AI over Large Formal Knowledge Bases: The First Decade

Josef Urban

ICIS, RU Nijmegen

Abstract: In March 2003, the first version of the Mizar Problems for Theorem Proving (MPTP) was released. In the past ten years, such large formal knowledge bases have started to provide an interesting playground for combining deductive and inductive AI methods. The talk will discuss three related areas of application in which machine learning and general AI have been recently experimented with: (i) premise selection for theorem proving over large formal libraries built with systems like Mizar, HOL Light, and Isabelle (ii) advising and tuning first-order automated theorem provers such as E and leanCoP, and (iii) building larger inductive/deductive AI systems such as MaLARea. Here I focus on the wider motivation for this work.

1 Why AI over Large Formal Knowledge Bases

1.1 The Three Semantic Dreams

There are three powerful and old dreams that came with the invention of computers by Turing, von Neuman, and others. (i) The dream of general Artificial Intelligence, and in particular AI that helps with general scientific reasoning and research. (ii) The dream of Automated Reasoning, and in particular Automated Theorem Proving that could attack nontrivial mathematical theorems. And (iii) the QED dream, of all of mathematics (and perhaps also programming and all exact sciences backed by mathematics) being done in a form that is fully understandable to computers, verified for correctness by computers, and assisted in various semantic ways by computers.

People coming from various backgrounds may be more interested in a particular dream. For example, some ITP (Interactive Theorem Proving) people may welcome the use of ATPs if it helps the QED/verification progress. And they may also welcome the use of arbitrary AI methods, if it strengthens the automation. My basic motivation is the dream of scientific AI. I cannot imagine any other future than the one in which computers will to a very large extent semantically understand, assist, and develop not just mathematics and computer science, but also physics, chemistry, biology, medicine, engineering, economics, law, politics, etc. Here, *politics* might really seem a bit far-fetched. But no AI person smaller than John McCarthy dreamed exactly about that quite lucidly and publicly in connection with the QED dream (in which he was obviously involved too) already in 1999.[1] And with varying levels of detail, he was preceded by Leibniz and others.

1.2 How to Get There

There are two major obstacles to strong scientific AI, and I believe those two are related. One is to get computers fully *understand* all those large fields mentioned above, and the other one is to develop programs that can efficiently and usefully *reason* once the understanding has been provided.

It is not obvious that the two abilities imply one another. For example, in ITPs like Mizar, Isabelle, HOL (Light) and Coq, large corpora of mathematics are in some sense very well understood (all concepts are formally defined, all proofs explained and verified), but the reasoning power of those systems is on average far lower today than what trained humans can do. The people behind the QED Manifesto[2] were even afraid enough of falling again into what they perceived as the "AI/AR trap", and wrote: *It is the view of some of us that many people who could have easily contributed to project QED have been distracted away by the enticing lure of AI or AR.* And indeed, following the QED workshops challenge, the Mizar team formalized over 50% of the Compendium of Continuous Lattices textbook already by 2002, demonstrating that advanced verification is possible without strong reasoning. The same lack of strong reasoning holds for more recent projects, like Flyspeck and the formal proof of the Feit-Thompson theorem. These are impressive results, which have however been obtained with a lot of effort. The formal proofs are still often much more detailed than standard mathematics, and many shortcuts that mathematicians naturally invent and use have to be manually translated into the formal languages.

I believe that good understanding involves combination of knowledge, reasoning and pattern discovery. The larger our initial knowledge, and the stronger our capability to find analogies and to make useful inductive/deductive reasoning chains about possible explanations and disambiguations, the stronger our understanding. Understanding is also a dynamic process: in mathematics one can understand various phenomena using various insights, and eventually realize that some theorems and ideas are very easy when viewed through the right conceptual framework. Finding such conceptual frameworks is however the hard part, which involves both deductive and inductive reasoning.

So to get to strong AI for sciences, it should be good to study useful ways of deductive and inductive reasoning (and understanding) over large corpora of advanced human knowledge and reasoning that are as deeply accessible to computers as possible. In my opinion, nothing has so far beaten large corpora of formal (but largely human-written,

[1] http://www-formal.stanford.edu/jmc/future/objectivity.html

[2] http://www.rbjones.com/rbjpub/logic/qedres00.htm

and often aligned with informal texts) mathematical knowledge as a suitable resource combining these aspects, pushing the deductive/inductive reasoning research from small toy domains to much more realistic domains. If reasoning and understanding capabilities can be developed for mathematics to the level where full LaTeX-written books will be automatically understood, verified, and digested into strong AI systems that solve nontrivial mathematical problems, then a lot of motivation will be provided for similar semantic treatment of other sciences. In this sense, the dream of strong ATP and the QED dream are included in the AI-for-science dream. But as suggested above, I also think that the ATP and QED dreams can considerably profit from complementary AI methods like machine learning and pattern discovery applied to the large computer-understandable corpora of human thinking. At least when reasoning over large domains, such complementary methods already help a lot.

2 The First Decade

The last decade opened three large corpora to such experiments with automated theorem proving and related AI methods: the Mizar Mathematical Library (MML), the Isabelle/HOL library (and all developments in the Archive of Formal Proofs – AFP – based on Isabelle/HOL), and the HOL Light/Flyspeck libraries. This translation work continues, in some sense from both sides: better translation methods from the more expressive logics to first-order logic are very useful, while the ATP calculi and systems are equipped with more and more features that take them closer to the expressive logics (types, higher-order ATPs, SMTs).

A considerable amount of work has been done since 2003 on better ATP data structures and large-theory algorithms inside ATPs. The only way to do some ATP experiments was initially however only through external knowledge selection methods, either heuristic or learning. In the evaluation [7] done in 2010, only 2% of the large MML problems could be solved by unaided E, 6% could be solved by E using the freshly developed heuristic SInE filtering, and 15% when using the most restrictive SInE filtering. On the smaller MPTP2078 benchmark used for the Mizar@Turing 2012 CASC competition, machine learning from previous proofs further improved on SInE by 44% when using naive Bayes, and by 50% when using kernel-based learning [1]. This has been further improved by 10% by a simple ensemble approach combining SInE as a ranker with the kernel-based ranking [4]. Exploring a number of the most complementary learning, feature-characterization, and ATP methods has led to development of strong "hammer" systems for Isabelle [5, 3], Mizar [8] and HOL Light [2]. 39% of all top-level Flyspeck theorems (about 14.000) can be today proved in a push-button mode.

The AI-based methods that have already been implemented and tried inside ATPs include various guidance mechanisms in E, in particular goal-oriented heuristics that prefer clauses that are in various ways close to the conjecture. Similar mechanisms exist also in Prover9, however they have been so far used rather for proving deep results in small algebraic theories. A recent experiment with full learning-based guidance is the MaLeCoP system [10], which learns guidance from any closed tableaux branch, and applies the guidance practically at every inference point, pruning the average search about twenty times.

Such experiments easily lead to larger deductive/inductive AI metasystems. E.g., learning premise selection can be interleaved with proving, resulting in the MaLARea-style [9] loop, where the premise selection benefits from more proofs and counterexamples, and the proving benefits from better premise selection. Such AI loops can be augmented by introduction of suitable concepts, lemmas, and conjectures. ATPs may self-improve for long times in strategy-evolving loops [6], growing new strategies by genetic or iterative methods, focusing on specific types of problems, and eventually optimizing a global strategy scheduling meta-system as done in Vampire and E-MaLeS (again using learning). A new kind of an AI metasystem that should be attempted soon will use large aligned informal/formal corpora such as Flyspeck, and develop AI loops that gradually improve the understanding (translation) capability between the informal and formal texts, helping the understanding by the AI/ATP methods simultaneously trained on the already understood part.

References

[1] Jesse Alama, Tom Heskes, Daniel Kühlwein, Evgeni Tsivtsivadze, and Josef Urban. Premise Selection for Mathematics by Corpus Analysis and Kernel Methods. *CoRR*, abs/1108.3446, 2011.

[2] Cezary Kaliszyk and Josef Urban. Learning-assisted automated reasoning with Flyspeck. *CoRR*, abs/1211.7012, 2012.

[3] Daniel Kühlwein, Jasmin Christian Blanchette, Cezary Kaliszyk, and Josef Urban. MaSh: Machine Learning for Sledgehammer. 2013. Accepted to ITP 2013.

[4] Daniel Kühlwein, Twan van Laarhoven, Evgeni Tsivtsivadze, Josef Urban, and Tom Heskes. Overview and evaluation of premise selection techniques for large theory mathematics. In *IJCAR*, pages 378–392, 2012.

[5] Jia Meng and Lawrence C. Paulson. Translating higher-order clauses to first-order clauses. *J. Autom. Reasoning*, 40(1):35–60, 2008.

[6] Josef Urban. BliStr: The Blind Strategymaker. *CoRR*, abs/1301.2683, 2013.

[7] Josef Urban, Krystof Hoder, and Andrei Voronkov. Evaluation of automated theorem proving on the Mizar Mathematical Library. In *ICMS*, pages 155–166, 2010.

[8] Josef Urban, Piotr Rudnicki, and Geoff Sutcliffe. ATP and presentation service for Mizar formalizations. *J. Autom. Reasoning*, 50:229–241, 2013.

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

[10] Josef Urban, Jirí Vyskocil, and Petr Stepánek. MaLeCoP Machine Learning Connection Prover. In *TABLEAUX*, pages 263–277, 2011.