Automatic Mathematical Information Retrieval to perform translations up toComputer Algebra Systems

André Greiner-Petter
Information Science Group
University of Konstanz, Germany
andre.greiner-petter@uni-konstanz.de

Research Objectives and Plans

In mathematics, LaTeX is the de facto standard to prepare documents, e.g., scientific publications. While some formulae are still developed using pen and paper, more complicated mathematical expressions used more and more often with computer algebra systems. Mathematical expressions are often manually transcribed to computer algebra systems. The goal of my doctoral thesis is to improve the efficiency of this workflow. My envisioned method will automatically semantically enrich mathematical expressions so that they can be imported to computer algebra systems and other systems which can take advantage of the semantics, such as search engines or automatic plagiarism detection systems. These imports should preserve essential semantic features of the expression.

The translation process between semantic expressions and computer algebra systems was realized in my Master's thesis and the results of this work were published in paper [CSY$^+$17]. Therefore, I will focus on the semantic enrichment process of generic LaTeX expressions in my doctoral thesis.

To achieve this goal, I am presenting the multiple-scan approach with three parts: (1) narrow down possible meanings only from the expression itself, without referring to the context of the expression; (2) refine the process with conclusions from the nearby context of the expressions, and (3) improve the previous process by analyzing not only the nearby context but the overall topic of the whole scientific paper or book, its references and other publications by the authors.

Objective (1) concentrates on the expression itself, without extracting information from the context. My proposed approach is to exploit the coherence between the structure of a given formula and its meaning, constructing a Markov logic network to deduce possible semantic meanings. Therefore, each meaning gets a probability. If the highest probability is below a given threshold, it would be necessary to use (2) and (3) for improving the probabilities. Otherwise, the probability is sufficiently high for concluding a semantic information.

For example, consider the Jacobi polynomial

$$P_n^{(\alpha,\beta)}(\cos(a\Theta)). \tag{1}$$

The given expression has a superscript, a subscript and a following expression in parentheses. A leading expression in letters with a following expression in parentheses may lead us to the conclusion that the leading expression is the name of a function and the expression in parentheses is its argument. Additionally, the first symbol P has a superscript and a subscript. However, the Meixner-Pollaczek polynomial $P_n^{(\lambda)}(x;\phi)$ and the associated Legendre function of the first kind $P_\nu^\mu(x)$ are also referenced with P and all of these functions has a superscript, a subscript and an argument. But the Jacobi polynomial assumes a superscript of two parameters, while the Meixner-Pollaczek polynomial and the Legendre function just assume one parameter in the superscript.

Assume we cannot conclude a unique mathematical object only from the expression itself. In those cases, we will investigate the context of the input, such as that mentioned in objective (2). A large-scale corpus study showed that around 70 percent of the symbolic elements in scientific papers are denoted in the text. Therefore, the idea

In: O. Hasan, D. Gallois-Wong (eds.): Proceedings of CICM 2018, Hagenberg, Austria, 13-08-2018, published at http://ceur-ws.org

is to identify the symbols in a formula (called identifiers) and in the surrounding text and identify its describing key words (called definiens). Once we have extracted possible identifier-definiens pairs, we can score each pair to conclude the most likely pair. This approach is already published by M. Schubotz et al. [SGL+16] in 2016. The scoring process assumes that the chance for a correct combination of identifier and definiens depends on the distance between identifier and its definiens and the distance of the identifier to the closest formula that contains this identifier. I strongly believe we can improve the score process with the conclusions from my first objective above.

If the correct semantic information is still unsure, objective (3) is the last way to find a solution. Online compendia, such as arXiv, can be used to discover the overall topic of a scientific paper, the references and the area of research of the authors. There already exists engines that try to find dependencies between publications by examine the citations, titles and abstracts. I am planning to incorporated realized approaches to solve objective (3).

Completed & Remaining Research

Since objective (1) is part of the Part-of-Math (POM) tagger [You17], I focus on objectives (2) and (3), and support the progress of the POM-Tagger collaboratively with the DRMF project team. In my first contribution towards (2), I analyzed the capabilities of existing tools to perform conversions from plain LaTeX to content MathML [SGPS+18]. The developed gold standard is used to measure accuracies and identify weaknesses of state-of-the-art tools. First experiments have shown that we were able to increase the accuracies significantly by adding semantic LaTeX macros based on the results of the context analyzation process using [SGL+16].

The next project aims to solve the problem from a different perspective. Wikipedia as a highly frequently used lexicon has over 17 million edits every month. During the last two years, 7 million different formulae have been edited. Wikipedia uses TeX-Markup since 2003 for mathematical expressions. We consider the Wikipedia word processor as a highly suitable test environment to add a recommendation system to enhance mathematical LaTeX input with semantics. The idea is to provide recommendations for replacing the plain LaTeX input by a semantic version using macros. A recommendation will be given by a machine learning (ML) algorithm trained by the defined backward translation, i.e., from semantic macros to plain LaTeX. The algorithm will further learn supervised from the selections from an editor. An implemented would than slowly increase semantic information of mathematics in Wikipedia and improve the ML algorithm. Subsequently, the algorithm can be used for automatic translations.

Summary

I am still at the beginning of my doctoral research, and the described approaches are ambitious. However, our first contributions have shown valuable results, and the developed gold standard builds a fundamental construct for evaluating our upcoming projects.

References

[CSY+17] Howard S. Cohl, Moritz Schubotz, Abdou Youssef, André Greiner-Petter, Jrgen Gerhard, Bonita V. Saunders, Marjorie A. McClain, Joon Bang, and Kevin Chen. Semantic preserving bijective mappings of mathematical formulae between document preparation systems and computer algebra systems. In *Lecture Notes in Computer Science*, pages 115–131. Springer International Publishing, 2017.

[SGL+16] Moritz Schubotz, Alexey Grigorev, Marcus Leich, Howard S. Cohl, Norman Meuschke, Bela Gipp, Abdou S. Youssef, and Volker Markl. Semantification of identifiers in mathematics for better math information retrieval. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pages 135–144, New York, NY, USA, 2016. ACM.

[SGPS+18] Moritz Schubotz, Andre Greiner-Petter, Philipp Scharpf, Norman Meuschke, Howard Cohl, and Bela Gipp. Improving the representation and conversion of mathematical formulae by considering their textual context. In *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*, Fort Worth, USA, Jun. 2018.

[You17] Abdou Youssef. Part-of-math tagging and applications. In *Lecture Notes in Computer Science*, pages 356–374. Springer International Publishing, 2017.