Finding Old Answers to New Math Questions: The ARQMath Lab at CLEF 2020

Behrooz Mansouri[1(✉)], Anurag Agarwal[1], Douglas Oard[2], and Richard Zanibbi[1]

[1] Rochester Institute of Technology, Rochester, NY, USA
{bm3302,axasma,rxzvcs}@rit.edu
[2] University of Maryland, College Park, MD, USA
oard@umd.edu

Abstract. The ARQMath Lab at CLEF 2020 considers the problem of finding answers to *new* mathematical questions among posted answers on a community question answering site (Math Stack Exchange). Queries are question postings held out from the test collection, each containing both text and at least one formula. We expect this to be a challenging task, as both math and text may be needed to find relevant answer posts. While several models have been proposed for text question answering, math question answering is in an earlier stage of development. To advance math-aware search and mathematical question answering systems, we will create a standard test collection for researchers to use for benchmarking. ARQMath will also include a formula retrieval sub-task: individual formulas from question posts are used to locate formulas in earlier answer posts, with relevance determined by narrative fields created based on the original question. We will use these narrative fields to explore diverse information needs for formula search (e.g., alternative notation, applications in specific fields or definition).

Keywords: Community question answering · Formula retrieval · Mathematical Information Retrieval · Math-aware search

1 Introduction

In a recent study, Mansouri et al. found that 20% of mathematical queries in a general-purpose search engine were expressed as well-formed questions, a rate ten times higher than that for all queries submitted [7]. Results such as these and the presence of Community Question Answering sites such as Math Stack Exchange[1] (MSE) and Math Overflow [11] suggest that there is a great public interest in finding answers to mathematical questions posed in natural language, using *both* text and mathematical notation. Related to this, there has also been increasing

[1] https://math.stackexchange.com.

J. M. Jose et al. (Eds.): ECIR 2020, LNCS 12036, pp. 564–571, 2020.
https://doi.org/10.1007/978-3-030-45442-5_73

work on math retrieval and math question answering in both the Information Retrieval (IR) and Natural Language Processing (NLP) communities.

In light of this growing interest, we are organizing a new lab at the Conference and Labs of the Evaluation Forum (CLEF) on Answer Retrieval for Questions about Math (ARQMath).[2] Using the mathematics and free text in posts from Math Stack Exchange, participating systems will be given a question, and asked to return a ranked list of potential answers. Relevance will be determined by how well the returned posts answer the provided question. Through this task we will explore leveraging math notation together with text to improve the quality of retrieval results. This is one case of what we generically call math-aware information retrieval, in which the focus is on leveraging the ability to process mathematical notation to enhance, rather than to replace, other information retrieval techniques. We will also include a query-by-example task on formula retrieval in which relevance will be determined by the degree to which a retrieved formula is useful for the searcher's intended purpose.

Question answering (QA) was among the earliest target applications for Artificial Intelligence. Techniques for answering one specific type of mathematical question, automated theorem proving, date back to 1956 when Newell and Simon [9] introduced the *logic theorist* that proved theorems in symbolic logic. Three years later, they introduced the General Problem Solver [10], which attempted to mimic students' behavior in discovering proofs. For math QA, an important recent development was the work of Ling et al. [6], who solved algebraic word problems by generating answer rationales and human-readable mathematical expressions that derive a final answer along with a description of the method used to solve the problem. Kushman et al. [5] presented an approach for automatically learning to solve algebra word problems expressed in text *and* math by defining a joint log-linear distribution over full systems of equations and aligning their variables and numbers to the problem text.

More recently, machine learning has been applied to answering a broader range of questions. For example, the *Arosti* system of Clark et al. [3] achieved a score of over 90% on the (non-diagram) multiple choice portion of the New York Regents 8th Grade Science Exam, the first such system to pass this test. Natural Language Processing has, in recent years, focused on Reading Comprehension QA tasks requiring an answer to be located in a single document. One recent shared task that involved processing mathematical notation was Task 10 at SemEval 2019 [4], which provided a question set derived from the Math-SAT (Scholastic Achievement Test) practice exams that included 2,778 training questions and 1,082 test questions from three major categories: Closed Algebra, Open Algebra and Geometry. A majority of the questions were multiple choice, with only a minority having a numeric answer.

Within the IR community, much of the recent work on QA has focused on Community Question Answering (CQA), with the goal of augmenting human talent by finding earlier answers that people have already given that can serve as answers to newly asked questions. One important line of work enabled by CQA

[2] https://www.cs.rit.edu/~dprl/ARQMath.

is that it becomes possible not just to search directly for potential answers, but also to search for prior questions that could lead to potential answers. Because CQA systems include social media features such as voting for answer quality, some types of non-text features can also be leveraged.

Math-Aware Information Retrieval. The existing evaluation resources for math-aware information retrieval were initially developed over a five-year period at the National Institute of Informatics (NII) Testbeds and Community for Information Access Research (at NTCIR-10 [1], NTCIR-11 [2] and NTCIR-12 [12]). The NTCIR Mathematical Information Retrieval (MathIR) tasks developed evaluation methods and allowed participating teams to establish baselines for both "text + math" queries and isolated formula retrieval. NTCIR-12 ultimately made use of two collections, one a set of arXiv papers from physics that is split into paragraph-sized documents, and the other a set of articles from English Wikipedia. Interest in these tasks is global; at the NTCIR-12 MathIR task, for example, there were participating groups from around the world, including Europe (Czech Republic, Germany), Asia (China, India, Japan) and North America (Canada, USA). The NTCIR-12 isolated formula retrieval test collection was also later used by participants for the 2016 Competition on Recognition of Online Handwritten Mathematical Expressions (CROHME) [8] at the International Conference on Frontiers in Handwriting Recognition (ICFHR).

ARQMath Goals. The ARQMath lab will provide an opportunity to push mathematical question answering in a new direction, where informal language is frequently used, and where answers provided by a community are selected and

Table 1. Example queries and results for question answering and formula retrieval.

Question Answering	Formula Retrieval
Question	Query
I've spent the better part of this day trying to show from first principles that this sequence tends to 1. Could anyone give me an idea of how I can approach this problem? $\lim_{n\to+\infty} n^{\frac{1}{n}}$	$\lim_{n\to+\infty} n^{\frac{1}{n}}$
Relevant	Relevant
You can use AM $\geq$ GM. $\frac{1+1+\cdots+1+\sqrt{n}+\sqrt{n}}{n} \geq n^{1/n} \geq 1$ $1-\frac{2}{n}+\frac{2}{\sqrt{n}} \geq n^{1/n} \geq 1$	$\lim_{n\to\infty} \sqrt[n]{n}$
Non-Relevant	Non-Relevant
If you just want to show it converges, then the partial sums are increasing but the whole series is bounded above by $1+\int_1^{\infty} \frac{1}{x^2}dx = 2$	$\sum_{k=1} \frac{1}{k^2} = \frac{\pi^2}{6}$

ranked rather than generated. One goal is to produce test collections; a second is to drive innovation in evaluation methods; and a third is to drive innovation in the development of math-aware information retrieval systems.

2 Overview of Tasks

There will be two tasks in the first year: (1) a question answering task (QA), where systems are provided a question post from MSE and then return a ranked list of answer posts and (2) an isolated formula retrieval task. Table 1 illustrates these two tasks, and we provide details about each task below.

Finding Answers to Math Questions (Main Task). For the QA task, at least 50 questions from MSE will be sampled, with the requirement that each question contains both text and at least one formula. Participants will have the option to run queries using only the text or math portions in each question, or to use both math and text. We will ask participants to label each run with which of these conditions they chose. One challenge inherent in this design is that the expressive power of text and formulas are sometimes complementary; so although all topics will include both text and formula(s), some may be better suited to text-based or math-based retrieval. We plan to accommodate this by reporting results for all participants that are averaged over three topic sets: (1) all topics, (2) topics for which the assessor believes the text alone to be an adequate characterization of the topic, and (3) topics for which the assessor believes the formula(s) alone to be an adequate characterization of the topic.

Formula Search (Secondary Task). In this task individual formulas are used as queries, and systems return a ranked list of similar and/or related formulas. As with the NTCIR-12 Wikipedia Formula Browsing Task, this task has the goal of fostering development of component technology for computing math similarity. We envision two improvements over what was done at NTCIR: further developing the concept of "formula relevance" and creating a collection with a larger number of formula queries (NTCIR-12 has only 20 formula queries + 20 modified versions of the same formulas with wildcards added).

Each formula query will be a single formula extracted from a question used in the main task. For each query, annotators will write a short human-readable narrative field – not available to participating systems – that reflects their understanding of the type(s) of similarity the person who asked the original question would have found useful. This may include alternative notation, simplification, specialization, or applications in specific fields, and we expect to extend those categories further based on suggestions from participating teams. Because participating systems won't have access to this narrative field in for their "standard condition" run, we expect this task to support research on diversity ranking for formula retrieval. We are also aiming to have at least 50 formula queries in the first year, with the intent to expand both query sets in subsequent years.

3 The Math Stack Exchange Collection

Our collection will be comprised of question and answer postings from Math Stack Exchange (MSE). These postings are freely available as data dumps from the Internet Archive. At the time of this writing, there are 1.1 million questions.

Figure 1 (left) shows the distribution of the number of formulas per question post. For query development, only the 45% of questions containing at least one formula,[3] will be considered. As Fig. 1 (right) shows, question production on Math Stack Exchange has been fairly steady in recent years. We plan to release an annotated version of the complete Math Stack Exchange data dump containing questions and answers produced through December 2018 as the test collection, holding out questions submitted in 2019 for query development.

We plan to stratify the questions by predicted difficulty (in the opinion of the collection developers) so as to avoid, for example, having too many common questions that are nearly identical to questions from 2018 or earlier. Each question will include a unique ID, the asker-entered title for their question, the asker-entered body of the question, three formats for formulas found in the body of the question (LaTeX, Presentation MathML, and Content Math ML; see below), a list containing any edits to the question that were subsequently done by the asker, one or more asker-entered type tags (e.g., "calculus"), and comments on the question entered by other users, the average score assigned by other MSE users to the question (which one might interpret as a measure of how "good" a question it is), and the asker's reputation score (which is estimated from scores given to their prior questions). Only the first four of these (ID, title, body, formulas) will be used for the standard condition, but the additional features will be available for use in contrastive runs.

We will use open source tools such as LaTeXML,[4] to label and convert LaTeX formulas from posts and convert them to XML markup, including both Presentation (appearance-based) and Content (semantic) MathML. We will perform this extraction centrally and distribute the extracted formulas as standoff annotations with references to the location of the formula in each XML question or answer post. Converting LaTeX to Presentation MathML is a straightforward transformation between formula appearance representations (i.e., symbols on writing lines). Producing Content MathML from LaTeX requires inference, and is thus potentially errorful. However, Content MathML supports a higher level of abstraction by representing operator structure explicitly. Centralizing this conversion will remove one possible source of variation, but conversion scripts will also be made available to participants who wish to experiment with extended conversion capabilities.

[3] In Math Stack Exchange formulas almost invariably appear between two '$' signs in LaTeX notation (e.g., $a+b=c$).

[4] https://dlmf.nist.gov/LaTeXML/.

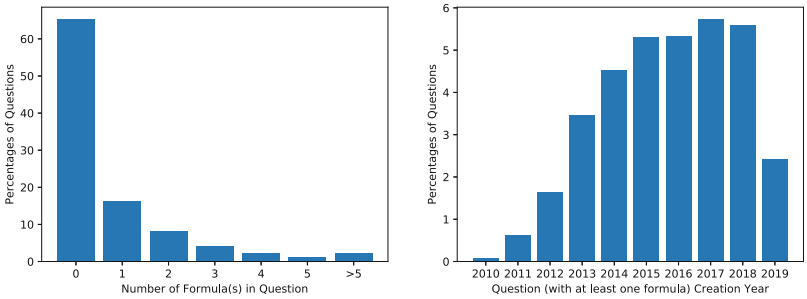

Fig. 1. Formulas in math stack exchange question postings. **Left:** formula counts for questions. **Right:** creation years for questions containing at least one formula.

4 Relevance Judgments

For both the QA and formula retrieval tasks, manual and automatic runs will be allowed. For each topic, the top-N (e.g., top-20) results from each participant run, along with additional manual runs conducted by the organizers, will be pooled. We will trade off pool depth and number of topics based on the available annotation resources.

Because specialized mathematical knowledge may be needed for assessment, the pooled documents will be assessed for relevance by volunteers from participating teams, augmented by assessors hired by the organizers. Evaluation will be performed using a web-based system (e.g., Sepia[5]). Assessors for the main task will be asked to identify relevant answers using pools from the main task. Assessors for the formula retrieval task will work with merged pools from both the formula retrieval task and (where appropriate for the question) from the main task to identify similar formulas. Most pools will be judged by a single assessor, but some will be dual-assessed to observe annotator agreement. For the formula retrieval task, queries will be selected for dual annotation using stratified random sampling so as to cover a broad range of similarity types. We also plan for some limited experimentation with alternative annotation strategies (e.g., additionally annotating the most useful parts of a relevant answer, or annotating the preference order between relevant answers) with the goal of informing evaluation design in future years.

We will use trec_eval to compute ranked document retrieval measures for each run for both tasks, with inferred Average Precision (infAP) as the standard measure for comparing systems. This choice of infAP is intended to provide results that can support future experimentation with the test collection by future systems that did not contribute to the judgment pools, but we will provide the full range of trec_eval measures to participating systems for use when different evaluation measures could provide additional insights.

[5] https://code.google.com/archive/p/sepia/.

5 Conclusion

The ARQMath lab at CLEF 2020 is the first in a three-year sequence of labs through which we aim to push the state of the art in evaluation design for math-aware IR, and in which we seek to support the development and ultimate deployment of new techniques for that task. We have chosen to focus on CQA, using Math Stack Exchange, both because that task models an actual employment scenario, and because the scale of the available collection is sufficient to also support the development of a second test collection more narrowly focused on formula retrieval. Math is, of course, only one example of structured notation, and we might reasonably hope to one day leverage similar ideas in other domains that also frequently use specialized notation, such as biology or chemistry.

Acknowledgements. This material is based upon work supported by the Alfred P. Sloan Foundation under Grant No. G-2017-9827 and the National Science Foundation (USA) under Grant No. IIS-1717997.

References

1. Aizawa, A., Kohlhase, M., Ounis, I.: NTCIR-10 math pilot task overview. In: NTCIR (2013)
2. Aizawa, A., Kohlhase, M., Ounis, I., Schubotz, M.: NTCIR-11 Math-2 task overview. In: NTCIR (2014)
3. Clark, P., et al.: From 'F' to 'A' on the NY regents science exams: An overview of the Aristo. arXiv preprint arXiv:1909.01958 (2019)
4. Hopkins, M., Le Bras, R., Petrescu-Prahova, C., Stanovsky, G., Hajishirzi, H., Koncel-Kedziorski, R.: SemEval-2019 Task 10: Math question answering. In: Proceedings of the 13th International Workshop on Semantic Evaluation (2019)
5. Kushman, N., Artzi, Y., Zettlemoyer, L., Barzilay, R.: Learning to automatically solve algebra word problems. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (2014)
6. Ling, W., Yogatama, D., Dyer, C., Blunsom, P.: Program induction by rationale generation: Learning to solve and explain algebraic word problems. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (2017)
7. Mansouri, B., Zanibbi, R., Oard, D.W.: Characterizing searches for mathematical concepts. In: Proceedings of JCDL (2019)
8. Mouchère, H., Viard-Gaudin, C., Zanibbi, R., Garain, U.: ICFHR 2016 CROHME: competition on recognition of online handwritten mathematical expressions. In: Proceedings of ICFHR (2016)
9. Newell, A., Simon, H.: The logic theory machine-A complex information processing system. IRE Trans. Inf. Theor. **2**, 61–79 (1956)
10. Newell, A., Shaw, J.C., Simon, H.A.: Report on a general problem solving program. In: IFIP Congress (1959)

11. Tausczik, Y.R., Kittur, A., Kraut, R.E.: Collaborative problem solving: a study of MathOverflow. In: CSCW 2014, pp. 355–367 (2014)
12. Zanibbi, R., Aizawa, A., Kohlhase, M., Ounis, I., Topic, G., Davila, K.: NTCIR-12 MathIR task overview. In: NTCIR (2016)