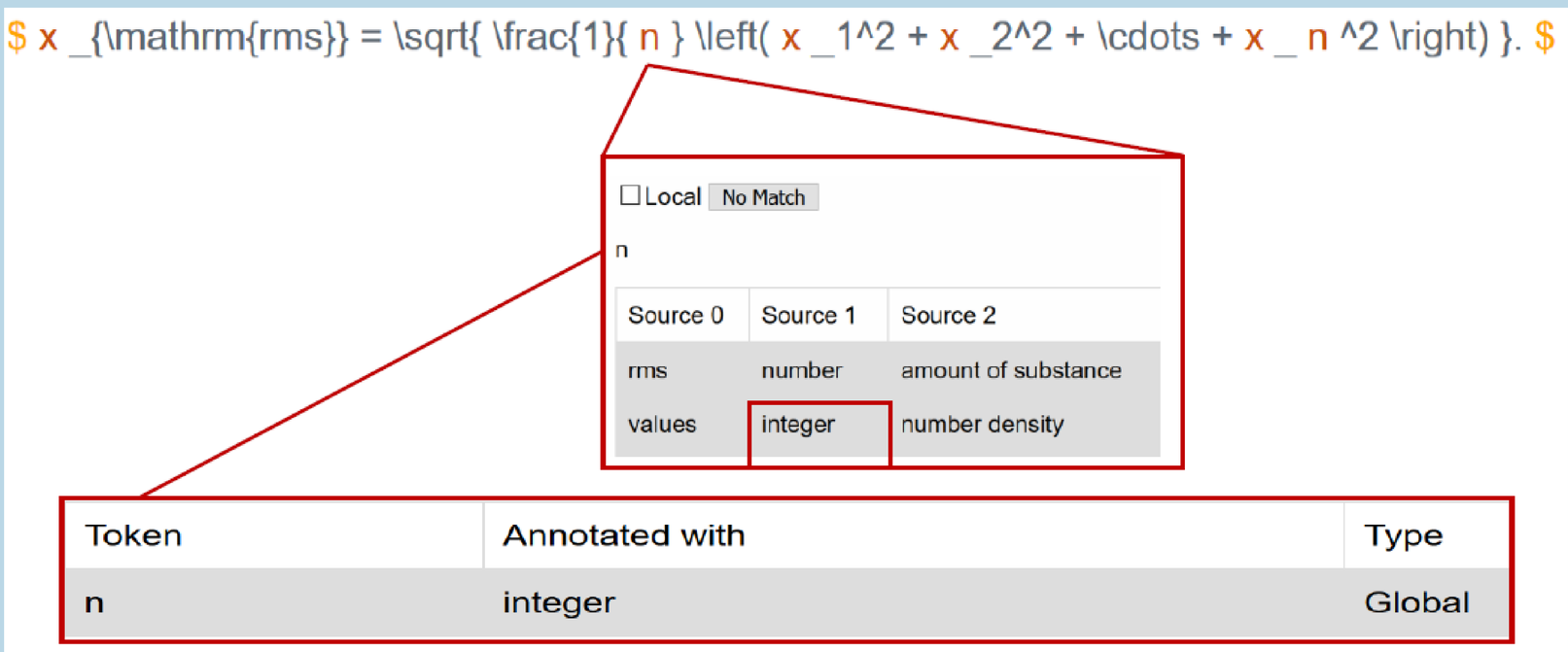
AnnoMathTeX - a Formula Identifier Annotation Recommender System for STEM Documents

Philipp Scharpf[1], Ian Mackerracher[1], Moritz Schubotz[2], Joeran Beel[3], Corinna Breitinger[1], Bela Gipp[1,2]
[1]first.last@uni-konstanz.de, [2]last@uni-wuppertal.de, [3]joeran.beel@tcd.ie

www.isg.uni.kn

AnnoMathTeX



System hosted by Wikimedia Foundation at `annomathtex.wmflabs.org`.
Demo video available at `bit.ly/annomathtex`.

Abstract

STEM documents often contain a **large number of mathematical formulae**.

The occuring formula symbols **(identifiers) must be disambiguated**.

Manual annotation can be very **time-consuming**.

We present a Recommender System that accelerates formula annotation by displaying the **most likely candidates** for **formula and identifier names**.

A first evaluation shows that in total, **78%** of the formula identifier name **recommendations were accepted** by the user.

Furthermore, document-wide annotation **saved the user the annotation of ten times more** other identifier occurrences.

Introduction

The **semantics of formulae** are crucial to understanding a **STEM document**. [1]

If for example the formula

$$S = 1 - \frac{|R| \cdot |U|}{|I|} \tag{1}$$

was annotated $\{S : sparsity,\ R : ratings,\ I : items,\ U : users\}$, the symbols (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*.

Prior research has aimed to **extract** the identifier meaning from the **text that surrounds** the formula [3, 4], but **lack the quality control** of a human expert verifier.

Workflow

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.).

If the user clicks on a formula identifier, AnnoMathTeX presents recommendations for its name, which we extracted using four different sources:

- **arXiv** - candidates[a] extracted from the surrounding text of 60 M formulae
- **Wikipedia** - candidates[b] extracted from definitions in mathematical English articles
- **Wikidata** - candidates retrieved via a SPARQL query[c]
- a **surrounding text window** of ±5 words around the formula

. The recommendations are then generated from **static dump lists** and textbf by the **occurrence frequency** in their sources.

In 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**.

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**.

[a] `http://ntcir-math.nii.ac.jp/data`
[b] `https://en.wikipedia.org/wiki/User:Physikerwelt`
[c] `https://query.wikidata.org`

Evaluation

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.

Outlook

As a next step, we will implement the possibility to further deepen the annotation by mathematical referencing [5]. The user will be able to **link formulae and identifiers** to **items** of the semantic knowledge-base **Wikidata**.

Our long-term aim is to **directly integrate** our annotation recommender into the **editing or composing** views of both **Wikipedia and Overleaf**.

Acknowledgements

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] 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.

[4] Giovanni Yoko Kristianto, Goran Topic, and Akiko Aizawa. Extracting textual descriptions of mathematical expressions in scientific papers. *D-Lib Magazine*, 20(11/12), 2014.

[5] Michael Kohlhase. Math object identifiers - towards research data in mathematics. In *LWDA*, volume 1917 of *CEUR Workshop Proceedings*, page 241. CEUR-WS.org, 2017.