Developing Corpus-based Translation Methods between Informal and Formal Mathematics

Cezary Kaliszyk Josef Urban Jiří Vyskočil Herman Geuvers

University of Innsbruck, Austria

Radboud University, Nijmegen

Czech Technical University

10 July 2014, Coimbra, CICM

Two Obstacles to Strong Computer Support for Math

There are two major obstacles preventing strong computer support for math and sciences:

- Low reasoning power of automated reasoning methods, particularly over large complex theories
- Lack of computer understanding of current human-level (math and exact science) knowledge

The two are related: human-level math may require nontrivial reasoning to become fully explained

Fully Computer-Understandable Math and Its History

- The Dream of Formal Math is far older than Internet, Wikipedia (2000) and even TEX(1978)
- SAM: 1965, Automath: 1968, Mizar: 1973
- some 200 articles in the formal Mizar library already in 1991

Formal Corpora in 2014

- The Mizar Mathematical Library: some 60,000 theorems (most of them rather small lemmas), 10,000 definitions
- HOL Light and Flyspeck: some 22,000 theorems
- Isabelle and the Archive of Formal Proofs: some 50,000 theorems
- Coq: several large projects (Feit-Thompson theorem)

Informal and Semiformal Corpora in 2014

- Arxiv.org: 1M articles collected over some 20 years (not just math)
- Wikipedia: 25,000 articles in 2010 - collected over 10 years only
- Semiformal and informal corpora have grown one or two orders of magnitude faster than formal ones
- We should use this energy

Attempts at auto-formalization

- Claus Zinn and others:
- manual translators from latex to formal math, failing for several reasons:
- lack of the vast background knowledge that the mathematicians use for gap-filling
- lack of decent automated reasoning methods over such vast corpora of math knowledge
- lack of translation methods that can automatically adapt to large corpora, using automated self-improvement

But this has been changing in the last decade!

- we started to have reasonably big formal corpora of common math
- we have developed reasonably strong automated reasoning methods over them
- and a large part of the latter was thanks to learning methods (40% of Mizar theorems automatically provable today)
- and we are even getting some aligned informal/formal corpora: Flyspeck, Compendium of Continuous Lattices, Feith-Thompson
- so let's use what works: statistical machine translation combined with strong learning-assisted automated reasoning over large libraries providing the common background!

What have we done so far

- Mathifier at RU Nijmegen: 75% of disambiguations can be guessed using very simple statistical methods
- Extracted 596 formulas from the Flyspeck book using LaTeXML currently parsin and typing 17% with very little effort
- All formal HOL Light/Flyspeck formulas exported into Lisp and Prolog formats, experiments with parsing them without knowing the HOL Light's parsing conventions and with forgetting some casting functors (using the Stanford parser and a custom CYK (charter) parser)
- Combination of LaTeX with natural language in the semi-formal corpora. Opaquified each proof sentence like this `Let MyTrmOrFla be MyRef of MyTrmOrFla.` The first 100 most frequent opaque patterns cover already half of all 42,931 ProofWiki sentences.

We need aligned or almost-formal corpora!

- If you know of any, tell us!