ARQMath-3@CLEF 2022

Answer Retrieval for Questions on Math

Participant Guidelines (v. 5)

Feb 1st, 2022

Task coordinators: Behrooz Mansouri, Vítek Novotný, Anurag Agarwal, Douglas W. Oard, and Richard Zanibbi.

Changes (from Version 4)

- Changes to runs and output format for new Task 3 (Open Domain QA)
- Minor wording changes

Overview

This document provides guidelines for participation in the ARQMath 2020 shared task at CLEF 2022, the third edition of the task. The **schedule and deadlines** for the task may be found on the ARQMath web page, including any updates.

Registration. Registration for the task should be done online through CLEF 2020, at http://clef2022-labs-registration.dei.unipd.it. Registration closes on April 22, 2022, which is a **firm** deadline for all CLEF 2022 tasks.

Communication. Communication between organizers and participants is being arranged through Google Groups, in the ARQMath forum. Please visit the forum link to request access and join in discussions about the task. Additional information about the task is available on the ARQMath web page.

Tasks

There are three (3) tasks in ARQMath-3.

- **Task 1: Answer Retrieval.** In the first task, a question post is provided as a query, and participant systems search a collection of question and answer posts for answers to the question.
- **Task 2: Formula Retrieval.** In the second task, a formula within a post acts as the query, and formulas in posts from the collection are returned.
- **(New) Task 3: Open Domain Question Answering.** In the new Open Domain Question Answering pilot task for ARQMath-3, systems return a single answer for a question, which may be an ARQMath post, posts or excerpts from outside ARQMath, or a generated answer (e.g., using a neural net).

Collection and Topics

The collection and topics (once available) are posted on the ARQMath Google Drive directory. The collection is adapted from the community question-answering forum Math Stack Exchange, whose data is provided free for non-commercial use as snapshots on the Internet Archive. The snapshot used is from mid-2019. The main collection consists of question and answer posts from 2010-2018. **Additional details about the collection can be found in the README files provided with the collection.**

For ARQMath-3, the question posts used for topics in Tasks 1 and 3 are taken from 2021. Formula queries for Task 2 are selected from question posts used for topics in Task 1.

TSV Run File Formats

Details of the **submission format for Tasks 1 and 2** may be found in the evaluation protocols document in the ARQMath Google Drive. The evaluation protocol for Tasks 1 and 2 may be found in the ARQMath-2 (2021) overview paper.

Valid Hits: Note carefully the following constraints for Tasks 1, 2, and 3. **Hits that do not adhere to these formatting constraints will be filtered during assessment.**

Task 1: The *Post_id* field for each hit must be an **answer post** (i.e., it cannot be a question post, a comment, or any other element type).

Task 2: The *Post_id* field for each *Formula_Id* in the results must belong to a **question post *or* answer post** (and not a comment or other element type). The provided formula index files contain the *Post_id* and post type associated with each formula in the collection.

Task 3 (New for ARQMath-3): Answers are unicode (UTF-8) strings containing between 0 and 1200 unicode characters, which may contain a combination of plain text and LaTeX formulas (demarcated with single (`$`) or double (`$$`) dollar signs). **Answers not adhering to these format and length restrictions will be removed from the evaluation pool.** Participants may visually check answer strings interactively using online tools such as this one or the Math Stack Exchange ask tool, or simply generate HTML documents with MathJax support.

Because only a single answer is generated per question, and because answers may be selected or automatically generated for this task, the submission TSV (Tab Separated Variable) format is slightly different, with four fields:

```
(1) Query_Id (2) Rank (3) Score (4) Run_Id (5) Sources (6) Answer
```

1. `Query_Id` is the topic identifier.
2. `Rank` is the rank (for this task, this should be 1 in all cases, as only one answer is returned)
3. `Score` is a user generated score for the returned answer (this is not used in evaluation).
4. `Run_Id` is a participant-chosen name for the run.
5. `Sources` is an optional annotation string to describe source(s) used in forming the answer.
6. The `Answer` field contains a **plain text (unicode, UTF-8) string that may include demarcated LaTeX formulas** that begin and end with single ($) or double ($$) dollar signs). Answer strings must be between 0 and 1200 unicode characters in length.

Here is an example run file for Task 3:

```
    A.1    1 0.95    Run_0    "Quora"     "This can be solved by..."
    A.2    1 0.90    Run_0    "Generated" "Consider $x$ ..."
    ...
```

Naming and Submitting Run Files

Manual and automatic runs are permitted for all tasks. In a manual run, results may be produced by any means, including means that involve manual intervention for query formulation, result selection, or any other purpose. Automatic runs must be produced algorithmically by computer, without user intervention.

Participants should provide runs in tab-separated variable (TSV) using the formats described above, and name their run files using the following convention:

```
[group]-[task]-[id]-[run-type]-[data-used]-[*ans-type]-[eval].tsv

    * [group]: CLEF-registered team name.
    * [task]: task1 / task2 / task3
    * [id]: identifier (e.g., identify algorithms and parameters)
    * [run-type]: manual / auto
            (manual or automatic)
    * [data-used]: text / math / both
            (text-only, formula-only, formula and text)
    * [*ans-type] extract / generate
            (extracted passages only vs. using generated text)
            **Task 3 only**
    * [eval]: P / A
              (primary run, alternate run)

Examples:  TeamX-task1-bertA-auto-both-P.tsv
           TeamX-task1-bm25v5-auto-text-A.tsv
           TeamA-task2-browser_search-manual-both-P.tsv

           TeamB-task3-gpt3-auto-both-generate-P.tsv
           TeamB-task3-arqmath-auto-both-extract-A.tsv
           TeamZ-task3-mantest-manual-both-generate-A.tsv
```

Important Notes:

- **Task 3 File Name Examples:** for the Task 3 example file names above, the first run generates answers, the second run returns text from the first answers retrieved in a Task 1 run, and the 3rd Task 3 run is a manual run.

- **Use of Formula Identifiers in Task 2.** Participants should index formulas using the provided formula identifiers in the ARQMath corpus. These will be used to evaluate formulas in-context (i.e., within their associated post and thread).

- **Task 3: 'Extractive' vs. 'Generative' Runs.** An 'extractive' run returns text excerpts from sources (i.e., the text is taken from sources), while a 'generative' run creates text (e.g., using a recurrent neural network). If a systems combines both approaches, it should be labelled as a 'generative' run.

Submitting Runs. Runs should be uploaded to the Google Drive directory that will be provided for participants, using the appropriate naming of run files as given above.

Up to a **maximum of 5 runs** may be submitted for evaluation by each group for all three Tasks. **We guarantee that the run identified as 'primary' will contribute to the judgment pools,** with the inclusion of alternate runs in the judgment pools depending upon the number of submissions. Participants should therefore make sure to submit a *single* 'best' run as primary (see file naming convention above), to insure that they are included in the evaluation.