# Advancing Math-Aware Search: The ARQMath-3 Lab at CLEF 2022

Behrooz Mansouri[1], Anurag Agarwal[1]
Douglas W. Oard[2], and Richard Zanibbi[1]

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

**Abstract.** ARQMath-3 is the third edition of the Answer Retrieval for Questions on Math lab at CLEF. In addition to the two main tasks from previous years, an interesting new pilot task will also be run. The main tasks include: (1) Answer Retrieval, returning posted answers to mathematical questions taken from a community question answering site (Math Stack Exchange (MSE)), and (2) Formula Retrieval, returning formulas and their associated question/answer posts in response to a query formula taken from a question. The previous ARQMath labs created a large new test collection, new evaluation protocols for formula retrieval, and established baselines for both main tasks. This year we will pilot a new *open domain* question answering task as Task 3, where questions from Task 1 may be answered using passages from documents from outside of the ARQMath collection, and/or that are generated automatically.

**Keywords:** Community Question Answering, Formula Retrieval, Mathematical Information Retrieval, Math-Aware Search, Open Domain QA

## 1 Introduction

Effective question answering systems for math would be valuable for math Community Question Answering (CQA) forums, and more broadly for the Web at large. Community Question Answering sites for mathematics such as Math Stack Exchange[3] (MSE) and Math Overflow [12] are widely-used resources. This indicates that there is great interest in finding answers to mathematical questions posed in natural language, using *both* text and mathematical notation.

The ARQMath lab [6, 17] was established to support research into retrieval models that incorporate mathematical notation. With a number of Math Information Retrieval (MIR) systems having been introduced recently [1, 5, 8, 9, 15], a standard MIR benchmark is essential for understanding the behavior of retrieval models and implementations. To that end, the previous ARQMath collections produced a new collection, assessment protocols, parsing and evaluation tools,

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

**Table 1.** Example ARQMath-2 Queries and Results.

| **Question Answering (Task 1)** | **Formula Retrieval (Task 2)** |
|---|---|
| QUESTION (TOPIC A.220)<br>I'm having a difficult time understanding how to give a combinatorics proof of the identity<br>$$\sum_{k=0}^{n} \binom{x+k}{k} = \binom{x+n+1}{n}$$ | FORMULA QUERY (TOPIC B.220) I'm having a difficult time understanding how to give a combinatorics proof of the identity<br>$$\sum_{k=0}^{n} \binom{x+k}{k} = \binom{x+n+1}{n}$$ |
| RELEVANT<br>The right side is the number of ways of choosing $n$ elements from $\{1, 2, 3, ..., 2n\}$. The number of ways of choosing n elements from that set that starting with $1, 2, ..., n-k$ and not containing $n-k+1$ is $\binom{n+k-1}{k}$. | RELEVANT<br>Question: prove by induction on n+m the combinatoric identity:<br>$$\sum_{k=0}^{n} \binom{m+k}{k} = \binom{m+n+1}{n}$$<br>I've tried to do on both $n$ and $m$<br>... |
| NON-RELEVANT<br>Hint: Find a combinatorial argument for which<br>$$\sum_{k=0}^{n} \binom{n}{k}\binom{k}{2} = \binom{n}{2} 2^{n-2}$$<br>then use the previous identity.<br>... | NON-RELEVANT<br>Hint<br>$$\sum_{k=0}^{n} \binom{n}{k} x^k = (1+x)^n$$<br>Integrate twice both rhs and lhs with respect to x and when finished, plug $x = 1$ in your result. |

and a benchmark containing over 140 annotated topics for each of two tasks: math question answer retrieval, and formula retrieval.[4]

ARQMath is the first shared-task evaluation on question answering for math. Using formulae and text in posts from Math Stack Exchange (MSE), participating systems are given a question and asked to return potential answers. Relevance is determined by how well returned posts answer the provided question. The left column of Table 1 shows an example topic from Task 1 (ARQMath-2 Topic A.220), showing one answer assessed as relevant, and another assessed as non-relevant. The goal of Task 2 is retrieving relevant formulae for a formula query taken from a question post (e.g., shown in blue in Table 1), where relevance is determined *in-context*, based on the question post for the query formula and the question/answer posts in which retrieved formulae appears. This task is illustrated in the right column of Table 1 (ARQMath-2 Topic B.220).

Before ARQMath, early benchmarks for math-aware search were developed through the National Institute of Informatics (NII) Testbeds and Community for Information Access Research (at NTCIR-10 [2], NTCIR-11 [3] and NTCIR-12 [16]). The Mathematical Information Retrieval (MathIR) evaluations at NTCIR included tasks for both structured "text + math" queries and isolated formula retrieval, using collections created from arXiv and Wikipedia. ARQMath complements the NTCIR test collections by introducing additional test collections

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

based on naturally occurring questions, by assessing formula relevance in context, and by substantially increasing the number of topics.

In this paper, we summarize existing data and tools, the second edition of the ARQMath lab, and planned changes for ARQMath-3. Briefly, ARQMath-3 will reuse the ARQMath collection, which consists of MSE posts from 2010 to 2018. The most substantial change is the addition of a new open domain question answering task as a pilot task.[5] Unlike our ongoing answer retrieval task, in which the goal is to return existing answers, for the new open domain question answering task systems may retrieve and/or generate answers, such as was done previously for the Question Answering tracks at TREC-8 [13] through TREC-13 [14], and the Conversational Question Answering Challenge [10].

## 2 ARQMath Tasks

ARQMath-3 will include the same two tasks as ARQMath-1 and -2, and it introduces a pilot task on open domain question answering for math, where external knowledge sources may be used to find, filter, and even generate answers.

### 2.1 Task 1: Answer Retrieval

The primary task for the ARQMath labs is answer retrieval, in which participants are presented with a question posted on MSE **after** 2018, and are asked to return a ranked list of up to 1,000 answers from **prior years** (2010-2018). In each lab, the participating teams ranked answer posts for 100 topics. In ARQMath-1 77 and in ARQMath-2 71 topics were assessed and used for the evaluation. In ARQMath-1, for primary runs the pooling depth was 50 and 20 for other runs. In ARQMath-2, these values were adjusted to 45 and 15 because the number of runs doubled, and participating teams also nearly doubled.

Table 2 summarizes the graded relevance scale used for assessment. System results ('runs') were evaluated using the nDCG′ measure (read as "nDCG-prime"), introduced by Sakai [11] as the primary measure. nDCG′ is simply normalized Discounted Cumulative Gain (nDCG), but with unjudged documents removed before scoring. Two additional measures, mAP′ and P′@10, were also reported using binarized relevance judgments. In both labs, participants were allowed to submit up to 5 runs, with at least one designated as primary.

### 2.2 Task 2: Formula Retrieval

The ARQMath formula retrieval task has some similarity to the Wikipedia Formula Browsing Task from NTCIR-12 [16]. In the NTCIR-12 task, given a single query formula, similar formulae in a collection were to be returned. The NTCIR-12 formula browsing task test collection had only 20 formula queries (plus 20 modified versions with wildcards added), whereas in ARQMath-1, 74 queries

[5] As proposed by Vít Novotný at CLEF 2021.

(45 for evaluation + 29 additional for future training) and in ARQMath-2, 70 queries (58 for evaluation + 12 additional) were assessed.

ARQMath has introduced two innovations for formula search evaluation. First, in ARQMath, relevance is decided by context, whereas in NTCIR-12, formula queries were compared by assessors with retrieved formula instances without consideration of the context for either. Second, in NTCIR-12 systems could receive credit for finding formula instances, whereas in ARQMath systems receive credit for finding *visually distinct* formulae. In other words, an NTCIR-12 system that found identical formulae in two different documents and returned that formula twice would get credit twice, whereas an ARQMath system would receive credit only once for each visually distinct formula. Deduplication of visually identical/near-identical formulae was done using Symbol Layout Trees produced from Presentation MathML by Tangent-S [4] where possible, and by comparing LaTeX strings otherwise. In ARQMath-1, this clustering was done *post hoc* on submitted runs; for ARQMath-2 this clustering was done *a priori* on the full collection and shared with participating teams. In ARQMath-3 the cluster ids will again be provided with the collection. For efficiency reasons, we have limited the number of instances of any visually distinct formula that were assessed to 5 in ARQMath-1 and -2, and expect the same for ARQMath-3.

The relevance of a *visually distinct* formula is defined by the maximum relevance for any of its pooled instances, based on the associated question/answer post for each instance. Table 2 summarizes the graded relevance scale used for assessment. Here relevance is interpreted as the likelihood of a retrieved formula being *associated with* information that helps answer the question in which a formula query appeared. There is an important difference in relevance assessment for Task 2 in ARQMath-1 and -2: although the relevance scale shown in Table 2 was unchanged between ARQMath-1 and ARQMath-2, we did change how the table was interpreted for ARQMath-2. In ARQMath-1, only the context in the question post associated with the query formula was considered, with ARQMath-1 assessors instructed to mark exact matches as relevant. This was changed when we noticed that visually identical formulas at times had no bearing on the information represented by a query formula within its associated question post. As an example, we can have two visually identical formulae, but where one represents operations on sets, and the other operations on integers.

### 2.3 Task 3 (Pilot): Open Domain QA for Math

In this pilot task, participants are given Task 1 topics and asked to provide a single answer for each question that must not exceed a fixed maximum length. Unlike Task 1 where answers are taken from the MSE collection, answers may be produced using any technique, and any available knowledge sources (with the exception of MSE answers from 2019 to the present). For example, responses may be a new machine-generated response, a single passage or complete answer post from MSE or another CQA platform (e.g., Math Overflow), or some combination of generated and existing content. For relevance assessment, responses from open domain QA systems will be included in the Task 1 pools. Rankings obtained

**Table 2.** Relevance Scores, Ratings, and Definitions for Tasks 1 and 2.

| Task 1: Question Answering | | |
|---|---|---|
| Score | Rating | Definition |
| 3 | High | Sufficient to answer the complete question on its own |
| 2 | Medium | Provides some path towards the solution. This path might come from clarifying the question, or identifying steps towards a solution |
| 1 | Low | Provides information that could be useful for finding or interpreting an answer, or interpreting the question |
| 0 | Not Relevant | Provides no information pertinent to the question or its answers. A post that restates the question without providing any new information is considered non-relevant |
| **Task 2: Formula Retrieval** | | |
| Score | Rating | Definition |
| 3 | High | Just as good as finding an exact match to the query formula would be |
| 2 | Medium | Useful but not as good as the original formula would be |
| 1 | Low | There is some chance of finding something useful |
| 0 | Not Relevant | Not expected to be useful |

from the Task 1 relevance measures will be compared with rankings produced by automated answer quality measures (e.g., derived from BLEU [7]) to assess whether these measures may be used reliably to evaluate future systems. Task 3 answers will be further assessed separately for aspects such as fluency, and whether answers appear to be human-generated or machine-generated (for this, we may include MSE posts alongside Task 3 submissions).

## 3 The ARQMath Test Collection

ARQMath uses Math Stack Exchange (MSE) as its collection, which is freely available through the Internet Archive. The ARQMath collection contains MSE posts published from 2010 to 2018, with a total of 1 million questions and 1.4 million answers. In ARQMath-1, posts from 2019, and in ARQMath-2 posts from 2020 were used for topic construction. For ARQMath-3, posts from 2021 will be used.[6] Topic questions must contain at least one formula; with this constraint, 89,905 questions are available for ARQMath-3 topic development.

**Topics.** In previous ARQMath labs, topics were annotated with three categories: complexity, dependency, and type. In ARQMath-1, more than half of the Task 1 topics were categorized as questions seeking a proof. We aimed to better balance across question categories in ARQMath-2, but when category combinations are considered the Task 1 topic set still exhibited considerable skew towards a few combinations. In ARQMath-3, we introduce a fourth category, *parts*, which indicates whether a topic question calls for an answer that has a single part, or whether it contains sub-questions that each call for answers.[7] We do see that different systems seem to be doing better on different ARQMath-1 and ARQMath-2 question categories, so in ARQMath-3 we continue to aim to balance the topic selection process across combinations of question categories as best we can, including the new *parts* category.

[6] from a September 7, 2021 snapshot.
[7] This is based on a suggestion at CLEF 2021 from Frank Tompa.

**Formulae.** In the Internet Archive version of the collection, formulae appear between two '$' or '$$' signs, or inside a 'math-container' tag. For ARQMath, all posts (and all MSE comments on those posts) have been processed to extract formulae, assigning a unique identifier to each formula instance. Each formula is provided in three encodings: (a) as LaTeX strings, (b) as (appearance-based) Presentation MathML, and (c) as (operator tree) Content MathML.

The open source LaTeXML[8] tool we use for converting LaTeX to MathML fails for some MSE formulae. Moreover, producing Content MathML from LaTeX requires inference, and is thus potentially errorful. As a result, the coverage of Presentation MathML for detected formulae in the ARQMath-1 collection was 92%, and the coverage for Content MathML was 90%. For ARQMath-2, after LaTeXML updates the error rate was reduced to less than one percent for both representations, reducing the need for participating systems to fall back to using LaTeX. However, there are some remaining MathML encoding issues and formula parsing/clustering failures in the ARQMath-2 collection that we plan to correct in ARQMath-3.

**Files.** As with any CQA task, the ARQMath collection contains more than just question and answer posts. We distribute the collection as four main files:

- **Posts.** The post file contains a unique identifier for each question or answer post, along with information such as creation date and creator. Question posts contain a title and a body (with the body being the question), while answer posts have a body and the unique identifier of the associated question.
- **Comments.** Any post can have one or more comments, each having a unique id and the unique identifier of the associated post.
- **Votes.** This file records positive and negative votes for posts, along with additional annotations such as 'offensive' or 'spam.'
- **Users.** Posters of questions and answers have a unique User ID and a reputation score.

## 4 Conclusion

For ARQMath-3, we will continue our focus on answering math questions (Tasks 1 and 3), with formula search as the secondary task (Task 2). For question answering, we are adding a new pilot task for open domain QA (Task 3) alongside the answer retrieval task (Task 1). A single Math Stack Exchange (MSE) collection will again be used. This is both because MSE models an actual usage scenario, and because we expect that reusing MSE will facilitate training and refinement of increasingly capable systems.

**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. We thank Vítek Novotný for providing details for Task 3.

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

# Bibliography


[1] Ahmed S, Davila K, Setlur S, Govindaraju V (2021) Equation attention relationship network (EARN): A geometric deep metric framework for learning similar math expression embedding. In: 2020 25th International Conference on Pattern Recognition (ICPR), IEEE, pp 6282–6289

[2] Aizawa A, Kohlhase M, Ounis I (2013) NTCIR-10 math pilot task overview. In: Proceedings of the 10th NTCIR Conference, pp 654–661

[3] Aizawa A, Kohlhase M, Ounis I, Schubotz M (2014) NTCIR-11 math-2 task overview. In: Proceedings of the 11th NTCIR Conference, pp 88–98

[4] Davila K, Zanibbi R (2017) Layout and semantics: Combining representations for mathematical formula search. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 1165–1168

[5] Mansouri B, Zanibbi R, Oard DW (2021) Learning to rank for mathematical formula retrieval. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, p 952–961

[6] Mansouri B, Zanibbi R, Oard DW, Agarwal A (2021) Overview of ARQMath-2 (2021): Second CLEF lab on answer retrieval for questions on math. In: Proceedings of the 11th International Conference of the CLEF Association (CLEF 2020), Springer, pp 215–238

[7] Papineni K, Roukos S, Ward T, Zhu WJ (2002) Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp 311–318

[8] Peng S, Yuan K, Gao L, Tang Z (2021) MathBERT: A pre-trained model for mathematical formula understanding. arXiv preprint arXiv:210500377

[9] Pfahler L, Morik K (2020) Semantic search in millions of equations. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp 135–143

[10] Reddy S, Chen D, Manning CD (2019) CoQA: A conversational question answering challenge. Transactions of the Association for Computational Linguistics 7:249–266, DOI 10.1162/tacl_a_00266

[11] Sakai T (2007) Alternatives to bpref. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 71–78

[12] Tausczik YR, Kittur A, Kraut RE (2014) Collaborative problem solving: A study of MathOverflow. In: Proceedings of the 17th ACM conference on Computer Supported Cooperative Work & Social Computing, pp 355–367

[13] Voorhees EM (1999) The TREC-8 question answering track report. In: Proceedings of the Eighth Text REtrieval Conference (TREC 8), pp 77–82

[14] Voorhees EM (2005) Overview of the TREC 2004 question answering track. In: Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004), pp 52–62

[15] Wang Z, Lan A, Baraniuk R (2021) Mathematical formula representation via tree embeddings. Online: URL https://people.umass.edu/~andrewlan/papers/preprint-forte.pdf
[16] Zanibbi R, Aizawa A, Kohlhase M, Ounis I, Topic G, Davila K (2016) NTCIR-12 MathIR task overview. In: Proceedings of the 12th NTCIR Conference, pp 299–308
[17] Zanibbi R, Oard DW, Agarwal A, Mansouri B (2020) Overview of ARQ-Math 2020: CLEF lab on answer retrieval for questions on math. In: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, pp 169–193