CLEF 2020 Lab Proposal for ARQMath: Answer Retrieval for Mathematical Questions

Coordinators: **Richard Zanibbi**[*] **(contact), Douglas Oard**[†] **and Anurag Agarwal**[‡]
Email: rxzvcs@rit.edu, oard@umd.edu, axasma@rit.edu

We propose a new lab, Answer Retrieval for Questions about Math (ARQMath). Using the mathematics and free text in posts from an online community question answering platform (Math StackExchange[1] (MSE)), participating systems are given a question, and must then return a ranked list of answers. Relevance is determined by how well the returned posts answer the provided question. We will also run a secondary task on formula retrieval (query-by-example), where relevance will be determined by the visual and semantic similarity between query and retrieved formulas.

Through this task we will explore leveraging math notation together with text to improve the quality of retrieval results for math-based queries. We call this **math-aware information retrieval** because our focus is on leveraging the ability to process mathematical notation to enhance, rather than to replace, other information retrieval techniques. We also hope to foster the development and comparison of math-aware search engines, and advance the semantic analysis of mathematical notation and texts.

There has been growing interest in answering mathematical questions from problem sets and tests within the Natural Language Processing (NLP) community, and more generally machine reading and comprehension (RC) and question answering (Q&A). This lab provides an opportunity to push mathematical question answering in a new direction, where informal language is frequently used, and where answers provided by a community are selected and ranked rather than generated. Further, this task would produce a test collection with a good chance of being widely used in future IR and NLP research.

Techniques developed are likely to advance multi-modal search engines for other domains that also frequently use specialized notation, such as biology and chemistry.

Formula Search. In general, searching for mathematical formulae is a non-trivial problem – especially if we want to be able to search for occurrences of parts of formulae.

1. For example, binomial coefficients come in variety of notations depending on the context: $\binom{n}{k}$, C^n_k, $C(n,k)$ or ${}^n_kC$, but they all mean the same thing. So it will be desirable to retrieve all forms, irrespective of the notation used.
2. Another challenge is that formulas may be polysemic (i.e., have multiple interpretations). For example, $\int f(x)dx$ represents a Riemann Integral, as well as Lebesgue Integral, as well as an anti-derivative of the function $f(x)$.
3. Suppose we get a question containing $a^2+b^2=c^2$. In mathematics we consider $a^2+b^2=c^2$ and $x^2+y^2=z^2$ to differ only in variable names. However a standard search engine may only return results containing $a^2+b^2=c^2$, ignoring relevant results containing $x^2+y^2=z^2$.

Effective math-aware search may enable new solutions to problems by exposing math solution methods, alternative mathematical models, or work-arounds to using a certain equation. It also stands to deepen our understanding of technical literature, by helping people identify new connections and patterns, analogous systems, and alternative perspectives (e.g., regarding loss functions used in Machine Learning). Math-aware search may even help find connections between seemingly unrelated fields; for example, a heart researcher might find the same equations describing cardiac electrical signals turning up in the work of astronomers studying solar flares, where a key technical problem has already been solved.

Further, if entities represented by formulas and text can be identified, then math can be used to find text and vice versa. For example, queries containing ‘Pythagorean theorem’ or variations of its formula could conceivably return the same passages about the theorem. These are things that existing search engines cannot do.

Related Labs/Tasks. Previously there were math retrieval tasks at NTCIR-10, NTCIR-11 and NTCIR-12, with the last held in 2016. The NTCIR tasks significantly advanced the state-of-the-art for systems

[*] Dept. Computer Science, Rochester Inst. Technology, 102 Lomb Memorial Drive, Rochester, NY 14623
[†] College of Information Studies (iSchool) & UMIACS, University of Maryland, 1109C Patuxent Building, College Park, MD 20742
[‡] School of Mathematical Sciences, Rochester Inst. Technology, 84 Lomb Memorial Drive Rochester, NY 14623
[1] https://math.stackexchange.com

and evaluation of text + math queries and isolated formula retrieval. NTCIR-12 made use of two corpora, one a set of arXiv papers from physics split into paragraph-sized documents, along with articles from English Wikipedia. Additionally, NTCIR-11 and NTCIR-12 provided two benchmarks for isolated formula retrieval (query-by-example). The NTCIR-12 formula collection was also later used for the CROHME 2016 handwritten math recognition competition at ICFHR 2016.[2] Some key differences between the NTCIR math tasks and our proposed lab are:

- *Realism.* The question answering task models an actual application in which the questions to be answered were generated by real users performing an actual task. In the NTCIR tasks, topics were manually constructed to explore the ability of systems to handle representative phenomena.
- *Reliability.* The interaction style is conversational, with the content to be searched containing both answers to questions and discussion of those answers. NTCIR's focus on published materials provided less scaffolding for the human relevance assessment process.
- *Reusability.* The answers to be found naturally occur as short passages, which avoids the complexity of reusing judgments for passages extracted differently from longer documents by different systems.
- *Scale.* While we will still check inter-annotator agreement on some topics, we will use single assessors in most cases in order to assess a sufficiently large number of topics to support statistical significance testing. The largest NTCIR-12 collection for text + math queries had only 30 topics (for the Wikipedia text+math query task).

A more distantly related task that also featured highly structured representations was a Chemical IR Track at TREC in 2011 that sought synergy between text-based retrieval and recognizing chemical diagram images in patents. There have also been community question answering tracks in English (TREC LiveQA in 2015, 2016, and 2017) and Japanese (NTCIR OpenLiveQ in 2017 and 2019) that, taken together, attracted participation from more than two dozen different research teams. Notably, they achieved this level of participation despite including a "live lab" component that made participation more challenging than for our more conventional test collection design. We take this as a sign of strong interest in the information retrieval community in community question answering as a research setting.

Tasks

There will be two main tasks in the first year. First, a question answering task (Q&A), where systems are provided a question post from MSE and then need to return a ranked list of answer posts. Second, an isolated formula retrieval task.

Main Task (Q&A). For the Q&A task, at least 100 questions from Math StackExchange will be sampled, with the requirement that each question contains both text and at least one formula. Participants will have the option to run queries using only the text or math portions in each question, or to use both math and text, and we will ask them to label each run with which of those conditions they chose. One challenge inherent in this design is that the expressive power of text and formulas are sometimes complementary; so although all topics will include both text and formula(s), some may be better suited to text-based or math-based retrieval. We plan to accommodate this by reporting results for all participants that are averaged over three topic sets: (1) all topics, (2) topics for which the assessor believes the text alone to be an adequate characterization of the topic, and (3) topics for which the assessor believes the formula(s) alone to be an adequate characterization of the topic.

Secondary Task (Formula retrieval). In this task individual formulas are used as queries, and systems must return a ranked list of similar formulas. As with the NTCIR-12 Wikipedia Formula Browsing Task, this task has the goal of fostering development of component technology for computing math similarity. We envision two improvements over what was done at NTCIR: further developing the concept of "formula relevance" and creating a collection with a larger number of formula queries (NTCIR-12 has only 20 formula queries + 20 simplified versions of the same formulas with wildcards added). Each formula query will be a single formula extracted from a question being used in the main task. For each such query, we will ask the annotator to write a short human-readable narrative field – not available to participating systems – that reflects their understanding of the type(s) of similarity the person who asked the original question would have found useful. This may include alternative notation, simplification, specialization, or applications in specific fields, and we expect to extend those categories further based on suggestions from participating teams. Because participating systems won't have access to this narrative field, we expect this task to support research on diversity ranking for formula retrieval. We are aiming to have at least 50 formula queries in the first year, with the intent to expand the collection in subsequent years.

[2] `http://www.cs.rit.edu/~rlaz/files/icfhr_crohme2016.pdf`

Corpus: Math StackExchange

Our collection will be comprised of question and answer postings from Math StackExchange. These postings are freely available, and we will use the publicly available data dumps on the Internet Archive[3] to produce our collection. At the time of this writing, there are 1.1 million questions on the site.

We plan to represent question and answer posts by their content and (for optional use by participating teams) by the content of comments that subsequently may have been made to a post, but not their metadata or other aspect of the way they are displayed on the MSE website. Questions and answers will be separated into independent documents, without representation of the ordering of the answers or votes/ratings for questions and responses. Answer ordering and and votes will be stored separately from their associated postings. During the task participants will not have access to this information, but we will make it available to assessors during evaluation.

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

Evaluation/Assessment

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

Because specialized mathematical and computational knowledge may be needed for assessment, the pooled documents will be assessed for relevance by volunteers from participating teams, augmented by assessors hired by the organizers using funds provided for this purpose by the NSF.[5] Evaluation will be performed using a web-based system (e.g., Sepia[6] was used for the MathIR task at NTCIR-12). Assessors for the main task will be asked to identify relevant answers using pools from the main task. Assessors for the formula retrieval task will work with merged pools from both the main task and the formula retrieval task to identify similar formulas. Most pools will be judged by a single assessor, but some will be dual-assessed to observe annotator agreement. For the formula retrieval task, queries will be selected for dual annotation using stratified random sampling so as to cover a broad range of similarity types.

Based on our previous experience at NTCIR-12, and the distribution of assessments across voluntary annotation by participants and hired individuals, we expect that assessment will take 4-6 weeks to complete for an estimated 200-400 hours of annotation effort for the main task, and roughly 75-100 hours of annotation effort for the formula retrieval task. Included in this estimate is some limited experimentation with alternative annotation strategies (e.g., additionally annotating the most useful parts of a relevant answer, or annotating the preference order between relevant answers) with the goal of informing evaluation design in future years.

We will use *trec_eval* to compute ranked document retrieval measures for each run, with infAP as the official measure used to rank systems for both tasks. We have chosen infAP with the goal of facilitating comparison to future systems that did not contribute to the judgment pools.

[3] https://archive.org/details/stackexchange

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

[5] https://nsf.gov/awardsearch/showAward?AWD_ID=1717997

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

Organization and Participants

We expect the first ARQMath lab to run for a full-day at CLEF 2020. The meeting will include an initial brief tutorial on math retrieval for the benefit of interested nonparticipants, presentations by the organizers and participating teams, break-out sessions focused on the main and secondary task, a panel discussion regarding the task design and results, and planning for the next lab. The lab organizers will set the schedule and run the tutorial, organizer talks, panel discussion, and planning sessions.

Participation: At the previous NTCIR-12 MathIR task, there were six participating groups from around the world, including researchers from Asia (Japan, China, and India), North America (USA and Canada), and Europe (Germany and the Czech Republic). We plan to reach out to both the IR and NLP communities for ARQMath, and anticipate more participation as a result. Below we list groups who have expressed an interest in ARQMath, along with friends willing to spread the word, and others we know have recently published work related to math-aware search.

Interested Groups: National Institute of Informatics (Japan): Akiko Aizawa; Masaryk University (Czech Republic): Petr Sojka; Old Dominion University, (USA): Jian Wu; Peking University (China): Liangcai Gao; Penn State (USA): Lee Giles, Shaurya Rohatgi; Rochester Institute of Technology (USA): Behrooz Mansouri, Wei Zhong; University at Buffalo (USA): Kenny Davila; University of Glasgow (Scotland): Iadh Ounis; University of Wuppertal (Germany): Moritz Schubotz; University of Waterloo (Canada): Frank Tompa. **Willing to spread the word:** Trinity College Dublin (Ireland): Joeran Beel. **Others working in math-aware search:** Yale University (USA): John Lafferty, Michihiro Yasunaga; Columbia University (USA): David Blei; University of Konstanz (Germany): Bela Gipp.

Draft Schedule

Please note: this is a preliminary schedule, which we expect will change.

Date	Description
Aug. 19, 2019	Start creation of posting data set
Sept. 9-12, 2019	Task presentation and organizational meeting/discussion at CLEF 2019 (D. Oard)
Oct. 15, 2019	Short lab description for ECIR 2020
Nov. 1, 2019	Release of posting data set and sample queries
Nov. 5, 2019	CLEF 2020 Lab registration opens
Jan. 15, 2020	Release of test query set
Apr. 14-17, 2020	Report at ECIR 2020
May 1, 2020	Participant submissions close
July 1, 2020	Assessments for main and secondary tasks complete
August 14, 2020	Final lab report complete
Sept. 22-25, 2020	ARQMath Lab at CLEF 2020

Organizers

Richard Zanibbi is a Professor of Computer Science and Director of the Document and Pattern Recognition Lab at the Rochester Institute of Technology (RIT). He has published extensively on the recognition and retrieval of mathematical notation, and received the Best Applications Paper Award with Wei Zhong at ECIR 2019. He was also an organizer for the MathIR task at NTCIR-12.

Douglas W. Oard is a Professor in the iSchool and UMIACS at the University of Maryland, College Park. He has co-organized the TREC Arabic CLIR track, the TREC Legal track, the CLEF Interactive CLIR task, the CLEF Cross-Language Speech Retrieval task, and the FIRE Spoken Web task. He serves on the TREC program committee, and he has served as a general co-chair for NTCIR.

Anurag Agarwal is an Associate Professor of Mathematics at RIT. His research interests lie in Algebraic Number Theory and Mathematical Cryptography. He has collaborated on search tasks related to keyword spotting and improving accuracy of relevance assessment for math search. He contributed to query design for the NTCIR-12 Math IR Task, and is rated as an ‘expert’ poster in the Math StackExchange system.