MizAR 60 for Mizar 50

Jan Jakubův[1], Karel Chvalovský[1], Zarathustra Goertzel[1], Cezary Kaliszyk[2], Mirek Olšák[4], Bartosz Piotrowski[1], Stephan Schulz[3], Martin Suda[1], and Josef Urban[1]

[1] Czech Technical University in Prague, Prague, Czech Republic
[2] University of Innsbruck, Innsbruck, Austria and INDRC, Prague, Czech Republic
[3] DHBW Stuttgart, Stuttgart, Germany
[4] Institut des Hautes Études Scientifiques, Paris, France

Introduction: Mizar, MML, Hammers and AITP. As a present to Mizar [8] on its 50th anniversary, we develop an AI/TP system that automatically proves about 60 % of the Mizar theorems in the hammer setting. We also automatically prove 75 % of the Mizar theorems when the automated provers are helped by using only the premises used in the human-written Mizar proofs. We describe the methods and large-scale experiments leading to these results. This includes in particular the E [15, 16] and Vampire [13] provers, their ENIGMA [7, 10] and Deepire [17, 18] learning modifications, a number of learning-based premise selection methods, and the incremental loop that interleaves growing a corpus of millions of ATP proofs with training increasingly strong AI/TP systems on them.[1]

In recent years, methods that combine machine learning (ML), artificial intelligence (AI) and automated theorem proving (ATP) [14] have been considerably developed, primarily targeting large libraries of formal mathematics developed by the ITP community. This ranges from *premise selection* methods [1] and *hammer* [4] systems to developing and training learning-based *internal guidance* of ATP systems such as E and Vampire on the thousands to millions of problems extracted from the ITP libraries. Such large ITP corpora have further enabled research topics such as *automated strategy invention* [21] and *tactical guidance* [6], learning-based *conjecturing* [22], *autoformalization* [12, 24], and development of metasystems that combine learning and reasoning in various feedback loops [23].

Starting with the March 2003 release of the MPTP system [19] and the first ML/TP and hammer experiments over it [20], the Mizar Mathematical Library [2,3,8] (MML) and its subsets have as of 2023 been used for twenty years for this research, making it perhaps the oldest and most researched AI/TP resource in the last two decades.

Contributions. The last large *Mizar40* evaluation [11] of the AI/TP methods over MML was done almost ten years ago, on the occasion of 40 years of Mizar. Since then, a number of strong methods have been developed in areas such as premise selection and internal guidance of ATPs. In this work, we therefore evaluate these methods in a way that can be compared to the Mizar40 evaluation, providing an overall picture of how far the field has moved. Our main results are:

1. Over 75 % of the Mizar toplevel lemmas can today be proved by AI/TP systems when the premises for the proof can be selected from the library either by a human or a machine. This should be compared to 56 % in Mizar40 achieved on the same version of the MML. Over 200 examples of the automatically obtained proofs are analyzed on our web page.[2]

[1] The full paper, recently accepted to ITP'23, is available on arXiv [9].
[2] https://github.com/ai4reason/ATP_Proofs

2. 58.4 % of the Mizar toplevel lemmas can be proved today without any help from the users, i.e., in the large-theory (hammering) mode. This should be compared to about 40.6 % achieved on the same version of the MML in Mizar40. In both cases, this is done by a large portfolio of AI/TP methods which is limited to 420 s of CPU time.

3. Our strongest single AI/TP method alone now proves in 30 s 40 % of the lemmas in the hammering mode, i.e., reaching the same strength as the full 420 s portfolio in Mizar40.

4. Our strongest *single* AI/TP method now proves in 120 s 60 % of the toplevel lemmas in the human-premises (*bushy*) mode, i.e., outperforming the union of *all* methods developed in Mizar40 (56 %).

5. We show that our strongest method transfers to a significantly newer version of the MML which contains a lot of new terminology and lemmas. In particular, on the new 13 370 theorems coming from the new 242 articles in MML version 1382, our strongest method outperforms standard E prover by 58.2 %, while this is only 56.1 % on the Mizar40 version of the library where we do the training and experiments. This is thanks to our development and use of *anonymous* [10] logic-aware ML methods that learn only from the structure of mathematical problems. This is unusual in today's machine learning which is dominated by large language models that typically struggle on new terminology.

The central methods in this evaluation are internal guidance provided by the ENIGMA (and later also Deepire) system, and premise selection methods. We have also used several additional approaches such as many previously invented strategies and new methods for constructing their portfolios, efficient methods for large-scale training on millions of ATP proofs, methods that interleave multiple runs of ATPs with restarts on ML-based selection of the best inferred clauses (*leapfrogging*), and methods for minimizing the premises needed for the problems by decomposition into many ATP subproblems.

Conclusion: AI/TP Bet Completed. In 2014, after the 40 % numbers were obtained by Kaliszyk and Urban both on the Flyspeck and Mizar corpora, the last author publicly announced three AI/TP bets[3] in a talk at Institut Henri Poincare and offered to bet up to 10 000 EUR on them. Part of the second bet said that by 2024, 60 % of the MML and Flyspeck toplevel theorems will be provable automatically when using the same setting as in 2014. In the HOL setting, this was done as early as 2017/18 by the TacticToe system, which achieved 66.4 % on the HOL library in 60 s and 69 % in 120 s [5, 6]. One could however argue that TacticToe introduced a new kind of ML-guided tactical prover that considerably benefits from targeted, expert-written procedures tailored to the corpora. This in particular showed in the large boost on HOL problems that required induction, on which standard higher-order ATPs traditionally struggled.

In this work, we largely completed this part of the second AI/TP bet also for the Mizar library. The main caveat is our use of more modern hardware, in particular many ENIGMAs using the GPU server for clause evaluation. It is however clear (both from the LightGBM experiments and from the very efficient and CPU-based Deepire experiments) that this is not a major issue. While it is today typically easier to use dedicated hardware in ML-based experiments, there is also growing research in the extraction of faster predictors from those trained on GPUs that can run more efficiently on standard hardware.

[3] http://ai4reason.org/aichallenges.html

References

[1] Jesse Alama, Tom Heskes, Daniel Kühlwein, Evgeni Tsivtsivadze, and Josef Urban. Premise selection for mathematics by corpus analysis and kernel methods. *J. Autom. Reasoning*, 52(2):191–213, 2014.

[2] Grzegorz Bancerek, Czeslaw Bylinski, Adam Grabowski, Artur Kornilowicz, Roman Matuszewski, Adam Naumowicz, and Karol Pak. The role of the Mizar Mathematical Library for interactive proof development in Mizar. *J. Autom. Reason.*, 61(1-4):9–32, 2018.

[3] Grzegorz Bancerek, Czeslaw Bylinski, Adam Grabowski, Artur Kornilowicz, Roman Matuszewski, Adam Naumowicz, Karol Pak, and Josef Urban. Mizar: State-of-the-art and beyond. In Manfred Kerber, Jacques Carette, Cezary Kaliszyk, Florian Rabe, and Volker Sorge, editors, *Intelligent Computer Mathematics - International Conference, CICM 2015, Washington, DC, USA, July 13-17, 2015, Proceedings*, volume 9150 of *Lecture Notes in Computer Science*, pages 261–279. Springer, 2015.

[4] Jasmin Christian Blanchette, Cezary Kaliszyk, Lawrence C. Paulson, and Josef Urban. Hammering towards QED. *J. Formalized Reasoning*, 9(1):101–148, 2016.

[5] Thibault Gauthier, Cezary Kaliszyk, and Josef Urban. TacticToe: Learning to reason with HOL4 tactics. In Thomas Eiter and David Sands, editors, *LPAR-21, 21st International Conference on Logic for Programming, Artificial Intelligence and Reasoning, Maun, Botswana, May 7-12, 2017*, volume 46 of *EPiC*, pages 125–143. EasyChair, 2017.

[6] Thibault Gauthier, Cezary Kaliszyk, Josef Urban, Ramana Kumar, and Michael Norrish. Tactictoe: Learning to prove with tactics. *J. Autom. Reason.*, 65(2):257–286, 2021.

[7] Zarathustra Amadeus Goertzel, Karel Chvalovský, Jan Jakubuv, Miroslav Olsák, and Josef Urban. Fast and slow Enigmas and parental guidance. In *FroCoS*, volume 12941 of *Lecture Notes in Computer Science*, pages 173–191. Springer, 2021.

[8] Adam Grabowski, Artur Korniłowicz, and Adam Naumowicz. Mizar in a nutshell. *J. Formalized Reasoning*, 3(2):153–245, 2010.

[9] Jan Jakubuv, Karel Chvalovský, Zarathustra Amadeus Goertzel, Cezary Kaliszyk, Mirek Olsák, Bartosz Piotrowski, Stephan Schulz, Martin Suda, and Josef Urban. Mizar 60 for mizar 50. *CoRR*, abs/2303.06686, 2023.

[10] Jan Jakubuv, Karel Chvalovský, Miroslav Olsák, Bartosz Piotrowski, Martin Suda, and Josef Urban. ENIGMA anonymous: Symbol-independent inference guiding machine (system description). In *IJCAR (2)*, volume 12167 of *Lecture Notes in Computer Science*, pages 448–463. Springer, 2020.

[11] Cezary Kaliszyk and Josef Urban. MizAR 40 for Mizar 40. *J. Autom. Reasoning*, 55(3):245–256, 2015.

[12] Cezary Kaliszyk, Josef Urban, and Jirí Vyskocil. Automating formalization by statistical and semantic parsing of mathematics. In *ITP*, volume 10499 of *Lecture Notes in Computer Science*, pages 12–27. Springer, 2017.

[13] Laura Kovács and Andrei Voronkov. First-order theorem proving and Vampire. In Natasha Sharygina and Helmut Veith, editors, *CAV*, volume 8044 of *LNCS*, pages 1–35. Springer, 2013.

[14] John Alan Robinson and Andrei Voronkov, editors. *Handbook of Automated Reasoning (in 2 volumes)*. Elsevier and MIT Press, 2001.

[15] Stephan Schulz. System description: E 1.8. In Kenneth L. McMillan, Aart Middeldorp, and Andrei Voronkov, editors, *LPAR*, volume 8312 of *LNCS*, pages 735–743. Springer, 2013.

[16] Stephan Schulz, Simon Cruanes, and Petar Vukmirović. Faster, higher, stronger: E 2.3. In Pascal Fontaine, editor, *Proc. of the 27th CADE, Natal, Brasil*, number 11716 in LNAI, pages 495–507. Springer, 2019.

[17] Martin Suda. Improving ENIGMA-style clause selection while learning from history. In André Platzer and Geoff Sutcliffe, editors, *Automated Deduction - CADE 28 - 28th International Conference on Automated Deduction, Virtual Event, July 12-15, 2021, Proceedings*, volume 12699 of

Lecture Notes in Computer Science, pages 543–561. Springer, 2021.

[18] Martin Suda. Vampire with a brain is a good ITP hammer. In Boris Konev and Giles Reger, editors, *Frontiers of Combining Systems - 13th International Symposium, FroCoS 2021, Birmingham, UK, September 8-10, 2021, Proceedings*, volume 12941 of *Lecture Notes in Computer Science*, pages 192–209. Springer, 2021.

[19] J. Urban. Translating Mizar for First Order Theorem Provers. In A. Asperti, B. Buchberger, and J.H. Davenport, editors, *Proceedings of the 2nd International Conference on Mathematical Knowledge Management*, number 2594 in LNCS, pages 203–215. Springer, 2003.

[20] Josef Urban. MPTP – Motivation, Implementation, First Experiments. *J. Autom. Reasoning*, 33(3-4):319–339, 2004.

[21] Josef Urban. BliStr: The Blind Strategymaker. In Georg Gottlob, Geoff Sutcliffe, and Andrei Voronkov, editors, *Global Conference on Artificial Intelligence, GCAI 2015, Tbilisi, Georgia, October 16-19, 2015*, volume 36 of *EPiC Series in Computing*, pages 312–319. EasyChair, 2015.

[22] Josef Urban and Jan Jakubuv. First neural conjecturing datasets and experiments. In Christoph Benzmüller and Bruce R. Miller, editors, *Intelligent Computer Mathematics - 13th International Conference, CICM 2020, Bertinoro, Italy, July 26-31, 2020, Proceedings*, volume 12236 of *Lecture Notes in Computer Science*, pages 315–323. Springer, 2020.

[23] Josef Urban, Geoff Sutcliffe, Petr Pudlák, and Jiří Vyskočil. MaLARea SG1 – Machine Learner for Automated Reasoning with Semantic Guidance. In *IJCAR*, pages 441–456, 2008.

[24] Qingxiang Wang, Cezary Kaliszyk, and Josef Urban. First experiments with neural translation of informal to formal mathematics. In Florian Rabe, William M. Farmer, Grant O. Passmore, and Abdou Youssef, editors, *11th International Conference on Intelligent Computer Mathematics (CICM 2018)*, volume 11006 of *LNCS*, pages 255–270. Springer, 2018.