Extracting Definienda in Mathematical Scholarly Articles with Transformers

Shufan Jiang[1,2] Pierre Senellart[1,3]

[1]DI ENS, ENS, PSL University, CNRS & Inria, Paris, France

[2]FIZ Karlsruhe, Germany

[3]Institut Universitaire de France

October 27, 2023

The 2nd WIESP @ IJCNLP-AACL 2023

Context

- Mathematical scholarly articles contain mathematical statements such as axioms, theorems, proofs, etc.
- Semantic knowledge in these articles are not captured by traditional ways of navigating the scientific literature, e.g., keyword search.
- We aim to propose a better knowledge discovery from mathematical papers, especially those with PDF versions only.

Extraction of Definienda from Math. Definitions

Mathematical definition in PDF

$\downarrow$

Text of the definition

$\downarrow$

definienda (terms defined within)

Extraction of Definienda from Math. Definitions

Mathematical definition in PDF

$\downarrow$

Text of the definition

$\downarrow$

definienda (terms defined within)

Definition 2.1. *Let V be a vector space over the field F. We say that the collection σ of subspaces is a* spread *if (1) $A, B \in \sigma$, $A \neq B$ then $V = A \oplus B$, and (2) every nonzero vector $x \in V$ lies in a unique member of σ. The members of σ are the* components *of the spread.*

```
\begin{definition}
Let $V$ be a vector space over the field $F$. We say that the
collection $\sigma$ of subspaces is a \emph{spread} if (1) $A,B
\in \sigma$, $A\neq B$ then $V=A\oplus B$, and (2) every nonzero
vector $x\in V$ lies in a unique member of $\sigma$. The members
of $\sigma$ are the \emph{components} of the spread.
\end{definition}
```

Extraction of Definienda from Math. Definitions

Mathematical definition in PDF

$\downarrow$

Text of the definition

$\downarrow$

definienda (terms defined within)

`Definition 1. Let V be a vector space over the field F . We say that the collection` σ `of subspaces is a spread if (1) A, B` $\in \sigma$`, A/= B then V = A` $\oplus$ `B, and (2) every nonzero vector x` $\in$ `V lies in a unique member of` σ`. The members of` σ `are the components of the spread.`

Extraction of Definienda from Math. Definitions

Mathematical definition in PDF

$\downarrow$

Text of the definition

$\downarrow$

definienda (terms defined within)

Definition 1. Let V be a vector space over the field F . We say that the collection σ of subspaces is a spread if (1) A, B $\in$ σ, A/= B then V = A $\oplus$ B, and (2) every nonzero vector x $\in$ V lies in a unique member of σ. The members of σ are the components of the spread.

spread components

Labeled Dataset Construction

High-quality (but not complete!) dataset construction:

1. Collected the LaTeX source of all 28 477 arXiv papers in Combinatorics published before 1st Jan 2020

Labeled Dataset Construction

High-quality (but not complete!) dataset construction:

1. Collected the LaTeX source of all 28 477 arXiv papers in Combinatorics published before 1st Jan 2020
2. Extracted definitions within `definition` environments

Labeled Dataset Construction

High-quality (but not complete!) dataset construction:

1. Collected the LaTeX source of all 28 477 arXiv papers in Combinatorics published before 1st Jan 2020
2. Extracted definitions within `definition` environments
3. Use arguments of `\textit{}` and `\emph{}` commands within definitions, as well as optional arguments to `\begin{definition}[]` as definienda

Labeled Dataset Construction

High-quality (but not complete!) dataset construction:

1. Collected the LaTeX source of all 28 477 arXiv papers in Combinatorics published before 1st Jan 2020
2. Extracted definitions within `definition` environments
3. Use arguments of `\textit{}` and `\emph{}` commands within definitions, as well as optional arguments to `\begin{definition}[]` as definienda
4. Clean up some common noise

Labeled Dataset Construction

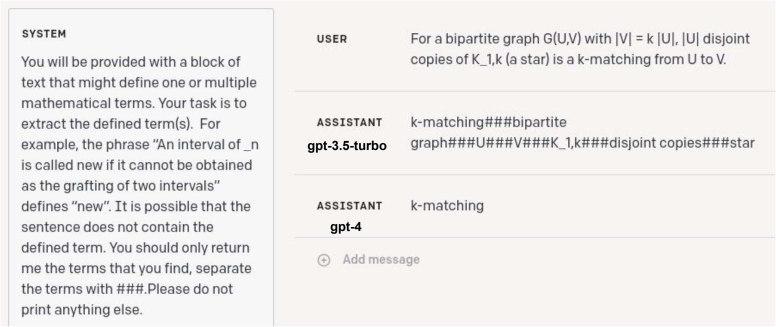
High-quality (but not complete!) dataset construction:

1. Collected the LaTeX source of all 28 477 arXiv papers in Combinatorics published before 1st Jan 2020
2. Extracted definitions within `definition` environments
3. Use arguments of `\textit{}` and `\emph{}` commands within definitions, as well as optional arguments to `\begin{definition}[]` as definienda
4. Clean up some common noise
5. Examine by hand 1 024 labeled entries. Only 30 annotated texts out of 1 024 were incorrectly labeled. Manually corrected, to obtain a test data set of 999 labeled texts with 1552 definienda. (The rest of the dataset, not manually checked, becomes training data.)

Fine-tuning Pre-trained Language Models for Token Classification

- We experimented with an out-of-the-box and general language model Roberta-base (Liu et al., 2019) and a domain-specific model cc_math_roberta (Mishra et al., 2023).
- We experimented with 1 024, 2 048, and 10 240 samples to see the performance of the classifiers with low resources.
- To evaluate the predictions, we used the predicted tag of the first word piece of each word and regrouped the IOB2-tagged word into definienda.

Alternative Approach: Querying GPT



Experimental Results

Model	**GPT-3.5**	**GPT-4**	**cc_ep01**	**cc_ep10**	**Rob.**
Training data	1	1	10240	10240	10240
Precision	0.1929	0.6248	0.420	0.652	0.697
Recall	0.8312	0.8821	0.473	0.743	0.794
F_1	0.3131	0.7315	0.442	0.692	0.742

Conclusion

- Fine-tuned classifiers have more balanced precision and recall and much smaller cost
- GPT's answers have better recall but much poorer precision than fine-tuned models
- GPT-3.5 tends to over predict formulas and mathematical expressions, while GPT-4 shows an impressive capacity to understand mathematical texts with only one example in the prompt

Future Work

- Test on a broader, more diverse, dataset of PDF papers (but if no LaTeX source available, no automatic construction of a labeled dataset)
- Extract terms elsewhere in the paper to link them back to their original definition
- Improve the robustness of domain-specific language models over different NLP tasks beyond extraction of definienda

Thank you for your attention!

Related Work

Name	Dataset	Remarks
ArGot (Berlioz, 2021)	mathematical arXiv papers	Use static word embeddings and hand-codes features. Mathematical expressions and formulas are masked out.
Scholarphi (Head et al., 2021)	papers in general domain	Use transformer-based architecture syntactic features & heuristic rules. Only processes papers with LaTeX sources.
NaturalProofs (Welleck et al., 2021)	mathematical papers & textbook	Extract definitions with hand-crafted rules. Definienda are not annotated.

Results of Fine-tuning PLM

Model	**cc_ep01**	**cc_ep10**	**Rob.**
Extracted	2093.0	1710.8	1764.2
True positive	514.9	881.2	934.2
TP+Split Term	693.8	1056.5	1127.5
Too Long	170.2	209.1	268.8
Cut Off	522.6	405.2	326.1
Precision	0.354	0.623	0.646
Recall	0.447	0.681	0.726
F_1	0.383	0.647	0.679

Results of Fine-tuning PLM with more training data

Model	cc_ep01	cc_ep10	Rob.
Extracted	1775.2	1779.2	1770.5
True positive	540.3	972.6	1082.6
TP+Split Term	733.9	1152.5	1232
Too Long	143.5	201.3	233.7
Cut Off	509.6	438.2	274.1
Precision	0.420	**0.652**	**0.697**
Recall	0.473	0.743	0.794
F_1	0.442	0.692	**0.742**

Evaluation of GPT's answers

Model	**GPT-3.5**	**GPT-4**
Extracted	6867	2245
True Positive	1072	942
TP+Split Term	1315	1383
Too Long	379	595
Cut Off	656	138
Precision	0.1929	0.6248
Recall	**0.8312**	**0.8821**
F_1	0.3131	**0.7315**

References

- Luis Berlioz. *ArGoT: A Glossary of Terms extracted from the arXiv*. *Electronic Proceedings in Theoretical Computer Science*, 342:14–21, 2021.
- Andrew Head et al. *Augmenting scientific papers with just-in-time, position-sensitive definitions of terms and symbols*. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2021.
- Sean Welleck et al. *NaturalProofs: Mathematical Theorem Proving in Natural Language*. In *The 35th NeurIPS, Datasets and Benchmarks Track (Round 1)*, 2021.
- Yinhan Liu et al. *Roberta: A robustly optimized BERT pretraining approach*. *arXiv preprint arXiv:1907.11692*, 2019.
- Shrey Mishra, Antoine Gauquier, and Pierre Senellart. *Multimodal Machine Learning for Extraction of Theorems and Proofs in the Scientific Literature*. *arXiv preprint arXiv:2307.09047*, 2023.