Disambiguating Symbolic Expressions in Informal Documents

Dennis Müller[1] Cezary Kaliszyk[2]

Computer Science, FAU Erlangen-Nürnberg

Computational Logic, Universität Innsbruck

March 25, 2021

Disambiguation

Disambiguation: Constructing the abstract syntax tree of a symbolic expression and associating each symbol with its precise *semantics*.

⇒ Meaning of an expression becomes unambiguous

What does $a^2 + b^2 = c^2$ mean?

$\cdot^2$: squaring or upper indices of a sequence $(a^i)_{i \in I}$?

$+$: Addition of numbers? What number space? Arbitrary monoid/group/ring/field/vector space? List/string concatenation?

a, b, c: Constants? Variables? Ranging over which space? Related to +

$=$: What kind of equality? Up to isomorphism? Syntactic equality?

- $a^2 + (b^2 = c^2)$?

sTEX

A LATEX-package for (among other things) writing symbolic expressions in a disambiguated manner [3]:

LATEX

```
Multiplication on natural numbers
is defined via $x\cdot0=0$ and...
```

sTEX

```
Multiplication on natural numbers
is defined via $\eq{
  \nattimes{x}{0}}{0}$ and...
```

Both yield:

"Multiplication on natural numbers is defined via $x \cdot 0 = 0$ *and..."*

Task: Translate LATEX to sTEX

sTEX can be translated to OMDoc/OpenMath [2] and imported by the MMTsystem [8, 7]

Datasets

For training, we need a parallel dataset.
Note: sTEX⇒ LATEX is easy, so we only need sTEX datasets. (just macro expansion)

Available sTEX Datasets:

- **SMGloM** [4]: *Semantic Multilingual Glossary of Mathematics*
  - Dictionary style entries Mostly definitions, few theorems, no proofs

  ⇒ (introduces, and hence) covers many mathematical symbols
  But: few symbols referenced more than once
- **MiKoMH**: CS Lecture notes by Michael Kohlhase Author of sTEX

  ⇒ uses only few symbols in the SMGloM, almost no (higher, pure) mathematics.

All documents split into ≈500 character *sentences* and expanded to plain LATEXfor a parallel dataset.
⇒ Small, heavily biased dataset.

Synthesizing sTEX Sentences

We use MMT to synthesize additional data:

- Align sTeX symbols with symbols in a strongly typed formal library [6, 5]

  The *Math-in-the-Middle (MitM)* library
- Generate well typed MitM-expressions with free variables

  ⇒ syntactically well-formed
- Translate generated expressions to sTeX and *verbalize* free variables and their types

Example:

```
Whenever we have some positive natural number $\varepsilon$, any integer
$\ell$ and a real number $\livar{\mathcal C}{2}$, then it follows that
$\realtimes{\livar{\mathcal C}{2},\livar{\mathcal C}{2},\realplus{
\realuminus{\ell},\natsucc{\varepsilon}}}$.
```

Parallel Dataset

In total:

	SMGloM	MiKoMH	Synthesized
# Sentences:	911	9200	23,000

$\Rightarrow \approx$ 33,000 sentences.
Additionally, we extract symbolic expressions in both sTeX and LaTeX, yielding quadruples $(S_{\text{s\TeX}}, S_{\text{\LaTeX}}, (m_{\text{s\TeX},i})_{i\leq n}, (m_{\text{\LaTeX}})_{i\leq n_S})$

Evaluation set written by hand (both sTeX and LaTeX, 161 symbolic expressions).

Task-specific Peculiarities

Neural Machine Translation (NMT) has been proven to be a successful approach in *autoformalization* (e.g. [1, 11, 10]).

Our translation task has unique properties and challenges:

1. Only a small, biased dataset.
2. **But** translation is the identity everywhere except for symbolic expressions.
3. **But also** document context required for disambiguation
4. Domain and target language (i.e. plain LaTeX and sTeX) share a huge amount of syntax and structure

Basic latex macro syntax

All natural language grammar + semantics

All required context in 3. is shared

⇒ We can exploit 2. and 4.

Our Approach

Dataset too small for off-the-shelf NMT models

⇒ Pretrain a GPT-2 language model [9] on existing LaTeX corpora

obtained from arXiv.org

- Finetuned on inputs of the form

$$S_{\LaTeX}\ \texttt{<s>}\ m_{\LaTeX,i}\ \texttt{<s>}\ m_{\mathrm{sTeX},i}\ \texttt{<s>}$$

e.g.

```
Multiplication on natural numbers is defined via $x\cdot 0=0$ and
            ...<s>$x\cdot 0=0$<s>$\eq{\nattimes{x,0}}{0}$<s>
```

- For translation we use text generation on inputs

$$S_{\LaTeX}\ \texttt{<s>}\ m_{\LaTeX,i}\ \texttt{<s>}$$

Evaluation and Results

We use MMT integration for evaluation.

Of the results:

- 96.9% are syntactically valid LaTeX.
- 64% are syntactically equal to the input after expanding sTeX macros.
  ⇒ preserve presentation
- 60.2% are disambiguated. use sTeX macros everywhere
- 47.2% are string-equal to the expected labels. ⇒ correctly disambiguated
- 59.6% can be type checked. after translation to MitM ⇒ well-typed

References I

C. Kaliszyk, J. Urban, and J. Vyskočil.
System description: Statistical parsing of informalized Mizar formulas.
In T. Jebelean, V. Negru, D. Petcu, D. Zaharie, T. Ida, and S. M. Watt, editors, *19th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, SYNASC 2017*, pages 169–172. IEEE Computer Society, 2017.

M. Kohlhase.
OMDoc: An Open Markup Format for Mathematical Documents (Version 1.2).
Number 4180 in Lecture Notes in Artificial Intelligence. Springer, 2006.

M. Kohlhase.
Using LaTeX as a Semantic Markup Format.
Mathematics in Computer Science, 2(2):279–304, 2008.

M. Kohlhase.
A data model and encoding for a semantic, multilingual terminology of mathematics.
In S. Watt, J. Davenport, A. Sexton, P. Sojka, and J. Urban, editors, *Intelligent Computer Mathematics 2014*, number 8543 in LNCS, pages 169–183. Springer, 2014.

D. Müller.
Mathematical Knowledge Management Across Formal Libraries.
PhD thesis, Informatics, FAU Erlangen-Nürnberg, 10 2019.

D. Müller, C. Rothgang, Y. Liu, and F. Rabe.
Alignment-based Translations Across Formal Systems Using Interface Theories.
In C. Dubois and B. Woltzenlogel Paleo, editors, *Proof eXchange for Theorem Proving*, pages 77–93. Open Publishing Association, 2017.

[II

References]References

F. Rabe.
How to Identify, Translate, and Combine Logics?
Journal of Logic and Computation, 27(6):1753–1798, 2017.

F. Rabe and M. Kohlhase.
A Scalable Module System.
Information and Computation, 230(1):1–54, 2013.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever.
Language models are unsupervised multitask learners.
2019.

Q. Wang, C. E. Brown, C. Kaliszyk, and J. Urban.
Exploration of neural machine translation in autoformalization of mathematics in Mizar.
In J. Blanchette and C. Hritcu, editors, *Proceedings of the 9th ACM SIGPLAN International Conference on Certified Programs and Proofs, CPP 2020, New Orleans, LA, USA, January 20-21, 2020*, pages 85–98. ACM, 2020.

Q. Wang, C. Kaliszyk, and J. Urban.
First experiments with neural translation of informal to formal mathematics.
In F. Rabe, W. M. Farmer, G. O. Passmore, and A. Youssef, editors, *11th International Conference on Intelligent Computer Mathematics (CICM 2018)*, volume 11006 of *LNCS*, pages 255–270. Springer, 2018.