MathFoldr: A Search Tool for Mathematics

Project Summary

Overview

The mathematical literature has grown far too vast for even an expert to keep track of it all, with over 120,000 math papers now published each year. Our infrastructure for organizing and communicating these results has not kept up. The ramifications are significant: wasted search time, duplication of research, and missed connections between fields. MathFoldr is a mathematics-specific search tool to address this problem.

The project consists of three work packages: data curation, knowledge extraction, and tool deployment. Data curation assembles clean, organized, and open-access corpora of mathematical papers and other mathematical text. Knowledge extraction uses deep learning and symbolic AI to build semantic representations of mathematical knowledge from these corpora, including word embeddings and ontologies. Tool deployment makes this knowledge available by a web browser-based front end, in particular allowing text based search.

The objectives are to:

1. Curate high-quality, up-to-date corpora of mathematical research papers and text.
2. Develop and deploy MathFoldr, a state-of-the-art search tool indexing over 500,000 pure mathematical papers.
3. Ensure MathFoldr serves over 5000 users a week, and promotes easier access to research across the mathematical community.

Intellectual Merit

The MathFoldr project advances knowledge in three distinct ways. First, through clean, organized corpora of mathematical text and papers it provides new raw resources and benchmarks for applying machine learning and natural language processing to math.

Second, it provides new artifacts that represent mathematical knowledge, including word embeddings (queryable semantic representations of mathematical concepts and papers), a new, open-access ontology of mathematical concepts, and a search engine. This provides a new way to explore the literature, improving speed and relevance of access, and opening the path for new, analysis of large scale patterns in mathematical knowledge.

Finally, it provides a blueprint for building practical scientific knowledge management tools based on machine learning and natural language processing, including both explainable symbolic methods and robust deep learning methods.

Broader Impacts

MathFoldr increases access to mathematical research, provides a resource for new mathematical discoveries, and lays a foundation for changing the publishing landscape. Most directly, it will mean that both math researchers and those seeking to apply math research can more quickly and effectively locate relevant results. This will increase the rate of mathematical discoveries, strengthen links between academia and industry, and lead to new technologies that impact everyday lives.

Project Description

Introduction

Math is not easy to search. A striking example comes from an episode reported by *Quanta Magazine* in November 2019 [1]. A group of physicists discovered a useful identity relating eigenvectors and eigenvalues, and did not know if it was novel. To check, they emailed a number of mathematicians, including Fields Medallist Terence Tao. Despite believing the result was "so short and simple – it should have been in textbooks already", Tao had not previously heard of it. This led to a paper submitted for publication and, soon after, the article in *Quanta*. In the weeks after the story emerged, more than three dozen previously published instances of the result were reported, dating back to 1934. *How can it be that even eminent mathematicians cannot find a widely published, basic result within their field of expertise?*

The simple answer is that the mathematical literature has grown far too vast for even an expert to keep track of it all. A recent analysis finds over 120,000 math papers published in 2017 alone, with this rate growing exponentially at 3% a year (see fig. 1)[2]. Our infrastructure for organizing and communicating these results has not kept up. The ramifications are significant: wasted search time, duplication of research, and missed connections between fields.

To address this, we are building MathFoldr: a search tool for mathematics. Our goals are to:

1. Curate high-quality, up-to-date corpora of mathematical research papers and text.
2. Develop and deploy MathFoldr, a state-of-the-art search tool indexing over 500,000 pure mathematical papers.
3. Ensure MathFoldr serves over 5000 users a week, and promotes easier access to research across the mathematical community.

Our approach is based on combining recent breakthroughs in transformer models from deep learning with symbolic and logical methods for analysing mathematical text.

To achieve these goals, we build on our previous work building natural language processing (NLP), domain-specific scientific search, and data integration tools, with key team members including:

- Dr Valeria de Paiva, an expert in NLP and mathematical logic with 20 years experience building commercial NLP and search tools at labs including Xerox PARC, Cuil, and Samsung.
- Dr Amalie Trewartha, lead developer of MatScholar [3] and COVIDScholar [4], similar field-specific academic literature curation and NLP tools.
- Antonin Delpeuch, lead contributor and maintainer of OpenRefine [5], an open source tool for cleaning and organizing messy data, under active development since 2010 and now available in 15 languages.

Our team in particular comprises senior members of the international category theory community, which provides both the domain expertise and community engagement required to ensure a product that serves the needs of mathematical researchers.

In achieving these goals, we will lay critically needed infrastructure for navigating mathematical knowledge, and a foundation for future discoveries. For example, our users will be able to locate literature by brief description of the underlying structural ideas, rather than requiring previous

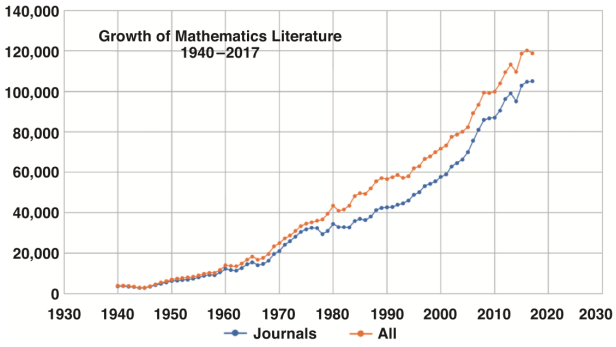

Figure 1: The growth of the mathematical literature, taken from Dunne [2]. New infrastructure is needed to help researchers navigate the vast quantities of new knowledge.

experience with the jargon of a subcommunity. Concretely, a sample goal is that search query "*classifying finite sets with invertible binary operation*" should return the Gorenstein–Lyons–Solomon monographs on the classification of finite simple groups.

More significantly, NLP tools have shown an impressive ability to not just connect researchers, but guide further breakthroughs, such as the discovery of novel thermoelectric materials [6]. By organizing mathematical knowledge, we hope to create new insights discoverable only through the accumulation of the knowledge present in thousands of papers. Indeed, existing attempts to organize the mathematical literature already yield such results, such as is N. Zeilberger's work showing a general correspondence between linear lambda terms with a context of free variables and rooted trivalent maps with a boundary of free edges [7]. These results, linking logic to geometrical insights on graphs, were only possible by matching two sequences in the Online Encyclopedia of Integer Sequences (OEIS). We anticipate that MathFoldr, which will add cutting-edge NLP techniques such as transformer models [8] to the organization of mathematical knowledge, will facilitate many more such discoveries.

Background

The dream of organizing humanity's mathematical knowledge is not new. Indeed, as far back as 1920, in a review of Dickson's three volume *History of the Theory of Numbers*, Lehmer writes [9]:

> There is the greatest need for just such a piece of work to promote efficiency among the professional workers in this field and to prevent them from wasting their time on problems that have already been adequately treated, and also to suggest other problems which still defy analysis.

What is new, however, is the opportunity afforded by the recent emergence of powerful NLP tools such as BERT [8] and GPT-3 [10], which for the first time have surpassed human performance on NLP benchmarks such as GLUE [11]. MathFoldr is a novel application of these tools to mathematical knowledge management, grounded in the successes of similar field-specific tools MatScholar [6] and COVIDScholar [4].

Mathematical Knowledge Management In 2014, the National Research Council of the US National Academies of Science and Engineering released the comprehensive study *Developing a 21st Century Global Library for Mathematics Research* [12], which argues persuasively for developing not just a search tool, but a robust platform that integrates and goes beyond the present functionalities of Google Scholar, the OEIS, WolframAlpha, libraries of formalized proofs, and other artifacts sharing mathematical knowledge. While critically important, the vastness of this project presents

a significant barrier to the development of useful tools. Our goal is to rapidly create an initial piece of infrastructure, which can then inform the design and construction of, and eventually be integrated with, future tools.

A major focus of previous work in this direction has been direct semantic representation of mathematical knowledge [13, 14, 15]. Such work has the extremely exciting potential to create error-free, formally-verified artifacts combining the work of many mathematicians. Yet while a fascinating and important problem, there is a justified diversity of approaches to formalization—not least the existing variety of proof assistants, with flagship systems including Agda, Coq, HOL, Isabelle, Lean, and Mizar. As the 2016 white paper arising from the Semantic Representation of Mathematical Knowledge Workshop shows, consensus around development of formal translation systems (eg. via the OpenTheory project [16]) or standards for mathematical writing has been difficult [9].

Moreover, even beyond designing automated translations between formal languages or gaining consensus around a standard, there remains the difficulty of performing the formal encoding itself. This requires individualized attention from mathematical authors themselves, who may not be interested in learning to use a proof assistant. One option that has been explored is the development of controlled natural languages, such as OMDoc/OpenMath, MathLang, and OntoMath. Yet gaining traction is still difficult here. For example, in 2017, the Formal Abstracts (FAbstracts) project proposed creating a controlled natural language to at the very least create human- and machine-readable summaries of papers, but even this initial step has been slow going [17].

Another approach to mathematical knowledge management is open-access, collaboratively-constructed databases of mathematical structures, such as

- The On-Line Encyclopedia of Integer Sequences, containing more than 335,000 sequences;
- The Wolfram Functions Site with 320,000 identities and about 10,000 visualizations;
- NIST's Digital Library of Mathematical Functions with about 35,000 identities.

These are all extremely useful, but not yet connected to a description of the body of mathematics that can help us search this body. Moreover, they again require the additional labor of mathematicians to record and curate structures in these databases.

Our approach is complementary to these formal and structured approaches, seeking to organize mathematics by directly extracting knowledge from standard, natural language, mathematical papers. This is, of course, the way mathematicians themselves learn mathematics. As such, it is easier to perform in a modular way, as it does not require any change on the part of authors. It is the same principle followed by the material scientists, chemists, and the COVID-19 researchers in building similar, successful NLP tools.

Deep learning Practical mathematical knowledge management using NLP is made possible by the advent of transformer models such as BERT [8] and GPT [10] over the past two years.

Transformer models fall within the paradigm of word embedding models [18, 19]. A word embedding model seeks to represent the meaning of each word as a vector. Semantic analysis of text can then be performed by various operations on these vectors. Transformer models provide a way to construct word embeddings that take into account the context in which words are used, so that the word 'space' would be represented by a different vector if it appeared in a topology rather than a linear algebra paper. This dramatically increases the utility of these models.

Such methods are already employed by Google and other major search tools. Yet while Google Scholar, Microsoft Academic, Semantic Scholar, and other tools already curate large collections of

mathematical papers and use modern deep learning methods to search over them, the specialized nature of mathematical text significantly hinders the efficacy of these tools. Indeed, these tools begin with models trained on corpora of standard English text, or at best scientific text. On the other hand, mathematical English is much like a foreign language masquerading as English.

In any corpus including non-mathematical text, key mathematical vocabulary is likely to be used with meaning significantly different to their technical mathematical usage. For example, words like 'group' and 'ring' have a variety of existing English meanings that bear little semantic resemblance to their mathematical definitions. As such, non-specialized deep learning models can make little sense of standard mathematical sentences such as 'every ring has an underlying abelian group'. Worse still, mathematical text often implictly employs standard mathematical facts of this sort, expecting the reader to infer use of common knowledge from an undergraduate curriculum.

This mathematical use of English means that off-the-shelf deep learning models are insufficient for practical, research-enhancing exploration of the mathematical literature. Such models are trained on corpora in which the incidence of the mathematical usage of the word 'group' makes up an miniscule fraction of all usages of the word 'group', and hence these models do not do well in representing meanings of mathematical terms. Instead, there is a need for specific corpora of mathematical texts to create mathematics-specific semantic representations. This domain specific training is known to outperform generic training significantly [20].

Scientific search tools Statistical machine learning systems such as transformer systems are extremely sophisticated and robust in the results they produce. However, they are for the most part opaque: one cannot tell why they produce the results they do and one cannot modify them in desired directions. By contrast, a symbolic AI system built on a knowledge graph is much more explainable and modifiable. If catastrophic bugs are discovered, we can try to modify the correlated inferences. Moreover, knowledge graphs and ontologies for mathematics can be used to improve the parsing and tagging of text used to train word embeddings.

Our recent work gives prototypes for building hybrid systems for natural language inference tuned for both performance [21] and explainability [22]. We design our MathFoldr system in the same way, keeping the best of both worlds: the precision and robustness of the purely machine learned system, as well as the explainability and the ability to correct mistakes afforded by the knowledge-graph system.

Beyond theoretical demonstrations, these principles are informed by the recent deployment of practical, discipline-specific AI-based search engines that have garnered everyday usage in their disciplines. Two examples are COVIDScholar and MatScholar, built by teams led by our lead developer Amalie Trewartha. These search engines serve over 4000 users each week, and lead directly to new results in their relevant fields [4, 3]. Similarly, in chemistry, field specific NLP tools have led to new insights about the patterns of the discovery of new chemical compounds [23].

Proposed Work

The construction of MathFoldr is divided into three distinct Work Packages (WP):

- WPI: Data Curation,
- WPII: Knowledge Extraction (WPIIa: Symbolic AI and WPIIb: Deep Learning),
- WPIII: Tool Deployment.

Data curation concerns the assembly and cleaning of text corpora, and pipelines for maintaining its relevance as new research emerges. Symbolic AI concerns to the development of mathematical

ontologies and logical relationships in text. Deep learning concerns the creation of word-embedding models using recent, state-of-the-art NLP techniques such as transformer models. Finally, tool deployment concerns the development of an online, freely accessible search engine for mathematics powered by these efforts.

Prototyping phase Each work package will be developed in two phases: first, a specialized prototype based on category theory. Second, a full mathematical tool. A prototyping phase allows us to quickly deploy a tool and serve a community we know well, to gather feedback on our methods without committing an overly large effort to certain approaches before fully understanding their end practicality. Category theory is particularly well-suited to this prototyping phase because:

- Category theory is a foundational part of mathematics, and its literature spans and incorporates significant technical vocabulary across algebra, topology, geometry, and logic.
- Category theory has a long history of high quality open access resources for community knowledge management, including open access journals such as the leading journal *Theory and Application of Categories*, the comprehensive and up-to-date research wiki nLab (15k+ pages), and the 30-year-old category theory mailing list (18k+ messages).
- The category theory community is very active online, for example through the Category Theory Community Server (1500+ users), providing a good resource for test deployment and rapid feedback.
- All members of our team have a shared technical background in category theory, allowing us to make quick assessments of the quality of our results.

WPI. Data Curation

Successful machine learning requires high quality data. The goal of the Data Curation Work Package is to create a pipeline for maintaining automated, up-to-date corpora of mathematical artifacts. These corpora will be cleaned and processed into standardized formats to provide easy use for user search, ontology extraction, training ML models, and benchmarking.

Our corpora come in two sorts: corpora of mathematical papers, and corpora of mathematical community communication, such as that from Wikipedia or MathOverflow. We also aim for two features: breadth, to facilitate search, and quality, to facilitate knowledge extraction.

When assembling corpora of papers, we will initially focus on abstracts and associated metadata. Indeed, parsing mathematical text itself presents difficulties for any computer-mediated system. Mathematical papers contain heavy use of formulas and diagrams, and simple symbols like a variable x can change in meaning multiple times even within the same page of text. Worse still, many texts are typeset in LaTeX, and are only available in pdf format, stripping out much of the syntactic clues that could help extract the structure of the text. Instead of developing tools to extract structure from full-text, our approach is to work primarily with abstracts, which are easier to source, written to be clear to a broad audience, more information dense, and have a lower incidence of LaTeX and symbolic mathematics. Moreover, abstracts by-and-large have more permissible copyright restrictions, allowing us to open-source more of our data.

This aligns with the approach of MatScholar and COVIDScholar, which have found that in a large collection, abstracts contain ample information for the extraction of new domain-specific insights [4, 6, 3].

As the project extends, we will improve our corpora by incorporating full text of papers, drawing on LaTeX source where available (eg. through the arXiv), and otherwise using recent OCR tools for extracting mathematical content, such as Mathpix.

Category theory corpora We have already built tools for constructed, cleaning, and automatically maintaining the following corpora, specializing in category theory as our prototype:

- arXiv papers categorized under math.CT (5786 entries)
- *Theory and Applications of Categories* papers (773 entries, 1995–present)
- *Applied Categorical Structures* papers (987 entries, 1993–present)
- *Cahiers de Topologie et Géométrie Différentielle Catégoriques* papers (587 entries, 1957–present)
- *n*Lab community wiki pages (15224 entries)

These corpora represent a significant core of published category theory research. To facilitate easy use of these corpora, we clean them, do part-of-speech tagging, and parse them. We assemble them into a single corpus for search and analysis, deduplicating entries based on title and DOIs, with the journal versions having priority over arXiv versions.

As part of the *n*Lab corpus, we not only extract the page text and metadata, but also exhibit the link structure between the pages, providing a directed graph of pages that can be used for further analysis.

Mathematics corpora We have also constructed a corpus of mathematics articles on Wikipedia (9217 entries, identified through crawling various lists of math-related articles).

At present, there are no existing, high quality, available corpora of mathematical papers suitable for our purposes. That said, led by organizations such as OpenAIRE and the AMS, there are now many initiatives building databases and open science graphs, which seek to organize research artifacts and the relationships between them, as well as provide fair and open access to these graphs. Different collections may focus on different goals and characteristics, with major examples including: Google Scholar, Scopus, Web of Science, Mathematical Reviews Database, Microsoft Academic Graph, FREYA PID Graph, Research Graph Foundation, and Semantic Scholar. These knowledge graphs provide structured way for us to explore and extract papers of interest.

To the full extent permitted under copyright and licencing agreements, our data sets will be published to facilitate future use. Open access, high quality datasets for testing and benchmarking has in particular been an important first step in advancing novel or domain-specific NLP tasks. Our *n*Lab, Wikipedia, and arXiv corpora, among others, will be prepared for these purposes.

WPI Deliverables:

- Robust software pipeline for maintaining, for the purposes of MathFoldr, an up-to-date, complete index of mathematics papers.
- Clean, high-quality open-access datasets of mathematical text for knowledge extraction.
- Clean, high-quality open-access datasets of mathematical papers for benchmarking.

WPIIa. Symbolic AI

The corpora we construct contain a wealth of mathematical knowledge. To assist with extracting this knowledge, we will also build and maintain an open-access ontology of research-level mathematical concepts.

An ontology, or knowledge graph, for mathematics is a list of mathematical objects, their defining properties, and the pairwise relationships between them. For example, our ontology contains the concepts 'function' and 'continuous function', together with the relationship that a continuous function is a type of function. While an ontology for mathematics has been previously discussed, there is no existing open-access ontology for research mathematics [24]. The closest existing tool is OntoMathPro [25, 26], a project based at Kazan Federal University, Russia. That project, however, has recently focussed on education and secondary school mathematics.

Nonetheless, Wikidata, a well-established open-access tool for building ontologies and linked data (over 91 million data items) [27], has via its links with Wikipedia good coverage of basic mathematical concepts. Our approach for creating a mathematical ontology is to improve the coverage of research level mathematics inside Wikidata. Advantages of this approach include the robustness of the Wikidata platform, its existing familiarity with the Semantic Web community, and its commitment to permanent, open-access, and collaboratively created knowledge. In particular, Wikidata editing is public and permissionless, allowing specialist mathematicians to improve the coverage of our ontology in their area of expertise as necessary.

We have already significantly enhanced the coverage of category theoretic concepts inside Wikidata (now 3864 concepts), again intending to take advantage of the fact that category theory was conceived as a language for bridging between mathematical disciplines, and thus with the belief that it will prove a good source of structure for other areas too.

To create entries for our ontology, we draw from three sources:

- Entries of online encyclopedia sources, like the nLab, Wikipedia, and PlanetMath.
- Author-identified keywords from our arXiv and category theory journal corpora.
- Automated keyword extraction from text.

The tool OpenRefine, developed by team member Antonin Delpeuch [5], is used to manage the complexity of merging these diverse sources of data, and the export to Wikidata. Challenges of the merging process include:

- Ambiguity: Many mathematical concepts, such as 'sheaf' (nothing to do with wheat), 'product' (not for sale), and 'injection' (does not prevent COVID-19), are ambiguous words, that must be disambiguated by context.
- Level of granularity: Some concepts are better understood as built from others. For example, while a 'fully faithful functor' is a central notion in category theory, it is best simply understood as a functor that is both 'full' and 'faithful'. We address this by the 'wikification' of a text, a process used by IBM Watson, to transform text into bite-size dictionary entries.
- Significance: Research is an ongoing process, and not every concept defined joins our community collection of mathematical objects. To ensure concepts meet an appropriate threshhold of significance, we cross-check these sources against each other, and use methods like PageRank on the link graph of online encyclopedias to rank concepts in terms of importance

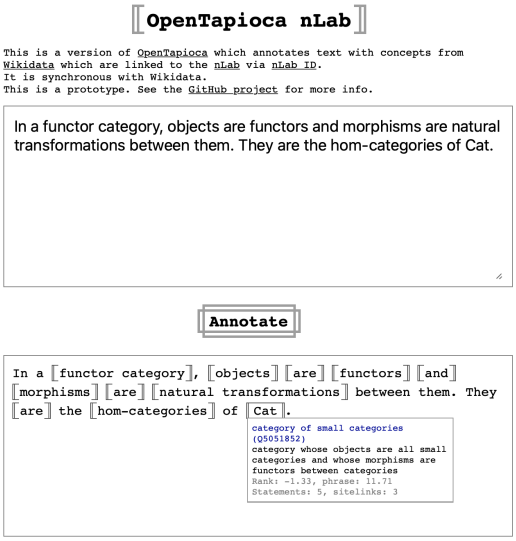

Figure 2: Our prototype named-entity recognizer (NER) tool is available online [29]. Observe how it identifies mathematical concepts such as 'the category of categories' within text, providing a link with the corresponding concept entry in Wikidata.

As a result, a small amount of manual fine-tuning has been important as a final quality control.

Note that the construction of our ontology is language independent: the entries are tied to mathematical concepts themselves, not the English words used to express them. Through projects such as team member Tim Hosgood's Maths Dictionary Project [28], translations of mathematical terms are also added to Wikidata.

This ontology can then be used for a number of purposes, beginning with named-entity recognition (NER) in text. We have built a prototype of an NER tool that identifies category theoretic concepts in text, based on team member Antonin Delpeuch's work on the OpenTapioca project (see fig. 2). The prototype is available online [29].

Further applications of the ontology are to improve tokenizing and parsing of text for machine learning, and for search query expansion.

WPIIa Deliverables:

- Open-access ontology of mathematics embedded in the Wikidata Semantic Web.
- NER toolkit for identifying mathematical entities and their relationships in natural language text.

WPIIb. Deep Learning

We use deep learning to construct word and document embeddings that represent the semantics of our text. These embeddings will drive our search tool MathFoldr, with uses including automated keyword extraction, document similarity ranking, classification, and query expansion. Moreover, we will provide tools for exploring the embeddings themselves, which we expect to provide new insights into connections between concepts and papers (see fig. 3).

A challenge for producing useful word embeddings for domain-specific text is simply the volume of available training data. To overcome this, we begin by restricting to a vocabulary relevant to the domain. To generate our vocabulary for our category theory specific system, for example, we combine our set of category theory concepts from our ontology with unsupervised tokenizing methods that pick up high frequency expressions, like byte-pair encoding [30], that are trained on our category theory corpora.

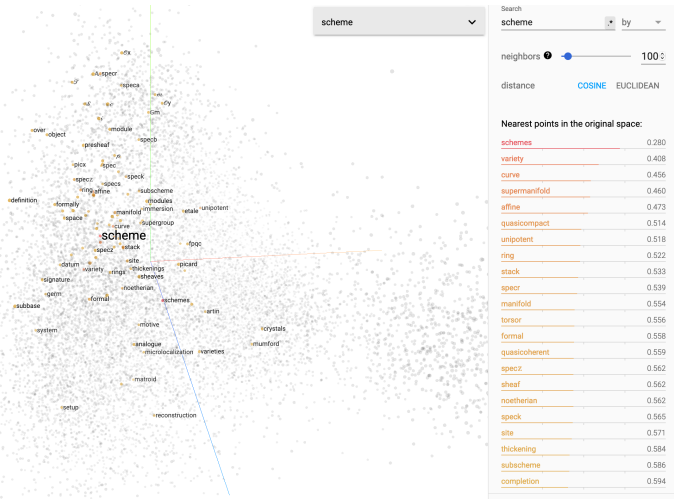

Figure 3: A visual exploration of our category theory word embedding, highlighting semantic neighbors of the concept 'scheme'. Word embeddings provide a new way to explore relationships between mathematical ideas, and provide the backbone of modern search techniques.

We then produce our word embeddings in two steps. First, we pretrain on a larger corpus of mathematical text, such as our Wikipedia corpus, in which words like 'category' will largely, but not always, be reserved for use in the mathematical sense. Second, we then fine-tune using a category theory specific corpus. Previous work has shown such a two-step approach yields high-quality word embeddings even from a small domain-specific training corpus [31, 32].

Recently constructed open-source libraries like the HuggingFace `Transformers` library provide a fast and ready toolkit for performing these tasks [33]. This has yielded encouraging results so far, with word embeddings that reproduce simple statements of similarity such as 'a group is more like a ring than a function', and analogies such as 'pullback : pushout :: equalizer : coequalizer'.

To extract additional information about mathematical concepts, we will also use methods like node2vec for constructing word embeddings from graphs, like the link graphs of nLab or Wikipedia, or our ontology.

From these word embeddings, we will then construct document embeddings by experimenting with a number of methods, including transformer models. This allows us to compare entire passages of text, such as abstracts, by similarity.

WPIIb Deliverables:

- Word embeddings for every concept in our ontology.
- Document embeddings for every paper in our text corpora.

WPIII. MathFoldr: An Integrated Tool

Users will engage with our text corpora and the extracted semantic models through MathFoldr, a browser-based front-end, much like Google Scholar or COVIDScholar.

To build this front-end, we will use Vespa.ai, a framework for low latency computation over large datasets, designed to facilitate the question-answering tasks that are the core of MathFoldr. Again, this follows our previous work on COVIDScholar (fig. 4).

In Work Packages IIa and IIb we construct two ways to analyze the semantics of mathematical text: logical relationships expressed in our ontology, and statistical relationships expressed in our word embeddings. Both are powerful ways to analyze mathematical text, and we expect different queries to be more amenable to one method or the other. To give the user control over which

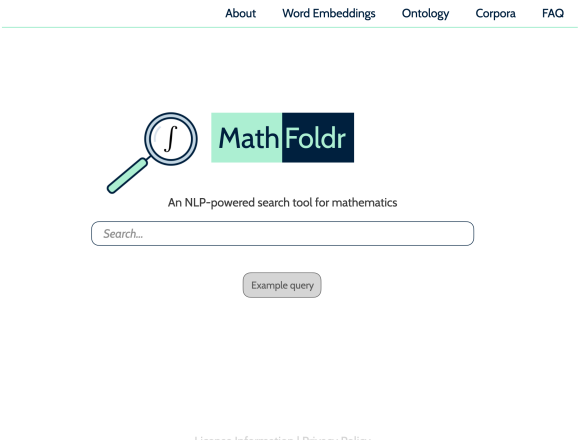

Figure 4: MathFoldr will be made available to users via a simple and easy-to-use web interface.

method suits their inquiries best, we will build on our previous xplainNLI and Hy-NLI prototypes which utilize both methods, and return information on which method was used [22, 21] (see fig. 5).

We will also provide an interface for examining our semantic embeddings directly, as in fig. 3. In other fields, publicly-available domain-specific word embeddings have been exploited by researchers to achieve state of the art performance on machine learning tasks, for example the predicting of synthesizable materials or materials properties from chemical formulas alone [34, 35].

Internally, evaluation will take place continuously through a collection of test queries and answers, as well as a regression testsuite, as we have pioneered in previous work [36]. This testsuite will begin with simple, factoid queries, ("*which TAC paper discusses 'bicycles' as a mathematical concept?*"), before progressing to increasingly complex multi-step queries which need to broken down into semantic pieces and transformed. An example of a multi-step query is the aforementioned "*classify finite sets with invertible binary operation*", in which "finite sets with invertible binary operation" should first be transformed into "finite groups", and the system re-queried with "*classify finite groups*" to arrive at the desired Gorenstein–Lyons–Solomon papers.

Our ultimate criterion for success, however, is the number of users we help access math research. Thus the deployment strategy is also critical. Our initial prototype will focus on serving the category theory community, and we will engage initial users through the Category Theory Community Server, a Zulip chat room with over 1500 participants. This will provide timely user feedback about whether our tools are serving the needs of active researchers. Once we have a working service and stable user

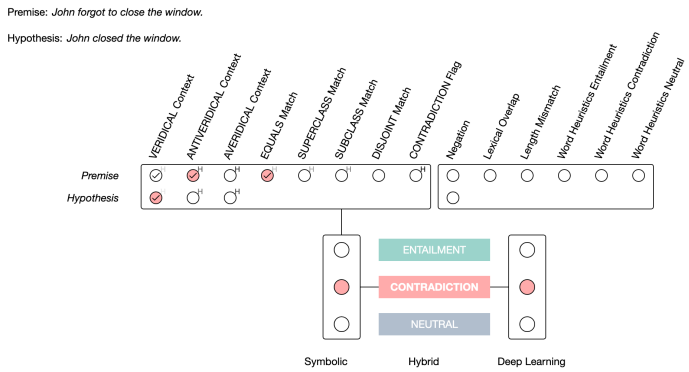

Figure 5: The xplainNLI tool provides a prototype for integrating symbolic and statistical NLP systems.

base, we will set up our own forum for communication with users, and expand through adjacent research communities by word-of-mouth and targeted advertising.

To help measure the success of the tool, we will retain basic, anonymized data regarding number of users and number of search queries over time.

WPIII Deliverables:

- A state-of-the-art search tool for math papers, incorporating both symbolic and statistical methods.
- Online interface for exploring word and document embeddings.
- Community portal for engaging with MathFoldr users.

Broader Impacts

MathFoldr creates new information infrastructure that increase access to mathematical research, provide a resource for new discoveries, and lay a foundation for changing the publishing landscape.

Scientists spend 23% of their time searching the literature [37]; the figure is likely similar for mathematicians. By providing a tool to more rapidly and reliably find relevant papers, we will save mathematicians a significant amount of their research time, leading to a more efficient research process.

Perhaps even more significantly, very often present day searches do not bear fruit. For example, a recent paper in the Applied Category Theory 2020 conference counts nine different introductions of the concept of a *cartesian monoid*, most in well-established literature [38]. The result is that many ideas have to be rediscovered over and over, and opportunities for collaboration between different researchers and subfields are missed. MathFoldr, as a semantic search tool, will identify similar concepts even if they use divergent terminology, thus making these new connections. Again, we believe this will increase the efficiency of mathematical research.

The difficulty of search further forms a barrier to application of mathematical ideas and construction of technologies based on them. By providing easier access to the literature, we believe that we will speed up technology transfer from universities, and strengthen links between academia and industry.

Search also forms a barrier, albeit one of very many, to diversity and inclusion in our field. With the present difficulty of locating results in the literature, professional networks—the ability to ask an expert colleague or friend for a literature reference—become a critical resource. This sort of network is often more available to those who already feel more included in the mathematical community, and thus conveys advantages to such people. While there is much more work to do with regard to diversity, inclusion, and equity, we hope that MathFoldr will contribute to ensuring that the ability to effectively search for mathematical knowledge is available to all.

In assembling digital artifacts collating mathematical knowledge, we also hope to open up new avenues and methods for research. Our cleaned corpora and word embedding models of mathematical knowledge will permit for the first time semantic meta-analysis of research papers, allowing discovery of large-scale patterns in mathematical knowledge. In chemistry this has led to insights about the chemical research process [23], while in materials science this has led to the discovery of thermoelectrics [6].

Many of the techniques we are developing are applicable for scientific knowledge management in general, with applications to other sciences and social sciences. All code will be open source, and to the full extent allowed by publisher agreements, all datasets too.

Finally, MathFoldr will approach the construction of a Global Digital Library of Mathematics from a new angle. Future development of the tool will incorporate complementary semantic search strategies based on formal encoding of mathematics, through integration with proof assistants and similar methods from the Mathematical Knowledge Management community. Our vision is to provide infrastructure for direct publishing of formally encoded mathematical statements and proofs.

Long-Term Maintainence

The project will be based at Topos Institute. Topos Institute is a new, nonprofit mathematics and computer science research institute, dedicated to building public infrastructure for integrating information based on advanced mathematics. It is located in Berkeley, California, and opens doors in January 2021, having raised over $1.6M in start-up funding from government, industry, and philanthropic sources.

MathFoldr is the first phase of a long-term Topos research focus on computer-assisted mathematical knowledge manangement, a program that will later tie in aspects of formal representation of mathematics as described in the Background section above. The requested grant funding is to kick-start development of a practical, widely-used search tool. Once built, this tool will continue to be maintained by Topos personnel and funds as part of our mathematical knowledge manangement program.

Personnel And Roles

Our team brings together significant, complementary experience and skills balanced across all aspects of the MathFoldr project, from data science to NLP to deep learning to software engineering to web-development and community manangement. Critically, our broad range of mathematical expertise is well-suited to creating a tool that serves all of mathematical sciences.

Relevant experience and roles are as follows.

Valeria de Paiva (PI) Dr Valeria de Paiva is an incoming Principal Research Scientist at Topos Institute. Dr de Paiva leads the MathFoldr project overall, specializing in the symbolic AI (WPIIa) and tool deployment (WPIII) aspects.

Dr de Paiva brings to this role 20 years of experience building commercial NLP and search tools at industry labs including Xerox PARC, Cuil, Nuance Communications, and Samsung Research America. During this time, her work includes creation of the Portuguese OpenWordNet-PT, used by Google Translate as its Portuguese lexical resource, developing systems for evaluating the performance of NLP systems [36], a patent for a new approach to content detection in documents, and hybrid systems for explainable natural language inference [22, 21].

De dr Paiva holds a PhD in Mathematical Logic (Cambridge); is an Honorary Research Fellow at the School of Computer Science, University of Birmingham (previously tenured Assistant Professor); a Member of the Industry Advisory Board of the Natural Language Processing Masters' Program at UC Santa Cruz; and a Member of the Scientific Advisory Committee of the Institute for Logic, Language and Information at the University of Amsterdam. She is a member of the editoral boards of journals *Logical Methods in Computer Science*, *Theory and Applications of Categories*, *Logica Universalis*, and *Compositionality*.

Brendan Fong (co-PI) Dr Brendan Fong is CEO and Research Scientist at Topos Institute. Dr Fong leads the deep learning work package (WPIIa), brings domain experience in category theory,

and does project management.

His work on MathFoldr is informed by previous work on category theoretic and structural approaches to deep learning [39, 40].He also brings significant project management and community building expertise, as co-founder and CEO of Topos Institute, a founding executive editor of the open-access journal *Compositionality*, organizer of the inauguaral Annual International Conference on Applied Category Theory, now in its fourth year, and inaugural organizer and steering board member of The Adjoint School, an annual community building research school in applied category theory. This expertise is critical to managing our distributed team, and to deploying a tool that will be used by the wider mathematical community.

Dr Fong holds a PhD in Computer Science (Oxford), and previously completed postdoctoral appointments at the Department of Mathematics, MIT, and the Department of Electrical and Systems Engineering, University of Pennsylvania. He is coauthor of the popular textbook *An Invitation to Applied Category Theory* [41], published by Cambridge University Press.

Evan Patterson (co-PI) Dr Evan Patterson is a Research Scientist at Topos Institute. Dr Patterson leads the data curation work package (WPI).

Dr Patterson's main role is to build and maintain the data pipeline. He brings to this significant experience in data science and software engineering, as key personnel on the Data Science Ontology project and its application to semantic models of data science code [42, 43], and as lead developer of the AlgebraicJulia project and Catlab framework, which provide novel approaches to scientific computing based on applied category theory [44].

Dr Patterson holds a PhD in Statistics (Stanford).

Amalie Trewartha (Lead Developer) Dr Amalie Trewartha is a Postdoctoral Scholar at Lawrence Berkeley National Laboratory, and academic NLP consultant on the MathFoldr project. Dr Trewartha leads the development team, building on previous experience leading the development teams of MatScholar [3] and COVIDScholar [4].

Informed by building these related tools, she brings expertise on recent practical breakthroughs in deep learning, how they apply to scientific knowledge management, and the latest software libraries and platforms for building user-facing tools.

Dr Trewartha holds a PhD in theoretical and computational physics (Adelaide).

Tim Hosgood (Software Engineer) Dr Tim Hosgood is a Postdoctoral Scholar at the Centre for Quantum Technologies, Singapore, and research affiliate at Topos Institute. Dr Hosgood is the main software engineer on the project. His tasks include working under the direction of Dr Trewatha and Dr Fong to experiment with and tune hyper-parameters for deep learning models, and build and maintain the front-end web application for MathFoldr.

Dr Hosgood holds a PhD in Complex Geometry (Aix-Marseille), and has done extensive work on mathematical translations, leading to the MathsDictionary project and Wikidata integration [28].

Antonin Delpeuch (Software Engineer) Antonin Delpeuch is a final year PhD candidate in computer science at the University of Oxford. On the MathFoldr project, he is the software engineer responsible for tasks in WPIIa (Symbolic AI), involving extracting ontologies from our corpora and integrating them with Wikidata.

This is a direct application of his work as Software Engineer with OpenRefine, an open-source tool for cleaning, transforming, and extending messy data, and lead developer of OpenTapioca, an open-source named-entity linking system for Wikidata [45].

Timeline

This project is a 24 month project, with each work package split into category theory (CT) and general mathematical stages. We anticipate the category theory MathFoldr prototype will be fully functional and serving the category theory community by Month 9 of the project. The category theory phase of the data curation work package (WPI) is already complete. We plan the following schedule of work:

Month	0	3	6	9	12	15	18	21	24
WPI: Data Curation			WPI (Math)						
WPIIa: Symbolic AI	WPIIa (CT)			WPIIa (Math)					
WPIIb: Deep Learning	WPIIb (CT)			WPIIb (Math)					
WPIII: Tool Deployment	WPIII (CT)				WPIII (Math)				

Results From Prior NSF Support

No PI or co-PI has received NSF funding in the past five years.

References Cited

[1] N. Wolchover, "Neutrinos lead to unexpected discovery in basic math," *Quanta Magazine*, 2019. [Online]. Available: https://www.quantamagazine.org/neutrinos-lead-to-unexpected-discovery-in-basic-math-20191113/

[2] E. Dunne, "Looking at the mathematics literature," *Notices of the American Mathematical Society*, vol. 66, no. 2, pp. 227–230, 2019. [Online]. Available: https://www.ams.org/journals/notices/201902/rnoti-p227.pdf

[3] L. Weston, V. Tshitoyan, J. Dagdelen, O. Kononova, A. Trewartha, K. A. Persson, G. Ceder, and A. Jain, "Named entity recognition and normalization applied to large-scale information extraction from the materials science literature," *Journal of chemical information and modeling*, vol. 59, no. 9, pp. 3692–3702, 2019.

[4] A. Trewartha, J. Dagdelen, H. Huo, K. Cruse, Z. Wang, T. He, A. Subramanian, Y. Fei, B. Justus, K. Persson, and G. Ceder, "Covidscholar: An automated covid-19 research aggregation and analysis platform," 2020.

[5] A. Delpeuch, "A survey of openrefine reconciliation services," *arXiv preprint arXiv:1906.08092*, 2019.

[6] V. Tshitoyan, J. Dagdelen, L. Weston, A. Dunn, Z. Rong, O. Kononova, K. A. Persson, G. Ceder, and A. Jain, "Unsupervised word embeddings capture latent knowledge from materials science literature," *Nature*, vol. 571, no. 7763, pp. 95–98, 2019.

[7] N. Zeilberger, "Linear lambda terms as invariants of rooted trivalent maps," *CoRR*, vol. abs/1512.06751, 2015. [Online]. Available: http://arxiv.org/abs/1512.06751

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://www.aclweb.org/anthology/N19-1423

[9] P. Ion, M. Trott, E. Weisstein, and F. Wiedijk, "White paper of the semantic representation of mathematical knowledge workshop," *Wolfram Foundation*, 2016. [Online]. Available: https://www.wolframfoundation.org/programs/SemanticWorkshopWhitePaper.pdf

[10] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.

[11] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," 2019, in the Proceedings of ICLR.

[12] N. R. Council *et al.*, *Developing a 21st century global library for mathematics research*. National Academies Press, 2014.

[13] J. Carette, W. M. Farmer, and R. O'Connor, "Mathscheme: project description," in *International Conference on Intelligent Computer Mathematics*. Springer, 2011, pp. 287–288.

[14] J. Carette, W. M. Farmer, Y. Sharoda, K. Bercic, M. Kohlhase, D. Müller, and F. Rabe, "The space of mathematical software systems," 2020. [Online]. Available: https://arxiv.org/abs/2002.04955

[15] M. Kohlhase, "Mathematical knowledge management: transcending the one-brain-barrier with theory graphs," *European Mathematical Society (EMS) Newsletter*, vol. 92, pp. 22–27, 2014.

[16] J. Hurd, "The opentheory standard theory library," pp. 177–191. [Online]. Available: http://www.gilith.com/papers

[17] T. Hales. (2020) Formal abstracts. [Online]. Available: https://formalabstracts.github.io

[18] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model," *Journal of machine learning research*, vol. 3, no. Feb, pp. 1137–1155, 2003.

[19] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[20] M. C. Iordan, T. Giallanza, C. T. Ellis, N. Beckage, and J. D. Cohen, "Context matters: Recovering human semantic structure from machine learning analysis of large-scale text corpora," *arXiv preprint arXiv:1910.06954*, 2019.

[21] A.-L. Kalouli, R. Crouch, and V. de Paiva, "Hy-nli: a hybrid system for natural language inference," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 5235–5249.

[22] A.-L. Kalouli, R. Sevastjanova, V. de Paiva, R. Crouch, and M. El-Assady, "Xplainli: Explainable natural language inference through visual analytics," in *Proceedings of the 28th International Conference on Computational Linguistics: System Demonstrations*, 2020, pp. 48–52.

[23] E. J. Llanos, W. Leal, D. H. Luu, J. Jost, P. F. Stadler, and G. Restrepo, "Exploration of the chemical space and its three historical regimes," *Proceedings of the National Academy of Sciences*, vol. 116, no. 26, pp. 12 660–12 665, 2019.

[24] C. Lange, "Ontologies and languages for representing mathematical knowledge on the semantic web," *Semantic Web*, vol. 4, no. 2, pp. 119–158, 2013.

[25] Intelligent Search Systems and Semantic Technologies Laboratory at Kazan Federal University. (2020) Ontomath pro. [Online]. Available: http://ontomathpro.org

[26] O. A. Nevzorova, N. Zhiltsov, A. Kirillovich, and E. Lipachev, "Ontomath pro ontology: a linked data hub for mathematics," in *International Conference on Knowledge Engineering and the Semantic Web*. Springer, 2014, pp. 105–119.

[27] T. W. Foundation. (2020) Wikidata. [Online]. Available: https://www.wikidata.org/

[28] T. Hosgood. (2020) Maths dictionary. [Online]. Available: https://thosgood.com/maths-dictionary/

[29] A. Delpeuch. (2020) Opentapioca nlab. [Online]. Available: https://nlab.opentapioca.org/

[30] P. Gage, "A new algorithm for data compression," *C Users Journal*, vol. 12, no. 2, pp. 23–38, 1994.

[31] I. Beltagy, K. Lo, and A. Cohan, "Scibert: A pretrained language model for scientific text," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 3606–3611.

[32] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, and D. Zhi, "Med-bert: pre-trained contextualized embeddings on large-scale structured electronic health records for disease prediction," *arXiv preprint arXiv:2005.12833*, 2020.

[33] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Huggingface's transformers: State-of-the-art natural language processing," *ArXiv*, pp. arXiv–1910, 2019.

[34] M. Aykol, V. Ishwar Hegde, L. Hung, S. Suram, P. Herring, C. Wolverton, and J. Hummelshøj, "Network analysis of synthesizable materials discovery," *Nature Communications*, vol. 10, 05 2019.

[35] R. E. A. Goodall and A. A. Lee, "Predicting materials properties without crystal structure: Deep representation learning from stoichiometry," 2020.

[36] V. De Paiva and T. H. King, "Designing testsuites for grammar-based systems in applications," in *Coling 2008: Proceedings of the workshop on Grammar Engineering Across Frameworks*, 2008, pp. 49–56.

[37] K. E. Hubbard and S. D. Dunbar, "Perceptions of scientific research literature and strategies for reading papers depend on academic career stage," *PloS one*, vol. 12, no. 12, p. e0189753, 2017. [Online]. Available: https://doi.org/10.1371/journal.pone.0189753

[38] R. Statman, "Products in a category with only one object," in *Electronic Proceedings in Theoretical Computer Science 328*, D. Spivak and J. Vicary, Eds. [Online]. Available: https://cgi.cse.unsw.edu.au/~eptcs/paper.cgi?ACT2020:10

[39] B. Fong, D. Spivak, and R. Tuyéras, "Backprop as functor: A compositional perspective on supervised learning," in *2019 34th Annual ACM/IEEE Symposium on Logic in Computer Science (LICS)*. IEEE, 2019, pp. 1–13.

[40] B. Fong and M. Johnson, "Lenses and learners," in *Proceedings of the Eighth International Workshop on Bidirectional Transformations (Bx 2019)*, J. Cheney and H.-S. Ko, Eds., 2019.

[41] B. Fong and D. I. Spivak, *An Invitation to Applied Category Theory*. CUP, 2019.

[42] E. Patterson, R. McBurney, H. Schmidt, I. Baldini, A. Mojsilović, and K. R. Varshney, "Dataflow representation of data analyses: Toward a platform for collaborative data science," *IBM Journal of Research and Development*, vol. 61, no. 6, pp. 9–1, 2017.

[43] E. Patterson, I. Baldini, A. Mojsilovic, and K. R. Varshney, "Teaching machines to understand data science code by semantic enrichment of dataflow graphs," *arXiv preprint arXiv:1807.05691*, 2018.

[44] M. Halter, E. Patterson, A. Baas, and J. Fairbanks, "Compositional scientific computing with catlab and semanticmodels," 2020.

[45] A. Delpeuch, "Opentapioca: Lightweight entity linking for wikidata," *arXiv preprint arXiv:1904.09131*, 2019.