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

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

[1] Computer Science, FAU Erlangen-Nürnberg
[2] LRI, Université Paris Sud

Abstract

Statistical machine learning techniques have proved very successful recently, including applications in logic. As logic has predominantly been based on exact symbolic methods, the question arises how to combine the strengths of the approaches.

We present MathHub, which aggregates formal libraries including those of most leading proof assistants. All these libraries are available in a standardized and easily machine-readable format, making it an ideal starting point for machine learning applications. Our contribution consists of posing the question given in the title, i.e., we do not provide an answer and instead hope discussions at the workshop will result in insights and collaborations towards future investigations.

Combinations of statistical and symbolic approaches to formal logic offer potential for groundbreaking innovations in artificial intelligence. However, a major impediment to large applications is that the currently most successful statistical methods are based on supervised machine learning and tend to require large sets of training data. And the most successful symbolic approaches to formalizing mathematical and related formal knowledge are based on interactive theorem proving, which requires human input for knowledge creation. Moreover, existing proof assistant are highly incompatible and do not allow easily merging their existing libraries into a single large one. Consequently, current applications have to focus on niches where sufficiently large datasets are in fact available. The most important such example is selecting axioms for reducing the search space of automated provers as in [KU15].

This was part of the motivation of the authors' MathHub project. MathHub collects libraries of mathematical knowledge in all forms. Its scope is not limited to logic, and includes also libraries of computation system, mathematical databases, and informal narrative texts. Technically, it is based on a GitHub-like repository management software with free access for researchers[1].

Its crucial and unique feature is the use of a single representation language for all knowledge, specifically the OMDoc language [RK13]. Thus, all libraries are not only available to be processed through third-party tools, but this processing can be done uniformly for all libraries. Moreover, mature software support is available for managing and reading MathHub repositories.

To make this possible, a huge effort is needed for each library, and we have done that for several major theorem provers, such as for Mizar in [Ian+13], HOL Light in [KR14], PVS in [Koh+17], IMPS [BK18] and very recently for Isabelle.[2] In these translations, great care has been taken to preserve — as much as possible — the original human-authored structure while also including the machine-inferred internal representation. Other MathHub libraries of interest to theorem proving are the LATIN logic library [Cod+11] and Math-in-the-Middle library currently built in the OpenDreamKit project. Figure 1 gives an overview of the sizes of these MathHub libraries. We expect many interesting sets of training data can be gleaned from these

[1] Available at `https://gl.mathhub.info`

[2] To be published. See `https://sketis.net/2018/isabelle-mmt-export-of-isabelle-theories-and-import-as-omdoc-content`.

System	**Library**	**Modules**	**Declarations/Theorems**
MMT	Math-in-the-Middle	183	826
Twelf	LATIN	529	2824
PVS	Prelude	226	3841
PVS	NASA	748	20243
Isabelle	Distribution	2308	484419
Isabelle	AFP	7245	987861
HOL Light	Basic	190	4707
IMPS	Library	64	8573
Mizar	MML	1194	69710

Figure 1: An Overview of the Available Archives on MathHub

and future MathHub libraries. However, transforming such libraries of formal declarations and expressions into the vectorized representations needed by standard machine learning algorithms is itself very difficult and an active research question. We do not offer a solution to this problem, but rather present and offer our library to the community with the hope of engaging in such experiments in future collaborations with machine learning experts.

References

[BK18] J. Betzendahl and M. Kohlhase. "Translating the IMPS theory library to OMDoc/MMT". In: *Intelligent Computer Mathematics (CICM) 2018*. Ed. by F. Rabe, W. Farmer, A. Youssef, and ... LNAI. in press. Springer, 2018. URL: `http://kwarc.info/kohlhase/papers/cicm18-imps.pdf`.

[Cod+11] M. Codescu, F. Horozal, M. Kohlhase, T. Mossakowski, and F. Rabe. "Project Abstract: Logic Atlas and Integrator (LATIN)". In: *Intelligent Computer Mathematics*. Ed. by J. Davenport, W. Farmer, F. Rabe, and J. Urban. Springer, 2011, pp. 289–291.

[Ian+13] M. Iancu, M. Kohlhase, F. Rabe, and J. Urban. "The Mizar Mathematical Library in OMDoc: Translation and Applications". In: *Journal of Automated Reasoning* 50.2 (2013), pp. 191–202.

[Koh+17] M. Kohlhase, D. Müller, S. Owre, and F. Rabe. "Making PVS Accessible to Generic Services by Interpretation in a Universal Format". In: *Interactive Theorem Proving*. Ed. by M. Ayala-Rincon and C. Munoz. Springer, 2017, pp. 319–335.

[KR14] C. Kaliszyk and F. Rabe. "Towards Knowledge Management for HOL Light". In: *Intelligent Computer Mathematics*. Ed. by S. Watt, J. Davenport, A. Sexton, P. Sojka, and J. Urban. Springer, 2014, pp. 357–372.

[KU15] C. Kaliszyk and J. Urban. "HOL(y)Hammer: Online ATP Service for HOL Light". In: *Mathematics in Computer Science* 9.1 (2015), pp. 5–22.

[RK13] F. Rabe and M. Kohlhase. "A Scalable Module System". In: *Information and Computation* 230.1 (2013), pp. 1–54.