Advancing Math-Aware Search: The ARQMath-2 Lab at CLEF 2021

Behrooz Mansouri[1(✉)], Anurag Agarwal[1], Douglas W. 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. ARQMath-2 is a continuation of the ARQMath Lab at CLEF 2020, with two main tasks: (1) finding answers to mathematical questions among posted answers on a community question answering site (Math Stack Exchange), and (2) formula retrieval, where formulae in question posts serve as queries for formulae in earlier question and answer posts; the relevance of retrieved formulae considers the context of the posts in which query and retrieved formulae appear. The 2020 Lab created a large new test collection and established strong baselines for both tasks. Plans for ARQMath-2 includes extending the same test collection with additional topics, provision of standard components for optional use by teams new to the task, and post-hoc evaluation scripts to support tuning of new systems that did not contribute to the 2020 judgment pools.

Keywords: Community Question Answering · Formula retrieval · Mathematical Information Retrieval · Math-aware search

1 Introduction

The ARQMath lab [15] was established to support research on search using mathematical notation. With a number of Math Information Retrieval (MIR) systems having been introduced recently [4,7,10,13,16], a standard MIR benchmark is essential for understanding the behavior of their retrieval models and implementations. To that end, the first ARQMath produced a new collection, assessment protocols, parsing and evaluation tools, and a benchmark containing over 70 annotated topics for each of two tasks: math question answer retrieval, and formula retrieval.[1]

Effective question answering systems for math would be highly valuable for both 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[2] (MSE) and Math Overflow [12] are widely-used resources.

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

D. Hiemstra et al. (Eds.): ECIR 2021, LNCS 12657, pp. 631–638, 2021.
https://doi.org/10.1007/978-3-030-72240-1_74

This indicates that there is great interest in finding answers to mathematical questions posed in natural language, using *both* text and mathematical notation. Moreover, a recent study found that retrieval effectiveness for mathematical queries submitted to a general-purpose search engine was much lower than for other queries [6].

ARQMath is the first shared-task evaluation of 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. Table 1 (left column) shows an example topic from Task 1, showing one answer assessed as relevant, and another assessed as non-relevant. The goal of Task 2 in ARQMath is retrieval of *visually distinct* formulae in decreasing relevance order, where the relevance of a visually distinct formula is the highest relevance of any assessed instance of that formula when judged in context. This task is illustrated in the right column of Table 1.

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 [1], NTCIR-11 [2] and NTCIR-12 [14]). The Mathematical Information Retrieval (MathIR) 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 based on naturally occurring questions, by assessing formula relevance in context, and by substantially increasing the number of topics.

ARQMath-2 will re-use the ARQMath 2020 collection, which consists of MSE posts from 2010 to 2018. ARQMath-1 topics disproportionately sampled commonly asked questions; in ARQMath-2 we plan to better balance topic development to include a greater range of novel questions. To facilitate participation of new teams we will provide some standard components (e.g., for computing formula similarity) that can easily be integrated with existing systems for ranked retrieval. ARQMath scoring in 2020 was designed for systems that had contributed to the judgment pools, but we are reworking the evaluation scripts to generate comparable scores for unjudged runs to support training and tuning learning to rank systems. We summarize the existing data and tools, the first edition of the ARQMath task, and planned changes for ARQMath-2 in the remainder of the paper.

2 The ARQMath Test Collection

The collection to be searched is comprised of question and answer posts from Math Stack Exchange (MSE). These postings are freely available as data dumps from the Internet Archive. The collection contains posts published from 2010 to 2018, a total of 1 million questions and 1.4 million answers. In ARQMath-1, posts from 2019 were used as a basis for topic construction. For ARQMath-2, posts from 2020 will be used for that purpose. The first criterion for selecting

Table 1. Example ARQMath queries and results.

Question answering (task 1)	**Formula retrieval (task 2)**
QUESTION (TOPIC A.4) I have the sum $$\sum_{k=0}^{n} \binom{n}{k} k$$ I know the result is $n2^{n-1}$ but I don't know how you get there. How does one even begin to simplify a sum like this that has binomial coefficients.	FORMULA QUERY (TOPIC B.4) $$\sum_{k=0}^{n} \binom{n}{k} k$$
RELEVANT (✓) You have to take the derivative of $$\sum_{i=0}^{n} \binom{n}{k} x^k = (1+x)^n$$ and then set x=1 in $$\sum_{i=0}^{n} k\binom{n}{k} x^{k-1} = n(1+x)^{n-1}$$	RELEVANT (✓) ... which can be obtained by manipulating the second derivative of $$\sum_{k=0}^{n} \binom{n}{k} z^k$$ and let $z = p/(1-p)$ ...
NON-RELEVANT (X) By your example, it seems that you're computing all the combinations of k elements of a set X having n elements. Intuitively, you wrote all possible strings, without considering the order (i.e. ab=ba as string) with the elements of X. Observe also that $\sum_{k=0}^{n} \binom{n}{k} = 2^n$, i.e. all the possible subsets of X.	NON-RELEVANT (X) Yes, it is in fact possible to sum this. The answer is $$\sum_{k=0}^{n} \binom{n}{k}\binom{m}{k} = \binom{m+n}{n}$$ assuming that $n \leq m$. This comes from the fact that ...

a topic is that the question contains at least one formula; with that constraint, nearly 240K questions are available for ARQMath-2 topic development.

Formulae. In the Internet Archive version of the collection, formulae are located 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 represented in three ways to facilitate participation by teams without specialized expertise in mathematical notation processing: (a) as LaTeX strings, (b) as (appearance-based) Presentation MathML, and (c) as (operator tree) Content MathML.

The open source LaTeXML[3] tool used for converting LaTeX to MathML fails on some MSE formulae. Moreover, producing Content MathML from LaTeX requires inference, and is thus potentially errorful. As a result, the coverage

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

of Presentation MathML for detected formulae in the ARQMath-1 collection was 92%, and the coverage for Content MathML was 90%. For ARQMath-2 we reduced the error rate to less than a percent for both representations, thus reducing the need for participating systems to fall back to processing the LaTeX string.

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 additional information such as creation date and creator (see Users below). Question posts contain both a title and a body (with the body holding the question itself) 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 provides information about positive or negative reactions to a post. Interestingly, no participating team in ARQMath-1 found this information to be helpful in their ranking algorithm.
- **Users.** Each poster or a question or an answer has a unique User ID and a reputation score.

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

3 Previous ARQMath Edition

The ARQMath-1 lab was part of the 2020 Conference and Labs of the Evaluation Forum (CLEF) [5,15].

3.1 Finding Answers to Math Questions

The primary task for ARQMath 2020 was answer retrieval, in which participants were presented with a question that had actually been asked on MSE in 2019, and were asked to return a ranked list of up to 1,000 answers from prior years (2010–2018). Participating teams ranked answer posts for 100 topics, 74 of which were assessed and used for the evaluation of participating systems. System results ('runs') were evaluated using the nDCG′ measure (read as "nDCG-prime") introduced by Sakai and Kando [11] as the primary measure for the task. This measure is simply Normalized Discounted Cumulative Gain (nDCG), but with unjudged documents removed before scoring. Table 2 summarizes the graded relevance scale used for assessment. Two additional measures, mAP′ and P@10, were also reported using binarized relevance judgments.

Five teams participated in ARQMath-1 task 1. Teams submitted up to 5 runs, with at least one designated as primary. For each primary run, for 5 additional organizer-provided baseline runs, and for any manual runs among those not designated as primary, the pooling depth was set to 50. A pool depth of 20 was used for other runs. The highest nDCG′ value (0.345) was achieved by the MathDowsers [9] team, while the highest mAP′ and P@10 was achieved by an oracle baseline built using links to related posts in the MSE collection (which were not available to the participating teams).

3.2 Formula Search

Formula search was run as an experimental task in ARQMath-1. The intent of the formula search task was similar to the Wikipedia Formula Browsing Task from NTCIR-12 [14], but with two novel innovations. First, relevance is defined differently: in NTCIR-12, formula queries were compared by assessors with retrieved formula instances, in isolation (i.e., the relevance of a retrieved formula was judged without access to the context in which that formula was found). In ARQMath, by contrast, both the formula query and a retrieved formula instance were presented to the assessor in context (in the question post and in an answer post, respectively). Second, in NTCIR-12 systems could receive credit for finding formula instances, whereas in ARQMath systems received 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 (or be penalized twice), whereas an ARQMath system would receive credit (or be penalized) only once for each visually distinct formula that was retrieved. We implemented this by deduplicating submitted ranked lists based on the linearized Symbol Layout Trees produced from Presentation MathML by Tangent-S [3] where possible, and by comparing LaTeX strings otherwise.

Notably, the NTCIR-12 formula browsing task test collection had only 20 formula queries (plus 20 modified versions of the same formulae with wildcards added), whereas ARQMath-1 generated relevance judgments for 74 queries (45 of which were used for evaluation, with both those and the remaining 29 available for training future systems).

Table 2 also summarizes the graded relevance scale used for assessment. In this case, however, assessors are asked to assess formula instances, drawing upon the context provided by the question post from which the formula query was selected and a specific answer post in which the formula was found. The relevance of a *visually distinct* formula is then computed as the maximum over all assessed instances of that visually distinct formula. For efficiency reasons, we limit the number of instances of any visually distinct formula that were assessed to 5.

Four teams participated in ARQMath-1 Task 2, with submission and pooling protocols similar to those for Task 1. The single baseline system provided by the organizers (Tangent-S [3]) achieved the highest nDCG′ value, while the DPRL team [8] obtained the highest mAP′ and P@10 scores.

4 Changes for ARQMath-2

ARQMath-2 will include the same two tasks as ARQMath-1, with formula retrieval (Task 2) being promoted from an experimental task to a full task now that the evaluation details have been fully worked out.

For ARQMath-1 we restricted our selection of question posts for topic construction to those with at least one related post link to a question in the collection to be searched.[4] We did this to minimize the risk of investing assessment effort on topics that yielded no relevant documents. For ARQMath-2 we plan to remove this restriction, and instead guard against wasted assessment effort by doing a limited amount of pre-assessment for the results of an ARQMath-1 baseline system.

The scoring scripts for ARQMath-1 were designed to score participating systems, but to support training of ARQMath-2 systems we need to change the order of some of the processing. Rather than deduplicating by clustering submitted runs, we will instead cluster all formula instances in the collection, and then score every run using that single clustering. This will permit accurate *post hoc* assessment. We also plan to extend the number of submitted formula instances beyond 1000 so that adequately deep lists of visually distinct formulae will remain after deduplication. As with all of the tools and collections used in the lab, the new Task 2 scoring script will be available on the AQRMath GitHub page.[5]

[4] These links were not available to participants, although they were used to construct the oracle baseline system.

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

5 Conclusion

The ARQMath-2 lab at CLEF 2021 will be the second in what we plan to be a three-year series of labs aiming to advance the state-of-the-art for math-aware IR. As in the first edition, we have chosen to focus on answer retrieval for math questions as the first task, and formula search for the second. The same Math Stack Exchange collection will be used, both because the first task models an actual employment scenario, and because we expect that the continuity provided by that consistency 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.

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. Davila, K., Zanibbi, R.: 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)
4. Fraser, D., Kane, A., Tompa, F.W.: Choosing math features for BM25 ranking with Tangent-L. In: Proceedings of the ACM Symposium on Document Engineering (2018)
5. Mansouri, B., Agarwal, A., Oard, D., Zanibbi, R.: Finding old answers to new math questions: the ARQMath lab at CLEF 2020. In: Jose, J.M., et al. (eds.) ECIR 2020. LNCS, vol. 12036, pp. 564–571. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-45442-5_73
6. Mansouri, B., Zanibbi, R., Oard, D.W.: Characterizing searches for mathematical concepts. IEEE (2019)
7. Mansouri, B., Rohatgi, S., Oard, D.W., Wu, J., Giles, C.L., Zanibbi, R.: Tangent-CFT: an embedding model for mathematical formulas. In: Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval (2019)
8. Mansouri, B., Oard, D.W., Zanibbi, R.: DPRL systems in the CLEF 2020 ARQMath lab. In: International Conference of the Cross-Language Evaluation Forum for European Languages (2020)
9. Yin Ki, N.G., et al.: Dowsing for answers with Tangent-L. In: International Conference of the Cross-Language Evaluation Forum for European Languages (2020)
10. Pfahler, L., Morik, K.: Semantic search in millions of equations. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (2020)
11. Sakai, T., Kando, N.: On information retrieval metrics designed for evaluation with incomplete relevance assessments. Inf. Retr. **11**, 447–470 (2008)
12. Tausczik, Y.R., Kittur, A., Kraut, R.E.: Collaborative problem solving: a study of MathOverflow. In: CSCW (2014)

13. Yasunaga, M., Lafferty, J.D.: TopicEq: a joint topic and mathematical equation model for scientific texts. In: Proceedings of the AAAI Conference on Artificial Intelligence (2019)
14. Zanibbi, R., Aizawa, A., Kohlhase, M., Ounis, I., Goran, T., Davila, K.: NTCIR-12 MathIR task overview. In: NTCIR (2016)
15. Zanibbi, R., Oard, D.W., Agarwal, A., Mansouri, B.: Overview of ARQMath 2020: CLEF lab on answer retrieval for questions on math. CLEF 2020. LNCS, vol. 12260, pp. 169–193. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58219-7_15
16. Zhong, W., Rohatgi, S., Wu, J., Giles, C.L., Zanibbi, R.: Accelerating substructure similarity search for formula retrieval. In: Jose, J.M., et al. (eds.) ECIR 2020. LNCS, vol. 12035, pp. 714–727. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-45439-5_47