Overview of ARQMath-2 (2021): Second CLEF Lab on Answer Retrieval for Questions on Math (Working Notes Version)

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

[1]Rochester Institute of Technology, NY, USA
[2]University of Maryland, College Park, USA

Abstract

This paper provides an overview of the second year of the Answer Retrieval for Questions on Math (ARQMath-2) lab, run as part of CLEF 2021. The goal of ARQMath is to advance techniques for mathematical information retrieval, in particular retrieving answers to mathematical questions (Task 1), and formula retrieval (Task 2). Eleven groups participated in ARQMath-2, submitting 36 runs for Task 1 and 17 runs for Task 2. The results suggest that some combination of experience with the task design and the training data available from ARQMath-1 was beneficial, with greater improvements in ARQMath-2 relative to baselines for both Task 1 and Task 2 than for ARQMath-1 relative to those same baselines. Tasks, topics, evaluation protocols, and results for each task are presented in this lab overview.

Keywords

Community Question Answering (CQA), Mathematical Information Retrieval (MIR), Math-aware search, Math formula search

1. Introduction

This second Answer Retrieval for Questions on Math (ARQMath-2) lab[1] at the Conference and Labs of the Evaluation Forum (CLEF) continues a multi-year effort to build new test collections for Mathematics Information Retrieval (Math IR) from content found on Math Stack Exchange,[2] a Community Question Answering (CQA) forum. Using the question posts from Math Stack Exchange, participating systems are given a question or a formula from a question, and asked to return a ranked list of either potential answers to the question or potentially useful formulae (in the case of a formula query). Relevance is determined by the expected utility of each returned item. These tasks allow participating teams to explore leveraging math notation together with text to improve the quality of retrieval results. Table 1 illustrates these two tasks, and Figure 1 shows the topic format for each task.

For the CQA task, 146,989 questions from 2020 that contained some text and at least one formula were considered as search topics, from which 100 were selected for use in ARQMath-2.

CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania
bm3302@rit.edu (B. Mansouri); rxzvcs@rit.edu (R. Zanibbi); oard@umd.edu (D. W. Oard); axasma@rit.edu (A. Agarwal)

[1]https://www.cs.rit.edu/~dprl/ARQMath
[2]https://math.stackexchange.com

Table 1
Examples of relevant and not-relevant results for tasks 1 and 2 [1]. For Task 2, formulae are associated with posts, indicated with ellipses at right (see Figure 1 for more details). Query formulae are from question posts (here, the question at left), and retrieved formulae are from either an answer or a question post.

Task 1: Question Answering	Task 2: Formula Retrieval
Question I have 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? $\lim_{n\to+\infty} n^{\frac{1}{n}}$	Query Formula $\ldots \quad \lim_{n\to+\infty} n^{\frac{1}{n}} \quad \ldots$
Relevant You can use AM $\geq$ GM. $\frac{1+1+\cdots+1+\sqrt{n}+\sqrt{n}}{n} \geq n^{1/n} \geq 1$ $1-\frac{2}{n}+\frac{2}{\sqrt{n}} \geq n^{1/n} \geq 1$	Relevant $\ldots \quad \lim_{n\to\infty} \sqrt[n]{n} \quad \ldots$
Not Relevant If you just want to show it converges, then the partial sums are increasing but the whole series is bounded above by $1+\int_1^{\infty} \frac{1}{x^2} dx = 2$	Not Relevant $\ldots \quad \sum_{k=1} \frac{1}{k^2} = \frac{\pi^2}{6} \quad \ldots$

For the question answering task, the title and body of the question were provided to participating teams, although other associated data (e.g., comments, answers, and links to related questions) were excluded. For the formula search task, an individual formula from the question post is specified as the query, and systems return a ranked list of other potentially useful instances of formulae found in the collection. Each of the 58 formula queries is a single formula extracted from a question used in the CQA task. For both tasks, participating teams had the option to construct queries using only the text or math portions of each question, or to use both math and text. Following convention, we refer to both questions and formula queries as *topics*.

The ARQMath labs have three objectives:

1. Create test collections for training and evaluating Math IR systems.
2. Establish state-of-the-art results on those test collections to which future researchers can compare their results.
3. Foster the growth of the Math IR research community.

ARQMath-2 saw progress on each of these goals, roughly doubling the size of the available test collections, nearly doubling the number of participating teams, and demonstrating that substantial improvements over the results reported in ARQMath-1 are possible.

```
<Topics>
  ...
  <Topic number="A.9">
    <Title>Simplifying this series</Title>
    <Question>
      I need to write the series
      <span class=``math-container'' id=``q_52''>
        $$\sum_{n=0}^N nx^n$$
      </span>
      in a form that does not involve the summation
      notation, for example
      <span class=``math-container'' id=``q_53''>
        $\sum_{i=0}^n i^2 = \frac{(n^2+n)(2n+1)}{6}$
      </span>
      Does anyone have any idea how to do this?
      I have attempted multiple ways including using
      generating functions however no luck.
    </Question>
    <Tags>sequences-and-series</Tags>
  </Topic>
  ...
</Topics>
```

Task 2: Formula Retrieval

```
<Topics>
  ...
  <Topic number="B.9">
    <Formula_Id>q_52</Formula_Id>
    <Latex>\sum_{n=0}^N nx^n</Latex>
    <Title>Simplifying this series</Title>
    <Question>
      ...
    </Question>
    <Tags>sequences-and-series</Tags>
  </Topic>
  ...
</Topics>
```

Figure 1: XML Topic File Formats for Tasks 1 and 2. Formula queries in Task 2 are taken from questions for Task 1. Here, ARQMath-1 formula topic B.9 is a copy of ARQMath-1 question topic A.9 with two additional tags for the query formula identifier and LaTeX before the question post.

2. Related Work

Math IR shares many commonalities with information retrieval more generally. For example, both exploratory search and refinding are common tasks, and query autocompletion and diversity ranking can be useful capabilities. Math IR is a special case of cross-modal retrieval, since both text and math can be used to express the same idea, and those two modalities can be productively used together in the query, the document, or both.

The nature of mathematics, however, introduces some unique challenges. Here we need to distinguish between mathematics as a field and mathematical notation as a language. The notion of relevance in Math IR is grounded in mathematics as a field, whereas many of the implementation details are grounded in mathematical notation as a language. To see the difference, consider the notion of equality: many people would consider that 3 + 2, 2 + 3, and 5 express the same idea, being equally happy to find formulae that contain any of those. However, many might regard $cos^2(x) + sin^2(x)$ and 1 as different, despite their equality, because the first

is specific to some uses of mathematics, and thus not an appropriate formulation for others.

Indeed, thinking of mathematics as a field is itself too reductionist – mathematics is used in many disciplines (e.g., computer science, physics, chemistry, quantum mechanics, economics, and nearly every branch of engineering). In some cases, relevance may be defined within one of those disciplines, with economists looking for other work in economics, for example. In other cases, relevance might be defined in a way that spans disciplines, such as when an engineer might be looking for the work of mathematicians that can help them to solve some specific problem, even when they don't know the notation the mathematicians would have used in formulating or solving that problem.

No single evaluation design could possibly model the full complexity of information needs for Math IR, so every evaluation has been specialized in some way. Mathematicians naturally find Math IR potentially interesting, and one motivation for early work on Math IR has been to support mathematics education. Students can use search engines to find references for assignments, to solve problems, increase knowledge, or clarify concepts. In general math-aware search can be used to find similarities between a piece of mathematics being developed, on the one hand, and proved theorems and well-developed theories in the same or different parts of mathematics, on the other hand.

Complementing this somewhat underdeveloped focus on task design among Math IR researchers is a quite well developed lower-level focus on mathematical notation within the Mathematical Knowledge Management (MKM) community, where the representations of, and operations on math notation have been carefully considered. Among other accomplishments, their activities informed the development of MathML[3] for math on the Web, and novel techniques for math representation and applications such as theorem proving. This community meets annually at the Conference on Intelligent Computer Mathematics (CICM) [2].

Math IR naturally draws on both of these traditions. Math formula search has been studied since the mid-1990's for use in solving integrals, and publicly available math+text search engines have been around since the DLMF[4] system in the early 2000's [3, 4]. Prior to ARQMath, the most widely used 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 [5], NTCIR-11 [6] and NTCIR-12 [7]). NTCIR-12 used 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. The NTCIR Mathematical Information Retrieval (MathIR) tasks developed evaluation methods and allowed participating teams to establish baselines for both "text + math" queries (i.e., keywords and formulae) and isolated formula queries.

At NTCIR-11 and NTCIR-12, formula retrieval was considered in a variety of settings, including the use of wildcards and constraints on symbols or subexpressions (e.g., requiring matched argument symbols to be variables or constants). Our Task 2, Formula Retrieval, has similarities in design to the NTCIR-12 Wikipedia Formula Browsing task, but differs in how queries are defined and how evaluation is performed. In particular, relevance is defined contextually in ARQMath, and ARQMath evaluation is based on *visually distinct* formulae, rather than all (possibly

[3] https://www.w3.org/Math
[4] https://dlmf.nist.gov

identical) formula instances, as had been done in NTCIR-12. The NTCIR-12 formula retrieval test collection also had a smaller number of queries, with 20 fully specified formula queries (plus 20 variants of those same queries with subexpressions replaced by wildcard characters). NTCIR-11 also had a formula retrieval task, with 100 queries, but in that case systems searched only for exact matches [8].

Another related effort was the SemEval 2019 [9] question answering task. Question sets from MathSAT (Scholastic Achievement Test) practice exams in three categories were used: Closed Algebra, Open Algebra and Geometry. A majority of the questions were multiple choice, with some having numeric answers. This is a valuable parallel development; the questions considered in the CQA task of ARQMath are more informal and open-ended, and selected from actual Math Stack Exchange user posts (a larger and less constrained set).

3. The ARQMath Stack Exchange Collection

For ARQMath-2, we reused the test collection from the first ARQMath. The test collection was constructed using the March 1st, 2020 Math Stack Exchange snapshot from the Internet Archive.[5] Questions and answers from 2010-2018 are included in the collection. The ARQMath test collection contains roughly 1 million questions and 28 million formulae. Formulae in the collection are annotated using `<span>` tags with the class attribute `math-container`, and a unique integer identifier given in the `id` attribute. Formulae are also provided separately in three index files for different formula representations (LaTeX, Presentation MathML, and Content MathML), which we describe in more detail below.

HTML views of question threads, similar to those on the Math Stack Exchange web site (a question, along with answers and other related information) are also included in the ARQMath test collection. The threads are constructed automatically from Math Stack Exchange snapshot XML files. The threads are intended for those performing manual runs, or who wish to examine search results (on queries other than evaluation queries) for formative evaluation purposes. These threads are also used by assessors during evaluation. The HTML thread files were intended only for viewing threads; participants were asked to use provided XML and formula index files to train their models.

Questions posted after 2018 are used to create test topics: questions from 2019 were used for the first ARQMath, and questions from 2020 are used for ARQMath-2. Additional details may be found in the ARQMath-1 task overview paper [10].

Formula Index Files and Visually Distinct Formulae. In addition to LaTeX, it is common for math-aware information retrieval systems to represent formulae as one or both of two types of rooted trees. Appearance is represented by the spatial arrangement of symbols on writing lines (in Symbol Layout Trees (SLTs)), and mathematical syntax (sometimes referred to as (shallow) semantics) is represented using a hierarchy of operators and arguments (in Operator Trees (OPTs)) [11, 12, 13]. The standard representations for these are Presentation MathML (SLT) and Content MathML (OPT).

To reduce effort for participants, and to maximize comparability across submitted runs, we

[5] https://archive.org/download/stackexchange

used LaTeXML[6] 0.8.5 to generate Presentation MathML and Content MathML from LaTeX for each formula in the ARQMath collection. Some LaTeX formulae were malformed, and LaTeXML has some processing limitations, resulting in conversion failures for 0.14% of both SLTs and OPTs.[7] Participants could elect to do their own formula extraction and conversions, although the formulae that could be submitted in system runs for Task 2 were limited to those with identifiers in the provided LaTeX formula index file.

During evaluation we learned that LaTeX formulae that could not be processed by LaTeXML had their visual identifiers assigned incorrectly, and this may have affected adjacent formulae in the formula index files. This had a small effect on evaluation metrics (our best estimate is that no more than 1.3 visually distinct formulae in the pool for each topic were affected).

ARQMath formulae are provided in LaTeX, SLT, and OPT representations, as Tab Separated Value (TSV) index files. Each line of a TSV file represents a single instance of a formula, containing the formula id, the id of the post in which the formula instance appeared, the id of the thread in which the post is located, a post type (title, question, answer or comment), and the formula representation in either LaTeX, SLT (Presentation MathML), or OPT (Content MathML).

For ARQMath-2, in the formula TSV index files we added a new field for *visually distinct* formula identifiers used in evaluation for Task 2 (Formula Retrieval).[8] The idea is to identify formulae sharing the same appearance. So for example, two occurrences of x^2 in a TSV formula index have different formula *instance* identifiers, but the same *visually distinct* formula identifier. All ARQMath-2 formula index files provide visually distinct identifiers for each formula in the collection.

There are three sets of formula index files: one set is for the collection (i.e., for posts from 2018 and before), while the second and third sets are for search topics from 2020 (ARQMath-2), and 2019 (ARQMath-1). Only the collection index files have visually distinct formula identifiers.

Distribution. The Math Stack Exchange test collection was distributed to participants as XML files on Google Drive.[9] To facilitate local processing, the organizers provided python code on GitHub[10] for reading and iterating over the XML data, and generating the HTML question threads.

4. Task 1: Answer Retrieval

The main task in ARQMath is the answer retrieval task. Participating systems are given a Math Stack Exchange question post from 2019, and return a ranked list of up to 1,000 answer posts from 2010-2018. System results ('runs') are evaluated using rank quality measures that characterize the extent to which annotated answers with higher relevance come before answers with lower relevance (e.g., nDCG′). This makes Task 1 a ranking task rather than a set retrieval task.

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

[7] We thank Deyan Ginev and Vit Novotny for helping reduce LaTeXML failures: for ARQMath-1 conversion failures affected 8% of SLTs, and 10% of OPTs.

[8] We thank Frank Tompa for sharing this suggestion at CLEF 2020.

[9] https://drive.google.com/drive/folders/1ZPKIWDnhMGRaPNVLi1reQxZWTfH2R4u3

[10] https://github.com/ARQMath/ARQMathCode

In the following we describe the Task 1 search topics, runs from participant and baseline systems, the assessment and evaluation procedures used, and a summary of the results.

4.1. Topics

In Task 1, participants were given 100 Math Stack Exchange questions posted in 2020 as topics. We used a sampling strategy similar to ARQMath-1, where we chose from questions containing text and at least one formula. To help ensure that most topics had relevant answers available in the collection, we calculated the number of duplicate and related posts for each question, and then chose the majority of topics (89 out of 100) from those with at least one duplicate or related post.[11] To increase the difficulty and diversity of topics, we selected the remaining topics from those without annotated duplicates or related posts.[12]

Because we were interested in a diverse range of search tasks, we also calculated the number of formulae for each question. Finally, we noted the asker's reputation and the tags assigned for each question. We manually drew a sample of 100 questions stratified along those dimensions. In the end, pools for 71 of these questions were evaluated and found to have a sufficient number of relevant responses, and thus were included in the ARQMath-2 test collection.

The topics were selected from various domains to capture a broad spectrum of mathematical areas. The difficulty level of the topics spanned from easy problems that a beginning undergraduate student might be interested in to difficult problems that would be of interest to more advanced users. The bulk of the topics were aimed at the level of undergraduate math majors (in their 3rd or 4th year) or engineering majors fulfilling their math requirements.

As organizers, we labeled each question with one of three broad categories, *computation*, *concept* or *proof*. Out of the 71 assessed questions, 25 were categorized as *computation*, 19 as *concept*, and 27 as *proof*. We also categorized questions based on their perceived difficulty level, with 32 categorized as easy, 20 as medium, and 19 as hard. Our last categorization was based on whether a question is dependent on text, formula or both. 10 questions were (in our opinion) dependent on text, 21 on formula and 40 on both.

The topics were published as an XML file with the format shown in Figure 1, where the topic number is an attribute of the Topic tag, and the Title, Question and asker-provided Tags are from the Math Stack Exchange question post. To facilitate system development, we provided python code that participants could use to load the topics. As in the collection, the formulae in the topic file are placed in '`math-container`' tags, with each formula instance represented by a unique identifier and its LaTeX representation. And, as with the collection, we provided three TSV files, one each for the LaTeX, OPT and SLT representations of the formulae, in the same format as the collection's TSV files.

4.2. Participant Runs

Participating teams submitted runs using Google Drive. A total of 36 runs were received from a total of 9 teams. Of these, 28 runs were declared to be automatic, meaning that queries were automatically processed from the topic file, that no changes to the system had been made

[11] Participating systems did not have access to this information.

[12] In ARQMath-1, all topics had links to at least one duplicate or related post that were available to the organizers.

Table 2
Submitted Runs for Task 1 (36 runs) and Task 2 (17 runs). Additional baselines for Task 1 (4 runs) and Task 2 (1 run) were also generated by the organizers.

	Automatic		Manual	
	Primary	Alternate	Primary	Alternate
Task 1: Answers				
Baselines	2	2		
Approach0			1	4
BetterThanG		2	1	2
DPRL	1	2		
GoogolFuel	1	4		
MathDowsers	1	1		
MIRMU	1	4		
MSM	1	4		
PSU	1			
TU_DBS	1	4		
Task 2: Formulas				
Baseline	1			
Approach0			1	4
DPRL	1	3		
MathDowsers	1	1		
NLP-NITS	1			
TU_DBS	1	3		
XY_PHOC_DPRL	1			

after seeing the queries, and that ranked lists for each query were produced with no human intervention. 8 runs were declared to be manual, meaning that there was some type of human involvement in generating the ranked list for each query. Manual runs can contribute diversity to the pool of documents that are judged for relevance, since their error characteristics can differ from those of automatic runs. The teams and submissions are shown in Table 2. Please see the participant papers in the working notes for descriptions of the systems that generated these runs.

Important Note: All participant and baseline runs are available online.[13]

4.3. Baseline Runs: TF-IDF, Tangent-S, Linked Posts

The organizers ran four baseline systems for Task 1.[14] These baselines were also run for ARQMath 2020, and we re-ran them on the same systems as last year, obtaining very similar run-times [10]. Here is a description of our baseline runs.

1. **TF-IDF.** A term frequency, inverse document frequency model implementation provided by the Terrier system [14]. Formulae are represented using their LaTeX strings. Default

[13] https://drive.google.com/drive/u/1/folders/1l1c2O06gfCk2jWOixgBXI9hAlATybxKv

[14] Source code and instructions for running baselines available from GitLab (Tangent-S: https://gitlab.com/dprl/tangent-s) and GoogleDrive (Terrier: https://drive.google.com/drive/u/0/folders/1YQsFSNoPAFHefweaN01Sy2ryJjb7XnKF)

parameters from Terrier were used.

2. **Tangent-S**. Formula search engine using SLT and OPT formula representations [11]. One formula was selected from each Task 1 question title if possible; if there was no formula in the title, then one formula was instead chosen from the question's body. If there were multiple formulae in the selected field, a formula with the largest number of symbols (nodes) in its SLT representation was chosen; if more than one formula had the largest number of symbols, we chose randomly between them.
3. **TF-IDF + Tangent-S.** Averaging similarity scores from the TF-IDF and Tangent-S baselines. The relevance scores from both systems were normalized in [0,1] using min-max normalization, and then combined using an unweighted average.
4. **Linked Math Stack Exchange Posts.** This is a simple oracle "system" that is able to see duplicate post links from 2020 in the Math Stack Exchange collection (which were not available to participants). It returns *all* answer posts from 2018 or earlier that were in threads that Math Stack Exchange moderators had marked as duplicating the topic question post. Answer posts are sorted in descending order by their vote scores.

4.4. Assessment

Pooling. Participants were asked to rank up to 1,000 answer posts for each topic, which were then sampled for assessment using Top-k pooling. The top 45 results were combined from all primary runs. To this, we added the top 15 results from each alternate run. The baseline systems, TF-IDF + Tangent-S and Linked Math Stack Exchange Posts, were considered as primary runs and the other two (TF-IDF and Tangent-S) were considered as alternative. Duplicates were then deleted, and the resulting pool was sorted randomly for display to assessors. The pooling depth was designed to identify as many relevant answer posts as possible given our assessment resources. On average, the pools contained 448.12 answers per topic.

Relevance definition. We used the same relevance definitions created for ARQMath-1. To avoid assessors needing to guess about the level of mathematical knowledge available to the Math Stack Exchange users who originally posted the questions, we asked assessors to base their judgments on the degree of usefulness for an expert (modeled in this case as a math professor), who might then try to use that answer to help the person who had asked the original question. We defined four levels of relevance, as shown in Table 3.

Assessors were allowed to consult external sources to familiarize themselves with the topic of a question, but relevance judgments were made using only information available within the ARQMath test collection. For example, if an answer contained a Math Stack Exchange link such as https://math.stackexchange.com/questions/163309/pythagorean-theorem, they could follow that link to better understand the intent of the person writing the answer, but an external link to the Wikipedia page https://en.wikipedia.org/wiki/Pythagorean_theorem would not be followed.

Training. Unlike ARQMath-1, for ARQMath-2 participants could use the 77 annotated topics for ARQMath-1 Task 1 as a training set [10, 15]. For sanity checking results and comparison, results were collected from participants for both the ARQMath-1 and ARQMath-2 topics, and results for both training (ARQMath-1) and testing (ARQMath-2) are provided at the end this document.

Table 3
Relevance Assessment Criteria for Tasks 1 and 2.

Score	Rating	Definition
Task 1: Answer Retrieval		
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		
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

Assessment System. For ARQMath-2, assessments were again performed using Turkle[15], a locally installed system with functionality similar to Amazon Mechanical Turk. As Figure 4 at the end of this document illustrates, there were two panels in the Turkle user interface. The question was shown on the left panel, with the Title on top in a grey bar; below that was the question body. There was also a Thread link, through which assessors could access the Math Stack Exchange post in context, with the question and all answers given for this question (in 2020). In the right panel, the answer to be judged was shown at the top, along with another thread link that allows assessors to view the original thread in which the answer post appeared, which could be helpful for clarifying the context of the answer post, for example by viewing the original question to which it had been posted as a response. Finally, below the answer in the right panel was where assessors selected relevance ratings.

In addition to four levels of relevance, two additional choices were available: 'System failure' indicated system issues such as unintelligible rendering of formulae, or the thread link not working (when it was essential for interpretation). If after viewing the threads, the assessors were still not able to decide the relevance degree, they were asked to choose 'Do not know'. The organizers asked the assessors to leave a comment in the event of a system failure or a 'Do not know' selection. As it happened, the ARQMath-2 assessors did not use these options for Task 1; for each answer, they decided a relevance degree.

Assessor Training. Seven paid undergraduate and graduate mathematics and computer science students from RIT and St. John Fisher College were paid to perform relevance judgments. One assessor had worked with us previously on ARQMath-1. Due to the COVID pandemic, all training sessions were performed remotely over Zoom. For ARQMath-1, relevance criteria had been developed interactively with assessors, leading to four rounds of training; we found the resulting relevance criteria worked well, and so we reused them for ARQMath-2. This allowed us to reduce assessor training time: the four assessors who worked exclusively on Task 1 participated in three meetings, and just two rounds of training assessments. The remaining

[15]https://github.com/hltcoe/turkle

three assessors initially worked on Task 2, and were later moved to Task 1 after Task 2 assessment was completed. Those three assessors had an initial training meeting when they returned to Task 1 to introduce the task, and then they performed a single round of training assessments, with subsequent discussion at a second meeting. One of those three assessors had previously worked on both Task 1 and Task 2 assessments for ARQMath-1.

At the first assessor meeting, the lab and administrative details were discussed. After this, the assessors were divided into two groups, for Task 1 and Task 2. After this we began training sessions. In the first Task 1 training session, the task was explained, making reference to specific topics and previously assessed answers from ARQMath-1. For each training/practice assessment round, the same 7 topics were assigned to every assessor and the assessors worked independently, thus permitting inter-assessor agreement measures to be computed. After completing training assessments, a meeting was held to discuss disagreements in relevance scores between with the organizers, along with clarifications of the relevance criteria. The assessors discussed the reasoning for their choices, with the fourth author of this paper (an expert Math Stack Exchange user) sharing their own assessment and reasoning. The primary goal of training was to help assessors make self-consistent annotations, as question interpretations will vary across individuals.

Some of the question topics would not be typically covered in regular undergraduate courses, so that was a challenge that required the assessors to get a basic understanding of those topics before they could do the assessment. The assessors found the question threads made available in the Turkle interface helpful in this regard (see Figure 4).

Figure 2 shows agreement between assessors in our two groups over the course of the training process. As shown, collapsing relevance to be binary by considering high and medium as relevant and low and not-relevant as a not-relevant (henceforth “H+M binarization") yielded better agreement among the assessors. [16]

Assessment. A total of 81 topics were assessed for Task 1. 10 judgment pools (for topics A.208, A.215, A.216, A.221, A.230, A.236, A.266, A.277, A.278 and A.280) had zero or one posts with relevance levels of high or medium; these topics were removed from the collection because topics with no relevant posts cannot be used to distinguish between ranked retrieval systems, and because topics with only a single relevant post result in coarsely quantized values for the evaluation measures that we report. For the remaining 71 topics, an average of 447.7 answers were assessed, with an average assessment time of 83.3 seconds per answer post. The average number of answers labeled with any degree of relevance (high, medium, or low; henceforth “H+M+L binarization”) was 49.0 per question, with the highest number of relevant answers being 134 (for topic A.237) and the lowest being 4 (for topic A.227).

Post Assessment. After assessments were completed for Task 1, each assessor was assigned one topic that had originally been completed by another assessor.[17] We were particularly interested in confirming cases in which non relevant documents were found, so for each

[16] H+M binarization corresponds to the definition of relevance usually used in the Text Retrieval Conference (TREC). The TREC definition is “If you were writing a report on the subject of the topic and would use the information contained in the document in the report, then the document is relevant. Only binary judgments (‘relevant’ or ‘not relevant’) are made, and a document is judged relevant if any piece of it is relevant (regardless of how small the piece is in relation to the rest of the document).” (source: https://trec.nist.gov/data/reljudge_eng.html)

[17] One assessor (with id 7) was not able to continue assessment.

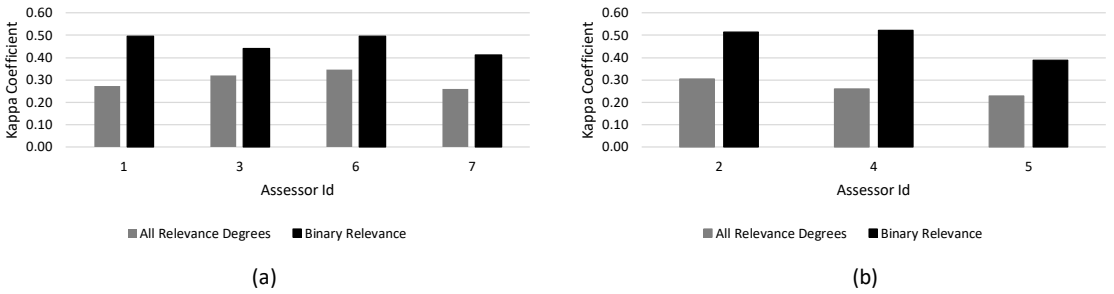

Figure 2: Inter-annotator agreement (Average Cohen' kappa) over 7 assessors during the last Task 1 training (7 topics from ARQMath-1); four-way classification (gray) and two-way (H+M binarized) classification (black). Chart (a) shows the agreement between the assessors who did only Task1 and had an additional training session. Chart (b) shows the agreement between the assessors who started with Task 2, and then moved to Task1.

assessor we selected the topic with the fewest relevant topics. Among the 6 dual-assessed topics, 4 had no high or medium relevant answers according to at least one of the two assessors[18]; meaningful values of kappa for binary relevance can not be calculated in such cases. Averaged over the remaining two questions, kappa was 0.21 on the four-way assessment task, and using H+M binarization it was 0.32.

4.5. Evaluation Measures

For a complex task where rich training data is not yet available, it is possible that a large number of relevant answers may be missed during assessment. Measures which treat unjudged documents as not relevant can be used when directly comparing systems contributing to the judgment pools, but non-contributing systems can be disadvantaged by treating unjudged documents as not relevant, which may prove to be relevant in later analysis. We therefore chose the nDCG′ measure (read as "nDCG-prime") introduced by Sakai and Kando [16] as the primary measure for the task.

nDCG is a widely used measure for graded relevance judgments, used to produce a single figure of merit over a set of ranked lists. For ARQMath, each retrieved document earns a gain value (relevance score) $g \in \{0, 1, 2, 3\}$, discounted by a slowly decaying function of the rank position of each result. Discounted gain values are accumulated and then normalized to [0,1] by dividing by the maximum possible Discounted Cumulative Gain (i.e., from all relevant documents sorted in decreasing order of gain value). This results in normalized Discounted Cumulative Gain (nDCG).

The only difference when computing nDCG′ is that unjudged documents are removed from the ranked list before performing the computation. It has been shown that nDCG′ has somewhat better discriminative power and somewhat better system ranking stability (with judgement ablation) than the bpref measure [17] used recently for formula search (e.g., [12]). Moreover, nDCG′ yields a single-valued measure with graded relevance, whereas bpref, Precision@k, and

[18]Two of the 4 dual-assessed topics had no high or medium relevant answers found by by either assessor

Mean Average Precision (MAP) all require binarized relevance judgments. In addition to nDCG′, we also compute Mean Average Precision (MAP) with unjudged posts removed (thus MAP′), and Precision at 10 with unjudged posts removed (P′@10).[19] For MAP′ and P′@10 we used H+M binarization.

The ARQMath Task 1 evaluation script removes unjudged posts as a preprocessing step where required, and then computes evaluation measures using `trec_eval`.[20]

4.6. Results

Table 4 in the appendix shows the results for baselines along with teams and their systems ranked by nDCG′. nDCG′ values can be interpreted as the average (over topics) of the fraction of the score for the best possible that was actually achieved. As can be seen, the best nDCG′ value that was achieved was 0.434, by the MathDowsers team. MAP′ with H+M binarization generally ranks systems in the same order as nDCG′ does with graded relevance judgments. However, the results for P′@10 with H+M binarization differ, the TU_DBS team doing best among the participating teams by that measure (exceeded only by the *Linked MSE posts* baseline, which uses human-built links that were not available to participating teams). There are some noticeable differences in system orderings for several participating teams when using ARQMath-2 topics compared with what was seen when those same teams used the same systems (in 2021) on ARQMath-1 topics.

Now comparing results from 2021 with results from 2020, we see that the best improvement over the strongest fully automated baseline in both years (*TF-IDF + Tangent-S*) was substantially larger in 2021 than in 2020. Specifically, in 2020 the MathDowsers team outperformed that baseline by 39% as measured by nDCG′; in 2021 they outperformed that same baseline by 116% as measured by nDCG′.

5. Task 2: Formula Retrieval

In the formula retrieval task, participants were presented with one formula from a 2020 question used in Task 1, and asked to return a ranked list of up to 1,000 formula instances from questions or answers from the evaluation epoch (2018 or earlier). Formulae were returned by their identifiers in `math-container` tags and the companion TSV LaTeX formula index file, along with their associated post identifiers.

As with Task 1, ranked lists were evaluated using rank quality measures, making this a ranking task rather than a set retrieval task. Three key details differentiate Task 2 from Task 1:

1. Unlike Task 1, in Task 2 the goal is not answering questions, but to instead show the searcher formulae that might be useful as they seek to satisfy their information need. Task 2 is thus still grounded in the question, but the relevance of a retrieved formula

[19] Pooling to at least depth 10 ensures that there are no unjudged posts above rank 10 for any baseline, primary, or alternative run. Note that P′@10 cannot achieve a value of 1 because some topics have fewer than 10 relevant posts.

[20] https://github.com/usnistgov/trec_eval

is defined by a formula's expected utility, not just the post in which any one formula instance was found.
2. In Task 1 only answer posts were returned, but for Task 2 the formulae may appear in answer posts or in question posts.
3. For Task 2 we distinguish visually distinct formulae from instances of those formulae, and evaluate by the ranking of visually distinct formulae returned. We call formulae appearing in posts *formula instances*, and of course the same formula may appear in more than one post. By *a visually distinct formula* we mean a set of formula instances that are visually identical when viewed in isolation. For example, x^2 is a formula, $x \cdot x$ is a different visually distinct formula, and each time x^2 appears is an instance of the visually distinct formula x^2. Although systems in Task 2 rank formula instances in order to support the relevance judgment process, the evaluation measure for Task 2 is based on the ranking of visually distinct formulae.

The remainder of this section describes for Task 2 the search topics, the submissions and baselines, the process used for creating relevance judgments, the evaluation measures, and the results.

Important Note: run data and code for baseline systems may be found at links provided in the previous Section.

5.1. Topics

In Task 2, participating teams were given 100 mathematical formulae, each found in a different Math Stack Exchange question from Task 1 (posted in 2020). They were asked to find relevant formulae instances from either question or answer posts in the test collection (from 2018 and earlier). The topics for Task 2 were provided in an XML file similar to those of Task 1, in the format shown in Figure 1. Task 2 topics differ from their corresponding Task 1 topics in three ways:

1. **Topic number.** For Task 2, topic ids are in the form "B.x" where x is the topic number. There is a correspondence between topic id in tasks 1 and 2. For instance, topic id "B.209" indicates the formula is selected from topic "A.209" in Task 1, and both topics include the same question post (see Figure 1).
2. **Formula_Id**. This added field specifies the unique identifier for the query formula instance. There may be other formulae in the Title or Body of the question post, but the query is only the formula instance specified by this Formula_Id.
3. **LaTeX.** This added field is the LaTeX representation of the query formula instance as found in the question post.

Because query formulae are drawn from Task 1 question posts, the same LaTeX, SLT and OPT TSV files that were provided for the Task 1 topics can be consulted when SLT or OPT representations for a query formula are needed.

Formulae for Task 2 were manually selected using a heuristic approach to stratified sampling over three criteria: complexity, elements, and text dependence. Formulae complexity was labeled

low, medium or high by the fourth author. For example, $\frac{df}{dx} = f(x+1)$ is low complexity, $\sum_{k=0}^{n} \binom{n}{k} k$ is medium complexity, and

$$x - \frac{x^3}{3 \times 3!} + \frac{x^5}{5 \times 5!} - \frac{x^7}{7 \times 7!} + \cdots = \sum_{n=0}^{\infty} (-1)^n \frac{x^{(2n+1)}}{(2n+1) \times (2n+1)!}$$

is high complexity.

Text dependence reflected the first author's opinion of the degree to which text in the Title and Question fields were likely to yield related search results. For instance, for one Task 2 topic, the query formula is $\frac{df}{dx} = f(x+1)$ whereas the complete question is: "How to solve differential equations of the following form: $\frac{df}{dx} = f(x+1)$." When searching for this formula, perhaps the surrounding text could safely be ignored. At most one formula was selected from each Task 1 question topic to produce Task 2 topics. For cases in which suitable formulae were present in both the title and the body of the Task 1 question, we selected the Task 2 formula query from the title.

5.2. Participant Runs

A total of 17 runs were received for Task 2 from a total of six teams, as shown in Table 2. Each run contained at most 1,000 formula instances for each topic, ranked in decreasing order of system-estimated relevance to that query. For each formula instance in a ranked list, participating teams provided the `formula_id` and the associated `post_id` for that formula. Please see the participant papers in the working notes for descriptions of the systems that generated these runs.

5.3. Baseline Run: Tangent-S

We again used Tangent-S [11] as our baseline. Unlike Task 1, a single formula is specified for each Task 2 query, so no formula selection step was needed. This Tangent-S baseline makes no use of the question text. Timing was similar to the use of Tangent-S in ARQMath-1.

5.4. Assessment

Pooling. The retrieved items for Task 2 are formula instances, but pooling was done based on the visually distinct formulae, and not individual formula instances. Visually distinct formulae were identified by clustering all formula instances in the collection.[21] Pooling was performed by then proceeding down each results list until at least one instance of some number of visually distinct formulae had been seen. For primary runs and for the baseline run, the pool depth was the rank of the first instance of the 20th visually distinct formula; for alternate runs the pool depth was the rank of the first instance of the 10th visually distinct formulae. [22]

[21] This differs from the approach used for ARQMath-1, when only submitted formula instances were clustered. For ARQMath-2 the full formula collection was clustered to facilitate post hoc use of the resulting test collection.

[22] In ARQMath-2, Task 1 pools were not used to seed Task 2 pools.

Clustering of visually distinct formulae instances was performed using the SLT representation when possible,[23] and the LaTeX representation otherwise. We first converted the Presentation MathML representation to a string representation using Tangent-S, which performed a depth-first traversal of the SLT, with each SLT node and edge generating a single character of the SLT string. Formula instances with identical SLT strings were considered to be the same formula; note that this ignores differences in font. For formula instances with no Tangent-S SLT string available, we removed the white space from their LaTeX strings and grouped formula instances with identical strings. This process is simple and appears to be reasonably robust, but it is possible that some visually identical formula instances were not captured due to LaTeXML conversion failures, or where different LaTeX strings produce the same formula (e.g., if subscripts and superscripts appear in a different order in LaTeX).

Assessment was done on formula instances: for each visually distinct formula we selected at most five instances to assess. We did this differently than last year; in order to prefer highly-ranked instances and instances returned in multiple runs, we selected the 5 instances using a simple voting protocol, where each instance votes by the sum of its reciprocal ranks within each run, breaking ties randomly. Out of 8,129 visually distinct formulae that were assessed, 117 (1.4%) had instances in more than 5 pooled posts.[24]

Relevance definition. The relevance judgment task was defined for assessors as follows: for a formula query, if a search engine retrieved one or more instances of this retrieved formula, would that have been expected to be useful for the task that the searcher was attempting to accomplish?

Assessors were presented with formula instances in context (i.e., in the question or answer in which they had been found). They were then asked to decide their relevance by considering whether retrieving either that instance or some other instance of that formula could have helped the searcher to address their information need. To make this judgment, they were shown the query formula within the question post where it appeared. Each formula instance in the judgment pool was assigned one of four relevance levels as defined in Table 3.

For example, if the formula query was $\sum \frac{1}{n^{2+\cos n}}$, and the formula instance to be judged is $\sum_{n-1}^{\infty} \frac{1}{n^2}$, the assessors could look at the formula's associated post, compare the formula's variable types and operations with the query, identify the area of mathematics it concerns, and then decide whether finding the second formula rather than the first would be expected to yield good results. Further, they could consider the content of the question post containing the query (and, optionally, the thread containing that question post) in order to understand the searcher's information need. Thus the question post fills a role akin to Borlund's simulated work task [18], although in this case the title, body and tags from the question post are included in the topic and thus can optionally be used by the retrieval system.

The assessor could also consult the post containing a retrieved formula instance (which may be another question post, or an answer post) along with the associated thread, to see if in that case the formula instance would indeed have been a useful basis for a search. Note, however,

[23] For ARQMath-1, 92% of formula instances had an SLT representation; for ARQMath-2 we reparsed the collection and improved this to 99.9%.

[24] As mentioned in Section 3, a relatively small number of formulae per topic had incorrectly generated visual ids. In 6 cases assessors indicated that a pooled formula for a single visual id was 'not matching' the other formulae in hits grouped for a visual id, rather than assign a relevance score for the formula.

that the assessment task is not to determine whether the specific post containing the retrieved formula instance is useful, but rather to use that context as a basis for estimating the degree to which useful content would likely be found if this or other instances of the retrieved formula were returned by a search engine.

Although this definition of relevance was unchanged between ARQMath-1 and ARQMath-2, we did make one potentially significant change to the way this relevance definition was interpreted for ARQMath-2. In ARQMath-2, Task 2 assessment was done based on the context in which the formulas appear, exactly as described above. This means that it is possible that an exact match with the query formula may in some cases (depending on context) be considered not relevant. For example, for the formula query $x^n + y^n + z^n$ (B.289) x, y, and z could be any real numbers in the original question post in which that formula had originally appeared. The assessors considered all exact matches in the pooled posts in which x, y, and z referred not to real numbers but specifically to integers as not relevant. On the other hand, formulas that do not share the same appearance or syntax with the query might be considered relevant. This is usually the case where they are both referring to the same concept. For the query formula $\frac{S}{n} \geq \sqrt[n]{P}$ (B.277), formula $\frac{1+2+3+...+n}{n} \geq \sqrt[n]{n!}$ has medium relevance. Both formulae are referring to the AM-GM inequality (of Arithmetic and Geometric Means).

It should be noted that the last author (a mathematician) reviewed these two examples and agreed with the assessors. In 2020, by contrast, assessors had been instructed during training that if the query and candidate formulae were the same (in appearance), then the candidate was certainly highly relevant. During assessor training in 2021 this issue received considerable attention and discussion, and we ultimately concluded that our guidance in 2020 had not been fully consistent with our relevance definition. We therefore clarified the interpretation of 'exact match' for ARQMath-2 annotations in 2021 to take the formula semantics and context directly into account even in the case of identical formulae (so for example, variables of different types would not be considered the same, even if variable names are identical). This change may affect the utility of some ARQMath-1 relevance judgments for training systems that will be evaluated using ARQMath-2 (or subsequent ARQMath) relevance judgments.

As in 2020, we defined the relevance score for a formula to be the maximum relevance score for any judged instance of that formula. This relevance definition essentially asks "if instances of this formula were returned, would we reasonably expect some of those instances to be useful?"

Assessment System. We again used Turkle to build the assessment system for Task 2. As shown in Figure 4 (at the end of this document), there are two main panels. In the left panel, the question is shown as in Task 1, but now with the formula query highlighted in yellow. In the right panel, up to five retrieved posts (question posts or answer posts) containing instances of the same retrieved formula are displayed, with the retrieved formula instance highlighted in each case. For example, the formula $\sum_{n=1}^{\infty} a_n$ shown in Figure 4 was retrieved in two question posts. As in Task 1, buttons are provided for the assessor to record their judgment; unlike Task 1, judgments for each instance of the same retrieved formula (up to 5) are recorded separately, and later used to produce a *single* maximum score for each visually distinct formula.

Assessor training. Three assessors were assigned to to perform relevance judgements for Task 2, one of whom had also assessed Task 2 for ARQMath-1 in 2020. Three rounds of training were performed.

In the first training round, the assessors were familiarized with the task. To illustrate how for-

mula search might be used, we interactively demonstrated formula suggestion in MathDeck [19] and the formula search capability of Approach0 [13]. Then the task was defined using examples, showing a formula query with some retrieved results, talking through the relevance definitions and how to apply those definitions in specific cases. Two topics from ARQMath-1 (B.1, B.18) were selected as examples. During the training session, the assessors saw different example results for topics and discussed their relevance based on criteria defined for them with the organizers. These examples were manually selected from ARQMath-1 relevance judgments having different relevance degrees, and included examples from dual-assessed topics that 2020 assessors had disagreements on. The assessors also received feedback from the fourth author of this paper, an expert Math Stack Exchange user.

All three assessors were then assigned 7 other Task 2 topics from ARQMath-1 (B.29, B.32, B.33, B.41, B.59, B.62, B.70) to independently assess. The formulae to assess were chosen manually using the same process as the first training round. After assessment, the assessors and organizers met by Zoom to discuss and resolve disagreements. The assessors used this opportunity to refine their understanding of the relevance criteria, and the application of those criteria to specific cases. Assessor agreement was found to be fairly good (kappa=0.281 over four relevance levels and kappa=0.417 with H+M binary relevance). The assessors were then each assigned another 7 ARQMath-1 topics (B.8, B.69, B.83, B.89, B.92, B.95, B.98) and a third round of assessment practice followed by discussion was performed. The average kappa on the these topics was 0.467 over four relevance levels, and 0.565 for H+M binary relevance, agreement levels consistent with those observed at the end of Task 2 assessor training in 2020 [10]. Figure 3.(a) shows the Cohen's kappa coefficient values for each assessor in the last training round.

Assessment. A total of 60 topics were assessed for Task 2. Two queries (B.243 and B.266) had fewer than two relevant answers after H+M binarization and were removed. Of the remaining 58 queries, an average of 140.0 visually distinct formulae were assessed per topic, with an average assessment time of 39.5 seconds per formulae. The average number of formula instances labeled as relevant after H+M binarization was 30.9 per topic, with the highest being 107 for topic B.296 and the lowest being 3 for topics B.211 and B.255.

Post Assessment. After assessment for Task 2 was completed, each of the three assessors were assigned two topics, one of which had been assessed by each of the other two assessors. Figure 3 shows the Cohen's kappa coefficient values for each assessor. A kappa of 0.329 was achieved on the four-way assessment task, and with H+M binarization the average kappa value was 0.694.

5.5. Evaluation Measures

As for Task 1, the primary evaluation measure for Task 2 is nDCG′, and MAP′ and P′@10 were also computed. Participants submitted ranked lists of formula instances, but we computed these measures over visually distinct formulae. The ARQMath-2 Task 2 evaluation script replaces each formula instance with its associated visually distinct formula, and then deduplicates from the top of the list downward, producing a ranked list of visually distinct formulae, from which our prime evaluation measures are then computed using `trec_eval`.

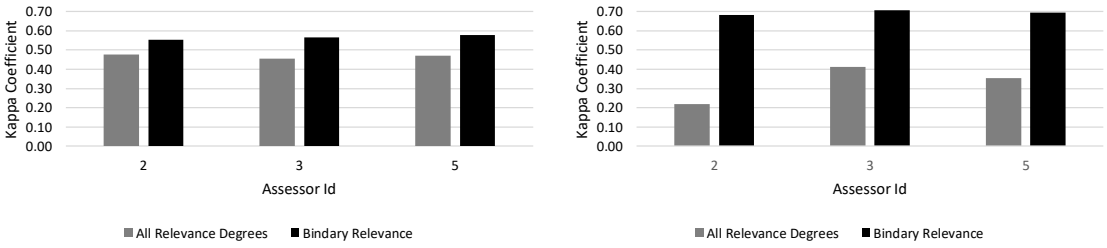

Figure 3: Inter-assessor agreement (Cohen's kappa) over 3 assessors. Chart (a) shows the agreement on the last training round (7 topics from ARQMath-1). Chart (b) shows the agreement after official Task 2 assessment. Each assessor evaluated two topics, one by each the other two assessors. Shown are four-way classification (gray) and two-way (H+M binarized) classification (black).

5.6. Results

Table 5 in the appendix shows the results, with the baseline run shown first, and then teams and their systems ranked by nDCG′. For ARQMath 2 topics, we see that the best results by nDCG′ were achieved by the Approach0 team, with the MathDowsers team doing almost as well by that measure, and the XY-PHOC-DPRL team a close third. The order between the best runs from each of those three teams is the same when evaluated on ARQMath-2 topics using MAP′ and P′@10.

Comparing ARQMath-2 results from 2021 with the last year's (2020) ARQMath-1 results, we see that (as with Task 1) for Task 2 the performance relative to the baseline is substantially improved in 2021 over 2020. Specifically, in 2020 the team with the best nDCG′ (DPRL) was 15% below the Tangent-S baseline by that measure; in 2021 the team with the best nDCG′ (Approach0) outperformed the Tangent-S baseline by 13%, as measured by nDCG′.

We note, however, a substantial drop in MAP′ and NDCG′ between ARQMath-1 topics and ARQMath-2 topics for the Tangent-S baseline on Task 2. Interestingly, the P′@10 results are much less adversely affected than the rank-based metrics when moving from the ARQMath-1 to the ARQMath-2 Task 2 topics. It is possible that these differences might result from the different instructions that we gave relevance assessors in 2021 regarding assessment of identical formulae, since it seems that the 2021 Task 2 assessors did indeed take the question context and formula semantics more directly into account when performing assessment for Task 2 of ARQMath-2 than did the ARQMath-1 assessors in 2020.

Tangent-S linearly combines six similarity scores computed from OPTs and SLTs, with weights fit using the NTCIR-12 math formula search topics. NTCIR-12 assessors had considered both visual and operator syntax similarity when comparing formulas retrieved in isolation, but without checking the types and domains for variables and operators associated with query and candidate formulas. This approach had transferred reasonably well when evaluated using ARQMath-1 formula queries, but the additional variation in the way identical formulae were assessed in ARQMath-2 may have made the problem more challenging for Tangent-S. In a quick experiment with retuning Tangent-S, we found that simply using the absolute value of

the combination weights (so that no similarity scores acted as penalties) increased nDCG' to 0.536. Without reading too much into that one simple experiment, we can at least say that the Tangent-S baseline that we report does not seem to be optimally tuned to the ARQMath-2 collection.

6. Conclusion

This second year of ARQMath resulted in an improved test collection, more participation, and better results. We anticipate continuing ARQMath for a third year, with participants in ARQMath benefiting from a mature evaluation infrastructure, a larger (and perhaps now also somewhat better) set of relevance judgements on which to train, and a larger and more diverse community of researchers with whom to share ideas.

Acknowledgments

We thank our student assessors from RIT and St. John Fisher College: Josh Anglum, Dominick Banasick, Aubrey Marcsisin, Nathalie Petruzelli, Siegfried Porterfield, Chase Shuster, and Freddy Stock. This material is based upon work supported by the National Science Foundation (USA) under Grant No. IIS-1717997 and the Alfred P. Sloan Foundation under Grant No. G-2017-9827.

References

[1] B. Mansouri, A. Agarwal, D. Oard, R. Zanibbi, Finding old answers to new math questions:the ARQMath lab at CLEF 2020, in: European Conference on Information Retrieval, 2020.

[2] C. Kaliszyk, E. C. Brady, A. Kohlhase, C. S. Coen (Eds.), Intelligent Computer Mathematics - 12th International Conference, CICM 2019, Prague, Czech Republic, July 8-12, 2019, Proceedings, volume 11617 of *Lecture Notes in Computer Science*, Springer, 2019.

[3] F. Guidi, C. S. Coen, A survey on retrieval of mathematical knowledge, in: CICM, volume 9150 of *Lecture Notes in Computer Science*, Springer, 2015, pp. 296–315.

[4] R. Zanibbi, D. Blostein, Recognition and retrieval of mathematical expressions, International Journal on Document Analysis and Recognition (IJDAR) 15 (2012) 331–357.

[5] A. Aizawa, M. Kohlhase, I. Ounis, NTCIR-10 math pilot task overview., in: NTCIR, 2013.

[6] A. Aizawa, M. Kohlhase, I. Ounis, M. Schubotz, NTCIR-11 Math-2 task overview., in: NTCIR, volume 11, 2014, pp. 88–98.

[7] R. Zanibbi, A. Aizawa, M. Kohlhase, I. Ounis, G. Topic, K. Davila, NTCIR-12 MathIR task overview, in: NTCIR, 2016.

[8] M. Schubotz, A. Youssef, V. Markl, H. S. Cohl, Challenges of mathematical information retrieval in the NTCIR-11 Math Wikipedia Task, in: SIGIR, ACM, 2015, pp. 951–954.

[9] M. Hopkins, R. Le Bras, C. Petrescu-Prahova, G. Stanovsky, H. Hajishirzi, R. Koncel-Kedziorski, SemEval-2019 Task 10: Math Question Answering, in: Proceedings of the 13th International Workshop on Semantic Evaluation, 2019.

[10] R. Zanibbi, D. W. Oard, A. Agarwal, B. Mansouri, Overview of ARQMath 2020: CLEF lab on answer retrieval for questions on math, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, H. Joho, C. Lioma, C. Eickhoff, A. Névéol, L. Cappellato, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction - 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22-25, 2020, Proceedings, volume 12260 of *Lecture Notes in Computer Science*, Springer, 2020, pp. 169–193. URL: https://doi.org/10.1007/978-3-030-58219-7_15. doi:10.1007/978-3-030-58219-7_15.

[11] K. Davila, R. Zanibbi, 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, 2017, pp. 1165–1168.

[12] B. Mansouri, S. Rohatgi, D. W. Oard, J. Wu, C. L. Giles, R. Zanibbi, Tangent-CFT: An embedding model for mathematical formulas, in: Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR), 2019, pp. 11–18.

[13] W. Zhong, R. Zanibbi, Structural similarity search for formulas using leaf-root paths in operator subtrees, in: European Conference on Information Retrieval, Springer, 2019, pp. 116–129.

[14] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, D. Johnson, Terrier information retrieval platform, in: European Conference on Information Retrieval, Springer, 2005, pp. 517–519.

[15] B. Mansouri, D. W. Oard, R. Zanibbi, DPRL systems in the CLEF 2020 ARQMath lab, in: L. Cappellato, C. Eickhoff, N. Ferro, A. Névéol (Eds.), Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020, volume 2696 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020. URL: http://ceur-ws.org/Vol-2696/paper_223.pdf.

[16] T. Sakai, N. Kando, On information retrieval metrics designed for evaluation with incomplete relevance assessments, Information Retrieval 11 (2008) 447–470.

[17] C. Buckley, E. M. Voorhees, Retrieval evaluation with incomplete information, in: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2004, pp. 25–32.

[18] P. Borlund, The IIR evaluation model: a framework for evaluation of interactive information retrieval systems, Information Research 8 (2003) 8–3.

[19] G. Nishizawa, J. Liu, Y. Diaz, A. Dmello, W. Zhong, R. Zanibbi, MathSeer: A math-aware search interface with intuitive formula editing, reuse, and lookup, in: European Conference on Information Retrieval, Springer, 2020, pp. 470–475.

Table 4
ARQMath-2 Task 1 (CQA) results. **P** indicates a primary run, **M** indicates a manual run, and **(✓)** indicates a baseline pooled at the primary run depth. For Precision′@10 and MAP′, H+M binarization was used. The best baseline results are in parentheses.

		Run Type		ARQMath-1 77 Topics			ARQMath-2 71 Topics		
Run	Data	P	M	nDCG′	MAP′	P′@10	nDCG′	MAP′	P′@10
Baselines									
Linked MSE posts	n/a	(✓)		(0.279)	(0.194)	(**0.386**)	0.203	0.120	**(0.282)**
TF-IDF + Tangent-S	Both	(✓)		0.248	0.047	0.073	0.201	0.045	0.086
TF-IDF	Both			0.204	0.049	0.074	0.185	0.046	0.063
Tangent-S	Math			0.158	0.033	0.051	0.111	0.027	0.052
MathDowsers									
primary	Both	✓		**0.433**	0.191	0.249	**0.434**	**0.169**	0.211
proximityReRank	Both			0.373	0.117	0.131	0.335	0.081	0.049
DPRL									
QASim	Both			0.417	0.234	0.369	0.388	0.147	0.193
RRF	Both	✓		0.422	**0.247**	**0.386**	0.347	0.101	0.132
Math Stack Exchange	Both			0.409	0.232	0.322	0.323	0.083	0.078
TU_DBS									
TU_DBS_P	Both	✓		0.380	0.198	0.316	0.377	0.158	**0.227**
TU_DBS_A2	Both			0.356	0.173	0.291	0.367	0.147	0.217
TU_DBS_A3	Both			0.359	0.173	0.299	0.357	0.141	0.194
TU_DBS_A1	Both			0.362	0.178	0.304	0.353	0.132	0.180
TU_DBS_A4	Both			0.045	0.016	0.071	0.028	0.004	0.009
Approach0									
B60	Both		✓	0.364	0.173	0.256	0.351	0.137	0.189
B60RM3	Both		✓	0.360	0.168	0.252	0.349	0.137	0.192
B55	Both	✓	✓	0.364	0.173	0.251	0.344	0.135	0.180
A55	Both		✓	0.364	0.171	0.256	0.343	0.134	0.194
P50	Both		✓	0.361	0.171	0.255	0.327	0.122	0.155
MIRMU									
WIBC	Both			0.381	0.135	0.161	0.332	0.087	0.106
RBC	Both	✓		0.392	0.153	0.220	0.322	0.088	0.132
IBC	Both			0.338	0.114	0.153	0.286	0.073	0.117
CompuBERT	Both			0.304	0.114	0.207	0.262	0.083	0.135
SCM	Both			0.324	0.119	0.156	0.250	0.059	0.072
MSM									
MG	Both	✓		0.310	0.114	0.170	0.278	0.077	0.127
PZ	Both			0.336	0.126	0.181	0.275	0.085	0.124
MP	Both			0.203	0.059	0.094	0.154	0.036	0.047
MH	Both			0.184	0.057	0.108	0.131	0.028	0.037
LM	Both			0.178	0.058	0.107	0.128	0.029	0.048
PSU									
PSU	Both	✓		0.317	0.116	0.165	0.242	0.065	0.110
GoogolFuel									
2020S41R71	Both	✓		0.292	0.086	0.153	0.203	0.050	0.092
2020S41R81	Both			0.290	0.085	0.153	0.203	0.050	0.089
2020S41R91	Both			0.289	0.084	0.157	0.203	0.050	0.089
2020S51R71	Both			0.288	0.082	0.140	0.202	0.049	0.089
2020S41	Both			0.281	0.076	0.135	0.201	0.048	0.080
BetterThanG									
Combiner1vs1	Both	✓	✓	0.233	0.046	0.073	0.157	0.031	0.051
Combiner2vs1	Both		✓	0.229	0.044	0.069	0.153	0.030	0.054
CombinerNorm	Both		✓	0.215	0.045	0.073	0.141	0.026	0.042
LuceneBM25	Text			0.179	0.052	0.079	0.119	0.025	0.032
Tangent-S	Math			0.158	0.033	0.051	0.110	0.026	0.061

Table 5
ARQMath-2 Task 2 (Formula Retrieval) results, computed over visually distinct formulae. **P** indicates a primary run, and **(✓)** shows the baseline pooled at the primary run depth. For MAP′ and P′@10, relevance was thresholded H+M binarization. All runs were automatic. Baseline results are in parentheses.

				ARQMath-1 45 Topics			ARQMath-2 58 Topics		
		Run Type							
Run	Data	P	M	nDCG′	MAP′	P′@10	nDCG′	MAP′	P′@10
Baseline									
Tangent-S	Math	(✓)		(0.692)	(0.446)	(0.453)	(0.492)	(0.272)	(0.419)
Approach0									
P300	Math		✓	0.507	0.342	0.441	**0.555**	**0.361**	**0.488**
B	Math		✓	0.493	0.340	0.425	0.519	0.336	0.461
B30	Math		✓	0.527	0.358	0.446	0.516	0.295	0.393
C30	Math		✓	0.527	0.358	0.446	0.516	0.295	0.393
P30	Math	✓	✓	0.527	0.358	0.446	0.505	0.284	0.371
MathDowsers									
formulaBase	Both	✓		0.562	0.370	0.447	0.552	0.333	0.450
docBase	Both			0.404	0.251	0.386	0.433	0.257	0.359
XY-PHOC-DPRL									
XY-PHOC	Math	✓		0.611	0.423	0.478	0.548	0.323	0.433
DPRL									
ltr29	Math			0.736	0.522	0.520	0.454	0.221	0.317
ltrall	Math	✓		**0.738**	**0.525**	**0.542**	0.445	0.216	0.333
TangentCFT2-TED	Math			0.648	0.480	0.502	0.410	0.253	0.464
TangentCFT-2	Math			0.607	0.437	0.480	0.338	0.188	0.297
TU_DBS									
TU_DBS_A3	Math			0.426	0.298	0.386	-	-	-
TU_DBS_A1	Math			0.396	0.271	0.391	-	-	-
TU_DBS_A2	Math			0.157	0.085	0.122	0.154	0.071	0.217
TU_DBS_P	Both	✓		0.152	0.080	0.122	0.153	0.069	0.216
NLP_NITS									
FormulaEmbedding_P	Math	✓		0.233	0.140	0.271	0.161	0.059	0.197
FormulaEmbedding_A	Math			-	-	-	0.114	0.039	0.152
Baseline	Math			-	-	-	0.091	0.032	0.151

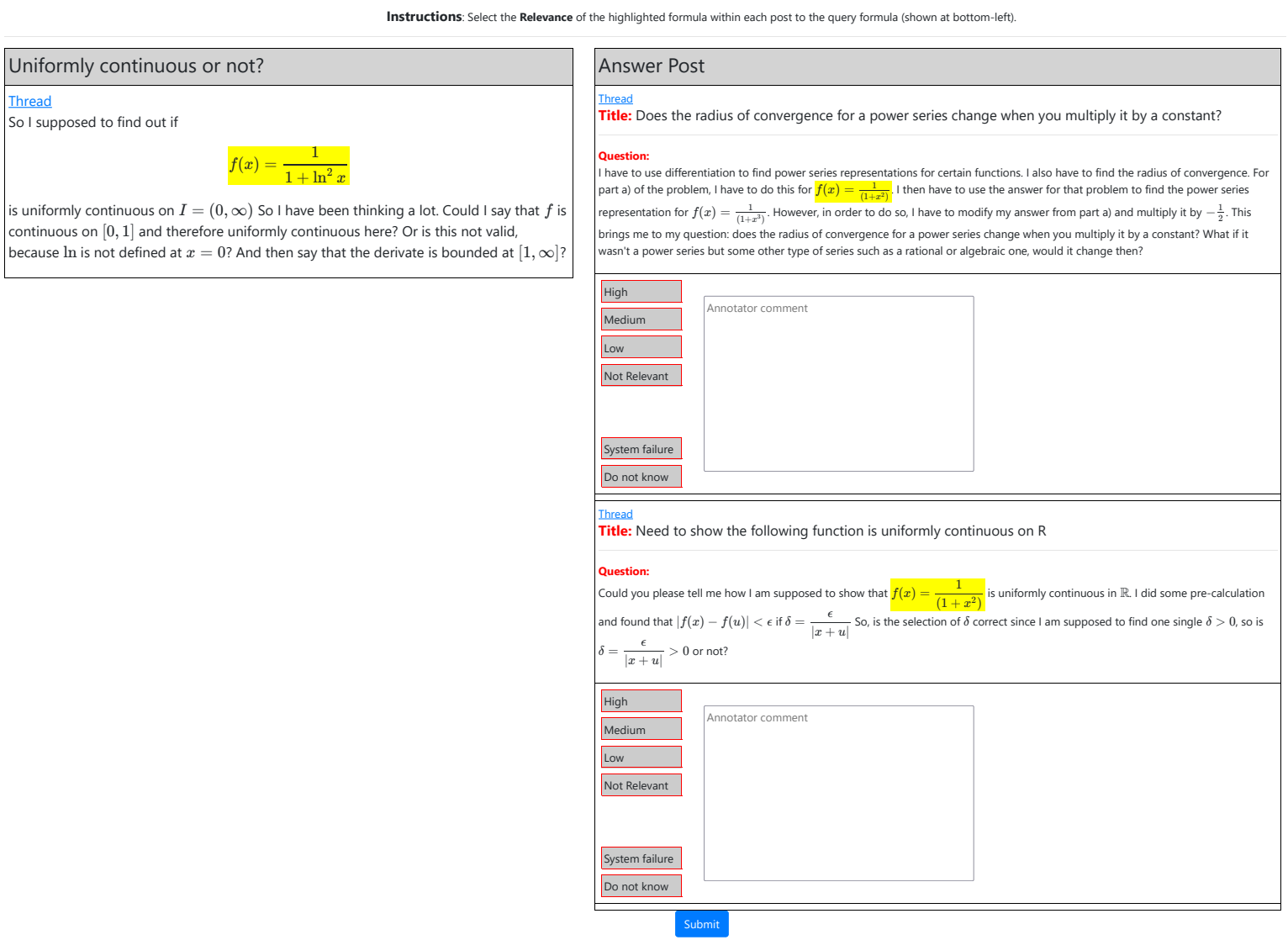

Figure 4: Turkle Assessment Interface. Shown are hits for Formula Retrieval (Task 2). In the left panel, the formula query is highlighted. In the right panel, two question posts containing the same retrieved formula are shown. For Task 1, a similar interface was used, but without formula highlighting, and just one returned answer post viewed at a time.