SCHOOL OF COMPUTATION,
INFORMATION AND TECHNOLOGY —
INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Bachelor's Thesis in Informatics

An Approach to Coreference Resolution and Formula Grounding for Mathematical Identifiers using Large Language Models

Aamin Dev

SCHOOL OF COMPUTATION, INFORMATION AND TECHNOLOGY — INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Bachelor's Thesis in Informatics

An Approach to Coreference Resolution and Formula Grounding for Mathematical Identifiers using Large Language Models

Ein Ansatz zur Auflösung von Koreferenzen und zur Ermittlung von Formeln für mathematische Symbole mit Hilfe von Large Language Models

Author:	Aamin Dev
Supervisor:	Prof. Dr. Georg Groh, Prof. Dr. Yusuke Miyao
Advisor:	Miriam Anschütz, Takuto Asakura
Submission Date:	2023-09-15

I confirm that this bachelor's thesis is my own work and I have documented all sources and material used.

Garching bei. München, 2023-09-15 Aamin Dev

Acknowledgments

I would like to thank all the members of Panikecke[1] for being there for me all the time and supporting me in my personal life.

[1] `https://discord.gg/RURaVwRe`

Abstract

TODO: KEEP UPDATING ABSTRACT WHILE WRITING

This thesis introduces a novel method for the automated annotation of mathematical identifiers in scientific papers, leveraging Large Language Models (LLMs) such as GPT-3.5 and GPT-4. The approach addresses the challenges of co-reference resolution and formula grounding, traditionally handled by human annotators through costly and time-intensive procedures. Our study utilises a Math Identifier-oriented Grounding Annotation Tool (MioGatto), and explores the potential of integrating Part-of-Speech (POS) tagging and other technologies assisting in the process. The key component of this research is the development of a procedure for generating a dictionary of mathematical identifiers, contextualising their various meanings, and enabling the language model to select the most accurate definition based on the given context. This method demonstrates the impressive capability of LLMs to disambiguate meanings based on context, a vital task due to the inherently polysemous nature of mathematical identifiers. The preliminary results of Computational Natural Language Learning (CoNLL) scores of 82.36 position our approach as a potential game-changer in mathematical text annotation, offering a significant reduction in time and financial costs. The findings underscore the promise of the untapped potential of general purpose Large Language Models in specific mathematical language understanding.

Contents

1 Introduction

Mathematical formulas, integral to STEM papers, often present a challenge to readers due to the ambiguous use of variables, or identifiers, and their associated meanings or definitions. The constraints of the English and Greek alphabets, being finite, frequently lead to the reuse of the same identifiers with varying definitions contingent upon context. This ambiguity can be particularly daunting for those unfamiliar with the subject matter.

MioGatto, a Math Identifier-oriented Grounding Annotation Tool, was conceived to address this very issue. However, a new challenge arises the need for manual annotation of papers to define each identifier. This process is not only time-consuming but also resource-intensive, often requiring a full day for annotation depending on the paper's length.

The concept of grounding mathematical formulas [Asa+20] offers a promising solution. By automating this process, we can significantly expedite the annotation process. This thesis adopts a predominantly data-driven approach to develop an automation tool, which will be realised through the following stages:

1. **Detecting/Retrieving:** This phase entails the transformation of the LaTeX source into a machine-readable HTML/XML format using LaTeXML [1] [Gin+11].

2. **Dictionary Generation:** Leveraging LLMs, short text clustering techniques will be employed to construct a comprehensive dictionary of all mathematical identifiers.

3. **Association of Each Occurrence:** This step will involve associating every instance of a mathematical identifier with its corresponding definition, drawing inspiration from MathAlign [Ale+20].

1.1 Motivation and Problem

The process of annotating mathematical identifiers in scientific papers is a cornerstone for enhancing comprehension. Traditionally, this task has been performed manually, a method that, while effective, is fraught with challenges:

- **Time-Consuming:** Manual annotation is inherently labour-intensive, often requiring hours or even days for a single paper.

[1] https://math.nist.gov/~BMiller/LaTeXML/

- **Accessibility:** The expertise and resources required for manual annotation are not universally available, limiting its reach.
- **Cost Implications:** The adage "time is money" holds here. The extended hours spent on manual annotation translate to increased financial costs.

Given these constraints, there's a pressing need for a solution that's both efficient and universally accessible. Automation emerges as a promising alternative, with the potential to reduce annotation time from days to mere minutes, democratising access for a wider audience.

However, the path to automation is not without its hurdles. Traditional Natural Language Processing (NLP) techniques, such as part-of-speech tagging or the establishment of formal grammar, tend to oversimplify the problem. These methods often provide a generalised solution, covering only a fraction of the diverse challenges presented by mathematical annotations.

In recent years, LLMs have shown immense promise in various NLP tasks. Their capacity to understand and generate context-rich text suggests they could be pivotal in automating the annotation process, provided they are harnessed effectively.

1.2 Research Questions

The primary objective of this research is to explore the feasibility and effectiveness of using LLMs for the automation of mathematical identifier annotations in scientific papers. To guide this investigation, the following research questions have been formulated:

1. **Efficacy of LLMs:** How effective are LLMs, specifically GPT-3.5 [Ope23a] and GPT-4 [Ope23b] and some Open Source LLMs, in generating accurate annotations for mathematical identifiers compared to traditional manual methods?
2. **Contextual Understanding:** To what extent can LLMs disambiguate mathematical identifiers based on context, given the inherent polysemy of these identifiers?
3. **Coverage of Annotation:** What percentage of a scientific paper can LLMs effectively annotate?
4. **Accuracy concerning Ground Truth:** How closely do the annotations generated by LLMs align with the ground truth provided by manual annotations?
5. **Efficiency:** How does the automation process using LLMs impact the time required for annotating scientific papers, and what are the implications for cost savings?
6. **Limitations of Automation:** What are the potential pitfalls or limitations of using LLMs for this automation task?

These questions aim to provide a comprehensive understanding of the potential and challenges of using LLMs for the automation of mathematical identifier annotations.

1.3 Contributions

The journey of this research has led to several significant advancements in the realm of automated mathematical identifier annotations. The contributions of this thesis can be enumerated as follows:

1. **Integration with MioGatto:** Successfully incorporated GPT-based annotation capabilities into the MioGatto platform, enhancing its potential for automated paper annotations.
2. **Extensive Annotations:** Employed a range of LLMs, including GPT-3.5-turbo, GPT-3.5-turbo-16k, GPT-4, vicuna-33b [Zhe+23] [2], and StableBeluga2 [Mah+; Tou+23; Muk+23] [3], to annotate a curated set of 40 scientific papers. This extensive annotation process serves as a comprehensive evaluation of the capabilities of these models.
3. **Performance Evaluation:** Conducted a thorough evaluation and analysis of the annotation performances of GPT, vicuna-33b, and StableBeluga2 LLMs, providing insights into their strengths and limitations.
4. **Ground Truth Annotation:** Personally annotated a subset of papers to establish a ground truth, ensuring a reliable benchmark for evaluating the automated annotations.
5. **CoNLL Score Approximation:** Developed a novel formula to approximate the expected CoNLL [Pra+12] score of a paper when annotated using GPT, offering a predictive tool for assessing annotation quality.

These contributions not only advance the field of automated annotation but also lay the groundwork for future research endeavours in this domain.

1.4 Outline

This thesis is structured to provide a comprehensive understanding of the challenges, methodologies, and outcomes associated with automating mathematical identifier annotations using Large Language Models. The subsequent chapters are organised as follows:

[2] `https://huggingface.co/TheBloke/Vicuna-33B-1-3-SuperHOT-8K-GPTQ`

[3] `https://huggingface.co/TheBloke/StableBeluga2-70B-GPTQ`

1. **Introduction:** This chapter sets the stage by introducing the motivation, research questions, and contributions of the study, offering readers a contextual foundation for the subsequent chapters.
2. **Related Work:** While the concept explored in this thesis is novel, this chapter delves into the limited existing literature that shares thematic similarities, providing a backdrop against which the current research can be contrasted.
3. **Methods:** This chapter chronicles the journey of methodological exploration. It begins with initial attempts using traditional techniques like parts of speech tagging and transitions into the more successful strategies involving GPT, detailing the various prompts and markers employed to optimise results.
4. **Results:** Here, the empirical outcomes of the research are presented. The chapter elucidates the scores achieved by each LLM, the methodologies used to derive these scores, and the rationale behind selecting specific methods.
5. **Analysis:** This chapter delves deep into the interpretation of the results. It provides insights into the significance of the outcomes, and their correlation with other findings, and introduces the novel formula developed during the research.
6. **Future Works:** While the research has achieved significant milestones, this chapter outlines potential avenues for further exploration and improvement, such as model combinations or the incorporation of emerging techniques.
7. **Conclusion:** The thesis culminates with a synthesis of the research findings, answering the research questions posed at the outset and drawing conclusions on the study's implications and contributions.

This structured approach ensures a logical flow, guiding readers from the foundational concepts to the conclusions, and offering a holistic understanding of the research journey.

2 Related Work

Mathematical Language Processing (MLP) is an evolving domain with increasing attention being accorded to the intricacies of mathematical text understanding, annotation, and disambiguation. This chapter critically reviews prominent works, establishing the trajectory of the domain and elucidating its relevance to the present investigation.

2.1 Mathematical Text and Quantitative Reasoning

Meadows' research [MF22] places a spotlight on the crucial role of informal mathematical text in quantitative reasoning. The study advocates for the utility of transformer models, like GPT, in formula retrieval and solving math word problems. It asserts that the marriage of linguistic context with structured mathematical knowledge databases can profoundly improve comprehension.

2.2 Deciphering Mathematical Formulae

Within the ambit of the "Mathematical Language Processing (MLP) project" [PS14], a comprehensive analysis of the semantics underlying identifiers in mathematical formulae is undertaken. The paper juxtaposes traditional pattern-matching techniques with a novel MLP approach that employs part-of-speech (POS) tag-based distances to deduce identifier-definition probabilities. While the method illustrated the potential of improving user interactivity via tooltips showcasing probable definitions, its efficacy was curtailed in scenarios marked by the intricate natural language descriptions of mathematical identifiers.

2.3 Interpreting Symbolic Expressions

Grigore's work [GWK09] delves into the complexity of comprehending symbolic expressions present in mathematical narratives. The paper underscores the merit of harnessing linguistic context to fathom the semantics of these expressions, especially when confronted with symbol overloading— a limitation that deterred the adaptation of certain pre-existing tools.

2.4 Innovations in Data Annotation

Ding's investigation [Din+22] champions the role of GPT-3 in revolutionizing data annotation, positing a potential reduction in annotation overheads. The study conceptualizes the process as funnelling the expansive knowledge of GPT-3 into nimble networks apt for production landscapes.

2.5 The Advent of Automated Extraction and Annotation Systems

Schubotz's exploration [Sch+17] around the automated extraction of mathematical identifier definitions sets the stage for innovative annotation frameworks. The methodology and insights from this study have been instrumental in shaping our approach.

2.6 Harnessing Machine Reading for Mathematical Concepts

The treatise by Alexeeva et al. [Ale+20] underscores the transformative power of machine reading in extracting mathematical concepts. The study introduces a judicious rule-based strategy that seamlessly extracts LaTeX representations of formula identifiers and aligns them with their textual counterparts.

2.7 Laying the Groundwork for Mathematical Formulae

The pioneering work by Asakura et al. [Asa+20] delineates the importance of anchoring mathematical formulae. The authors champion the indispensable role of MLP in deciphering STEM manuscripts and introduce MioGatto, a cutting-edge annotation tool.

2.8 Evolving Pre-trained Models for Formula Insight

The innovative "MathBERT" framework [Pen+21] emerges as a pre-trained model fine-tuned for decoding mathematical formulas. Its training regimen, which couples formulas with their contextual narratives, attests to the centrality of context in deciphering mathematical formulae. While transformer models like MathBERT are of value, our research necessitated models that could craft comprehensive identifier description dictionaries. Moreover, the absence of BERT variants accommodating expansive context windows, spanning up to 1000 tokens, prompted us to adopt different strategies.

2.9 Redefining Annotation with Large Language Models

His seminal research [He+23] evaluates the efficacy of GPT-3.5 as a robust annotator. The paper interrogates if GPT-3.5's vast training could usher in a paradigm shift, replacing traditional crowdsourced annotators. Their findings accentuate the promise of deploying GPT for annotation endeavours.

2.10 Conclusion

The scholarship reviewed herein illuminates the myriad challenges and innovations characterising mathematical language processing. As the landscape evolves, these foundational works offer invaluable insights and lessons, setting the stage for the next wave of advancements in the field.

3 Methods

4 Results

5 Analysis

6 Future Works

7 Conclusion

Abbreviations

LLMs Large Language Models

NLP Natural Language Processing

POS Part-of-Speech

CoNLL Computational Natural Language Learning

MioGatto Math Identifier-oriented Grounding Annotation Tool

List of Figures

List of Tables

Bibliography

[Ale+20] M. Alexeeva, R. Sharp, M. A. Valenzuela-Escárcega, J. Kadowaki, A. Pyarelal, and C. Morrison. “MathAlign: Linking formula identifiers to their contextual natural language descriptions.” In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. 2020, pp. 2204–2212.

[Asa+20] T. Asakura, A. Greiner-Petter, A. Aizawa, and Y. Miyao. “Towards grounding of formulae.” In: *Proceedings of the First Workshop on Scholarly Document Processing*. 2020, pp. 138–147.

[Din+22] B. Ding, C. Qin, L. Liu, L. Bing, S. Joty, and B. Li. “Is gpt-3 a good data annotator?” In: *arXiv preprint arXiv:2212.10450* (2022).

[Gin+11] D. Ginev, H. Stamerjohanns, B. R. Miller, and M. Kohlhase. “The LATEXML daemon: Editable math on the collaborative web.” In: *Intelligent Computer Mathematics: 18th Symposium, Calculemus 2011, and 10th International Conference, MKM 2011, Bertinoro, Italy, July 18-23, 2011. Proceedings 4*. Springer. 2011, pp. 292–294.

[GWK09] M. Grigore, M. Wolska, and M. Kohlhase. “Towards context-based disambiguation of mathematical expressions.” In: *The joint conference of ASCM*. 2009.

[He+23] X. He, Z. Lin, Y. Gong, A. Jin, H. Zhang, C. Lin, J. Jiao, S. M. Yiu, N. Duan, W. Chen, et al. “Annollm: Making large language models to be better crowdsourced annotators.” In: *arXiv preprint arXiv:2303.16854* (2023).

[Mah+] D. Mahan, R. Carlow, L. Castricato, N. Cooper, and C. Laforte. *Stable Beluga models*.

[MF22] J. Meadows and A. Freitas. “A survey in mathematical language processing.” In: *arXiv preprint arXiv:2205.15231* (2022).

[Muk+23] S. Mukherjee, A. Mitra, G. Jawahar, S. Agarwal, H. Palangi, and A. Awadallah. “Orca: Progressive learning from complex explanation traces of gpt-4.” In: *arXiv preprint arXiv:2306.02707* (2023).

[Ope23a] OpenAI. *GPT-3.5*. 2023. URL: `https://platform.openai.com/docs/models` (visited on 09/15/2023).

[Ope23b] OpenAI. *GPT-4 Technical Report*. 2023. eprint: `arXiv:2303.08774`.

[Pen+21] S. Peng, K. Yuan, L. Gao, and Z. Tang. "Mathbert: A pre-trained model for mathematical formula understanding." In: *arXiv preprint arXiv:2105.00377* (2021).

[Pra+12] S. Pradhan, A. Moschitti, N. Xue, O. Uryupina, and Y. Zhang. "CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes." In: *Joint conference on EMNLP and CoNLL-shared task*. 2012, pp. 1–40.

[PS14] R. Pagael and M. Schubotz. "Mathematical language processing project." In: *arXiv preprint arXiv:1407.0167* (2014).

[Sch+17] M. Schubotz, L. Krämer, N. Meuschke, F. Hamborg, and B. Gipp. "Evaluating and improving the extraction of mathematical identifier definitions." In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11–14, 2017, Proceedings 8*. Springer. 2017, pp. 82–94.

[Tou+23] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. "Llama 2: Open foundation and fine-tuned chat models." In: *arXiv preprint arXiv:2307.09288* (2023).

[Zhe+23] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, et al. "Judging LLM-as-a-judge with MT-Bench and Chatbot Arena." In: *arXiv preprint arXiv:2306.05685* (2023).