Erschienen in: RecSys '19 : Proceedings of the 13th ACM Conference on Recommender Systems : September 16-20, 2019, Copenhagen, Denmark. - New York : ACM Press, 2019. - S. 532-533. - ISBN 978-1-4503-6243-6
https://dx.doi.org/10.1145/3298689.3347042

AnnoMathTeX - a Formula Identifier Annotation Recommender System for STEM Documents

Philipp Scharpf
University of Konstanz, Germany
philipp.scharpf@uni-konstanz.de

Ian Mackerracher
University of Konstanz, Germany
ian.mackerracher@uni-konstanz.de

Moritz Schubotz
University of Wuppertal, Germany
schubotz@uni-wuppertal.de

Joeran Beel
Trinity College Dublin, Ireland
joeran.beel@tcd.ie

Corinna Breitinger
University of Konstanz, Germany
corinna.breitinger@uni-konstanz.de

Bela Gipp
University of Konstanz & University of Wuppertal, Germany
gipp@uni-wuppertal.de

ABSTRACT

Documents from science, technology, engineering and mathematics (STEM) often contain a large number of mathematical formulae alongside text. Semantic search, recommender, and question answering systems require the occurring formula constants and variables (identifiers) to be disambiguated. We present a first implementation of a recommender system that enables and accelerates formula annotation by displaying the most likely candidates for formula and identifier names from four different sources (arXiv, Wikipedia, Wikidata, or the surrounding text). A first evaluation shows that in total, 78% of the formula identifier name recommendations were accepted by the user as a suitable annotation. Furthermore, document-wide annotation saved the user the annotation of ten times more other identifier occurrences. Our long-term vision is to integrate the annotation recommender into the edit-view of Wikipedia and the online LaTeX editor Overleaf.

KEYWORDS

Information Retrieval, Mathematical Information Retrieval, Recommender Systems, Semantification, Wikipedia/Wikidata

1 INTRODUCTION

Documents from Science, Technology, Engineering, and Mathematics (STEM) often contain numerous mathematical formulae [1], which are crucial to understanding the semantics of the text. If the formula characters (constants or variables) are not annotated, the mathematical statement of a formula cannot be understood and queried. However, if for example the formula $S = 1 - \frac{|R| \cdot |U|}{|I|}$ was annotated $\{S : sparsity,\ R : ratings,\ I : items,\ U : users\}$, the characters (in the following referred to as formula identifiers) are translated into words that represent their meaning.

This enables semantic search, recommender and mathematical question answering systems [2] to find documents with formulae that for example

- allow calculating *sparsity* or
- allow calculating *sparsity*, given *ratings*, *items*, and *users* or
- contain specific variables, such as *ratings* and *items* or
- relate *ratings* and *users*.

These are examples of structured queries, which require machine-interpretability of mathematical documents to approach Mathematical Language Understanding (MLU). A large part of the mathematical knowledge today is either contained within research papers (LaTeX) or in condensed form in Wikipedia articles (Wikitext).

Wikipedia articles are only semi-structured (linked). For the direct retrieval of specific facts and systematic queries, Wikidata was launched in 2012 [3]. Language-independent items (identified by a unique ID) are linked by properties.

In addition to natural language statements, mathematical formulae were transferred from Wikipedia [2] as items with a "defining formula" property that allows a LaTeX formula string as value. However, only a few formulae contain their identifier names. Thus, a large part is not machine-interpretable (=allowing structured queries).

Prior research has aimed to extract the identifier meaning from the text that surrounds the formula [4, 5], but all approaches lack an important element: the quality control afforded by a human expert verifier. Annotating multitudinous formulae can be tedious. Since the identifier annotation in a document must be globally consistent, annotating each instance individually should be avoided.

We address these shortcomings by introducing an annotation recommender system[1, 2] for formula identifiers at the document level. We evaluate our system's performance while comparing the user's acceptance of recommendations from four different sources.

[1]System hosted by Wikimedia Foundation at annomathtex.wmflabs.org
[2]Demo video available at bit.ly/annomathtex

2 ANNOMATHTEX

The workflow of our system is as follows: a user uploads a mathematical document in Wikitext or LaTeX format. The system displays the text while highlighting formulae and identifiers. The formulae are located by searching for their environment tags (`<math>, $, \{equation}, \{align}`, etc.). Parsing the formulae yields their identifiers, which are then shown in color. If the user clicks on a formula identifier, AnnoMathTeX presents recommendations for its name, which we extracted using four different sources: 1) *arXiv* - candidates[3] extracted from the surrounding text of 60 M formulae 2) *Wikipedia* - candidates[4] extracted from definitions in mathematical English articles 3) *Wikidata* - candidates retrieved via a SPARQL query[5] 4) a *surrounding text window* of ±5 words around the formula. The recommendations are then generated from static dump lists and ranked by the occurrence frequency in their sources.

Figure 1 shows the recommendation table/matrix. Each column corresponds to one source and is presented to the user in a shuffled order and using anonymous labels to avoid bias. If no recommendation matches, the user can type in the correct identifier name directly. By default, identifiers are annotated globally and automatically annotated at any further occurrence within the document to enable significant time savings. In the rare case of a double meaning within the same document, a locally different annotation is possible. All annotations made by the user are shown as rows at the top of the document and saved in a separate annotation file. Finally, the user's selection is stored in an evaluation file to compare the usefulness of the four sources.

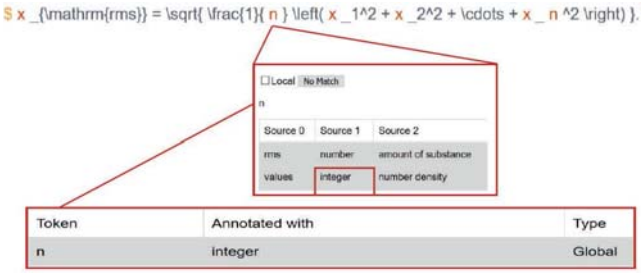

Figure 1: AnnoMathTeX recommendations for formula identifier annotation.

3 EVALUATION

As a proof-of-concept, we evaluate the performance of recommendations for formula identifiers comparing the four sources. Annotating a sample of 100 identifiers from 10 different Wikipedia articles, we find that the acceptance distribution (item coverage) of the sources is {*arXiv*: 35%, *Wikipedia*: 16%, *Wikidata*: 13%, *WordWindow*: 35%}. Overall, 82% of the recommendations were accepted. On average, the accepted recommendation was ranked third (3.0) out of ten, with a ranking distribution of {*arXiv*: 2.3, *Wikipedia*: 4.0, *Wikidata*: 2.5, *WordWindow*: 3.1}.

We conclude that in most cases, the recommendations are useful, and thus, the system can significantly speed up the annotation process. Furthermore, 99% of the identifiers could be annotated globally, saving the user 1045 annotations - on average 105 per document and 10 per identifier.

4 CONCLUSION & OUTLOOK

We demonstrated a first recommender for mathematical identifier annotation. Our presented system enables researchers to quickly disambiguate formula identifiers, and thus contributes significantly towards the aim of making mathematical documents machine-interpretable.

Converting mathematical language statements encoded in formulae into natural language is a crucial task for enabling semantic search queries, and for improving mathematical recommender and question answering systems. In a preliminary evaluation, our system suggested correct names for 78% of the examined identifier instances.

As a next step, we will implement the possibility to further deepen the annotation by mathematical referencing [6]. The user will be able to link formulae and identifiers to items of the semantic knowledge-base Wikidata. Subsequently, we plan to carry out a large-scale user study in which we will evaluate the *formula name* recommendations from the following sources: 1) surrounding text 2) a history of manual inserts 3) a self created database of annotated formulae, and 4) Wikidata.

Our long-term aim is to directly integrate our annotation recommender into the editing or composing views of both Wikipedia and Overleaf. This would allow for the Wikipedia and research communities to be directly included in the semantification process of mathematical articles and research papers.

ACKNOWLEDGMENTS

This work was supported by the German Research Foundation (DFG grant GI-1259-1). We thank the Wikimedia Foundation for hosting the system.

REFERENCES

[1] Radu Hambasan and Michael Kohlhase. Faceted search for mathematics. In *LWA*, volume 1458 of *CEUR Workshop Proceedings*, pages 33–44. CEUR-WS.org, 2015.
[2] Moritz Schubotz, Philipp Scharpf, Kaushal Dudhat, Yash Nagar, Felix Hamborg, and Bela Gipp. Introducing mathqa: a math-aware question answering system. *Information Discovery and Delivery*, 46(4):214–224, 2018.
[3] Denny Vrandecic and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85, 2014.
[4] 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 *SIGIR*, pages 135–144. ACM, 2016.
[5] Giovanni Yoko Kristianto, Goran Topic, and Akiko Aizawa. Extracting textual descriptions of mathematical expressions in scientific papers. *D-Lib Magazine*, 20(11/12), 2014.
[6] Michael Kohlhase. Math object identifiers - towards research data in mathematics. In *LWDA*, volume 1917 of *CEUR Workshop Proceedings*, page 241. CEUR-WS.org, 2017.

[3] http://ntcir-math.nii.ac.jp/data
[4] https://en.wikipedia.org/wiki/User:Physikerwelt
[5] https://query.wikidata.org