Developing Corpus-based Translation Methods between Informal and Formal Mathematics

Cezary Kaliszyk (supported by FWF grant P26201), Josef Urban, Jiří Vyskočil, Herman Geuvers

universität innsbruck

CZECH TECHNICAL UNIVERSITY IN PRAGUE

Radboud University Nijmegen

Goal

- Learn the translation of informal math to formal from aligned informal/semiformal/formal corpora
- Complement statistical approaches by strong ATP systems (*hammers*) for large formal libraries
- Complement human annotation by feedback loops between translation guessing and theorem proving

Example: ProofWiki vs Mizar vs Mizar-style automated proof

== Theorem ==
Let $(S, \circ)$ be an [[Definition:Algebraic Structure|algebraic structure]] that has a [[Definition:Zero Element|zero element]] $z \in S$. Then z is unique.
== Proof ==
Suppose z_1 and z_2 are both zeroes of $(S, \circ)$.
Then by the definition of [[Definition:Zero Element|zero element]]:
$z_2 \circ z_1 = z_1$ by dint of z_1 being a zero;
$z_2 \circ z_1 = z_2$ by dint of z_2 being a zero.
So $z_1 = z_2 \circ z_1 = z_2$.
So $z_1 = z_2$ and there is only one zero after all.
{{qed}}
// *NB: Informal proofs are buggy!*

```
Th9:  e1 is_a_left_unity_wrt o &
e2 is_a_right_unity_wrt o implies e1 = e2
proof
assume that A1:  e1 is_a_left_unity_wrt o and
A2:  e2 is_a_right_unity_wrt o;
thus e1 = o.(e1,e2) by A2,Def6 .= e2 by A1,Def5;
end;

z1 is_a_unity_wrt o & z2 is_a_unity_wrt o
implies z1 = z2 proof
assume that A1:  z1 is_a_unity_wrt o and
A2:  z2 is_a_unity_wrt o;
A3:  o.(z2,z1) = z1 by Th3,A2; ::[ATP]
A4:  o.(z2,z1) = z2 by Def 6,Def 7,A1,A3; ::[ATP]
hence z1 = z2 by Th9,A1,Def 7,A2; ::[ATP]
end;
```

Formal Corpora

- **Mizar** and the Mizar Mathematical Library (MML): ca 60,000 theorems, designed by a mathematician and a linguist to resemble math papers/books, human-readable formulas and natural deduction proof style, possible disadvantage (or at least research topic): complex overloading/disambiguation mechanisms
- **HOL Light** and Flyspeck: 22,000 theorems, less human-oriented but easy to work with, a lot of analysis and topology recently covered (Multivariate), quite minimal overloading mechanisms, proofs unreadable but recent attempts at their transformation
- **Isabelle** and the Archive of Formal Proofs (AFP): ca 50,000 theorems, somewhere between HOL (Light) and Mizar in the presentation/proof style, AFP not as organized as MML but growing quite fast (probability, etc)
- **Coq** and Feit-Thompson: large piece of advanced math, overloading/disambiguation mechanisms similar to Mizar, the logic quite non-standard (*proof programming*), semi-declarative proofs, no strong *hammers* yet

(Info|Semi)Formal Corpora

- **Flyspeck book** by Hales: *Dense Sphere Packings A Blueprint for Formal Proofs*, about 250-400 theorems mapped and over 200 concepts aligned by Hales with the HOL Light formalization (a simple wiki for extending this)
- the book **Compendium of Continuous Lattices** (CCL) by Gierz et all.: about 60% of the book formalized in Mizar, some high-level alignment of concepts and theorems available
- parts of Engelking's **General Topology** mapped to Mizar by Bancerek
- high-level mapping of the two graduate books leading to the **Feit-Thompson theorem** by Gonthier
- **ProofWiki**: very detailed Mizar-style (but informal) proofs and symbol linking, quite a lot of general topology – probably can be aligned with some Mizar theorems/proofs
- **PlanetMath**: similar to ProofWiki, not so detailed and unified proof style
- **Wikipedia**: a lot of mathematics, again not so unified and detailed proof style as ProofWiki

Relevant NLP and AI/ATP methods and tools

- **Machine translation**: algorithms that directly translate between two languages, statistical machine translation recently very successful - needs large aligned corpora (n-grams etc.)
- **Word-sense disambiguation**: algorithms that determine the exact meaning of (sets of) words in sentences: the Wikifier system (Dan Roth@UIUC): infer exact disambiguations (meanings) by statistical methods from context - a kind of *probabilistic semantic parsing* (somewhat similar to contextual disambiguation of overloading in Mizar and Coq)
- Part-of-speech tagging, phrasal and dependency parsing (Stanford parser, MBT, Moses, ...)
- However, none of the existing NLP research is grounded in complete semantics like formal mathematics has (one cannot really try to prove things and learn from the successes)
- How we differ from previous informal2formal attempts:
  - Large-theory automated reasoning became a reality in the last decade: we have large libraries covering common math facts together with quite strong gap-filling reasoning methods over them
  - We do not intend to design the translations (completely) by hand, in the same way as we do not design the strongest large-theory reasoning methods by hand – this does not scale to large corpora
  - Instead, we want learn the translations from the aligned corpora and let them self-improve by feedback loops between the guessing and deductive confirmation (this is also how the strongest large-theory ATP methods work today)

First Experiments

- The Mathifier student project at RU: we have used the simplest statistical disambiguation method on the semi-formal PlanetMath and ProofWiki corpora and obtained 75% success rate. This means that even very simple statistical methods can be usable.
- We have extracted 596 formulas from the Flyspeck book using LaTeXML and tried their simple translation (a small table of most common meanings) to HOL Light, i.e., parsing into preterms and their type-checking (Hindley-Milner). Currently 17% success, growing with each new symbol in the table.
- We have exported all formal HOL Light/Flyspeck formulas into Lisp and Prolog formats on which we do experiments with parsing them without knowing the HOL Light's parsing conventions and with forgetting some casting functors. For this we so far use a custom implementation of the CYK charter parser, and we also experiment with the Stanford parser.
- We are also looking at the combination of LaTeX with natural language in the semi-formal corpora. We have opaquified each proof sentence by replacing LaTeX math with `MyTrmOrFla` and replacing links with `MyRef`. The first 100 most frequent opaque patterns cover already half of all 42,931 ProofWiki sentences. The most frequent ones are as follows (mapping easily to Mizar-style constructs):

```
5829 Let MyTrmOrFla be MyRef.
2688 Let MyTrmOrFla.
774 Then MyTrmOrFla is MyRef.
736 Let MyTrmOrFla be MyRef of MyTrmOrFla.
724 Let MyTrmOrFla and MyTrmOrFla be MyRef.
578 Let MyTrmOrFla be the MyRef of MyTrmOrFla.
555 Let MyTrmOrFla be the MyRef.
```