# Learning alignment between formal & informal mathematics

Kshitij Bansal[1] and Christian Szegedy[2]

[1] Google Research
kbk@google.com
[2] Google Research
szegedy@google.com

## 1 Introduction

In this talk, we explore the possibility of training an alignment model between informal and formal mathematical corpora in a semi-supervised manner. Though there is a lot of informal mathematics available in natural language (textbooks, papers), the fully formalized and computer checked mathematical content is limited. Availability of alignment information between the two is even further limited. That said, an alignment model between formal and informal mathematics would be essential for the task of autoformalization [3] and could result in dramatically growing the corpus of formalized mathematics. This could open up the possibility for an open-endedly improving system by training proof-guidance and alignment models in lockstep. We look into the currently available resources for bootstrapping such a system, and share our findings.

## 2 Learning an alignment model

Unsupervised (and weakly-supervised) neural approaches to machine translation relying on learning semantic representations for languages and an alignment model between them have shown great promise (e.g. [4]). We look into various aspects from the point of view of incorporating such ideas for learning an alignment model between formal and informal mathematics.

One of the key aspects is learning semantic representations from large unstructured corpora in a self-supervised manner. On the natural language side (generally, not specifically for mathematics) this is a well-studied area with a lot of progress over the past few years alone [2,6,8]. In general, research has established that training current deep neural network based models on proxy-tasks for natural language modeling can be fine-tuned to several downstream tasks such as machine translation, semantic search, sentiment analysis and question answering. Moreover, these tasks did not need as large amounts of data, yet yielded significant gains. For mathematics, on the informal side, there is also significant semantic information in the formulas, equations, diagrams, etc. which would be crucial to leverage for autoformalization work. The availability of large (unlabeled) corpora of informal mathemetics is not necessarily an issue, even if work is required for collecting such datasets for our puprose.

Perhaps less systematically explored and established, nevertheless, various works on theorem proving using neural approaches have looked into learning semantic representations on the formal side. Examples include tasks such as predicting the relevance of premises for proving a statement [1], predicting latent representations of rewrites [5], and labeling a formula with symbols using its structure alone [7]. One can argue that formal mathematical content is even more amenable to unsupervised pretraining as there is a larger number of conceivable self-supervised tasks than in the case of natural language processing. For that, we can leverage the well-defined graph structure of formulas and the ability to systematically transform them (using, say, rewrite rules and substitutions).

Given the success of unsupervised pretraining on the natural language side and encouraging initial results of semantic embeddings of formal mathematical content, the main task that remains is to train an alignment model between the two sets of embeddings. One key idea is to use cycle consistency [10]. We are especially inspired by its use for learning machine translation models on non-aligned corpora [4]. We propose a similar approach in conjunction with requiring that the translations should utilize similar notions. We explore models that translate natural language text to formal mathematical content (in HOL Light) and vice versa, with several constraints: after back and forth translation the embedding of the resulting statement should stay close to the input in the embedding space; put a loss on the network to enforce that the set of notions referred to by the two translations contain similar notions; and we maximize the probability of the translated sentence looking natural (or being a valid formal sentence). Using these constraints (that is, a combination of the associated losses) we have trained sequence-to-sequence models based on the transformer network [9] with end-to-end backpropagation.

To summarize, using corpora derived from formalization efforts in HOL Light proof assistant on the formal side, we will dicuss the different aspects of the approach:

- sources of datasets,
- language models for informal mathematics including formulas/equations,
- semantic embedding models and discussion of training tasks for formal mathemetics,
- training translation models with the various (cycle consistency, notion-similarity and naturality) requirements,
- neural network architecture choices and
- qualitative evaluation of our first alignment and translation models.

## References

[1] Alex A Alemi, Francois Chollet, Niklas Een, Geoffrey Irving, Christian Szegedy, and Josef Urban. Deepmath-deep sequence models for premise selection. *arXiv preprint arXiv:1606.04442*, 2016.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[3] Cezary Kaliszyk, Josef Urban, and Jiří Vyskočil. Automating formalization by statistical and semantic parsing of mathematics. In *International Conference on Interactive Theorem Proving*, pages 12–27. Springer, 2017.

[4] Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*, 2017.

[5] Dennis Lee, Christian Szegedy, Markus N Rabe, Sarah M Loos, and Kshitij Bansal. Mathematical reasoning in latent space. *arXiv preprint arXiv:1909.11851*, 2019.

[6] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[7] Miroslav Olšák, Cezary Kaliszyk, and Josef Urban. Property invariant embedding for automated reasoning. *arXiv preprint arXiv:1911.12073*, 2019.

[8] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[10] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.