Automating Formalization of Mathematics with Machine Learning and Data Mining: A PhD Progress Report*

Qingxiang, Wang[1,2]

University of Innsbruck Czech Technical University in Prague

The progress in foundations of mathematics since the turn of the 20th century has enabled the creation of a precise and unified formal language for mathematics to be based on. Although never written down in full, it is a consensus among mathematicians that ultimate certitude of their mathematics can be achieved by using the language of set theory or any alternatives. It is a vision set out by the QED Manifesto [1] that all the existing mathematics can be formalized and their proofs formally verified.

Since 1960s various interactive proof assistants have been developed. Each has its own formal library covering a portion of mathematics. Because of difference in mathematical foundations and idiosyncratic design features, those libraries are not compatible with each other. As a result the formalization community has been segregated into different groups based on proof assistant. Meanwhile, formalization is still a task that requires considerable amount of intellectual effort: to be a qualified formalizer, one needs to be an expert in the proof assistant being used, its underlying mathematical foundations as well as the mathematical theorems being formalized. This makes the mainstream mathematics community shunning formalization of mathematics.

As mathematics has become unmanageably vast, it is now a pressing issue for the speed of formalization of mathematics to catch up with the speed of mathematics development. However, due to the reasons above formalization of mathematics is still largely done manually, requiring dedicated skill sets and without the participation of the mainstream mathematics community. To realize the dream of the QED Manifesto, it is inevitable that some form of automation should be involved in the process of formalization.

In recent years deep learning has gained a lot of attention in both industry and academia. With large high-quality datasets, deep learning models are able to generate translators that show competitive results comparing to human translators. As formalization can be considered as a process of translating from one language (the informal natural language used in mathematics) to another (the formal language used by a proof assistant), it is tempting to think that machine learning could be useful to automate formalization of mathematics. It is under this line of thought that the author started this 4-year PhD program 18 months ago. Till now we have conducted several experiments along this direction:

1. *Supervised neural machine translation on synthetic LaTeX-Mizar dataset.* Using Luong et al's sequence-to-sequence model [2] we were able to quickly test out its efficacy on translating from informal LaTeX sentences to formal sentences. To cope with the lack of aligned dataset, we adopted a deterministic back-translation tool [3] to generate readable LaTeX sentences from corresponding Mizar sentences. We have thoroughly experimented with different network parameters. The results have been published in last year's CICM [4].

2. *Unsupervised neural machine translation model on different datasets.* We digested and adapted Lample et al's unsupervised neural machine translation model [5] to the theorem-level aligned ProofWiki-Mizar dataset as well as our previous synthetic LaTeX-Mizar dataset. For the synthetic dataset we found that even when the LaTeX sentences and the Mizar sentences are not aligned, for short sentences, unsupervised model still has a chance to generate correct Mizar translations given LaTeX input.

3. *Data augmentation with type-checking on supervised model.* We keep usingg Luong's supervised model but

In: C. Kaliszyk, E. Brady, J. Davenport, W.M. Farmer, A. Kohlhase, M. Kohlhase, D. Müller, K. Pąk, and C. Sacerdoti Coen (eds.): Joint Proceedings of the FMM and LML Workshops, Doctoral Program and Work in Progress at the Conference on Intelligent Computer Mathematics 2019 co-located with the 12th Conference on Intelligent Computer Mathematics (CICM 2019), Prague, Czech Republic, July 8–12, 2019, published at http://ceur-ws.org

the training data we provide contains aligned Mizar-TPTP sentence pairs. A type-checking utility [1] takes the translated TPTP sentences and conducts an "elaboration" process which tries to fill in the missing Mizar type information. A completely elaborated TPTP sentence can be translated deterministically back to Mizar sentence. We divide our training phase into iterations. Each iteration generates multiple TPTP translations from each of Mizar sentences. The elaborated and back-translated new sentence pairs will be added to the training dataset for the next iteration. In this experiment we found that translation quality can be improved in initial three iterations, but performance stabilizes in later iterations.

4. *The ArXiv dataset has been collected.* Several tools and techniques have been employed, including LaTeXML, GNU Parallel and Apache Spark, to extract more than one-hundred billion sentences out of 29 years of submissions in mathematics. This provides us a sufficiently large amount of informal data to conduct further experiments.

In summary, the first supervised learning experiment is by far the most successful experiments we have conducted. It proves that given sufficient amount of informal-formal sentence pairs as training data, a neural network is able to generate nearly accurate formal translations from informal natural language statements, thereby having the potential to speed up the process of formalization. The unsupervised model also shows promising results, though further experiments and customizations need to be done to evaluate its overall potential. The initial gains of data augmentation using type-checking is noticeable and there is also potential to incorporate this mechanism into the LaTeX to Mizar translation. In addition to those we have collected abundant informal mathematics statements from ArXiv and will seek further ways to eventually bootstrap a formalizer that has the potential to extract a significant amount of logical information out of informal mathematical literature.

Future works will be focused on the following directions:

1. Apply natural language processing on the ArXiv corpus to iteratively normalize the informal mathematics data. Use techniques from descriptive and inferential statistics to extract inherit logic information from syntactically parsed natural language sentences.

2. Further explore the unsupervised model and enrich the informal corpus with the ArXiv data.

References

[1] The qed manifesto. In *Proceedings of the 12th International Conference on Automated Deduction*, CADE-12, pages 238–251, London, UK, UK, 1994. Springer-Verlag.

[2] Minh-Thang Luong, Eugene Brevdo, and Rui Zhao. Neural machine translation (seq2seq) tutorial. *https://github.com/tensorflow/nmt*, 2017.

[3] Grzegorz Bancerek. Automatic translation in formalized mathematics. *Mechanized Mathematics and Its Applications*, 5(2):19–31, 2006.

[4] 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.

[5] Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

[1] Unpublished work by Chad E. Brown at CTU in Prague in 2017