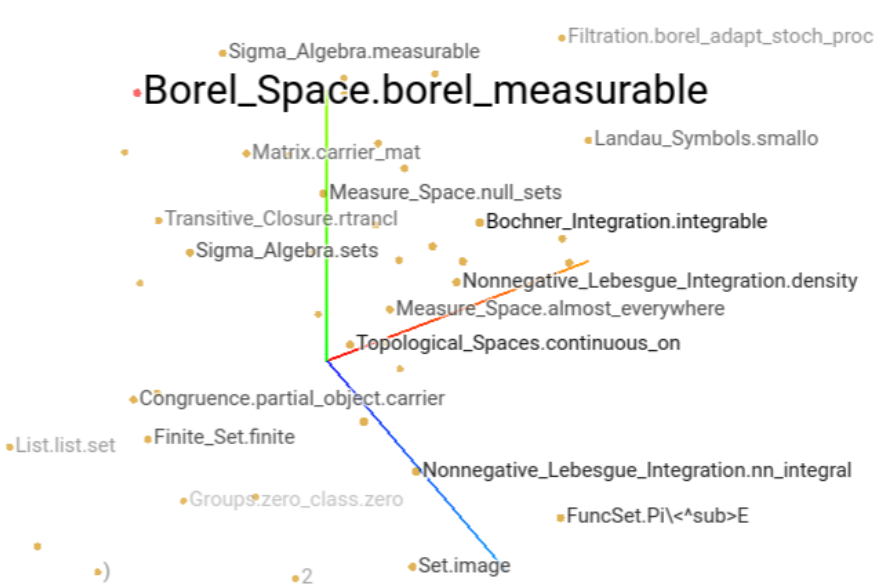
We mined the Archieve of Formal Proofs and the standard library of Isabelle/HOL:

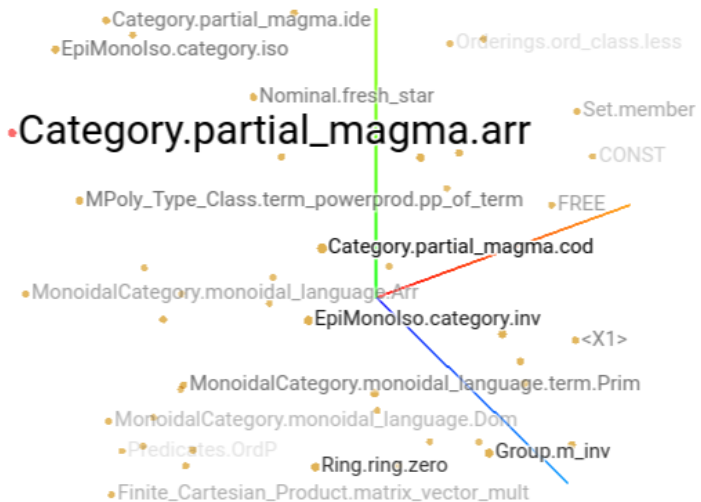
- About 1.2 million raw data points from 204K lemmas
- After filtering, the final split is 820K, 5K, 5K for the training, validation, and test set
- Average length of input is 310, and the average length of output is 43.
- The vocabulary size is about 30K.

Results

Model	Top-1 Acc.		Top-10 Acc.		BLEU	
	Base	+**F.5**	Base	+**F.5**	Base	+**F.5**
RNNSearch	13.0	16.7	26.2	32.2	42.3	52.2
Transformer	20.4	22.1	33.1	34.6	59.6	62.9
HAT	**22.8**	**24.3**	**35.2**	**37.2**	**61.8**	**65.7**

Embedding space





An example

Source (pretty printed for readability): $? \wedge x_{57} \subseteq x_{39} \Rightarrow x_{70} \in x_{39}$

The model gives: $x_{70} \in x_{57}$

The derivation involves the following lemma:

$$x \in A, A \subseteq B \vdash x \in B.$$

Some robustness

Source: $? \wedge \cancel{x_{57} \subseteq x_{39}}\ x_{39} \subseteq x_{57} \Rightarrow x_{70} \in x_{39}$

The model gives: $\cancel{x_{70} \in x_{57}}\ x_{70} \in x_{39}$

The previous lemma $x \in A, A \subseteq B \vdash x \in B$ no longer applies – to prove $x_{70} \in x_{39}$, we can only discharge itself.

Thanks for your attention.