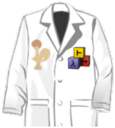
Improving Proof Assistant User Productivity Using Language Models

Pengyu Nie[1], **Karl Palmskog**[2],
Junyi Jessy Li[1], and Milos Gligoric[1]



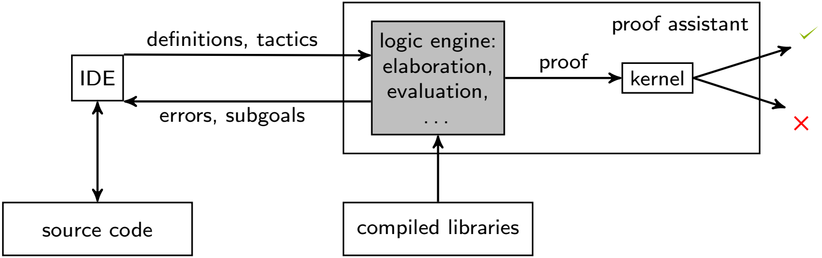
[1] The University of Texas at Austin

[2] KTH Royal Institute of Technology

Partially supported by: UT Austin Graduate School Continuing Fellowship; NSF Grant Nos. CCF-1652517, IIS-1850153; Trustfull project

Context: Verification using Proof Assistants

1. encode definitions in higher-order language
2. prove specifications **semi-interactively**
3. soundness of every low-level step checked by **kernel**



Examples: Coq, Lean, Isabelle/HOL, HOL4, ...

Verification Projects Keep Growing in Size

proof assistants are increasingly used formalize advanced mathematics and develop large trustworthy software systems

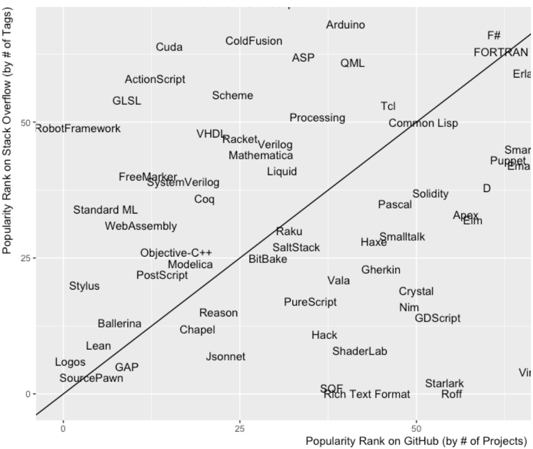
Project	**Domain**	**Assistant**	**LOC**
CompCert	compiler	Coq	100k+
MathComp	math	Coq	100k+
Verdi Raft	systems	Coq	50k+
seL4	kernel	Isabelle/HOL	200k+
CakeML	compiler	HOL4	200k+
Mathlib	math	Lean	300k+

challenges similar to in large software projects:
engineer productivity, tools, ecosystem, evolution

Example: The Coq Ecosystem

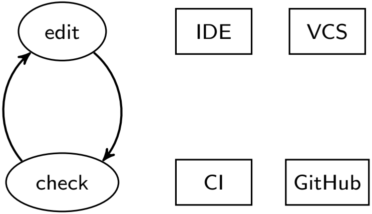
- the Coq Proof Assistant and its standard library
- the Coq Platform of 25+ curated libraries
- Coq's continuous integration suite of 2.5M+ LOC
- coq-community on GitHub with 40+ projects and 700k+ LOC
- millions of LOC in projects on GitHub and GitLab

Proof Assistants on GitHub and StackOverflow



https://redmonk.com/sogrady/2021/08/05/language-rankings-6-21/

Our Research at a Glance



- improving the feedback loop for proof engineers
- leveraging of techniques from software engineering
- Coq used for evaluation and case studies
- previously: regression proof selection, mutation proving
- today: support for **adhering to coding conventions**

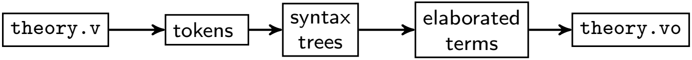
Key idea: apply *language models* to Coq code

Language Models and Programming Languages

- language models capture **regularities** in large text corpora
- developed initially for Natural Language Processing
- programming languages are very repetitive (have high **naturalness**)
- naturalness recently used to learn from code corpora and perform tasks:
  - predict the next token in a sequence
  - translate to other language
  - summarize functionality
  - ...
- Coq code has high naturalness (Hellendoorn et al., FSE New Ideas and Emerging Results, 2018)

Coq in a Nutshell

- choice of IDEs: CoqIDE, VS Code, Emacs, Vim
- Vernacular **command language**
- Gallina, extensible **specification language**
- foundational calculus (CIC) of **terms** and **types**
- **checker** that a term has a given type
- implemented in OCaml



Coq's Zoo of Languages

```
Require Import List.
Require Import ListUtil.
Import ListNotations.

Definition dedup :=
fix dedup_aux A A_eq_dec (xs : list A) : list A :=
match xs with
| [] => []
| x :: xs =>
  if in_dec A_eq_dec x xs then dedup_aux A A_eq_dec xs
  else x :: dedup_aux A A_eq_dec xs
end.

Lemma remove_dedup :
 forall A A_eq_dec (x : A) xs,
  remove A_eq_dec x (dedup A A_eq_dec xs) =
  dedup A A_eq_dec (remove A_eq_dec x xs).
Proof.
induction xs; intros; auto; simpl.
repeat (try case in_dec; try case A_eq_dec;
 simpl; intuition); auto using f_equal.
- exfalso. apply n0. apply remove_preserve; auto.
- exfalso. apply n. apply in_remove in i; intuition.
Qed.
```

Coq's Zoo of Languages

```
Require Import List.
Require Import ListUtil.
Import ListNotations.

Definition dedup :=
fix dedup_aux A A_eq_dec (xs : list A) : list A :=
match xs with
| [] => []
| x :: xs =>
  if in_dec A_eq_dec x xs then dedup_aux A A_eq_dec xs
  else x :: dedup_aux A A_eq_dec xs
end.

Lemma remove_dedup :
 forall A A_eq_dec (x : A) xs,
  remove A_eq_dec x (dedup A A_eq_dec xs) =
  dedup A A_eq_dec (remove A_eq_dec x xs).
Proof.
induction xs; intros; auto; simpl.
repeat (try case in_dec; try case A_eq_dec;
 simpl; intuition); auto using f_equal.
- exfalso. apply n0. apply remove_preserve; auto.
- exfalso. apply n. apply in_remove in i; intuition.
Qed.
```

Vernacular commands — always end in a period.

Coq's Zoo of Languages

```
Require Import List.
Require Import ListUtil.
Import ListNotations.

Definition dedup :=
fix dedup_aux A A_eq_dec (xs : list A) : list A :=
match xs with
| [] => []
| x :: xs =>
  if in_dec A_eq_dec x xs then dedup_aux A A_eq_dec xs
  else x :: dedup_aux A A_eq_dec xs
end.

Lemma remove_dedup :
 forall A A_eq_dec (x : A) xs,
  remove A_eq_dec x (dedup A A_eq_dec xs) =
  dedup A A_eq_dec (remove A_eq_dec x xs).
Proof.
induction xs; intros; auto; simpl.
repeat (try case in_dec; try case A_eq_dec;
 simpl; intuition); auto using f_equal.
- exfalso. apply n0. apply remove_preserve; auto.
- exfalso. apply n. apply in_remove in i; intuition.
Qed.
```

Gallina definition of a recursive function to remove duplicate list elements.

Coq's Zoo of Languages

```
Require Import List.
Require Import ListUtil.
Import ListNotations.

Definition dedup :=
fix dedup_aux A A_eq_dec (xs : list A) : list A :=
match xs with
| [] => []
| x :: xs =>
  if in_dec A_eq_dec x xs then dedup_aux A A_eq_dec xs
  else x :: dedup_aux A A_eq_dec xs
end.

Lemma remove_dedup :
 forall A A_eq_dec (x : A) xs,
  remove A_eq_dec x (dedup A A_eq_dec xs) =
  dedup A A_eq_dec (remove A_eq_dec x xs).
Proof.
induction xs; intros; auto; simpl.
repeat (try case in_dec; try case A_eq_dec;
 simpl; intuition); auto using f_equal.
- exfalso. apply n0. apply remove_preserve; auto.
- exfalso. apply n. apply in_remove in i; intuition.
Qed.
```

Statement (type) of a lemma in Gallina.

Coq's Zoo of Languages

```
Require Import List.
Require Import ListUtil.
Import ListNotations.

Definition dedup :=
fix dedup_aux A A_eq_dec (xs : list A) : list A :=
match xs with
| [] => []
| x :: xs =>
  if in_dec A_eq_dec x xs then dedup_aux A A_eq_dec xs
  else x :: dedup_aux A A_eq_dec xs
end.

Lemma remove_dedup :
 forall A A_eq_dec (x : A) xs,
  remove A_eq_dec x (dedup A A_eq_dec xs) =
  dedup A A_eq_dec (remove A_eq_dec x xs).
Proof.
induction xs; intros; auto; simpl.
repeat (try case in_dec; try case A_eq_dec;
 simpl; intuition); auto using f_equal.
- exfalso. apply n0. apply remove_preserve; auto.
- exfalso. apply n. apply in_remove in i; intuition.
Qed.
```

Proof script in **Ltac** – executed to produce a **term**.

Problem: Hard-coded Lemma Naming Conventions

CONTRIBUTIONS.md in MathComp, 50+ entries

Naming conventions for lemmas (non exhaustive)

Names in the library usually obey one of the following conventions

- `(condition_)?mainSymbol_suffixes`
- `mainSymbol_suffixes(_condition)?` Or in the presence of a property denoted by an n-ary or unary predicate:
- `naryPredicate_mainSymbol+`
- `mainSymbol_unaryPredicate`

Where:

- `mainSymbol` is the most meaningful part of the lemma. It generally is the head symbol of the right-hand side of an equation or the head symbol of a theorem. It can also simply be the main object of study, head symbol or not. It is usually either
  - one of the main symbols of the theory at hand. For example, it will be `opp`, `add`, `mul`, etc., or
  - a special "canonical" operation, such as a ring morphism or a subtype predicate. e.g. `linear`, `raddf`, `rmorph`, `rpred`, etc.
- "condition" is used when the lemma applies under some hypothesis.
- "suffixes" are there to refine what shape and/or what other symbols the lemma has. It can either be the name of a symbol ("add", "mul", etc.), or the (short) name of a predicate ("`inj`" for "`injectivity`", "`id`" for "identity", etc.) or an abbreviation. Abbreviations are in the header of the file which introduces them. We list here the main abbreviations.
- `A` -- associativity, as in `andbA : associative andb.`
- `AC` -- right commutativity.
- `ACA` -- self-interchange (inner commutativity), e.g., `orbACA : (a || b) || (c || d) = (a || c) || (b || d).`
- `b` -- a boolean argument, as in `andbb : idempotent andb.`
- `C` -- commutativity, as in `andbC : commutative andb.` -- alternatively, predicate or set complement, as in `predC.`

Many Naming Inconsistencies in Large Projects

math-comp / **math-comp**

Code | Issues 69 | Pull requests 29 | Actions | Projects | Wiki | Security | Insights

mem_imset mem_map naming/statement inconsistency #508

Open **chdoc** opened this issue on 14 May · 9 comments

chdoc commented on 14 May — Contributor

while cleaning, I noticed the following oddity:

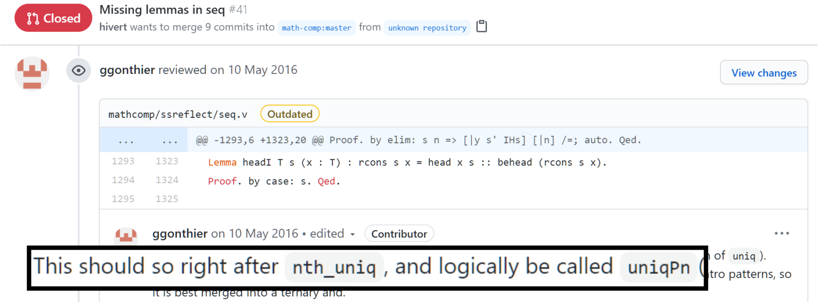
```
Lemma mem_imset (aT rT : finType) (f : aT -> rT) (D : {pred aT}) (x : aT) :
  x \in D -> f x \in [set f x | x in D]

Lemma map_f (T1 T2 : eqType) (f : T1 -> T2) (s : seq T1) (x : T1) :
   x \in s -> f x \in [seq f i | i <- s]
Lemma mem_map (T1 T2 : eqType) (f : T1 -> T2),
  injective f -> forall (s : seq T1) (x : T1), (f x \in [seq f i | i <- s]) = (x \in s)
```

Wouldn't it be more consistent to rename `mem_imset` to `imset_f`

```
Lemma mem_imset (aT rT : finType) (f : aT -> rT) (A : {set aT}) (x : aT) :
  injective f -> (f x \in f @: A) = (x \in A).
```

👍 1

Manually Checking and Enforcing Names



Our Recent Work

- **Roosterize**: toolchain for learning and suggesting lemma names
  - Code review process
  - Interactive development
  - Batch mode

Our Recent Work

- **Roosterize**: toolchain for learning and suggesting lemma names
  - Code review process
  - Interactive development
  - Batch mode

- Novel **generation models** based on multi-input encoder-decoder neural networks leveraging **elaborated terms**

Our Recent Work

- **Roosterize**: toolchain for learning and suggesting lemma names
  - Code review process
  - Interactive development
  - Batch mode

- Novel **generation models** based on multi-input encoder-decoder neural networks leveraging **elaborated terms**

- A **corpus** of 164k LOC high quality Coq code

Our Recent Work

- **Roosterize**: toolchain for learning and suggesting lemma names
  - Code review process
  - Interactive development
  - Batch mode
- Novel **generation models** based on multi-input encoder-decoder neural networks leveraging **elaborated terms**
- A **corpus** of 164k LOC high quality Coq code
- An extensive **evaluation** on our corpus via automated metrics

Our Recent Work

- **Roosterize**: toolchain for learning and suggesting lemma names
  - Code review process
  - Interactive development
  - Batch mode
- Novel **generation models** based on multi-input encoder-decoder neural networks leveraging **elaborated terms**
- A **corpus** of 164k LOC high quality Coq code
- An extensive **evaluation** on our corpus via automated metrics
- A qualitative **case study** on a project outside corpus

Our Recent Work

- **Roosterize**: toolchain for learning and suggesting lemma names
  - Code review process
  - Interactive development
  - Batch mode
- Novel **generation models** based on multi-input encoder-decoder neural networks leveraging **elaborated terms**
- A **corpus** of 164k LOC high quality Coq code
- An extensive **evaluation** on our corpus via automated metrics
- A qualitative **case study** on a project outside corpus
- **Integration** of Roosterize with VS Code editor

Running Example: A Lemma from `reglang` Project

- A lemma from a project on the theory of regular languages
- **M**ost **g**eneral classifiers can be casted to **eq**uivalent languages

```
Lemma mg_eq_proof L1 L2 (N1 : mgClassifier L1) : L1 =i L2 -> nerode L2 N1.
Proof.
move=> eq_L u v.
split=> [/nerodeP eq_in w|eq_in].
- by rewrite -!eq_L.
- apply/nerodeP=> w.
  by rewrite !eq_L.
Qed.
```

Running Example: A Lemma from `reglang` Project

- A lemma from a project on the theory of regular languages
- **M**ost **g**eneral classifiers can be casted to **eq**uivalent languages

Lemma name

```
Lemma mg_eq_proof L1 L2 (N1 : mgClassifier L1) : L1 =i L2 -> nerode L2 N1.
Proof.
move=> eq_L u v.
split=> [/nerodeP eq_in w|eq_in].
- by rewrite -!eq_L.
- apply/nerodeP=> w.
  by rewrite !eq_L.
Qed.
```

Running Example: A Lemma from `reglang` Project

- A lemma from a project on the theory of regular languages
- **M**ost **g**eneral classifiers can be casted to **eq**uivalent languages

Lemma statement

```
Lemma mg_eq_proof L1 L2 (N1 : mgClassifier L1) : L1 =i L2 -> nerode L2 N1.
Proof.
move=> eq_L u v.
split=> [/nerodeP eq_in w|eq_in].
- by rewrite -!eq_L.
- apply/nerodeP=> w.
  by rewrite !eq_L.
Qed.
```

Running Example: A Lemma from reglang Project

- A lemma from a project on the theory of regular languages
- **M**ost **g**eneral classifiers can be casted to **eq**uivalent languages

Proof script

```
Lemma mg_eq_proof L1 L2 (N1 : mgClassifier L1) : L1 =i L2 -> nerode L2 N1.
Proof.
move=> eq_L u v.
split=> [/nerodeP eq_in w|eq_in].
- by rewrite -!eq_L.
- apply/nerodeP=> w.
  by rewrite !eq_L.
Qed.
```

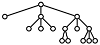
Roosterize Toolchain

`Lemma mg_eq_proof L1 L2 (N1 : mgClassifier L1) : L1 =i L2 -> nerode L2 N1.`

Lemma statement

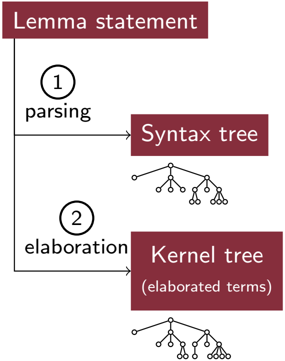
1
parsing

Syntax tree



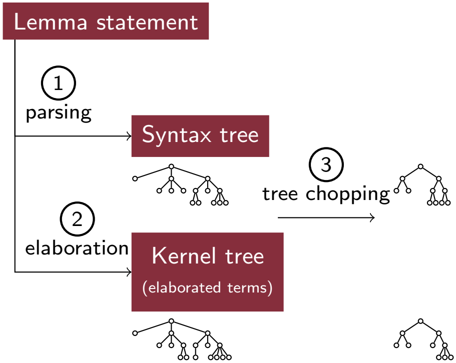
Roosterize Toolchain

```
Lemma mg_eq_proof L1 L2 (N1 : mgClassifier L1) : L1 =i L2 -> nerode L2 N1.
```



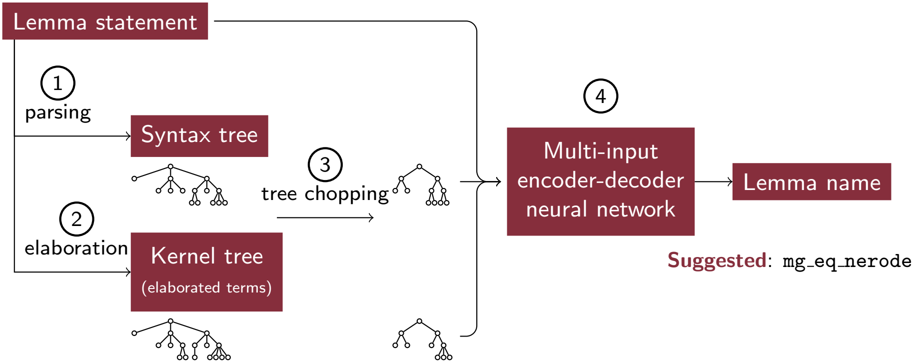
Roosterize Toolchain

```
Lemma mg_eq_proof L1 L2 (N1 : mgClassifier L1) : L1 =i L2 -> nerode L2 N1.
```



Roosterize Toolchain

```
Lemma mg_eq_proof L1 L2 (N1 : mgClassifier L1) : L1 =i L2 -> nerode L2 N1.
```



Model Input: Lemma Statement

```
Lemma mg_eq_proof L1 L2 (N1 : mgClassifier L1) : L1 =i L2 -> nerode L2 N1.
```

```
(Sentence((IDENT Lemma)(IDENT mg_eq_proof)(IDENT L1)(IDENT L2)
  (KEYWORD"(")(IDENT N1)(KEYWORD :)(IDENT mgClassifier)
  (IDENT L1)(KEYWORD")")(KEYWORD :)(IDENT L1)(KEYWORD =i)(IDENT L2)
  (KEYWORD ->)(IDENT nerode)(IDENT L2)(IDENT N1)(KEYWORD .)))
```

S-expression

- In lexing phase
- **Surface syntax** level information

Model Input: Syntax Tree

```
Lemma mg_eq_proof L1 L2 (N1 : mgClassifier L1) : L1 =i L2 -> nerode L2 N1.
```

```
(VernacExpr()(VernacStartTheoremProof Lemma (Id mg_eq_proof)
 (((CLocalAssum(Name(Id L1))(CLocalAssum(Name(Id L2)))
   (CLocalAssum(Name(Id N1))(CApp(CRef(Ser_Qualid(DirPath())(Id mgClassifier)))
    (CRef(Ser_Qualid(DirPath())(Id L1))))))
  (CNotation(InConstrEntrySomeLevel"_ -> _")
   (CNotation(InConstrEntrySomeLevel"_ =i _")
    (CRef(Ser_Qualid(DirPath())(Id L1)))(CRef(Ser_Qualid(DirPath())(Id L2))))
   (CApp(CRef(Ser_Qualid(DirPath())(Id nerode)))
    (CRef(Ser_Qualid(DirPath())(Id L2)))(CRef(Ser_Qualid(DirPath())(Id N1))))))))
```

- In parsing phase
- **Surface syntax** level information

Model Input: Kernel Tree

```
Lemma mg_eq_proof L1 L2 (N1 : mgClassifier L1) : L1 =i L2 -> nerode L2 N1.
```

```
(Prod (Name (Id char)) ... (Prod (Name (Id L1)) ...
  (Prod (Name (Id L2)) ... (Prod (Name (Id N1)) ...
   (Prod Anonymous (App (Ref (DirPath ((Id ssrbool) (Id ssr) (Id Coq))) (Id eq_mem)) ...
     (Var (Id L1)) ... (Var (Id L2)))
    (App (Ref (DirPath ((Id myhill_nerode) (Id RegLang))) (Id nerode)) ...
     (Var (Id L2)) ... (Var (Id N1))))))))
```

- In elaboration phase
- **Semantic** level information

Model Input: Kernel Tree

```
Lemma mg_eq_proof L1 L2 (N1 : mgClassifier L1) : L1 =i L2 -> nerode L2 N1.
```

```
(Prod (Name (Id char)) ... (Prod (Name (Id L1)) ...
  (Prod (Name (Id L2)) ... (Prod (Name (Id N1)) ...
   (Prod Anonymous (App (Ref (DirPath ((Id ssrbool) (Id ssr) (Id Coq))) (Id eq_mem)) ...
     (Var (Id L1)) ... (Var (Id L2)))
    (App (Ref (DirPath ((Id myhill_nerode) (Id RegLang))) (Id nerode)) ...
     (Var (Id L2)) ... (Var (Id N1))))))))
```

- In elaboration phase
- **Semantic** level information
    - Add implicit terms

Model Input: Kernel Tree

```
Lemma mg_eq_proof L1 L2 (N1 : mgClassifier L1) : L1 =i L2 -> nerode L2 N1.
```

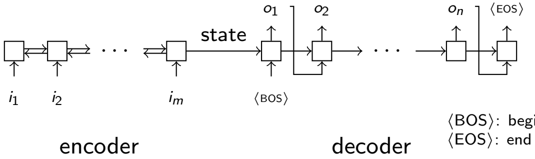
```
(Prod (Name (Id char)) ... (Prod (Name (Id L1)) ...
  (Prod (Name (Id L2)) ... (Prod (Name (Id N1)) ...
   (Prod Anonymous (App (Ref (DirPath ((Id ssrbool) (Id ssr) (Id Coq))) (Id eq_mem)) ...
     (Var (Id L1)) ... (Var (Id L2)))
    (App (Ref (DirPath ((Id myhill_nerode) (Id RegLang))) (Id nerode)) ...
     (Var (Id L2)) ... (Var (Id N1))))))))
```

- In elaboration phase
- **Semantic** level information
  - Add implicit terms
  - Translate operators to their kernel names

Lemma Naming as a Transduction Task

- **Encoder-decoder neural network**: specifically designed for **transduction tasks** (e.g., machine translation, summarization, question answering)

input
lemma statement
syntax tree
kernel tree

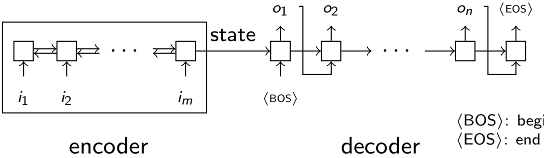


output
lemma name

⟨BOS⟩: begin of sequence
⟨EOS⟩: end of sequence

- **Encoder-decoder neural network**: specifically designed for **transduction tasks** (e.g., machine translation, summarization, question answering)

input
lemma statement
syntax tree
kernel tree



output
lemma name

⟨BOS⟩: begin of sequence
⟨EOS⟩: end of sequence

Lemma Naming as a Transduction Task

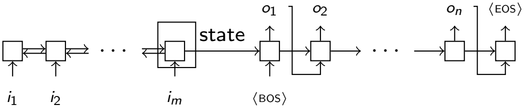
- **Encoder-decoder neural network**: specifically designed for **transduction tasks** (e.g., machine translation, summarization, question answering)

input

lemma statement

syntax tree

kernel tree



encoder

decoder

output

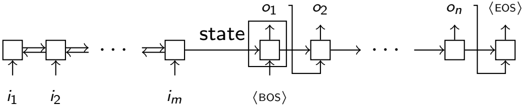
lemma name

⟨BOS⟩: begin of sequence

⟨EOS⟩: end of sequence

- **Encoder-decoder neural network**: specifically designed for **transduction tasks** (e.g., machine translation, summarization, question answering)

input
lemma statement
syntax tree
kernel tree



encoder

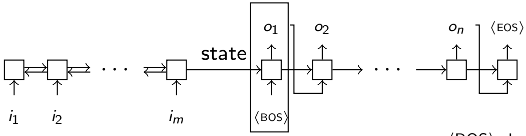
decoder

output
lemma name

⟨BOS⟩: begin of sequence
⟨EOS⟩: end of sequence

- **Encoder-decoder neural network**: specifically designed for **transduction tasks** (e.g., machine translation, summarization, question answering)

input
lemma statement
syntax tree
kernel tree



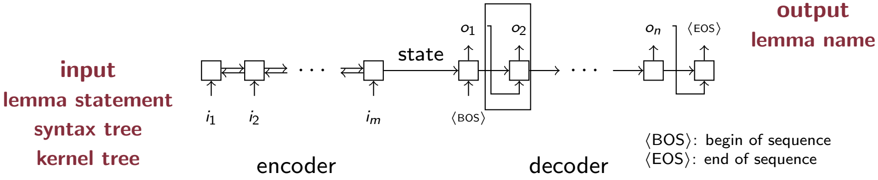
encoder

decoder

output
lemma name

⟨BOS⟩: begin of sequence
⟨EOS⟩: end of sequence

- **Encoder-decoder neural network**: specifically designed for **transduction tasks** (e.g., machine translation, summarization, question answering)



Lemma Naming as a Transduction Task

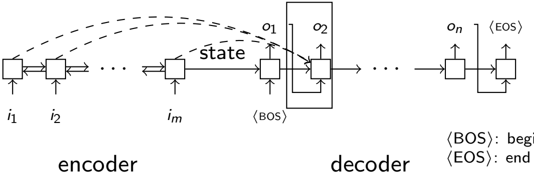
- **Encoder-decoder neural network**: specifically designed for **transduction tasks** (e.g., machine translation, summarization, question answering)
- **Attention mechanism**: decoder can "pay attention to" different parts of the inputs at each time step

input

lemma statement

syntax tree

kernel tree



output

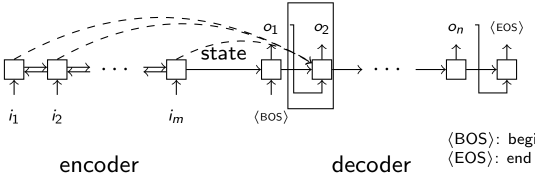
lemma name

⟨BOS⟩: begin of sequence
⟨EOS⟩: end of sequence

Lemma Naming as a Transduction Task

- **Encoder-decoder neural network**: specifically designed for **transduction tasks** (e.g., machine translation, summarization, question answering)
- **Attention mechanism**: decoder can "pay attention to" different parts of the inputs at each time step
- **Copy mechanism**: decoder can choose between copying from a sub-token in the input $\{i_1, \ldots, i_m\}$ or generating a sub-token

input
lemma statement
syntax tree
kernel tree



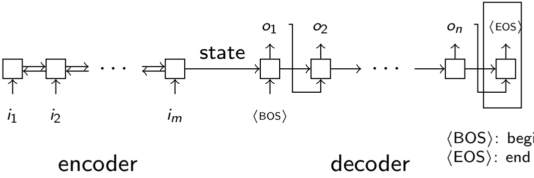
output
lemma name

⟨BOS⟩: begin of sequence
⟨EOS⟩: end of sequence

Lemma Naming as a Transduction Task

- **Encoder-decoder neural network**: specifically designed for **transduction tasks** (e.g., machine translation, summarization, question answering)
- **Attention mechanism**: decoder can "pay attention to" different parts of the inputs at each time step
- **Copy mechanism**: decoder can choose between copying from a sub-token in the input $\{i_1, \ldots, i_m\}$ or generating a sub-token

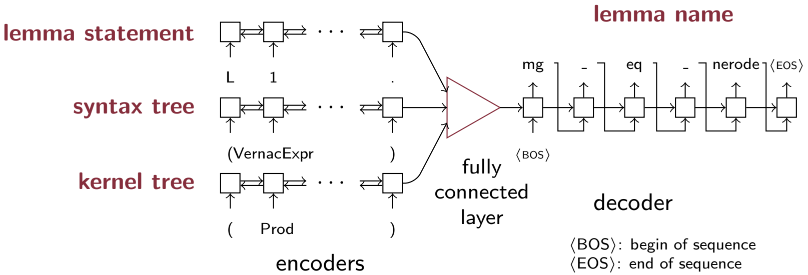
input
lemma statement
syntax tree
kernel tree



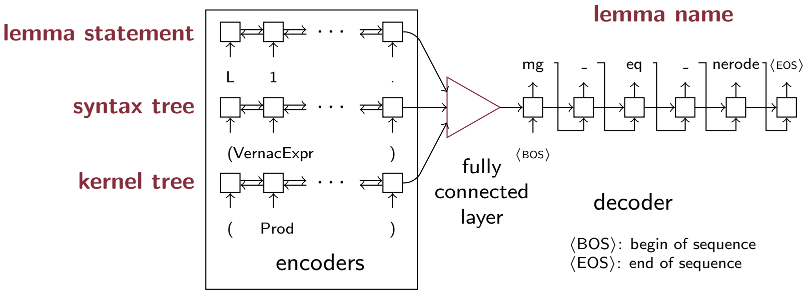
output
lemma name

⟨BOS⟩: begin of sequence
⟨EOS⟩: end of sequence

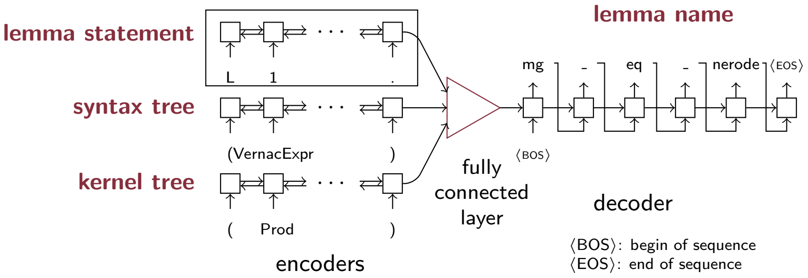
Multi-input Encoder-decoder Neural Network Architecture



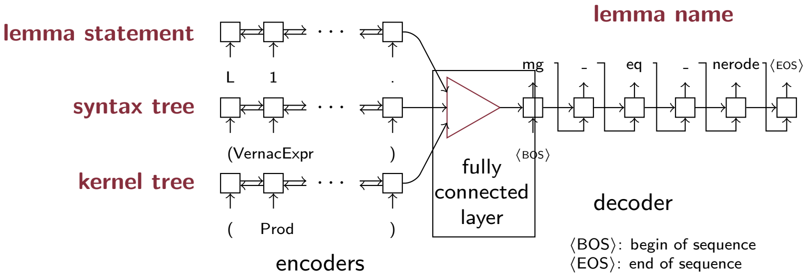
Multi-input Encoder-decoder Neural Network Architecture



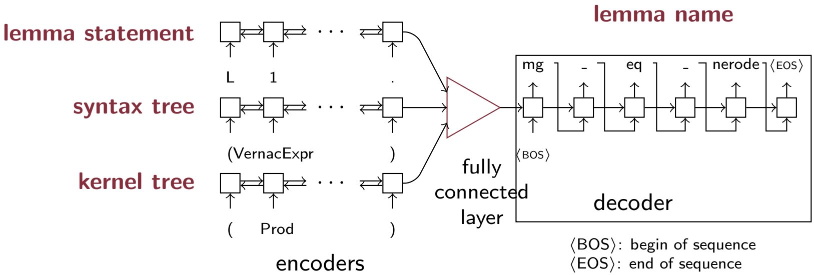
Multi-input Encoder-decoder Neural Network Architecture



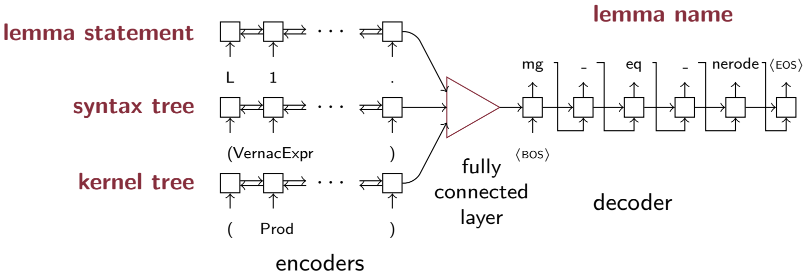
Multi-input Encoder-decoder Neural Network Architecture



Multi-input Encoder-decoder Neural Network Architecture

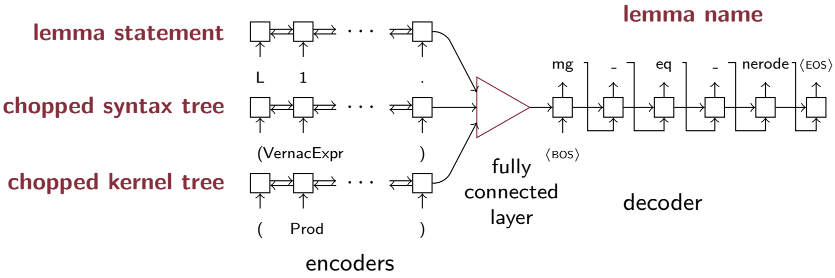


Tree Chopping



- Syntax and kernel trees can be large, which prevents the neural networks to learn effectively
- Some parts are irrelevant for naming and can be “chopped”

Tree Chopping



- Syntax and kernel trees can be large, which prevents the neural networks to learn effectively
- Some parts are irrelevant for naming and can be "chopped"
- Tree chopping heuristics:
  1. Replace the fully qualified name sub-trees with only the last component of the name
  2. Remove the location information
  3. Extract the singletons

Example Tree Chopping

- Before chopping

```
(Prod Anonymous (App (Ref (DirPath ((Id ssrbool) (Id ssr) (Id Coq))) (Id eq_mem))
  ... ((App (Ref ... ))) ... ))
```

Example Tree Chopping

- Before chopping

#1 prefixes in a fully-qualified name:
usually related to directory paths and likely not relevant

```
(Prod Anonymous (App (Ref (DirPath ((Id ssrbool) (Id ssr) (Id Coq))) (Id eq_mem))
  ... ((App (Ref ... ))) ... ))
```

#3 singleton: unnecessarily increase tree size

Example Tree Chopping

- Before chopping

#1 prefixes in a fully-qualified name: usually related to directory paths and likely not relevant

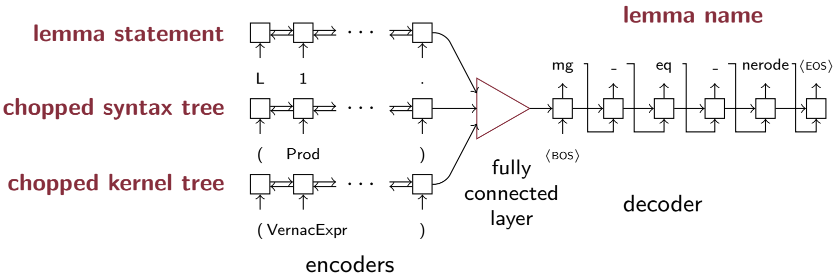
```
(Prod Anonymous (App (Ref (DirPath ((Id ssrbool) (Id ssr) (Id Coq))) (Id eq_mem))
  ... ((App (Ref ... ))) ... ))
```

#3 singleton: unnecessarily increase tree size

- After chopping

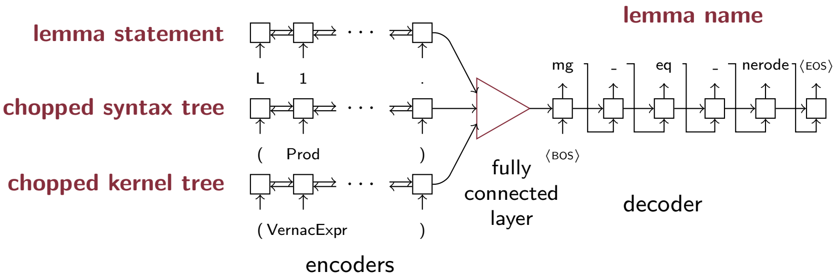
```
(Prod Anonymous (App eq_mem ... (App (Ref ... )) ... ))
```

Sub-tokenization



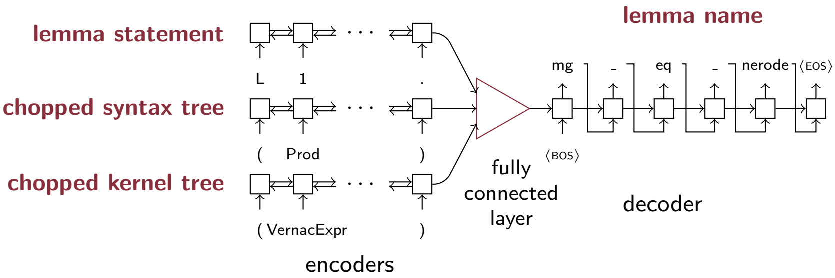
- Coq names have multiple components (e.g., prefixes and suffixes), making the vocabulary large and sparse

Sub-tokenization



- Coq names have multiple components (e.g., prefixes and suffixes), making the vocabulary large and sparse
- All inputs and outputs are **sub-tokenized** (e.g., `extprod_mulgA` $\rightarrow$ `extprod`, `_`, `mul`, `g`, and `A`)

Sub-tokenization



- Coq names have multiple components (e.g., prefixes and suffixes), making the vocabulary large and sparse
- All inputs and outputs are **sub-tokenized** (e.g., `extprod_mulgA` $\rightarrow$ `extprod`, `_`, `mul`, `g`, and `A`)
- Reduces the sparsity of the vocabulary and improves the performance of the model

Corpus: MathComp Family of Projects

- We constructed a corpus of four large Coq projects from the MathComp family, totaling **164k lines of code**
- High quality and stringent adherence to coding conventions

Project	SHA	#Files	#Lemmas	#Toks	LOC Spec.	LOC Proof
finmap	27642a8	4	940	78,449	4,260	2,191
fourcolor	0851d49	60	1,157	560,682	9,175	27,963
math-comp	748d716	89	8,802	1,076,096	38,243	46,470
odd-order	ca602a4	34	367	519,855	11,882	24,243
Avg.	N/A	46.75	2,816.50	558,770.50	15,890.00	25,216.75
Σ	N/A	187	**11,266**	2,235,082	63,560	100,867

Evaluation: Setup

- Randomly split corpus files into training, validation and test sets which contain 80%, 10%, 10% of the files, respectively

	#Files	#Lemmas	Name		Lemma Statement	
			#Char	#SubToks	#Char	#SubToks
training	152	**8,861**	10.14	4.22	44.16	19.59
validation	18	**1,085**	9.20	4.20	38.28	17.30
test	17	**1,320**	9.76	4.34	48.49	23.20

Evaluation: Setup

- Randomly split corpus files into training, validation and test sets which contain 80%, 10%, 10% of the files, respectively

	#Files	#Lemmas	Name #Char	Name #SubToks	Lemma Statement #Char	Lemma Statement #SubToks
training	152	**8,861**	10.14	4.22	44.16	19.59
validation	18	**1,085**	9.20	4.20	38.28	17.30
test	17	**1,320**	9.76	4.34	48.49	23.20

- Train Roosterize using training and validation sets
- Apply Roosterize on test set, and evaluate generated lemma names against the reference lemma names (as written by developers)

Evaluation: Automated Metrics

- **BLEU**
- **Fragment accuracy**
- **Top-1 accuracy**
- **Top-5 accuracy**

Evaluation: Automated Metrics

- **BLEU**: range 0–100, percentage of 1–4-grams overlap between the characters of the generated name and the reference name
- **Fragment accuracy**
- **Top-1 accuracy**
- **Top-5 accuracy**

$$\text{BLEU}(\texttt{card\_Syl\_dvd}, \texttt{card\_Syl\_dvd}) = 100$$
$$\text{BLEU}(\texttt{card\_Syl\_dvd}, \texttt{card\_dvd\_Syl}) = 81.9$$
$$\text{BLEU}(\texttt{card\_Syl\_dvd}, \texttt{card\_dvd}) = 52.7$$
$$\text{BLEU}(\texttt{card\_Syl\_dvd}, \texttt{Frattini\_arg}) = 14.7$$

Evaluation: Automated Metrics

- **BLEU**: range 0–100, percentage of 1–4-grams overlap between the characters of the generated name and the reference name
- **Fragment accuracy**: accuracy of generated names on the fragment level (defined by splitting the name by "_")
- **Top-1 accuracy**
- **Top-5 accuracy**

Evaluation: Automated Metrics

- **BLEU**: range 0–100, percentage of 1–4-grams overlap between the characters of the generated name and the reference name
- **Fragment accuracy**: accuracy of generated names on the fragment level (defined by splitting the name by "_")
- **Top-1 accuracy**: frequency of the reference name fully matches the generated name
- **Top-5 accuracy**

Evaluation: Automated Metrics

- **BLEU**: range 0–100, percentage of 1–4-grams overlap between the characters of the generated name and the reference name
- **Fragment accuracy**: accuracy of generated names on the fragment level (defined by splitting the name by "_")
- **Top-1 accuracy**: frequency of the reference name fully matches the generated name
- **Top-5 accuracy**: frequency of the reference name is one of the top-5 generated names

Evaluation: Results

- Key results: Roosterize significantly outperforms baselines
- Ablation studies:
  - **Tree chopping** effectively improves performance
  - Roosterize's tree chopping is better than variants
  - Using **kernel trees** in inputs effectively improves performance (i.e., **semantics** information helps naming)

Evaluation: Key Results

Model	BLEU	Frag.Acc.	Top-1	Top-5
Roosterize	**47.2**	**24.9%**	**9.6%**	**18.0%**
Baseline neural network based model	20.0	4.7%	0.1%	0.3%
Baseline retrieval-based model	28.3	10.0%	0.2%	0.3%

- Baseline neural network based model: using only lemma statement as input, w/o attention mechanism
- Baseline retrieval-based model: *details in the paper*

Evaluation: Key Results

Model	BLEU	Frag.Acc.	Top-1	Top-5
Roosterize	**47.2**	**24.9%**	**9.6%**	**18.0%**
Baseline neural network based model	20.0	4.7%	0.1%	0.3%
Baseline retrieval-based model	28.3	10.0%	0.2%	0.3%

- Baseline neural network based model: using only lemma statement as input, w/o attention mechanism
- Baseline retrieval-based model: *details in the paper*
- Roosterize, using **lemma statement and chopped kernel tree** as inputs, obtained the best performance
  - 20+ points in BLEU better than baselines
  - statistically significantly better than all other model variants

Ablation Study: Tree Chopping

Model	BLEU	Frag.Acc.	Top-1	Top-5
ChopKnlTree+attn+copy	42.9	19.8%	5.0%	11.7%
KnlTree+attn+copy	37.0	14.2%	2.2%	8.4%
ChopSynTree+attn+copy	39.8	18.3%	6.8%	12.2%
SynTree+attn+copy	31.0	10.8%	2.8%	6.1%

- **Tree chopping** improves performance by 6 points in BLEU for kernel tree and 9 points in BLEU for syntax tree

Ablation Study: Tree Chopping

Model	BLEU	Frag.Acc.	Top-1	Top-5
ChopKnlTree+attn+copy	42.9	19.8%	5.0%	11.7%
KnlTree+attn+copy	37.0	14.2%	2.2%	8.4%
ChopSynTree+attn+copy	39.8	18.3%	6.8%	12.2%
SynTree+attn+copy	31.0	10.8%	2.8%	6.1%

- **Tree chopping** improves performance by 6 points in BLEU for kernel tree and 9 points in BLEU for syntax tree
- The size of the original trees and a lot of irrelevant data in those trees hurt the performance

Ablation Study: Tree Chopping Variants

Model	BLEU	Frag.Acc.	Top-1	Top-5
Roosterize Chopping	47.2	24.9%	9.6%	18.0%
Keep-category Chopping	46.8	25.3%	9.5%	19.0%
Rule-based Chopping	37.0	17.7%	5.9%	10.5%
Random Chopping	37.7	19.2%	6.7%	10.9%

Ablation Study: Tree Chopping Variants

Model	BLEU	Frag.Acc.	Top-1	Top-5
Roosterize Chopping	47.2	24.9%	9.6%	18.0%
Keep-category Chopping	46.8	25.3%	9.5%	19.0%
Rule-based Chopping	37.0	17.7%	5.9%	10.5%
Random Chopping	37.7	19.2%	6.7%	10.9%

- **Keep-category chopping** = Roosterize chopping, but keeps the category of referenced name in kernel trees, since that semantic information could be relevant for naming

Ablation Study: Tree Chopping Variants

Model	BLEU	Frag.Acc.	Top-1	Top-5
Roosterize Chopping	47.2	24.9%	9.6%	18.0%
Keep-category Chopping	46.8	25.3%	9.5%	19.0%
Rule-based Chopping	37.0	17.7%	5.9%	10.5%
Random Chopping	37.7	19.2%	6.7%	10.9%

- **Keep-category chopping** = Roosterize chopping, but keeps the category of referenced name in kernel trees, since that semantic information could be relevant for naming
- **Rule-based chopping** chops all nodes after depth 10, similar to the proof kernel tree processing heuristics used in ML4PG

Ablation Study: Tree Chopping Variants

Model	BLEU	Frag.Acc.	Top-1	Top-5
Roosterize Chopping	47.2	24.9%	9.6%	18.0%
Keep-category Chopping	46.8	25.3%	9.5%	19.0%
Rule-based Chopping	37.0	17.7%	5.9%	10.5%
Random Chopping	37.7	19.2%	6.7%	10.9%

- **Keep-category chopping** = Roosterize chopping, but keeps the category of referenced name in kernel trees, since that semantic information could be relevant for naming
- **Rule-based chopping** chops all nodes after depth 10, similar to the proof kernel tree processing heuristics used in ML4PG
- **Random chopping** chops random 91.4% nodes from the kernel tree to match the average number of nodes of Roosterize chopped trees, as the “dumb” baseline

Ablation Study: Inputs

Inputs Combinations	BLEU	Frag.Acc.	Top-1	Top-5
Stmt+ChopKnlTree+ChopSynTree+attn+copy	45.4	22.2%	7.5%	16.5%
Stmt+ChopKnlTree+attn+copy	**47.2**	**24.9%**	**9.6%**	**18.0%**
Stmt+ChopSynTree+attn+copy	37.7	18.1%	6.1%	10.6%
ChopKnlTree+ChopSynTree+attn+copy	45.4	22.9%	7.6%	15.3%
ChopKnlTree+attn+copy	42.9	19.8%	5.0%	11.7%
ChopSynTree+attn+copy	39.8	18.3%	6.8%	12.2%
Stmt+attn+copy	38.9	19.4%	6.9%	11.6%

- The inputs combination of **lemma statement and chopped kernel tree** works the best

Ablation Study: Inputs

Inputs Combinations	BLEU	Frag.Acc.	Top-1	Top-5
Stmt+ChopKnlTree+ChopSynTree+attn+copy	45.4	22.2%	7.5%	16.5%
Stmt+ChopKnlTree+attn+copy	**47.2**	**24.9%**	**9.6%**	**18.0%**
Stmt+ChopSynTree+attn+copy	37.7	18.1%	6.1%	10.6%
ChopKnlTree+ChopSynTree+attn+copy	45.4	22.9%	7.6%	15.3%
ChopKnlTree+attn+copy	42.9	19.8%	5.0%	11.7%
ChopSynTree+attn+copy	39.8	18.3%	6.8%	12.2%
Stmt+attn+copy	38.9	19.4%	6.9%	11.6%

- The inputs combination of **lemma statement and chopped kernel tree** works the best
- Lemma statement and syntax tree do not work well together because the two representations contain mostly the same information

Ablation Study: Inputs

Inputs Combinations	BLEU	Frag.Acc.	Top-1	Top-5
Stmt+ChopKnlTree+ChopSynTree+attn+copy	45.4	22.2%	7.5%	16.5%
Stmt+ChopKnlTree+attn+copy	**47.2**	**24.9%**	**9.6%**	**18.0%**
Stmt+ChopSynTree+attn+copy	37.7	18.1%	6.1%	10.6%
ChopKnlTree+ChopSynTree+attn+copy	45.4	22.9%	7.6%	15.3%
ChopKnlTree+attn+copy	42.9	19.8%	5.0%	11.7%
ChopSynTree+attn+copy	39.8	18.3%	6.8%	12.2%
Stmt+attn+copy	38.9	19.4%	6.9%	11.6%

- The inputs combination of **lemma statement and chopped kernel tree** works the best
- Lemma statement and syntax tree do not work well together because the two representations contain mostly the same information
- Multiple inputs $\geq$ single input most of the times

Case Study: Setup

- Motivation: generated lemma names may not match the manually written ones in the corpus, but can still be **semantically valid**, which is not reflected in our automated evaluation metrics

Case Study: Setup

- Motivation: generated lemma names may not match the manually written ones in the corpus, but can still be **semantically valid**, which is not reflected in our automated evaluation metrics

- Apply Roosterize to a project outside of our corpus: the PCM library (#Files = 12, #Lemmas = 690)

Case Study: Setup

- Motivation: generated lemma names may not match the manually written ones in the corpus, but can still be **semantically valid**, which is not reflected in our automated evaluation metrics

- Apply ROOSTERIZE to a project outside of our corpus: the PCM library (#Files = 12, #Lemmas = 690)

- Automated evaluation metrics: BLEU = 36.3, fragment accuracy = 17%, Top-1 accuracy = 5% (i.e., **36 lemmas** match exactly)

Case Study: Setup

- Motivation: generated lemma names may not match the manually written ones in the corpus, but can still be **semantically valid**, which is not reflected in our automated evaluation metrics
- Apply Roosterize to a project outside of our corpus: the PCM library (#Files = 12, #Lemmas = 690)
- Automated evaluation metrics: BLEU = 36.3, fragment accuracy = 17%, Top-1 accuracy = 5% (i.e., **36 lemmas** match exactly)
- We asked the maintainer of the PCM library to evaluate **the remaining 654 lemma names** that do not match exactly and send us feedback

Case Study: Findings

- The maintainer provided comments on 150 suggested names
- 20% were of good quality, out of which more than half were of high quality
  recall the analysis was of top-1 suggestions excluding exact matches

Case Study: Findings

- The maintainer provided comments on 150 suggested names
- 20% were of good quality, out of which more than half were of high quality
  recall the analysis was of top-1 suggestions excluding exact matches
- Other suggested names tend to be "too generic"
- Unsuitable suggestions may contain useful parts

Case Study: Examples

Lemma statement: `g e k v f : path ord k (supp f) ->`
`foldfmap g e (ins k v f) = g (k, v) (foldfmap g e f)`
Hand-written: `foldf_ins`
Roosterize: `foldfmap_ins`
Comment: ✓ The whole function name is used in the suggested name.

Case Study: Examples

Lemma statement: `g e k v f : path ord k (supp f) ->`
`foldfmap g e (ins k v f) = g (k, v) (foldfmap g e f)`
Hand-written: `foldf_ins`
Roosterize: `foldfmap_ins`
Comment: ✓ The whole function name is used in the suggested name.

Lemma statement: `: transitive (@ord T)`
Hand-written: `trans`
Roosterize: `ord_trans`
Comment: ✓ Useful to add the `ord` prefix to the name.

Case Study: Examples

Lemma statement: `g e k v f : path ord k (supp f) ->`
`foldfmap g e (ins k v f) = g (k, v) (foldfmap g e f)`
Hand-written: `foldf_ins`
Roosterize: `foldfmap_ins`
Comment: ✓ The whole function name is used in the suggested name.

Lemma statement: `: transitive (@ord T)`
Hand-written: `trans`
Roosterize: `ord_trans`
Comment: ✓ Useful to add the ord prefix to the name.

Lemma statement: `p1 p2 s : kfilter (predI p1 p2) s =`
`kfilter p1 (kfilter p2 s)`
Hand-written: `kfilter_predI`
Roosterize: `eq_kfilter`
Comment: ✗ The suggested name is too generic.

Expanded Corpus: Tiers

- **21** Coq projects related to the MathComp family, totaling over **297k LOC**

Expanded Corpus: Tiers

- **21** Coq projects related to the MathComp family, totaling over **297k LOC**
- **Tier 1**: the 4 core MathComp projects used in the original corpus

Expanded Corpus: Tiers

- **21** Coq projects related to the MathComp family, totaling over **297k LOC**
- **Tier 1**: the 4 core MathComp projects used in the original corpus
- **Tier 2**: 9 projects, where each project
  - has a main contributor who is also a significant contributor to one of the tier 1 projects, and
  - follows the coding conventions specified for MathComp to a significant degree

Expanded Corpus: Tiers

- **21** Coq projects related to the MathComp family, totaling over **297k LOC**
- **Tier 1**: the 4 core MathComp projects used in the original corpus
- **Tier 2**: 9 projects, where each project
  - has a main contributor who is also a significant contributor to one of the tier 1 projects, and
  - follows the coding conventions specified for MathComp to a significant degree
- **Tier 3**: 8 projects, follow MathComp coding conventions but do not fulfill the tier 2 criteria

Expanded Corpus: Statistics

	Project	SHA	#Files	#Lemmas	#Toks	LOC Spec.	LOC Proof
Tier 1	finmap	27642a8	4	940	78,449	4,260	2,191
	fourcolor	0851d49	60	1,157	560,682	9,175	27,963
	math-comp	748d716	89	8,802	1,076,096	38,243	46,470
	odd-order	ca602a4	34	367	519,855	11,882	24,243
Tier 2	analysis	9e5fe1d	17	969	152,542	5,553	6,186
	bigenough	5f79a32	1	4	731	70	8
	elliptic-curves	631af89	18	625	110,480	3,298	6,298
	grobner	dfa54f9	1	81	15,656	312	1,018
	multinomials	691d795	5	831	83,438	3,699	3,664
	real-closed	495a1fa	10	561	108,925	4,348	4,577
	robot	b341ad1	13	864	130,024	3,881	7,249
Tier 3	two-square	1c09aca	2	200	20,326	413	1,308
	bits	3cd298a	10	411	40,420	1,578	2,463
	comp-dec-pdl	c1f813b	16	494	61,731	2,305	2,114
	disel	e8aa80c	20	256	51,473	2,575	1,898
	fcsl-pcm	eef4503	12	690	70,273	2,937	2,852
	games	3d3bd31	12	231	43,438	1,450	3,503
	monae	9d0e461	18	349	73,578	3,422	3,233
	reglang	da333e0	12	230	41,327	1,299	1,734
	toychain	97bd697	14	67	61,997	1,747	3,528
Left-out	infotheo	6c17242	81	1,891	463,593	12,517	29,778
	Avg.	N/A	21.38	953.33	179,287.33	5,474.48	8,679.90
	Σ	N/A	449	20,020	3,765,034	114,964	182,278

Left-out: the project is taken out of the tier 2 projects, and will not be used in the standard training/validation/test experiment, but will be used for the generalizability study

Evaluation on the Expanded Corpus: Setup

- Randomly split corpus files into training, validation and test sets which contain 80%, 10%, 10% of the files, respectively

	#Files	#Lemmas	Name		Lemma Statement	
			#Char	#SubToks	#Char	#SubToks
training	302	**15,011**	9.99	4.12	47.93	21.20
validation	36	**1,556**	9.20	4.08	41.44	18.65
test	30	**1,562**	9.68	4.26	49.19	23.21

Evaluation on the Expanded Corpus: Results

Model	BLEU	Frag.Acc.	Top-1	Top-5
Stmt+ChopKnlTree+ChopSynTree+attn+copy	43.2	23.2%	7.1%	15.5%
Stmt+ChopKnlTree+attn+copy	**47.2**	**26.1%**	**10.3%**	**19.0%**
Stmt+ChopSynTree+attn+copy	34.9	18.0%	4.9%	10.7%
ChopKnlTree+ChopSynTree+attn+copy	44.2	22.2%	7.4%	14.8%
ChopKnlTree+attn+copy	44.1	20.9%	5.8%	13.1%
ChopSynTree+attn+copy	39.0	19.1%	7.9%	13.3%
Stmt+attn+copy	39.7	20.8%	7.5%	13.6%
Baseline neural network based model	20.3	5.4%	0.3%	0.5%
Baseline retrieval-based model	28.2	10.9%	0.6%	0.8%

- Similar finding: ROOSTERIZE using lemma statement and chopped kernel tree as inputs obtained the best performance
- Metrics are slightly higher compared to the results on the original (tier 1) corpus, because more data is used for training the model

Generalizability Study: Setup

- Goal: to evaluate
  - whether ROOSTERIZE can be used on a project outside of the training corpus
  - whether additional training (using the data from that project) is needed

Generalizability Study: Setup

- Goal: to evaluate
  - whether ROOSTERIZE can be used on a project outside of the training corpus
  - whether additional training (using the data from that project) is needed
- The left-out project (infotheo): 81 Coq files, 1,891 lemmas

Generalizability Study: Setup

- Goal: to evaluate
  - whether ROOSTERIZE can be used on a project outside of the training corpus
  - whether additional training (using the data from that project) is needed
- The left-out project (infotheo): 81 Coq files, 1,891 lemmas
- Split the files into training, validation, and test sets which contain 40%, 10%, 50% of the files (#files: 580, 144, 1,167)

Generalizability Study: Setup

- Goal: to evaluate
  - whether ROOSTERIZE can be used on a project outside of the training corpus
  - whether additional training (using the data from that project) is needed
- The left-out project (infotheo): 81 Coq files, 1,891 lemmas
- Split the files into training, validation, and test sets which contain 40%, 10%, 50% of the files (#files: 580, 144, 1,167)
- Applying ROOSTERIZE on infotheo
  - **without additional training**
  - **with additional training** using {25%, 50%, 75%, 100%} of the training set

Generalizability Study: Results

#Lemmas	BLEU	Frag.Acc.	Top-1	Top-5
0	33.9	21.3%	4.4%	8.9%
105	32.6	21.5%	3.3%	5.3%
223	34.1	22.7%	3.8%	6.9%
505	35.7	24.3%	5.0%	8.7%
580	**37.4**	**26.5%**	**7.4%**	**12.5%**

- Applying ROOSTERIZE without additional training achieves moderate performance (BLEU $= 33.9$)
- With some additional training, performance can be markedly improved (up to a BLEU score of 37.4 when training on 580 lemmas)

More Details in Our IJCAR Paper

- Implementation details of ROOSTERIZE toolchain
- Ablation study of more variants of ROOSTERIZE
- Evaluation results with different combinations of training, validation, and test sets
- `https://arxiv.org/abs/2006.16743`

Conclusions

- **Rooserize**: toolchain for learning and suggesting Coq lemma names, based on multi-input encoder-decoder neural networks
- **Kernel trees** provides important semantics context for lemma naming
- **Tree chopping** helps our models to effectively handle long inputs
- Evaluated on a **corpus** of 164k LOC high quality Coq code
- Case study shows ROOSTERIZE can provide useful suggestions in practice for a project outside our corpus

Tools and data:

- Roosterize: `https://github.com/EngineeringSoftware/roosterize`
- Corpus: `https://github.com/EngineeringSoftware/math-comp-corpus`
- VS Code: `https://github.com/EngineeringSoftware/roosterize-vscode`

Outlook

- growth of Coq ecosystem leads to better model training
- Coq serialization toolchain continuously improving
- applications in Coq code formatting, completion, ...

Pengyu Nie (pynie@utexas.edu)
Karl Palmskog (palmskog@kth.se)

TEXAS
The University of Texas at Austin

KTH
VETENSKAP OCH KONST

Backup Slides After This Point

Ablation Study: Copy Mechanism

Model	BLEU	Frag.Acc.	Top-1	Top-5
Stmt+ChopKnlTree+attn+copy	**47.2**	**24.9%**	**9.6%**	**18.0%**
Stmt+ChopKnlTree+attn	25.6	8.5%	0.9%	1.7%

- **Copy mechanism** improves performance by 22 points in BLEU
- Many sub-tokens are specific to the file context and do not appear in the fixed vocabulary of the training set