# Extracting Mathematical Concepts with Large Language Models

Valeria de Paiva
(joint work with Qiyue Gao, Pavel Kovalev, and Larry Moss)

MathUI Workshop, CICM 2023

7 September, 2023

TOPOS INSTITUTE

# Reading Mathematics

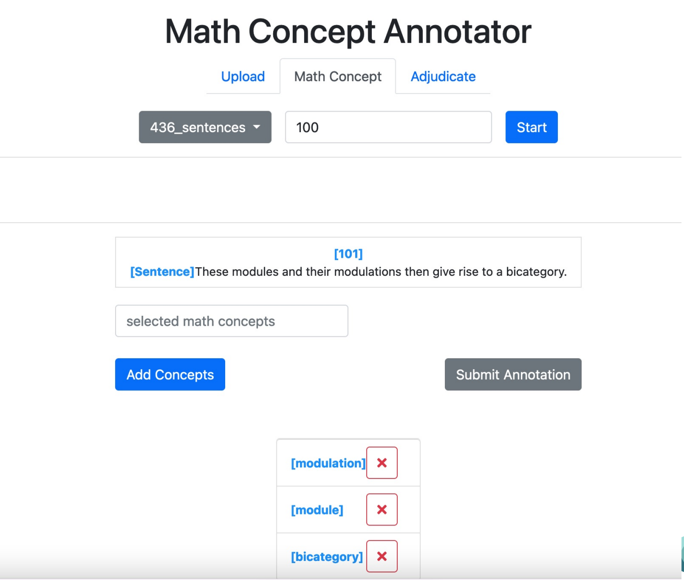
When you read some mathematics you're not familiar with...

We define the notion of a torsor for an inverse semigroup, which is based on semigroup actions, and prove that this is precisely the structure classified by the topos associated with an inverse semigroup. Unlike in the group case, not all set-theoretic torsors are isomorphic: we shall give a complete description of the category of torsors...

Want to create a personal *index* automatically and on the fly.

Can we do it?

# Maybe, maybe not...



# Networked Mathematics


- A project of the Topos Institute: a non-profit start-up about math for the benefit of society
- multidisciplinary research is **hard** to fund!
- Still: in EuroProofNet workshop NOW, Lucy Horowitz on **MathGloss**, connecting electronic glossaries
- Next week, Jacob Collard on **Parmesan** Mathematical Concept Extraction for Education, arxiv 2307.06699
- This work, a collaboration with Indiana U, using chatGPT
- Next year, Hausdorff Institute of Mathematics: can connect to formalizations?

TOPOS INSTITUTE

# Extracting mathematical concepts

- NLP tools are improving incredibly **fast**
- Can we extract mathematical concepts from any text and link them to some math ontology?
- How well 'automatic term extractors' (ATE) systems work for mathematical text?
- No ATE system dedicated to mathematics?
- checking baselines is HARD, as they don't exist.
- no human annotated datasets, on math concepts? 2017 scientific concepts, no transfer
- Results with *silversets* (Collard et al 2022) very low
- Now Parmesan `http://www.jacobcollard.com/parmesan2/` math term search

TOPOS INSTITUTE

# Goals of this work

- Make math more accessible to everyone: students, mathematicians from different areas, scientists, interested laypeople
- a WordNet for mathematics
- Produce indices for monographs, books, self-study more easily
- Make sure concepts extracted are correct and as complete as possible
- Understand opportunities and limitations of large language models – Old English similar problem
- Help start building a Mathematical Knowledge Graph
- (eventually) Mathematical Natural Language Inference (mathNLI)

# Mathematical Vernacular our way

- Many mathematical terms are NOT concepts: *characterization,conjecture, consequence, counter-example, paper, etc*
- Traditional issues of terminologies:
  - plurals or not? (*presheaves or presheaf*?),
  - adjectives without a noun? (*2-categorical refinement*?),
  - adjective is related to noun/verb, which one? (*interpolated function, interpolate, interpolation*?)
- constructions with two or more adjectives, add sub-expressions *differential graded category, graded category*?
- expressions with prepositions, e.g. *sheaf of germs of analytic functions*
- proper nouns in adjective-like positions as *Shanin's method, Lagrange interpolation*

TOPOS INSTITUTE

# Annotation guidelines

- Treat math concepts as black-boxes, as possible
- Use singular for concepts
- convention: Terms like theorem, corollary, conjecture, paper are not concepts
- convention: Mathematicians are not mathematical concepts
- If a long span is a concept, add important subspans e.g. *enriched accessible category, accessible category*
- avoid prepositions inside of concepts *sheaf of germs of analytic functions* we add instead *sheaf, germ, analytic function*

TOPOS INSTITUTE

# Corpus description

- We use an already prepared corpus, consisting of 755 abstracts from the open source journal *Theory and Application of Categories*, around 2020, at `https://github.com/ToposInstitute/tac-corpus`.
- The abstracts were processed with the BERT version of spaCy, from 2022.
- We also use the corpus of nLab entries, already processed with spaCy and available at `https://github.com/ToposInstitute/nlab-corpus`.
- Instead of using an ATE system we want to see how chatGPT can do in the task of extracting concepts
- We use sanitized sentences for that, i.e. sentences of medium length and with no LaTeX, after LateXML (Miller).

TOPOS INSTITUTE

# Bringing in chatGPT

- Three experiments: 100 sentences Pilot, 436TAC, 55KnLab
- for Pilot100 we have three mathematicians and chatGPT is the fourth annotator
- For 436TAC only one annotator and chatGPT, plus adjucation
- For 55KnLab no human evaluation, so far

---

Given the following Context, extract the words that denote Math concepts.

Here are some examples:
{*in-context example*}

Now please solve the following problem.

Context: {*math_sentence*}
Concepts:

---

The initial prompt template, where *math_sentence* denotes the sentence from which we wish to extract math concepts.

# chatGPT prompts

Given the following Context, extract the words that denote Math concepts.

Be sure to make the concept words singular. For example when we see 'functors' in a sentence, we would extract 'functor' rather than 'functors' or when we see 'categories' we would like to extract 'category' instead of 'categories'!

Here are some examples:
{*in-context example*}

Also note that we are looking for concepts like modulation, enriched orthogonality, holonomy, localization, variety, but not words shown in daily English sentences like 'future work', 'conclusion', 'this property'!

We don't want a person's name to be extracted as a math concept although we understand that a person's name could be part of the phrase that denotes a math concept.

Now please solve the following problem.

Context: {*math_sentence*}
Concepts:

Table: The updated prompt template with specialized instructions shown in different colors.

TOPOS INSTITUTE

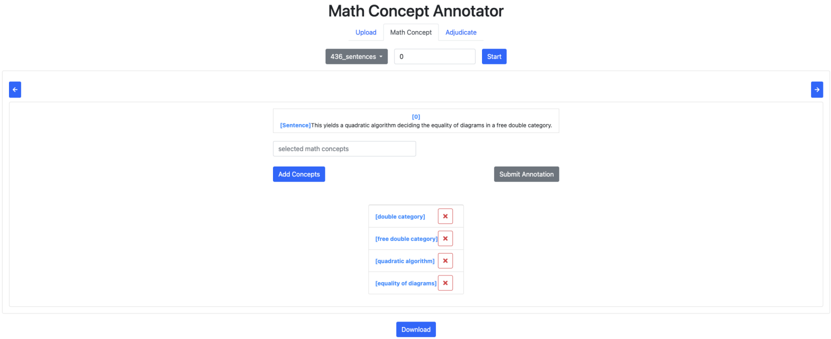
# Experiments and Evaluation

- Pilot experiment: 100 sentences from TAC abstracts
- Examples of sentences:

  | |
  |---|
  | We check these extra assumptions in several categories with pre-topologies.<br>We show that both approaches give equivalent bicategories.<br>In this paper, we use the language of operads to study open dynamical systems. |

- Many 'minor mistakes' were made by the annotators
- NLI annotation tool repurposed

# Annotation Tool



# Results

## Results on experiment 1

| Annotators Being Compared | Jaccard Score |
|---|---|
| annotator 1 and annotator 2 | 0.753 |
| annotator 1 and annotator 3 | 0.794 |
| annotator 2 and annotator 3 | 0.746 |
| ChatGPT and annotator 1 | 0.485 |
| ChatGPT and annotator 2 | 0.518 |
| ChatGPT and annotator 3 | 0.505 |
| ChatGPT and union of the humans | 0.45 |
| ChatGPT (after filtering) and union of the humans | 0.5 |

**Figure 3:** Comparison of Jaccard similarities in experiment 1.

## Results on experiment 2

| Measurement | Jaccard Score |
|---|---|
| Between Human and Un-filtered ChatGPT | 0.531 |
| Between Human and Filtered ChatGPT | 0.631 |

TOPOS INSTITUTE

# Contributions of this work

- discuss the problems of mathematical term extraction
- discuss inter-annotator (dis)agreements
- guidelines to standardize the process
- an annotation tool to help humans with ATE
- best practices for prompts to ChatGPT
- discussing whether ChatGPT could be used as an annotator similar to a human expert

# Conclusions

- Results over surveyable collections of data are not too good
- With judicious prompting, ChatGPT can find math terms, but misses some important ones
- Also it extracts terms that are not “mathematical terms” e.g. *conjecture*
- ChatGPT may be used very easily and gives acceptable preliminary results
- we need to measure the quality of these LLMs results
- we have first numbers, hopefully we can build on these

Thanks!

# Some References


- Collard et al, *Extracting Mathematical Concepts from Text.* In W-NUT 2022, arxiv 2208.13830.
- Collard et al, *Parmesan: mathematical concept extraction for education* arxiv 2307.06699.
- Horowitz, de Paiva, *MathGloss: Linked Undergraduate Math Concepts*, EuroProofNet Workshop 2023.
- *spaCy* Industrial-Strength Natural Language Processing, `https://spacy.io/`
- *Universal Dependencies*, `https://universaldependencies.org/`
- *LateXML*, `https://math.nist.gov/~BMiller/LaTeXML/`
- *Math Concept Annotator*, `https://gaoq111.github.io/math_concept_annotation/`


TOPOS INSTITUTE

# Extra

Here are some of the concepts which all three human annotators found but ChatGPT missed:

- Grothendieck's six operations
- Lie algebra
- arithmetic variety
- cartesian closed category
- closed category
- graph rewriting
- group
- homotopy
- left proper model structure
- morphism axiom
- pointed regular protomodular category
- probability distribution
- quotient triangulated category
- representation
- smooth stack
- sup-lattice
- topology
- triple category

Here are some of the concepts found by ChatGPT which none of the three human annotators found:

- Grothendieck
- acyclic models method
- algebraic context
- analysis
- application
- approach
- axiom
- balanced category
- calculate
- categorical
- categorical property
- category theory
- characterization
- closed monoidal
- consequence
- construction
- corollary
- definition
- example
- issue
- language
- localize
- locally presentable
- mathematically natural
- motivation
- non-abelian
- open question
- perspective
- previous work
- proof
- property
- prove
- result
- six operation
- structure
- symmetric monoidal
- tame
- theory
- trivial
- uniqueness statement

TOPOS INSTITUTE