Learning Semantic Annotations for LaTeX Documents

Dennis Müller*and Cezary Kaliszyk

Department of Computer Science
University of Innsbruck, Austria

1 Introduction

In the last decades, the formalization of mathematical knowledge, and the verification and automation of formal proofs, has become increasingly popular. Formal methods nowadays are not just used by computer scientists to verify software and hardware as well as in program synthesis, but ave also drawn the interest of an increasing number of research mathematicians. By now, there is a plurality of systems available, each with its own growing library of formalized mathematics.

However, many mathematicians complain that

- formal systems are difficult to learn and use, even if one is well acquainted with the (informal) mathematics involved,
- they require a level of detail in proofs that is prohibitive even for "obvious" conclusions,
- their libraries are difficult to grasp without already being familiar with the system's language, conventions and functionalities.

Consequently, the utility of formalizing mathematical results can be too easily (and too often *is*) dismissed in light of the additional time and work required for non-experts. This is despite the fact that many services available for formal mathematics are already enabled by *semi*-formal (or *flexiformal*) representations, such as semantic annotations in natural language texts, or formal representations containing opaque informal expressions (see e.g. [Koh13], [Lan11a], [Ian17], [Koh+17b], [CS17], [Deh+16]). Therefore, we need to invest into methods for bridging the gap between informal mathematical practice and (semi-)formal mathematics.

We want to contribute to such a bridge between informal and (semi-)formal documents, by **developing a framework** using symbolic and machine learning techniques that

1. **automatically adds formal semantic annotations** to informal mathematics where possible, and
2. **highlights ambiguities** where not, in order to encourage clarification from a user.

Michael Kohlhase developed the sTeX package [Koh08] for LaTeX, specifically for annotating mathematical documents with structural and formal semantics. In particular, sTeX is based on an OMDoc [Koh06] ontology, which is foundation-agnostic in the sense that it does not favor a specific foundation (such as type or set theories) over any other. This approach is consequently best suited for semantifying informal documents, where foundations are often unspecified, left implicit or switched fluently. Furthermore, sTeX allows markup both on the level of mathematical expressions as well as on a structural level, such as declarations, definienda/definientia and theorems. Consequently, sTeX can serve as an ideal target for this goal.

*The first author and this work are supported by a postdoc fellowship of the German Academic Exchange Service (DAAD)

As a first approach, we will use the *SMGloM* [Koh14] semantic glossary of mathematics, which contains hundreds of sTEX-annotated concepts and definitions, providing LATEX-macros for their symbolic *notations* (i.e. presentation as pure LATEX) as well as introducing logical identifiers for semantically referencing concepts in natural language texts.

Individual entries in the glossary are collected in individual, .tex-files, which can be compiled into (disambiguated) OMDoc. The individual files are connected via a module system provided by the sTEX-package using the logical identifiers.

Consequently, the *SMGloM* library can serve as an ideal data set for supervised learning to 1. disambiguate formal expressions in LATEX using SMGloM macros, and 2. automatically reference SMGloM entries in natural language paragraphs.

sTEX declaration	% equality as a flexary infix operator \symdef[name=equal,gfc=N2]{eqFN}{\mathrel{=}} \symdef[name=equal,assocarg=1]{eq}[1]{\assoc[p=300]\eqFN{#1}}
sTEX references	We call two mathematical objects a and b \trefi{equal}, (written $\eq{a,b}$), iff there are no properties that discern them.
OMDoc for \eq{a,b}	<OMA> <OMS cd="http://mathhub.info/smglom/mv/equal.omdoc?equal" name="equal"/> <OMV name="a"/> <OMV name="b"/> </OMA>

`\symdef` introduces a new mathematical concept with globally unique identifier (see third row), `\trefi` allows for referencing it, the formal expression $a = b$ is disambiguated in the resulting OMDoc.

sTEX itself is integrated, and shares an underlying OMDoc ontology, with the Mmt system [RK13; HKR12; Rab17] – a foundation-independent meta-framework and API for knowledge management services. This integration makes the generic services provided by MMT available to informal mathematical texts. As a next step, we will explore the possibility of using Mmt's generic type checking component to formally verify the disambiguated expressions obtained from informal mathematical texts in the step above. This would result in a rudimentary type checker integrated into LATEX, similar to *Naproche* [Cra+09] and related systems.

Additionally, several theorem prover libraries have been translated to OMDoc and integrated in the Mmt system, e.g. [Koh+17a; MRS19] (for a detailed overview, see [Mül19] and [KR20]). This allows extending our training data to existing data sets for automated formalization (e.g. [KUV17a; KUV17b; WKU18]), potentially extending the SMGloM automatically, and provides an attractive avenue for subsequent research by using *alignments* [Mül19; Mül+17] between SMGloM and formal libraries to verify informal mathematics using several state-of-the-art theorem prover systems.

We expect the work to result in a deeper integration of formal methods in the workflows of working mathematicians (e.g. via proper integration in LATEX-IDEs), making formal methods and their advantages accessible to non-experts in STEM fields. Hopefully, this will vastly increase both their ubiquity outside the formal mathematics community and the general amount of formal mathematics available, thus also benefiting e.g. the formal abstracts and related projects.

References

[Cra+09] M. Cramer, B. Fisseni, P. Koepke, D. Kühlwein, B. Schröder, and J. Veldman. "The Naproche Project Controlled Natural Language Proof Checking of Mathematical Texts". In: *Controlled Natural Language*. Ed. by N. Fuchs. Springer, 2009, pp. 170–186.

[CS17] J. Corneli and M. Schubotz. "math.wikipedia.org: A vision for a collaborative semi-formal, language independent math(s) encyclopedia". English. In: *AITP 2017. The Second Conference on Artificial Intelligence and Theorem Proving*. 2017, pp. 28–31.

[Deh+16] P.-O. Dehaye et al. "Interoperability in the OpenDreamKit Project: The Math-in-the-Middle Approach". In: *Intelligent Computer Mathematics 2016*. Conferences on Intelligent Computer Mathematics (Bialystok, Poland, July 25, 2016–July 29, 2016). Ed. by M. Kohlhase, M. Johansson, B. Miller, L. de Moura, and F. Tompa. LNAI 9791. Springer, 2016. URL: https://github.com/OpenDreamKit/OpenDreamKit/blob/master/WP6/CICM2016/published.pdf.

[Geu+17] H. Geuvers, M. England, O. Hasan, F. Rabe, and O. Teschke, eds. *Intelligent Computer Mathematics*. Conferences on Intelligent Computer Mathematics. LNAI 10383. Springer, 2017. DOI: 10.1007/978-3-319-62075-6.

[HKR12] F. Horozal, M. Kohlhase, and F. Rabe. "Extending MKM Formats at the Statement Level". In: *Intelligent Computer Mathematics*. Ed. by J. Campbell, J. Carette, G. Dos Reis, J. Jeuring, P. Sojka, V. Sorge, and M. Wenzel. Springer, 2012, pp. 64–79.

[Ian17] M. Iancu. "Towards Flexiformal Mathematics". PhD thesis. Bremen, Germany: Jacobs University, 2017. URL: https://opus.jacobs-university.de/frontdoor/index/index/docId/721.

[Koh06] M. Kohlhase. *OMDoc: An Open Markup Format for Mathematical Documents (Version 1.2)*. Lecture Notes in Artificial Intelligence 4180. Springer, 2006.

[Koh08] M. Kohlhase. "Using LaTeX as a Semantic Markup Format". In: *Mathematics in Computer Science* 2.2 (2008), pp. 279–304.

[Koh13] M. Kohlhase. "The Flexiformalist Manifesto". In: *14th International Workshop on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC 2012)*. Ed. by A. Voronkov, V. Negru, T. Ida, T. Jebelean, D. Petcu, S. M. Watt, and D. Zaharie. Timisoara, Romania: IEEE Press, 2013, pp. 30–36. URL: http://kwarc.info/kohlhase/papers/synasc13.pdf.

[Koh14] M. Kohlhase. "A Data Model and Encoding for a Semantic, Multilingual Terminology of Mathematics". In: *Intelligent Computer Mathematics 2014*. Conferences on Intelligent Computer Mathematics (Coimbra, Portugal, July 7, 2014–July 11, 2014). Ed. by S. Watt, J. Davenport, A. Sexton, P. Sojka, and J. Urban. LNCS 8543. Springer, 2014, pp. 169–183. URL: http://kwarc.info/kohlhase/papers/cicm14-smglom.pdf.

[Koh+17a] M. Kohlhase, D. Müller, S. Owre, and F. Rabe. "Making PVS Accessible to Generic Services by Interpretation in a Universal Format". In: *Interactive Theorem Proving*. Ed. by M. Ayala-Rincón and C. A. Muñoz. Vol. 10499. LNCS. Springer, 2017. URL: http://kwarc.info/kohlhase/submit/itp17-pvs.pdf.

[Koh+17b] M. Kohlhase, T. Koprucki, D. Müller, and K. Tabelow. "Mathematical models as research data via flexiformal theory graphs". In: *Intelligent Computer Mathematics (CICM) 2017*. Conferences on Intelligent Computer Mathematics. Ed. by H. Geuvers, M. England, O. Hasan, F. Rabe, and O. Teschke. LNAI 10383. Springer, 2017. DOI: 10.1007/978-3-319-62075-6. URL: http://kwarc.info/kohlhase/papers/cicm17-models.pdf.

[KR20] M. Kohlhase and F. Rabe. "Experiences from Exporting Major Proof Assistant Libraries". 2020. URL: https://kwarc.info/people/frabe/Research/KR_oafexp_20.pdf.

[KUV17a] C. Kaliszyk, J. Urban, and J. Vyskočil. "Automating Formalization by Statistical and Semantic Parsing of Mathematics". In: *Interactive Theorem Proving*. Ed. by M. Ayala-Rincón and C. A. Muñoz. Cham: Springer International Publishing, 2017, pp. 12–27.

[KUV17b] C. Kaliszyk, J. Urban, and J. Vyskocil. "System Description: Statistical Parsing of Informalized Mizar Formulas". In: 2017. DOI: 10.1109/synasc.2017.00036.

[Lan11a] C. Lange. "Enabling Collaboration on Semiformal Mathematical Knowledge by Semantic Web Integration". Also available as a book [Lan11b]. PhD thesis. Jacobs University Bremen, 2011. URL: https://svn.kwarc.info/repos/swim/doc/phd/phd.pdf.

[Lan11b] C. Lange. *Enabling Collaboration on Semiformal Mathematical Knowledge by Semantic Web Integration*. Studies on the Semantic Web 11. Heidelberg and Amsterdam: AKA Verlag and IOS Press, 2011. URL: http://www.semantic-web-studies.net.

[MRS19] D. Müller, F. Rabe, and C. Sacerdoti Coen. "The Coq Library as a Theory Graph". accepted at CICM 2019. 2019.

[Mül+17] D. Müller, T. Gauthier, C. Kaliszyk, M. Kohlhase, and F. Rabe. "Classification of Alignments between Concepts of Formal Mathematical Systems". In: *Intelligent Computer Mathematics (CICM) 2017*. Conferences on Intelligent Computer Mathematics. Ed. by H. Geuvers, M. England, O. Hasan, F. Rabe, and O. Teschke. LNAI 10383. Springer, 2017. DOI: 10.1007/978-3-319-62075-6. URL: http://kwarc.info/kohlhase/papers/cicm17-alignments.pdf.

[Mül19] D. Müller. "Mathematical Knowledge Management Across Formal Libraries". PhD thesis. Informatics, FAU Erlangen-Nürnberg, Oct. 2019. URL: https://kwarc.info/people/dmueller/pubs/thesis.pdf.

[Rab17] F. Rabe. "How to Identify, Translate, and Combine Logics?" In: *Journal of Logic and Computation* 27.6 (2017), pp. 1753–1798.

[RK13] F. Rabe and M. Kohlhase. "A Scalable Module System". In: *Information and Computation* 230.1 (2013), pp. 1–54.

[WKU18] Q. Wang, C. Kaliszyk, and J. Urban. "First Experiments with Neural Translation of Informal to Formal Mathematics". In: *CoRR* abs/1805.06502 (2018). arXiv: 1805.06502. URL: http://arxiv.org/abs/1805.06502.