How to Leverage a Large Dataset of Formalized Mathematics with Machine Learning?

Dennis Müller[1] Michael Kohlhase[1] Florian Rabe[1,2]

Computer Science, FAU Erlangen-Nürnberg

LRI, Université Paris Sud

April 10, 2019

So, how?

I'm not here to answer this question.

I'm here to pose it.

And collaborate on finding an answer!

Background

To apply machine learning to a problem you need two things:

- Expertise in machine learning
- Huge sets of training data

We lack the expertise
but we have the data!

Training Data for ATP Applications

To train e.g. a neural network, you need huge data sets

The more the better

But: Most theorem prover libraries contain only $\approx 10^4$, maybe 10^5 declarations.

Furthermore, libraries in surface syntax are often

- Difficult to parse without access to the internals of the system
- Incomplete TCCs, implicit arguments, notational ambiguity...
- Specific to one system $\Rightarrow$ Results hardly reusable

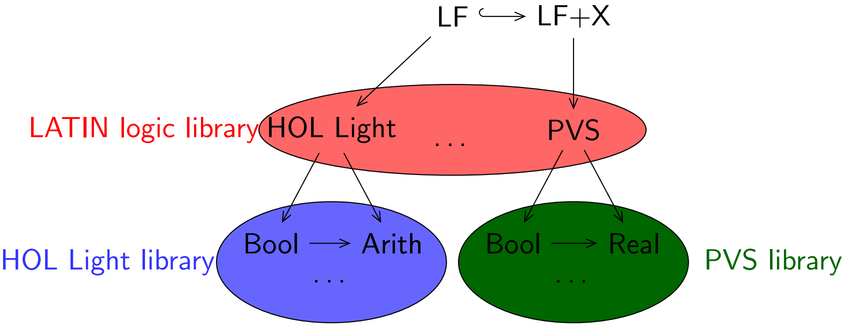
An Open Archive of Formalizations (OAF)

Represent math libraries in a universal framework:

- Use logical frameworks to represent Logics ⇒ Includes Type and Proof system
- Standardized XML Syntax (OMDoc) ⇒ Easily parsable
- High-Level API (MMT) ⇒ Allows generic services across systems

Imported libraries: Mizar, HOL Light, Isabelle, Coq, PVS, Sage, GAP, LMFDB, OEIS…

The OAF Methodology



Logical frameworks represented in MMT
Logics manually defined in a framework
Libraries imported from respective systems

MMT

A framework and Scala API for formal knowledge

allows integrating formal systems

- Parser
- type checking/inference — for any formal system
- Simplifier/Rewriter
- "Prover"

  very simple, but can e.g. be replaced by an external system
- Backend/Physical storage — e.g. resolves logical identifiers
- Knowledge Management Service

  Search, IDE, Refactoring, Web server...
- Flexible API and plugin architecture

`http://uniformal.github.io`

Available Libraries

System	Library	Modules	Declarations/Theorems
MMT	Math-in-the-Middle	183	826
Twelf	LATIN	529	2,824
PVS	Prelude	226	3,841
PVS	NASA	748	20,243
Isabelle	Distribution	2,308	484,419
Isabelle	AFP	7,245	987,861
HOL Light	Basic	189	22,830
IMPS	Library	64	8,573
Mizar	MML	1,194	69,710
Coq	49 Packages	1,979	383,500

Enough for Across-system machine learning applications?

`https://gl.mathhub.info`

Demo

Questions

- What services can we offer using ML?
- Which functions can we try to learn?
- How to vectorize our content?

We have students to do it and are happy to collaborate